HHS Public Access

Author manuscript

Math Program. Author manuscript; available in PMC 2021 April 07.

Published in final edited form as:

Math Program. 2019 July ; 176(1-2): 5–37. doi:10.1007/s10107-019-01363-6.

Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval

Yuxin Chen[*], **Yuejie Chi**[†], **Jianqing Fan**[‡], **Cong Ma**[‡]

[*]Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

[†]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[‡]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

Abstract

This paper considers the problem of solving systems of quadratic equations, namely, recovering an object of interest $\boldsymbol{x}^\natural \in \mathbb{R}^n$ from m quadratic equations/samples $y_i = \left(a_i^\top \boldsymbol{x}^\natural\right)^2, 1 \le i \le m$. This problem, also dubbed as phase retrieval, spans multiple domains including physical sciences and machine learning.

We investigate the efficacy of gradient descent (or Wirtinger flow) designed for the nonconvex least squares problem. We prove that under Gaussian designs, gradient descent — when randomly initialized — yields an ϵ-accurate solution in $O(\log n + \log(1/\epsilon))$ iterations given nearly minimal samples, thus achieving near-optimal computational and sample complexities at once. This provides the first global convergence guarantee concerning vanilla gradient descent for phase retrieval, without the need of (i) carefully-designed initialization, (ii) sample splitting, or (iii) sophisticated saddle-point escaping schemes. All of these are achieved by exploiting the statistical models in analyzing optimization algorithms, via a leave-one-out approach that enables the decoupling of certain statistical dependency between the gradient descent iterates and the data.

1 Introduction

Suppose we are interested in learning an unknown object $\boldsymbol{x}^\natural \in \mathbb{R}^n$, but only have access to a few quadratic equations of the form

$$y_i = \left(a_i^\top \boldsymbol{x}^\natural\right)^2, \qquad 1 \le i \le m, \tag{1}$$

where y_i is the sample we collect and $\boldsymbol{a}_i$ is the design vector known *a priori*. Is it feasible to reconstruct $\boldsymbol{x}^\natural$ in an accurate and efficient manner?

The problem of solving systems of quadratic equations (1) is of fundamental importance and finds applications in numerous contexts. Perhaps one of the best-known applications is the

yuxin.chen@princeton.edu.

so-called *phase retrieval* problem arising in physical sciences [CESV13,SEC+15]. In X-ray crystallography, due to the ultra-high frequency of the X-rays, the optical sensors and detectors are incapable of recording the phases of the diffractive waves; rather, only intensity measurements are collected. The phase retrieval problem comes down to reconstructing the specimen of interest given intensity-only measurements. If one thinks of $\boldsymbol{x}^\natural$ as the specimen under study and uses $\{y_i\}$ to represent the intensity measurements, then phase retrieval is precisely about inverting the quadratic system (1).

Moving beyond physical sciences, the above problem also spans various machine learning applications. One example is *mixed linear regression*, where one wishes to estimate two unknown vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ from unlabeled linear measurements [CYC14]. The acquired data $\{\boldsymbol{a}_i, b_i\}_{1\le i\le m}$ take the form of either $b_i \approx a_i^\top \beta_1$ or $b_i \approx a_i^\top \beta_2$, without knowing which of the two vectors generates the data. In a simple symmetric case with $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{x}^\natural$ (so that $b_i \approx \pm a_i^\top \boldsymbol{x}^\natural$), the squared measurements $y_i = b_i^2 \approx \left(a_i^\top \boldsymbol{x}^\natural\right)^2$ become the sufficient statistics, and hence mixed linear regression can be converted to learning $\boldsymbol{x}^\natural$ from $\{\boldsymbol{a}_i, y_i\}$. Furthermore, the quadratic measurement model in (1) allows to represent a single neuron associated with a quadratic activation function, where $\{\boldsymbol{a}_i, y_i\}$ are the data and $\boldsymbol{x}^\natural$ encodes the parameters to be learned. As described in [SJL17, LMZ17], *learning neural nets with quadratic activations* involves solving systems of quadratic equations.

1.1 Nonconvex optimization via gradient descent

A natural strategy for inverting the system of quadratic equations (1) is to solve the following nonconvex least squares estimation problem

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

Under Gaussian designs where $a_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, the solution to (2) is known to be exact — up to some global sign — with high probability, as soon as the number m of equations (samples) exceeds the order of the number n of unknowns [BCMN14]. However, the loss function in (2) is highly nonconvex, thus resulting in severe computational challenges. With this issue in mind, can we still hope to find the global minimizer of (2) via low-complexity algorithms which, ideally, run in time proportional to that taken to read the data?

Fortunately, in spite of nonconvexity, a variety of optimization-based methods are shown to be effective in the presence of proper statistical models. Arguably, one of the simplest algorithms for solving (2) is vanilla gradient descent (GD), which attempts recovery via the update rule

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

with η_t being the stepsize/learning rate. The above iterative procedure is also dubbed *Wirtinger flow* for phase retrieval, which can accommodate the complex-valued case as well [CLS15]. This simple algorithm is remarkably efficient under Gaussian designs: in conjunction with carefully-designed initialization and stepsize rules, GD provably converges

to the truth $\boldsymbol{x}^\natural$ at a linear rate[1], provided that the ratio m/n of the number of equations to the number of unknowns exceeds some logarithmic factor [CLS15,Sol14,MWCC17].

One crucial element in prior convergence analysis is initialization. In order to guarantee linear convergence, prior works typically recommend spectral initialization or its variants [CLS15,CC17,WGE17,ZZLC17,MWCC17,LL17,MM17]. Specifically, the spectral method forms an initial estimate $\boldsymbol{x}^0$ using the (properly scaled) leading eigenvector of a certain data matrix. Two important features are worth emphasizing:

- $\boldsymbol{x}^0$ convexity; falls within a local ℓ_2-ball surrounding $\boldsymbol{x}^\natural$ with a reasonably small radius, where $f(\cdot)$ enjoys strong convexity;
- $\boldsymbol{x}^0$ is incoherent with all the design vectors $\{\boldsymbol{a}_i\}$ — in the sense that $\left|a_i^\top \boldsymbol{x}^0\right|$ is reasonably small for all $1 \le i \le m$ — and hence $\boldsymbol{x}^0$ falls within a region where $f(\cdot)$ enjoys desired smoothness conditions.

These two properties taken collectively allow gradient descent to converge rapidly from the very beginning.

1.2 Random initialization?

The enormous success of spectral initialization gives rise to a curious question: is carefully-designed initialization necessary for achieving fast convergence? Obviously, vanilla GD cannot start from arbitrary points, since it may get trapped in undesirable stationary points (e.g. saddle points). However, is there any *simpler* initialization approach that avoids such stationary points and works equally well as spectral initialization?

A strategy that practitioners often like to employ is to initialize GD randomly. The advantage is clear: compared with spectral methods, random initialization is model-agnostic and is usually more robust visa-vis model mismatch. Despite its wide use in practice, however, GD with random initialization is poorly understood in theory. One way to study this method is through a geometric lens [SQW16]: under Gaussian designs, the loss function $f(\cdot)$ (cf. (2)) does not have any spurious local minima as long as the sample size m is on the order of $n \log^3 n$. Moreover, all saddle points are strict [GHJY15], meaning that the associated Hessian matrices have at least one negative eigenvalue if they are not local minima. Armed with these two conditions, the theory of Lee et al. [LSJR16] implies that vanilla GD converges *almost surely* to the truth. However, the convergence rate remains unsettled. In fact, we are not aware of any theory that guarantees polynomial-time convergence of vanilla GD for phase retrieval in the absence of carefully-designed initialization.

Motivated by this, we aim to pursue a formal understanding about the convergence properties of GD with random initialization. Before embarking on theoretical analyses, we first assess its practical efficiency through numerical experiments. Generate the true object $\boldsymbol{x}^\natural$ and the initial guess $\boldsymbol{x}^0$ randomly as

[1]An iterative algorithm is said to enjoy linear convergence if the iterates $\{\boldsymbol{x}^t\}$ converge geometrically fast to the minimizer $\boldsymbol{x}^\natural$.

$$\boldsymbol{x}^{\natural} \sim \mathcal{N}\left(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n\right) \quad \text{and} \quad \boldsymbol{x}^0 \sim \mathcal{N}\left(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n\right).$$

We vary the number n of unknowns (i.e. $n = 100, 200, 500, 800, 1000$), set $m = 10n$, and take a constant stepsize $\eta_t \equiv 0.1$. Here the measurement vectors are generated from Gaussian distributions, i.e. $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for $1 \le i \le m$. The relative ℓ_2 errors $\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^{\natural})/\|\boldsymbol{x}^{\natural}\|_2$ of the GD iterates in a random trial are plotted in Figure 1, where

$$\text{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^{\natural}\right) := \min\left\{\left\|\boldsymbol{x}^t - \boldsymbol{x}^{\natural}\right\|_2, \left\|\boldsymbol{x}^t + \boldsymbol{x}^{\natural}\right\|_2\right\} \tag{4}$$

represents the ℓ_2 distance between $\boldsymbol{x}^t$ and $\boldsymbol{x}^{\natural}$ modulo the unrecoverable global sign.

In all experiments carried out in Figure 1, we observe two stages for GD: (1) Stage 1: the relative error of $\boldsymbol{x}^t$ stays nearly flat; (2) Stage 2: the relative error of $\boldsymbol{x}^t$ experiences geometric decay. Interestingly, Stage 1 lasts only for a few tens of iterations. These numerical findings taken together reveal appealing computational efficiency of GD in the presence of random initialization — it attains 5-digit accuracy within about 200 iterations!

To further illustrate this point, we take a closer inspection of the signal component $\langle \boldsymbol{x}^t, \boldsymbol{x}^{\natural}\rangle \boldsymbol{x}^{\natural}$ and the orthogonal component $\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^{\natural}\rangle \boldsymbol{x}^{\natural}$, where we normalize $\|\boldsymbol{x}^{\natural}\|_2 = 1$ for simplicity. Denote by $\|\boldsymbol{x}^t_{\perp}\|_2$ the ℓ_2 norm of the orthogonal component. We highlight two important and somewhat surprising observations that allude to why random initialization works.

- *The strength ratio of the signal to the orthogonal components grows exponentially.* The ratio, $\left|\left\langle \boldsymbol{x}^t, \boldsymbol{x}^{\natural}\right\rangle\right| / \|\boldsymbol{x}^t_{\perp}\|_2$, grows exponentially fast throughout the execution of the algorithm, as demonstrated in Figure 2(a). This metric $\left|\left\langle \boldsymbol{x}^t, \boldsymbol{x}^{\natural}\right\rangle\right| / \|\boldsymbol{x}^t_{\perp}\|_2$ in some sense captures the signal-to-noise ratio of the running iterates.
- *Exponential growth of the signal strength in Stage 1.* While the ℓ_2 estimation error of $\boldsymbol{x}^t$ may not drop significantly during Stage 1, the size $|\langle \boldsymbol{x}^t, \boldsymbol{x}^{\natural}\rangle|$ of the signal component increases exponentially fast and becomes the dominant component within several tens of iterations, as demonstrated in Figure 2(b). This helps explain why Stage 1 lasts only for a short duration.

The central question then amounts to whether one can develop a mathematical theory to interpret such intriguing numerical performance. In particular, how many iterations does Stage 1 encompass, and how fast can the algorithm converge in Stage 2?

1.3 Main findings

The objective of the current paper is to demystify the computational efficiency of GD with random initialization, thus bridging the gap between theory and practice. Assuming a tractable random design model in which $\boldsymbol{a}_i$'s follow Gaussian distributions, our main findings are summarized in the following theorem. Here and throughout, the notation $f(n) \lesssim$

$g(n)$ or $f(n) = O(g(n))$ (resp. $f(n) \gtrsim g(n)$, $f(n) \asymp g(n)$) means that there exist constants c_1, $c_2 > 0$ such that $f(n) \leq c_1 g(n)$ (resp. $f(n) \geq c_2 g(n)$, $c_1 g(n) \leq f(n) \leq c_2 g(n)$).

Theorem 1—*Fix* $\boldsymbol{x}^\natural \in \mathbb{R}^n$ *with* $\|\boldsymbol{x}^\natural\|_2 = 1$. *Suppose that* $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ *for* $1 \leq i \leq m$, $\boldsymbol{x}^0 \sim \mathcal{N}\left(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n\right)$, *and* $\eta_t \equiv \eta = c/\|\boldsymbol{x}^\natural\|_2^2$ *for some sufficiently small constant* $c > 0$. *Then with probability approaching one, there exist some sufficiently small constant* $0 < \gamma < 1$ *and* $T_\gamma \lesssim \log n$ *such that the GD iterates (3) obey*

$$\operatorname{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \leq \gamma(1-\rho)^{t-T_\gamma}, \qquad \forall t \geq T_\gamma$$

for some absolute constant $0 < \rho < 1$, *provided that the sample size* $m \gtrsim n \operatorname{poly}\log(m)$.

Remark 1—The readers are referred to Theorem 2 for a more general statement.

Here, the stepsize is taken to be a fixed constant throughout all iterations, and we reuse the same data across all iterations (i.e. no sample splitting is needed to establish this theorem). The GD trajectory is divided into 2 stages: (1) Stage 1 consists of the first T_γ iterations, corresponding to the first tens of iterations discussed in Section 1.2; (2) Stage 2 consists of all remaining iterations, where the estimation error contracts linearly. Several important implications/remarks follow immediately.

