Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval

Zhe Gao

School of Management
University of Science and Technology of China

18, April, 2025

Outline

1. Introduction
2. Main Result
3. Conclusion

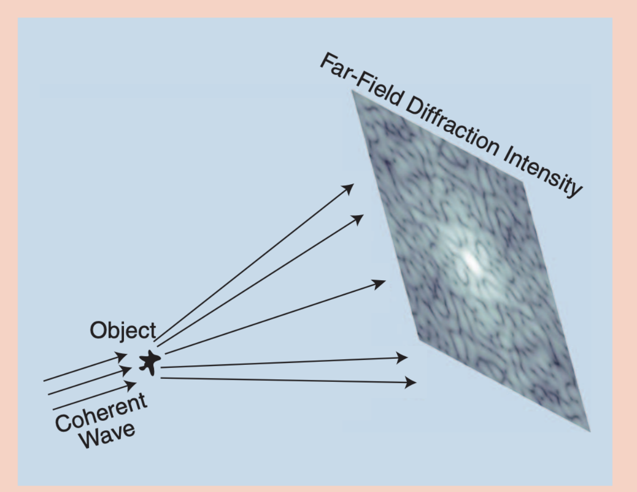
Phase retrieval

- The problem of phase retrieval, i.e., the recovery of a function given the magnitude of its Fourier transform, arises in various fields of science and engineering, including electron microscopy, crystallography, astronomy, and optical imaging.

Coherent diffractive imaging

- In the basic CDI setup (forward scattering), an object is illuminated by a quasi-monochromatic coherent wave and the diffracted intensity is measured.
- When the object is small and the intensity is measured far away, the measured intensity is proportional to the magnitude of the Fourier transform of the wave at the object plane with appropriate spatial scaling.

Coherent diffractive imaging



Coherent diffractive imaging

- Consider the discretized one-dimensional real-space distribution function of an object: $x \in \mathbb{C}^N$, which corresponds to the transmittance function of the object.
- The fact that x is generally complex corresponds physically to the fact that the electromagnetic field emanating from different points on the object has not only magnitude but also phase (as is always the case, for example, when 3-D objects are illuminated and light is reflected from points at different planes).

Coherent diffractive imaging

- The 1-D discrete Fourier transform (DFT) of x is given by

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi\frac{kn}{N}}, \quad k = 0, 1, \ldots, N-1.$$

- The term oversampled DFT used in this article will refer to an M point DFT of $x \in \mathbb{C}^N$ with $M > N$

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi\frac{kn}{M}}, \quad k = 0, 1, \ldots, M-1.$$

Coherent diffractive imaging

- The recovery of x from measurement of $\mathbf{X}$ can be achieved by simply applying the inverse-DFT operator. Writing $X[k] = |X[k]| \cdot e^{j\phi[k]}$, the Fourier phase-retrieval problem is to recover x when only the magnitude of X is measured, i.e., to recover $x[n]$ given $|X[k]|$.
- Since the DFT operator is bijective, this is equivalent to recovering the phase of $X[k]$, i.e., $\phi[k]$-hence the term phase retrieval. Denote by $\hat{\mathbf{x}}$ the vector x after padding with $N-1$ zeros. The autocorrelation sequence of $\hat{\mathbf{x}}$ is then defined as

$$g[m] = \sum_{i=\max\{1,m+1\}}^{N} \hat{x}_i \overline{\hat{x}_{i-m}}, \quad m = -(N-1), \ldots, N-1.$$

It is well known that the DFT of $g[m]$, denoted by $G[k]$, satisfies $G[k] = |X[k]|^2$. Thus, the problem of recovering a signal from its Fourier magnitude is equivalent to recovering a signal from its autocorrelation sequence.

Phase retrieval

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

$$y_i = \left(\boldsymbol{a}_i^{\top} \boldsymbol{x}^{\natural}\right)^2, \quad 1 \leq i \leq m,$$

where y_i is the sample we collect and $\boldsymbol{a}_i$ is the design vector known a priori. Here $\boldsymbol{a}_i = [e^{-j2\pi \frac{i}{m}}, \ldots, e^{-j2\pi \frac{in}{m}}]$.

