Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion and Blind Deconvolution

Cong Ma* Kaizheng Wang* Yuejie Chi† Yuxin Chen‡

November 2017; Revised April 2019

Abstract

Recent years have seen a flurry of activities in designing provably efficient nonconvex procedures for solving statistical estimation problems. Due to the highly nonconvex nature of the empirical loss, state-of-the-art procedures often require proper regularization (e.g. trimming, regularized cost, projection) in order to guarantee fast convergence. For vanilla procedures such as gradient descent, however, prior theory either recommends highly conservative learning rates to avoid overshooting, or completely lacks performance guarantees.

This paper uncovers a striking phenomenon in nonconvex optimization: even in the absence of explicit regularization, gradient descent enforces proper regularization implicitly under various statistical models. In fact, gradient descent follows a trajectory staying within a basin that enjoys nice geometry, consisting of points incoherent with the sampling mechanism. This "implicit regularization" feature allows gradient descent to proceed in a far more aggressive fashion without overshooting, which in turn results in substantial computational savings. Focusing on three fundamental statistical estimation problems, i.e. phase retrieval, low-rank matrix completion, and blind deconvolution, we establish that gradient descent achieves near-optimal statistical and computational guarantees without explicit regularization. In particular, by marrying statistical modeling with generic optimization theory, we develop a general recipe for analyzing the trajectories of iterative algorithms via a leave-one-out perturbation argument. As a byproduct, for noisy matrix completion, we demonstrate that gradient descent achieves near-optimal error control — measured entrywise and by the spectral norm — which might be of independent interest.

Contents

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: {congm, kaizheng}@princeton.edu.

†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Email: yuejiechi@cmu.edu.

‡Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA; Email: yuxin.chen@princeton.edu.

1 Introduction

1.1 Nonlinear systems and empirical loss minimization

A wide spectrum of science and engineering applications calls for solutions to a nonlinear system of equations. Imagine we have collected a set of data points $\boldsymbol{y} = \{y_j\}_{1\le j\le m}$, generated by a nonlinear sensing system,

$$y_j \approx \mathcal{A}_j(\boldsymbol{x}^\star), \quad 1 \le j \le m,$$

where $\boldsymbol{x}^\star$ is the unknown object of interest, and the $\mathcal{A}_j$'s are certain nonlinear maps known *a priori*. Can we reconstruct the underlying object $\boldsymbol{x}^\star$ in a faithful yet efficient manner? Problems of this kind abound in information and statistical science, prominent examples including low-rank matrix recovery [KMO10a,CR09], robust principal component analysis [CSPW11, CLMW11], phase retrieval [CSV13, JEH15], neural networks [SJL19, ZSJ+17], to name just a few.

In principle, it is possible to attempt reconstruction by searching for a solution that minimizes the empirical loss, namely,

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \sum_{j=1}^{m} \big|y_j - \mathcal{A}_j(\boldsymbol{x})\big|^2. \tag{1}$$

Unfortunately, this empirical loss minimization problem is, in many cases, nonconvex, making it NP-hard in general. This issue of non-convexity comes up in, for example, several representative problems that epitomize the structures of nonlinear systems encountered in practice.[1]

- **Phase retrieval / solving quadratic systems of equations.** Imagine we are asked to recover an unknown object $\boldsymbol{x}^\star \in \mathbb{R}^n$, but are only given the square modulus of certain linear measurements about the object, with all sign / phase information of the measurements missing. This arises, for example, in X-ray crystallography [CESV13], and in latent-variable models where the hidden variables are captured by the missing signs [CYC14]. To fix ideas, assume we would like to solve for $\boldsymbol{x}^\star \in \mathbb{R}^n$ in the following quadratic system of m equations

  $$y_j = \big(\boldsymbol{a}_j^\top \boldsymbol{x}^\star\big)^2, \qquad 1 \le j \le m,$$

  where $\{\boldsymbol{a}_j\}_{1\le j\le m}$ are the known design vectors. One strategy is thus to solve the following problem

  $$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{j=1}^{m}\Big[y_j - \big(\boldsymbol{a}_j^\top \boldsymbol{x}\big)^2\Big]^2. \tag{2}$$

- **Low-rank matrix completion.** In many scenarios such as collaborative filtering, we wish to make predictions about all entries of an (approximately) low-rank matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ (e.g. a matrix consisting of users' ratings about many movies), yet only a highly incomplete subset of the entries are revealed to us [CR09]. For clarity of presentation, assume $\boldsymbol{M}^\star$ to be rank-r ($r \ll n$) and positive semidefinite (PSD), i.e. $\boldsymbol{M}^\star = \boldsymbol{X}^\star\boldsymbol{X}^{\star\top}$ with $\boldsymbol{X}^\star \in \mathbb{R}^{n\times r}$, and suppose we have only seen the entries

  $$Y_{j,k} = M^\star_{j,k} = (\boldsymbol{X}^\star\boldsymbol{X}^{\star\top})_{j,k}, \qquad (j,k) \in \Omega$$

  within some index subset Ω of cardinality m. These entries can be viewed as nonlinear measurements about the low-rank factor $\boldsymbol{X}^\star$. The task of completing the true matrix $\boldsymbol{M}^\star$ can then be cast as solving

  $$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{n^2}{4m}\sum_{(j,k)\in\Omega}\big(Y_{j,k} - \boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k\big)^2, \tag{3}$$

  where the $\boldsymbol{e}_j$'s stand for the canonical basis vectors in $\mathbb{R}^n$.

[1] Here, we choose different pre-constants in front of the empirical loss in order to be consistent with the literature of the respective problems. In addition, we only introduce the problem in the noiseless case for simplicity of presentation.

- **Blind deconvolution / solving bilinear systems of equations.** Imagine we are interested in estimating two signals of interest $\boldsymbol{h}^\star, \boldsymbol{x}^\star \in \mathbb{C}^K$, but only get to collect a few bilinear measurements about them. This problem arises from mathematical modeling of blind deconvolution [ARR14, LLSW18], which frequently arises in astronomy, imaging, communications, etc. The goal is to recover two signals from their convolution. Put more formally, suppose we have acquired m bilinear measurements taking the following form

$$y_j = \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j, \qquad 1 \le j \le m,$$

  where $\boldsymbol{a}_j, \boldsymbol{b}_j \in \mathbb{C}^K$ are distinct design vectors (e.g. Fourier and/or random design vectors) known *a priori*, and $\boldsymbol{b}_j^{\mathsf{H}}$ denotes the conjugate transpose of $\boldsymbol{b}_j$. In order to reconstruct the underlying signals, one asks for solutions to the following problem

$$\text{minimize}_{\boldsymbol{h},\boldsymbol{x}\in\mathbb{C}^K} \quad f(\boldsymbol{h},\boldsymbol{x}) = \sum_{j=1}^{m} \left| y_j - \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j \right|^2.$$

1.2 Nonconvex optimization via regularized gradient descent

First-order methods have been a popular heuristic in practice for solving nonconvex problems including (1). For instance, a widely adopted procedure is gradient descent, which follows the update rule

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f\big(\boldsymbol{x}^t\big), \qquad t \ge 0, \tag{4}$$

where η_t is the learning rate (or step size) and $\boldsymbol{x}^0$ is some proper initial guess. Given that it only performs a single gradient calculation $\nabla f(\cdot)$ per iteration (which typically can be completed within near-linear time), this paradigm emerges as a candidate for solving large-scale problems. The concern is: whether $\boldsymbol{x}^t$ converges to the global solution and, if so, how long it takes for convergence, especially since (1) is highly nonconvex.

Fortunately, despite the worst-case hardness, appealing convergence properties have been discovered in various statistical estimation problems; the blessing being that the statistical models help rule out ill-behaved instances. For the average case, the empirical loss often enjoys benign geometry, in a *local* region (or at least along certain directions) surrounding the global optimum. In light of this, an effective nonconvex iterative method typically consists of two stages:

1. a carefully-designed initialization scheme (e.g. spectral method);
2. an iterative refinement procedure (e.g. gradient descent).

This strategy has recently spurred a great deal of interest, owing to its promise of achieving computational efficiency and statistical accuracy at once for a growing list of problems (e.g. [KMO10a, JNS13, CW15, SL16, CLS15, CC17, LLSW18, LLB17]). However, rather than directly applying gradient descent (4), existing theory often suggests enforcing proper regularization. Such explicit regularization enables improved computational convergence by properly "stabilizing" the search directions. The following regularization schemes, among others, have been suggested to obtain or improve computational guarantees. We refer to these algorithms collectively as *Regularized Gradient Descent*.

- *Trimming / truncation*, which discards/truncates a subset of the gradient components when forming the descent direction. For instance, when solving quadratic systems of equations, one can modify the gradient descent update rule as

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \mathcal{T}\left(\nabla f\big(\boldsymbol{x}^t\big)\right), \tag{5}$$

  where $\mathcal{T}$ is an operator that effectively drops samples bearing too much influence on the search direction. This strategy [CC17, ZCL16, WGE17] has been shown to enable exact recovery with linear-time computational complexity and optimal sample complexity.

- *Regularized loss*, which attempts to optimize a regularized empirical risk

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \left(\nabla f\big(\boldsymbol{x}^t\big) + \nabla R\big(\boldsymbol{x}^t\big)\right), \tag{6}$$

  where $R(\boldsymbol{x})$ stands for an additional penalty term in the empirical loss. For example, in low-rank matrix completion $R(\cdot)$ imposes penalty based on the ℓ_2 row norm [KMO10a, SL16] as well as the Frobenius norm [SL16] of the decision matrix, while in blind deconvolution, it penalizes the ℓ_2 norm as well as certain component-wise incoherence measure of the decision vectors [LLSW18, HH17, LS17].

Table 1: Prior theory for gradient descent (with spectral initialization)

	Vanilla gradient descent			Regularized gradient descent		
	sample complexity	iteration complexity	step size	sample complexity	iteration complexity	type of regularization
Phase retrieval	$n \log n$	$n \log \frac{1}{\epsilon}$	$\frac{1}{n}$	n	$\log \frac{1}{\epsilon}$	trimming [CC17, ZCL16]
Matrix completion	n/a	n/a	n/a	nr^7	$\frac{n}{r} \log \frac{1}{\epsilon}$	regularized loss [SL16]
				nr^2	$r^2 \log \frac{1}{\epsilon}$	projection [CW15, ZL16]
Blind deconvolution	n/a	n/a	n/a	$K \operatorname{poly} \log m$	$m \log \frac{1}{\epsilon}$	regularized loss & projection [LLSW18]

- *Projection*, which projects the iterates onto certain sets based on prior knowledge, that is,

$$\boldsymbol{x}^{t+1} = \mathcal{P}\left(\boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right)\right), \tag{7}$$

  where $\mathcal{P}$ is a certain projection operator used to enforce, for example, incoherence properties. This strategy has been employed in both low-rank matrix completion [CW15, ZL16] and blind deconvolution [LLSW18].

Equipped with such regularization procedures, existing works uncover appealing computational and statistical properties under various statistical models. Table 1 summarizes the performance guarantees derived in the prior literature; for simplicity, only orderwise results are provided.

Remark 1. There is another role of regularization commonly studied in the literature, which exploits prior knowledge about the structure of the unknown object, such as sparsity to prevent overfitting and improve statistical generalization ability. This is, however, not the focal point of this paper, since we are primarily pursuing solutions to (1) without imposing additional structures.

1.3 Regularization-free procedures?

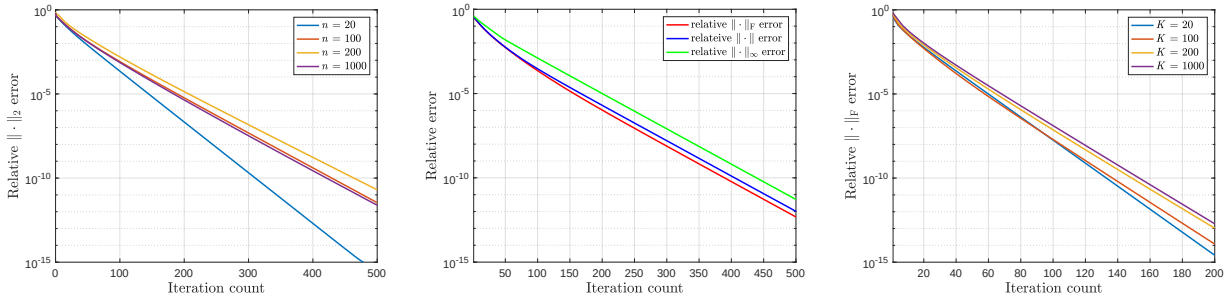
The regularized gradient descent algorithms, while exhibiting appealing performance, usually introduce more algorithmic parameters that need to be carefully tuned based on the assumed statistical models. In contrast, vanilla gradient descent (cf. (4)) — which is perhaps the very first method that comes into mind and requires minimal tuning parameters — is far less understood (cf. Table 1). Take matrix completion and blind deconvolution as examples: to the best of our knowledge, there is currently no theoretical guarantee derived for vanilla gradient descent.

The situation is better for phase retrieval: the local convergence of vanilla gradient descent, also known as Wirtinger flow (WF), has been investigated in [CLS15, SWW17]. Under i.i.d. Gaussian design and with near-optimal sample complexity, WF (combined with spectral initialization) provably achieves ϵ-accuracy (in a relative sense) within $O\big(n \log(1/\varepsilon)\big)$ iterations. Nevertheless, the computational guarantee is significantly outperformed by the regularized version (called truncated Wirtinger flow [CC17]), which only requires $O\big(\log(1/\varepsilon)\big)$ iterations to converge with similar per-iteration cost. On closer inspection, the high computational cost of WF is largely due to the vanishingly small step size $\eta_t = O\big(1/(n\|\boldsymbol{x}^\star\|_2^2)\big)$ — and hence slow movement — suggested by the theory [CLS15]. While this is already the largest possible step size allowed in the theory published in [CLS15], it is considerably more conservative than the choice $\eta_t = O\big(1/\|\boldsymbol{x}^\star\|_2^2\big)$ theoretically justified for the regularized version [CC17, ZCL16].

The lack of understanding and suboptimal results about vanilla gradient descent raise a very natural question: *are regularization-free iterative algorithms inherently suboptimal when solving nonconvex statistical estimation problems of this kind?*

1.4 Numerical surprise of unregularized gradient descent

To answer the preceding question, it is perhaps best to first collect some numerical evidence. In what follows, we test the performance of vanilla gradient descent for phase retrieval, matrix completion, and blind



(a) phase retrieval (b) matrix completion (c) blind deconvolution

Figure 1: (a) Relative ℓ_2 error of $\boldsymbol{x}^t$ (modulo the global phase) vs. iteration count for phase retrieval under i.i.d. Gaussian design, where $m = 10n$ and $\eta_t = 0.1$. (b) Relative error of $\boldsymbol{X}^t\boldsymbol{X}^{t\top}$ (measured by $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_\infty$) vs. iteration count for matrix completion, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. (c) Relative error of $\boldsymbol{h}^t\boldsymbol{x}^{t\,\mathsf{H}}$ (measured by $\|\cdot\|_{\mathrm{F}}$) vs. iteration count for blind deconvolution, where $m = 10K$ and $\eta_t = 0.5$.

deconvolution, using a *constant* step size. For all of these experiments, the initial guess is obtained by means of the standard spectral method. Our numerical findings are as follows:

- *Phase retrieval.* For each n, set $m = 10n$, take $\boldsymbol{x}^\star \in \mathbb{R}^n$ to be a random vector with unit norm, and generate the design vectors $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le j \le m$. Figure 1(a) illustrates the relative ℓ_2 error $\min\{\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|_2, \|\boldsymbol{x}^t + \boldsymbol{x}^\star\|_2\}/\|\boldsymbol{x}^\star\|_2$ (modulo the unrecoverable global phase) vs. the iteration count. The results are shown for $n = 20, 100, 200, 1000$, with the step size taken to be $\eta_t = 0.1$ in all settings.

- *Matrix completion.* Generate a random PSD matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ with dimension $n = 1000$, rank $r = 10$, and all nonzero eigenvalues equal to one. Each entry of $\boldsymbol{M}^\star$ is observed independently with probability $p = 0.1$. Figure 1(b) plots the relative error $\|\|\boldsymbol{X}^t\boldsymbol{X}^{t\top} - \boldsymbol{M}^\star\|\|/\|\|\boldsymbol{M}^\star\|\|$ vs. the iteration count, where $\|\|\cdot\|\|$ can either be the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the spectral norm $\|\cdot\|$, or the entrywise ℓ_∞ norm $\|\cdot\|_\infty$. Here, we pick the step size as $\eta_t = 0.2$.

- *Blind deconvolution.* For each $K \in \{20, 100, 200, 1000\}$ and $m = 10K$, generate the design vectors $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K) + i\mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K)$ for $1 \le j \le m$ independently,[2] and the $\boldsymbol{b}_j$'s are drawn from a partial Discrete Fourier Transform (DFT) matrix (to be described in Section 3.3). The underlying signals $\boldsymbol{h}^\star, \boldsymbol{x}^\star \in \mathbb{C}^K$ are produced as random vectors with unit norm. Figure 1(c) plots the relative error $\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\|_{\mathrm{F}}/\|\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\|_{\mathrm{F}}$ vs. the iteration count, with the step size taken to be $\eta_t = 0.5$ in all settings.

In all of these numerical experiments, vanilla gradient descent enjoys remarkable linear convergence, always yielding an accuracy of 10^{-5} (in a relative sense) within around 200 iterations. In particular, for the phase retrieval problem, the step size is taken to be $\eta_t = 0.1$ although we vary the problem size from $n = 20$ to $n = 1000$. The consequence is that the convergence rates experience little changes when the problem sizes vary. In comparison, the theory published in [CLS15] seems overly pessimistic, as it suggests a diminishing step size inversely proportional to n and, as a result, an iteration complexity that worsens as the problem size grows.

In addition, it has been empirically observed in prior literature [CC17, ZZLC17, LLSW18] that vanilla gradient descent performs comparably with the regularized counterpart for phase retrieval and blind deconvolution. To complete the picture, we further conduct experiments on matrix completion. In particular, we follow the experimental setup for matrix completion used above. We vary p from 0.01 to 0.1 with 51 logarithmically spaced points. For each p, we apply vanilla gradient descent, projected gradient descent [CW15] and gradient descent with additional regularization terms [SL16] with step size $\eta = 0.2$ to 50 randomly generated instances. Successful recovery is declared if $\|\boldsymbol{X}^t\boldsymbol{X}^{t\top} - \boldsymbol{M}^\star\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}} \le 10^{-5}$ in 10^4 iterations. Figure 2 reports the success rate vs. the sampling rate. As can be seen, the phase transition of vanilla GD and that

[2]Here and throughout, i represents the imaginary unit.

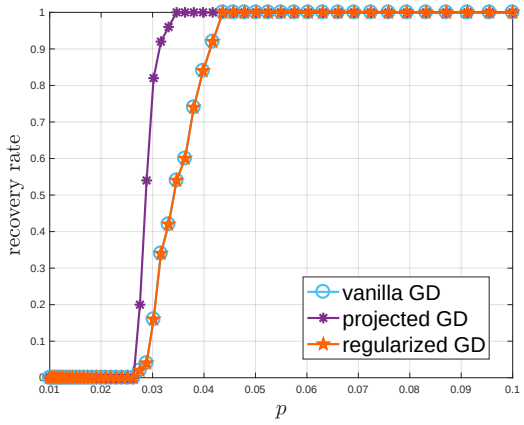


Figure 2: Success rate vs. sampling rate p over 50 Monte Carlo trials for matrix completion with $n = 1000$ and $r = 10$.

of GD with regularized cost are almost identical, whereas projected GD performs slightly better than the other two.

In short, the above empirical results are surprisingly positive yet puzzling. Why was the computational efficiency of vanilla gradient descent unexplained or substantially underestimated in prior theory?

1.5 This paper

The main contribution of this paper is towards demystifying the "unreasonable" effectiveness of regularization-free nonconvex iterative methods. As asserted in previous work, regularized gradient descent succeeds by properly enforcing/promoting certain incoherence conditions throughout the execution of the algorithm. In contrast, we discover that

> *Vanilla gradient descent automatically forces the iterates to stay incoherent with the measurement mechanism, thus implicitly regularizing the search directions.*

This "implicit regularization" phenomenon is of fundamental importance, suggesting that vanilla gradient descent proceeds as if it were properly regularized. This explains the remarkably favorable performance of unregularized gradient descent in practice. Focusing on the three representative problems mentioned in Section 1.1, our theory guarantees both statistical and computational efficiency of vanilla gradient descent under random designs and spectral initialization. With near-optimal sample complexity, to attain ϵ-accuracy,

- **Phase retrieval (informal)**: vanilla gradient descent converges in $O\big(\log n \log \frac{1}{\epsilon}\big)$ iterations;
- **Matrix completion (informal)**: vanilla gradient descent converges in $O\big(\log \frac{1}{\epsilon}\big)$ iterations;
- **Blind deconvolution (informal)**: vanilla gradient descent converges in $O\big(\log \frac{1}{\epsilon}\big)$ iterations.

