Spectral Initialization for Phase Retrieval: Theory and Proofs

Soo Min Kwon, Nitya Sathyavageeswaran

Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey

May 9, 2022

Learning objectives of this project

The main objectives of our project are

1. To use concentration inequalities to show that an average of random matrices is close to its expectation
2. To show how we can condition on high probability events in order to use concentration inequalities
3. To use concepts that we learned such as rotational invariance, union bounds and matrix Bernstein's inequality

What is phase retrieval?

Mathematically, phase retrieval is the problem of recovering a complex signal $\mathbf{x} \in \mathbb{C}^n$ given measurements $\mathbf{a}_i \in \mathbb{C}^n$ and observations

$$y_i = |\mathbf{a}_i^\top \mathbf{x}|^2, \quad i = 1, \ldots, m.$$

- This problem is also referred to as quadratic sensing or non-linear compressed sensing
- Generally "harder" to theoretically analyze than linear compressed sensing
- Occurs in many imaging domains such as X-ray crystallography and Fourier ptychography

Why do we care about phase retrieval?

For Fourier ptychography:

- Used to solve challenges regarding microscopes (tradeoff between resolution and field of view)
- Microscopes capture the intensity of the parts of an image given by the Fourier spectrum
- Cannot capture the complex values

For X-ray crystallography:

- Exposes crystals to x-rays to capture diffracted patterns
- Sensing apparatus is only able to observe the amplitude of the intensities (or patterns)

What are some algorithms for phase retrieval?

There are a lot of existing algorithms:

- Convex: PhaseLift [Candés et al. 2015]
- Non-convex: PhaseCut [Waldspurger et al. 2013], AltMinPhase [Netrapalli et al. 2015], Wirtinger Flow [Candés et al. 2015], Truncated Wirtinger Flow [Chen et al. 2015], Reshaped Wirtinger Flow [Zhang et al. 2016], ...

We will try to understand the theoretical guarantees of AltMinPhase

Intuition behind AltMinPhase

Recall that the goal is to recover $\mathbf{x} \in \mathbb{C}^n$ given $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{A} \in \mathbb{C}^{n \times m}$.

- The main issue is that we do not have access to the true phases of $\mathbf{y}$.
- What would we do if we did have access to the phases of y_i (i.e. $c_i = \text{Phase}(\langle \mathbf{a}_i, \mathbf{x} \rangle)$)?
- Then the problem just simplifies to solving a least squares problem:

$$\mathbf{C}\mathbf{y} = \mathbf{A}^\top \mathbf{x},$$

where $\mathbf{C} := \text{Diag}(\mathbf{c})$ is a diagonal matrix.

Intuition behind AltMinPhase

Of course, we do not have the phase matrix $\mathbf{C}$:

- We can try and estimate $\mathbf{C}$ given a good estimate of $\mathbf{x}$.
- Given a good estimate of $\mathbf{C}$, we can estimate $\mathbf{x}$ using least squares.
- Then, we can alternately update $\mathbf{C}$ and $\mathbf{x}$ until we get a "good enough" solution for $\mathbf{x}$.

But wait, we said that we can estimate $\mathbf{C}$ given a good estimate of $\mathbf{x}$.

Q: How can we obtain a good initial estimate of $\mathbf{x}$?

A: Use spectral initialization to initialize $\mathbf{x}$!

What is spectral initialization?

- Spectral initialization is just a fancy way of saying that we can find an initial estimate of $\mathbf{x}$ that is close to the true $\mathbf{x}^*$ with high probability.
- The term "spectral" comes from the use eigenvectors of properly designed matrices from data
- This just means we can construct a matrix from $\mathbf{y}$ and $\mathbf{A}$ and use the top eigenvector of the matrix as our initial estimate
- We see spectral initialization a lot in non-convex optimization problems [Chen et al. 2021]

Spectral initialization for AltMinPhase

The spectral initialization step for our problem involves **taking the top eigenvector of**

$$\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{a}_i \mathbf{a}_i^\top .$$

