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

Cong Ma [1] **Kaizheng Wang** [1] **Yuejie Chi** [2] **Yuxin Chen** [3]

Abstract

Recent years have seen a flurry of activities in designing provably efficient nonconvex optimization procedures for solving statistical estimation problems. For various problems like phase retrieval or low-rank matrix completion, state-of-the-art nonconvex procedures require proper regularization (e.g. trimming, regularized cost, projection) in order to guarantee fast convergence. When it comes to 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 several nonconvex problems: even in the absence of explicit regularization, 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 two statistical estimation problems, i.e. solving random quadratic systems of equations and low-rank matrix completion, we establish that gradient descent achieves near-optimal statistical and computational guarantees without explicit regularization. As a byproduct, for noisy matrix completion, we demonstrate that gradient descent enables optimal control of both entrywise and spectral-norm errors.

[1]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA [2]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA [3]Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. Correspondence to: Cong Ma <congm@princeton.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

1. Introduction

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,

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

where $\boldsymbol{x}^\natural$ is the unknown object of interest, and the $\mathcal{A}_j$'s are certain nonlinear maps known *a priori*. Can we hope to reconstruct the underlying object $\boldsymbol{x}^\natural$ in a faithful yet efficient manner? Problems of this kind abound in information and statistical science, prominent examples including low-rank matrix recovery (Keshavan et al., 2010; Candès & Recht, 2009), phase retrieval (Candès et al., 2013; Jaganathan et al., 2015), and learning neural networks (Soltanolkotabi et al., 2017; Zhong et al., 2017), to name just a few.

In principle, one can attempt reconstruction by seeking a solution that minimizes the empirical loss, namely,

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

Unfortunately, this empirical loss minimization problem is, in many cases, highly nonconvex, making it NP-hard in general. For example, this non-convexity issue comes up in:

- *Solving random quadratic systems of equations (a.k.a. phase retrieval)*: where one wishes to solve for $\boldsymbol{x}^\natural$ in m quadratic equations $y_j = \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2$, $1 \le j \le m$, with $\{\boldsymbol{a}_j\}_{1\le j\le m}$ denoting the known design vectors. In this case, the empirical risk minimization is given by

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

- *Low-rank matrix completion*: which aims to predict all entries of a low-rank matrix $\boldsymbol{M}^\natural = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ from partial entries (those from an index subset Ω), where $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$ $(r \ll n)$. Here, the nonconvex problem to solve is

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}} \ f(\boldsymbol{X}) = \frac{n^2}{4m}\sum_{(j,k)\in\Omega}\left(M_{j,k}^\natural - \boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k\right)^2.$$

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 e.g. (Chen & Candès, 2017)
Matrix completion	n/a	n/a	n/a	nr^7	$\frac{n}{r} \log \frac{1}{\epsilon}$	regularized loss e.g. (Sun & Luo, 2016)
				nr^2	$r^2 \log \frac{1}{\epsilon}$	projection e.g. (Chen & Wainwright, 2015)

1.1. Nonconvex Optimization via Regularized GD

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 (GD), which follows the update rule

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f\big(\boldsymbol{x}^t\big), \qquad t \geq 0, \tag{3}$$

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 natural questions are: 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, particularly in a *local* region surrounding the global optimum. In light of this, an effective nonconvex iterative method typically consists of two parts:

1. an initialization scheme (e.g. spectral methods);
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. (Keshavan et al., 2010; Jain et al., 2013; Chen & Wainwright, 2015; Sun & Luo, 2016; Candès et al., 2015; Chen & Candès, 2017). However, rather than directly applying vanilla GD (3), 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 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

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \mathcal{T}\left(\nabla f\big(\boldsymbol{x}^t\big)\right), \tag{4}$$

  where $\mathcal{T}$ is an operator that effectively drops samples bearing too much influence on the search direction (Chen & Candès, 2017; Zhang et al., 2016b; Wang et al., 2017).
- *Regularized loss*, which attempts to optimize a regularized empirical risk function through