In words, gradient descent provably achieves (nearly) linear convergence in all of these examples. Throughout this paper, an algorithm is said to *converge (nearly) linearly* to $\boldsymbol{x}^\star$ in the noiseless case if the iterates $\{\boldsymbol{x}^t\}$ obey

$$\mathrm{dist}(\boldsymbol{x}^{t+1}, \boldsymbol{x}^\star) \le (1-c)\,\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star), \qquad \forall t \ge 0$$

for some $0 < c \le 1$ that is (almost) independent of the problem size. Here, $\mathrm{dist}(\cdot,\cdot)$ can be any appropriate discrepancy measure.

As a byproduct of our theory, gradient descent also provably controls the *entrywise* empirical risk uniformly across all iterations; for instance, this implies that vanilla gradient descent controls entrywise estimation error for the matrix completion task. Precise statements of these results are deferred to Section 3 and are briefly summarized in Table 2.

Table 2: Prior theory vs. our theory for vanilla gradient descent (with spectral initialization)

	Prior theory			Our theory		
	sample complexity	iteration complexity	step size	sample complexity	iteration complexity	step size
Phase retrieval	$n \log n$	$n \log(1/\varepsilon)$	$1/n$	$n \log n$	$\log n \log(1/\varepsilon)$	$1/\log n$
Matrix completion	n/a	n/a	n/a	$nr^3 \mathrm{poly} \log n$	$\log(1/\varepsilon)$	1
Blind deconvolution	n/a	n/a	n/a	$K \mathrm{poly} \log m$	$\log(1/\varepsilon)$	1

Notably, our study of implicit regularization suggests that the behavior of *nonconvex optimization* algorithms for statistical estimation needs to be examined in the context of *statistical models*, which induces an objective function as a finite sum. Our proof is accomplished via a leave-one-out perturbation argument, which is inherently tied to statistical models and leverages homogeneity across samples. Altogether, this allows us to localize benign landscapes for optimization and characterize finer dynamics not accounted for in generic gradient descent theory.

1.6 Notations

Before continuing, we introduce several notations used throughout the paper. First of all, boldfaced symbols are reserved for vectors and matrices. For any vector $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|_2$ to denote its Euclidean norm. For any matrix $\boldsymbol{A}$, we use $\sigma_j(\boldsymbol{A})$ and $\lambda_j(\boldsymbol{A})$ to denote its jth largest singular value and eigenvalue, respectively, and let $\boldsymbol{A}_{j,\cdot}$ and $\boldsymbol{A}_{\cdot,j}$ denote its jth row and jth column, respectively. In addition, $\|\boldsymbol{A}\|$, $\|\boldsymbol{A}\|_{\mathrm{F}}$, $\|\boldsymbol{A}\|_{2,\infty}$, and $\|\boldsymbol{A}\|_{\infty}$ stand for the spectral norm (i.e. the largest singular value), the Frobenius norm, the ℓ_2/ℓ_∞ norm (i.e. the largest ℓ_2 norm of the rows), and the entrywise ℓ_∞ norm (the largest magnitude of all entries) of a matrix $\boldsymbol{A}$. Also, $\boldsymbol{A}^\top$, $\boldsymbol{A}^{\mathsf{H}}$ and $\overline{\boldsymbol{A}}$ denote the transpose, the conjugate transpose, and the entrywise conjugate of $\boldsymbol{A}$, respectively. $\boldsymbol{I}_n$ denotes the identity matrix with dimension $n \times n$. The notation $\mathcal{O}^{n\times r}$ represents the set of all $n \times r$ orthonormal matrices. The notation $[n]$ refers to the set $\{1, \cdots, n\}$. Also, we use $\mathrm{Re}(x)$ to denote the real part of a complex number x. Throughout the paper, we use the terms "samples" and "measurements" interchangeably.

Additionally, the standard notation $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \le c|g(n)|$, $f(n) \gtrsim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \ge c\,|g(n)|$, and $f(n) \asymp g(n)$ means that there exist constants $c_1, c_2 > 0$ such that $c_1|g(n)| \le |f(n)| \le c_2|g(n)|$. Also, $f(n) \gg g(n)$ means that there exists some large enough constant $c > 0$ such that $|f(n)| \ge c\,|g(n)|$. Similarly, $f(n) \ll g(n)$ means that there exists some sufficiently small constant $c > 0$ such that $|f(n)| \le c\,|g(n)|$.

2 Implicit regularization – a case study

To reveal reasons behind the effectiveness of vanilla gradient descent, we first examine existing theory of gradient descent and identify the geometric properties that enable linear convergence. We then develop an understanding as to why prior theory is conservative, and describe the phenomenon of implicit regularization that helps explain the effectiveness of vanilla gradient descent. To facilitate discussion, we will use the problem of solving random quadratic systems (phase retrieval) and Wirtinger flow as a case study, but our diagnosis applies more generally, as will be seen in later sections.

2.1 Gradient descent theory revisited

In the convex optimization literature, there are two standard conditions about the objective function — strong convexity and smoothness — that allow for linear convergence of gradient descent.

Definition 1 (Strong convexity)**.** *A twice continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be α-strongly convex for $\alpha > 0$ if*

$$\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}_n, \qquad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Definition 2 (Smoothness). *A twice continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be β-smooth for $\beta > 0$ if*

$$\left\|\nabla^2 f(\boldsymbol{x})\right\| \le \beta, \qquad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

It is well known that for an unconstrained optimization problem, if the objective function f is both α-strongly convex and β-smooth, then vanilla gradient descent (4) enjoys ℓ_2 error contraction [Bub15, Theorem 3.12], namely,

$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\right\|_2 \le \left(1-\frac{2}{\beta/\alpha+1}\right)\left\|\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2, \quad \text{and} \quad \left\|\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2 \le \left(1-\frac{2}{\beta/\alpha+1}\right)^t\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2, \quad t \ge 0, \tag{8}$$

as long as the step size is chosen as $\eta_t = 2/(\alpha+\beta)$. Here, $\boldsymbol{x}^\star$ denotes the global minimum. This immediately reveals the iteration complexity for gradient descent: the number of iterations taken to attain ϵ-accuracy (in a relative sense) is bounded by

$$O\left(\frac{\beta}{\alpha}\log\frac{1}{\epsilon}\right).$$

In other words, the iteration complexity is dictated by and scales linearly with the condition number — the ratio β/α of smoothness to strong convexity parameters.

Moving beyond convex optimization, one can easily extend the above theory to *nonconvex* problems with *local* strong convexity and smoothness. More precisely, suppose the objective function f satisfies

$$\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I} \qquad \text{and} \qquad \left\|\nabla^2 f(\boldsymbol{x})\right\| \le \beta$$

over a local ℓ_2 ball surrounding the global minimum $\boldsymbol{x}^\star$:

$$\mathcal{B}_\delta(\boldsymbol{x}) := \left\{\boldsymbol{x} \mid \|\boldsymbol{x}-\boldsymbol{x}^\star\|_2 \le \delta\|\boldsymbol{x}^\star\|_2\right\}. \tag{9}$$

Then the contraction result (8) continues to hold, as long as the algorithm is seeded with an initial point that falls inside $\mathcal{B}_\delta(\boldsymbol{x})$.

2.2 Local geometry for solving random quadratic systems

To invoke generic gradient descent theory, it is critical to characterize the local strong convexity and smoothness properties of the loss function. Take the problem of solving random quadratic systems (phase retrieval) as an example. Consider the i.i.d. Gaussian design in which $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le j \le m$, and suppose without loss of generality that the underlying signal obeys $\|\boldsymbol{x}^\star\|_2 = 1$. It is well known that $\boldsymbol{x}^\star$ is the unique minimizer — up to global phase — of (2) under this statistical model, provided that the ratio m/n of equations to unknowns is sufficiently large. The Hessian of the loss function $f(\boldsymbol{x})$ is given by

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^m \left[3\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top. \tag{10}$$

- *Population-level analysis.* Consider the case with an infinite number of equations or samples, i.e. $m \to \infty$, where $\nabla^2 f(\boldsymbol{x})$ converges to its expectation. Simple calculation yields that

$$\mathbb{E}\left[\nabla^2 f(\boldsymbol{x})\right] = 3\left(\|\boldsymbol{x}\|_2^2\boldsymbol{I}_n + 2\boldsymbol{x}\boldsymbol{x}^\top\right) - \left(\boldsymbol{I}_n + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right).$$

  It it straightforward to verify that for any sufficiently small constant $\delta > 0$, one has the crude bound

$$\boldsymbol{I}_n \preceq \mathbb{E}\left[\nabla^2 f(\boldsymbol{x})\right] \preceq 10\boldsymbol{I}_n, \qquad \forall \boldsymbol{x} \in \mathcal{B}_\delta(\boldsymbol{x}) : \left\|\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2 \le \delta\left\|\boldsymbol{x}^\star\right\|_2,$$

  meaning that f is 1-strongly convex and 10-smooth within a local ball around $\boldsymbol{x}^\star$. As a consequence, when we have infinite samples and an initial guess $\boldsymbol{x}^0$ such that $\|\boldsymbol{x}^0-\boldsymbol{x}^\star\|_2 \le \delta\left\|\boldsymbol{x}^\star\right\|_2$, vanilla gradient descent with a constant step size converges to the global minimum within logarithmic iterations.

- *Finite-sample regime with $m \asymp n\log n$.* Now that f exhibits favorable landscape in the population level, one thus hopes that the fluctuation can be well-controlled so that the nice geometry carries over to the finite-sample regime. In the regime where $m \asymp n\log n$ (which is the regime considered in [CLS15]), the local strong convexity is still preserved, in the sense that

$$\nabla^2 f(\boldsymbol{x}) \succeq (1/2)\cdot \boldsymbol{I}_n, \qquad \forall \boldsymbol{x}:\ \big\|\boldsymbol{x}-\boldsymbol{x}^\star\big\|_2 \le \delta\big\|\boldsymbol{x}^\star\big\|_2$$

occurs with high probability, provided that $\delta > 0$ is sufficiently small (see [Sol14, SWW17] and Lemma 1). The smoothness parameter, however, is not well-controlled. In fact, it can be as large as (up to logarithmic factors)[3]

$$\big\|\nabla^2 f(\boldsymbol{x})\big\| \lesssim n$$

even when we restrict attention to the local ℓ_2 ball (9) with $\delta > 0$ being a fixed small constant. This means that the condition number β/α (defined in Section 2.1) may scale as $O(n)$, leading to the step size recommendation

$$\eta_t \asymp 1/n,$$

and, as a consequence, a high iteration complexity $O\big(n\log(1/\epsilon)\big)$. This underpins the analysis in [CLS15].

In summary, the geometric properties of the loss function — even in the local ℓ_2 ball centering around the global minimum — is not as favorable as one anticipates, in particular in view of its population counterpart. A direct application of generic gradient descent theory leads to an overly conservative step size and a pessimistic convergence rate, unless the number of samples is enormously larger than the number of unknowns.

Remark 2. Notably, due to Gaussian designs, the phase retrieval problem enjoys more favorable geometry compared to other nonconvex problems. In matrix completion and blind deconvolution, the Hessian matrices are rank-deficient even at the population level. In such cases, the above discussions need to be adjusted, e.g. strong convexity is only possible when we restrict attention to certain directions.

2.3 Which region enjoys nicer geometry?

Interestingly, our theory identifies a local region surrounding $\boldsymbol{x}^\star$ with a large diameter that enjoys much nicer geometry. This region does not mimic an ℓ_2 ball, but rather, the intersection of an ℓ_2 ball and a polytope. We term it the *region of incoherence and contraction* (RIC). For phase retrieval, the RIC includes all points $\boldsymbol{x} \in \mathbb{R}^n$ obeying

$$\big\|\boldsymbol{x}-\boldsymbol{x}^\star\big\|_2 \le \delta\big\|\boldsymbol{x}^\star\big\|_2 \qquad \text{and} \tag{11a}$$

$$\max_{1\le j\le m} \big|\boldsymbol{a}_j^\top\big(\boldsymbol{x}-\boldsymbol{x}^\star\big)\big| \lesssim \sqrt{\log n}\,\big\|\boldsymbol{x}^\star\big\|_2, \tag{11b}$$

where $\delta > 0$ is some small numerical constant. As will be formalized in Lemma 1, with high probability the Hessian matrix satisfies

$$(1/2)\cdot \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log n)\cdot \boldsymbol{I}_n$$

simultaneously for $\boldsymbol{x}$ in the RIC. In words, the Hessian matrix is nearly well-conditioned (with the condition number bounded by $O(\log n)$), as long as (i) the iterate is not very far from the global minimizer (cf. (11a)), and (ii) the iterate remains incoherent[4] with respect to the sensing vectors (cf. (11b)). Another way to interpret the incoherence condition (11b) is that the empirical risk needs to be well-controlled uniformly across all samples. See Figure 3(a) for an illustration of the above region.

The following observation is thus immediate: one can safely adopt a far more aggressive step size (as large as $\eta_t = O(1/\log n)$) to achieve acceleration, as long as the iterates stay within the RIC. This, however, fails to be guaranteed by generic gradient descent theory. To be more precise, if the current iterate $\boldsymbol{x}^t$ falls within the desired region, then in view of (8), we can ensure ℓ_2 error contraction after one iteration, namely,

$$\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\|_2 \le \|\boldsymbol{x}^t-\boldsymbol{x}^\star\|_2$$

[3] To demonstrate this, take $\boldsymbol{x} = \boldsymbol{x}^\star + (\delta/\|\boldsymbol{a}_1\|_2)\cdot\boldsymbol{a}_1$ in (10), one can easily verify that, with high probability, $\big\|\nabla^2 f(\boldsymbol{x})\big\| \ge \big|3(\boldsymbol{a}_1^\top\boldsymbol{x})^2 - y_1\big|\,\big\|\boldsymbol{a}_1\boldsymbol{a}_1^\top\big\|/m - O(1) \gtrsim \delta^2 n^2/m \asymp \delta^2 n/\log n$.

[4] If $\boldsymbol{x}$ is aligned with (and hence very coherent with) one vector $\boldsymbol{a}_j$, then with high probability one has $\big|\boldsymbol{a}_j^\top\big(\boldsymbol{x}-\boldsymbol{x}^\star\big)\big| \gtrsim \big|\boldsymbol{a}_j^\top\boldsymbol{x}\big| \asymp \sqrt{n}\|\boldsymbol{x}\|_2$, which is significantly larger than $\sqrt{\log n}\|\boldsymbol{x}\|_2$.

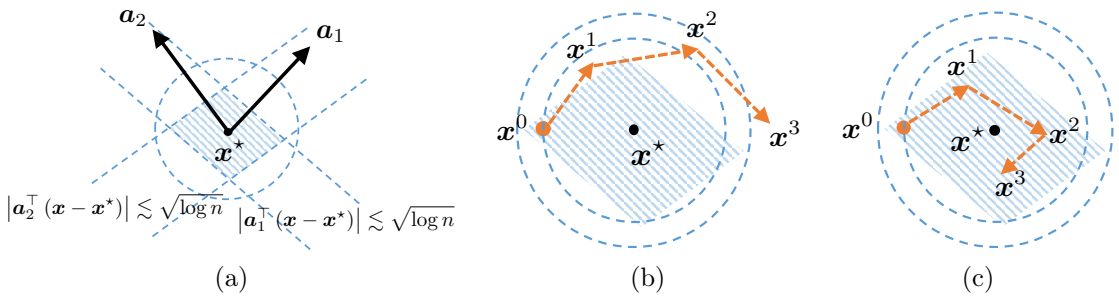


Figure 3: (a) The shaded region is an illustration of the incoherence region, which satisfies $\left|\boldsymbol{a}_j^\top (\boldsymbol{x} - \boldsymbol{x}^\star)\right| \lesssim \sqrt{\log n}$ for all points $\boldsymbol{x}$ in the region. (b) When $\boldsymbol{x}^0$ resides in the desired region, we know that $\boldsymbol{x}^1$ remains within the ℓ_2 ball but might fall out of the incoherence region (the shaded region). Once $\boldsymbol{x}^1$ leaves the incoherence region, we lose control and may overshoot. (c) Our theory reveals that with high probability, all iterates will stay within the incoherence region, enabling fast convergence.

and hence $\boldsymbol{x}^{t+1}$ stays within the local ℓ_2 ball and hence satisfies (11a). However, it is not immediately obvious that $\boldsymbol{x}^{t+1}$ would still stay incoherent with the sensing vectors and satisfy (11b). If $\boldsymbol{x}^{t+1}$ leaves the RIC, it no longer enjoys the benign local geometry of the loss function, and the algorithm has to slow down in order to avoid overshooting. See Figure 3(b) for a visual illustration. In fact, in almost all regularized gradient descent algorithms mentioned in Section 1.2, one of the main purposes of the proposed regularization procedures is to enforce such incoherence constraints.

2.4 Implicit regularization

However, is regularization really necessary for the iterates to stay within the RIC? To answer this question, we plot in Figure 4(a) (resp. Figure 4(b)) the incoherence measure $\frac{\max_j \left|\boldsymbol{a}_j^\top \boldsymbol{x}^t\right|}{\sqrt{\log n}\|\boldsymbol{x}^\star\|_2}$ (resp. $\frac{\max_j \left|\boldsymbol{a}_j^\top (\boldsymbol{x}^t - \boldsymbol{x}^\star)\right|}{\sqrt{\log n}\|\boldsymbol{x}^\star\|_2}$) vs. the iteration count in a typical Monte Carlo trial, generated in the same way as for Figure 1(a). Interestingly, the incoherence measure remains bounded by 2 for all iterations $t > 1$. This important observation suggests that one may adopt a substantially more aggressive step size throughout the whole algorithm.

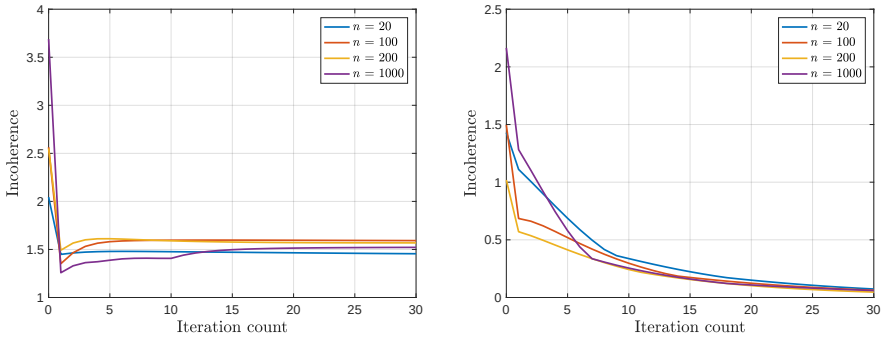


Figure 4: The incoherence measure $\frac{\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \boldsymbol{x}^t\right|}{\sqrt{\log n}\|\boldsymbol{x}^\star\|_2}$ (in (a)) and $\frac{\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top (\boldsymbol{x}^t - \boldsymbol{x}^\star)\right|}{\sqrt{\log n}\|\boldsymbol{x}^\star\|_2}$ (in (b)) of the gradient iterates vs. iteration count for the phase retrieval problem. The results are shown for $n \in \{20, 100, 200, 1000\}$ and $m = 10n$, with the step size taken to be $\eta_t = 0.1$. The problem instances are generated in the same way as in Figure 1(a).