This is mainly because the expectation of $\mathbf{S}$ (assuming that $\mathbf{a}$ is Gaussian) is

$$\mathbb{E}[\mathbf{S}] = 2\mathbf{x}\mathbf{x}^\top + \|\mathbf{x}\|_2^2 .$$

Spectral initialization for AltMinPhase

So then why is taking the top eigenvector $\mathbf{S}$ a good estimate of $\mathbf{x}$?
Let's look at what taking the top eigenvector of $\mathbb{E}[\mathbf{S}]$ gives us:

$$\begin{aligned}
\mathbb{E}[\mathbf{S}]\mathbf{u} &= \lambda\mathbf{u} \\
(2\mathbf{x}\mathbf{x}^\top + \|\mathbf{x}\|_2^2\mathbf{I})\mathbf{u} &= \lambda\mathbf{u} \\
\mathbf{x}^\top(2\mathbf{x}\mathbf{x}^\top + \|\mathbf{x}\|_2^2\mathbf{I})\mathbf{u} &= \mathbf{x}^\top\lambda\mathbf{u} \\
2\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top\mathbf{u} + \mathbf{x}^\top\mathbf{x}^\top\mathbf{x}\mathbf{u} &= \lambda\mathbf{x}^\top\mathbf{u} \\
2\mathbf{x}^\top\mathbf{x}(\mathbf{x}^\top\mathbf{u}) + \mathbf{x}^\top\mathbf{x}(\mathbf{x}^\top\mathbf{u}) &= \lambda(\mathbf{x}^\top\mathbf{u}) \\
3\|\mathbf{x}\|_2^2 &= \lambda.
\end{aligned}$$

The leading eigenvector $\mathbf{u}_1$ is equivalent to $\mathbf{u}_1 = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ with eigenvalue $\lambda = 3\|\mathbf{x}\|_2^2$!

Theoretical guarantees for spectral initialization

- All we said so far was that given enough samples m, we can find a good initial estimate $\mathbf{x}$
- We want to make this argument more rigorous using the tools from high-dimensional probability!

Theorem (Netrapalli et al. 2015)

There exists a constant C_1 such that if $m \geq \frac{C_1}{c^2} n \log^3 n$, then the spectral initialization of AltMinPhase guarantees that

$$\mathit{dist}(\mathbf{x}^0, \mathbf{x}^*) \leq \sqrt{c}$$

with probability greater than $1 - \frac{4}{m^2}$, where

$$\mathit{dist}(\mathbf{w}_1, \mathbf{w}_2) := \sqrt{1 - \left| \frac{\langle \mathbf{w}_1, \mathbf{w}_2 \rangle}{\|\mathbf{w}_1\|_2 \|\mathbf{w}_2\|_2} \right|^2}.$$

Proof outline

The goal is to show that given sufficiently large m, the top eigenvector of $\mathbf{S}$ is close to $\mathbf{x}^*$ w.h.p:

1. We want to first show that $\mathbf{S}$ is close to $\mathbb{E}[\mathbf{S}]$ w.h.p using matrix Bernstein's inequality
2. Show that if $\mathbf{S}$ is close to $\mathbb{E}[\mathbf{S}]$, then the top eigenvector of $\mathbf{S}$ is close to $\mathbf{x}^*$

We will prove this under the Gaussian assumption, i.e. $\mathbf{a}_i \sim \mathcal{CN}(0, \mathbf{I})$.

Proof outline

The goal is to show that given sufficiently large m, the top eigenvector of $\mathbf{S}$ is close to $\mathbf{x}^*$ w.h.p:

1. **We want to first show that $\mathbf{S}$ is close to $\mathbb{E}[\mathbf{S}]$ w.h.p using matrix Bernstein's inequality**
2. Show that if $\mathbf{S}$ is close to $\mathbb{E}[\mathbf{S}]$, then the top eigenvector of $\mathbf{S}$ is close to $\mathbf{x}^*$

We will prove this under the Gaussian assumption, i.e. $\mathbf{a}_i \sim \mathcal{CN}(0, \mathbf{I})$.

