ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

Phase retrieval

Yuejie Chi

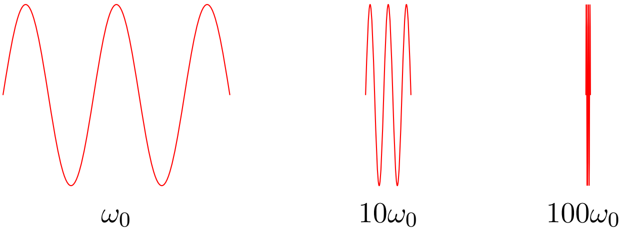
Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2018

Phase retrieval: the missing phase problem

In high-frequency (e.g. optical) applications, the (optical) detection devices [e.g., CCD cameras, photosensitive films, and the human eye] **cannot** measure the phase of a light wave.



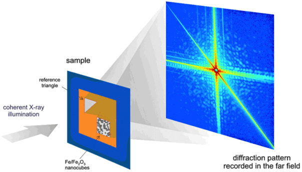
- Optical devices measure the *photon flux* (no. of photons per second per unit area), which is proportional to the magnitude.
- This leads to the so-called *phase retrieval* problem — inference with only intensity measurements.

Coherent diffraction imaging

Detectors record intensities of diffracted rays

- electric field $x(t_1, t_2) \longrightarrow$ Fourier transform $\hat{x}(f_1, f_2)$

Fig credit: Stanford SLAC



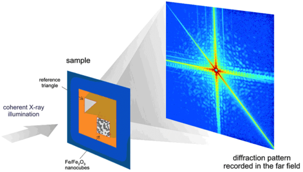
intensity of electrical field: $\left|\hat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Coherent diffraction imaging

Detectors record intensities of diffracted rays

- electric field $x(t_1, t_2) \longrightarrow$ Fourier transform $\hat{x}(f_1, f_2)$

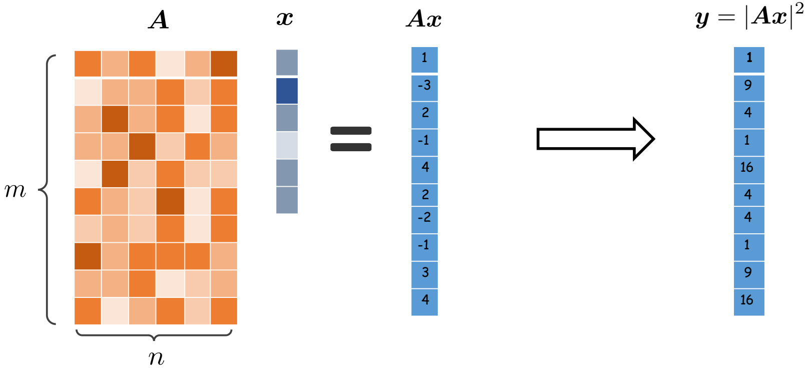
Fig credit: Stanford SLAC



intensity of electrical field: $\left|\hat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Phase retrieval: recover signal $x(t_1, t_2)$ from intensity $\left|\hat{x}(f_1, f_2)\right|^2$

Mathematical setup



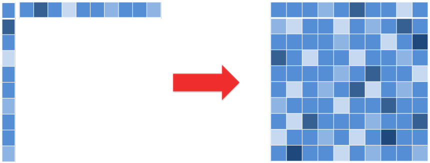
Recover $x^\natural \in \mathbb{R}^n$ from m random quadratic measurements

$$y_k = |a_k^\top x^\natural|^2, \qquad k = 1, \dots, m \tag{10.1}$$

An equivalent view: low-rank factorization

Lifting: Introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ to linearize constraints

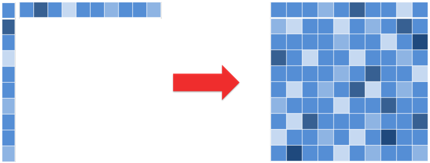
$$y_k \approx |\boldsymbol{a}_k^\top \boldsymbol{x}|^2 = \boldsymbol{a}_k^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} \qquad \Longrightarrow \qquad y_k \approx \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k$$



An equivalent view: low-rank factorization

Lifting: Introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ to linearize constraints

$$y_k \approx |\boldsymbol{a}_k^\top \boldsymbol{x}|^2 = \boldsymbol{a}_k^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} \qquad \Longrightarrow \qquad y_k \approx \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k$$

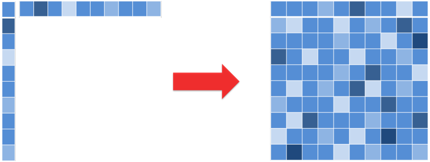


$$\begin{array}{ll} \text{find} & \boldsymbol{X} \\ \text{s.t.} & y_k \approx \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k, \qquad k = 1, \cdots, m \\ & \color{red}{\text{rank}(\boldsymbol{X}) = 1} \\ & \boldsymbol{X} \succeq 0 \end{array}$$