The main objective of this paper is thus to provide a theoretical validation of the above empirical observation. As we will demonstrate shortly, with high probability all iterates along the execution of the algorithm (as well as the spectral initialization) are provably constrained within the RIC, implying fast convergence of

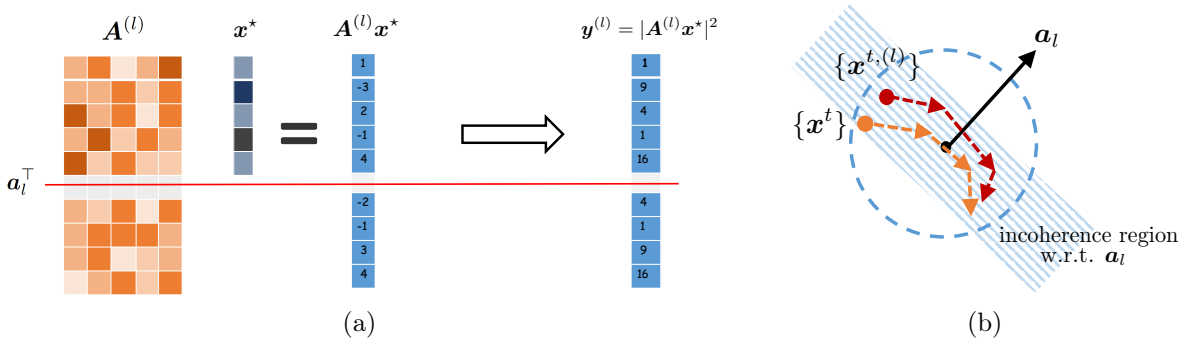


Figure 5: Illustration of the leave-one-out sequence w.r.t. $\boldsymbol{a}_l$. (a) The sequence $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$ is constructed without using the lth sample. (b) Since the auxiliary sequence $\{\boldsymbol{x}^{t,(l)}\}$ is constructed without using $\boldsymbol{a}_l$, the leave-one-out iterates stay within the incoherence region w.r.t. $\boldsymbol{a}_l$ with high probability. Meanwhile, $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{x}^{t,(l)}\}$ are expected to remain close as their construction differ only in a single sample.

vanilla gradient descent (cf. Figure 3(c)). The fact that the iterates stay incoherent with the measurement mechanism automatically, without explicit enforcement, is termed "implicit regularization".

2.5 A glimpse of the analysis: a leave-one-out trick

In order to rigorously establish (11b) for all iterates, the current paper develops a powerful mechanism based on the leave-one-out perturbation argument, a trick rooted and widely used in probability and random matrix theory. Note that the iterate $\boldsymbol{x}^t$ is statistically dependent with the design vectors $\{\boldsymbol{a}_j\}$. Under such circumstances, one often resorts to generic bounds like the Cauchy-Schwarz inequality, which would not yield a desirable estimate. To address this issue, we introduce a sequence of auxiliary iterates $\{\boldsymbol{x}^{t,(l)}\}$ for each $1 \leq l \leq m$ (for analytical purposes only), obtained by running vanilla gradient descent using all but the lth sample. As one can expect, such auxiliary trajectories serve as extremely good surrogates of $\{\boldsymbol{x}^t\}$ in the sense that

$$\boldsymbol{x}^t \approx \boldsymbol{x}^{t,(l)}, \qquad 1 \leq l \leq m, \quad t \geq 0, \tag{12}$$

since their constructions only differ by a single sample. Most importantly, since $\boldsymbol{x}^{t,(l)}$ is independent with the lth design vector, it is much easier to control its incoherence w.r.t. $\boldsymbol{a}_l$ to the desired level:

$$\left|\boldsymbol{a}_l^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right)\right| \lesssim \sqrt{\log n}\, \left\|\boldsymbol{x}^\star\right\|_2. \tag{13}$$

Combining (12) and (13) then leads to (11b). See Figure 5 for a graphical illustration of this argument. Notably, this technique is very general and applicable to many other problems. We invite the readers to Section 5 for more details.

3 Main results

This section formalizes the implicit regularization phenomenon underlying unregularized gradient descent, and presents its consequences, namely near-optimal statistical and computational guarantees for phase retrieval, matrix completion, and blind deconvolution. Note that the discrepancy measure $\mathrm{dist}\,(\cdot,\cdot)$ may vary from problem to problem.

3.1 Phase retrieval

Suppose the m quadratic equations

$$y_j = \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\star\right)^2, \qquad j = 1, 2, \ldots, m \tag{14}$$

are collected using random design vectors, namely, $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and the nonconvex problem to solve is

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) := \frac{1}{4m}\sum_{j=1}^{m}\left[\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j\right]^2. \tag{15}$$

The Wirtinger flow (WF) algorithm, first introduced in [CLS15], is a combination of spectral initialization and vanilla gradient descent; see Algorithm 1.

Algorithm 1 Wirtinger flow / gradient descent for phase retrieval

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m}$ and $\{y_j\}_{1\le j\le m}$.
Spectral initialization: Let $\lambda_1(\boldsymbol{Y})$ and $\widetilde{\boldsymbol{x}}^0$ be the leading eigenvalue and eigenvector of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{j=1}^{m} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top, \tag{16}$$

respectively, and set $\boldsymbol{x}^0 = \sqrt{\lambda_1(\boldsymbol{Y})/3}\,\widetilde{\boldsymbol{x}}^0$.
Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right). \tag{17}$$

Recognizing that the global phase / sign is unrecoverable from quadratic measurements, we introduce the ℓ_2 distance modulo the global phase as follows

$$\text{dist}(\boldsymbol{x}, \boldsymbol{x}^\star) := \min\left\{\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2, \|\boldsymbol{x} + \boldsymbol{x}^\star\|_2\right\}. \tag{18}$$

Our finding is summarized in the following theorem.

Theorem 1. *Let $\boldsymbol{x}^\star \in \mathbb{R}^n$ be a fixed vector. Suppose $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for each $1 \le j \le m$ and $m \ge c_0 n \log n$ for some sufficiently large constant $c_0 > 0$. Assume the step size obeys $\eta_t \equiv \eta = c_1/\left(\log n \cdot \|\boldsymbol{x}_0\|_2^2\right)$ for any sufficiently small constant $c_1 > 0$. Then there exist some absolute constants $0 < \varepsilon < 1$ and $c_2 > 0$ such that with probability at least $1 - O\left(mn^{-5}\right)$, Algorithm 1 satisfies that for all $t \ge 0$,*

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \varepsilon(1 - \eta\|\boldsymbol{x}^\star\|_2^2/2)^t \|\boldsymbol{x}^\star\|_2, \tag{19a}$$

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \le c_2\sqrt{\log n}\|\boldsymbol{x}^\star\|_2. \tag{19b}$$

Theorem 1 reveals a few intriguing properties of Algorithm 1.

- **Implicit regularization:** Theorem 1 asserts that the incoherence properties are satisfied throughout the execution of the algorithm (see (19b)), which formally justifies the implicit regularization feature we hypothesized.
- **Near-constant step size:** Consider the case where $\|\boldsymbol{x}^\star\|_2 = 1$. Theorem 1 establishes near-linear convergence of WF with a substantially more aggressive step size $\eta \asymp 1/\log n$. Compared with the choice $\eta \lesssim 1/n$ admissible in [CLS15, Theorem 3.3], Theorem 1 allows WF / GD to attain ϵ-accuracy within $O(\log n \log(1/\epsilon))$ iterations. The resulting computational complexity of the algorithm is

$$O\left(mn \log n \log\frac{1}{\epsilon}\right),$$

which significantly improves upon the result $O\left(mn^2 \log(1/\epsilon)\right)$ derived in [CLS15]. As a side note, if the sample size further increases to $m \asymp n\log^2 n$, then a constant step size $\eta \asymp 1$ is also feasible, resulting in an iteration complexity $\log(1/\epsilon)$. This follows since with high probability, the entire trajectory resides within a more refined incoherence region $\max_j \left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \lesssim \|\boldsymbol{x}^\star\|_2$. We omit the details here.

- **Incoherence of spectral initialization:** We have also demonstrated in Theorem 1 that the initial guess $\boldsymbol{x}^0$ falls within the RIC and is hence nearly orthogonal to all design vectors. This provides a finer characterization of spectral initialization, in comparison to prior theory that focuses primarily on the ℓ_2 accuracy [NJS13, CLS15]. We expect our leave-one-out analysis to accommodate other variants of spectral initialization studied in the literature [CC17, CLM+16, WGE17, LL17, MM17].

Remark 3. As it turns out, a carefully designed initialization is not pivotal in enabling fast convergence. In fact, randomly initialized gradient descent provably attains ε-accuracy in $O(\log n + \log \frac{1}{\varepsilon})$ iterations; see [CCFM18] for details.

3.2 Low-rank matrix completion

Let $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix[5] with rank r, and suppose its eigendecomposition is

$$\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{U}^{\star\top}, \tag{20}$$

where $\boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$ consists of orthonormal columns, and $\boldsymbol{\Sigma}^\star$ is an $r\times r$ diagonal matrix with eigenvalues in a descending order, i.e. $\sigma_{\max} = \sigma_1 \geq \cdots \geq \sigma_r = \sigma_{\min} > 0$. Throughout this paper, we assume the condition number $\kappa := \sigma_{\max}/\sigma_{\min}$ is bounded by a fixed constant, independent of the problem size (i.e. n and r). Denoting $\boldsymbol{X}^\star = \boldsymbol{U}^\star(\boldsymbol{\Sigma}^\star)^{1/2}$ allows us to factorize $\boldsymbol{M}^\star$ as

$$\boldsymbol{M}^\star = \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}. \tag{21}$$

Consider a random sampling model such that each entry of $\boldsymbol{M}^\star$ is observed independently with probability $0 < p \leq 1$, i.e. for $1 \leq j \leq k \leq n$,

$$Y_{j,k} = \begin{cases} M^\star_{j,k} + E_{j,k}, & \text{with probability } p, \\ 0, & \text{else,} \end{cases} \tag{22}$$

where the entries of $\boldsymbol{E} = [E_{j,k}]_{1\leq j\leq k\leq n}$ are independent sub-Gaussian noise with sub-Gaussian norm σ (see [Ver12, Definition 5.7]). We denote by Ω the set of locations being sampled, and $\mathcal{P}_\Omega(\boldsymbol{Y})$ represents the projection of $\boldsymbol{Y}$ onto the set of matrices supported in Ω. We note here that the sampling rate p, if not known, can be faithfully estimated by the sample proportion $|\Omega|/n^2$.

To fix ideas, we consider the following nonconvex optimization problem

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) := \frac{1}{4p}\sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k}\right)^2. \tag{23}$$

The vanilla gradient descent algorithm (with spectral initialization) is summarized in Algorithm 2.

Algorithm 2 Vanilla gradient descent for matrix completion (with spectral initialization)

Input: $\boldsymbol{Y} = [Y_{j,k}]_{1\leq j,k\leq n}$, r, p.

Spectral initialization: Let $\boldsymbol{U}^0\boldsymbol{\Sigma}^0\boldsymbol{U}^{0\top}$ be the rank-r eigendecomposition of

$$\boldsymbol{M}^0 := \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{Y}) = \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{M}^\star + \boldsymbol{E}\right),$$

and set $\boldsymbol{X}^0 = \boldsymbol{U}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla f\left(\boldsymbol{X}^t\right). \tag{24}$$

Before proceeding to the main theorem, we first introduce a standard incoherence parameter required for matrix completion [CR09].

[5]Here, we assume $\boldsymbol{M}^\star$ to be positive semidefinite to simplify the presentation, but note that our analysis easily extends to asymmetric low-rank matrices.

Definition 3 (Incoherence for matrix completion)**.** *A rank-r matrix $\boldsymbol{M}^\star$ with eigendecomposition $\boldsymbol{M}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{U}^{\star\top}$ is said to be μ-incoherent if*

$$\|\boldsymbol{U}^\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n}}\,\|\boldsymbol{U}^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}. \tag{25}$$

In addition, recognizing that $\boldsymbol{X}^\star$ is identifiable only up to orthogonal transformation, we define the optimal transform from the tth iterate $\boldsymbol{X}^t$ to $\boldsymbol{X}^\star$ as

$$\widehat{\boldsymbol{H}}^t := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{X}^t\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{26}$$

where $\mathcal{O}^{r\times r}$ is the set of $r\times r$ orthonormal matrices. With these definitions in place, we have the following theorem.

Theorem 2. *Let $\boldsymbol{M}^\star$ be a rank r, μ-incoherent PSD matrix, and its condition number κ is a fixed constant. Suppose the sample size satisfies $n^2p \ge C\mu^3r^3n\log^3 n$ for some sufficiently large constant $C>0$, and the noise satisfies*

$$\sigma\sqrt{\frac{n}{p}} \ll \frac{\sigma_{\min}}{\sqrt{\kappa^3\mu r\log^3 n}}. \tag{27}$$

With probability at least $1-O\left(n^{-3}\right)$, the iterates of Algorithm 2 satisfy

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \left(C_4\rho^t\mu r\frac{1}{\sqrt{np}} + C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{28a}$$

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{2,\infty} \le \left(C_5\rho^t\mu r\sqrt{\frac{\log n}{np}} + C_8\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{28b}$$

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\| \le \left(C_9\rho^t\mu r\frac{1}{\sqrt{np}} + C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\| \tag{28c}$$

for all $0\le t\le T = O(n^5)$, where C_1, C_4, C_5, C_8, C_9 and C_{10} are some absolute positive constants and $1-(\sigma_{\min}/5)\cdot\eta \le \rho < 1$, provided that $0<\eta_t\equiv\eta\le 2/(25\kappa\sigma_{\max})$.

Theorem 2 provides the first theoretical guarantee of unregularized gradient descent for matrix completion, demonstrating near-optimal statistical accuracy and computational complexity.

- **Implicit regularization:** In Theorem 2, we bound the ℓ_2/ℓ_∞ error of the iterates in a uniform manner via (28b). Note that $\left\|\boldsymbol{X}-\boldsymbol{X}^\star\right\|_{2,\infty} = \max_j\left\|\boldsymbol{e}_j^\top\left(\boldsymbol{X}-\boldsymbol{X}^\star\right)\right\|_2$, which implies the iterates remain incoherent with the sensing vectors throughout and have small incoherence parameters (cf. (25)). In comparison, prior works either include a penalty term on $\{\|\boldsymbol{e}_j^\top\boldsymbol{X}\|_2\}_{1\le j\le n}$ [KMO10a,SL16] and/or $\|\boldsymbol{X}\|_{\mathrm{F}}$ [SL16] to encourage an incoherent and/or low-norm solution, or add an extra projection operation to enforce incoherence [CW15, ZL16]. Our results demonstrate that such explicit regularization is unnecessary.
- **Constant step size:** Without loss of generality we may assume that $\sigma_{\max} = \|\boldsymbol{M}^\star\| = O(1)$, which can be done by choosing proper scaling of $\boldsymbol{M}^\star$. Hence we have a constant step size $\eta_t \asymp 1$. Actually it is more convenient to consider the scale invariant parameter ρ: Theorem 2 guarantees linear convergence of the vanilla gradient descent at a constant rate ρ. Remarkably, the convergence occurs with respect to three different unitarily invariant norms: the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the ℓ_2/ℓ_∞ norm $\|\cdot\|_{2,\infty}$, and the spectral norm $\|\cdot\|$. As far as we know, the latter two are established for the first time. Note that our result even improves upon that for regularized gradient descent; see Table 1.
- **Near-optimal sample complexity:** When the rank $r = O(1)$, vanilla gradient descent succeeds under a near-optimal sample complexity $n^2p \gtrsim n\mathrm{poly}\log n$, which is statistically optimal up to some logarithmic factor.

- **Near-minimal Euclidean error:** In view of (28a), as t increases, the Euclidean error of vanilla GD converges to

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|_{\mathrm{F}}\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{29}$$

  which coincides with the theoretical guarantee in [CW15, Corollary 1] and matches the minimax lower bound established in [NW12, KLT11].

- **Near-optimal entrywise error:** The ℓ_2/ℓ_∞ error bound (28b) immediately yields entrywise control of the empirical risk. Specifically, as soon as t is sufficiently large (so that the first term in (28b) is negligible), we have

$$\begin{aligned}\left\|\boldsymbol{X}^t\boldsymbol{X}^{t\top}-\boldsymbol{M}^\star\right\|_\infty&\le\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\big(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\big)^\top\right\|_\infty+\left\|\big(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\big)\boldsymbol{X}^{\star\top}\right\|_\infty\\&\le\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\right\|_{2,\infty}\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|_{2,\infty}+\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\\&\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{M}^\star\right\|_\infty,\end{aligned}$$

  where the last line follows from (28b) as well as the facts that $\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\|_{2,\infty}\le\|\boldsymbol{X}^\star\|_{2,\infty}$ and $\|\boldsymbol{M}^\star\|_\infty=\|\boldsymbol{X}^\star\|_{2,\infty}^2$. Compared with the Euclidean loss (29), this implies that when $r=O(1)$, the entrywise error of $\boldsymbol{X}^t\boldsymbol{X}^{t\top}$ is uniformly spread out across all entries. As far as we know, this is the first result that reveals near-optimal entrywise error control for noisy matrix completion using nonconvex optimization, without resorting to sample splitting.

Remark 4. Theorem 2 remains valid if the total number T of iterations obeys $T=n^{O(1)}$. In the noiseless case where $\sigma=0$, the theory allows arbitrarily large T.

Finally, we report the empirical statistical accuracy of vanilla gradient descent in the presence of noise. Figure 6 displays the squared relative error of vanilla gradient descent as a function of the signal-to-noise ratio (SNR), where the SNR is defined to be

$$\mathsf{SNR}:=\frac{\sum_{(j,k)\in\Omega}\left(M^\star_{j,k}\right)^2}{\sum_{(j,k)\in\Omega}\mathsf{Var}\left(E_{j,k}\right)}\approx\frac{\|\boldsymbol{M}^\star\|_{\mathrm{F}}^2}{n^2\sigma^2}, \tag{30}$$

and the relative error is measured in terms of the square of the metrics as in (28) as well as the squared entrywise prediction error. Both the relative error and the SNR are shown on a dB scale (i.e. $10\log_{10}(\mathrm{SNR})$ and $10\log_{10}(\text{squared relative error})$ are plotted). As one can see from the plot, the squared relative error scales inversely proportional to the SNR, which is consistent with our theory.[6]

3.3 Blind deconvolution

Suppose we have collected m bilinear measurements

$$y_j=\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j,\qquad 1\le j\le m, \tag{31}$$

where $\boldsymbol{a}_j$ follows a complex Gaussian distribution, i.e. $\boldsymbol{a}_j\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)+i\mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)$ for $1\le j\le m$, and $\boldsymbol{B}:=[\boldsymbol{b}_1,\cdots,\boldsymbol{b}_m]^{\mathsf{H}}\in\mathbb{C}^{m\times K}$ is formed by the first K columns of a unitary discrete Fourier transform (DFT) matrix $\boldsymbol{F}\in\mathbb{C}^{m\times m}$ obeying $\boldsymbol{F}\boldsymbol{F}^{\mathsf{H}}=\boldsymbol{I}_m$ (see Appendix D.3.2 for a brief introduction to DFT matrices). This setup models blind deconvolution, where the two signals under convolution belong to known low-dimensional subspaces of dimension K [ARR14][7]. In particular, the partial DFT matrix $\boldsymbol{B}$ plays an important role in image blind deblurring. In this subsection, we consider solving the following nonconvex optimization problem

$$\text{minimize}_{\boldsymbol{h},\boldsymbol{x}\in\mathbb{C}^K}\quad f\left(\boldsymbol{h},\boldsymbol{x}\right)=\sum_{j=1}^m\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j-y_j\right|^2. \tag{32}$$

[6]Note that when $\boldsymbol{M}^\star$ is well-conditioned and when $r=O(1)$, one can easily check that $\mathsf{SNR}\approx\left(\|\boldsymbol{M}^\star\|_{\mathrm{F}}^2\right)/\left(n^2\sigma^2\right)\asymp\sigma_{\min}^2/(n^2\sigma^2)$, and our theory says that the squared relative error bound is proportional to $\sigma^2/\sigma_{\min}^2$.

[7]For simplicity, we have set the dimensions of the two subspaces equal, and it is straightforward to extend our results to the case of unequal subspace dimensions.

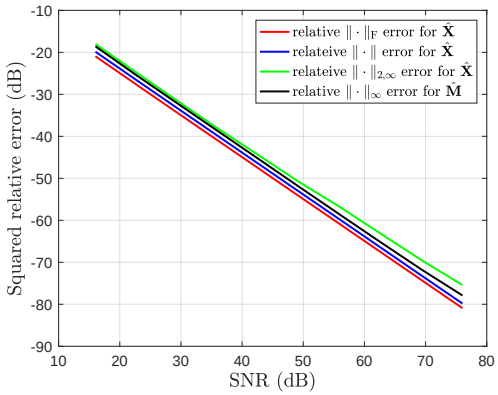


Figure 6: Squared relative error of the estimate $\widehat{\boldsymbol{X}}$ (measured by $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_{2,\infty}$ modulo global transformation) and $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ (measured by $\|\cdot\|_\infty$) vs. SNR for noisy matrix completion, where $n = 500$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. Here $\widehat{\boldsymbol{X}}$ denotes the estimate returned by Algorithm 2 after convergence.

