Non-Convex Phase Synchronization via Projected Gradient Ascent with Provable Estimation and Convergence Guarantees

Anthony Man-Cho So
Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong (CUHK)

(Joint Work with Huikang Liu and Man-Chung Yue)

The Greater Bay Area Optimization Day
The Hong Kong Polytechnic University

24 February 2018

Phase Synchronization

- Goal: Estimate an unknown phase vector $z^\star \in \mathbb{T}^n = \{z \in \mathbb{C}^n : |z_1| = \cdots = |z_n| = 1\}$ from the noisy measurements

$$C_{j\ell} = z_j^\star \bar{z}_\ell^\star + \Delta_{j\ell} \quad \text{for } 1 \le j < \ell \le n.$$

  Here, we assume that the noise satisfies $\Delta_{jj} = 0$ and $\Delta_{j\ell} = \bar{\Delta}_{\ell j}$. Then, we can write $C = z^\star {z^\star}^H + \Delta$ with $C_{jj} = 1$ and $C, \Delta \in \mathbb{H}^n$.

- Various applications
  - lensless imaging
  - clock synchronization in wireless networks
  - localization on circle
  - ranking of items based on noisy pairwise comparisons

Phase Synchronization

- Consider the following natural least-squares formulation:

$$\hat{z} \in \arg\min_{z\in\mathbb{T}^n} \sum_{1\le j<\ell\le n} \left|C_{j\ell} - z_j\bar{z}_\ell\right|^2 = \arg\max_{z\in\mathbb{T}^n} \left\{ f(z) = z^H C z \right\}. \qquad \text{(QP)}$$

  - Note that if $\hat{z}$ is an optimal solution, then so is $e^{i\theta}\hat{z}$ for any $\theta \in [0, 2\pi)$. Thus, we can at most identify $z^\star$ up to a global phase.
  - In the Gaussian noise setting (i.e., $\Delta = \sigma W$, where $\sigma^2 > 0$ is the noise power and W is a Wigner matrix), any optimal solution to (QP) is a maximum likelihood estimator of $z^\star$.

Phase Synchronization

- Given the estimation problem (QP), we are interested in the following:
  - Is $\hat{z}$ (optimal solution) close to $z^\star$ (ground truth)?
  - Can we find $\hat{z}$ efficiently?

- Recall that the value $z^H C z$ is invariant under multiplication of a common phase (unit-modulus complex number).
  - This motivates the following distance measure for two points w, z on $\mathbb{T}^n$:

$$d_2(w, z) = \min_{\theta \in [0, 2\pi)} \|w - e^{i\theta} z\|_2.$$

- **Fact: [Bandeira-Boumal-Singer'17]** (Estimation Error of $\hat{z}$)

$$d_2(\hat{z}, z^\star) \leq \frac{4\|\Delta\|_{\mathrm{op}}}{\sqrt{n}}.$$

- Thus, we shall mainly focus on the second question.

Phase Synchronization

- Given an arbitrary C, Problem (QP) is NP-hard **[Toker-Özbay'98]**.

- However, we have a highly structured C. Perhaps we can find $\hat{z}$ efficiently?

- A natural apprach: SDP relaxation

$$\min_{Z \in \mathbb{H}^n} \operatorname{Tr}(CZ) \quad \text{subject to} \quad \operatorname{diag}(Z) = \boldsymbol{e},\ Z \succeq \mathbf{0}. \tag{SDP}$$

  – **Fact: [Bandeira-Boumal-Singer'17]** Under the Gaussian noise setting, if $\sigma = O(n^{1/4})$, then (SDP) admits a unique optimal solution $\hat{Z}$ that is of rank-one.

- Despite its theoretical appeal, solving large instances of (SDP) is considered to be computationally expensive.

Can we develop fast methods for solving (QP) while
still being able to establish some theoretical guarantee?