An equivalent view: low-rank factorization

Lifting: Introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ to linearize constraints

$$y_k \;\approx\; |\boldsymbol{a}_k^\top \boldsymbol{x}|^2 = \boldsymbol{a}_k^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} \qquad \Longrightarrow \qquad y_k \approx \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k$$



$$\begin{array}{lll} \text{find} & \boldsymbol{X} & \\ \text{s.t.} & y_k \;\approx\; \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k, & k = 1, \cdots, m \\ & \textcolor{red}{\mathsf{rank}(\boldsymbol{X}) = 1} & \\ & \boldsymbol{X} \succeq 0 & \end{array}$$

Solving quadratic systems is essentially low-rank matrix completion

Solving quadratic systems is NP-complete ***in general***

The stone assignment problem (assign stones of weight w_i into two groups of equal weight) is NP-hard. Let

$$x_i^2 = 1; \forall i; (w_1 x_1 + w_2 x_2 + \cdots + w_n x_n)^2 = 0.$$

"I can't find an efficient algorithm, but neither can all these people."

figure credit: coding horror

Convex Relaxation

Rank-one measurements

Measurements: see (10.1)

$$y_i = \boldsymbol{a}_i^\top \underbrace{\boldsymbol{x}\boldsymbol{x}^\top}_{:=\boldsymbol{M}} \boldsymbol{a}_i = \langle \underbrace{\boldsymbol{a}_i\boldsymbol{a}_i^\top}_{:=\boldsymbol{A}_i}, \boldsymbol{M} \rangle, \qquad 1 \le i \le m$$

Define the measurement operator $\mathcal{A}$:

$$\mathcal{A}(\boldsymbol{X}) = \begin{bmatrix} \langle \boldsymbol{A}_1, \boldsymbol{X} \rangle \\ \langle \boldsymbol{A}_2, \boldsymbol{X} \rangle \\ \vdots \\ \langle \boldsymbol{A}_m, \boldsymbol{X} \rangle \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{a}_1\boldsymbol{a}_1^\top, \boldsymbol{X} \rangle \\ \langle \boldsymbol{a}_2\boldsymbol{a}_2^\top, \boldsymbol{X} \rangle \\ \vdots \\ \langle \boldsymbol{a}_m\boldsymbol{a}_m^\top, \boldsymbol{X} \rangle \end{bmatrix}$$

Rank-one measurements: $\boldsymbol{A}_i = \boldsymbol{a}_i\boldsymbol{a}_i^\top$ are rank-one!

Do rank-one measurements satisfy RIP?

Suppose $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

- If $\boldsymbol{x}$ is independent of $\{\boldsymbol{a}_i\}$, then

$$\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{x}\boldsymbol{x}^\top \rangle = \left|\boldsymbol{a}_i^\top \boldsymbol{x}\right|^2 \asymp \|\boldsymbol{x}\|^2 \;\Rightarrow\; \left\|\mathcal{A}(\boldsymbol{x}\boldsymbol{x}^\top)\right\|_{\mathrm{F}} \asymp \sqrt{m}\|\boldsymbol{x}\boldsymbol{x}^\top\|_{\mathrm{F}}$$

- Consider $\boldsymbol{A}_i = \boldsymbol{a}_i\boldsymbol{a}_i^\top$:

$$\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{A}_i \rangle = \|\boldsymbol{a}_i\|^4 \approx n\|\boldsymbol{a}_i\boldsymbol{a}_i^\top\|_{\mathrm{F}}$$

$$\Longrightarrow \quad \|\mathcal{A}(\boldsymbol{A}_i)\|_{\mathrm{F}} \geq |\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{A}_i \rangle| \approx n\|\boldsymbol{A}_i\|_{\mathrm{F}}$$

Do rank-one measurements satisfy RIP?

Suppose $\boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$

- If sample size $m \asymp n$ (information limit), then

$$\frac{\max_{\boldsymbol{X}:\,\mathsf{rank}(\boldsymbol{X})=1} \frac{\|\mathcal{A}(\boldsymbol{X})\|_{\mathrm{F}}}{\|\boldsymbol{X}\|_{\mathrm{F}}}}{\min_{\boldsymbol{X}:\,\mathsf{rank}(\boldsymbol{X})=1} \frac{\|\mathcal{A}(\boldsymbol{X})\|_{\mathrm{F}}}{\|\boldsymbol{X}\|_{\mathrm{F}}}} \gtrsim \frac{n}{\sqrt{m}} \gtrsim \sqrt{n}$$

$$\frac{\max_{\boldsymbol{X}:\,\mathsf{rank}(\boldsymbol{X})=1} \frac{\|\mathcal{A}(\boldsymbol{X})\|_{\mathrm{F}}}{\|\boldsymbol{X}\|_{\mathrm{F}}}}{\min_{\boldsymbol{X}:\,\mathsf{rank}(\boldsymbol{X})=1} \frac{\|\mathcal{A}(\boldsymbol{X})\|_{\mathrm{F}}}{\|\boldsymbol{X}\|_{\mathrm{F}}}} \gtrsim \sqrt{n} \gg 1$$

  - Violate RIP condition in Theorem **??**

Why do we lose RIP?