The (Wirtinger) gradient descent algorithm (with spectral initialization) is summarized in Algorithm 3; here, $\nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{x})$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{h}, \boldsymbol{x})$ stand for the Wirtinger gradient and are given in (77) and (78), respectively; see [CLS15, Section 6] for a brief introduction to Wirtinger calculus.

It is self-evident that $\boldsymbol{h}^\star$ and $\boldsymbol{x}^\star$ are only identifiable up to global scaling, that is, for any nonzero $\alpha \in \mathbb{C}$,

$$\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} = \frac{1}{\alpha}\boldsymbol{h}^\star (\alpha \boldsymbol{x}^\star)^{\mathsf{H}}.$$

In light of this, we will measure the discrepancy between

$$\boldsymbol{z} := \begin{bmatrix}\boldsymbol{h}\\ \boldsymbol{x}\end{bmatrix} \in \mathbb{C}^{2K} \qquad \text{and} \qquad \boldsymbol{z}^\star := \begin{bmatrix}\boldsymbol{h}^\star\\ \boldsymbol{x}^\star\end{bmatrix} \in \mathbb{C}^{2K} \tag{33}$$

via the following function

$$\operatorname{dist}(\boldsymbol{z}, \boldsymbol{z}^\star) := \min_{\alpha\in\mathbb{C}} \sqrt{\left\|\frac{1}{\alpha}\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2^2 + \|\alpha\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2}. \tag{34}$$

Algorithm 3 Vanilla gradient descent for blind deconvolution (with spectral initialization)

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m}, \{\boldsymbol{b}_j\}_{1\le j\le m}$ and $\{y_j\}_{1\le j\le m}$.

Spectral initialization: Let $\sigma_1(\boldsymbol{M})$, $\check{\boldsymbol{h}}^0$ and $\check{\boldsymbol{x}}^0$ be the leading singular value, left and right singular vectors of

$$\boldsymbol{M} := \sum_{j=1}^m y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}},$$

respectively. Set $\boldsymbol{h}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\,\check{\boldsymbol{h}}^0$ and $\boldsymbol{x}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\,\check{\boldsymbol{x}}^0$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{bmatrix}\boldsymbol{h}^{t+1}\\ \boldsymbol{x}^{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{h}^{t}\\ \boldsymbol{x}^{t}\end{bmatrix} - \eta \begin{bmatrix}\frac{1}{\|\boldsymbol{x}^t\|_2^2}\nabla_{\boldsymbol{h}} f(\boldsymbol{h}^t, \boldsymbol{x}^t)\\ \frac{1}{\|\boldsymbol{h}^t\|_2^2}\nabla_{\boldsymbol{x}} f(\boldsymbol{h}^t, \boldsymbol{x}^t)\end{bmatrix}. \tag{35}$$

Before proceeding, we need to introduce the incoherence parameter [ARR14, LLSW18], which is crucial for blind deconvolution, whose role is similar to the incoherence parameter (cf. Definition 3) in matrix completion.

Definition 4 (Incoherence for blind deconvolution)**.** *Let the incoherence parameter μ of $\boldsymbol{h}^\star$ be the smallest number such that*

$$\max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right| \le \frac{\mu}{\sqrt{m}}\left\|\boldsymbol{h}^\star\right\|_2. \tag{36}$$

The incoherence parameter describes the spectral flatness of the signal $\boldsymbol{h}^\star$. With this definition in place, we have the following theorem, where for identifiability we assume that $\|\boldsymbol{h}^\star\|_2 = \|\boldsymbol{x}^\star\|_2$.

Theorem 3. *Suppose the number of measurements obeys $m \ge C\mu^2 K\log^9 m$ for some sufficiently large constant $C>0$, and suppose the step size $\eta>0$ is taken to be some sufficiently small constant. Then there exist constants $c_1, c_2, C_1, C_3, C_4 > 0$ such that with probability exceeding $1 - c_1 m^{-5} - c_1 m e^{-c_2 K}$, the iterates in Algorithm 3 satisfy*

$$\operatorname{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le C_1\left(1-\frac{\eta}{16}\right)^t \frac{1}{\log^2 m}\left\|\boldsymbol{z}^\star\right\|_2, \tag{37a}$$

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\alpha^t\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \le C_3\frac{1}{\log^{1.5} m}\left\|\boldsymbol{x}^\star\right\|_2, \tag{37b}$$

$$\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^t\right| \le C_4\frac{\mu}{\sqrt{m}}\log^2 m\left\|\boldsymbol{h}^\star\right\|_2 \tag{37c}$$

for all $t\ge 0$. Here, we denote α^t as the alignment parameter*,*

$$\alpha^t := \arg\min_{\alpha\in\mathbb{C}}\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}^t - \boldsymbol{h}^\star\right\|_2^2 + \left\|\alpha\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2^2. \tag{38}$$

Theorem 3 provides the first theoretical guarantee of unregularized gradient descent for blind deconvolution at a near-optimal statistical and computational complexity. A few remarks are in order.

- **Implicit regularization:** Theorem 3 reveals that the unregularized gradient descent iterates remain incoherent with the sampling mechanism (see (37b) and (37c)). Recall that prior works operate upon a regularized cost function with an additional penalty term that regularizes the global scaling $\{\|\boldsymbol{h}\|_2, \|\boldsymbol{x}\|_2\}$ and the incoherence $\{|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|\}_{1\le j\le m}$ [LLSW18, HH17, LS17]. In comparison, our theorem implies that it is unnecessary to regularize either the incoherence or the scaling ambiguity, which is somewhat surprising. This justifies the use of regularization-free (Wirtinger) gradient descent for blind deconvolution.

- **Constant step size:** Compared to the step size $\eta_t \lesssim 1/m$ suggested in [LLSW18] for regularized gradient descent, our theory admits a substantially more aggressive step size (i.e. $\eta_t \asymp 1$) even without regularization. Similar to phase retrieval, the computational efficiency is boosted by a factor of m, attaining ϵ-accuracy within $O\left(\log(1/\epsilon)\right)$ iterations (vs. $O\left(m\log(1/\epsilon)\right)$ iterations in prior theory).

- **Near-optimal sample complexity:** It is demonstrated that vanilla gradient descent succeeds at a near-optimal sample complexity up to logarithmic factors, although our requirement is slightly worse than [LLSW18] which uses explicit regularization. Notably, even under the sample complexity herein, the iteration complexity given in [LLSW18] is still $O\left(m/\mathrm{poly}\log(m)\right)$.

- **Incoherence of spectral initialization:** As in phase retrieval, Theorem 3 demonstrates that the estimates returned by the spectral method are incoherent with respect to both $\{\boldsymbol{a}_j\}$ and $\{\boldsymbol{b}_j\}$. In contrast, [LLSW18] recommends a projection operation (via a linear program) to enforce incoherence of the initial estimates, which is dispensable according to our theory.

- **Contraction in $\|\cdot\|_{\mathrm{F}}$:** It is easy to check that the Frobenius norm error satisfies $\left\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \lesssim \operatorname{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right)$, and therefore Theorem 3 corroborates the empirical results shown in Figure 1(c).

4 Related work

Solving nonlinear systems of equations has received much attention in the past decade. Rather than directly attacking the nonconvex formulation, convex relaxation lifts the object of interest into a higher dimensional

space and then attempts recovery via semidefinite programming (e.g. [RFP10, CSV13, CR09, ARR14]). This has enjoyed great success in both theory and practice. Despite appealing statistical guarantees, semidefinite programming is in general prohibitively expensive when processing large-scale datasets.

Nonconvex approaches, on the other end, have been under extensive study in the last few years, due to their computational advantages. There is a growing list of statistical estimation problems for which nonconvex approaches are guaranteed to find global optimal solutions, including but not limited to phase retrieval [NJS13, CLS15, CC17], low-rank matrix sensing and completion [TBS+16, BNS16, CW15, ZL15, GLM16], blind deconvolution and self-calibration [LLSW18, LS17, LLB17, LLJB17], dictionary learning [SQW17], tensor decomposition [GM17], joint alignment [CC18], learning shallow neural networks [SJL19, ZSJ+17], robust subspace learning [NNS+14, MZL19, LM14, CJN17]. In several problems [SQW16, SQW17, GM17, GLM16, LWL+16, LT16, MBM18, MZL19, DDP17], it is further suggested that the optimization landscape is benign under sufficiently large sample complexity, in the sense that all local minima are globally optimal, and hence nonconvex iterative algorithms become promising in solving such problems. See [CLC18] for a recent overview. Below we review the three problems studied in this paper in more details. Some state-of-the-art results are summarized in Table 1.

- *Phase retrieval.* Candès et al. proposed *PhaseLift* [CSV13] to solve the quadratic systems of equations based on convex programming. Specifically, it lifts the decision variable $\boldsymbol{x}^\star$ into a rank-one matrix $\boldsymbol{X}^\star = \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$ and translates the quadratic constraints of $\boldsymbol{x}^\star$ in (14) into linear constraints of $\boldsymbol{X}^\star$. By dropping the rank constraint, the problem becomes convex [CSV13, CL14, CCG15, CZ15, Tro15a]. Another convex program PhaseMax [GS18, BR17, HV16, DTL17] operates in the natural parameter space via linear programming, provided that an anchor vector is available. On the other hand, alternating minimization [NJS13] with sample splitting has been shown to enjoy much better computational guarantee. In contrast, Wirtinger Flow [CLS15] provides the first global convergence result for nonconvex methods without sample splitting, whose statistical and computational guarantees are later improved by [CC17] via an adaptive truncation strategy. Several other variants of WF are also proposed [CLM+16, KÖ16, Sol19], among which an amplitude-based loss function has been investigated [WGE17, ZZLC17, WZG+18, WGSC17]. In particular, [ZZLC17] demonstrates that the amplitude-based loss function has a better curvature, and vanilla gradient descent can indeed converge with a constant step size at the order-wise optimal sample complexity. A small sample of other nonconvex phase retrieval methods include [SBE14, SR15, CL16, CFL15, DR18, GX16, Wei15, BEB17, TV17, CLW17, QZEW17], which are beyond the scope of this paper.

- *Matrix completion.* Nuclear norm minimization was studied in [CR09] as a convex relaxation paradigm to solve the matrix completion problem. Under certain incoherence conditions imposed upon the ground truth matrix, exact recovery is guaranteed under near-optimal sample complexity [CT10, Gro11, Rec11, Che15, DR16]. Concurrently, several works [KMO10a, KMO10b, LB10, JNS13, HW14, HMLZ15, ZWL15, JN15, TW16, JKN16, WCCL16, ZWL15] tackled the matrix completion problem via nonconvex approaches. In particular, the seminal work by Keshavan et al. [KMO10a, KMO10b] pioneered the two-stage approach that is widely adopted by later works. Sun and Luo [SL16] demonstrated the convergence of gradient descent type methods for noiseless matrix completion with a regularized nonconvex loss function. Instead of penalizing the loss function, [CW15, ZL16] employed projection to enforce the incoherence condition throughout the execution of the algorithm. To the best of our knowledge, no rigorous guarantees have been established for matrix completion without explicit regularization. A notable exception is [JKN16], which uses unregularized stochastic gradient descent for matrix completion in the online setting. However, the analysis is performed with fresh samples in each iteration. Our work closes the gap and makes the first contribution towards understanding implicit regularization in gradient descent without sample splitting. In addition, entrywise eigenvector perturbation has been studied by [JN15, AFWZ17, CCF18] in order to analyze the spectral algorithms for matrix completion, which helps us establish theoretical guarantees for the spectral initialization step. Finally, it has recently been shown that the analysis of nonconvex gradient descent in turn yields near-optimal statistical guarantees for convex relaxation in the context of noisy matrix completion; see [CCF+19].

- *Blind deconvolution.* In [ARR14], Ahmed et al. first proposed to invoke similar lifting ideas for blind deconvolution, which translates the bilinear measurements (31) into a system of linear measurements of

a rank-one matrix $\boldsymbol{X}^\star = \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}$. Near-optimal performance guarantees have been established for convex relaxation [ARR14]. Under the same model, Li et al. [LLSW18] proposed a regularized gradient descent algorithm that directly optimizes the nonconvex loss function (32) with a few regularization terms that account for scaling ambiguity and incoherence. In [HH17], a Riemannian steepest descent method is developed that removes the regularization for scaling ambiguity, although they still need to regularize for incoherence. In [AAH17], a linear program is proposed but requires exact knowledge of the signs of the signals. Blind deconvolution has also been studied for other models – interested readers may refer to [Chi16, LS17, LLJB17, LS15, LTR16, ZLK$^+$17, WC16].

On the other hand, our analysis framework is based on a leave-one-out perturbation argument. This technique has been widely used to analyze high-dimensional problems with random designs, including but not limited to robust M-estimation [EKBB$^+$13, EK15], statistical inference for sparse regression [JM$^+$18], likelihood ratio test in logistic regression [SCC17], phase synchronization [ZB18, AFWZ17], ranking from pairwise comparisons [CFMW17], community recovery [AFWZ17], and covariance sketching [LMCC18]. In particular, this technique results in tight performance guarantees for the generalized power method [ZB18], the spectral method [AFWZ17, CFMW17], and convex programming approaches [EK15, ZB18, SCC17, CFMW17], however it has not been applied to analyze nonconvex optimization algorithms.

Finally, we note that the notion of implicit regularization — broadly defined — arises in settings far beyond the models and algorithms considered herein. For instance, it has been conjectured that in matrix factorization, over-parameterized stochastic gradient descent effectively enforces certain norm constraints, allowing it to converge to a minimal-norm solution as long as it starts from the origin [GWB$^+$17]. The stochastic gradient methods have also been shown to implicitly enforce Tikhonov regularization in several statistical learning settings [LCR16]. More broadly, this phenomenon seems crucial in enabling efficient training of deep neural networks [ZBH$^+$17, SHS17].

5 A general recipe for trajectory analysis

In this section, we sketch a general recipe for establishing performance guarantees of gradient descent, which conveys the key idea for proving the main results of this paper. The main challenge is to demonstrate that appropriate incoherence conditions are preserved throughout the trajectory of the algorithm. This requires exploiting statistical independence of the samples in a careful manner, in conjunction with generic optimization theory. Central to our approach is a leave-one-out perturbation argument, which allows to decouple the statistical dependency while controlling the component-wise incoherence measures.

General Recipe (a leave-one-out analysis)

Step 1: characterize restricted strong convexity and smoothness of f, and identify the region of incoherence and contraction (RIC).

Step 2: introduce leave-one-out sequences $\{\boldsymbol{X}^{t,(l)}\}$ and $\{\boldsymbol{H}^{t,(l)}\}$ for each l, where $\{\boldsymbol{X}^{t,(l)}\}$ (resp. $\{\boldsymbol{H}^{t,(l)}\}$) is independent of any sample involving $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$);

Step 3: establish the incoherence condition for $\{\boldsymbol{X}^t\}$ and $\{\boldsymbol{H}^t\}$ via induction. Suppose the iterates satisfy the claimed conditions in the tth iteration:

(a) show, via restricted strong convexity, that the true iterates $(\boldsymbol{X}^{t+1}, \boldsymbol{H}^{t+1})$ and the leave-one-out version $(\boldsymbol{X}^{t+1,(l)}, \boldsymbol{H}^{t+1,(l)})$ are exceedingly close;

(b) use statistical independence to show that $\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star$ (resp. $\boldsymbol{H}^{t+1,(l)} - \boldsymbol{H}^\star$) is incoherent w.r.t. $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$), namely, $\|\boldsymbol{\phi}_l^{\mathsf{H}}(\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star)\|_2$ and $\|\boldsymbol{\psi}_l^{\mathsf{H}}(\boldsymbol{H}^{t+1,(l)} - \boldsymbol{H}^\star)\|_2$ are both well-controlled;

(c) combine the bounds to establish the desired incoherence condition concerning $\max_l \|\boldsymbol{\phi}_l^{\mathsf{H}}(\boldsymbol{X}^{t+1} - \boldsymbol{X}^\star)\|_2$ and $\max_l \|\boldsymbol{\psi}_l^{\mathsf{H}}(\boldsymbol{H}^{t+1} - \boldsymbol{H}^\star)\|_2$.

5.1 General model

Consider the following problem where the samples are collected in a bilinear/quadratic form as

$$y_j = \boldsymbol{\psi}_j^{\mathsf{H}} \boldsymbol{H}^\star \boldsymbol{X}^{\star\mathsf{H}} \boldsymbol{\phi}_j, \qquad 1 \le j \le m, \tag{39}$$

where the objects of interest $\boldsymbol{H}^\star, \boldsymbol{X}^\star \in \mathbb{C}^{n\times r}$ or $\mathbb{R}^{n\times r}$ might be vectors or tall matrices taking either real or complex values. The design vectors $\{\boldsymbol{\psi}_j\}$ and $\{\boldsymbol{\phi}_j\}$ are in either $\mathbb{C}^n$ or $\mathbb{R}^n$, and can be either random or deterministic. This model is quite general and entails all three examples in this paper as special cases:

- *Phase retrieval*: $\boldsymbol{H}^\star = \boldsymbol{X}^\star = \boldsymbol{x}^\star \in \mathbb{R}^n$, and $\boldsymbol{\psi}_j = \boldsymbol{\phi}_j = \boldsymbol{a}_j$;
- *Matrix completion*: $\boldsymbol{H}^\star = \boldsymbol{X}^\star \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\psi}_j, \boldsymbol{\phi}_j \in \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n\}$;
- *Blind deconvolution*: $\boldsymbol{H}^\star = \boldsymbol{h}^\star \in \mathbb{C}^K$, $\boldsymbol{X}^\star = \boldsymbol{x}^\star \in \mathbb{C}^K$, $\boldsymbol{\phi}_j = \boldsymbol{a}_j$, and $\boldsymbol{\psi}_j = \boldsymbol{b}_j$.

For this setting, the empirical loss function is given by

$$f(\boldsymbol{Z}) := f(\boldsymbol{H}, \boldsymbol{X}) = \frac{1}{m}\sum_{j=1}^{m} \Big|\boldsymbol{\psi}_j^{\mathsf{H}} \boldsymbol{H}\boldsymbol{X}^{\mathsf{H}} \boldsymbol{\phi}_j - y_j\Big|^2,$$

where we denote $\boldsymbol{Z} = (\boldsymbol{H}, \boldsymbol{X})$. To minimize $f(\boldsymbol{Z})$, we proceed with vanilla gradient descent

$$\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t - \eta \nabla f\big(\boldsymbol{Z}^t\big), \qquad \forall t \ge 0$$

following a standard spectral initialization, where η is the step size. As a remark, for complex-valued problems, the gradient (resp. Hessian) should be understood as the Wirtinger gradient (resp. Hessian).

It is clear from (39) that $\boldsymbol{Z}^\star = (\boldsymbol{H}^\star, \boldsymbol{X}^\star)$ can only be recovered up to certain global ambiguity. For clarity of presentation, we assume in this section that such ambiguity has already been taken care of via proper global transformation.

5.2 Outline of the recipe

We are now positioned to outline the general recipe, which entails the following steps.

- **Step 1: characterizing local geometry in the RIC.** Our first step is to characterize a region $\mathcal{R}$ — which we term as the *region of incoherence and contraction* (RIC) — such that the Hessian matrix $\nabla^2 f(\boldsymbol{Z})$ obeys strong convexity and smoothness,

$$\boldsymbol{0} \prec \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{Z}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{Z} \in \mathcal{R}, \tag{40}$$

  or at least along certain directions (i.e. restricted strong convexity and smoothness), where β/α scales slowly (or even remains bounded) with the problem size. As revealed by optimization theory, this geometric property (40) immediately implies linear convergence with the contraction rate $1 - O(\alpha/\beta)$ for a properly chosen step size η, as long as all iterates stay within the RIC.