- *Stage 1 takes* $O(\log n)$ *iterations.* When seeded with a random initial guess, GD is capable of entering a local region surrounding $\boldsymbol{x}^\natural$ within $T_\gamma \lesssim \log n$ iterations, namely,

  $$\operatorname{dist}\left(\boldsymbol{x}^{T_\gamma}, \boldsymbol{x}^\natural\right) \leq \gamma$$

  for some sufficiently small constant $\gamma > 0$. Even though Stage 1 may not enjoy linear convergence in terms of the estimation error, it is of fairly short duration.
- *Stage 2 takes* $O(\log(1/\epsilon))$ *iterations.* After entering the local region, GD converges linearly to the ground truth $\boldsymbol{x}^\natural$ with a contraction rate $1-\rho$. This tells us that GD reaches ϵ-accuracy (in a relative sense) within $O(\log(1/\epsilon))$ iterations.
- *Near linear-time computational complexity.* Taken collectively, these imply that the iteration complexity of GD with random initialization is

  $$O\left(\log n + \log\frac{1}{\epsilon}\right).$$

  Given that the cost of each iteration mainly lies in calculating the gradient $\nabla f(\boldsymbol{x}^t)$, the whole algorithm takes nearly linear time, namely, it enjoys a computational complexity proportional to the time taken to read the data (modulo some logarithmic factor).

- *Near-minimal sample complexity.* The preceding computational guarantees occur as soon as the sample size exceeds $m \gtrsim n$ poly log(m). Given that one needs at least n samples to recover n unknowns, the sample complexity of randomly initialized GD is optimal up to some logarithmic factor.
- *Saddle points?* The GD iterates never hit the saddle points (see Figure 3 for an illustration). In fact, after a constant number of iterations at the very beginning, GD will follow a path that increasingly distances itself from the set of saddle points as the algorithm progresses. There is no need to adopt sophisticated saddle-point escaping schemes developed in generic optimization theory (e.g. cubic regularization [NP06], perturbed GD [JGN$^+$17]).
- *Weak dependency w.r.t. the design vectors.* As we will elaborate in Section 4, the statistical dependency between the GD iterates $\{\boldsymbol{x}^t\}$ and certain components of the design vectors $\{\boldsymbol{a}_i\}$ stays at an exceedingly weak level. Consequently, the GD iterates $\{\boldsymbol{x}^t\}$ proceed *as if* fresh samples were employed in each iteration. This statistical observation plays a crucial role in characterizing the dynamics of the algorithm without the need of sample splitting.

It is worth emphasizing that the entire trajectory of GD is automatically confined within a certain region enjoying favorable geometry. For example, the GD iterates are always incoherent with the design vectors, stay sufficiently away from any saddle point, and exhibit desired smoothness conditions, which we will formalize in Section 4. Such delicate geometric properties underlying the GD trajectory are not explained by prior papers. In light of this, convergence analysis based on global geometry [SQW16] — which provides valuable insights into algorithm designs with *arbitrary* initialization — results in suboptimal (or even pessimistic) computational guarantees when analyzing a specific algorithm like GD. In contrast, the current paper establishes near-optimal performance guarantees by paying particular attention to finer dynamics of the algorithm. As will be seen later, this is accomplished by heavily exploiting the statistical properties in each iterative update.

2 Why random initialization works?

Before diving into the proof of the main theorem, we pause to develop intuitions regarding why gradient descent with random initialization is expected to work. We will build our understanding step by step: (i) we first investigate the dynamics of the population gradient sequence (the case where we have infinite samples); (ii) we then turn to the finite-sample case and present a heuristic argument assuming independence between the iterates and the design vectors; (iii) finally, we argue that the true trajectory is remarkably close to the one heuristically analyzed in the previous step, which arises from a key property concerning the "near-independence" between $\{\boldsymbol{x}^t\}$ and the design vectors $\{\boldsymbol{a}_i\}$.

Without loss of generality, we assume $\boldsymbol{x}^\natural = \boldsymbol{e}_1$ throughout this section, where $\boldsymbol{e}_1$ denotes the first standard basis vector. For notational simplicity, we denote by

$$\boldsymbol{x}_{\|}^t := \boldsymbol{x}_1^t \qquad \text{and} \qquad \boldsymbol{x}_{\perp}^t := \left[\boldsymbol{x}_i^t\right]_{2 \le i \le n} \tag{5}$$

the first entry and the 2nd through the nth entries of $\boldsymbol{x}^t$, respectively. Since $\boldsymbol{x}^\natural = \boldsymbol{e}_1$, it is easily seen that

$$\underbrace{x_{\|}^t e_1 = \left\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \right\rangle \boldsymbol{x}^\natural}_{\text{signal component}} \qquad \text{and} \qquad \underbrace{\begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix} = \boldsymbol{x}^t - \left\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \right\rangle \boldsymbol{x}^\natural}_{\text{orthogonal component}} \tag{6}$$

represent respectively the components of $\boldsymbol{x}^t$ along and orthogonal to the signal direction. In what follows, we focus our attention on the following two quantities that reflect the sizes of the preceding two components[2]

$$\alpha_t := \boldsymbol{x}_{\|}^t \qquad \text{and} \qquad \beta_t := \left\| \boldsymbol{x}_\perp^t \right\|_2 . \tag{7}$$

Without loss of generality, assume that $\alpha_0 > 0$.

2.1 Population dynamics

To start with, we consider the unrealistic case where the iterates $\{\boldsymbol{x}^t\}$ are constructed using the population gradient (or equivalently, the gradient when the sample size m approaches infinity), i.e.

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla F\left(\boldsymbol{x}^t\right).$$

Here, $\nabla F(\boldsymbol{x})$ represents the population gradient given by

$$\nabla F(\boldsymbol{x}) := \left(3\|\boldsymbol{x}\|_2^2 - 1\right)\boldsymbol{x} - 2\left(\boldsymbol{x}^{\natural\top} \boldsymbol{x}\right)\boldsymbol{x}^\natural,$$

which can be computed by $\nabla F(\boldsymbol{x}) = \mathbb{E}[\nabla f(\boldsymbol{x})] = \mathbb{E}\left[\left\{\left(a_i^\top \boldsymbol{x}\right)\boldsymbol{x}^2 - \left(\mathbf{a}_i^\top \boldsymbol{x}^\natural\right)^2\right\} a_i a_i^\top \boldsymbol{x}\right]$ assuming that $\boldsymbol{x}$ and the $\boldsymbol{a}_i$'s are independent. Simple algebraic manipulation reveals the dynamics for both the signal and the orthogonal components:

$$\boldsymbol{x}_{\|}^{t+1} = \left\{1 + 3\eta\left(1 - \|\boldsymbol{x}^t\|_2^2\right)\right\} x_{\|}^t; \tag{8a}$$

$$\boldsymbol{x}_\perp^{t+1} = \left\{1 + \eta\left(1 - 3\|\boldsymbol{x}^t\|_2^2\right)\right\} \boldsymbol{x}_\perp^t . \tag{8b}$$

Assuming that η is sufficiently small and recognizing that $\|\boldsymbol{x}^t\|_2^2 = \alpha_t^2 + \beta_t^2$, we arrive at the following population-level state evolution for both α_t and β_t (cf. (7)):

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right]\right\}\alpha_t; \tag{9a}$$

[2]Here, we do not take the absolute value of $\boldsymbol{x}_{\|}^t$. As we shall see later, the $\boldsymbol{x}_{\|}^t$'s are of the same sign throughout the execution of the algorithm.

$$\beta_{t+1} = \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right]\right\}\beta_t . \tag{9b}$$

This recursive system has three *fixed points*:

$$(\alpha, \beta) = (1, 0), \qquad (\alpha, \beta) = (0, 0), \qquad \text{and} \qquad (\alpha, \beta) = (0, 1/\sqrt{3}),$$

which correspond to the global minimizer, the local maximizer, and the saddle points, respectively, of the population objective function.

We make note of the following key observations in the presence of a randomly initialized $\boldsymbol{x}^0$, which will be formalized later in Lemma 1:

- the ratio α_t/β_t of the size of the signal component to that of the orthogonal component increases exponentially fast;
- the size α_t of the signal component keeps growing until it plateaus around 1;
- the size β_t of the orthogonal component eventually drops towards zero.

In other words, when randomly initialized, (α^t, β^t) converges to (1, 0) rapidly, thus indicating rapid convergence of $\boldsymbol{x}^t$ to the truth $\boldsymbol{x}^\natural$, without getting stuck at any undesirable saddle points. We also illustrate these phenomena numerically. Set $n = 1000$, $\eta_t \equiv 0.1$ and $\boldsymbol{x}^0 \sim \mathcal{N}\left(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n\right)$. Figure 4 displays the dynamics of α_t/β_t, α_t, and β_t, which are precisely as discussed above.

2.2 Finite-sample analysis: a heuristic treatment

We now move on to the finite-sample regime, and examine how many samples are needed in order for the population dynamics to be reasonably accurate. Notably, the arguments in this subsection are heuristic in nature, but they are useful in developing insights into the true dynamics of the GD iterates.

Rewrite the gradient update rule (3) as

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta\nabla f\left(\boldsymbol{x}^t\right) = \boldsymbol{x}^t - \eta\nabla F\left(\boldsymbol{x}^t\right) - \underbrace{\eta\left(\nabla f\left(\boldsymbol{x}^t\right) - \nabla F\left(\boldsymbol{x}^t\right)\right)}_{:\,= r\left(\boldsymbol{x}^t\right)}, \tag{10}$$

where $\nabla f(\boldsymbol{x}) = m^{-1}\sum_{i=1}^m \left[\left(a_i^\top \boldsymbol{x}\right)^2 - \left(a_i^\top \boldsymbol{x}^\natural\right)^2\right] a_i a_i^\top \boldsymbol{x}$. Assuming (unreasonably) that the iterate $\boldsymbol{x}^t$ is *independent of* $\{\boldsymbol{a}_i\}$, the central limit theorem (CLT) allows us to control the size of the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$. Take the signal component as an example: simple calculations give

$$x_{\|}^{t+1} = x_{\|}^t - \eta\left(\nabla F\left(\boldsymbol{x}^t\right)\right)_1 - \eta r_1\left(\boldsymbol{x}^t\right),$$

where

$$r_1(\boldsymbol{x}) := \frac{1}{m}\sum_{i=1}^{m}\left[\left(a_i^\top \boldsymbol{x}\right)^3 - a_{i,1}^2\left(a_i^\top \boldsymbol{x}\right)\right]a_{i,1} - \mathbb{E}\left[\left\{\left(a_i^\top \boldsymbol{x}\right)^3 - a_{i,1}^2\left(a_i^\top \boldsymbol{x}\right)\right\}a_{i,1}\right] \tag{11}$$

with $a_{i,1}$ the first entry of $\boldsymbol{a}_i$. Owing to the preceding independence assumption, r_1 is the sum of m i.i.d. zero-mean random variables. Assuming that $\boldsymbol{x}^t$ never blows up so that $\|\boldsymbol{x}^t\|_2 = O(1)$, one can apply the CLT to demonstrate that

$$\left|r_1(\boldsymbol{x}^t)\right| \lesssim \sqrt{\mathrm{Var}(r_1(\boldsymbol{x}^t))\mathrm{polylog}(m)} \lesssim \sqrt{\frac{\mathrm{polylog}(m)}{m}} \tag{12}$$

with high probability, which is often negligible compared to the other terms. For instance, for the random initial guess $\boldsymbol{x}^0 \sim \mathcal{N}\left(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n\right)$ one has $\left|\boldsymbol{x}_{\|}^0\right| \gtrsim 1/\sqrt{n\log n}$ with probability approaching one, telling us that

$$\left|r_1\left(\boldsymbol{x}^0\right)\right| \lesssim \sqrt{\frac{\mathrm{polylog}(m)}{m}} \ll \left|\boldsymbol{x}_{\|}^0\right|$$

as long as $m \gtrsim n\,\mathrm{poly}\log(m)$. This combined with the fact that $\left|\boldsymbol{x}_{\|}^0 - \eta\left(\nabla F\left(\boldsymbol{x}^0\right)\right)_1\right| \asymp \left|\boldsymbol{x}_{\|}^0\right|$ reveals $\left|r_1\left(\boldsymbol{x}^0\right)\right| \lesssim \left|\boldsymbol{x}_{\|}^0 - \eta\left(\nabla F\left(\boldsymbol{x}^0\right)\right)_1\right|$. Similar observations hold true for the orthogonal component $\boldsymbol{x}_{\perp}^t$.

In summary, by assuming independence between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_i\}$, we arrive at an approximate state evolution for the finite-sample regime:

$$\alpha_{t+1} \approx \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right]\right\}\alpha_t; \tag{13a}$$

$$\beta_{t+1} \approx \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right]\right\}\beta_t, \tag{13b}$$

with the proviso that $m \gtrsim n\,\mathrm{poly}\log(m)$.

2.3 Key analysis ingredients: near-independence and leave-one-out tricks

The preceding heuristic argument justifies the approximate validity of the population dynamics, under an independence assumption that never holds unless we use fresh samples in each iteration. On closer inspection, what we essentially need is the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$ (cf. (10)) being well-controlled. For instance, when focusing on the signal component, one need $\left|r_1(\boldsymbol{x}^t)\right| \ll \left|\boldsymbol{x}_{\|}^t\right|$ for all $t \geq 0$. In particular, in the beginning iterations, $\left|\boldsymbol{x}_{\|}^t\right|$ is as small as $O(1/\sqrt{n})$. Without the independence assumption, the CLT types of results fail to hold due to the complicated dependency between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_i\}$. In fact, one can easily find many points that result in much larger remainder terms (as large as $O(1)$) and that violate the approximate state evolution (13). See Figure 5 for a caricature of the region where the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$ is well-controlled. As can be seen, it only occupies a tiny fraction of the neighborhood of $\boldsymbol{x}^\natural$.