Is it feasible to reconstruct $x^{\natural}$ in an accurate and efficient manner?

Optimization Problem

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

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

- Fortunately, in spite of nonconvexity, a variety of optimization-based methods are shown to be effective in the presence of proper statistical models.
- Gradient descent (GD):

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right), \quad t = 0, 1, \ldots$$

with η_t being the stepsize/learning rate.

Optimization Problem

- The above iterative procedure is also dubbed Wirtinger flow for phase retrieval, which can accommodate the complex-valued case as well.
- 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, provided that the ratio m/n of the number of equations to the number of unknowns exceeds some logarithmic factor.

Initialization

One crucial element in prior convergence analysis is initialization. In order to guarantee linear convergence, prior works typically recommend spectral initialization or its variants. 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:

- $\boldsymbol{x}^0$ 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 $\left|\boldsymbol{a}_i^\top \boldsymbol{x}^0\right|$ is reasonably small for all $1 \leq i \leq m$-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.

Initialization

- A strategy that practitioners often like to employ is to initialize GD randomly.
- Random initialization is model- agnostic and is usually more robust vis-a-vis model mismatch.
- GD with random initialization is poorly understood in theory.

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.

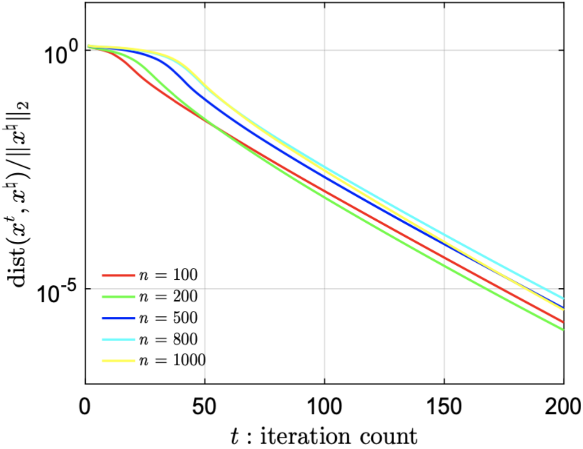
Reviews

- Convex relaxation (computationally prohibitive for solving large-scale problems);
- Wirtinger flow algorithm with spectral initialization

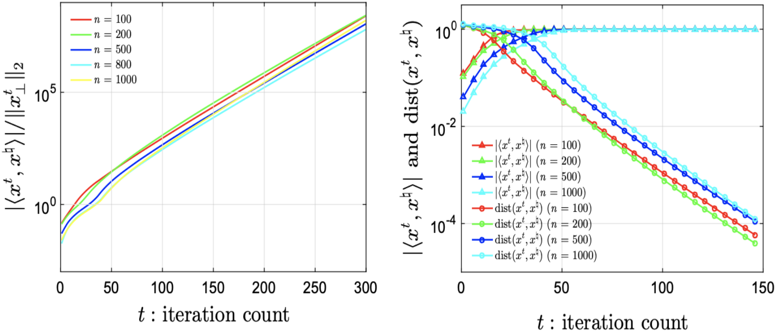
Require carefully-designed initialization to guarantee a sufficiently accurate initial point. (But if initial signal strength is vanishingly small)

- Global convergence of alternating minimization / projection with random initialization,
- Saddle-point escaping algorithms (high computational complexity).

Toy Example


10^0
10^{-5}
$\mathrm{dist}(x^t, x^\natural)/\|x^\natural\|_2$
$n = 100$
$n = 200$
$n = 500$
$n = 800$
$n = 1000$
0
50
100
150
200
t : iteration count

Toy Example



(a)

(b)

Toy Example

- We observe two stages for GD: (1) Stage 1: the relative error of x^t stays nearly flat; (2) Stage 2: the relative error of x^t experiences geometric decay.
- The strength ratio of the signal to the orthogonal components grows exponentially.
- Exponential growth of the signal strength in Stage 1.