  A natural question then arises: what does the RIC $\mathcal{R}$ look like? As it turns out, the RIC typically contains all points such that the ℓ_2 error $\|\boldsymbol{Z} - \boldsymbol{Z}^\star\|_{\mathrm{F}}$ is not too large and

$$\text{(\textbf{incoherence})} \qquad \max_j \big\|\boldsymbol{\phi}_j^{\mathsf{H}}(\boldsymbol{X} - \boldsymbol{X}^\star)\big\|_2 \ \text{ and } \ \max_j \big\|\boldsymbol{\psi}_j^{\mathsf{H}}(\boldsymbol{H} - \boldsymbol{H}^\star)\big\|_2 \text{ are well-controlled.} \tag{41}$$

  In the three examples, the above incoherence condition translates to:

  - *Phase retrieval*: $\max_j \big|\boldsymbol{a}_j^\top(\boldsymbol{x} - \boldsymbol{x}^\star)\big|$ is well-controlled;
  - *Matrix completion*: $\big\|\boldsymbol{X} - \boldsymbol{X}^\star\big\|_{2,\infty}$ is well-controlled;
  - *Blind deconvolution*: $\max_j \big|\boldsymbol{a}_j^\top(\boldsymbol{x} - \boldsymbol{x}^\star)\big|$ and $\max_j \big|\boldsymbol{b}_j^\top(\boldsymbol{h} - \boldsymbol{h}^\star)\big|$ are well-controlled.

- **Step 2: introducing the leave-one-out sequences.** To justify that no iterates leave the RIC, we rely on the construction of auxiliary sequences. Specifically, for each l, produce an auxiliary sequence $\{\boldsymbol{Z}^{t,(l)} = (\boldsymbol{X}^{t,(l)}, \boldsymbol{H}^{t,(l)})\}$ such that $\boldsymbol{X}^{t,(l)}$ (resp. $\boldsymbol{H}^{t,(l)}$) is independent of any sample involving $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$). As an example, suppose that the $\boldsymbol{\phi}_l$'s and the $\boldsymbol{\psi}_l$'s are independently and randomly generated. Then for each l, one can consider a leave-one-out loss function

$$f^{(l)}(\boldsymbol{Z}) := \frac{1}{m}\sum_{j:j\neq l}\left|\boldsymbol{\psi}_j^{\mathsf{H}}\boldsymbol{H}\boldsymbol{X}^{\mathsf{H}}\boldsymbol{\phi}_j - y_j\right|^2$$

  that discards the lth sample. One further generates $\{\boldsymbol{Z}^{t,(l)}\}$ by running vanilla gradient descent w.r.t. this auxiliary loss function, with a spectral initialization that similarly discards the lth sample. Note that this procedure is only introduced to facilitate analysis and is never implemented in practice.

- **Step 3: establishing the incoherence condition.** We are now ready to establish the incoherence condition with the assistance of the auxiliary sequences. Usually the proof proceeds by induction, where our goal is to show that the next iterate remains within the RIC, given that the current one does.
  - **Step 3(a): proximity between the original and the leave-one-out iterates.** As one can anticipate, $\{\boldsymbol{Z}^t\}$ and $\{\boldsymbol{Z}^{t,(l)}\}$ remain "glued" to each other along the whole trajectory, since their constructions differ by only a single sample. In fact, as long as the initial estimates stay sufficiently close, their gaps will never explode. To intuitively see why, use the fact $\nabla f(\boldsymbol{Z}^t) \approx \nabla f^{(l)}(\boldsymbol{Z}^t)$ to discover that

$$\begin{aligned}\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^{t+1,(l)} &= \boldsymbol{Z}^t - \eta\nabla f(\boldsymbol{Z}^t) - \big(\boldsymbol{Z}^{t,(l)} - \eta\nabla f^{(l)}\big(\boldsymbol{Z}^{t,(l)}\big)\big)\\ &\approx \boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)} - \eta\nabla^2 f(\boldsymbol{Z}^t)\big(\boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)}\big),\end{aligned}$$

    which together with the strong convexity condition implies ℓ_2 contraction

$$\big\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^{t+1,(l)}\big\|_{\mathrm{F}} \approx \Big\|\big(\boldsymbol{I} - \eta\nabla^2 f(\boldsymbol{Z}^t)\big)\big(\boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)}\big)\Big\|_{\mathrm{F}} \le \big\|\boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)}\big\|_2.$$

    Indeed, (restricted) strong convexity is crucial in controlling the size of leave-one-out perturbations.
  - **Step 3(b): incoherence condition of the leave-one-out iterates.** The fact that $\boldsymbol{Z}^{t+1}$ and $\boldsymbol{Z}^{t+1,(l)}$ are exceedingly close motivates us to control the incoherence of $\boldsymbol{Z}^{t+1,(l)} - \boldsymbol{Z}^\star$ instead, for $1 \le l \le m$. By construction, $\boldsymbol{X}^{t+1,(l)}$ (resp. $\boldsymbol{H}^{t+1,(l)}$) is statistically *independent* of any sample involving the design vector $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$), a fact that typically leads to a more friendly analysis for controlling $\big\|\boldsymbol{\phi}_l^{\mathsf{H}}\big(\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star\big)\big\|_2$ and $\big\|\boldsymbol{\psi}_l^{\mathsf{H}}\big(\boldsymbol{H}^{t+1,(l)} - \boldsymbol{H}^\star\big)\big\|_2$.
  - **Step 3(c): combining the bounds.** With these results in place, apply the triangle inequality to obtain

$$\big\|\boldsymbol{\phi}_l^{\mathsf{H}}\big(\boldsymbol{X}^{t+1} - \boldsymbol{X}^\star\big)\big\|_2 \le \|\boldsymbol{\phi}_l\|_2\big\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^{t+1,(l)}\big\|_{\mathrm{F}} + \big\|\boldsymbol{\phi}_l^{\mathsf{H}}\big(\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star\big)\big\|_2,$$

    where the first term is controlled in Step 3(a) and the second term is controlled in Step 3(b). The term $\big\|\boldsymbol{\psi}_l^{\mathsf{H}}\big(\boldsymbol{H}^{t+1} - \boldsymbol{H}^\star\big)\big\|_2$ can be bounded similarly. By choosing the bounds properly, this establishes the incoherence condition for all $1 \le l \le m$ as desired.

6 Analysis for phase retrieval

In this section, we instantiate the general recipe presented in Section 5 to phase retrieval and prove Theorem 1. Similar to the Section 7.1 in [CLS15], we are going to use $\eta_t = c_1/(\log n \cdot \|\boldsymbol{x}^\star\|_2^2)$ instead of $c_1/(\log n \cdot \|\boldsymbol{x}_0\|_2^2)$ as the step size for analysis. This is because with high probability, $\|\boldsymbol{x}_0\|_2$ and $\|\boldsymbol{x}^\star\|_2$ are rather close in the relative sense. Without loss of generality, we assume throughout this section that $\big\|\boldsymbol{x}^\star\big\|_2 = 1$ and

$$\mathrm{dist}(\boldsymbol{x}^0, \boldsymbol{x}^\star) = \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2 \le \|\boldsymbol{x}^0 + \boldsymbol{x}^\star\|_2. \tag{42}$$

In addition, the gradient and the Hessian of $f(\cdot)$ for this problem (see (15)) are given respectively by

$$\nabla f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^m\Big[\big(\boldsymbol{a}_j^\top\boldsymbol{x}\big)^2 - y_j\Big]\big(\boldsymbol{a}_j^\top\boldsymbol{x}\big)\,\boldsymbol{a}_j, \tag{43}$$

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m}\left[3\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top, \tag{44}$$

which are useful throughout the proof.

6.1 Step 1: characterizing local geometry in the RIC

6.1.1 Local geometry

We start by characterizing the region that enjoys both strong convexity and the desired level of smoothness. This is supplied in the following lemma, which plays a crucial role in the subsequent analysis.

Lemma 1 (Restricted strong convexity and smoothness for phase retrieval)**.** *Fix any sufficiently small constant $C_1 > 0$ and any sufficiently large constant $C_2 > 0$, and suppose the sample complexity obeys $m \geq c_0 n \log n$ for some sufficiently large constant $c_0 > 0$. With probability at least $1 - O(mn^{-10})$,*

$$\nabla^2 f(\boldsymbol{x}) \succeq (1/2)\cdot \boldsymbol{I}_n$$

holds simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq 2C_1$; and

$$\nabla^2 f(\boldsymbol{x}) \preceq (5C_2(10 + C_2)\log n)\cdot \boldsymbol{I}_n$$

holds simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n$ obeying

$$\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq 2C_1, \tag{45a}$$

$$\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^\top(\boldsymbol{x} - \boldsymbol{x}^\star)\right| \leq C_2\sqrt{\log n}. \tag{45b}$$

Proof. See Appendix A.1. □

In words, Lemma 1 reveals that the Hessian matrix is positive definite and (almost) well-conditioned, if one restricts attention to the set of points that are (i) not far away from the truth (cf. (45a)) and (ii) incoherent with respect to the measurement vectors $\{\boldsymbol{a}_j\}_{1\leq j\leq m}$ (cf. (45b)).

6.1.2 Error contraction

As we point out before, the nice local geometry enables ℓ_2 contraction, which we formalize below.

Lemma 2. *There exists an event that does not depend on t and has probability $1 - O(mn^{-10})$, such that when it happens and $\boldsymbol{x}^t$ obeys the conditions (45), one has*

$$\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\right\|_2 \leq (1 - \eta/2)\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \tag{46}$$

provided that the step size satisfies $0 < \eta \leq 1/[5C_2(10 + C_2)\log n]$.

Proof. This proof applies the standard argument when establishing the ℓ_2 error contraction of gradient descent for strongly convex and smooth functions. See Appendix A.2. □

With the help of Lemma 2, we can turn the proof of Theorem 1 into ensuring that the trajectory $\{\boldsymbol{x}^t\}_{0\leq t\leq n}$ lies in the RIC specified by (47).[8] This is formally stated in the next lemma.

Lemma 3. *Suppose for all $0 \leq t \leq T_0 := n$, the trajectory $\{\boldsymbol{x}^t\}$ falls within the region of incoherence and contraction (termed the RIC), namely,*

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \leq C_1, \tag{47a}$$

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \leq C_2\sqrt{\log n}, \tag{47b}$$

then the claims in Theorem 1 hold true. Here and throughout this section, $C_1, C_2 > 0$ are two absolute constants as specified in Lemma 1.

Proof. See Appendix A.3. □

[8] Here, we deliberately change $2C_1$ in (45a) to C_1 in the definition of the RIC (47a) to ensure the correctness of the analysis.

6.2 Step 2: introducing the leave-one-out sequences

In comparison to the ℓ_2 error bound (47a) that captures the overall loss, the incoherence hypothesis (47b) — which concerns sample-wise control of the empirical risk — is more complicated to establish. This is partly due to the statistical dependence between $\boldsymbol{x}^t$ and the sampling vectors $\{\boldsymbol{a}_l\}$. As described in the general recipe, the key idea is the introduction of a *leave-one-out* version of the WF iterates, which removes a single measurement from consideration.

To be precise, for each $1 \le l \le m$, we define the leave-one-out empirical loss function as

$$f^{(l)}(\boldsymbol{x}) := \frac{1}{4m} \sum_{j:j\neq l} \left[\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j \right]^2, \tag{48}$$

and the auxiliary trajectory $\left\{\boldsymbol{x}^{t,(l)}\right\}_{t\ge 0}$ is constructed by running WF w.r.t. $f^{(l)}(\boldsymbol{x})$. In addition, the spectral initialization $\boldsymbol{x}^{0,(l)}$ is computed based on the rescaled leading eigenvector of the leave-one-out data matrix

$$\boldsymbol{Y}^{(l)} := \frac{1}{m} \sum_{j:j\neq l} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top. \tag{49}$$

Clearly, the entire sequence $\left\{\boldsymbol{x}^{t,(l)}\right\}_{t\ge 0}$ is independent of the lth sampling vector $\boldsymbol{a}_l$. This auxiliary procedure is formally described in Algorithm 4.

Algorithm 4 The lth leave-one-out sequence for phase retrieval

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m, j\neq l}$ and $\{y_j\}_{1\le j\le m, j\neq l}$.

Spectral initialization: let $\lambda_1\big(\boldsymbol{Y}^{(l)}\big)$ and $\widetilde{\boldsymbol{x}}^{0,(l)}$ be the leading eigenvalue and eigenvector of

$$\boldsymbol{Y}^{(l)} = \frac{1}{m} \sum_{j:j\neq l} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top,$$

respectively, and set

$$\boldsymbol{x}^{0,(l)} = \begin{cases} \sqrt{\lambda_1\left(\boldsymbol{Y}^{(l)}\right)/3}\, \widetilde{\boldsymbol{x}}^{0,(l)}, & \text{if } \left\|\widetilde{\boldsymbol{x}}^{0,(l)} - \boldsymbol{x}^\star\right\|_2 \le \left\|\widetilde{\boldsymbol{x}}^{0,(l)} + \boldsymbol{x}^\star\right\|_2, \\ -\sqrt{\lambda_1\left(\boldsymbol{Y}^{(l)}\right)/3}\, \widetilde{\boldsymbol{x}}^{0,(l)}, & \text{else.} \end{cases}$$

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{x}^{t+1,(l)} = \boldsymbol{x}^{t,(l)} - \eta_t \nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big). \tag{50}$$

6.3 Step 3: establishing the incoherence condition by induction

As revealed by Lemma 3, it suffices to prove that the iterates $\{\boldsymbol{x}^t\}_{0\le t\le T_0}$ satisfies (47) with high probability. Our proof will be inductive in nature. For the sake of clarity, we list all the induction hypotheses:

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \le C_1, \tag{51a}$$

$$\max_{1\le l\le m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \le C_3 \sqrt{\frac{\log n}{n}} \tag{51b}$$

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \le C_2 \sqrt{\log n}. \tag{51c}$$

Here $C_3 > 0$ is some universal constant. For any $t \ge 0$, define $\mathcal{E}_t$ to be the event where the conditions in (51) hold for the t-th iteration. According to Lemma 2, there exists some event $\mathcal{E}$ with probability $1 - O(mn^{-10})$ such that on $\mathcal{E}_t \cap \mathcal{E}$ one has

$$\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\right\|_2 \le C_1. \tag{52}$$

This subsection is devoted to establishing (51b) and (51c) for the $(t+1)$th iteration, assuming that (51) holds true up to the tth iteration. We defer the justification of the base case (i.e. initialization at $t=0$) to Section 6.4.

- **Step 3(a): proximity between the original and the leave-one-out iterates.** The leave-one-out sequence $\{\boldsymbol{x}^{t,(l)}\}$ behaves similarly to the true WF iterates $\{\boldsymbol{x}^t\}$ while maintaining statistical independence with $\boldsymbol{a}_l$, a key fact that allows us to control the incoherence of lth leave-one-out sequence w.r.t. $\boldsymbol{a}_l$. We will formally quantify the gap between $\boldsymbol{x}^{t+1}$ and $\boldsymbol{x}^{t+1,(l)}$ in the following lemma, which establishes the induction in (51b).

  Lemma 4. *Suppose that the sample size obeys $m \geq Cn\log n$ for some sufficiently large constant $C>0$ and that the stepsize obeys $0<\eta<1/[5C_2(10+C_2)\log n]$. Then on some event $\mathcal{E}_{t+1,1}\subseteq\mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t\cap\mathcal{E}_{t+1,1}^c)=O(mn^{-10})$, one has*

$$\max_{1\leq l\leq m}\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right\|_2\leq C_3\sqrt{\frac{\log n}{n}}. \tag{53}$$

  Proof. The proof relies heavily on the restricted strong convexity (see Lemma 1) and is deferred to Appendix A.4. □

- **Step 3(b): incoherence of the leave-one-out iterates.** By construction, $\boldsymbol{x}^{t+1,(l)}$ is statistically independent of the sampling vector $\boldsymbol{a}_l$. One can thus invoke the standard Gaussian concentration results and the union bound to derive that on an event $\mathcal{E}_{t+1,2}\subseteq\mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t\cap\mathcal{E}_{t+1,2}^c)=O(mn^{-10})$,

$$\begin{aligned}\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right| &\leq 5\sqrt{\log n}\left\|\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^\star\right\|_2\\ &\overset{\text{(i)}}{\leq} 5\sqrt{\log n}\left(\left\|\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1}\right\|_2+\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\right\|_2\right)\\ &\overset{\text{(ii)}}{\leq} 5\sqrt{\log n}\left(C_3\sqrt{\frac{\log n}{n}}+C_1\right)\\ &\leq C_4\sqrt{\log n}\end{aligned} \tag{54}$$

  holds for some constant $C_4\geq 6C_1>0$ and n sufficiently large. Here, (i) comes from the triangle inequality, and (ii) arises from the proximity bound (53) and the condition (52).

- **Step 3(c): combining the bounds.** We are now prepared to establish (51c) for the $(t+1)$th iteration. Specifically,

$$\begin{aligned}\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\right)\right| &\leq \max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right)\right|+\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right|\\ &\overset{\text{(i)}}{\leq}\max_{1\leq l\leq m}\|\boldsymbol{a}_l\|_2\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right\|_2+C_4\sqrt{\log n}\\ &\overset{\text{(ii)}}{\leq}\sqrt{6n}\cdot C_3\sqrt{\frac{\log n}{n}}+C_4\sqrt{\log n}\leq C_2\sqrt{\log n},\end{aligned} \tag{55}$$

  where (i) follows from the Cauchy-Schwarz inequality and (54), the inequality (ii) is a consequence of (53) and (98), and the last inequality holds as long as $C_2/(C_3+C_4)$ is sufficiently large. From the deduction above we easily get $\mathbb{P}(\mathcal{E}_t\cap\mathcal{E}_{t+1}^c)=O(mn^{-10})$.

Using mathematical induction and the union bound, we establish (51) for all $t\leq T_0=n$ with high probability. This in turn concludes the proof of Theorem 1, as long as the hypotheses are valid for the base case.

6.4 The base case: spectral initialization

In the end, we return to verify the induction hypotheses for the base case ($t = 0$), i.e. the spectral initialization obeys (51). The following lemma justifies (51a) by choosing δ sufficiently small.

Lemma 5. *Fix any small constant $\delta > 0$, and suppose $m > c_0 n \log n$ for some large constant $c_0 > 0$. Consider the two vectors $\boldsymbol{x}^0$ and $\widetilde{\boldsymbol{x}}^0$ as defined in Algorithm 1, and suppose without loss of generality that (42) holds. Then with probability exceeding $1 - O(n^{-10})$, one has*

$$\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\| \le \delta, \tag{56}$$

$$\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2 \le 2\delta \qquad \text{and} \qquad \left\|\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2 \le \sqrt{2}\delta. \tag{57}$$

Proof. This result follows directly from the Davis-Kahan sinΘ theorem. See Appendix A.5. □

We then move on to justifying (51b), the proximity between the original and leave-one-out iterates for $t = 0$.

Lemma 6. *Suppose $m > c_0 n \log n$ for some large constant $c_0 > 0$. Then with probability at least $1 - O(mn^{-10})$, one has*

$$\max_{1\le l\le m} \left\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\right\|_2 \le C_3\sqrt{\frac{\log n}{n}}. \tag{58}$$

Proof. This is also a consequence of the Davis-Kahan sinΘ theorem. See Appendix A.6. □

The final claim (51c) can be proved using the same argument as in deriving (55), and hence is omitted.

7 Analysis for matrix completion

In this section, we instantiate the general recipe presented in Section 5 to matrix completion and prove Theorem 2. Before continuing, we first gather a few useful facts regarding the loss function in (23). The gradient of it is given by

$$\nabla f(\boldsymbol{X}) = \frac{1}{p}\mathcal{P}_\Omega\left[\boldsymbol{X}\boldsymbol{X}^\top - (\boldsymbol{M}^\star + \boldsymbol{E})\right]\boldsymbol{X}. \tag{59}$$

We define the expected gradient (with respect to the sampling set Ω) to be

$$\nabla F(\boldsymbol{X}) = \left[\boldsymbol{X}\boldsymbol{X}^\top - (\boldsymbol{M}^\star + \boldsymbol{E})\right]\boldsymbol{X}$$

and also the (expected) gradient without noise to be

$$\nabla f_{\text{clean}}(\boldsymbol{X}) = \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\boldsymbol{X} \qquad \text{and} \qquad \nabla F_{\text{clean}}(\boldsymbol{X}) = \left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\boldsymbol{X}. \tag{60}$$

In addition, we need the Hessian $\nabla^2 f_{\text{clean}}(\boldsymbol{X})$, which is represented by an $nr \times nr$ matrix. Simple calculations reveal that for any $\boldsymbol{V} \in \mathbb{R}^{n\times r}$,

$$\text{vec}(\boldsymbol{V})^\top \nabla^2 f_{\text{clean}}(\boldsymbol{X})\,\text{vec}(\boldsymbol{V}) = \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{V}^\top\right)\right\|_{\text{F}}^2 + \frac{1}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right), \boldsymbol{V}\boldsymbol{V}^\top\right\rangle, \tag{61}$$

where $\text{vec}(\boldsymbol{V}) \in \mathbb{R}^{nr}$ denotes the vectorization of $\boldsymbol{V}$.