Problem:

- Low-rank matrices $\boldsymbol{X}$ (e.g. $\boldsymbol{a}_i\boldsymbol{a}_i^\top$) might be too aligned with some rank-one measurements
  - loss of incoherence in some measurements
- Some measurements $\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle$ might have too high of a leverage on $\mathcal{A}(\boldsymbol{X})$ when measured in $\|\cdot\|_{\mathrm{F}}$
  - Change $\|\cdot\|_{\mathrm{F}}$ to other norms!

Mixed-norm RIP

Solution: modify RIP appropriately ...

Definition 10.1 (RIP-ℓ_2/ℓ_1)

Let $\xi_r^{\mathrm{ub}}(\mathcal{A})$ and $\xi_r^{\mathrm{lb}}(\mathcal{A})$ be smallest quantities s.t.

$$(1-\xi_r^{\mathrm{lb}})\|\boldsymbol{X}\|_{\mathsf{F}} \leq \|\mathcal{A}(\boldsymbol{X})\|_1 \leq (1+\xi_r^{\mathrm{ub}})\|\boldsymbol{X}\|_{\mathsf{F}}, \qquad \forall \boldsymbol{X} : \mathrm{rank}(\boldsymbol{X}) \leq r$$

Analyzing phase retrieval via RIP-ℓ_2/ℓ_1

Theorem 10.2 (Chen, Chi, Goldsmith '15)

Suppose $\mathrm{rank}(\boldsymbol{M}) = r$. *For any fixed integer* $K > 0$, *if* $\frac{1+\delta_{Kr}^{\mathrm{ub}}}{1-\delta_{(2+K)r}^{\mathrm{lb}}} < \sqrt{\frac{K}{2}}$, *then nuclear norm minimization is exact.*

- Follows same proof/form as for Theorem 6.9, except that $\|\cdot\|_{\mathrm{F}}$ (highlighted in red) is replaced by $\|\cdot\|_1$.

Analyzing phase retrieval via RIP-ℓ_2/ℓ_1

Theorem 10.2 (Chen, Chi, Goldsmith '15)

Suppose $\mathrm{rank}(\boldsymbol{M}) = r$. *For any fixed integer* $K > 0$, *if* $\frac{1+\delta_{Kr}^{\mathrm{ub}}}{1-\delta_{(2+K)r}^{\mathrm{lb}}} < \sqrt{\frac{K}{2}}$, *then nuclear norm minimization is exact.*

- Back to the example in Slide 9:
  - If $\boldsymbol{x}$ is independent of $\{\boldsymbol{a}_i\}$, then

$$\left\langle \boldsymbol{a}_i \boldsymbol{a}_i^\top, \boldsymbol{x}\boldsymbol{x}^\top \right\rangle = \left|\boldsymbol{a}_i^\top \boldsymbol{x}\right|^2 \asymp \|\boldsymbol{x}\|^2 \quad \Rightarrow \quad \left\|\mathcal{A}\big(\boldsymbol{x}\boldsymbol{x}^\top\big)\right\|_1 \asymp m\|\boldsymbol{x}\boldsymbol{x}^\top\|_{\mathrm{F}}$$

  - $\|\mathcal{A}(\boldsymbol{A}_i)\|_1 = |\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{A}_i\rangle| + \sum_{j:j\neq i} |\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{A}_j\rangle| \approx (n+m)\|\boldsymbol{A}_i\|_{\mathrm{F}}$
  - For both cases, $\frac{\|\mathcal{A}(\boldsymbol{X})\|_1}{\|\boldsymbol{X}\|_{\mathrm{F}}}$ are of same order

Analyzing phase retrieval via RIP-ℓ_2/ℓ_1

A debiased operator satisfies RIP condition of Theorem 10.2 when $m \gtrsim nr$

$$\mathcal{B}(\boldsymbol{X}) := \begin{bmatrix} \langle \boldsymbol{A}_1 - \boldsymbol{A}_2, \boldsymbol{X} \rangle \\ \langle \boldsymbol{A}_3 - \boldsymbol{A}_4, \boldsymbol{X} \rangle \\ \vdots \end{bmatrix} \in \mathbb{R}^{m/2}$$

- Debiasing is crucial when $r \gg 1$
- A consequence of Hanson-Wright inequality for quadratic form (Hanson & Wright '71, Rudelson & Vershynin '03)