Fortunately, despite the complicated dependency across iterations, one can provably guarantee that $\boldsymbol{x}^t$ always stays within the preceding desirable region in which $\boldsymbol{r}(\boldsymbol{x}^t)$ is well-controlled. The key idea is to exploit a certain "near-independence" property between $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{a}_i\}$. Towards this, we make use of a leave-one-out trick proposed in [MWCC17] for analyzing nonconvex iterative methods. In particular, we construct auxiliary sequences that are

1. independent of *certain components* of the design vectors $\{\boldsymbol{a}_i\}$; and
2. extremely close to the original gradient sequence $\{\boldsymbol{x}^t\}_{t\geq 0}$.

As it turns out, we need to construct several auxiliary sequences $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$, $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}_{t\geq 0}$ and $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}_{t\geq 0}$, where $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$ is independent of the lth sampling vector $\boldsymbol{a}_l$, $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}_{t\geq 0}$ is independent of the sign information of the first entries of all $\boldsymbol{a}_i$'s, and $\{\boldsymbol{x}^{t,\mathrm{sgn};(l)}\}$ is independent of both. In addition, these auxiliary sequences are constructed by slightly perturbing the original data (see Figure 6 for an illustration), and hence one can expect all of them to stay close to the original sequence throughout the execution of the algorithm. Taking these two properties together, one can propagate the above statistical independence underlying each auxiliary sequence to the true iterates $\{\boldsymbol{x}^t\}$, which in turn allows us to obtain near-optimal control of the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$. The details are postponed to Section 4.

3 Related work

Solving systems of quadratic equations, or phase retrieval, has been studied extensively in the recent literature; see [SEC+15] for an overview. One popular method is convex relaxation (e.g. *PhaseLift* [CSV13]), which is guaranteed to work as long as m/n exceeds some large enough constant [CL14,DH14,CCG15,CZ15,KRT17]. However, the resulting semidefinite program is computationally prohibitive for solving large-scale problems. To address this issue, [CLS15] proposed the Wirtinger flow algorithm with spectral initialization, which provides the first convergence guarantee for nonconvex methods without sample splitting. Both the sample and computation complexities were further improved by [CC17] with an adaptive truncation strategy. Other nonconvex phase retrieval methods include [NJS13,CLM16,Sol17,WGE17,ZZLC17,WGSC17,CL16,DR17,GX16,CFL15,Wei15,BEB17,TV17,CLW17,ZWGC17,QZEW17,ZCL16,YYF+17,CWZG17,Zha17,MXM18,CLC18]. Almost all of these nonconvex methods require carefully-designed initialization to guarantee a sufficiently accurate initial point. One exception is the approximate message passing algorithm proposed in [MXM18], which works as long as the correlation between the truth and the initial signal is bounded away from zero. This, however, does not accommodate the case when the initial signal strength is vanishingly small (like random initialization). Other works [Zha17,LGL15] explored the global convergence of alternating minimization/projection with random initialization which, however, require fresh samples at least in each of the first $O(\log n)$ iterations in order to enter the local basin. In addition, [LMZ17] explored low-rank recovery from quadratic measurements with near-zero initialization. Using a truncated least-squares objective, [LMZ17] established approximate (but non-exact) recovery of over-parametrized GD. Notably, if we do not over-parametrize the phase retrieval problem, then GD with near-zero initialization is (nearly) equivalent to running the power method for spectral initialization[3], which can be understood using prior theory.

Another related line of research is the design of generic saddle-point escaping algorithms, where the goal is to locate a second-order stationary point (i.e. the point with a vanishing gradient and a positive-semidefinite Hessian). As mentioned earlier, it has been shown by [SQW16] that as soon as $m \gg \log^3 n$, all local minima are global and all the saddle points are strict. With these two geometric properties in mind, saddle-point escaping algorithms are guaranteed to converge globally for phase retrieval. Existing saddle-point escaping algorithms include but are not limited to Hessian-based methods [NP06,SQW16] (see also [AAZB+16,AZ17,JGN+17] for some reviews), noisy stochastic gradient descent [GHJY15], perturbed gradient descent [JGN+17], and normalized gradient descent [MSK17]. On the one hand, the results developed in these works are fairly general: they establish polynomial-time convergence guarantees under a few generic geometric conditions. On the other hand, the iteration complexity derived therein may be pessimistic when specialized to a particular problem.

Take phase retrieval and the perturbed gradient descent algorithm [JGN+17] as an example. It has been shown in [JGN+17, Theorem 5] that for an objective function that is L-gradient Lipschitz, ρ-Hessian Lipschitz, (θ, γ, ζ)-strict saddle, and also locally α-strongly convex and β-smooth (see definitions in [JGN+17]), it takes[4]

$$O\left(\frac{L}{\left[\min\left(\theta, \gamma^2/\rho\right)\right]^2} + \frac{\beta}{\alpha}\log\frac{1}{\epsilon}\right) = O\left(n^3 + n\log\frac{1}{\epsilon}\right)$$

iterations (ignoring logarithmic factors) for perturbed gradient descent to converge to ϵ-accuracy. In fact, even with Nesterov's accelerated scheme [JNJ17], the iteration complexity for entering the local region is at least

$$O\left(\frac{L^{1/2}\rho^{1/4}}{\left[\min\left(\theta, \gamma^2/\rho\right)\right]^{7/4}}\right) = O\left(n^{2.5}\right).$$

Both of them are much larger than the $O(\log n + \log(1/\epsilon))$ complexity established herein. This is primarily due to the following facts: (i) the Lipschitz constants of both the gradients and the Hessians are quite large, i.e. $L \asymp n$ and $\rho \asymp n$ (ignoring log factors), which are, however, treated as dimension-independent constants in the aforementioned papers; (ii) the local condition number is also large, i.e. $\beta/\alpha \asymp n$. In comparison, as suggested by our theory, the GD iterates with random initialization are always confined within a restricted region enjoying much more benign geometry than the worst-case / global characterization.

Furthermore, the above saddle-escaping first-order methods are often more complicated than vanilla GD. Despite its algorithmic simplicity and wide use in practice, the convergence rate

[3]More specifically, the GD update $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - m^{-1}\eta_t \sum_{i=1}^m \left[\left(a_i^\top \boldsymbol{x}^t\right)^2 - y_i\right] a_i a_i^\top \boldsymbol{x}_t \approx \left(\boldsymbol{I} + m^{-1}\eta_t \sum_{i=1}^m y_i a_i a_i^\top\right)\boldsymbol{x}_t$ when $\boldsymbol{x}_t \approx 0$, which is equivalent to a power iteration (without normalization) w.r.t. the data matrix $\boldsymbol{I} + m^{-1}\eta_t \sum_{i=1}^m y_i a_i a_i^\top$.

[4]When applied to phase retrieval with $m \asymp n$ poly log n, one has $L \asymp n$, $\rho \asymp n$, $\theta \asymp \gamma \asymp 1$ (see [SQW16, Theorem 2.2]), $\alpha \asymp 1$, and $\beta \lesssim n$ (ignoring logarithmic factors).

of GD with random initialization remains largely unknown. In fact, Du et al. [DJL+17] demonstrated that there exist non-pathological functions such that GD can take exponential time to escape the saddle points when initialized randomly. In contrast, as we have demonstrated, saddle points are not an issue for phase retrieval; the GD iterates with random initialization never get trapped in the saddle points.

Finally, the leave-one-out arguments have been invoked to analyze other high-dimensional statistical inference problems including robust M-estimators [EKBB+13,EK15], and maximum likelihood theory for logistic regression [SCC18], etc. In addition, [ZB17,CFMW17,AFWZ17] made use of the leave-one-out trick to derive entrywise perturbation bounds for eigenvectors resulting from certain spectral methods. The techniques have also been applied by [MWCC17,LMCC18] to establish local linear convergence of vanilla GD for nonconvex statistical estimation problems in the presence of proper spectral initialization.

4 Analysis

In this section, we first provide a more general version of Theorem 1 as follows. It spells out exactly the conditions on $\boldsymbol{x}^0$ in order for vanilla GD with random initialization to succeed.

Theorem 2

Fix $\boldsymbol{x}^\natural \in \mathbb{R}^n$. *Suppose* $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)(1 \le i \le m)$ *and* $m \ge Cn \log^{13} m$ *for some sufficiently large constant* $C > 0$. *Assume that the initialization* $\boldsymbol{x}^0$ *is independent of* $\{\boldsymbol{a}_i\}$ *and obeys*

$$\frac{\left|\left\langle \boldsymbol{x}^0, \boldsymbol{x}^\natural \right\rangle\right|}{\|\boldsymbol{x}^\natural\|_2^2} \ge \frac{1}{\sqrt{n}\log n} \qquad \text{and} \qquad \left(1 - \frac{1}{\log n}\right)\|\boldsymbol{x}^\natural\|_2 \le \|\boldsymbol{x}^0\|_2 \le \left(1 + \frac{1}{\log n}\right)\|\boldsymbol{x}^\natural\|_2, \tag{14}$$

and that the stepsize satisfies $\eta_t \equiv \eta = c/\|\boldsymbol{x}^\natural\|_2^2$ *for some sufficiently small constant* $c > 0$. *Then there exist a sufficiently small absolute constant* $0 < \gamma < 1$ *and* $T_\gamma \lesssim \log n$ *such with probability at least* $1 - O(m^2 e^{-1.5n}) - O(m^{-9})$,

1. *the GD iterates (3) converge linearly to* $\boldsymbol{x}^\natural$ *after* $t \ge T_\gamma$, *namely,*

$$\operatorname{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \le \left(1 - \frac{\eta}{2}\|\boldsymbol{x}^\natural\|_2^2\right)^{t - T_\gamma} \cdot \gamma\|\boldsymbol{x}^\natural\|_2, \qquad \forall t \ge T_\gamma;$$

2. the strength ratio of the signal component $\frac{\left\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \right\rangle}{\|\boldsymbol{x}^\natural\|_2^2}\boldsymbol{x}^\natural$ to the orthogonal component $\boldsymbol{x}^t - \frac{\left\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \right\rangle}{\|\boldsymbol{x}^\natural\|_2^2}\boldsymbol{x}^\natural$ obeys

$$\frac{\left\| \frac{\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle}{\|\boldsymbol{x}^\natural\|_2^2} \boldsymbol{x}^\natural \right\|_2}{\left\| \boldsymbol{x}^t - \frac{\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle}{\|\boldsymbol{x}^\natural\|_2^2} \boldsymbol{x}^\natural \right\|_2} \gtrsim \frac{1}{\sqrt{n \log n}} \left(1 + c_1 \eta^2\right)^t, \qquad t = 0, 1, \cdots \tag{15}$$

for some constant $c_1 > 0$.

Several remarks regarding Theorem 2 are in order.

- Our current sample complexity reads $m \gtrsim n \log^{13} m$, which is optimal up to logarithmic factors. It is possible to further reduce the logarithmic factors using more refined probabilistic tools, which we leave for future work.
- We can also prove similar performance guarantees for noisy phase retrieval. For brevity, we do not provide the exact theorem and the detailed proofs. The readers will find them in the last author's Ph.D. thesis.
- The random initialization $\boldsymbol{x}^0 \sim \mathcal{N}\left(\boldsymbol{0}, n^{-1}\|\boldsymbol{x}^\natural\|_2^2 \boldsymbol{I}_n\right)$ obeys the condition (14) with probability exceeding $1 - O(1/\sqrt{\log n})$, which in turn establishes Theorem 1.
- Theorem 2 requires an initialization $\boldsymbol{x}^0$ which is independent of the data and the knowledge of $\|\boldsymbol{x}^\natural\|$, which is not practical. One possible method is to estimate it from the data, which results in an initial value that depends on the data. The following theorem demonstrate both independent initial value and known $\|\boldsymbol{x}^\natural\|$ are not necessary, resulting a practical algorithm.

Theorem 3

Let

$$\boldsymbol{x}^0 = \sqrt{\frac{1}{m} \sum_{i=1}^{m} y_i} \cdot \boldsymbol{u},$$

where $\mathbf{u}$ is uniformly distributed over the unit sphere. With probability at least $1 - O(1/\sqrt{\log n})$ all the claims in Theorem 2 continue to hold.

Proof. The proof is very similar to that of Theorem 2, with only a few changes. See Appendix N for detailed explanations. □

The remainder of this section is then devoted to proving Theorem 2. Without loss of generality[5], we will assume throughout that

$$\boldsymbol{x}^\natural = e_1 \qquad \text{and} \qquad \boldsymbol{x}_1^0 > 0. \tag{16}$$

Given this, one can decompose

[5]This is because of the rotational invariance of Gaussian distributions.

$$\boldsymbol{x}^t = \boldsymbol{x}^t_{\|} e_1 + \begin{bmatrix} 0 \\ \boldsymbol{x}^t_{\perp} \end{bmatrix} \tag{17}$$

where $\boldsymbol{x}^t_{\|} = \boldsymbol{x}^t_1$ and $\boldsymbol{x}^t_{\perp} = \left[\boldsymbol{x}^t_i\right]_{2 \le i \le n}$ as introduced in Section 2. For notational simplicity, we define

$$\alpha_t := \boldsymbol{x}^t_{\|} \qquad \text{and} \qquad \beta_t := \left\|\boldsymbol{x}^t_{\perp}\right\|_2 . \tag{18}$$

Intuitively, α_t represents the size of the signal component, whereas β_t measures the size of the component orthogonal to the signal direction. In view of (16), we have $\alpha_0 > 0$.