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

  where $R(\boldsymbol{x})$ stands for an additional penalty term in the empirical loss. For example, in matrix completion, $R(\cdot)$ penalizes the ℓ_2 row norm (Keshavan et al., 2010; Sun & Luo, 2016) as well as the Frobenius norm (Sun & Luo, 2016) of the decision matrix.
- *Projection*, which projects the iterates onto certain sets based on prior knowledge, that is,

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

  where $\mathcal{P}$ is a certain projection operator used to enforce, for example, incoherence properties. This strategy has been employed in low-rank matrix completion (Chen & Wainwright, 2015; Zheng & Lafferty, 2016).

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.

1.2. Regularization-free Procedures?

The regularized gradient descent algorithms, while exhibiting appealing performance, usually introduce more tuning parameters depending on the assumed statistical models. In contrast, vanilla gradient descent (cf. (3)) — 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 as an example: 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 (Candès et al., 2015).

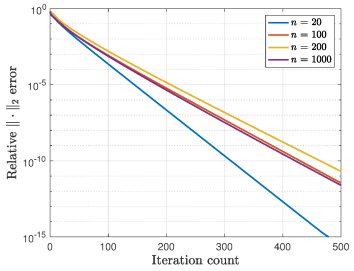

Figure 1. Relative ℓ_2 error of $\boldsymbol{x}^t$ (modulo the global phase) vs. iteration count for phase retrieval under i.i.d. Gaussian design, where $m = 10n$ and $\eta_t = 0.1$.

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 $O(n \log(1/\varepsilon))$ iterations. Nevertheless, the computational guarantee is significantly outperformed by the regularized version (called truncated Wirtinger flow (Chen & Candès, 2017)), which only requires $O(\log(1/\varepsilon))$ 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 $\eta_t = O(1/(n\|\boldsymbol{x}^\natural\|_2^2))$ — and hence slow movement — suggested by the theory (Candès et al., 2015). While this is already the largest possible step size allowed in the theory published in (Candès et al., 2015), it is considerably more conservative than the choice $\eta_t = O\big(1/\|\boldsymbol{x}^\natural\|_2^2\big)$ theoretically justified for the regularized version (Chen & Candès, 2017; Zhang et al., 2016b).

The lack of understanding and the suboptimal results about vanilla GD raise a natural question: *are regularization-free iterative algorithms inherently suboptimal for solving nonconvex statistical estimation problems?*

1.3. Numerical Surprise of Unregularized GD

To answer the preceding question, it is perhaps best to first collect some numerical evidence. In what follows, we test the performance of vanilla GD for solving random quadratic systems using a *constant* step size. The initial guess is obtained by means of the standard spectral method.

For each n, set $m = 10n$, take $\boldsymbol{x}^\natural \in \mathbb{R}^n$ 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 illustrates the relative ℓ_2 error $\min\{\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2, \|\boldsymbol{x}^t + \boldsymbol{x}^\natural\|_2\}/\|\boldsymbol{x}^\natural\|_2$ (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.

In all settings, vanilla gradient descent enjoys remarkable linear convergence, always yielding an accuracy of 10^{-5} (in a relative sense) within around 200 iterations. In particular, 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 (Candès et al., 2015) 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 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?

1.4. This Paper

The main contribution of this paper is towards demystifying the "unreasonable" effectiveness of regularization-free nonconvex gradient 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 two fundamental statistical estimation problems, 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, vanilla gradient descent converges in an almost dimension-free $O(\log(1/\epsilon))$ iterations, possibly up to a $\log n$ factor. As a byproduct of our theory, we show that gradient descent provably controls the *entrywise* and *spectral-norm* estimation errors for noisy matrix completion.