Generalized Power Method for Phase Synchronization

Consider the following projected gradient-type scheme for solving (QP):

Algorithm 1 Generalized Power Method for Problem (QP) **[Boumal'16]**

```
input: objective matrix $C \in \mathbb{H}^n$, step size $\alpha > 0$, initial point $z^0 \in \mathbb{T}^n$
for $k = 0, 1, \ldots$ do
    if termination criterion is met then
        return $z^k$
    else
        $w^k \leftarrow \left(I + \frac{\alpha}{n} C\right) z^k$    // gradient ascent
        $z^{k+1} \leftarrow \frac{w^k}{|w^k|}$    // projection onto $\mathbb{T}^n$
    end if
end for
```

- Here, $2Cz^k$ is the gradient of $z \mapsto z^H C z$ at z^k and

$$\left(\frac{v}{|v|} \right)_j = \begin{cases} \frac{v_j}{|v_j|} & \text{if } v_j \neq 0, \\ 1 & \text{otherwise} \end{cases}$$

is the projection of $v \in \mathbb{C}^n$ onto $\mathbb{T}^n$, which can be efficiently computed.

Generalized Power Method for Phase Synchronization

- When $\alpha \to \infty$, the update becomes $z^{k+1} \leftarrow \frac{Cz^k}{|Cz^k|}$, which resembles the power method for computing the dominant eigenvector of C, hence the name *generalized power method* (GPM) for Algorithm 1.

- Problem (QP) is non-convex in general. Thus, one expects that the initial point will affect the convergence performance of the algorithm.

- Consider the so-called spectral initialization $z^0 = v_C$, where $v_C = \frac{u}{|u|}$ and u is a dominant eigenvector of C.

- As it turns out, the spectral initialization has many nice properties.

Generalized Power Method for Phase Synchronization

- **Fact: [Boumal'16]** (Estimation Error of v_C)

$$d_2(v_C, z^\star) \leq \frac{8\|\Delta\|_{\mathrm{op}}}{\sqrt{n}}$$

  (recall that $d_2(\hat{z}, z^\star) \leq \frac{4\|\Delta\|_{\mathrm{op}}}{\sqrt{n}}$).

- **Fact: [Boumal'16]** Under the Gaussian noise setting, if $\sigma = O(n^{1/6})$ and $\alpha = O(n^{1/2})$, then the GPM with spectral initialization will converge to an optimal solution to (QP).

- The above results do not resolve three key issues:
  - Are the estimation errors of the iterates $\{z^k\}$ generated by the GPM monotonically decreasing?
  - The convergence result applies to the noise regime $\sigma = O(n^{1/6})$, while the SDP tightness result applies to the less restrictive noise regime $\sigma = O(n^{1/4})$. Could the convergence result for the GPM apply to the latter noise regime?
  - Can we establish the convergence rate of the GPM?

Estimation Performance of the GPM

- **Theorem: [Liu-Yue-S.'17]** (Estimation Errors of the GPM Iterates) Suppose that

  - $\|\Delta\|_{\rm op} \le \frac{n}{16}$,
  - $\alpha \ge 6$,
  - $z^0 = v_C$.

  Then, the iterates $\{z^k\}_{k\ge 0}$ generated by the GPM satisfy

$$d_2(z^{k+1}, z^\star) \le \mu^{k+1} \cdot d_2(z^0, z^\star) + \frac{\nu}{1-\mu} \cdot \frac{8\|\Delta\|_{\rm op}}{\sqrt{n}}$$

  for some $\mu \in (0,1)$ and $\nu > 0$.

- Curious observation: The above result does not assume the convergence of the GPM. It gives upper bounds on the estimation errors of the iterates, and the bounds are decreasing.

- Under the Gaussian noise setting, we have $\|\Delta\|_{\rm op} = O(\sigma n^{1/2})$ with high probability. Thus, the above result applies to the noise regime $\sigma = O(n^{1/2})$.