4.1 Outline of the proof

To begin with, it is easily seen that if α_t and β_t (cf. (18)) obey $|\alpha_t - 1| \le \gamma/2$ and $\beta_t \le \gamma/2$, then

$$\mathrm{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \le \left\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\right\|_2 \le |\alpha_t - 1| + |\beta_t| \le \gamma .$$

Therefore, our first step — which is concerned with proving $\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \le \gamma$ — comes down to the following two steps.

1. Show that if α_t and β_t satisfy the approximate state evolution (see (13)), then there exists some $T_\gamma = O(\log n)$ such that

$$\left|\alpha_{T_\gamma} - 1\right| \le \gamma/2 \qquad \text{and} \qquad \beta_{T_\gamma} \le \gamma/2, \tag{19}$$

   which would immediately imply that

$$\mathrm{dist}\left(\boldsymbol{x}^{T_\gamma}, \boldsymbol{x}^\natural\right) \le \gamma .$$

   Along the way, we will also show that the ratio α_t/β_t grows exponentially fast.

2. Justify that α_t and β_t satisfy the approximate state evolution with high probability, using (some variants of) leave-one-out arguments.

After $t \ge T_\gamma$, we can invoke prior theory [MWCC17] concerning local convergence to show that with high probability,

$$\mathrm{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \le (1-\rho)^{t - T_\gamma} \left\|\boldsymbol{x}^{T_\gamma} - \boldsymbol{x}^\natural\right\|_2, \qquad \forall t > T_\gamma$$

for some constant $0 < \rho < 1$ independent of n and m.

4.2 Dynamics of approximate state evolution

This subsection formalizes our intuition in Section 2: as long as the approximate state evolution holds, then one can find $T_\gamma \lesssim \log n$ obeying condition (19). In particular, the approximate state evolution is given by

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\zeta_t\right\}\alpha_t \tag{20a}$$

$$\beta_{t+1} = \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\rho_t\right\}\beta_t, \tag{20b}$$

where $\{\zeta_t\}$ and $\{\rho_t\}$ represent the perturbation terms. Our result is this:

Lemma 1—*Let $\gamma > 0$ be some sufficiently small constant, and consider the approximate state evolution* (20)*. Suppose the initial point obeys*

$$\alpha_0 \geq \frac{1}{\sqrt{n\log n}} \qquad \text{and} \qquad 1 - \frac{1}{\log n} \leq \sqrt{\alpha_0^2 + \beta_0^2} \leq 1 + \frac{1}{\log n}. \tag{21}$$

and the perturbation terms satisfy

$$\max\{|\zeta_t|, |\rho_t|\} \leq \frac{c_3}{\log n}, t = 0, 1, \cdots$$

for some sufficiently small constant $c_3 > 0$.

(a) Let

$$T_\gamma := \min\{t : |\alpha_t - 1| \leq \gamma/2 \text{ and } \beta_t \leq \gamma/2\}. \tag{22}$$

Then for any sufficiently large n and m and any sufficiently small constant $\eta > 0$, one has

$$T_\gamma \lesssim \log n, \tag{23}$$

and there exist some constants c_5, $c_{10} > 0$ independent of n and m such that

$$\frac{1}{2\sqrt{n\log n}} \leq \alpha_t \leq 2, \quad c_5 \leq \beta_t \leq 1.5 \quad \text{and} \quad \frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} \geq 1 + c_{10}\eta^2, \qquad 0 \leq t \leq T_\gamma \tag{24}$$

.

(b) If we define

$$T_0 := \min\left\{t : \alpha_{t+1} \geq c_6/\log^5 m\right\}, \tag{25}$$

$$T_1 := \min\{t : \alpha_{t+1} > c_4\}, \tag{26}$$

for some arbitrarily small constants c_4, $c_6 > 0$, then

1. $T_0 \le T_1 \le T_\gamma \lesssim \log n$; $T_1 - T_0 \lesssim \log\log m$; $T_\gamma - T_1 \lesssim 1$;
2. *For* $T_0 < t \le T_\gamma$, *one has* $\alpha_t \ge c_6 \lambda \log^5 m$.

Proof. See Appendix B. □

Remark 2—Recall that γ is sufficiently small and $(\alpha, \beta) = (1, 0)$ represents the global minimizer. Since $|\alpha_0 - 1| \approx 1$, one has $T_\gamma > 0$, which denotes the first time when the iterates enter the local region surrounding the global minimizer. In addition, the fact that $\alpha_0 \lesssim 1/\sqrt{n}$ gives $T_0 > 0$ and $T_1 > 0$, both of which indicate the first time when the signal strength is sufficiently large.

Lemma 1 makes precise that under the approximate state evolution, the first stage enjoys a fairly short duration $T_\gamma \lesssim \log n$. Moreover, the size of the signal component grows faster than that of the orthogonal component for any iteration $t < T_\gamma$, thus confirming the exponential growth of α_t/β_t.

In addition, Lemma 1 identifies two midpoints T_0 and T_1 when the sizes of the signal component α_t become sufficiently large. These are helpful in our subsequent analysis. In what follows, we will divide Stage 1 (which consists of all iterations up to T_γ) into two phases:

- *Phase I*: consider the duration $0 \le t \le T_0$;
- *Phase II*: consider all iterations with $T_0 < t \le T_\gamma$.

We will justify the approximate state evolution (20) for these two phases separately.

4.3 Motivation of the leave-one-out approach

As we have alluded in Section 2.3, the main difficulty in establishing the approximate state evolution (20) lies in controlling the perturbation terms to the desired orders (i.e. $|\zeta_t|, |\rho_t| \ll 1/\log n$ in Lemma 1). To achieve this, we advocate the use of (some variants of) leave-one-out sequences to help establish certain "near-independence" between $\boldsymbol{x}^t$ and certain components of $\{\boldsymbol{a}_i\}$.

We begin by taking a closer look at the perturbation terms. Regarding the signal component, it is easily seen from (11) that

$$\boldsymbol{x}_{\|}^{t+1} = \left\{1 + 3\eta\left(1 - \|\boldsymbol{x}^t\|_2^2\right)\right\}\boldsymbol{x}_{\|}^t - \eta r_1(\boldsymbol{x}^t),$$

where the perturbation term $r_1(\boldsymbol{x}^t)$ obeys

$$r_1(\boldsymbol{x}^t) = \underbrace{\left[1-\left(\boldsymbol{x}_{\|}^t\right)^2\right]\boldsymbol{x}_{\|}^t\left(\frac{1}{m}\sum_{i=1}^m a_{i,1}^4 - 3\right)}_{:=\boldsymbol{I}_1} + \underbrace{\left[1-3\left(\boldsymbol{x}_{\|}^t\right)^2\right]\frac{1}{m}\sum_{i=1}^m a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t}_{:=\boldsymbol{I}_2}$$
$$\underbrace{-3\boldsymbol{x}_{\|}^t\left(\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 a_{i,1}^2 - \left\|\boldsymbol{x}_\perp^t\right\|_2^2\right)}_{:=\boldsymbol{I}_3} - \underbrace{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 a_{i,1}}_{:=\boldsymbol{I}_4}. \tag{27}$$

Here and throughout the paper, for any vector $v \in \mathbb{R}^n$, $v_\perp \in \mathbb{R}^{n-1}$ denotes the 2nd through the nth entries of $\boldsymbol{v}$. Due to the dependency between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_i\}$, it is challenging to obtain sharp control of some of these terms.

In what follows, we use the term I_4 to explain and motivate our leave-one-out approach. As discussed in Section 2.3, I_4 needs to be controlled to the level $O(1/(\sqrt{n}\text{polylog}(n)))$. This precludes us from seeking a uniform bound on the function $h(\boldsymbol{x}) := m^{-1}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp\right)^3 a_{i,1}$ over all $\boldsymbol{x}$ (or even all $\boldsymbol{x}$ within the set $\mathscr{C}$ incoherent with $\{\boldsymbol{a}_i\}$), since the uniform bound $\sup_{\boldsymbol{x}\in\mathscr{C}}|h(\boldsymbol{x})|$ can be $O(\sqrt{n}/\text{polylog}(n))$ times larger than the desired order.

Algorithm 1

The lth leave-one-out sequence

Input: $\{\boldsymbol{a}_i\}_{1\le i\le m, i\ne l}$, $\{y_i\}_{1\le i\le m, i\ne l}$, and $\boldsymbol{x}^0$.

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

$$\boldsymbol{x}^{t+1,(l)} = \boldsymbol{x}^{t,(l)} - \eta_t \nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right), \tag{29}$$

where $\boldsymbol{x}^{0,(l)} = \boldsymbol{x}_0$ and $f^{(l)}(\boldsymbol{x}) = (1/4m)\cdot\sum_{i:i\ne l}\left[\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2\right]^2$.

In order to control I_4 to the desirable order, one strategy is to approximate it by a sum of independent variables and then invoke the CLT. Specifically, we first rewrite I_4 as

$$\boldsymbol{I}_4 = \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 |a_{i,1}|\xi_i$$

with $\xi_i := \text{sgn}(a_{i,1})$. Here $\text{sgn}(\cdot)$ denotes the usual sign function. To exploit the statistical independence between ξ_i and $\{|a_{i,1}|, \boldsymbol{a}_{i,\perp}\}$, we would like to identify some vector independent of ξ_i that well approximates $\boldsymbol{x}^t$. If this can be done, then one may treat I_4 as a weighted independent sum of $\{\xi_i\}$. Viewed in this light, our plan is the following:

1. Construct a sequence $\{\boldsymbol{x}^{t,\text{sgn}}\}$ independent of $\{\xi_i\}$ obeying $\boldsymbol{x}^{t,\text{sgn}} \approx \boldsymbol{x}^t$, so that

$$\boldsymbol{I}_4 \approx \frac{1}{m}\sum_{i=1}^m \underbrace{\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\text{sgn}}\right)^3 |a_{i,1}|}_{:=w_i}\xi_i.$$

One can then apply standard concentration results (e.g. the Bernstein inequality) to control I_4, as long as none of the weight w_i is exceedingly large.

2. Demonstrate that the weight w_i is well-controlled, or equivalently, $|a_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}| (1 \le i \le m)$ is not much larger than its typical size. This can be accomplished by identifying another sequence $\{\boldsymbol{x}^{t,(i)}\}$ independent of $\boldsymbol{a}_i$ such that $\boldsymbol{x}^{t,(i)} \approx \boldsymbol{x}^{t} \approx \boldsymbol{x}^{t,\mathrm{sgn}}$, followed by the argument:

$$\left|a_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right| \approx \left|a_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right| \approx \left|a_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,(i)}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,(i)}\right\|_2 \approx \sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2. \tag{28}$$

Here, the inequality follows from standard Gaussian tail bounds and the independence between $\boldsymbol{a}_i$ and $\boldsymbol{x}^{t,(i)}$. This explains why we would like to construct $\{\boldsymbol{x}^{t,(i)}\}$ for each $1 \le i \le m$.

As we will detail in the next subsection, such auxiliary sequences are constructed by leaving out a small amount of relevant information from the collected data before running the GD algorithm, which is a variant of the "leave-one-out" approach rooted in probability theory and random matrix theory.

4.4 Leave-one-out and random-sign sequences

We now describe how to design auxiliary sequences to help establish certain independence properties between the gradient iterates $\{\boldsymbol{x}^t\}$ and the design vectors $\{\boldsymbol{a}_i\}$. In the sequel, we formally define the three sets of auxiliary sequences $\{\boldsymbol{x}^{t,(l)}\}, \{\boldsymbol{x}^{t,\mathrm{sgn}}\}, \{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}$ as introduced in Section 2.3 and Section 4.3.

- *Leave-one-out sequences* $\{\boldsymbol{x}^{t,(l)}\}_{t\ge 0}$. For each $1 \le l \le m$, we introduce a sequence $\{\boldsymbol{x}^{t,(l)}\}$, which drops the lth sample and runs GD w.r.t. the auxiliary objective function

$$f^{(l)}(\boldsymbol{x}) = \frac{1}{4m} \sum_{i:\, i \ne l} \left[\left(a_i^{\top} \boldsymbol{x}\right)^2 - \left(a_i^{\top} \boldsymbol{x}^{\natural}\right)^2\right]^2. \tag{32}$$

See Algorithm 1 for details and also Figure 6(a) for an illustration. One of the most important features of $\{\boldsymbol{x}^{t,(l)}\}$ is that all of its iterates are statistically independent of $(\boldsymbol{a}_l, y_l)$, and hence are incoherent with $\boldsymbol{a}_l$ with high probability, in the sense that $\left|a_l^{\top} \boldsymbol{x}^{t,(l)}\right| \lesssim \sqrt{\log m}\|\boldsymbol{x}^{t,(l)}\|_2$. Such incoherence properties further allow us to control both $\left|a_l^{\top} \boldsymbol{x}^{t}\right|$ and $\left|a_l^{\top} \boldsymbol{x}^{t,\mathrm{sgn}}\right|$ (see (28)), which is crucial for controlling the size of the residual terms (e.g. $r_1(\boldsymbol{x}^t)$ as defined in (11)). Notably, the sequence $\{\boldsymbol{x}^{t,(l)}\}$ has also been applied by [MWCC17] to justify the success of GD with spectral initialization for several nonconvex statistical estimation problems.

Algorithm 2

The random-sign sequence

Input: $\{|a_{i,1}|\}_{1\le i\le m}$, $\{\boldsymbol{a}_{i,\perp}\}_{1\le i\le m}$, $\left\{\xi_i^{\mathrm{sgn}}\right\}_{1 \le i \le m}$, $\{y_i\}_{1\le i\le m}$, $\boldsymbol{x}^0$.