Theoretical guarantee for phase retrieval

$$\text{(PhaseLift)} \qquad \underset{\boldsymbol{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \underbrace{\mathrm{Tr}(\boldsymbol{X})}_{\|\cdot\|_* \text{ for PSD matrices}}$$

$$\text{s.t.} \quad y_i = \boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{a}_i, \quad 1 \le i \le m$$

$$\boldsymbol{X} \succeq \boldsymbol{0} \quad (\text{since } \boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top)$$

Theorem 10.3 (Candès et al. '13, Candès and Li '14)

Suppose $\boldsymbol{a}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. *With high prob., PhaseLift recovers* $\boldsymbol{x}\boldsymbol{x}^\top$ *exactly as soon as* $m \gtrsim n$.

Extension of phase retrieval to low-rank setting

Measurements:

$$y_i = \langle \boldsymbol{a}_i \boldsymbol{a}_i^\top, \boldsymbol{M} \rangle := \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle \qquad 1 \leq i \leq m$$

where $\boldsymbol{M} \succeq \boldsymbol{0}$ and rank$(\boldsymbol{M}) = r$.

Extension of phase retrieval to low-rank setting

Measurements:

$$y_i = \langle \boldsymbol{a}_i \boldsymbol{a}_i^\top, \boldsymbol{M} \rangle := \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle \qquad 1 \le i \le m$$

where $\boldsymbol{M} \succeq \boldsymbol{0}$ and $\mathrm{rank}(\boldsymbol{M}) = r$.

$$(\textbf{PhaseLift}) \qquad \underset{\boldsymbol{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \underbrace{\mathrm{Tr}(\boldsymbol{X})}_{\|\cdot\|_* \text{ for PSD matrices}}$$

$$\text{s.t.} \quad \boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{a}_i = \boldsymbol{a}_i^\top \boldsymbol{M} \boldsymbol{a}_i, \quad 1 \le i \le m$$

$$\boldsymbol{X} \succeq \boldsymbol{0}$$

Theorem 10.4 (Chen, Chi, Goldsmith '15, Cai, Zhang '15, Kueng, Rauhut, Terstiege '17)

Suppose $\boldsymbol{a}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. *With high prob., PhaseLift recovers* $\boldsymbol{M}$ *exactly as soon as* $m \gtrsim nr$.

Nonconvex Wirtinger flow

A natural least squares formulation

What nonconvex?

$$\text{given:} \qquad y_k \;=\; |\boldsymbol{a}_k^\top \boldsymbol{x}^\natural|^2, \quad 1 \leq k \leq m$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\Big[\big(\boldsymbol{a}_k^\top \boldsymbol{x}\big)^2 - y_k\Big]^2$$

A natural least squares formulation

What nonconvex?

$$\text{given:} \qquad y_k \;=\; |\boldsymbol{a}_k^\top \boldsymbol{x}^\natural|^2, \quad 1 \le k \le m$$

$$\Downarrow$$

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

- **pros:** often exact as long as sample size is sufficiently large

A natural least squares formulation

What nonconvex?

$$\text{given:} \qquad y_k \;=\; |\boldsymbol{a}_k^\top \boldsymbol{x}^\natural|^2, \quad 1 \le k \le m$$

$$\Downarrow$$

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

- **pros:** often exact as long as sample size is sufficiently large
- **cons:** $f(\cdot)$ is highly nonconvex

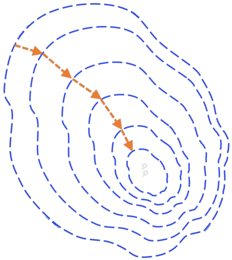
  $\longrightarrow$ *computationally challenging!*

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[\left(\boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

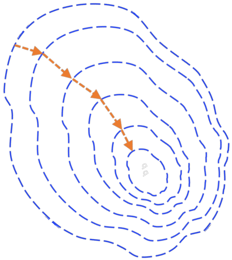
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



- **spectral initialization:** $\boldsymbol{x}^0 \leftarrow$ leading eigenvector of certain data matrix

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

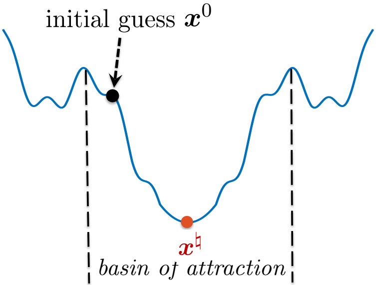
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



- **spectral initialization:** $\boldsymbol{x}^0 \leftarrow$ leading eigenvector of certain data matrix
- **gradient descent:**