2. Implicit Regularization – A Case Study

To reveal reasons behind the effectiveness of vanilla gradient descent, we first examine the 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 of equations (or phase retrieval) and Wirtinger flow as a case study, but our diagnosis applies more generally.

2.1. Gradient Descent Theory Revisited

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 (3) enjoys ℓ_2

error contraction (Bubeck, 2015), namely, for $t \geq 0$

$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{\natural}\right\|_2 \leq \left(1-\frac{2}{\beta/\alpha+1}\right)\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{\natural}\right\|_2, \tag{7}$$

as long as the step size is chosen as $\eta_t = 2/(\alpha+\beta)$. Here, $\boldsymbol{x}^{\natural}$ denotes the global minimum. This immediately reveals the iteration complexity for gradient descent: the number of iterations taken to attain ϵ-accuracy is bounded by $O((\beta/\alpha)\log(1/\epsilon))$. 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

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

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

$$\mathcal{B}_{\delta}(\boldsymbol{x}) := \left\{\boldsymbol{x} \mid \|\boldsymbol{x}-\boldsymbol{x}^{\natural}\|_2 \leq \delta\|\boldsymbol{x}^{\natural}\|_2\right\}. \tag{8}$$

The contraction result (7) continues to hold, as long as the algorithm starts with an initial point that falls inside $\mathcal{B}_{\delta}(\boldsymbol{x})$.

2.2. Local Geometry for Solving 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 as an example. Consider the i.i.d. Gaussian design in which $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \leq j \leq m$, and suppose without loss of generality that the underlying signal obeys $\|\boldsymbol{x}^{\natural}\|_2 = 1$.

In the regime where $m \asymp n\log n$ (which is the regime considered in (Candès et al., 2015)), local strong convexity is present, in the sense that $f(\cdot)$ as defined in (2) obeys

$$\nabla^2 f(\boldsymbol{x}) \succeq (1/2)\cdot \boldsymbol{I}_n, \qquad \forall \boldsymbol{x}: \left\|\boldsymbol{x}-\boldsymbol{x}^{\natural}\right\|_2 \leq \delta\left\|\boldsymbol{x}^{\natural}\right\|_2$$

with high probability, provided that $\delta > 0$ is sufficiently small (see (Soltanolkotabi, 2014; White et al., 2015) and (Ma et al., 2017)). The smoothness parameter, however, is not well-controlled. In fact, it can be as large as (up to logarithmic factors) $\left\|\nabla^2 f(\boldsymbol{x})\right\| \lesssim n$ even when we restrict attention to the local ℓ_2 ball (8) 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 $O(n\log(1/\epsilon))$. This underpins the analysis in (Candès et al., 2015).

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. A direct application of generic gradient descent theory leads to an overly conservative learning rate and a pessimistic convergence rate, unless the number of samples is enormously larger than the number of unknowns.

2.3. Which Region Enjoys Nicer Geometry?

Interestingly, our theory identifies a local region surrounding $\boldsymbol{x}^{\natural}$ 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

$$\left\|\boldsymbol{x}-\boldsymbol{x}^{\natural}\right\|_2 \leq \delta\left\|\boldsymbol{x}^{\natural}\right\|_2 \qquad \text{and} \tag{9a}$$

$$\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\top}\left(\boldsymbol{x}-\boldsymbol{x}^{\natural}\right)\right| \lesssim \sqrt{\log n}\left\|\boldsymbol{x}^{\natural}\right\|_2, \tag{9b}$$

where $\delta > 0$ is some small numerical constant. As will be formalized in (Ma et al., 2017), 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 all $\boldsymbol{x}$ 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. (9a)), and (ii) the iterate remains incoherent[1] with respect to the sensing vectors (cf. (9b)). See Figure 2(a) for an illustration.