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

$$\boldsymbol{x}^{t+1,\mathrm{sgn}} = \boldsymbol{x}^{t,\mathrm{sgn}} - \eta_t \nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right), \tag{30}$$

where $\boldsymbol{x}^{0,\mathrm{sgn}} = \boldsymbol{x}^0$, $f^{\mathrm{sgn}}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i=1}^{m}\left[\left(a_i^{\mathrm{sgn}\top}\boldsymbol{x}\right)^2 - \left(a_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural\right)^2\right]^2$ with $a_i^{\mathrm{sgn}} := \begin{bmatrix}\xi_i^{\mathrm{sgn}}|a_{i,1}| \\ a_{i,\perp}\end{bmatrix}$.

Algorithm 3

The lth leave-one-out and random-sign sequence

Input: $\{|a_{i,1}|\}_{1\le i\le m, i\neq l}$, $\{\boldsymbol{a}_{i,\perp}\}_{1\le i\le m, i\neq l}$, $\left\{\xi_i^{\mathrm{sgn}}\right\}_{1\le i\le m, i\neq l}$, $\{y_i\}_{1\le i\le m, i\neq l}$, $\boldsymbol{x}^0$.

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

$$\boldsymbol{x}^{t+1,\mathrm{sgn},(l)} = \boldsymbol{x}^{t,\mathrm{sgn},(l)} - \eta_t \nabla f^{\mathrm{sgn},(l)}\left(\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right), \tag{31}$$

where $\boldsymbol{x}^{0,\mathrm{sgn},(l)} = \boldsymbol{x}^0$, $f^{\mathrm{sgn},(l)}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i:i\neq l}\left[\left(a_i^{\mathrm{sgn}\top}\boldsymbol{x}\right)^2 - \left(a_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural\right)^2\right]^2$ with $a_i^{\mathrm{sgn}} := \begin{bmatrix}\xi_i^{\mathrm{sgn}}|a_{i,1}| \\ a_{i,\perp}\end{bmatrix}$.

- *Random-sign sequence* $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}_{t\ge 0}$. Introduce a collection of auxiliary design vectors $\left\{a_i^{\mathrm{sgn}}\right\}_{1\le i\le m}$ defined as

$$a_i^{\mathrm{sgn}} := \begin{bmatrix}\xi_i^{\mathrm{sgn}}|a_{i,1}| \\ a_{i,\perp}\end{bmatrix}, \tag{33}$$

where $\left\{\xi_i^{\mathrm{sgn}}\right\}_{1\le i\le m}$ is a set of Rademacher random variables independent of $\{\boldsymbol{a}_j\}$, i.e.

$$\xi_i^{\mathrm{sgn}} \overset{\text{i.i.d.}}{=} \begin{cases}1, & \text{with probability } 1/2,\\ -1, & \text{else}\end{cases} \quad 1\le i\le m. \tag{34}$$

In words, a_i^{sgn} is generated by randomly flipping the sign of the first entry of $\boldsymbol{a}_i$. To simplify the notations hereafter, we also denote

$$\xi_i = \mathrm{sgn}(a_{i,1}). \tag{35}$$

As a result, $\boldsymbol{a}_i$ and a_i^{sgn} differ only by a single bit of information. With these auxiliary design vectors in place, we generate a sequence $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ by running GD w.r.t. the auxiliary loss function

$$f^{\mathrm{sgn}}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i=1}^{m}\left[\left(a_i^{\mathrm{sgn}\top}\boldsymbol{x}\right)^2 - \left(a_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural\right)^2\right]^2. \tag{36}$$

One simple yet important feature associated with these new design vectors is that it produces the same measurements as $\{\boldsymbol{a}_i\}$:

$$\left(a_i^\top \boldsymbol{x}^\natural\right)^2 = \left(a_i^{\mathrm{sgn}\top} \boldsymbol{x}^\natural\right)^2 = \left|a_{i,1}\right|^2, 1 \le i \le m\,. \tag{37}$$

See Figure 6(b) for an illustration and Algorithm 2 for the detailed procedure. This sequence is introduced in order to "randomize" certain Gaussian polynomials (e.g. I_4 in (27)), which in turn enables optimal control of these quantities. This is particularly crucial at the initial stage of the algorithm.

- *Leave-one-out and random-sign sequences* $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}_{t\ge 0}$. Furthermore, we also need to introduce another collection of sequences $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}$ by simultaneously employing the new design vectors $\left\{a_i^{\mathrm{sgn}}\right\}$ and discarding a single sample $\left(a_l^{\mathrm{sgn}}, y_l^{\mathrm{sgn}}\right)$.. This enables us to propagate the kinds of independence properties across the above two sets of sequences, which is useful in demonstrating that $\boldsymbol{x}^t$ is jointly "nearly-independent" of both $\boldsymbol{a}_l$ and $\{\mathrm{sgn}(a_{i,1})\}$. See Algorithm 3 and Figure 6(c).

As a remark, all of these leave-one-out and random-sign procedures are assumed to start from the same initial point as the original sequence, namely,

$$\boldsymbol{x}^0 = \boldsymbol{x}^{0,(l)} = \boldsymbol{x}^{0,\mathrm{sgn}} = \boldsymbol{x}^{0,\mathrm{sgn},(l)}, 1 \le l \le m\,. \tag{38}$$

4.5 Justification of approximate state evolution for Phase I of Stage 1

Recall that Phase I consists of the iterations $0 \le t \le T_0$, where

$$T_0 = \min\left\{t : \alpha_{t+1} \ge \frac{c_6}{\log^5 m}\right\}. \tag{39}$$

Our goal here is to show that the approximate state evolution (20) for both the size α_t of the signal component and the size β_t of the orthogonal component holds true throughout Phase I. Our proof will be inductive in nature. Specifically, we will first identify a set of induction hypotheses that are helpful in proving the validity of the approximate state evolution (20), and then proceed by establishing these hypotheses via induction.

4.5.1 Induction hypotheses—For the sake of clarity, we first list all the induction hypotheses.

$$\max_{1\le l\le m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \le \beta_t\left(1+\frac{1}{\log m}\right)^t C_1\frac{\sqrt{n\log^5 m}}{m}, \tag{40a}$$

$$\max_{1\le l\le m} \left|\boldsymbol{x}_{\|}^t - \boldsymbol{x}_{\|}^{t,(l)}\right| \le \alpha_t\left(1+\frac{1}{\log m}\right)^t C_2\frac{\sqrt{n\log^{12} m}}{m}, \tag{40b}$$

$$\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\|_2 \le \alpha_t\left(1+\frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}}, \tag{40c}$$

$$\max_{1\le l\le m} \|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\|_2 \le \alpha_t\left(1+\frac{1}{\log m}\right)^t C_4\frac{\sqrt{n\log^9 m}}{m}, \tag{40d}$$

$$c_5 \le \|\boldsymbol{x}_\perp^t\|_2 \le \|\boldsymbol{x}^t\|_2 \le C_5, \tag{40e}$$

$$\|\boldsymbol{x}^t\|_2 \le 4\alpha_t\sqrt{n\log m}, \tag{40f}$$

where $C_1,\cdots, C_5$ and c_5 are some absolute positive constants.

Now we are ready to prove an immediate consequence of the induction hypotheses (40): if (40) hold for the t^{th} iteration, then α_{t+1} and β_{t+1} follow the approximate state evolution (see (20)). This is justified in the following lemma.

Lemma 2: *Suppose* $m \ge Cn\log^{11} m$ *for some sufficiently large constant* $C > 0$. *For any* $0 \le t \le T_0$ *(cf.* (39)*), if the* t^{th} *iterates satisfy the induction hypotheses* (40), *then with probability at least* $1 - O(me^{-1.5n}) - O(m^{-10})$,

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\zeta_t\right\}\alpha_t; \tag{41a}$$

$$\beta_{t+1} = \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\rho_t\right\}\beta_t \tag{41b}$$

hold for some $|\zeta_t| \ll 1/\log m$ *and* $|\rho_t| \ll 1/\log m$.

Proof. See Appendix C. □

It remains to inductively show that the hypotheses hold for all $0 \le t \le T_0$. Before proceeding to this induction step, it is helpful to first develop more understanding about the preceding hypotheses.

1. In words, (40a), (40b), (40c) specify that the leave-one-out sequences $\{\boldsymbol{x}^{t,(l)}\}$ and $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ are exceedingly close to the original sequence $\{\boldsymbol{x}^t\}$. Similarly, the difference between $\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}$ and $\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}$ is extremely small, as asserted in (40d). The hypothesis (40e) says that the norm of the iterates $\{\boldsymbol{x}^t\}$ is always bounded from above and from below in Phase I. The last one (40f) indicates that the size α_t of the signal component is never too small compared with $\|\boldsymbol{x}^t\|_2$.
2. Another property that is worth mentioning is the growth rate (with respect to t) of the quantities appeared in the induction hypotheses (40). For instance, $\left|\boldsymbol{x}_{\|}^t - \boldsymbol{x}_{\|}^{t,(l)}\right|$, $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\|_2$ and $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\|_2$ grow more or less at the same rate as α_t (modulo some $(1+1/\log m)^{T_0}$ factor). In contrast, $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\|2$

shares the same growth rate with β_t (modulo the $(1+1/\log m)^{T_0}$ factor). See Figure 7 for an illustration. The difference in the growth rates turns out to be crucial in establishing the advertised result.

3. Last but not least, we emphasize the sizes of the quantities of interest in (40) for $t = 1$ under the Gaussian initialization. Ignoring all of the log m terms and recognizing that $\alpha_1 \asymp 1/\sqrt{n}$ and $\beta_1 \asymp 1$, one sees that $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,(l)}\|_2 \lesssim 1/\sqrt{m}$, $\left|\boldsymbol{x}_{\|}^1 - \boldsymbol{x}_{\|}^{1,(l)}\right| \lesssim 1/m$, $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}}\|_2 \lesssim 1/\sqrt{m}$ and $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}} - \boldsymbol{x}^{1,(l)} + \boldsymbol{x}^{1,\mathrm{sgn},(l)}\|_2 \lesssim 1/m$. See Figure 7 for an illustration of the trends of the above four quantities.

Several consequences of (40) regarding the incoherence between $\{\boldsymbol{x}^t\}$, $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ and $\{\boldsymbol{a}_i\}$, $\{a_i^{\mathrm{sgn}}\}$ are immediate, as summarized in the following lemma.

Lemma 3: *Suppose that $m \geq Cn\log^6 m$ for some sufficiently large constant $C > 0$ and the t^{th} iterates satisfy the induction hypotheses (40) for $t \leq T_0$, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\begin{aligned}
&\max_{1\leq l\leq m}\left|a_l^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log m}\|\boldsymbol{x}^t\|_2;\\
&\max_{1\leq l\leq m}\left|a_{l,\perp}^\top \boldsymbol{x}_\perp^t\right| \lesssim \sqrt{\log m}\|\boldsymbol{x}_\perp^t\|_2;\\
&\max_{1\leq l\leq m}\left|a_l^\top \boldsymbol{x}^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\log m}\|\boldsymbol{x}^{t,\mathrm{sgn}}\|_2;\\
&\max_{1\leq l\leq m}\left|a_{l,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2;\\
&\max_{1\leq l\leq m}\left|a_l^{\mathrm{sgn}\,\top} \boldsymbol{x}^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\log m}\|\boldsymbol{x}^{t,\mathrm{sgn}}\|_2.
\end{aligned}$$

Proof. These incoherence conditions typically arise from the independence between $\{\boldsymbol{x}^{t,(l)}\}$ and a_l. For instance, the first line follows since

$$\left|a_l^\top \boldsymbol{x}^t\right| \approx \left|a_l^\top \boldsymbol{x}^{t,(l)}\right| \lesssim \sqrt{\log m}\|\boldsymbol{x}^{t,(l)}\|_2 \asymp \sqrt{\log m}\|\boldsymbol{x}^t\|_2.$$

See Appendix M for detailed proofs. □

4.5.2 Induction step—We then turn to showing that the induction hypotheses (40) hold throughout Phase I, i.e. for $0 \leq t \leq T_0$. The base case can be easily verified because of the identical initial points (38). Now we move on to the inductive step, i.e. we aim to show that if the hypotheses (40) are valid up to the t^{th} iteration for some $t \leq T_0$, then they continue to hold for the $(t+1)^{\mathrm{th}}$ iteration.

The first lemma concerns the difference between the leave-one-out sequence $x^{t+1,(l)}$ and the true sequence x^{t+1} (see (40a)).

Lemma 4: *Suppose $m \geq Cn\log^5 m$ for some sufficiently large constant $C > 0$. If the induction hypotheses (40) hold true up to the t^{th} iteration for some $t \leq T_0$, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\max_{1 \le l \le m} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\|_2 \le \beta_{t+1}\left(1 + \frac{1}{\log m}\right)^{t+1} C_1 \frac{\sqrt{n\log^5 m}}{m} \tag{43}$$

holds as long as $\eta > 0$ *is a sufficiently small constant and* $C_1 > 0$ *is sufficiently large.*

Proof. See Appendix D. □

The next lemma characterizes a finer relation between $\boldsymbol{x}^{t+1}$ and $\boldsymbol{x}^{t+1,(l)}$ when projected onto the signal direction (cf. (40b)).