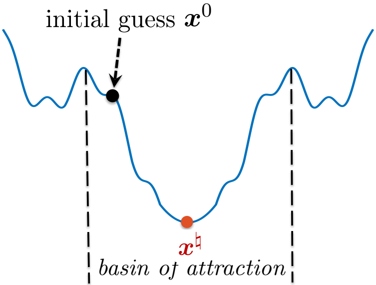
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta\,\nabla f(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

Rationale of two-stage approach



1. find an initial point within a local basin sufficiently close to $\boldsymbol{x}^\natural$

Rationale of two-stage approach



1. find an initial point within a local basin sufficiently close to $\boldsymbol{x}^\natural$
2. careful iterative refinement without leaving this local basin

Initialization via spectral method

$\boldsymbol{x}^0 \leftarrow$ leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{k=1}^{m} y_k\, \boldsymbol{a}_k \boldsymbol{a}_k^\top$$

- Intuition:

$$\mathbb{E}\left[\boldsymbol{Y}\right] = \mathbb{E}[(\boldsymbol{a}_k^\top \boldsymbol{x})^2 \boldsymbol{a}_k \boldsymbol{a}_k^\top] = \boldsymbol{I} + 2\boldsymbol{x}^\natural \boldsymbol{x}^{\natural\top}.$$

Computational cost

$$\boldsymbol{A}\boldsymbol{x} := \left[\boldsymbol{a}_k^\top \boldsymbol{x}\right]_{1\le k\le m}$$

- **Spectral initialization:** leading eigenvector $\rightarrow$ a few applications of $\boldsymbol{A}$ and $\boldsymbol{A}^\top$

$$\frac{1}{m}\sum_{k=1}^{m} y_k\, \boldsymbol{a}_k \boldsymbol{a}_k^\top = \frac{1}{m}\boldsymbol{A}^\top \,\mathrm{diag}\{y_k\}\, \boldsymbol{A}$$

Computational cost

$$\boldsymbol{A}\boldsymbol{x} := \left[\boldsymbol{a}_k^\top \boldsymbol{x}\right]_{1\le k\le m}$$

- **Spectral initialization:** leading eigenvector $\rightarrow$ a few applications of $\boldsymbol{A}$ and $\boldsymbol{A}^\top$

$$\frac{1}{m}\sum_{k=1}^{m} y_k\,\boldsymbol{a}_k\boldsymbol{a}_k^\top = \frac{1}{m}\boldsymbol{A}^\top\,\mathrm{diag}\{y_k\}\,\boldsymbol{A}$$

- **Iterations:** one application of $\boldsymbol{A}$ and $\boldsymbol{A}^\top$ per iteration

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

Performance guarantees of WF

First theory:

Theorem 10.5 (Candès, Li, Soltanolkotabi '14)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{4}\right)^{t/2} \|\boldsymbol{x}^\natural\|_2,$$

with high prob., provided that step size $\eta \lesssim 1/n$ *and sample size :* $m \gtrsim n \log n$

- Iteration complexity: $O\big(n \log \frac{1}{\epsilon}\big)$
- Sample complexity: $O(n \log n)$

Performance guarantees of WF

Improved theory:

Theorem 10.6 (Ma, Wang, Chi, Chen '17)

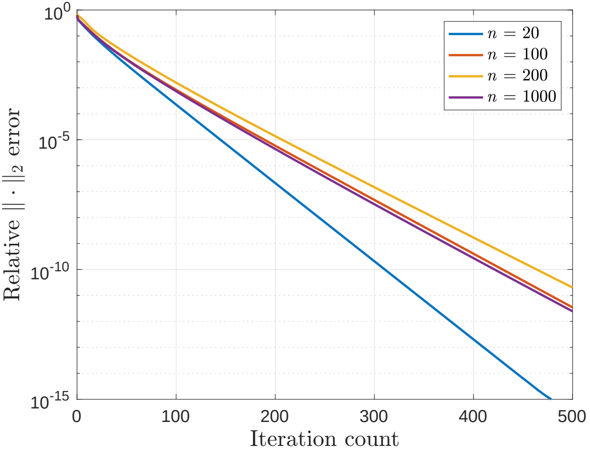
Under i.i.d. Gaussian design, WF with spectral initialization achieves

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$$

with high prob., provided that step size $\eta \asymp 1/\log n$ *and sample size* $m \gtrsim n \log n$.

- Iteration complexity: $O\big(n \log \frac{1}{\epsilon}\big) \searrow O\big(\log n \log \frac{1}{\epsilon}\big)$
- Sample complexity: $O(n \log n)$

Numerical surprise with $\eta_t = 0.1$

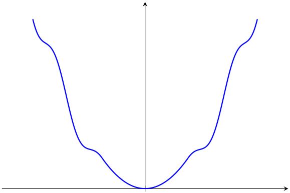


Vanilla GD (WF) converges fast!