Outline

1 Introduction

2 Main Result

3 Conclusion

Main Result

Theorem 2.1

Fix $\boldsymbol{x}^{\natural} \in \mathbb{R}^n$ *with* $\left\|\boldsymbol{x}^{\natural}\right\|_2 = 1$. *Suppose that* $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ *for* $1 \leq i \leq m, \boldsymbol{x}^0 \sim \mathcal{N}\left(\mathbf{0}, n^{-1}\boldsymbol{I}_n\right)$, *and* $\eta_t \equiv \eta = c / \left\|\boldsymbol{x}^{\natural}\right\|_2^2$ *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*

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

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

- The stepsize is taken to be a fixed constant throughout all iterations.
- We reuse the same data across all iterations (i.e. no sample splitting is needed to establish this theorem).

Remark

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; (2) Stage 2 consists of all remaining iterations, where the estimation error contracts linearly.

- Stage 1 takes $O(\log n)$ iterations. When seeded with a random initial guess, GD is capable of entering a local region surrounding $x^\natural$ 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 $\gamma > 0$.

Remark

- 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

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

- Near-minimal sample complexity. The preceding computational guarantees occur as soon as the sample size exceeds $m \gtrsim n \operatorname{poly}\log(m)$. 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.

Remark

- There is no need to adopt sophisticated saddle-point escaping schemes developed in generic optimization theory (e.g. cubic regularization, perturbed GD).
- 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 (no sample splitting).

Intuitions

- Investigate the dynamics of the population gradient sequence (the case where we have infinite samples);
- The finite-sample case and independence between the iterates and the design vectors;
- 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 $\{\boldsymbol{a}_i\}$.

Assumption

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

$$x_{\|}^{t} := x_1^{t} \quad \text{and} \quad \boldsymbol{x}_{\perp}^{t} := [x_i^{t}]_{2 \le i \le n} \tag{5}$$

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

$$\underbrace{x_{\|}^{t}\boldsymbol{e}_1 = \langle \boldsymbol{x}^t, \boldsymbol{x}^{\natural} \rangle \boldsymbol{x}^{\natural}}_{\text{signal component}} \quad \text{and} \quad \underbrace{\begin{bmatrix} 0 \\ \boldsymbol{x}_{\perp}^{t} \end{bmatrix} = \boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^{\natural} \rangle \boldsymbol{x}^{\natural}}_{\text{orthogonal component}} \tag{6}$$

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]

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

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

Population dynamics

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

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

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

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

which can be computed by $\nabla F(\boldsymbol{x}) = \mathbb{E}[\nabla f(\boldsymbol{x})] = \mathbb{E}\left[\left\{\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{\natural}\right)^2\right\} \boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{x}\right]$ assuming that $\boldsymbol{x}$ and the $\boldsymbol{a}_i$'s are independent.

Population dynamics

Note that

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

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

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

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

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

Population dynamics

This recursive system has three fixed points:

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

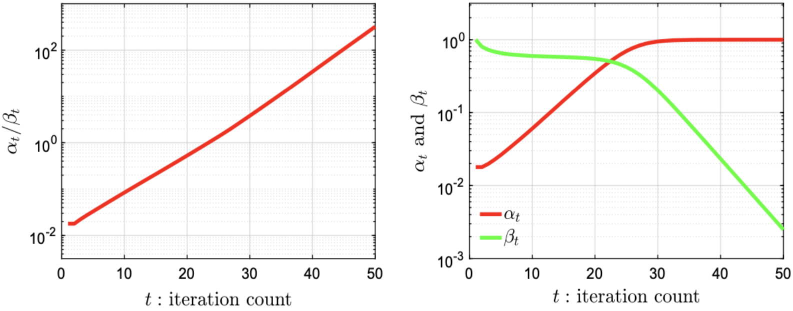
which correspond to the global minimizer, the local maximizer, and the saddle points, respectively, of the population objective function.