The main tool that we will need

Theorem (Matrix Bernstein (Tropp 2012))

Consider a finite sequence of $\mathbf{X}_i$ *of independent random matrices with dimensions* $n \times n$*. Assume that* $\mathbb{E}[\mathbf{X}_i] = 0$ *and* $\|\mathbf{X}_i\| \leq R$ *for all* i*, almost surely. Let* $\sigma^2 := \| \sum_i \mathbb{E}[\mathbf{X}_i^2] \|_2$*. Then, for all* $t \geq 0$*, the following holds:*

$$\mathbb{P}\left(\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{X}_i \right\|_2 \geq t \right) \leq 2n \exp\left(\frac{-m^2 t^2}{\sigma^2 + Rmt/3} \right).$$

Proof of initialization stage of AltMinPhase

Recall that we want to first show that $\mathbf{S}$ is close to $\mathbb{E}[\mathbf{S}]$, where

$$\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{a}_i \mathbf{a}_i^\top \tag{1}$$

$$= \frac{1}{m}\sum_{i=1}^{m} |\mathbf{a}_i^\top \mathbf{x}|^2 \mathbf{a}_i \mathbf{a}_i^\top . \tag{2}$$

- However, we want to show this **for all** $\mathbf{x}$.
- But remember that $\mathbf{a}_i \sim \mathcal{CN}(0, \mathbf{I}_n)$ and are **rotationally invariant**.
- That makes $\mathbf{S}$ rotationally invariant, which then we can set $\mathbf{x} = \mathbf{e}_1$, where $\mathbf{e}_1$ is the first elementary vector and simply rotate.

Attempting to apply matrix Bernstein

Now, let $\mathbf{S}$ be

$$\mathbf{S} = \frac{1}{m} \sum_{i=1}^{m} |\langle \mathbf{a}_i, \mathbf{e}_1 \rangle|^2 \mathbf{a}_i \mathbf{a}_i^\top \tag{3}$$

$$= \frac{1}{m} \sum_{i=1}^{m} |a_{1,i}|^2 \mathbf{a}_i \mathbf{a}_i^\top \tag{4}$$

and $\mathbf{S}_i = |\langle \mathbf{a}_i, \mathbf{e}_1 \rangle|^2 \mathbf{a}_i \mathbf{a}_i^\top$. To apply matrix Bernstein on $\mathbf{S}$, we need two assumptions to hold:

1. $\mathbb{E}[\mathbf{S}_i] = 0$
2. $\|\mathbf{S}_i\|_2 \leq R$ for all i, almost surely

Let's see if these assumptions hold.

Handling the first assumption

Recall that

$$\mathbf{S} = \frac{1}{m} \sum_{i=1}^{m} |a_{1,i}|^2 \mathbf{a}_i \mathbf{a}_i^\top \tag{5}$$

and $\mathbf{a}_i \sim \mathcal{CN}(0, \mathbf{I}_n)$.

Due to the term $|a_{1,i}|^2$,

$$\mathbf{S}_i = \begin{bmatrix} |a_{1,i}|^4 & & \\ & \ddots & \\ & & |a_{1,i}|^2 |a_{n,i}|^2 \end{bmatrix}. \tag{6}$$

Thus, $\mathbb{E}[\mathbf{S}_i] \neq 0$.

Handling the first assumption

Even though $\mathbb{E}[\mathbf{S}_i] \neq 0$, dealing with this is quite simple – we can just show

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{S}_i - \mathbb{E}[\mathbf{S}_i]\right\|_2 \geq t\right) \leq \alpha(-mt).$$

Note that this is actually what we wanted to show anyways, so it all works out!

Handling the second assumption

Now, what about the second assumption $\|\mathbf{S}_i\|_2 \leq R$ for all i? Recall that

$$\mathbf{S}_i = \begin{bmatrix} |a_{1,i}|^4 & & \\ & \ddots & \\ & & |a_{1,i}|^2|a_{n,i}|^2 \end{bmatrix}, \tag{7}$$

with $|a_{j,i}|$ being complex Gaussian. Since $|a_{j,i}|$ is Gaussian, it is unbounded almost surely.

What can we do to solve this issue?