Gradient descent theory revisited

Consider unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$

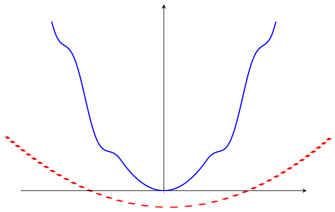


Two standard conditions that enable geometric convergence of GD

Gradient descent theory revisited

Consider unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$



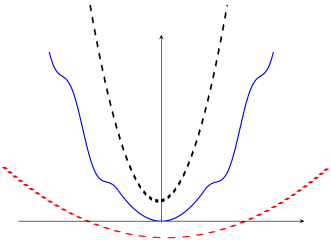
Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity (or regularity condition)

Gradient descent theory revisited

Consider unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x})$$



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity (or regularity condition)
- (local) smoothness

$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad \text{and} \quad \text{is well-conditioned}$$

Gradient descent theory revisited

f is said to be α-strongly convex and β-smooth if

$$\mathbf{0} \; \preceq \; \alpha \boldsymbol{I} \; \preceq \; \nabla^2 f(\boldsymbol{x}) \; \preceq \; \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

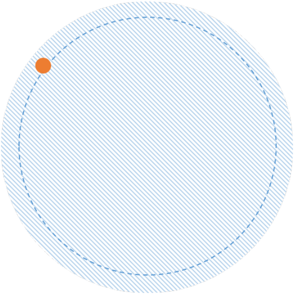
ℓ_2 **error contraction:** GD with $\eta = 1/\beta$ obeys

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

Gradient descent theory revisited

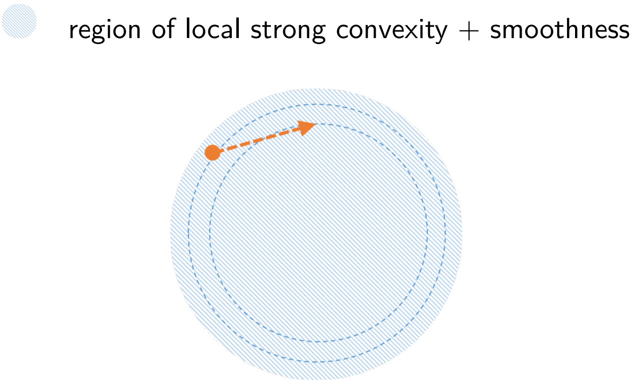
$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{\natural}\|_2 \leq (1 - \alpha/\beta)\, \|\boldsymbol{x}^{t} - \boldsymbol{x}^{\natural}\|_2$$

 region of local strong convexity + smoothness



Gradient descent theory revisited

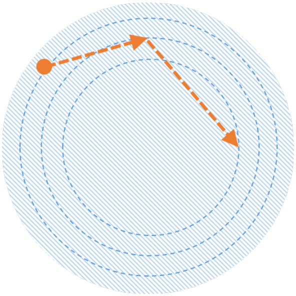
$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{\natural}\|_2 \leq (1 - \alpha/\beta)\, \|\boldsymbol{x}^{t} - \boldsymbol{x}^{\natural}\|_2$$



Gradient descent theory revisited

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{\natural}\|_2 \leq (1 - \alpha/\beta)\, \|\boldsymbol{x}^{t} - \boldsymbol{x}^{\natural}\|_2$$

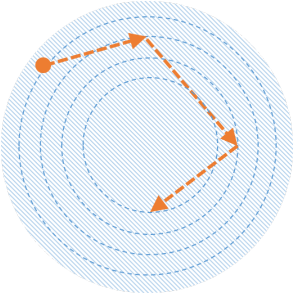
region of local strong convexity + smoothness



Gradient descent theory revisited

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\natural\|_2 \leq (1 - \alpha/\beta)\,\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2$$

region of local strong convexity + smoothness



Gradient descent theory revisited

$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x})$) with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\natural\|_2 \le \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2$$

- Condition number β/α determines rate of convergence

Gradient descent theory revisited

$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x})$) with $\eta = 1/\beta$ obeys

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

- Condition number β/α determines rate of convergence
- Attains ε-accuracy within $O(\frac{\beta}{\alpha} \log \frac{1}{\varepsilon})$ iterations

What does this optimization theory say about WF?

Gaussian designs: $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$

What does this optimization theory say about WF?

$$\textit{Gaussian designs: } \boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$$

Population level (infinite samples)

$$\mathbb{E}[\nabla^2 f(\boldsymbol{x})] = \underbrace{3\left(\|\boldsymbol{x}\|_2^2\,\boldsymbol{I} + 2\boldsymbol{x}\boldsymbol{x}^\top\right) - \left(\|\boldsymbol{x}^\natural\|_2^2\boldsymbol{I} + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)}_{\textit{locally}\text{ positive definite and well-conditioned}}$$

$$\boldsymbol{I}_n \preceq \mathbb{E}[\nabla^2 f(\boldsymbol{x})] \preceq 10\boldsymbol{I}_n \quad (\|\boldsymbol{x}^\natural\| = 1)$$

Consequence: Given good initialization, WF converges within $O\big(\log\frac{1}{\varepsilon}\big)$ iterations if sample size $m \to \infty$

What does this optimization theory say about WF?