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 (7), we can ensure ℓ_2 error contraction after one iteration, namely,

$$\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{\natural}\|_2 \leq \|\boldsymbol{x}^{t}-\boldsymbol{x}^{\natural}\|_2$$

and hence $\boldsymbol{x}^{t+1}$ stays within the local ℓ_2 ball and hence satisfies (9a). However, it is not immediately obvious that $\boldsymbol{x}^{t+1}$ would still stay incoherent with the sensing vectors and satisfy (9b). If $\boldsymbol{x}^{t+1}$ leaves the RIC, then 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 Fig. 2(b) for a visual illustration. In fact, in almost all regularized gradient descent algorithms mentioned in Section 1.1, the regularization procedures are mainly proposed 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

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

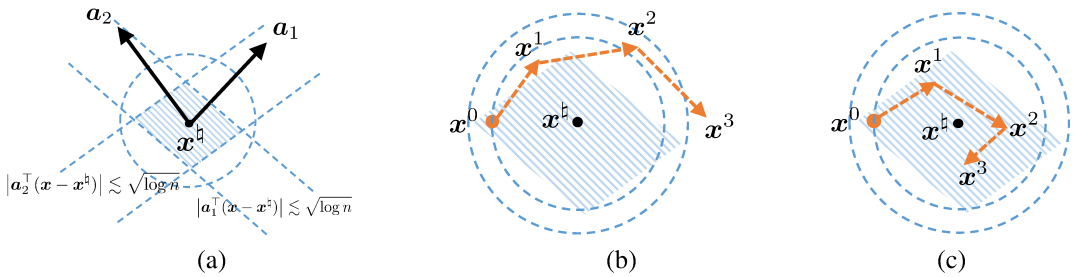

Figure 2. (a) The shaded region is an illustration of the incoherence region, which satisfies $\left|\boldsymbol{a}_j^\top(\boldsymbol{x}-\boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}$ for all points $\boldsymbol{x}$ in the region. (b) When $\boldsymbol{x}^0$ resides in the desired region, we 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.

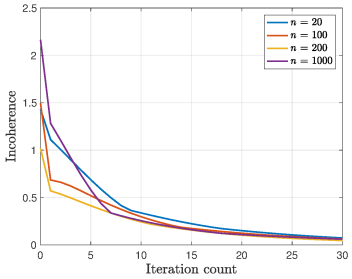

Figure 3. The incoherence measure $\frac{\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top(\boldsymbol{x}^t-\boldsymbol{x}^\natural)\right|}{\sqrt{\log n}\left\|\boldsymbol{x}^\natural\right\|_2}$ of the gradient iterates vs. iteration count for the phase retrieval problem. The results are shown for $n \in \{20, 100, 200, 1000\}$ 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.

in Fig. 3 the incoherence measure $\frac{\max_j\left|\boldsymbol{a}_j^\top(\boldsymbol{x}^t-\boldsymbol{x}^\natural)\right|}{\sqrt{\log n}\|\boldsymbol{x}^\natural\|_2}$ vs. the iteration count in a typical Monte Carlo trial, generated in the same way as for Figure 1. 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. 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 throughout the execution of the algorithm (as well as the spectral initialization) are provably constrained within the RIC, implying fast convergence of vanilla gradient descent (cf. Figure 2(c)). The fact that the iterates stay incoherent with the measurement mechanism automatically, without explicit enforcement, is termed "implicit regularization" in the current work.

2.5. A Glimpse of the Analysis: A Leave-one-out Trick

In order to rigorously establish (9b) 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 (El Karoui, 2015; Javanmard & Montanari, 2015; Sur et al., 2017; Zhong & Boumal, 2017; Chen et al., 2017; Abbe et al., 2017). Note that the iterate $\boldsymbol{x}^t$ is statistically dependent of the design vectors $\{\boldsymbol{a}_j\}$. Under such circumstances, one often resorts to generic bounds like the Cauchy-Schwarz inequality when bounding $\boldsymbol{a}_l^\top(\boldsymbol{x}^t-\boldsymbol{x}^\natural)$, which would not yield a desirable estimate. To address this issue, we introduce a sequence of auxiliary iterates $\{\boldsymbol{x}^{t,(l)}\}$ for each $1 \le l \le m$ (for analytical purposes only), obtained by running vanilla gradient descent using all but the lth sample. As one expects, such auxiliary trajectories serve as extremely good surrogates of $\{\boldsymbol{x}^t\}$ in the sense that

$$\boldsymbol{x}^t \approx \boldsymbol{x}^{t,(l)}, \qquad 1 \le l \le m, \quad t \ge 0, \tag{10}$$