Conditioning on high probability events

What we can do is **condition on the fact that $\mathbf{a}_i$ is well behaved**. Specifically, define

$$\underbrace{\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{S}_i - \mathbb{E}[\mathbf{S}_i]\right\|_2 \geq t\right)}_{E} \leq \alpha(-mt). \tag{8}$$

Now, define

$$\mathbb{P}\underbrace{\left(\|\mathbf{S}_i\|_2 \geq R\right)}_{A} \leq \alpha\left(\frac{1}{m}\right). \tag{9}$$

What we want to show is that

$$\mathbb{P}(E^C) = \mathbb{P}(E^C|A)\cdot\mathbb{P}(A) + \mathbb{P}(E^C|A^C)\cdot\mathbb{P}(A^C). \tag{10}$$

Conditioning on high probability events

Note that

$$\mathbb{P}(E^C) = \underbrace{\mathbb{P}(E^C|A)}_{\text{use concentration}} \cdot \underbrace{\mathbb{P}(A)}_{\leq 1} + \underbrace{\mathbb{P}(E^C|A^C)}_{\leq 1} \cdot \underbrace{\mathbb{P}(A^C)}_{\text{tiny}} \tag{11}$$

$$\leq \mathbb{P}(E^C|A) + \mathbb{P}(A^C) \tag{12}$$

$$\leq \mathbb{P}(E^C|A) + \alpha\left(\frac{1}{m}\right). \tag{13}$$

We pay a price of event A happening by conditioning.

We need to find out what $\mathbb{P}(A)$ actually is. To do this, we need to look at what $\|\mathbf{S}_i\|_2$ is.

Looking at the norm of $\mathbf{S}_i$

Some algebra shows that

$$\begin{aligned} \|\mathbf{S}_i\|_2 &= \sigma_{\mathsf{max}}(\mathbf{S}_i) && (14) \\ &\leq \mathsf{tr}(\mathbf{S}_i) && (15) \\ &= |a_{i,1}|^2 \|\mathbf{a}_i\|_2^2. && (16) \end{aligned}$$

Using this inequality, what we want to show is

$$\mathbb{P}(|a_{i,1}|^2 \|\mathbf{a}_i\|_2^2 \geq R) \leq \alpha \left(\frac{1}{m}\right), \tag{17}$$

and then use this to prove our result.

Some more tools that we will need

Lemma (Bound on the Spectral Norm)

Suppose $\mathbf{a}_i \sim \mathcal{CN}(0, \mathbf{I}_n)$ *and* $a_{1,i}$ *be the first element of* $\mathbf{a}_i$. *Then,*

$$\mathbb{P}(|a_{1,i}|^2 \|\mathbf{a}_i\|^2 \geq 12 \log m) \leq \frac{3}{m^2}.$$

The proof for this is provided at the end of these slides.

Applying matrix Bernstein via conditioning

Going back, so far we have

$$\mathbb{P}(E) \leq \mathbb{P}(E|A) + \frac{3}{m^2}. \tag{18}$$

For $\mathbb{P}(E|A)$, we can apply matrix Bernstein using $\|\mathbf{S}_i\|_2 \leq 12 \log m$:

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{S}_i - \mathbb{E}[\mathbf{S}_i]\right\|_2 \geq t\right) \leq 2n \exp\left(-\frac{m^2t^2}{4nmt\log m}\right). \tag{19}$$

Now, to match the probability of $\frac{4}{m^2}$ on their Theorem, we need to solve for t with

$$2n \exp\left(-\frac{m^2t^2}{4nmt\log m}\right) = \frac{1}{m^2}. \tag{20}$$

Solving for t

Some algebra shows that

$$t = \frac{4n \log(m + 2nm^2)}{m} \quad \text{(our term)} \tag{21}$$

$$\leq \frac{4n \log^{3/2}(m)}{\sqrt{m}} \quad \text{(their term).} \tag{22}$$

Thus, we have shown that w.p $1 - \frac{1}{m^2}$,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{S}_i - \mathbb{E}[\mathbf{S}_i] \right\|_2 \leq \frac{4n \log^{3/2}(m)}{\sqrt{m}}. \tag{23}$$

Solving for m

Given

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{S}_i - \mathbb{E}[\mathbf{S}_i] \right\|_2 \leq \frac{4n \log^{3/2}(m)}{\sqrt{m}}, \tag{24}$$

if we set the right hand side to ϵ, some algebra shows that

$$m \geq \frac{C}{\epsilon^2} n^2 \log^3 n \tag{25}$$

$$\geq \frac{C}{\epsilon^2} n \log^3 n, \tag{26}$$

which is the sample complexity we wanted to show.