Gaussian designs: $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$

Finite-sample level ($m \asymp n \log n$**)**

$$\nabla^2 f(\boldsymbol{x}) \ \succ \boldsymbol{0}$$

What does this optimization theory say about WF?

$$\textit{Gaussian designs: } \boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$$

Finite-sample level ($m \asymp n \log n$)

$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp n} \text{ (even locally)}$$

$$\frac{1}{2}\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)\boldsymbol{I}_n$$

What does this optimization theory say about WF?

$$\textit{Gaussian designs: } \boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$$

Finite-sample level ($m \asymp n \log n$)

$$\nabla^2 f(\boldsymbol{x}) \succ \mathbf{0} \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp n} \text{ (even locally)}$$

$$\frac{1}{2}\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)\boldsymbol{I}_n$$

Consequence (Candès et al '14): WF attains ε-accuracy within $O\left(n \log \frac{1}{\varepsilon}\right)$ iterations if $m \asymp n \log n$

A peek into the Hessian

The Hessian satisfies:

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m}\left[3\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_k^\top \boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top$$

$$= \underbrace{\frac{3}{m}\sum_{j=1}^{m}\left[\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top}_{:=\boldsymbol{\Lambda}_1}$$

$$+ \underbrace{\frac{2}{m}\sum_{j=1}^{m}\left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2\boldsymbol{a}_j\boldsymbol{a}_j^\top - 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)}_{:=\boldsymbol{\Lambda}_2} + \underbrace{2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)}_{:=\boldsymbol{\Lambda}_3},$$

Detour: some basic facts

Assume $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$ for every $1 \leq j \leq m$.

- With probability at least $1 - O(me^{-1.5n})$, $\{\boldsymbol{a}_j\}$ obey

$$\max_{1 \leq j \leq m} \|\boldsymbol{a}_j\|_2 \leq \sqrt{6n}$$

- With probability exceeding $1 - O(mn^{-10})$,

$$\max_{1 \leq j \leq m} \left|\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right| \leq 5\sqrt{\log n}$$

- Fix any small constant $\delta > 0$. With probability at least $1 - C_2 e^{-c_2 m}$, one has

$$\left\|\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{a}_j^\top - \boldsymbol{I}_n\right\| \leq \delta,$$

  as long as $m \geq c_0 n$ for some sufficiently large constant $c_0 > 0$.

Smoothness of Hessian

$$\boldsymbol{\Lambda}_2 = \frac{2}{m}\sum_{j=1}^{m}\left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2 \boldsymbol{a}_j \boldsymbol{a}_j^\top - 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural \boldsymbol{x}^{\natural\top}\right)$$

$$\boldsymbol{\Lambda}_3 = 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural \boldsymbol{x}^{\natural\top}\right)$$

Smoothness of Hessian

$$\boldsymbol{\Lambda}_2 = \frac{2}{m}\sum_{j=1}^{m}\left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2 \boldsymbol{a}_j\boldsymbol{a}_j^\top - 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)$$

$$\boldsymbol{\Lambda}_3 = 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)$$

- $\boldsymbol{\Lambda}_3$ is well-controlled:

$$\|\boldsymbol{\Lambda}_3\| \le 2\left(\|\boldsymbol{I}_n\| + 2\|\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\|\right) = 6$$

Smoothness of Hessian

$$\mathbf{\Lambda}_2 = \frac{2}{m}\sum_{j=1}^{m}\left(\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right)^2 \boldsymbol{a}_j\boldsymbol{a}_j^\top - 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)$$

$$\mathbf{\Lambda}_3 = 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)$$

- $\mathbf{\Lambda}_3$ is well-controlled:

$$\|\mathbf{\Lambda}_3\| \le 2\left(\|\boldsymbol{I}_n\| + 2\|\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\|\right) = 6$$

- When $n = O(n\log n)$, $\mathbf{\Lambda}_2$ is well-controlled:

$$\|\mathbf{\Lambda}_2\| \le 2\delta.$$

for arbitrary small δ for a fixed $\boldsymbol{x}^\natural$.