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

$$\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^\natural\right)\right| \lesssim \sqrt{\log n}\left\|\boldsymbol{x}^\natural\right\|_2. \tag{11}$$

Combining (10) and (11) then leads to (9b). See Figure 4 for a graphical illustration of this argument.

3. Main Results

This section formalizes the implicit regularization phenomenon underlying unregularized GD, and presents its consequences, namely near-optimal statistical and computational guarantees for phase retrieval and matrix completion. The complete proofs can be found in (Ma et al., 2017).

3.1. Solving Random Quadratic Systems / Phase Retrieval

Suppose the m quadratic equations

$$y_j = \left(\boldsymbol{a}_j^\top\boldsymbol{x}^\natural\right)^2, \qquad j = 1, 2, \ldots, m \tag{12}$$

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

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) := \frac{1}{4m}\sum_{j=1}^m\left[\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)^2 - y_j\right]^2. \tag{13}$$

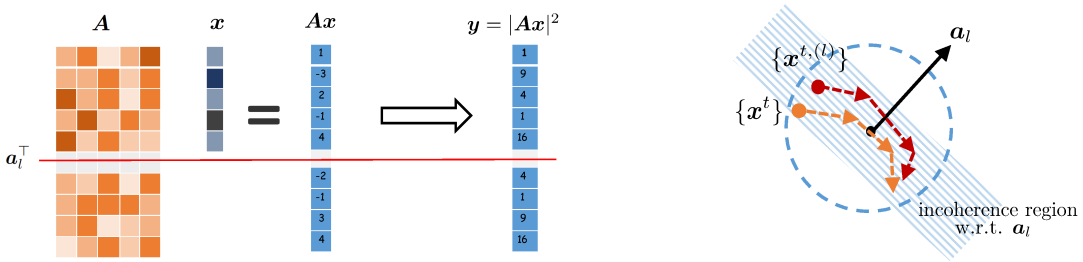

Figure 4. Illustration of the leave-one-out sequence w.r.t. $\boldsymbol{a}_l$. (a) The sequence $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$ is constructed without using the lth sample. (b) Since the auxiliary sequence $\{\boldsymbol{x}^{t,(l)}\}$ is constructed without using $\boldsymbol{a}_l$, the leave-one-out 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 one sample.

The Wirtinger flow (WF) algorithm, first introduced in (Candès et al., 2015), is a combination of spectral initialization and vanilla gradient descent; see Algorithm 1.

Algorithm 1 Wirtinger flow for phase retrieval

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

$$\boldsymbol{Y} = \frac{1}{m}\sum\nolimits_{j=1}^{m} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top, \tag{14}$$

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**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right). \tag{15}$$

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

$$\operatorname{dist}(\boldsymbol{x}, \boldsymbol{x}^\natural) := \min\left\{\|\boldsymbol{x}-\boldsymbol{x}^\natural\|_2, \|\boldsymbol{x}+\boldsymbol{x}^\natural\|_2\right\}. \tag{16}$$

Our finding is summarized in the following theorem.