Lemma 5: *Suppose* $m \ge Cn\log^6 m$ *for some sufficiently large constant* $C > 0$. *If the induction hypotheses (40) hold true up to the* t^{th} *iteration for some* $t \le T_0$, *then with probability at least* $1 - O(me^{-1.5n}) - O(m^{-10})$,

$$\max_{1 \le l \le m} \left|\boldsymbol{x}_{\|}^{t+1} - \boldsymbol{x}_{\|}^{t+1,(l)}\right| \le \alpha_{t+1}\left(1 + \frac{1}{\log m}\right)^{t+1} C_2 \frac{\sqrt{n\log^{12} m}}{m} \tag{44}$$

holds as long as $\eta > 0$ *is a sufficiently small constant and* $C_2 \gg C_4$.

Proof. See Appendix E. □

Regarding the difference between $\boldsymbol{x}^t$ and $\boldsymbol{x}^{t,\text{sgn}}$ (see (40c)), we have the following result.

Lemma 6: *Suppose* $m \ge Cn\log^5 m$ *for some sufficiently large constant* $C > 0$. *If the induction hypotheses (40) hold true up to the* t^{th} *iteration for some* $t \le T_0$, *then with probability at least* $1 - O(me^{-1.5n}) - O(m^{-10})$,

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,\text{sgn}}\|_2 \le \alpha_{t+1}\left(1 + \frac{1}{\log m}\right)^{t+1} C_3\sqrt{\frac{n\log^5 m}{m}} \tag{45}$$

holds as long as $\eta > 0$ *is a sufficiently small constant and* C_3 *is a sufficiently large positive constant.*

Proof. See Appendix F. □

We are left with the double difference $\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,\text{sgn}} - \boldsymbol{x}^{t+1,(l)} + \boldsymbol{x}^{t+1,\text{sgn},(l)}$ (cf. (40d)), for which one has the following lemma.

Lemma 7: *Suppose* $m \ge Cn\log^8 m$ *for some sufficiently large constant* $C > 0$. *If the induction hypotheses (40) hold true up to the* t^{th} *iteration for some* $t \le T_0$, *then with probability at least* $1 - O(me^{-1.5n}) - O(m^{-10})$,

$$\begin{aligned}&\max_{1 \le l \le m} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,\text{sgn}} - \boldsymbol{x}^{t+1,(l)} + \boldsymbol{x}^{t+1,\text{sgn},(l)}\|_2 \\ &\le \alpha_{t+1}\left(1 + \frac{1}{\log m}\right)^{t+1} C_4 \frac{\sqrt{n\log^9 m}}{m}\end{aligned} \tag{46}$$

holds as long as $\eta > 0$ *is a sufficiently small constant and* $C_4 > 0$ *is sufficiently large.*

Proof. See Appendix G. □

Assuming the induction hypotheses (40) hold up to the t^{th} iteration for some $t \le T_0$, we know from Lemma 2 that the approximate state evolution for both α_t and β_t (see (20)) holds up to $t+1$. As a result, the last two hypotheses (40e) and (40f) for the $(t+1)^{\text{th}}$ iteration can be easily verified.

4.6 Justification of approximate state evolution for Phase II of Stage 1

Recall from Lemma 1 that Phase II refers to the iterations $T_0 < t \le T_\gamma$ (see the definition of T_0 in Lemma 1), for which one has

$$\alpha_t \ge \frac{c_6}{\log^5 m} \tag{47}$$

as long as the approximate state evolution (20) holds. Here $c_6 > 0$ is the same constant as in Lemma 1. Similar to Phase I, we invoke an inductive argument to prove that the approximate state evolution (20) continues to hold for $T_0 < t \le T_\gamma$.

4.6.1 Induction hypotheses—In Phase I, we rely on the leave-one-out sequences and the random-sign sequences $\{\boldsymbol{x}^{t,(l)}\}$, $\{\boldsymbol{x}^{t,\text{sgn}}\}$ and $\{\boldsymbol{x}^{t,\text{sgn},(l)}\}$ to establish certain "near-independence" between $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{a}_l\}$, which in turn allows us to obtain sharp control of the residual terms $\boldsymbol{r}(\boldsymbol{x}^t)$ (cf. (10)) and $r_1(\boldsymbol{x}^t)$ (cf. (11)). As it turns out, once the size α_t of the signal component obeys $\alpha_t \gtrsim 1/\text{poly}\log(m)$, then $\{\boldsymbol{x}^{t,(l)}\}$ alone is sufficient for our purpose to establish the "near-independence" property. More precisely, in Phase II we only need to impose the following induction hypotheses.

$$\max_{1 \le l \le m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \le \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_6 \frac{\sqrt{n \log^{15} m}}{m}; \tag{48a}$$

$$c_5 \le \left\|\boldsymbol{x}^t_\perp\right\|_2 \le \left\|\boldsymbol{x}^t\right\|_2 \le C_5. \tag{48b}$$

A direct consequence of (48) is the incoherence between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_l\}$, namely,

$$\max_{1 \le l \le m} \left|a_l^\top{}_\perp \boldsymbol{x}^t_\perp\right| \lesssim \sqrt{\log m} \left\|\boldsymbol{x}^t_\perp\right\|_2; \tag{49a}$$

$$\max_{1 \le l \le m} \left|a_l^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log m} \left\|\boldsymbol{x}^t\right\|_2. \tag{49b}$$

To see this, one can use the triangle inequality to show that

$$\begin{aligned}\left|a_{l,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right| &\le \left|a_{l,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,(l)}\right| + \left|a_{l,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right)\right| \\ &\overset{\text{(i)}}{\lesssim} \sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 + \sqrt{n}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_2 \\ &\lesssim \sqrt{\log m}\left(\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2 + \left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_2\right) + \sqrt{n}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_2 \\ &\overset{\text{(ii)}}{\lesssim} \sqrt{\log m} + \frac{\sqrt{n\log^{15} m}}{m}\sqrt{n} \lesssim \sqrt{\log m},\end{aligned}$$

where (i) follows from the independence between $\boldsymbol{a}_l$ and $\boldsymbol{x}^{t,(l)}$ and the Cauchy-Schwarz inequality, and the last line (ii) arises from $(1+1/\log m)^t \lesssim 1$ for $t \le T_\gamma \lesssim \log n$ and $m \gg n \log^{15/2} m$. This combined with the fact that $\|\boldsymbol{x}_{\perp}^{t}\|_2 \ge c_5/2$ results in

$$\max_{1\le l\le m}\left|a_{l,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2. \tag{50}$$

The condition (49b) follows using nearly identical arguments, which are omitted here.

As in Phase I, we need to justify the approximate state evolution (20) for both α_t and β_t, given that the t^{th} iterates satisfy the induction hypotheses (48). This is stated in the following lemma.

Lemma 8: *Suppose $m \ge Cn\log^{13} m$ for some sufficiently large constant $C > 0$. If the t^{th} iterates satisfy the induction hypotheses (48) for $T_0 < t < T_\gamma$, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1-\left(\alpha_t^2+\beta_t^2\right)\right] + \eta\zeta_t\right\}\alpha_t; \tag{51a}$$

$$\beta_{t+1} = \left\{1 + \eta\left[1-3\left(\alpha_t^2+\beta_t^2\right)\right] + \eta\rho_t\right\}\beta_t, \tag{51b}$$

for some $|\zeta_t| \ll 1/\log m$ and $|\rho_t| \ll 1/\log m$

Proof. See Appendix H for the proof of (51a). The proof of (51b) follows exactly the same argument as in proving (41b), and is hence omitted. □

4.6.2 Induction step—We proceed to complete the induction argument. Towards this end, one has the following lemma in regard to the induction on $\max_{1\le l\le m}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\|_2$ (see (48a)).

Lemma 9: *Suppose $m \ge Cn\log^5 m$ for some sufficiently large constant $C > 0$, and consider any $T_0 < t < T_\gamma$. If the induction hypotheses (40) are valid throughout Phase I and (48) are valid from the T_0th to the t^{th} iterations, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\max_{1\le l\le m}\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right\|_2 \le \alpha_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1} C_6\frac{\sqrt{n\log^{13} m}}{m}$$

holds as long as $\eta > 0$ *is sufficiently small and* $C_6 > 0$ *is sufficiently large.*

Proof. See Appendix I. □

As in Phase I, since we assume the induction hypotheses (40) (resp. (48)) hold for all iterations up to the T_0th iteration (resp. between the T_0th and the t^{th} iteration), we know from Lemma 8 that the approximate state evolution for both α_t and β_t (see (20)) holds up to $t + 1$. The last induction hypothesis (48b) for the $(t+1)^{\text{th}}$ iteration can be easily verified from Lemma 1.

It remains to check the case when $t = T_0 + 1$. It can be seen from the analysis in Phase I that

$$\max_{1 \le l \le m} \left\| \boldsymbol{x}^{T_0+1} - \boldsymbol{x}^{T_0+1,(l)} \right\|_2 \le \beta_{T_0+1} \left(1 + \frac{1}{\log m}\right)^{T_0+1} C_1 \frac{\sqrt{n \log^5 m}}{m} \le \alpha_{T_0+1} \left(1 + \frac{1}{\log m}\right)^{T_0+1} C_6 \frac{\sqrt{n \log^{15} m}}{m},$$

for some constant condition $C_6 \gg 1$, where the second line holds since $\beta_{T_0+1} \le C_5$, $\alpha_{T_0+1} \ge c_6/\log^5 m$.

4.7 Analysis for Stage 2

Combining the analyses in *Phase I* and *Phase II*, we finish the proof of Theorem 2 for Stage 1, i.e. $t \le T_\gamma$. In addition to $\left(\boldsymbol{x}^{T_\gamma}, \boldsymbol{x}^\natural\right) \le \gamma$, we can also see from (49b) that

$$\max_{1 \le i \le m} \left| a_i^\top \boldsymbol{x}^{T_\gamma} \right| \lesssim \sqrt{\log m},$$

which in turn implies that

$$\max_{1 \le i \le m} \left| a_i^\top \left(\boldsymbol{x}^{T_\gamma} - \boldsymbol{x}^\natural \right) \right| \lesssim \sqrt{\log m}.$$

Armed with these properties, one can apply the arguments in [MWCC17, Section 6] to prove that for $t \ge T_\gamma + 1$,

$$\operatorname{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \le \left(1 - \frac{\eta}{2}\right)^{t - T_\gamma} \operatorname{dist}\left(\boldsymbol{x}^{T_\gamma}, \boldsymbol{x}^\natural\right) \le \left(1 - \frac{\eta}{2}\right)^{t - T_\gamma} \cdot \gamma. \tag{52}$$

Notably, the theorem therein [MWCC17, Theorem 1] works under the stepsize $\eta_t \equiv \eta \asymp c/\log n$ when $m \gg n \log n$. Nevertheless, as remarked by the authors, when the sample complexity exceeds $m \gg n \log^3 m$, a constant stepsize is allowed.

We are left with proving (15) for Stage 2. Note that we have already shown that the ratio α_t / β_t increases exponentially fast in Stage 1. Therefore,

$$\frac{\alpha_{T_1}}{\beta_{T_1}} \geq \frac{1}{\sqrt{2n\log n}}\left(1+c_{10}\eta^2\right)^{T_1}$$

and, by the definition of T_1 (see (26)) and Lemma 1, one has $\alpha_{T_1} \asymp \beta_{T_1} \asymp 1$ and hence

$$\frac{\alpha_{T_1}}{\beta_{T_1}} \asymp 1. \tag{53}$$

When it comes to $t > T_\gamma$, in view of (52), one has

$$\begin{aligned}
\frac{\alpha_t}{\beta_t} &\geq \frac{1-\mathrm{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right)}{\mathrm{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right)} \geq \frac{1-\gamma}{\left(1-\frac{\eta}{2}\right)^{t-T_\gamma}\cdot\gamma} \\
&\geq \frac{1-\gamma}{\gamma}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma} \overset{\text{(i)}}{\asymp} \frac{\alpha_{T_1}}{\beta_{T_1}}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma} \\
&\gtrsim \frac{1}{\sqrt{n\log n}}\left(1+c_{10}\eta^2\right)^{T_1}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma} \\
&\overset{\text{(ii)}}{\asymp} \frac{1}{\sqrt{n\log n}}\left(1+c_{10}\eta^2\right)^{T_\gamma}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma} \\
&\gtrsim \frac{1}{\sqrt{n\log n}}\left(1+c_{10}\eta^2\right)^{t},
\end{aligned}$$

where (i) arises from (53) and the fact that γ is a constant, (ii) follows since $T_\gamma - T_1 \asymp 1$ according to Lemma 1, and the last line holds as long as $c_{10} > 0$ and η are sufficiently small. This concludes the proof regarding the lower bound on α_t/β_t.

5 Discussions

The current paper justifies the fast global convergence of gradient descent with random initialization for phase retrieval. Specifically, we demonstrate that GD with random initialization takes only $O(\log n + \log(1/\epsilon))$ iterations to achieve a relative ϵ-accuracy in terms of the estimation error. It is likely that such fast global convergence properties also arise in other nonconvex statistical estimation problems. The technical tools developed herein may also prove useful for other settings. We conclude our paper with a few directions worthy of future investigation.

- *Sample complexity and phase transition.* We have proved in Theorem 2 that GD with random initialization enjoys fast convergence, with the proviso that $m \gg n \log^{13} m$. It is possible to improve the sample complexity via more sophisticated arguments. In addition, it would be interesting to examine the phase transition phenomenon of GD with random initialization.
- *Other nonconvex statistical estimation problems.* We use the phase retrieval problem to showcase the efficiency of GD with random initialization. It is certainly interesting to investigate whether this fast global convergence carries over to other nonconvex statistical estimation problems including *low-rank matrix and tensor recovery* [KMO10,SL16,CW15,TBS

+16,ZL16,ZWL15,MWCC17,CL17,CC18,CCF18,HZC18], *blind deconvolution* [LLSW18,MWCC17,HH17] and *neural networks* [SJL17,LMZ17,FCL18]. The leave-one-out sequences and the "near-independence" property introduced/identified in this paper might be useful in proving efficiency of randomly initialized GD for the aforementioned problems.