We make note of the following key observations in the presence of a randomly initialized $\boldsymbol{x}^0$,

- 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, $\left(\alpha^t,\beta^t\right)$ 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.

Population dynamics



(a) α_t/β_t

(b) α_t and β_t

Finite-sample analysis

We examine how many samples are needed in order for the population dynamics to be reasonably accurate. Rewrite the gradient update rule as

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

where $\nabla f(\boldsymbol{x}) = m^{-1} \sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2\right] \boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{x}$.

Assuming 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 $\boldsymbol{r}\left(\boldsymbol{x}^t\right)$.

Finite-sample analysis

Take the signal component as an example:

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

where

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

with $a_{i,1}$ the first entry of $\boldsymbol{a}_{i}$.

Finite-sample analysis

- 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 $\left\|x^t\right\|_2 = O(1)$, one can apply the CLT to demonstrate that

$$\left|r_1\left(\boldsymbol{x}^t\right)\right| \lesssim \sqrt{\operatorname{Var}\left(r_1\left(\boldsymbol{x}^t\right)\right)\ \text{poly}\ \log(m)} \lesssim \sqrt{\frac{\text{poly}\ \log(m)}{m}}$$

  with high probability.
- For instance, for the random initial guess $\boldsymbol{x}^0 \sim \mathcal{N}\left(\mathbf{0}, n^{-1}\boldsymbol{I}_n\right)$ one has $\left|x_{\|}^0\right| \gtrsim 1/\sqrt{n\log n}$ with probability approaching one, telling us that

$$\left|r_1\left(x^0\right)\right| \lesssim \sqrt{\frac{\text{poly}\log(m)}{m}} \ll \left|x_{\||}^0\right|$$

  as long as $m \gtrsim n$ poly $\log(m)$.

Finite-sample analysis

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:

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

with the proviso that $m \gtrsim n$ poly $\log(m)$.

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.
- Without the independence assumption, the CLT types of results fail to hold due to the complicated dependency between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_i\}$.

Near-independence and leave-one-out tricks

To exploit a certain “near-independence” property between $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{a}_i\}$, we make use of a leave-one-out trick.

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

As it turns out, we need to construct several auxiliary sequences $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$, $\{\boldsymbol{x}^{t,\text{sgn}}\}_{t\geq 0}$ and $\{\boldsymbol{x}^{t,\text{sgn},(l)}\}_{t\geq 0}$, where $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$ is independent of the l th sampling vector $\boldsymbol{a}_l$, $\{\boldsymbol{x}^{t,\text{sgn}}\}_{t\geq 0}$ is independent of the sign information of the first entries of all $\boldsymbol{a}_i$ ’s, and $\{\boldsymbol{x}^{t,\text{sgn},(l)}\}$ is independent of both.

These auxiliary sequences are constructed by slightly perturbing the original data, and hence one can expect all of them to stay close to the original sequence throughout the execution of the algorithm.

General Results

Theorem 2.2

Fix $\boldsymbol{x}^{\natural} \in \mathbb{R}^n$. *Suppose* $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)\ (1 \le i \le m)$ *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*

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

and that the stepsize satisfies $\eta_t \equiv \eta = c/\left\|\boldsymbol{x}^{\natural}\right\|_2^2$ *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 that with probability at least* $1 - O\left(m^2 e^{-1.5n}\right) - O\left(m^{-9}\right)$,

General Results

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

$$\mathrm{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \leq \left(1 - \frac{\eta}{2}\left\|\boldsymbol{x}^\natural\right\|_2^2\right)^{t-T_\gamma} \cdot \gamma \left\|\boldsymbol{x}^\natural\right\|_2, \quad \forall t \geq T_\gamma;$$

General Results (continue)

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

$$\boldsymbol{x}^t - \frac{\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle}{\|\boldsymbol{x}^\natural\|_2^2}\boldsymbol{x}^\natural \text{ obeys}$$

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

for some constant $c_1 > 0$.