Putting everything together

Combining everything together, we have shown that there exists some $C > 0$ such that if $m \geq \frac{C}{\epsilon^2} n \log^3 n$, then

$$\mathbb{P}\left(\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{S}_i - \mathbb{E}[\mathbf{S}_i] \right\|_2 \leq \epsilon \right) \geq 1 - \frac{4}{m^2}. \tag{27}$$

What we learned from this project

- We had a hard time understanding the authors' proofs – we should take the time to write clearer proofs for other people to understand more easily
- Conditioning on high probability events is useful in scenarios in which assumptions may not hold
- A nice example to use some of the tools that we learned from this course

Thank You!

Proofs of Lemmas

Lemma (Tail-Bound on Chi-Squared)

Suppose $x \sim \mathcal{CN}(0,1)$. *Then,*

$$\mathbb{P}(|x|^2 \geq 6\log m) = \frac{1}{m^3}.$$

Lemma (Tail-Bound on Norm of Complex Gaussian)

Suppose $\mathbf{x} \sim \mathcal{CN}(0,\mathbf{I})$. *Then,*

$$\mathbb{P}(\|\mathbf{x}\|^2 \geq 2n) \leq 2\exp(-n^2/2).$$

Proof of Lemma (Tail Bound on Chi-Squared)

$|x|^2$ is the sum of squares of two $\mathcal{N}(0,1)$ random variables, thus it is a chi-squared random variable with 2 degrees of freedom.
The CDF of a chi-squared random variable with 2 degrees of freedom is given by

$$F_X(x) = 1 - \exp\left(-x/2\right). \tag{28}$$

Hence we have,

$$\begin{aligned} \mathbb{P}(|x|^2 \geq 6\log m) &= 1 - F_X(6\log m) \\ &= \exp\left(-3\log m\right) \\ &= \left(\frac{1}{m^3}\right). \end{aligned} \tag{29}$$

Proof of Lemma (Bound on Norm of Complex Gaussian)

We know that, if $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$, then $\|\mathbf{x}\|^2$ is close to n and we have

$$\mathbb{P}(\|\mathbf{x}\|^2 \geq n) \leq \exp(-n^2/2). \tag{30}$$

Here, since $\mathbf{x} \sim \mathcal{CN}(0, \mathbf{I})$ we need to consider Re$\{\mathbf{x}\}$ and Im$\{\mathbf{x}\}$, and hence

$$\mathbb{P}(\|\mathbf{x}\|^2 \geq 2n) \leq 2\exp(-n^2/2). \tag{31}$$

Proof of Lemmas

Lemma (Bound on the Spectral Norm)

Suppose $\mathbf{a}_i \sim \mathcal{CN}(0, \mathbf{I}_n)$ *and* $a_{1,i}$ *be the first element of* $\mathbf{a}_i$. *Then,*

$$\mathbb{P}(|a_{1,i}|^2 \|\mathbf{a}_i\|^2 \geq 12 \log m) \leq \frac{3}{m^2}.$$

Proof of Lemma (Bound on the Spectral Norm)

From (29) and (31) we have

$$\mathbb{P}(|a_{1,i}|^2\|\mathbf{a}_i\|^2 \geq 12n\log m) \leq 2\exp\left(-n^2/2\right) + \frac{1}{m^3}. \tag{32}$$

However, this is true for only one of the a_i's. By applying a union bound over all i, we get

$$\begin{aligned}\mathbb{P}(\cup_{i=1}^m |a_{1,i}|^2\|\mathbf{a}_i\|^2 \geq 12n\log m) &\leq 2m\exp\left(-n^2/2\right) + \frac{1}{m^2}. \\ &= 2\exp\left(-n^2/2 + \log m\right) + \frac{1}{m^2}. \\ &\leq \frac{2}{m^2} + \frac{1}{m^2}. \\ &= \frac{3}{m^2}. \end{aligned} \tag{33}$$