Theorem 1. *Let $\boldsymbol{x}^\natural \in \mathbb{R}^n$ be a fixed vector. Suppose $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for each $1\leq j\leq m$ and $m \geq c_0 n\log n$ for some sufficiently large constant $c_0 > 0$. Assume the learning rate 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 $1 - O\left(mn^{-5}\right)$, the Wirtinger flow iterates (Algorithm 1) satisfy that for all $t \geq 0$,*

$$\operatorname{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \varepsilon(1-\eta\|\boldsymbol{x}^\natural\|_2^2/2)^t\|\boldsymbol{x}^\natural\|_2, \tag{17a}$$

$$\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}^t-\boldsymbol{x}^\natural\right)\right| \leq c_2\sqrt{\log n}\|\boldsymbol{x}^\natural\|_2. \tag{17b}$$

Theorem 1 reveals a few intriguing properties of WF.

- **Implicit regularization:** Theorem 1 asserts that the incoherence properties are satisfied throughout the execution of the algorithm, including the spectral initialization (see (17b)), which formally justifies the implicit regularization feature we hypothesized.
- **Near-constant step size:** Consider the case where $\|\boldsymbol{x}^\natural\|_2 = 1$. 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 (Candès et al., 2015), Theorem 1 allows WF to attain ϵ-accuracy within $O(\log n\log(1/\epsilon))$ iterations. The resulting computational complexity of WF is $O\left(mn\log n\log(1/\epsilon)\right)$, which significantly improves upon the result $O\left(mn^2\log\left(1/\epsilon\right)\right)$ derived in (Candès et al., 2015). As a side note, if the sample size further increases to $m \asymp n\log^2 n$, then $\eta \asymp 1$ 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 $\max_j\left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}^t-\boldsymbol{x}^\natural\right)\right| \lesssim \|\boldsymbol{x}^\natural\|_2$. We omit the details here.

Finally, we remark that similar implicit regularization phenomenon holds even in the presence of random initialization. See (Chen et al., 2018) for details.

3.2. Low-rank Matrix Completion

We move on to the low-rank matrix completion problem.

Let $\boldsymbol{M}^\natural \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix[2] with rank r, and suppose its eigendecomposition is

$$\boldsymbol{M}^\natural = \boldsymbol{U}^\natural\boldsymbol{\Sigma}^\natural\boldsymbol{U}^{\natural\top}, \tag{18}$$

where $\boldsymbol{U}^\natural \in \mathbb{R}^{n\times r}$ consists of orthonormal columns, and $\boldsymbol{\Sigma}^\natural$ 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

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

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

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

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

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

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**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla f\left(\boldsymbol{X}^t\right). \tag{21}$$

$\boldsymbol{X}^\natural = \boldsymbol{U}^\natural(\boldsymbol{\Sigma}^\natural)^{1/2}$ allows us to factorize $\boldsymbol{M}^\natural$ as

$$\boldsymbol{M}^\natural = \boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}. \tag{19}$$

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

$$Y_{j,k} = \begin{cases} M^\natural_{j,k} + E_{j,k} & \text{with probability } p, \\ 0, & \text{else,} \end{cases} \tag{20}$$

where the entries of $\boldsymbol{E} = [E_{j,k}]_{1\le j\le k\le n}$ are independent sub-Gaussian noise with sub-Gaussian norm σ (see (Vershynin, 2012)). 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

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

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

Before proceeding to the main theorem, we first introduce a standard incoherence parameter required for matrix completion (Candès & Recht, 2009).

Definition 1 (Incoherence for matrix completion)**.** *A rank-r matrix $\boldsymbol{M}^\natural$ with eigendecomposition $\boldsymbol{M}^\natural = \boldsymbol{U}^\natural\boldsymbol{\Sigma}^\natural\boldsymbol{U}^{\natural\top}$ is said to be μ-incoherent if*

$$\left\|\boldsymbol{U}^\natural\right\|_{2,\infty} \le \sqrt{\mu/n}\left\|\boldsymbol{U}^\natural\right\|_{\mathrm{F}} = \sqrt{\mu r/n}, \tag{22}$$

where $\|\cdot\|_{2,\infty}$ denotes the largest ℓ_2 norm of the rows.