Key Property of Projection onto $\mathbb{T}^n$

- The proof of the theorem relies crucially on the following result:

  Proposition: For any $w \in \mathbb{C}^n$ and $z \in \mathbb{T}^n$,

  $$\left\| \frac{w}{|w|} - z \right\|_2 \le 2\|w - z\|_2.$$

- Roughly speaking, it says that the projection operator onto $\mathbb{T}^n$ is not too expansive (recall that projections onto closed convex sets are always non-expansive).

Convergence Rate of the GPM

- From an optimization-theoretic viewpoint, we are still interested in knowing whether the GPM will converge or not, and if so, what is the rate of convergence.

- To answer these questions, we need to first get a handle on the set of potential optimal solutions to (QP).

Optimality Conditions of Problem (QP)

- Recall

$$\hat{z} \in \arg\max_{z \in \mathbb{T}^n} \left\{ f(z) = z^H C z \right\}. \tag{QP}$$

- Viewing $\mathbb{T}^n$ as a manifold, the tangent space to $\mathbb{T}^n$ at $z \in \mathbb{T}^n$ is given by

$$T_z\mathbb{T}^n = \{ w \in \mathbb{C}^n : \Re\{w_j \bar{z}_j\} = 0 \text{ for } j = 1, \ldots, n \}.$$

The projector onto $T_z\mathbb{T}^n$ can be computed as

$$\Pi_{T_z\mathbb{T}^n}(w) = w - \mathrm{Diag}(\Re\{z_j \bar{w}_j\})z.$$

This yields the first-order optimality condition of (QP):

$$\mathbf{0} = \mathrm{grad}\, f(z) = \Pi_{T_z\mathbb{T}^n}(2Cz) = -2S(z)z,$$

where $S(z) = \mathrm{Diag}(\Re((Cz)_j \bar{z}_j)) - C$. We call $\mathrm{grad}\, f(z)$ the Riemannian gradient of f at $z \in \mathbb{T}^n$.

Optimality Conditions of Problem (QP)

- We can also write down the second-order optimality condition. To do this, we need the Riemannian Hessian $\text{Hess}\, f(z)$ of f at $z \in \mathbb{T}^n$, which is obtained by projecting the directional derivatives of $\text{grad}\, f(z)$ onto the tangent space to $\mathbb{T}^n$ at $z \in \mathbb{T}^n$:

$$(\text{Hess}\, f(z))(w) = \Pi_{T_z\mathbb{T}^n}(D\, \text{grad}\, f(z))(w) = -\Pi_{T_z\mathbb{T}^n}(2S(z)w).$$

- Then, the second-order optimality condition is given by

$$w^H(\text{Hess}\, f(z))w = -2w^H S(z)w \le 0 \quad \text{for all } w \in T_z\mathbb{T}^n.$$

- **Fact:** Every optimal solution to (QP) is a second-order critical point; i.e., it satisfies both the first- and second-order optimality conditions.

- **Fact: [Boumal'16]** Every second-order critical point $\tilde{z} \in \mathbb{T}^n$ satisfies

$$\left(\text{Diag}(|\tilde{C}\tilde{z}|) - \tilde{C}\right)\tilde{z} = (\text{Diag}(|C\tilde{z}|) - C)\,\tilde{z} = \mathbf{0},$$

where $\tilde{C} = C + \frac{n}{\alpha}I$, for any $\alpha > 0$.

Proximity to Second-Order Criticality

- The earlier discussion motivates us to define the following:

$$\Sigma(z) = \mathrm{Diag}(|\tilde{C}z|) - \tilde{C}, \quad \rho(z) = \|\Sigma(z)z\|_2.$$

- Intuitively, the residual function ρ measures the proximity to the set of second-order critical points.
  - This suggests ρ can be used to measure the progress of the GPM.
  - Challenge: Although $\rho(\hat{z}) = 0$ for any optimal solution $\hat{z}$ to (QP), it is not clear the converse holds.