7.1 Step 1: characterizing local geometry in the RIC

7.1.1 Local geometry

The first step is to characterize the region where the empirical loss function enjoys restricted strong convexity and smoothness in an appropriate sense. This is formally stated in the following lemma.

Lemma 7 (Restricted strong convexity and smoothness for matrix completion). *Suppose that the sample size obeys $n^2 p \geq C\kappa^2 \mu r n \log n$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O\left(n^{-10}\right)$, the Hessian $\nabla^2 f_{\mathrm{clean}}(\boldsymbol{X})$ as defined in (61) obeys*

$$\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f_{\mathrm{clean}}\left(\boldsymbol{X}\right) \mathrm{vec}\left(\boldsymbol{V}\right) \geq \frac{\sigma_{\min}}{2} \left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2 \qquad \text{and} \qquad \left\|\nabla^2 f_{\mathrm{clean}}\left(\boldsymbol{X}\right)\right\| \leq \frac{5}{2}\sigma_{\max} \tag{62}$$

for all $\boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{H}_Y - \boldsymbol{Z}$, with $\boldsymbol{H}_Y := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{Y}\boldsymbol{R} - \boldsymbol{Z}\right\|_{\mathrm{F}}$, satisfying:

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_{2,\infty} \leq \epsilon \left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{63a}$$

$$\left\|\boldsymbol{Z} - \boldsymbol{X}^\star\right\| \leq \delta \|\boldsymbol{X}^\star\|, \tag{63b}$$

where $\epsilon \ll 1/\sqrt{\kappa^3 \mu r \log^2 n}$ and $\delta \ll 1/\kappa$.

Proof. See Appendix B.1. □

Lemma 7 reveals that the Hessian matrix is well-conditioned in a neighborhood close to $\boldsymbol{X}^\star$ that remains incoherent measured in the ℓ_2/ℓ_∞ norm (cf. (63a)), and along directions that point towards points which are not far away from the truth in the spectral norm (cf. (63b)).

Remark 5. The second condition (63b) is characterized using the spectral norm $\|\cdot\|$, while in previous works this is typically presented in the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. It is also worth noting that the Hessian matrix — even in the infinite-sample and noiseless case — is rank-deficient and cannot be positive definite. As a result, we resort to the form of strong convexity by restricting attention to certain directions (see the conditions on $\boldsymbol{V}$).

7.1.2 Error contraction

Our goal is to demonstrate the error bounds (28) measured in three different norms. Notably, as long as the iterates satisfy (28) at the tth iteration, then $\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\|_{2,\infty}$ is sufficiently small. Under our sample complexity assumption, $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t$ satisfies the ℓ_2/ℓ_∞ condition (63a) required in Lemma 7. Consequently, we can invoke Lemma 7 to arrive at the following error contraction result.

Lemma 8 (Contraction w.r.t. the Frobenius norm). *Suppose that $n^2 p \geq C\kappa^3\mu^3 r^3 n \log^3 n$ for some sufficiently large constant $C > 0$, the noise satisfies (27). There exists an event that does not depend on t and has probability $1 - O(n^{-10})$, such that when it happens and (28a), (28b) hold for the tth iteration, one has*

$$\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq C_4\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} + C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$$

provided that $0 < \eta \leq 2/(25\kappa\sigma_{\max})$, $1 - (\sigma_{\min}/4)\cdot\eta \leq \rho < 1$, and C_1 is sufficiently large.

Proof. The proof is built upon Lemma 7. See Appendix B.2. □

Further, if the current iterate satisfies all three conditions in (28), then we can derive a stronger sense of error contraction, namely, contraction in terms of the spectral norm.

Lemma 9 (Contraction w.r.t. the spectral norm). *Suppose $n^2 p \geq C\kappa^3\mu^3 r^3 n \log^3 n$ for some sufficiently large constant $C > 0$ and the noise satisfies (27). There exists an event that does not depend on t and has probability $1 - O(n^{-10})$, such that when it happens and (28) holds for the tth iteration, one has*

$$\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^\star\right\| \leq C_9\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\| + C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\| \tag{64}$$

provided that $0 < \eta \leq 1/\left(2\sigma_{\max}\right)$ and $1 - (\sigma_{\min}/3)\cdot\eta \leq \rho < 1$.

Proof. The key observation is this: the iterate that proceeds according to the population-level gradient reduces the error w.r.t. $\|\cdot\|$, namely,

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \eta\nabla F_{\text{clean}}\big(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\big) - \boldsymbol{X}^\star\right\| < \left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|,$$

as long as $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t$ is sufficiently close to the truth. Notably, the orthonormal matrix $\widehat{\boldsymbol{H}}^t$ is still chosen to be the one that minimizes the $\|\cdot\|_{\text{F}}$ distance (as opposed to $\|\cdot\|$), which yields a symmetry property $\boldsymbol{X}^{\star\top}\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t = \big(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\big)^\top\boldsymbol{X}^\star$, crucial for our analysis. See Appendix B.3 for details. □

7.2 Step 2: introducing the leave-one-out sequences

In order to establish the incoherence properties (28b) for the entire trajectory, which is difficult to deal with directly due to the complicated statistical dependence, we introduce a collection of *leave-one-out* versions of $\{\boldsymbol{X}^t\}_{t\geq 0}$, denoted by $\big\{\boldsymbol{X}^{t,(l)}\big\}_{t\geq 0}$ for each $1\leq l\leq n$. Specifically, $\big\{\boldsymbol{X}^{t,(l)}\big\}_{t\geq 0}$ is the iterates of gradient descent operating on the auxiliary loss function

$$f^{(l)}\left(\boldsymbol{X}\right) := \frac{1}{4p}\left\|\mathcal{P}_{\Omega^{-l}}\left[\boldsymbol{X}\boldsymbol{X}^\top - (\boldsymbol{M}^\star+\boldsymbol{E})\right]\right\|_{\text{F}}^2 + \frac{1}{4}\left\|\mathcal{P}_l\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\right\|_{\text{F}}^2. \tag{65}$$

Here, $\mathcal{P}_{\Omega_l}$ (resp. $\mathcal{P}_{\Omega^{-l}}$ and $\mathcal{P}_l$) represents the orthogonal projection onto the subspace of matrices which vanish outside of the index set $\Omega_l := \{(i,j)\in\Omega \mid i=l \text{ or } j=l\}$ (resp. $\Omega^{-l} := \{(i,j)\in\Omega \mid i\neq l, j\neq l\}$ and $\{(i,j) \mid i=l \text{ or } j=l\}$); that is, for any matrix $\boldsymbol{M}$,

$$\left[\mathcal{P}_{\Omega_l}\left(\boldsymbol{M}\right)\right]_{i,j} = \begin{cases} M_{i,j}, & \text{if } (i=l \text{ or } j=l) \text{ and } (i,j)\in\Omega, \\ 0, & \text{else,}\end{cases} \tag{66}$$

$$\left[\mathcal{P}_{\Omega^{-l}}\left(\boldsymbol{M}\right)\right]_{i,j} = \begin{cases} M_{i,j}, & \text{if } i\neq l \text{ and } j\neq l \text{ and } (i,j)\in\Omega \\ 0, & \text{else}\end{cases} \quad \text{and} \quad \left[\mathcal{P}_{l}\left(\boldsymbol{M}\right)\right]_{i,j} = \begin{cases} 0, & \text{if } i\neq l \text{ and } j\neq l, \\ M_{i,j}, & \text{if } i=l \text{ or } j=l.\end{cases} \tag{67}$$

The gradient of the leave-one-out loss function (65) is given by

$$\nabla f^{(l)}\left(\boldsymbol{X}\right) = \frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left[\boldsymbol{X}\boldsymbol{X}^\top - (\boldsymbol{M}^\star+\boldsymbol{E})\right]\boldsymbol{X} + \mathcal{P}_l\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\boldsymbol{X}. \tag{68}$$

The full algorithm to obtain the leave-one-out sequence $\{\boldsymbol{X}^{t,(l)}\}_{t\geq 0}$ (including spectral initialization) is summarized in Algorithm 5.

Algorithm 5 The lth leave-one-out sequence for matrix completion

Input: $\boldsymbol{Y} = [Y_{i,j}]_{1\leq i,j\leq n}, \boldsymbol{M}^\star_{\cdot,l}, \boldsymbol{M}^\star_{l,\cdot}, r, p$.

Spectral initialization: Let $\boldsymbol{U}^{0,(l)}\boldsymbol{\Sigma}^{(l)}\boldsymbol{U}^{0,(l)\top}$ be the top-r eigendecomposition of

$$\boldsymbol{M}^{(l)} := \frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left(\boldsymbol{Y}\right) + \mathcal{P}_l\left(\boldsymbol{M}^\star\right) = \frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left(\boldsymbol{M}^\star+\boldsymbol{E}\right) + \mathcal{P}_l\left(\boldsymbol{M}^\star\right)$$

with $\mathcal{P}_{\Omega^{-l}}$ and $\mathcal{P}_l$ defined in (67), and set $\boldsymbol{X}^{0,(l)} = \boldsymbol{U}^{0,(l)}\big(\boldsymbol{\Sigma}^{(l)}\big)^{1/2}$.

Gradient updates: **for** $t=0,1,2,\ldots,T-1$ **do**

$$\boldsymbol{X}^{t+1,(l)} = \boldsymbol{X}^{t,(l)} - \eta_t\nabla f^{(l)}\big(\boldsymbol{X}^{t,(l)}\big). \tag{69}$$

Remark 6. Rather than simply dropping all samples in the lth row/column, we replace the lth row/column with their respective population means. In other words, the leave-one-out gradient forms an unbiased surrogate for the true gradient, which is particularly important in ensuring high estimation accuracy.

7.3 Step 3: establishing the incoherence condition by induction

We will continue the proof of Theorem 2 in an inductive manner. As seen in Section 7.1.2, the induction hypotheses (28a) and (28c) hold for the $(t+1)$th iteration as long as (28) holds at the tth iteration. Therefore, we are left with proving the incoherence hypothesis (28b) for all $0 \le t \le T = O(n^5)$. For clarity of analysis, it is crucial to maintain a list of induction hypotheses, which includes a few more hypotheses that complement (28), and is given below.

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \left(C_4\rho^t\mu r\frac{1}{\sqrt{np}} + C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{70a}$$

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{2,\infty} \le \left(C_5\rho^t\mu r\sqrt{\frac{\log n}{np}} + C_8\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{70b}$$

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\| \le \left(C_9\rho^t\mu r\frac{1}{\sqrt{np}} + C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|, \tag{70c}$$

$$\max_{1\le l\le n}\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \le \left(C_3\rho^t\mu r\sqrt{\frac{\log n}{np}} + C_7\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{70d}$$

$$\max_{1\le l\le n}\left\|\left(\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2 \le \left(C_2\rho^t\mu r\frac{1}{\sqrt{np}} + C_6\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \tag{70e}$$

hold for some absolute constants $0 < \rho < 1$ and $C_1, \cdots, C_{10} > 0$. Here, $\widehat{\boldsymbol{H}}^{t,(l)}$ and $\boldsymbol{R}^{t,(l)}$ are orthonormal matrices defined by

$$\widehat{\boldsymbol{H}}^{t,(l)} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{71}$$

$$\boldsymbol{R}^{t,(l)} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R} - \boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\right\|_{\mathrm{F}}. \tag{72}$$

Clearly, the first three hypotheses (70a)-(70c) constitute the conclusion of Theorem 2, i.e. (28). The last two hypotheses (70d) and (70e) are auxiliary properties connecting the true iterates and the auxiliary leave-one-out sequences. Moreover, we summarize below several immediate consequences of (70), which will be useful throughout.

Lemma 10. *Suppose $n^2p \ge C\kappa^3\mu^2r^2n\log n$ for some sufficiently large constant $C > 0$ and the noise satisfies* (27)*. Under the hypotheses* (70)*, one has*

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}\right\|_{\mathrm{F}} \le 5\kappa\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}, \tag{73a}$$

$$\left\|\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \left\{2C_4\rho^t\mu r\frac{1}{\sqrt{np}} + 2C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{73b}$$

$$\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star\right\|_{2,\infty} \le \left\{\left(C_3+C_5\right)\rho^t\mu r\sqrt{\frac{\log n}{np}} + \left(C_8+C_7\right)\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right\}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{73c}$$

$$\left\|\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \boldsymbol{X}^\star\right\| \le \left\{2C_9\rho^t\mu r\frac{1}{\sqrt{np}} + 2C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^\star\right\|. \tag{73d}$$

In particular, (73a) *follows from hypotheses* (70c) *and* (70d)*.*

Proof. See Appendix B.4. □

In the sequel, we follow the general recipe outlined in Section 5 to establish the induction hypotheses. We only need to establish (70b), (70d) and (70e) for the $(t+1)$th iteration, since (70a) and (70c) have been established in Section 7.1.2. Specifically, we resort to the leave-one-out iterates by showing that: first, the true and the auxiliary iterates remain exceedingly close throughout; second, the lth leave-one-out sequence stays incoherent with $\boldsymbol{e}_l$ due to statistical independence.

- **Step 3(a): proximity between the original and the leave-one-out iterates.** We demonstrate that $\boldsymbol{X}^{t+1}$ is well approximated by $\boldsymbol{X}^{t+1,(l)}$, up to proper orthonormal transforms. This is precisely the induction hypothesis (70d) for the $(t+1)$th iteration.

  Lemma 11. *Suppose the sample complexity satisfies $n^2 p \geq C\kappa^4\mu^3 r^3 n\log^3 n$ for some sufficiently large constant $C>0$ and the noise satisfies (27). Let $\mathcal{E}_t$ be the event where the hypotheses in (70) hold for the tth iteration. Then on some event $\mathcal{E}_{t+1,1}\subseteq\mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t\cap\mathcal{E}_{t+1,1}^c)=O(n^{-10})$, we have*

$$\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}}\leq C_3\rho^{t+1}\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+C_7\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \tag{74}$$

  provided that $0<\eta\leq 2/(25\kappa\sigma_{\max})$, $1-(\sigma_{\min}/5)\cdot\eta\leq\rho<1$ and $C_7>0$ is sufficiently large.

  Proof. The fact that this difference is well-controlled relies heavily on the benign geometric property of the Hessian revealed by Lemma 7. Two important remarks are in order: (1) both points $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t$ and $\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}$ satisfy (63a); (2) the difference $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}$ forms a valid direction for restricted strong convexity. These two properties together allow us to invoke Lemma 7. See Appendix B.5. □

- **Step 3(b): incoherence of the leave-one-out iterates.** Given that $\boldsymbol{X}^{t+1,(l)}$ is sufficiently close to $\boldsymbol{X}^{t+1}$, we turn our attention to establishing the incoherence of this surrogate $\boldsymbol{X}^{t+1,(l)}$ w.r.t. $\boldsymbol{e}_l$. This amounts to proving the induction hypothesis (70e) for the $(t+1)$th iteration.

  Lemma 12. *Suppose the sample complexity meets $n^2 p\geq C\kappa^3\mu^3 r^3 n\log^3 n$ for some sufficiently large constant $C>0$ and the noise satisfies (27). Let $\mathcal{E}_t$ be the event where the hypotheses in (70) hold for the tth iteration. Then on some event $\mathcal{E}_{t+1,2}\subseteq\mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t\cap\mathcal{E}_{t+1,2}^c)=O(n^{-10})$, we have*

$$\left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}-\boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2\leq C_2\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+C_6\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \tag{75}$$

  so long as $0<\eta\leq 1/\sigma_{\max}$, $1-(\sigma_{\min}/3)\cdot\eta\leq\rho<1$, $C_2\gg\kappa C_9$ and $C_6\gg\kappa C_{10}/\sqrt{\log n}$.

  Proof. The key observation is that $\boldsymbol{X}^{t+1,(l)}$ is statistically independent from any sample in the lth row/column of the matrix. Since there are an order of np samples in each row/column, we obtain enough information that helps establish the desired incoherence property. See Appendix B.6. □

- **Step 3(c): combining the bounds.** The inequalities (70d) and (70e) taken collectively allow us to establish the induction hypothesis (70b). Specifically, for every $1\leq l\leq n$, write

$$\left(\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^\star\right)_{l,\cdot}=\left(\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}\right)_{l,\cdot}+\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}-\boldsymbol{X}^\star\right)_{l,\cdot},$$

  and the triangle inequality gives

$$\left\|\left(\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2\leq\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}\right\|_{\mathrm{F}}+\left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}-\boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2. \tag{76}$$

  The second term has already been bounded by (75). Since we have established the induction hypotheses (70c) and (70d) for the $(t+1)$th iteration, the first term can be bounded by (73a) for the $(t+1)$th iteration, i.e.

$$\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}\right\|_{\mathrm{F}}\leq 5\kappa\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}}.$$

  Plugging the above inequality, (74) and (75) into (76), we have

$$\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^\star\right\|_{2,\infty}\leq 5\kappa\left(C_3\rho^{t+1}\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+\frac{C_7}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\right)$$

$$+ C_2\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} + \frac{C_6}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}$$
$$\le C_5\rho^{t+1}\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} + \frac{C_8}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}$$

as long as $C_5/(\kappa C_3+C_2)$ and $C_8/(\kappa C_7+C_6)$ are sufficiently large. This establishes the induction hypothesis (70b). From the deduction above we see $\mathcal{E}_t \cap \mathcal{E}_{t+1}^c = O(n^{-10})$ and thus finish the proof.

7.4 The base case: spectral initialization

Finally, we return to check the base case, namely, we aim to show that the spectral initialization satisfies the induction hypotheses (70a)-(70e) for $t = 0$. This is accomplished via the following lemma.

Lemma 13. *Suppose the sample size obeys $n^2p \ge C\mu^2r^2n\log n$ for some sufficiently large constant $C > 0$, the noise satisfies (27), and $\kappa = \sigma_{\max}/\sigma_{\min} \asymp 1$. Then with probability at least $1 - O\left(n^{-10}\right)$, the claims in* (70a)-(70e) *hold simultaneously for $t = 0$.*

Proof. This follows by invoking the Davis-Kahan sinΘ theorem [DK70] as well as the entrywise eigenvector perturbation analysis in [AFWZ17]. We defer the proof to Appendix B.7. □

8 Analysis for blind deconvolution

In this section, we instantiate the general recipe presented in Section 5 to blind deconvolution and prove Theorem 3. Without loss of generality, we assume throughout that $\left\|\boldsymbol{h}^\star\right\|_2 = \left\|\boldsymbol{x}^\star\right\|_2 = 1$.