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

$$\widehat{\boldsymbol{H}}^t := \underset{\boldsymbol{R}\in\mathcal{O}^{r\times r}}{\operatorname{argmin}}\left\|\boldsymbol{X}^t\boldsymbol{R} - \boldsymbol{X}^\natural\right\|_{\mathrm{F}}, \tag{23}$$

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 $\boldsymbol{M}^\natural$ 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*

$$\sigma\sqrt{\frac{n}{p}} \ll \frac{\sigma_{\min}}{\sqrt{\kappa^3\mu r\log^3 n}}. \tag{24}$$

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

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

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

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

for all $0 \le t \le T = O(n^5)$,[3] 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, under near-minimal sample complexity.

- **Implicit regularization:** In Theorem 2, we bound the ℓ_2/ℓ_∞ error of the iterates in a uniform manner. Note that $\left\|\boldsymbol{X} - \boldsymbol{X}^\natural\right\|_{2,\infty} = \max_j\left\|\boldsymbol{e}_j^\top\left(\boldsymbol{X} - \boldsymbol{X}^\natural\right)\right\|_2$, which implies the iterates remain incoherent with the sensing vectors throughout and have small incoherence parameters, including the spectral initialization (cf. (22)). In comparison, prior works either include a penalty term on $\{\|\boldsymbol{e}_j^\top\boldsymbol{X}\|_2\}_{1\le j\le n}$ (Keshavan et al., 2010; Sun & Luo, 2016) and/or $\|\boldsymbol{X}\|_{\mathrm{F}}$ (Sun & Luo, 2016) to encourage an incoherent and/or low-norm solution, or add an extra projection operation to enforce incoherence (Chen & Wainwright, 2015; Zheng & Lafferty, 2016). Our results demonstrate that such explicit regularization is unnecessary for the success of gradient descent.
- **Constant step size:** Without loss of generality we may assume that $\sigma_{\max} = \|\boldsymbol{M}^\natural\| = O(1)$, which can be done by choosing proper scaling of $\boldsymbol{M}^\natural$. 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

[3] 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.

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 GD; see Table 1.

- **Near-minimal Euclidean error:** As the number of iterations t increases, the Euclidean error of vanilla GD converges to

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\natural\right\|_{\mathrm{F}}\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\natural\right\|_{\mathrm{F}}, \tag{25}$$

  which coincides with the theoretical guarantee in (Chen & Wainwright, 2015) and matches the minimax lower bound established in (Negahban & Wainwright, 2012; Koltchinskii et al., 2011).
- **Near-optimal entrywise error:** The ℓ_2/ℓ_∞ error bound immediately yields entrywise control of the empirical risk. Specifically, as soon as the number of iterations t is sufficiently large, we have

$$\left\|\boldsymbol{X}^t\boldsymbol{X}^{t\top}-\boldsymbol{M}^\natural\right\|_\infty\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{M}^\natural\right\|_\infty.$$

  Compared with the Euclidean loss (25), 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.

4. Related Work

Convex relaxations have received much attention for solving nonlinear systems of equations in the past decade. Instead of 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. (Recht et al., 2010; Candès et al., 2013; Candès & Recht, 2009)). This has enjoyed great success in both theory and practice. Despite appealing statistical guarantees, SDP is in general prohibitively expensive when processing large-scale datasets.

In comparison, nonconvex approaches 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 (Netrapalli et al., 2013; Candès et al., 2015; Chen & Candès, 2017), low-rank matrix sensing and completion (Tu et al., 2016; Bhojanapalli et al., 2016; Park et al., 2016; Chen & Wainwright, 2015; Zheng & Lafferty, 2015; Ge et al., 2016), dictionary learning (Sun et al., 2017), blind deconvolution (Li et al., 2016a; Cambareri & Jacques, 2016; Lee et al., 2017), tensor decomposition (Ge & Ma, 2017), joint alignment (Chen & Candès, 2018), learning shallow neural networks (Soltanolkotabi et al., 2017; Zhong et al., 2017). In several problems (Sun et al., 2016; 2017; Ge & Ma, 2017; Ge et al., 2016; Li et al., 2016b; Li & Tang, 2016; Mei et al., 2016; Maunu et al., 2017), 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.