Local Error Bound for Problem (QP)

- **Proposition: [Liu-Yue-S.'17]** (Local Error Bound for (QP)) Under the Gaussian noise setting, if $\sigma = O(n^{1/4})$ and $\alpha \geq 4$, then with high probability, we have

$$d_2(z, \hat{z}) \leq \frac{8}{n}\rho(z) \qquad \text{(EB)}$$

  for any $z \in \mathbb{T}^n$ satisfying $d_2(z, z^\star) \leq \frac{\sqrt{n}}{2}$ and any optimal solution $\hat{z}$ to (QP).

  - This shows that every second-order critical point of (QP) is optimal.
  - The above error bound allows us to follow rather standard arguments to establish the convergence rate of the GPM.

Convergence Rate of the GPM

- **Proposition: [Liu-Yue-S.'17]** Under the Gaussian noise setting, suppose that
  - $\sigma = O(n^{1/4})$,
  - $\alpha = O(n^{1/4})$,
  - $z^0 = v_C$.

  Then, for any optimal solution $\hat{z}$ to (QP), the iterates $\{z^k\}_{k \geq 0}$ generated by the GPM possess the following properties:

  - **(Sufficient Ascent)** There exists a constant $a_0 > 0$ such that
  $$f(z^{k+1}) - f(z^k) \geq a_0 \cdot \|z^{k+1} - z^k\|_2^2. \tag{A1}$$
  - **(Cost–to–Go Estimate)** There exists a constant $a_1 > 0$ such that
  $$f(\hat{z}) - f(z^k) \leq a_1 \cdot d_2(z^k, \hat{z})^2. \tag{A2}$$
  - **(Safeguard)** There exists a constant $a_2 > 0$ such that
  $$\rho(z^k) \leq a_2 \cdot \|z^{k+1} - z^k\|_2. \tag{A3}$$

Convergence Rate of the GPM

- By combining (EB) and (A1)–(A3), we have the following:

  Theorem: [Liu-Yue-S'.17] (Linear Convergence of the GPM) Under the above setting, for any optimal solution $\hat{z}$ to (QP), the iterates $\{z^k\}_{k \geq 0}$ generated by the GPM satisfy

$$\begin{aligned} f(\hat{z}) - f(z^k) &\leq \left(f(\hat{z}) - f(z^0) \right) \lambda^k, \\ d_2(z^k, \hat{z}) &\leq a \left(f(\hat{z}) - f(z^0) \right)^{1/2} \lambda^{k/2} \end{aligned}$$

  for some $a > 0$ and $\lambda \in (0, 1)$.

- The above result applies to the noise regime $\sigma = O(n^{1/4})$, which matches that required for the tightness of the SDP relaxation.

Extension to MIMO Detection

- The GPM is rather versatile and can be easily adapted to other problems.
- Take, for instance, MIMO detection, where we have the model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}^\star + \boldsymbol{\nu}.$$

  Here,

  - $\boldsymbol{y} \in \mathbb{C}^m$ is the received signal vector,
  - $\boldsymbol{H} \in \mathbb{C}^{m \times n}$ is the channel matrix,
  - $\boldsymbol{x}^\star \in \mathbb{C}^n$ is the transmitted symbol vector,
  - $\boldsymbol{\nu} \in \mathbb{C}^m$ is the noise vector.

- We assume that each symbol x_i is drawn from some discrete constellation $\mathcal{S}$, where $\mathcal{S}$ is either the $(4u^2)$-QAM constellation

$$\mathcal{Q}_u = \{z \in \mathbb{C} : \Re(z), \Im(z) = \pm 1, \pm 3, \cdots, \pm(2u-1)\}$$

  or the MPSK constellation

$$\mathcal{S}_M = \{\exp(2\pi i k/M) : k = 0, 1, \ldots, M-1\},$$

Extension to MIMO Detection

- We are interested in the following maximum likelihood (ML) estimation problem:

$$\min_{\boldsymbol{x} \in \mathcal{S}^n} \left\{ F(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 \right\}. \tag{ML}$$