Before presenting the analysis, we first gather some simple facts about the empirical loss function in (32). Recall the definition of $\boldsymbol{z}$ in (33), and for notational simplicity, we write $f(\boldsymbol{z}) = f(\boldsymbol{h},\boldsymbol{x})$. Since $\boldsymbol{z}$ is complex-valued, we need to resort to Wirtinger calculus; see [CLS15, Section 6] for a brief introduction. The Wirtinger gradient of (32) with respect to $\boldsymbol{h}$ and $\boldsymbol{x}$ are given respectively by

$$\nabla_{\boldsymbol{h}} f(\boldsymbol{z}) = \nabla_{\boldsymbol{h}} f(\boldsymbol{h},\boldsymbol{x}) = \sum_{j=1}^m \left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j - y_j\right)\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}; \tag{77}$$

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{z}) = \nabla_{\boldsymbol{x}} f(\boldsymbol{h},\boldsymbol{x}) = \sum_{j=1}^m \overline{(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j - y_j)}\boldsymbol{a}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}. \tag{78}$$

It is worth noting that the formal Wirtinger gradient contains $\nabla_{\overline{\boldsymbol{h}}} f(\boldsymbol{h},\boldsymbol{x})$ and $\nabla_{\overline{\boldsymbol{x}}} f(\boldsymbol{h},\boldsymbol{x})$ as well. Nevertheless, since $f(\boldsymbol{h},\boldsymbol{x})$ is a real-valued function, the following identities always hold

$$\nabla_{\boldsymbol{h}} f(\boldsymbol{h},\boldsymbol{x}) = \overline{\nabla_{\overline{\boldsymbol{h}}} f(\boldsymbol{h},\boldsymbol{x})} \qquad \text{and} \qquad \nabla_{\boldsymbol{x}} f(\boldsymbol{h},\boldsymbol{x}) = \overline{\nabla_{\overline{\boldsymbol{x}}} f(\boldsymbol{h},\boldsymbol{x})}.$$

In light of these observations, one often omits the gradient with respect to the conjugates; correspondingly, the gradient update rule (35) can be written as

$$\boldsymbol{h}^{t+1} = \boldsymbol{h}^t - \frac{\eta}{\left\|\boldsymbol{x}^t\right\|_2^2}\sum_{j=1}^m \left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}}\boldsymbol{a}_j - y_j\right)\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^t, \tag{79a}$$

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{\left\|\boldsymbol{h}^t\right\|_2^2}\sum_{j=1}^m \overline{(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}}\boldsymbol{a}_j - y_j)}\boldsymbol{a}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^t. \tag{79b}$$

We can also compute the Wirtinger Hessian of $f(\boldsymbol{z})$ as follows,

$$\nabla^2 f(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^{\mathsf{H}} & \overline{\boldsymbol{A}} \end{bmatrix}, \tag{80}$$

where

$$\boldsymbol{A}=\begin{bmatrix}\sum_{j=1}^m \left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}\right|^2 \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}} & \sum_{j=1}^m \left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j-y_j\right)\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\\ \sum_{j=1}^m \left[\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j-y_j\right)\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\right]^{\mathsf{H}} & \sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right|^2\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\end{bmatrix}\in\mathbb{C}^{2K\times 2K};$$

$$\boldsymbol{B}=\begin{bmatrix}\boldsymbol{0} & \sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\left(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}\right)^{\top}\\ \sum_{j=1}^m \boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}\left(\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right)^{\top} & \boldsymbol{0}\end{bmatrix}\in\mathbb{C}^{2K\times 2K}.$$

Last but not least, we say $(\boldsymbol{h}_1,\boldsymbol{x}_1)$ is aligned with $(\boldsymbol{h}_2,\boldsymbol{x}_2)$, if the following holds,

$$\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2+\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2=\min_{\alpha\in\mathbb{C}}\left\{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2^2+\|\alpha\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2\right\}.$$

To simplify notations, define $\widetilde{\boldsymbol{z}}^t$ as

$$\widetilde{\boldsymbol{z}}^t=\begin{bmatrix}\widetilde{\boldsymbol{h}}^t\\ \widetilde{\boldsymbol{x}}^t\end{bmatrix}:=\begin{bmatrix}\frac{1}{\overline{\alpha}^t}\boldsymbol{h}^t\\ \alpha^t\boldsymbol{x}^t\end{bmatrix} \tag{81}$$

with the alignment parameter α^t given in (38). Then we can see that $\widetilde{\boldsymbol{z}}^t$ is aligned with $\boldsymbol{z}^\star$ and

$$\operatorname{dist}\left(\boldsymbol{z}^t,\boldsymbol{z}^\star\right)=\operatorname{dist}\left(\widetilde{\boldsymbol{z}}^t,\boldsymbol{z}^\star\right)=\left\|\widetilde{\boldsymbol{z}}^t-\boldsymbol{z}^\star\right\|_2.$$

8.1 Step 1: characterizing local geometry in the RIC

8.1.1 Local geometry

The first step is to characterize the region of incoherence and contraction (RIC), where the empirical loss function enjoys restricted strong convexity and smoothness properties. To this end, we have the following lemma.

Lemma 14 (Restricted strong convexity and smoothness for blind deconvolution)**.** *Let $c>0$ be a sufficiently small constant and*

$$\delta=c/\log^2 m.$$

Suppose the sample size satisfies $m\geq c_0\mu^2K\log^9 m$ for some sufficiently large constant $c_0>0$. Then with probability $1-O\left(m^{-10}+e^{-K}\log m\right)$, the Wirtinger Hessian $\nabla^2 f(\boldsymbol{z})$ obeys

$$\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f(\boldsymbol{z})+\nabla^2 f(\boldsymbol{z})\,\boldsymbol{D}\right]\boldsymbol{u}\geq(1/4)\cdot\|\boldsymbol{u}\|_2^2\qquad\text{and}\qquad\left\|\nabla^2 f(\boldsymbol{z})\right\|\leq 3$$

simultaneously for all

$$\boldsymbol{z}=\begin{bmatrix}\boldsymbol{h}\\ \boldsymbol{x}\end{bmatrix}\qquad\text{and}\qquad\boldsymbol{u}=\begin{bmatrix}\boldsymbol{h}_1-\boldsymbol{h}_2\\ \boldsymbol{x}_1-\boldsymbol{x}_2\\ \overline{\boldsymbol{h}_1-\boldsymbol{h}_2}\\ \overline{\boldsymbol{x}_1-\boldsymbol{x}_2}\end{bmatrix}\qquad\text{and}\qquad\boldsymbol{D}=\begin{bmatrix}\gamma_1\boldsymbol{I}_K & & & \\ & \gamma_2\boldsymbol{I}_K & & \\ & & \gamma_1\boldsymbol{I}_K & \\ & & & \gamma_2\boldsymbol{I}_K\end{bmatrix},$$

where $\boldsymbol{z}$ satisfies

$$\max\left\{\|\boldsymbol{h}-\boldsymbol{h}^\star\|_2,\|\boldsymbol{x}-\boldsymbol{x}^\star\|_2\right\}\leq\delta; \tag{82a}$$

$$\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\boldsymbol{x}-\boldsymbol{x}^\star)\right|\leq 2C_3\frac{1}{\log^{3/2}m}; \tag{82b}$$

$$\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right|\leq 2C_4\frac{\mu}{\sqrt{m}}\log^2 m; \tag{82c}$$

$(\boldsymbol{h}_1,\boldsymbol{x}_1)$ is aligned with $(\boldsymbol{h}_2,\boldsymbol{x}_2)$, and they satisfy

$$\max\left\{\|\boldsymbol{h}_1-\boldsymbol{h}^\star\|_2,\|\boldsymbol{h}_2-\boldsymbol{h}^\star\|_2,\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|_2,\|\boldsymbol{x}_2-\boldsymbol{x}^\star\|_2\right\}\leq\delta; \tag{83}$$

and finally, $\boldsymbol{D}$ satisfies for $\gamma_1,\gamma_2\in\mathbb{R}$,

$$\max\left\{|\gamma_1-1|,|\gamma_2-1|\right\}\leq\delta. \tag{84}$$

Here, $C_3,C_4>0$ are numerical constants.

Proof. See Appendix C.1. □

Lemma 14 characterizes the restricted strong convexity and smoothness of the loss function used in blind deconvolution. To the best of our knowledge, this provides the first characterization regarding geometric properties of the Hessian matrix for blind deconvolution. A few interpretations are in order.

- The conditions (82) specify the region of incoherence and contraction (RIC). In particular, (82a) specifies a neighborhood that is close to the ground truth in ℓ_2 norm, and (82b) and (82c) specify the incoherence region with respect to the sensing vectors $\{\boldsymbol{a}_j\}$ and $\{\boldsymbol{b}_j\}$, respectively.
- Similar to matrix completion, the Hessian matrix is rank-deficient even at the population level. Consequently, we resort to a restricted form of strong convexity by focusing on certain directions. More specifically, these directions can be viewed as the difference between two pre-aligned points that are not far from the truth, which is characterized by (83).
- Finally, the diagonal matrix $\boldsymbol{D}$ accounts for scaling factors that are not too far from 1 (see (84)), which allows us to account for different step sizes employed for $\boldsymbol{h}$ and $\boldsymbol{x}$.

8.1.2 Error contraction

The restricted strong convexity and smoothness allow us to establish the contraction of the error measured in terms of $\mathrm{dist}(\cdot, \boldsymbol{z}^\star)$ as defined in (34) as long as the iterates stay in the RIC.

Lemma 15. *Suppose the number of measurements satisfies $m \geq C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$, and the step size $\eta > 0$ is some sufficiently small constant. There exists an event that does not depend on t and has probability $1 - O\left(m^{-10} + e^{-K}\log m\right)$, such that when it happens and*

$$\mathrm{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \leq \xi, \tag{85a}$$

$$\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \leq C_3 \frac{1}{\log^{1.5} m}, \tag{85b}$$

$$\max_{1\leq j\leq m} \left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \leq C_4 \frac{\mu}{\sqrt{m}} \log^2 m \tag{85c}$$

hold for some constants $C_3, C_4 > 0$, one has

$$\mathrm{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^\star\right) \leq (1 - \eta/16)\,\mathrm{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right).$$

Here, $\widetilde{\boldsymbol{h}}^t$ and $\widetilde{\boldsymbol{x}}^t$ are defined in (81), and $\xi \ll 1/\log^2 m$.

Proof. See Appendix C.2. □

As a result, if $\boldsymbol{z}^t$ satisfies the condition (85) for all $0 \leq t \leq T$, then

$$\mathrm{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \leq \rho\,\mathrm{dist}\left(\boldsymbol{z}^{t-1}, \boldsymbol{z}^\star\right) \leq \rho^t \mathrm{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) \leq \rho^t c_1, \qquad 0 < t \leq T,$$

where $\rho := 1 - \eta/16$. Furthermore, similar to the case of phase retrieval (i.e. Lemma 3), as soon as we demonstrate that the conditions (85) hold for all $0 \leq t \leq m$, then Theorem 3 holds true. The proof of this claim is exactly the same as for Lemma 3, and is thus omitted for conciseness. In what follows, we focus on establishing (85) for all $0 \leq t \leq m$.

Before concluding this subsection, we make note of another important result that concerns the alignment parameter α^t, which will be useful in the subsequent analysis. Specifically, the alignment parameter sequence $\{\alpha^t\}$ converges linearly to a constant whose magnitude is fairly close to 1, as long as the two initial vectors $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ have similar ℓ_2 norms and are close to the truth. Given that α^t determines the global scaling of the iterates, this reveals rapid convergence of both $\|\boldsymbol{h}^t\|_2$ and $\|\boldsymbol{x}^t\|_2$, which explains why there is no need to impose extra terms to regularize the ℓ_2 norm as employed in [LLSW18, HH17].

Lemma 16. *When $m > 1$ is sufficiently large, the following two claims hold true.*

- *If* $\big||\alpha^t|-1\big| \le 1/2$ *and* $\mathrm{dist}(\boldsymbol{z}^t, \boldsymbol{z}^\star) \le C_1/\log^2 m$, *then*

$$\left|\frac{\alpha^{t+1}}{\alpha^t} - 1\right| \le c\,\mathrm{dist}(\boldsymbol{z}^t, \boldsymbol{z}^\star) \le \frac{cC_1}{\log^2 m}$$

for some absolute constant $c > 0$;

- *If* $\big||\alpha^0|-1\big| \le 1/4$ *and* $\mathrm{dist}(\boldsymbol{z}^s, \boldsymbol{z}^\star) \le C_1(1-\eta/16)^s/\log^2 m$ *for all* $0 \le s \le t$, *then one has*

$$\big||\alpha^{s+1}|-1\big| \le 1/2, \qquad 0 \le s \le t.$$

Proof. See Appendix C.2. □

The initial condition $\big||\alpha^0|-1\big| < 1/4$ will be guaranteed to hold with high probability by Lemma 19.

8.2 Step 2: introducing the leave-one-out sequences

As demonstrated by the assumptions in Lemma 15, the key is to show that the whole trajectory lies in the region specified by (85a)-(85c). Once again, the difficulty lies in the statistical dependency between the iterates $\{\boldsymbol{z}^t\}$ and the measurement vectors $\{\boldsymbol{a}_j\}$. We follow the general recipe and introduce the *leave-one-out* sequences, denoted by $\big\{\boldsymbol{h}^{t,(l)}, \boldsymbol{x}^{t,(l)}\big\}_{t\ge 0}$ for each $1 \le l \le m$. Specifically, $\big\{\boldsymbol{h}^{t,(l)}, \boldsymbol{x}^{t,(l)}\big\}_{t\ge 0}$ is the gradient sequence operating on the loss function

$$f^{(l)}(\boldsymbol{h}, \boldsymbol{x}) := \sum_{j:j\neq l} \left|\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_j\right|^2. \tag{86}$$

The whole sequence is constructed by running gradient descent with spectral initialization on the leave-one-out loss (86). The precise description is supplied in Algorithm 6.

For notational simplicity, we denote $\boldsymbol{z}^{t,(l)} = \begin{bmatrix} \boldsymbol{h}^{t,(l)} \\ \boldsymbol{x}^{t,(l)} \end{bmatrix}$ and use $f(\boldsymbol{z}^{t,(l)}) = f(\boldsymbol{h}^{t,(l)}, \boldsymbol{x}^{t,(l)})$ interchangeably. Define similarly the alignment parameters

$$\alpha^{t,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}^{t,(l)} - \boldsymbol{h}^\star\right\|_2^2 + \left\|\alpha\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right\|_2^2, \tag{87}$$

and denote $\widetilde{\boldsymbol{z}}^{t,(l)} = \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t,(l)} \\ \widetilde{\boldsymbol{x}}^{t,(l)} \end{bmatrix}$ where

$$\widetilde{\boldsymbol{h}}^{t,(l)} = \frac{1}{\overline{\alpha^{t,(l)}}}\boldsymbol{h}^{t,(l)} \qquad \text{and} \qquad \widetilde{\boldsymbol{x}}^{t,(l)} = \alpha^{t,(l)}\boldsymbol{x}^{t,(l)}. \tag{88}$$

Algorithm 6 The lth leave-one-out sequence for blind deconvolution

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m, j\neq l}$, $\{\boldsymbol{b}_j\}_{1\le j\le m, j\neq l}$ and $\{y_j\}_{1\le j\le m, j\neq l}$.

Spectral initialization: Let $\sigma_1(\boldsymbol{M}^{(l)})$, $\check{\boldsymbol{h}}^{0,(l)}$ and $\check{\boldsymbol{x}}^{0,(l)}$ be the leading singular value, left and right singular vectors of

$$\boldsymbol{M}^{(l)} := \sum_{j:j\neq l} y_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}},$$

respectively. Set $\boldsymbol{h}^{0,(l)} = \sqrt{\sigma_1(\boldsymbol{M}^{(l)})}\,\check{\boldsymbol{h}}^{0,(l)}$ and $\boldsymbol{x}^{0,(l)} = \sqrt{\sigma_1(\boldsymbol{M}^{(l)})}\,\check{\boldsymbol{x}}^{0,(l)}$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{bmatrix} \boldsymbol{h}^{t+1,(l)} \\ \boldsymbol{x}^{t+1,(l)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}^{t,(l)} \\ \boldsymbol{x}^{t,(l)} \end{bmatrix} - \eta \begin{bmatrix} \frac{1}{\|\boldsymbol{x}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{h}} f^{(l)}\big(\boldsymbol{h}^{t,(l)}, \boldsymbol{x}^{t,(l)}\big) \\ \frac{1}{\|\boldsymbol{h}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{x}} f^{(l)}\big(\boldsymbol{h}^{t,(l)}, \boldsymbol{x}^{t,(l)}\big) \end{bmatrix}. \tag{89}$$

8.3 Step 3: establishing the incoherence condition by induction

As usual, we continue the proof in an inductive manner. For clarity of presentation, we list below the set of induction hypotheses underlying our analysis:

$$\operatorname{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le C_1 \frac{1}{\log^2 m}, \tag{90a}$$

$$\max_{1\le l\le m} \operatorname{dist}\big(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^t\big) \le C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}}, \tag{90b}$$

$$\max_{1\le l\le m} \left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\big)\right| \le C_3 \frac{1}{\log^{1.5} m}, \tag{90c}$$

$$\max_{1\le l\le m} \left|\boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t\right| \le C_4 \frac{\mu}{\sqrt{m}} \log^2 m, \tag{90d}$$

where $\widetilde{\boldsymbol{h}}^t$, $\widetilde{\boldsymbol{x}}^t$ and $\widetilde{\boldsymbol{z}}^t$ are defined in (81). Here, $C_1, C_3 > 0$ are some sufficiently small constants, while $C_2, C_4 > 0$ are some sufficiently large constants. We aim to show that if these hypotheses (90) hold up to the tth iteration, then the same would hold for the $(t+1)$th iteration with exceedingly high probability (e.g. $1 - O(m^{-10})$). The first hypothesis (90a) has already been established in Lemma 15, and hence the rest of this section focuses on establishing the remaining three. To justify the incoherence hypotheses (90c) and (90d) for the $(t+1)$th iteration, we need to leverage the nice properties of the leave-one-out sequences, and establish (90b) first. In the sequel, we follow the steps suggested in the general recipe.

- **Step 3(a): proximity between the original and the leave-one-out iterates.** We first justify the hypothesis (90b) for the $(t+1)$th iteration via the following lemma.

  Lemma 17. *Suppose the sample complexity obeys $m \ge C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$. Let $\mathcal{E}_t$ be the event where the hypotheses (90a)-(90d) hold for the tth iteration. Then on an event $\mathcal{E}_{t+1,1} \subseteq \mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1,1}^c) = O(m^{-10} + me^{-cK})$ for some constant $c > 0$, one has*

$$\max_{1\le l\le m} \operatorname{dist}\big(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\big) \le C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}}$$

$$\text{and} \qquad \max_{1\le l\le m} \big\|\widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1}\big\|_2 \lesssim C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}},$$

  provided that the step size $\eta > 0$ is some sufficiently small constant.

  Proof. As usual, this result follows from the restricted strong convexity, which forces the distance between the two sequences of interest to be contractive. See Appendix C.3. □

- **Step 3(b): incoherence of the leave-one-out iterate $\boldsymbol{x}^{t+1,(l)}$ w.r.t. $\boldsymbol{a}_l$.** Next, we show that the leave-one-out iterate $\widetilde{\boldsymbol{x}}^{t+1,(l)}$ — which is independent of $\boldsymbol{a}_l$ — is incoherent w.r.t. $\boldsymbol{a}_l$ in the sense that

$$\left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star\big)\right| \le 10 C_1 \frac{1}{\log^{3/2} m} \tag{91}$$

  with probability exceeding $1 - O\left(m^{-10} + e^{-K}\log m\right)$. To see why, use the statistical independence and the standard Gaussian concentration inequality to show that

$$\max_{1\le l\le m} \left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star\big)\right| \le 5\sqrt{\log m} \max_{1\le l\le m} \big\|\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star\big\|_2$$

  with probability exceeding $1 - O(m^{-10})$. It then follows from the triangle inequality that

$$\big\|\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star\big\|_2 \le \big\|\widetilde{\boldsymbol{x}}^{t+1,(l)} - \widetilde{\boldsymbol{x}}^{t+1}\big\|_2 + \big\|\widetilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^\star\big\|_2$$

$$\overset{\text{(i)}}{\leq} CC_2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}}+C_1\frac{1}{\log^2 m}$$
$$\overset{\text{(ii)}}{\leq} 2C_1\frac{1}{\log^2 m},$$

where (i) follows from Lemmas 15 and 17, and (ii) holds as soon as $m/(\mu^2\sqrt{K}\log^{13/2} m)$ is sufficiently large. Combining the preceding two bounds establishes (91).

- **Step 3(c): combining the bounds to show incoherence of $\boldsymbol{x}^{t+1}$ w.r.t. $\{\boldsymbol{a}_l\}$.** The above bounds immediately allow us to conclude that

$$\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^{t+1}-\boldsymbol{x}^\star\big)\right|\leq C_3\frac{1}{\log^{3/2} m}$$

with probability at least $1-O\left(m^{-10}+e^{-K}\log m\right)$, which is exactly the hypothesis (90c) for the $(t+1)$th iteration. Specifically, for each $1\leq l\leq m$, the triangle inequality yields

$$\begin{aligned}\left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^{t+1}-\boldsymbol{x}^\star\big)\right| &\leq \left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^{t+1}-\widetilde{\boldsymbol{x}}^{t+1,(l)}\big)\right|+\left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\big)\right|\\ &\overset{\text{(i)}}{\leq} \|\boldsymbol{a}_l\|_2\left\|\widetilde{\boldsymbol{x}}^{t+1}-\widetilde{\boldsymbol{x}}^{t+1,(l)}\right\|_2+\left|\boldsymbol{a}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\big)\right|\\ &\overset{\text{(ii)}}{\leq} 3\sqrt{K}\cdot CC_2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}+10C_1\frac{1}{\log^{3/2} m}\\ &\overset{\text{(iii)}}{\leq} C_3\frac{1}{\log^{3/2} m}.\end{aligned}$$

Here (i) follows from Cauchy-Schwarz, (ii) is a consequence of (190), Lemma 17 and the bound (91), and the last inequality holds as long as $m/(\mu^2K\log^6 m)$ is sufficiently large and $C_3\geq 11C_1$.