- *Noisy setting and other activation functions.* Throughout this paper, our focus is on inverting noiseless quadratic systems. Extensions to the noisy case is definitely worth investigating. Moving beyond quadratic samples, one may also study other activation functions, including but not limited to Rectified Linear Units (ReLU), polynomial functions and sigmoid functions. Such investigations might shed light on the effectiveness of GD with random initialization for training neural networks.
- *Other iterative optimization methods.* Apart from gradient descent, other iterative procedures have been applied to solve the phase retrieval problem. Partial examples include *alternating minimization*, *Kaczmarz algorithm,* and *truncated gradient descent (Truncated Wirtinger flow)*. In conjunction with random initialization, whether the iterative algorithms mentioned above enjoy fast global convergence is an interesting open problem. For example, it has been shown that truncated WF together with truncated spectral initialization achieves optimal sample complexity (i.e. $m \asymp n$) and computational complexity simultaneously [CC17]. Does truncated Wirtinger flow still enjoy optimal sample complexity when initialized randomly?
- *Beyond Gaussian sampling vectors.* In this work, we consider the Gaussian phase retrieval problem where the sampling vectors are i.i.d. Gaussian vectors. We expect our results to generalize to other sampling vectors. Experimentally, we can verify that random initialization also converges fast under a Rademacher sampling model; see Figure 8.
- *Applications of leave-one-out tricks.* In this paper, we heavily deploy the *leave-one-out* trick to demonstrate "near-independence" between the iterates $\boldsymbol{x}^t$ and the sampling vectors $\{\boldsymbol{a}_i\}$. The basic idea is to construct an auxiliary sequence that is (i) independent w.r.t. certain components of the design vectors, and (ii) extremely close to the original sequence. These two properties allow us to propagate the desired independence properties to $\boldsymbol{x}^t$. As mentioned in Section 3, the leave-one-out trick has served as a very powerful hammer for decoupling the dependency between random vectors in several high-dimensional estimation problems. We expect this powerful trick to be useful in broader settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ARO grant W911NF-18-1-0303, and by the Princeton SEAS innovation award. Y. Chi is supported in part by AFOSR under the grant

FA9550-15-1-0205, by ONR under the grant N00014-18-1-2142, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571 and CCF-1806154. J. Fan is supported in part by NSF grants DMS-1662139 and DMS-1712591 and NIH grant 2R01-GM072611-13.

References

[AAZB+16]. Agarwal N, Allen-Zhu Z, Bullins B, Hazan E, and Ma T Finding approximate local minima for nonconvex optimization in linear time. arXiv preprint arXiv:1611.01146, 2016.

[AFWZ17]. 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.

[AZ17]. Allen-Zhu Z Natasha 2: Faster non-convex optimization than sgd. arXiv preprint arXiv:1708.08694, 2017.

[BCMN14]. Bandeira AS, Cahill J, Mixon DG, and Nelson AA Saving phase: Injectivity and stability for phase retrieval. Applied and Computational Harmonic Analysis, 37(1):106–125, 2014.

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

[CC17]. Chen Y and Candès EJ Solving random quadratic systems of equations is nearly as easy as solving linear systems. Comm. Pure Appl. Math, 70(5):822–883, 2017.

[CC18]. 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.

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

[CCG15]. Chen Y, Chi Y, and Goldsmith AJ Exact and stable covariance estimation from quadratic sampling via convex programming. IEEE Transactions on Information Theory, 61(7):4034–4059, 2015.

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

[CFL15]. Chen P, Fannjiang A, and Liu G-R 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]. Chen Y, Fan J, Ma C, and Wang K Spectral method and regularized MLE are both optimal for top-K ranking. arXiv preprint arXiv:1707.09971, 2017.

[CL14]. Candès EJ and Li X Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. Foundations of Computational Mathematics, 14(5):1017–1026, 2014.

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

[CL17]. Chen J and Li X Memory-efficient kernel PCA via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. arXiv preprint arXiv:1711.01742, 2017.

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

[CLM16]. Cai TT, Li X, and Ma Z Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. The Annals of Statistics, 44(5):2221–2251, 2016.

[CLS15]. Candès EJ, Li X, and Soltanolkotabi M Phase retrieval via Wirtinger flow: Theory and algorithms. IEEE Transactions on Information Theory, 61(4):1985–2007, 4 2015.

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

[CSV13]. Candès EJ, 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.

[CW15]. Chen Y and Wainwright MJ Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv preprint arXiv:1509.03025, 2015.

[CWZG17]. Chen J, Wang L, Zhang X, and Gu Q Robust wirtinger flow for phase retrieval with arbitrary corruption. arXiv preprint arXiv:1704.06256, 2017.

[CYC14]. Chen Y, Yi X, and Caramanis C A convex formulation for mixed regression with two components: Minimax optimal rates. In Conference on Learning Theory, pages 560–604, 2014.

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

[DH14]. Demanet L and Hand P Stable optimizationless recovery from phaseless linear measurements. Journal of Fourier Analysis and Applications, 20(1):199–221, 2014.

[DJL+17]. Du SS, Jin C, Lee JD, Jordan MI, Singh A, and Poczos B Gradient descent can take exponential time to escape saddle points. In Advances in Neural Information Processing Systems, pages 1067–1077, 2017.

[DR17]. Duchi JC and Ruan F Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. arXiv preprint arXiv:1705.02356, 2017.

[EK15]. 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, pages 1–81, 2015.

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

[FCL18]. Fu H, Chi Y, and Liang Y Local geometry of one-hidden-layer neural networks for logistic regression. arXiv preprint arXiv:1802.06463, 2018.

[GHJY15]. Ge R, Huang F, Jin C, and Yuan Y Escaping from saddle points online stochastic gradient for tensor decomposition. In Conference on Learning Theory, pages 797–842, 2015.

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

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

[HZC18]. Hao B, Zhang A, and Cheng G Sparse and low-rank tensor estimation via cubic sketchings. arXiv preprint arXiv:1801.09326, 2018.

[JGN+17]. Jin C, Ge R, Netrapalli P, Kakade SM, and Jordan MI How to escape saddle points efficiently. arXiv preprint arXiv:1703.00887, 2017.

[JNJ17]. Jin C, Netrapalli P, and Jordan MI Accelerated gradient descent escapes saddle points faster than gradient descent. arXiv preprint arXiv:1711.10456, 2017.

[KMO10]. Keshavan RH, Montanari A, and Oh S Matrix completion from a few entries. IEEE Transactions on Information Theory, 56(6):2980–2998, 6 2010.

[KRT17]. Kueng R, Rauhut H, and Terstiege U Low rank matrix recovery from rank one measurements. Applied and Computational Harmonic Analysis, 42(1):88–116, 2017.

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

[LGL15]. Li G, Gu Y, and Lu YM Phase retrieval using iterative projections: Dynamics in the large systems limit. In Allerton Conference on Communication, Control, and Computing, pages 1114–1118. IEEE, 2015.

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

[LLSW18]. Li X, Ling S, Strohmer T, and Wei K Rapid, robust, and reliable blind deconvolution via nonconvex optimization. Applied and Computational Harmonic Analysis, 2018.

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

[LMZ17]. Li Y, Ma T, and Zhang H Algorithmic regularization in over-parameterized matrix recovery. arXiv preprint arXiv:1712.09203, 2017.

[LSJR16]. Lee JD, Simchowitz M, Jordan MI, and Recht B Gradient descent converges to minimizers. arXiv preprint arXiv:1602.04915, 2016.

[MM17]. Mondelli M and Montanari A Fundamental limits of weak recovery with applications to phase retrieval. arXiv preprint arXiv:1708.05932, 2017.

[MSK17]. Murray R, Swenson B, and Kar S Revisiting normalized gradient descent: Evasion of saddle points. arXiv preprint arXiv:1711.05224, 2017.

[MWCC17]. 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.

[MXM18]. Ma J, Xu J, and Maleki A Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization. arXiv preprint arXiv:1801.01170, 2018.

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

[NP06]. Nesterov Y and Polyak BT Cubic regularization of Newton method and its global performance. Mathematical Programming, 108(1):177–205, 2006.

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

[SCC18]. Sur P, Chen Y, and Candès EJ The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. accepted to Probability Theory and Related Fields, 2018.

[SEC+15]. Shechtman Y, Eldar YC, Cohen O, Chapman HN, Miao J, and Segev M Phase retrieval with application to optical imaging: a contemporary overview. IEEE signal processing magazine, 32(3):87–109, 2015.

[SJL17]. Soltanolkotabi M, Javanmard A, and Lee JD Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. arXiv preprint arXiv:1707.04926, 2017.

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

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

[Sol17]. Soltanolkotabi M Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. arXiv preprint arXiv:1702.06175, 2017.

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

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

[TBS+16]. 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, pages 964–973. JMLR. org, 2016.

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

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

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

[WGE17]. Wang G, Giannakis GB, and Eldar YC Solving systems of random quadratic equations via truncated amplitude flow. IEEE Transactions on Information Theory, 2017.

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

[YYF+17]. Yang Z, Yang LF, Fang EX, Zhao T, Wang Z, and Neykov M Misspecified nonconvex statistical optimization for phase retrieval. arXiv preprint arXiv:1712.06245, 2017.

[ZB17]. Zhong Y and Boumal N Near-optimal bounds for phase synchronization. arXiv preprint arXiv:1703.06605, 2017.

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

[Zha17]. Zhang T Phase retrieval using alternating minimization in a batch setting. arXiv preprint arXiv:1706.08167, 2017.

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

[ZWGC17]. Zhang L, Wang G, Giannakis GB, and Chen J Compressive phase retrieval via reweighted amplitude flow. arXiv preprint arXiv:1712.02426, 2017.

[ZWL15]. Zhao T, Wang Z, and Liu H A nonconvex optimization framework for low rank matrix estimation. In Advances in Neural Information Processing Systems, pages 559–567, 2015. [PubMed: 28316458]

[ZZLC17]. Zhang H, Zhou Y, Liang Y, and Chi Y A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. Journal of Machine Learning Research, 2017.

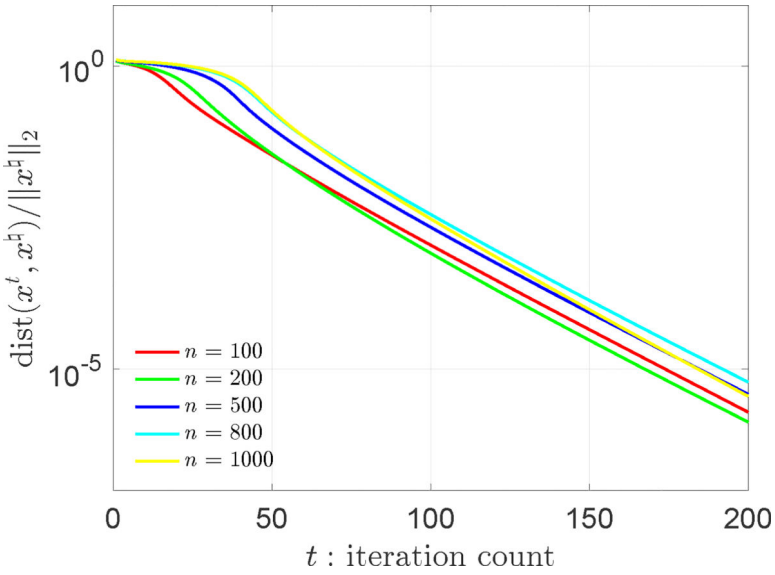


Figure 1:

The relative ℓ_2 error vs. iteration count for GD with random initialization, plotted semilogarithmically. The results are shown for $n = 100, 200, 500, 800, 1000$ with $m = 10n$ and $\eta_t \equiv 0.1$.

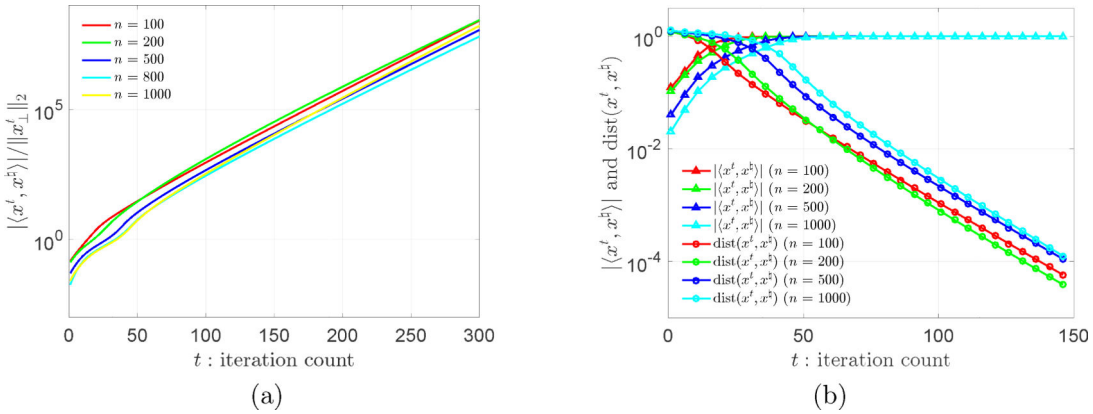


Figure 2:

(a) The ratio $\left|\left\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \right\rangle\right| / \|\boldsymbol{x}_\perp^t\|_2$, and (b) the size $\left|\left\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \right\rangle\right|$ of the signal component and the ℓ_2 error vs. iteration count, both plotted on semilogarithmic scales. The results are shown for $n =$ 100, 200, 500, 800, 1000 with $m = 10n$, $\eta_t \equiv 0.1$, and $\|\boldsymbol{x}^\natural\|_2 = 1$.