- We can then write down the GPM for solving (ML) as follows:

Algorithm 2 Generalized Power Method for MIMO Detection

input: initial point $\boldsymbol{x}^0 \in \mathcal{S}^n$ and step sizes $\{\alpha_k\}_{k \geq 0}$
if stopping criterion is not met **then**
    $\nabla F(\boldsymbol{x}^k) \longleftarrow 2\boldsymbol{H}^*(\boldsymbol{H}\boldsymbol{x}^k - \boldsymbol{y})$
    $\boldsymbol{x}^{k+1} \longleftarrow \Pi_{\mathcal{S}^n}\left(\boldsymbol{x}^k - \frac{\alpha_k}{m}\nabla F(\boldsymbol{x}^k)\right)$
    $k \longleftarrow k + 1$
end if

- Note that for $\mathcal{S} = \mathcal{Q}_u$ and $\mathcal{S} = \mathcal{S}_M$, the projection $\Pi_{\mathcal{S}^n}(w)$ can be efficiently computed.

Analysis of the GPM for MIMO Detection

- Interestingly, we can also establish some convergence guarantee for the GPM.
- Key to the analysis is the following result, which concerns the not–too–expansiveness of the projection operator $\Pi_{\mathcal{S}^n}$:

  Proposition: [Liu-Yue-S.-Ma'17] Consider the case where $\mathcal{S} = \mathcal{Q}_u$ or $\mathcal{S} = \mathcal{S}_M$. Let $\boldsymbol{z} \in \mathbb{C}^n$ and $\boldsymbol{x} \in \mathcal{S}^n$ be given. Then,

$$\|\Pi_{\mathcal{S}^n}(\boldsymbol{z}) - \boldsymbol{x}\|_2 \leq 2\|\boldsymbol{z} - \boldsymbol{x}\|_2.$$

Analysis of the GPM for MIMO Detection

- **Theorem: [Liu-Yue-S.-Ma'17]** Let $\{\boldsymbol{x}^k\}_{k\geq 0}$ be the sequence of iterates generated by the GPM with step sizes $\{\alpha_k\}_{k\geq 0}$ satisfying

$$\left\|\frac{2\alpha_k}{m}\boldsymbol{H}^*\boldsymbol{\nu}\right\|_\infty < \frac{1}{c} \quad \text{and} \quad \left\|\boldsymbol{I} - \frac{2\alpha_k}{m}\boldsymbol{H}^*\boldsymbol{H}\right\|_{\rm op} \leq \beta < \frac{1}{4}, \tag{P}$$

  where $c = \frac{4}{\min_{s\neq s'\in\mathcal{S}}|s-s'|} < \infty$ (hence, we have $c = 2$ for $\mathcal{S} = \mathcal{Q}_u$ and $c = \frac{2}{\sin(\pi/M)}$ for $\mathcal{S} = \mathcal{S}_M$). Then, we have

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|_2 \leq 4\beta\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|_2$$

  for all $k \geq 0$.

- In particular, after at most $k^\star = \left\lceil \ln\left(\frac{2}{c\|\boldsymbol{x}^0-\boldsymbol{x}^\star\|_2}\right) / \ln(4\beta) \right\rceil$ iterations, we have $\boldsymbol{x}^k = \boldsymbol{x}^\star$ for all $k \geq k^\star$; i.e., the GPM admits finite convergence.

- Under suitable probabilistic assumptions on $\boldsymbol{H}$ and $\boldsymbol{\nu}$, one can show that condition (P) will be satisfied with high probability.

Closing Remarks

- There has been much recent interest in the design and analysis of fast methods for structured non-convex optimization problems.
- Many such problems arise in machine learning and signal processing.
- Similar to the convex case, error bounds play a fundamental role in the convergence analyses of these methods.

Thank You!