- **Step 3(d): incoherence of $\boldsymbol{h}^{t+1}$ w.r.t. $\{\boldsymbol{b}_l\}$.** It remains to justify that $\boldsymbol{h}^{t+1}$ is also incoherent w.r.t. its associated design vectors $\{\boldsymbol{b}_l\}$. This proof of this step, however, is much more involved and challenging, due to the deterministic nature of the $\boldsymbol{b}_l$'s. As a result, we would need to "propagate" the randomness brought about by $\{\boldsymbol{a}_l\}$ to $\boldsymbol{h}^{t+1}$ in order to facilitate the analysis. The result is summarized as follows.

Lemma 18. *Suppose that the sample complexity obeys $m\geq C\mu^2K\log^9 m$ for some sufficiently large constant $C>0$. Let $\mathcal{E}_t$ be the event where the hypotheses (90a)-(90d) hold for the tth iteration. Then on an event $\mathcal{E}_{t+1,2}\subseteq\mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t\cap\mathcal{E}_{t+1,2}^c)=O(m^{-10})$, one has*

$$\max_{1\leq l\leq m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t+1}\right|\leq C_4\frac{\mu}{\sqrt{m}}\log^2 m$$

as long as C_4 is sufficiently large, and $\eta>0$ is taken to be some sufficiently small constant.

Proof. The key idea is to divide $\{1,\cdots,m\}$ into consecutive bins each of size $\operatorname{poly}\log(m)$, and to exploit the randomness (namely, the randomness from $\boldsymbol{a}_l$) within each bin. This binning idea is crucial in ensuring that the incoherence measure of interest does not blow up as t increases. See Appendix C.4. □

With these steps in place, we conclude the proof of Theorem 3 via induction and the union bound.

8.4 The base case: spectral initialization

In order to finish the induction steps, we still need to justify the induction hypotheses for the base cases, namely, we need to show that the spectral initializations $\boldsymbol{z}^0$ and $\left\{\boldsymbol{z}^{0,(l)}\right\}_{1\leq l\leq m}$ satisfy the induction hypotheses (90) at $t=0$.

To start with, the initializations are sufficiently close to the truth when measured by the ℓ_2 norm, as summarized by the following lemma.

Lemma 19. *Fix any small constant $\xi > 0$. Suppose the sample size obeys $m \geq C\mu^2 K \log^2 m/\xi^2$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O(m^{-10})$, we have*

$$\min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{ \left\|\alpha \boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha \boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2 \right\} \leq \xi \qquad \text{and} \tag{92}$$

$$\min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{ \left\|\alpha \boldsymbol{h}^{0,(l)} - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha \boldsymbol{x}^{0,(l)} - \boldsymbol{x}^\star\right\|_2 \right\} \leq \xi, \qquad 1 \leq l \leq m, \tag{93}$$

and $||\alpha_0| - 1| \leq 1/4$.

Proof. This follows from Wedin's sinΘ theorem [Wed72] and [LLSW18, Lemma 5.20]. See Appendix C.5. □

From the definition of $\mathrm{dist}(\cdot,\cdot)$ (cf. (34)), we immediately have

$$\begin{aligned}\mathrm{dist}\big(\boldsymbol{z}^0, \boldsymbol{z}^\star\big) &= \min_{\alpha\in\mathbb{C}} \sqrt{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2^2 + \|\alpha\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2} \overset{\text{(i)}}{\leq} \min_{\alpha\in\mathbb{C}} \left\{ \left\|\frac{1}{\overline{\alpha}}\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2 + \|\alpha\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \right\} \\ &\overset{\text{(ii)}}{\leq} \min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{ \left\|\alpha \boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha \boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2 \right\} \overset{\text{(iii)}}{\leq} C_1 \frac{1}{\log^2 m},\end{aligned} \tag{94}$$

as long as $m \geq C\mu^2 K \log^6 m$ for some sufficiently large constant $C > 0$. Here (i) follows from the elementary inequality that $a^2 + b^2 \leq (a+b)^2$ for positive a and b, (ii) holds since the feasible set of the latter one is strictly smaller, and (iii) follows directly from Lemma 19. This finishes the proof of (90a) for $t = 0$. Similarly, with high probability we have

$$\mathrm{dist}\big(\boldsymbol{z}^{0,(l)}, \boldsymbol{z}^\star\big) \leq \min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{ \left\|\alpha \boldsymbol{h}^{0,(l)} - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha \boldsymbol{x}^{0,(l)} - \boldsymbol{x}^\star\right\|_2 \right\} \lesssim \frac{1}{\log^2 m}, \quad 1 \leq l \leq m. \tag{95}$$

Next, when properly aligned, the true initial estimate $\boldsymbol{z}^0$ and the leave-one-out estimate $\boldsymbol{z}^{0,(l)}$ are expected to be sufficiently close, as claimed by the following lemma. Along the way, we show that $\boldsymbol{h}^0$ is incoherent w.r.t. the sampling vectors $\{\boldsymbol{b}_l\}$. This establishes (90b) and (90d) for $t = 0$.

Lemma 20. *Suppose that $m \geq C\mu^2 K \log^3 m$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O(m^{-10})$, one has*

$$\max_{1\leq l\leq m} \mathrm{dist}\big(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^0\big) \leq C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}} \tag{96}$$

and

$$\max_{1\leq l\leq m} \left|\boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^0\right| \leq C_4 \frac{\mu \log^2 m}{\sqrt{m}}. \tag{97}$$

Proof. The key is to establish that $\mathrm{dist}\big(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^0\big)$ can be upper bounded by some linear scaling of $\left|\boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^0\right|$, and vice versa. This allows us to derive bounds simultaneously for both quantities. See Appendix C.6. □

Finally, we establish (90c) regarding the incoherence of $\boldsymbol{x}^0$ with respect to the design vectors $\{\boldsymbol{a}_l\}$.

Lemma 21. *Suppose that $m \geq C\mu^2 K \log^6 m$ for some sufficiently large constant $C > 0$. Then with probability exceeding $1 - O(m^{-10})$, we have*

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_l^{\mathsf{H}} \big(\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\big)\right| \leq C_3 \frac{1}{\log^{1.5} m}.$$

Proof. See Appendix C.7. □

9 Discussions

This paper showcases an important phenomenon in nonconvex optimization: even without explicit enforcement of regularization, the vanilla form of gradient descent effectively achieves implicit regularization for a large family of statistical estimation problems. We believe this phenomenon arises in problems far beyond the three cases studied herein, and our results are initial steps towards understanding this fundamental phenomenon. There are numerous avenues open for future investigation, and we point out a few of them.

- *Improving sample complexity.* In the current paper, the required sample complexity $O\left(\mu^3 r^3 n \log^3 n\right)$ for matrix completion is sub-optimal when the rank r of the underlying matrix is large. While this allows us to achieve a dimension-free iteration complexity, it is slightly higher than the sample complexity derived for regularized gradient descent in [CW15]. We expect our results continue to hold under lower sample complexity $O\left(\mu^2 r^2 n \log n\right)$, but it calls for a more refined analysis (e.g. a generic chaining argument).

- *Leave-one-out tricks for more general designs.* So far our focus is on independent designs, including the i.i.d. Gaussian design adopted in phase retrieval and partially in blind deconvolution, as well as the independent sampling mechanism in matrix completion. Such independence property creates some sort of "statistical homogeneity", for which the leave-one-out argument works beautifully. It remains unclear how to generalize such leave-one-out tricks for more general designs (e.g. more general sampling patterns in matrix completion and more structured Fourier designs in phase retrieval and blind deconvolution). In fact, the readers can already get a flavor of this issue in the analysis of blind deconvolution, where the Fourier design vectors require much more delicate treatments than purely Gaussian designs.

- *Uniform stability.* The leave-one-out perturbation argument is established upon a basic fact: when we exclude one sample from consideration, the resulting estimates/predictions do not deviate much from the original ones. This leave-one-out stability bears similarity to the notion of uniform stability studied in statistical learning theory [BE02]. We expect our analysis framework to be helpful for analyzing other learning algorithms that are uniformly stable.

- *Other iterative methods and other loss functions.* The focus of the current paper has been the analysis of vanilla GD tailored to the natural squared loss. This is by no means to advocate GD as the top-performing algorithm in practice; rather, we are using this simple algorithm to isolate some seemingly pervasive phenomena (i.e. implicit regularization) that generic optimization theory fails to account for. The simplicity of vanilla GD makes it an ideal object to initiate such discussions. That being said, practitioners should definitely explore as many algorithmic alternatives as possible before settling on a particular algorithm. Take phase retrieval for example: iterative methods other than GD and / or algorithms tailored to other loss functions have been proposed in the nonconvex optimization literature, including but not limited to alternating minimization, block coordinate descent, and sub-gradient methods and prox-linear methods tailed to non-smooth losses. It would be interesting to develop a full theoretical understanding of a broader class of iterative algorithms, and to conduct a careful comparison regarding which loss functions lead to the most desirable practical performance.

- *Connections to deep learning?* We have focused on nonlinear systems that are bilinear or quadratic in this paper. Deep learning formulations/architectures, highly nonlinear, are notorious for their daunting nonconvex geometry. However, iterative methods including stochastic gradient descent have enjoyed enormous practical success in learning neural networks (e.g. [ZSJ$^+$17, SJL19, FMZ19]), even when the architecture is significantly over-parameterized without explicit regularization. We hope the message conveyed in this paper for several simple statistical models can shed light on why simple forms of gradient descent and variants work so well in learning complicated neural networks.

Finally, while the present paper provides a general recipe for problem-specific analyses of nonconvex algorithms, we acknowledge that a unified theory of this kind has yet to be developed. As a consequence, each problem requires delicate and somewhat lengthy analyses of its own. It would certainly be helpful if one could single out a few stylized structural properties / elements (like sparsity and incoherence in compressed sensing [CP11]) that enable near-optimal performance guarantees through an over-arching method of analysis; with this in place, one would not need to start each problem from scratch. Having said that, we believe

that our current theory elucidates on a few ingredients (e.g. the region of incoherence and leave-one-out stability) that might serve as crucial building blocks for such a general theory. We invite the interested readers to contribute towards this path forward.

Acknowledgements

Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ARO grant W911NF-18-1-0303, by the ONR grant N00014-19-1-2120, and by the Princeton SEAS innovation award. Y. Chi is supported in part by the grants AFOSR FA9550-15-1-0205, ONR N00014-18-1-2142, ARO W911NF-18-1-0303, NSF CCF-1826519, ECCS-1818571, CCF-1806154. Y. Chen would like to thank Yudong Chen for inspiring discussions about matrix completion.

References

[AAH17] A. Aghasi, A. Ahmed, and P. Hand. Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. *arXiv preprint arXiv:1702.04342*, 2017.

[AFWZ17] E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

[ARR14] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[AS08] N. Alon and J. H. Spencer. *The Probabilistic Method (3rd Edition)*. Wiley, 2008.

[BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

[BEB17] T. Bendory, Y. C. Eldar, and N. Boumal. Non-convex phase retrieval from STFT measurements. *IEEE Transactions on Information Theory*, 2017.

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[BR17] S. Bahmani and J. Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. In *Artificial Intelligence and Statistics*, pages 252–260, 2017.

[Bub15] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[CC17] Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883, 2017.

[CC18] Y. Chen and E. Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics*, 71(8):1648–1714, 2018.

[CCF18] Y. Chen, C. Cheng, and J. Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *arXiv preprint arXiv:1811.12804*, 2018.

[CCF+19] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv preprint arXiv:1902.07698*, 2019.

[CCFM18] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, pages 1–33, 2018.

[CCG15] 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.

[CESV13] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[CFL15] P. Chen, A. Fannjiang, and G.-R. Liu. Phase retrieval with one or two diffraction patterns by alternating projections with the null initialization. *Journal of Fourier Analysis and Applications*, pages 1–40, 2015.

[CFMW17] Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-k ranking. *arXiv:1707.09971, accepted to Annals of Statistics*, 2017.

[Che15] Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[Chi16] Y. Chi. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):782–794, 2016.

[CJN17] Y. Cherapanamjeri, P. Jain, and P. Netrapalli. Thresholding based outlier robust PCA. In *Conference on Learning Theory*, pages 593–628, 2017.

[CL14] 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.

[CL16] Y. Chi and Y. M. Lu. Kaczmarz method for solving quadratic equations. *IEEE Signal Processing Letters*, 23(9):1183–1187, 2016.

[CLC18] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

[CLM+16] T. T. Cai, X. Li, Z. Ma, et al. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(3):11:1–11:37, Jun 2011.

[CLS15] 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, April 2015.

[CLW17] J.-F. Cai, H. Liu, and Y. Wang. Fast rank one alternating minimization algorithm for phase retrieval. *arXiv preprint arXiv:1708.08751*, 2017.

[CP11] E. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, April 2009.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[CSV13] 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):1017–1026, 2013.

[CT10] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053 –2080, May 2010.

[CW15] 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.

[CYC14] 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.

[CZ15] T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[DDP17] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

[DK70] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[Dop00] F. M. Dopico. A note on $\sin\Theta$ theorems for singular subspace variations. *BIT*, 40(2):395–403, 2000.

[DR16] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[DR18] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference*, 2018.

[DTL17] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu. Phase retrieval via linear programming: Fundamental limits and algorithmic improvements. *arXiv preprint arXiv:1710.05234*, 2017.

[EK15] N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81, 2015.

[EKBB+13] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[FMZ19] J. Fan, C. Ma, and Y. Zhong. A selective overview of deep learning. *arXiv preprint arXiv:1904.05526*, 2019.

[GLM16] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[GM17] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems*, pages 3653–3663, 2017.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011.

[GS18] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, 2018.

[GWB+17] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[GX16] B. Gao and Z. Xu. Phase retrieval using Gauss-Newton method. *arXiv preprint arXiv:1606.08135*, 2016.

[HH17] W. Huang and P. Hand. Blind deconvolution by a steepest descent algorithm on a quotient manifold. *arXiv preprint arXiv:1710.03309*, 2017.

[Hig92] N. J. Higham. Estimating the matrix p-norm. *Numerische Mathematik*, 62(1):539–555, 1992.

[HKZ12] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 6, 2012.

[HMLZ15] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, 16:3367–3402, 2015.

[HV16] P. Hand and V. Voroninski. An elementary proof of convex phase retrieval in the natural parameter space via the linear program PhaseMax. *arXiv preprint arXiv:1611.03935*, 2016.

[HW14] M. Hardt and M. Wootters. Fast matrix completion without the condition number. *Conference on Learning Theory*, pages 638 – 678, 2014.

[JEH15] K. Jaganathan, Y. C. Eldar, and B. Hassibi. Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*, 2015.

[JKN16] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4520–4528, 2016.

[JM+18] A. Javanmard, A. Montanari, et al. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

[JN15] P. Jain and P. Netrapalli. Fast exact matrix completion with finite samples. In *Conference on Learning Theory*, pages 1007–1034, 2015.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM symposium on Theory of computing*, pages 665–674, 2013.

[KD09] K. Kreutz-Delgado. The complex gradient operator and the CR-calculus. *arXiv preprint arXiv:0906.4835*, 2009.

[KLT11] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[KMO10a] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980 –2998, June 2010.

[KMO10b] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

[KÖ16] R. Kolte and A. Özgür. Phase retrieval via incremental truncated Wirtinger flow. *arXiv preprint arXiv:1606.03196*, 2016.

[Kol11] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011.

[Lan93] S. Lang. Real and functional analysis. *Springer-Verlag, New York,*, 10:11–13, 1993.

[LB10] K. Lee and Y. Bresler. Admira: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

[LCR16] J. Lin, R. Camoriano, and L. Rosasco. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

[LL17] Y. M. Lu and G. Li. Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *arXiv preprint arXiv:1702.06435*, 2017.

[LLB17] Y. Li, K. Lee, and Y. Bresler. Blind gain and phase calibration for low-dimensional or sparse signal sensing via power iteration. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 119–123. IEEE, 2017.

[LLJB17] K. Lee, Y. Li, M. Junge, and Y. Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transactions on Information Theory*, 63(2):802–821, 2017.

[LLSW18] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 2018.

[LM14] G. Lerman and T. Maunu. Fast, robust and non-convex subspace recovery. *arXiv preprint arXiv:1406.6145*, 2014.

[LMCC18] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *arXiv preprint arXiv:1802.06286*, 2018.

[LS15] S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[LS17] S. Ling and T. Strohmer. Regularized gradient descent: A nonconvex recipe for fast joint blind deconvolution and demixing. *arXiv preprint arXiv:1703.08642*, 2017.

[LT16] Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv preprint arXiv:1611.03060*, 2016.

[LTR16] K. Lee, N. Tian, and J. Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *arXiv preprint arXiv:1610.06469*, 2016.

[LWL+16] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

[Mat90] R. Mathias. The spectral norm of a nonnegative matrix. *Linear Algebra Appl.*, 139:269–284, 1990.

[Mat93] R. Mathias. Perturbation bounds for the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 14(2):588–597, 1993.

[MBM18] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[MM17] M. Mondelli and A. Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, pages 1–71, 2017.

[MZL19] T. Maunu, T. Zhang, and G. Lerman. A well-tempered landscape for non-convex robust subspace recovery. *Journal of Machine Learning Research*, 20(37):1–59, 2019.

[NJS13] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[NNS+14] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[NW12] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, 2012.

[QZEW17] Q. Qing, Y. Zhang, Y. Eldar, and J. Wright. Convolutional phase retrieval via gradient descent. *Neural Information Processing Systems*, 2017.

[Rec11] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[RV+13] M. Rudelson, R. Vershynin, et al. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[SBE14] Y. Shechtman, A. Beck, and Y. C. Eldar. GESPAR: Efficient phase retrieval of sparse signals. *IEEE Transactions on Signal Processing*, 62(4):928–938, 2014.

[SCC17] P. Sur, Y. Chen, and E. J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *arXiv preprint arXiv:1706.01191, accepted to Probability Theory and Related Fields*, 2017.

[Sch92] B. A. Schmitt. Perturbation bounds for matrix square roots and Pythagorean sums. *Linear Algebra Appl.*, 174:215–227, 1992.

[SHS17] D. Soudry, E. Hoffer, and N. Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

[SJL19] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

[SL16] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[Sol14] M. Soltanolkotabi. *Algorithms and Theory for Clustering and Nonconvex Quadratic Programming*. PhD thesis, Stanford University, 2014.

[Sol19] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *IEEE Transactions on Information Theory*, 65(4):2374–2400, 2019.

[SQW16] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE, 2016.

[SQW17] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[SR15] P. Schniter and S. Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2015.

[SS12] W. Schudy and M. Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Symposium on Discrete Algorithms*, pages 437–446. ACM, New York, 2012.

[SWW17] 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.

[Tao12] T. Tao. *Topics in Random Matrix Theory*. Graduate Studies in Mathematics. American Mathematical Society, Providence, Rhode Island, 2012.

[tB77] J. M. F. ten Berge. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276, 1977.

[TBS+16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. JMLR. org, 2016.

[Tro15a] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[Tro15b] J. A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, May 2015.

[TV17] Y. S. Tan and R. Vershynin. Phase retrieval via randomized kaczmarz: Theoretical guarantees. *arXiv preprint arXiv:1706.09993*, 2017.

[TW16] J. Tanner and K. Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417–429, 2016.

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, pages 210 – 268, 2012.

[WC16] L. Wang and Y. Chi. Blind deconvolution from multiple sparse inputs. *IEEE Signal Processing Letters*, 23(10):1384–1388, 2016.

[WCCL16] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[Wed72] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[Wei15] K. Wei. Solving systems of phaseless equations via Kaczmarz methods: A proof of concept study. *Inverse Problems*, 31(12):125008, 2015.

[WGE17] 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.

[WGSC17] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen. Solving almost all systems of random quadratic equations. *arXiv preprint arXiv:1705.10407*, 2017.

[WZG+18] G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491, 2018.

[YWS15] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[ZB18] Y. Zhong and N. Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

[ZBH+17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017.

[ZCL16] H. Zhang, Y. Chi, and Y. Liang. Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. In *International conference on machine learning*, pages 1022–1031, 2016.

[ZL15] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

[ZL16] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[ZLK+17] Y. Zhang, Y. Lau, H.-w. Kuo, S. Cheung, A. Pasupathy, and J. Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4902, 2017.

[ZSJ+17] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning*, pages 4140–4149. JMLR. org, 2017.

[ZWL15] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

[ZZLC17] 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.