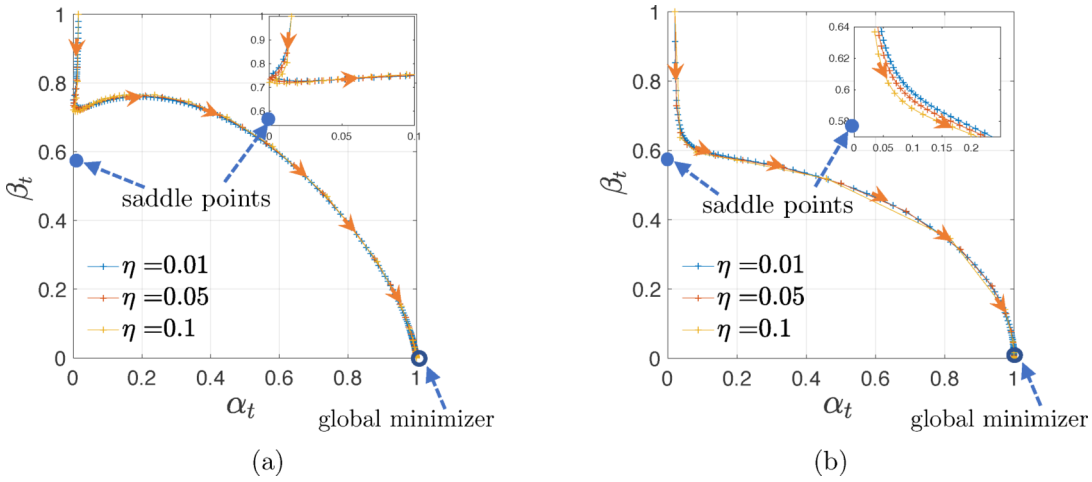
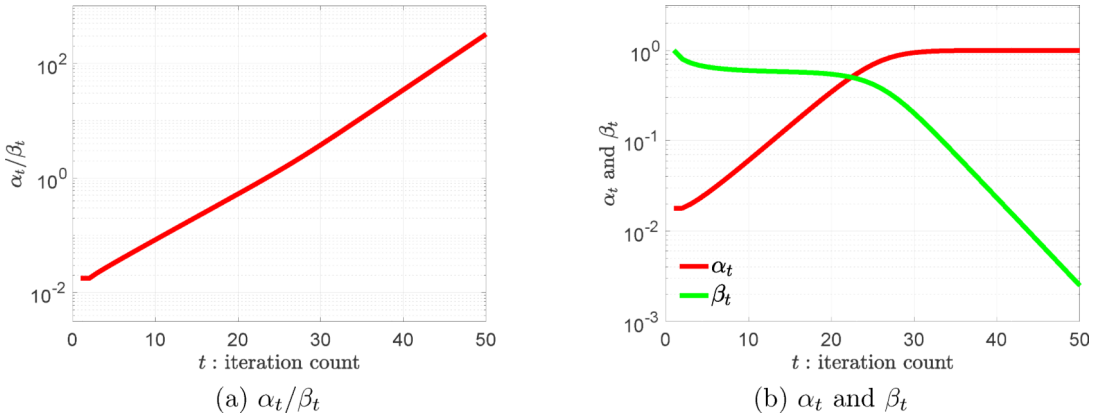


Figure 3:

The trajectory of (α_t, β_t), where $\alpha_t = |\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle|$ and $\beta_t = \|\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural\|_2$ represent respectively the size of the signal component and that of the orthogonal component of the GD iterates (assume $\|\boldsymbol{x}^\natural\|_2 = 1$). (a) The results are shown for $n = 1000$ with $m = 10n$, and $\eta_t = 0.01, 0.05, 0.1$. (b) The results are shown for $n = 1000$ with m approaching infinity, and $\eta_t = 0.01, 0.05, 0.1$. The blue filled circles represent the population-level saddle points, and the orange arrows indicate the directions of increasing t.



(a) α_t/β_t (b) α_t and β_t

Figure 4:

Population-level state evolution, plotted semilogarithmically: (a) the ratio α_t/β_t vs. iteration count, and (b) α_t and β_t vs. iteration count. The results are shown for $n = 1000$, $\eta_t \equiv 0.1$, and $\boldsymbol{x}^0 \sim \mathcal{N}\left(\mathbf{0}, n^{-1}\boldsymbol{I}_n\right)$ (assuming $\alpha_0 > 0$ though).

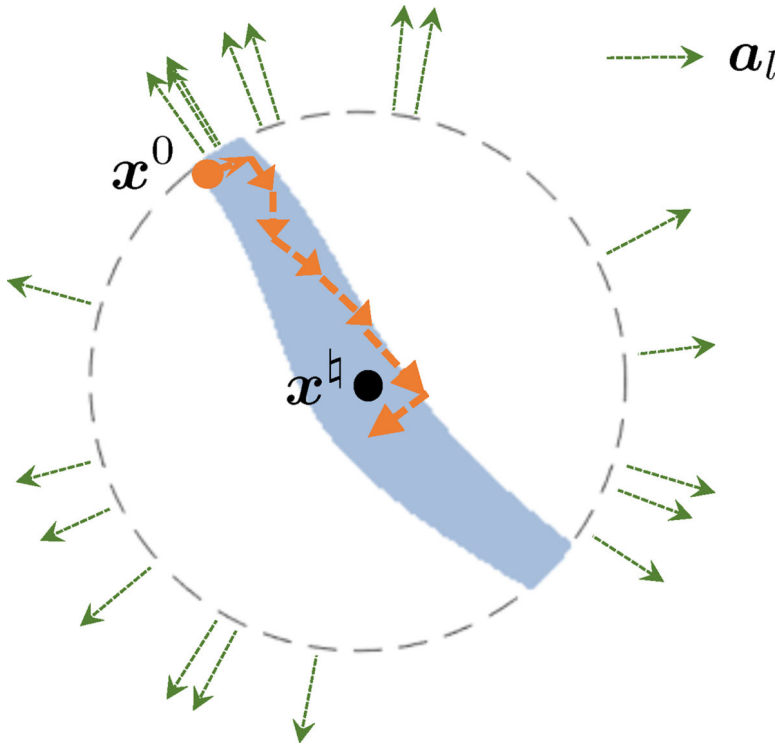


Figure 5:

Illustration of the region satisfying the "near-independence" property. Here, the green arrows represent the directions of $\{\boldsymbol{a}_i\}_{1\le i\le 20}$, and the blue region consists of all points such that the first entry $r_1(\boldsymbol{x})$ of the fluctuation $\boldsymbol{r}(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \nabla F(\boldsymbol{x})$ is bounded above in magnitude by $\left\|\boldsymbol{x}_{\|}^{\natural}\right\|/5 \left(\text{or} \left|\left\langle \boldsymbol{x}, \boldsymbol{x}^{\natural}\right\rangle\right|/5\right)$.

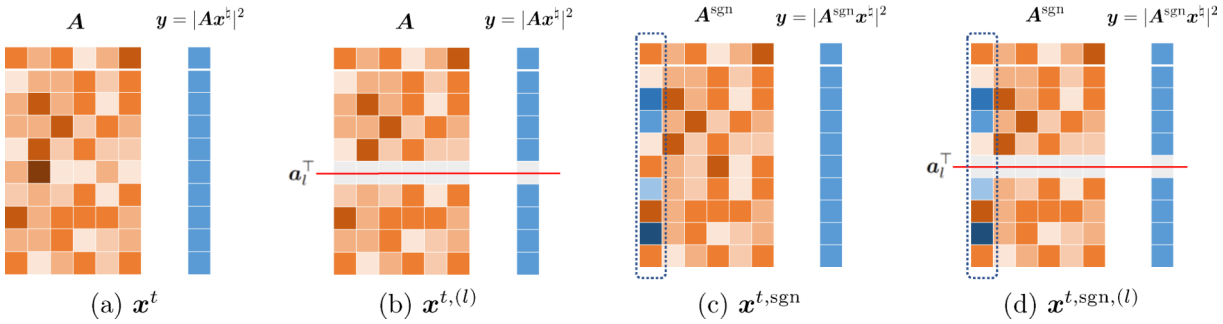


Figure 6:

Illustration of the leave-one-out and random-sign sequences. (a) $\{\boldsymbol{x}^t\}$ is constructed using all data $\{\boldsymbol{a}_i, y_i\}$; (b) $\{\boldsymbol{x}^{t,(l)}\}$ is constructed by discarding the lth sample $\{\boldsymbol{a}_l, y_l\}$; (c) $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ is constructed by using auxiliary design vectors $\{a_i^{\mathrm{sgn}}\}$, where a_i^{sgn} is obtained by randomly flipping the sign of the first entry of $\boldsymbol{a}_i$; (d) $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}$ is constructed by discarding the lth sample $\{a_l^{\mathrm{sgn}}, y_l\}$.

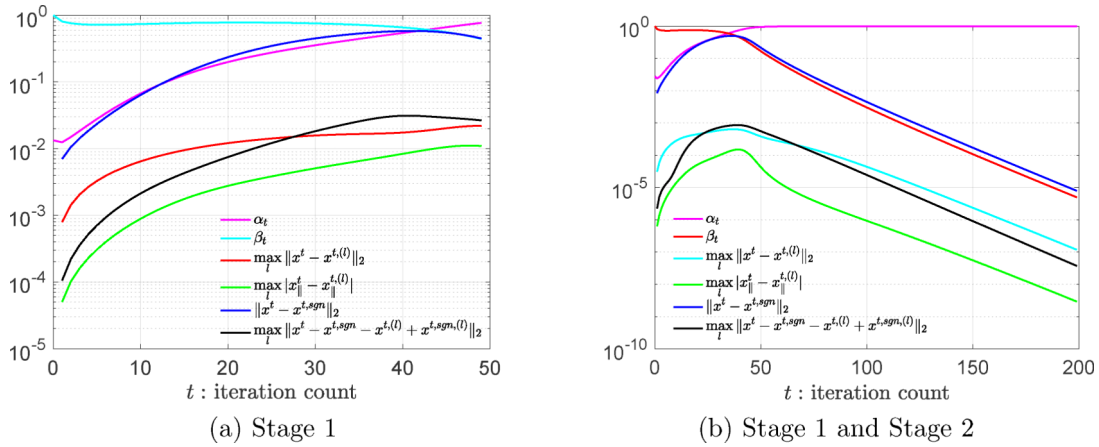


Figure 7:

Illustration of the differences among leave-one-out and original sequences vs. iteration count, plotted semilogarithmically. The results are shown for $n = 1000$ with $m = 10n$, $\eta_t \equiv 0.1$, and $\|\boldsymbol{x}^\natural\|_2 = 1$. (a) The four differences increases in Stage 1. From the induction hypotheses (40), our upper bounds on $\left|\boldsymbol{x}^t_{\|} - \boldsymbol{x}^{t,(l)}_{\|}\right|$, $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\|_2$ and $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\|_2$ scale linearly with α_t, whereas the upper bound on $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\|_2$ is proportional to β_t. In addition, $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,(l)}\|_2 \lesssim 1/\sqrt{m}$, $\left|\boldsymbol{x}^1_{\|} - \boldsymbol{x}^{1,(l)}_{\|}\right| \lesssim 1/m$, $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}}\|_2 \lesssim 1/\sqrt{m}$ and $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}} - \boldsymbol{x}^{1,(l)} + \boldsymbol{x}^{1,\mathrm{sgn},(l)}\|_2 \lesssim 1/m$. (b) The four differences converge to zero geometrically fast in Stage 2, as all the (variants of) leave-one-out sequences and the original sequence converge to the truth $\boldsymbol{x}^\natural$.

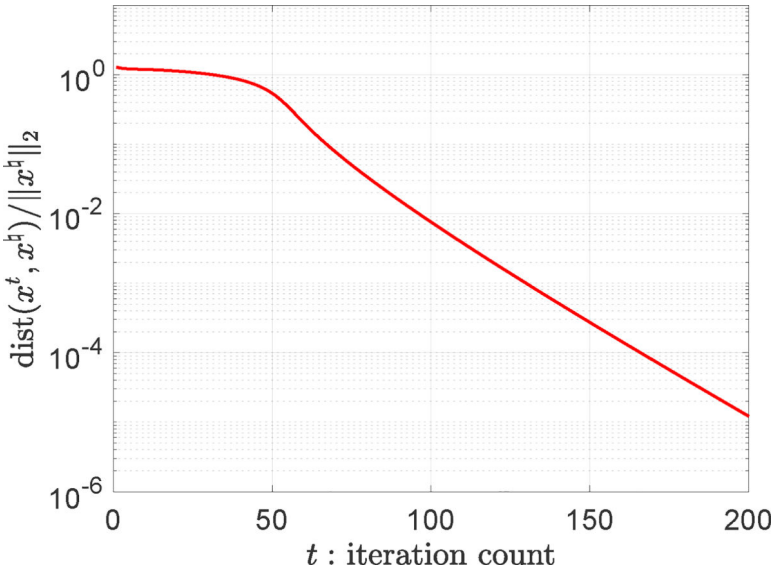


Figure 8:

The relative ℓ_2 error vs. iteration count for GD with random initialization, plotted semilogarithvectorsmically. The results are shown for $n = 1000$ with $m = 10n$ and $\eta t \equiv 0.1$. Here the entries of the sampling $\boldsymbol{a}_i$ are drawn *i.i.d.* from a Rademacher distribution.