When it comes to noisy matrix completion, to the best of our knowledge, no rigorous guarantees have been established for gradient descent without explicit regularization. A notable exception is (Jin et al., 2016), which studies unregularized stochastic gradient descent for online matrix completion with fresh samples used in each iteration.

Finally, we note that the notion of implicit regularization — broadly defined — arises in settings far beyond what considered herein. For instance, it has been 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 (Li et al., 2017; Gunasekar et al., 2017). The stochastic gradient methods have also been shown to implicitly enforce Tikhonov regularization in several statistical learning settings (Lin et al., 2016). More broadly, this phenomenon seems crucial in enabling efficient training of deep neural networks (Neyshabur et al., 2017; Zhang et al., 2016a; Soudry et al., 2017; Keskar et al., 2016).

5. 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 two cases studied herein, and our results are initial steps towards understanding this fundamental phenomenon. That being said, there are numerous avenues that remain open. For instance, it remains unclear how to generalize the proposed leave-one-out tricks for more general designs beyond the i.i.d. Gaussian design. It would also be interesting to see whether the message conveyed in this paper can shed light on why simple forms of gradient descent and variants work so well in learning complicated neural networks. We leave these for future investigation.

Acknowledgements

Y. Chi is supported in part by the grants AFOSR FA9550-15-1-0205, ONR N00014-18-1-2142, ARO grant W911NF-18-1-0303, NSF CCF-1527456 and ECCS-1818571. Y. Chen is supported by ARO grant W911NF-18-1-0303 and by Princeton SEAS innovation award.

References

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

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

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

Cambareri, V. and Jacques, L. A non-convex blind calibration method for randomised sensing strategies. *arXiv preprint arXiv:1605.02615*, 2016.

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

Candès, E. J., Strohmer, T., and Voroninski, V. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1017–1026, 2013.

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

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

Chen, Y. and Candès, E. J. 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.

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

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

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

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

Ge, R. and Ma, T. On the optimization landscape of tensor decompositions. *arXiv preprint arXiv:1706.05598*, 2017.

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

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. *arXiv preprint arXiv:1705.09280*, 2017.

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

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

Javanmard, A. and Montanari, A. De-biasing the lasso: Optimal sample size for Gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015.

Jin, C., Kakade, S. M., and Netrapalli, P. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *NIPS*, pp. 4520–4528, 2016.

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

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011. ISSN 0090-5364. doi: 10.1214/11-AOS894. URL http://dx.doi.org/10.1214/11-AOS894.

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

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

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

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

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix recovery. pp. arXiv preprint arXiv:1712.09203, 12 2017.

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

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

Maunu, T., Zhang, T., and Lerman, G. A well-tempered landscape for non-convex robust subspace recovery. *arXiv preprint arXiv:1706.03896*, 2017.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, 2012. ISSN 1532-4435.

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

Neyshabur, B., Tomioka, R., Salakhutdinov, R., and Srebro, N. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Park, D., Kyrillidis, A., Caramanis, C., and Sanghavi, S. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.

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

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

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

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

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. In *ISIT*, pp. 2379–2383, 2016.

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

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

Sur, P., Chen, Y., and Candès, E. J. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *arXiv preprint arXiv:1706.01191*, 2017.

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

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

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

White, C. D., Ward, R., and Sanghavi, S. The local convexity of solving quadratic equations. *arXiv preprint arXiv:1506.07868*, 2015.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016a.

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

Zheng, Q. and Lafferty, J. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *NIPS*, pp. 109–117, 2015.

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

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. *arXiv:1706.03175*, 2017.

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