Remark

- Our current sample complexity reads $m \gtrsim n \log^{13} m$, which is optimal up to logarithmic factors.
- We can also prove similar performance guarantees for noisy phase retrieval.
- The random initialization $\boldsymbol{x}^0 \sim \mathcal{N}\left(\mathbf{0}, n^{-1} \left\|\boldsymbol{x}^\natural\right\|_2^2 \boldsymbol{I}_n\right)$ 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 $\left\|x^\natural\right\|$, 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.

Initial value

Theorem 2.3

Let

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

where $\boldsymbol{u}$ *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 ideas

- $d(\boldsymbol{x}_t, \boldsymbol{x}^\natural) \leq |\alpha_t - 1| + |\beta_t| < \gamma$.
- Two stages convergence: $t < T_\gamma$, $t > T_\gamma$.
- Approximate state evolution:

$$\alpha_{t+1} = \{1 + 3\eta[1 - (\alpha_t^2 + \beta_t^2)] + \eta\zeta_t\}\alpha_t,$$

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

- Control the perturbation terms $\{\zeta_t\}, \{\rho_t\}$ with leave-one-out approach.

Decomposition

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

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

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

$$r_1(\boldsymbol{x}^t) = \underbrace{\left[1 - \left(x_{\|}^t\right)^2\right] x_{\|}^t \left(\frac{1}{m}\sum_{i=1}^m a_{i,1}^4 - 3\right)}_{:=I_1} + \underbrace{\left[1 - 3\left(x_{\|}^t\right)^2\right] \frac{1}{m}\sum_{i=1}^m a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t}_{:=I_2}$$

$$\underbrace{- 3x_{\|}^t \left(\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 a_{i,1}^2 - \|\boldsymbol{x}_\perp^t\|_2^2\right)}_{:=I_3} - \underbrace{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 a_{i,1}}_{:=I_4}. \qquad (27)$$

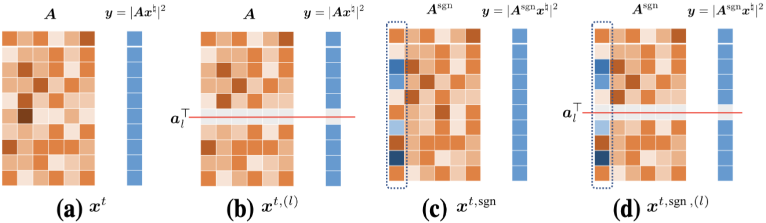
Control

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

$$I_4 = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)^{3}\left|a_{i,1}\right|\xi_i$$

with $\xi_i := \operatorname{sgn}(a_{i,1})$. Here $\operatorname{sgn}(\cdot)$ 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

Illustration of the leave-one-out and random-sign sequences


$\boldsymbol{A}$
$\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}^{\natural}|^2$
$\boldsymbol{A}$
$\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}^{\natural}|^2$
$\boldsymbol{a}_l^{\top}$
$\boldsymbol{A}^{\mathrm{sgn}}$
$\boldsymbol{y} = |\boldsymbol{A}^{\mathrm{sgn}}\boldsymbol{x}^{\natural}|^2$
$\boldsymbol{A}^{\mathrm{sgn}}$
$\boldsymbol{y} = |\boldsymbol{A}^{\mathrm{sgn}}\boldsymbol{x}^{\natural}|^2$
$\boldsymbol{a}_l^{\top}$

(a) $\boldsymbol{x}^{t}$ **(b)** $\boldsymbol{x}^{t,(l)}$ **(c)** $\boldsymbol{x}^{t,\mathrm{sgn}}$ **(d)** $\boldsymbol{x}^{t,\mathrm{sgn},(l)}$

Leave-one-out approach

Outline

1 Introduction

2 Main Result

3 Conclusion

Further investigation

- Sample complexity and phase transition.
- Other nonconvex statistical estimation problems. (low-rank matrix and tensor recovery, blind deconvolution and neural networks)
- Other iterative optimization methods. (alternating minimization, Kaczmarz algorithm, and truncated gradient descent)
- Beyond Gaussian sampling vectors. (Rademacher sampling model)
- Applications of leave-one-out tricks.