A peek into the smoothness of Hessian

The term $\mathbf{\Lambda}_1$ is problematic:

$$\|\mathbf{\Lambda}_1\| \leq \left\| \frac{3}{m} \sum_{j=1}^{m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural \right) \right| \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\|.$$

A peek into the smoothness of Hessian

The term $\boldsymbol{\Lambda}_1$ is problematic:

$$\|\boldsymbol{\Lambda}_1\| \le \left\| \frac{3}{m} \sum_{j=1}^{m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural\right)\right| \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural\right)\right| \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\|.$$

- In the local neighborhood $\|\boldsymbol{x} - \boldsymbol{x}^\natural\| \le \frac{1}{10}\|\boldsymbol{x}^\natural\| = \frac{1}{10}$, we have

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural\right)\right| \lesssim {\color{red}\sqrt{n}} \quad \text{by Cauchy-Schwartz}$$

$$\begin{aligned} \max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural\right)\right| &\le 2 \max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right| + \max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural\right)\right| \\ &\lesssim \sqrt{\log n} + \sqrt{n} \asymp \sqrt{n} \end{aligned}$$

(think when $\boldsymbol{x}$ is aligned with $\boldsymbol{a}_j$)

A peek into the smoothness of Hessian

The term $\boldsymbol{\Lambda}_1$ is problematic:

$$\|\boldsymbol{\Lambda}_1\| \leq \left\| \frac{3}{m} \sum_{j=1}^{m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural \right) \right| \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\|.$$

- In the local neighborhood $\|\boldsymbol{x} - \boldsymbol{x}^\natural\| \leq \frac{1}{10} \|\boldsymbol{x}^\natural\| = \frac{1}{10}$, we have

$$\max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \lesssim \sqrt{n} \quad \text{by Cauchy-Schwartz}$$

$$\begin{aligned} \max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural \right) \right| &\leq 2 \max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \boldsymbol{x}^\natural \right| + \max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \\ &\lesssim \sqrt{\log n} + \sqrt{n} \asymp \sqrt{n} \end{aligned}$$

(think when $\boldsymbol{x}$ is aligned with $\boldsymbol{a}_j$)

$\Longrightarrow$

$$\|\boldsymbol{\Lambda}_1\| \lesssim n \cdot \left\| \frac{1}{m} \sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\| \asymp n,$$

A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m}\left[3\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_k^\top \boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_k\boldsymbol{a}_k^\top$$

A second look at gradient descent theory

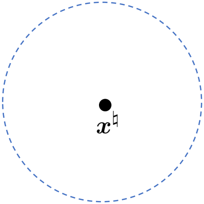
Which region enjoys both strong convexity and smoothness?

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m}\left[3\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_k^\top \boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_k\boldsymbol{a}_k^\top$$

- Not smooth if $\boldsymbol{x}$ and $\boldsymbol{a}_k$ are too close (coherent)

A second look at gradient descent theory

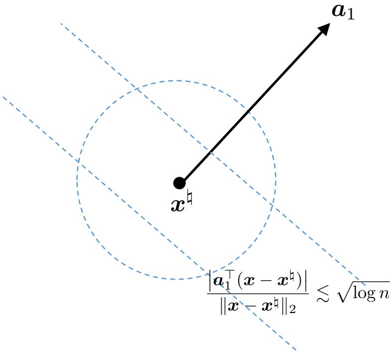
Which region enjoys both strong convexity and smoothness?



- x is not far away from $x^\natural$

A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

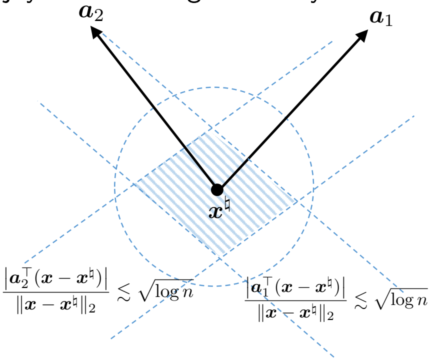


- $\boldsymbol{x}$ is not far away from $\boldsymbol{x}^\natural$
- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors (incoherence region)

$$(1/2)\cdot \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log n)\cdot \boldsymbol{I}_n$$

A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ is not far away from $\boldsymbol{x}^\natural$
- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors (incoherence region)

$$(1/2) \cdot \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log n) \cdot \boldsymbol{I}_n$$

Re-examine the Hessian in incoherence region

The term $\boldsymbol{\Lambda}_1$ is okay now:

$$\|\boldsymbol{\Lambda}_1\| \leq \left\| \frac{3}{m} \sum_{j=1}^{m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural \right) \right| \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\| .$$

Re-examine the Hessian in incoherence region

The term $\mathbf{\Lambda}_1$ is okay now:

$$\|\mathbf{\Lambda}_1\| \leq \left\| \frac{3}{m} \sum_{j=1}^{m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural \right) \right| \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\|.$$

- In the local neighborhood and incoherence region, we have

$$\max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \lesssim \sqrt{\log n} \quad \text{by Cauchy-Schwartz}$$

$$\begin{aligned} \max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural \right) \right| &\leq 2 \max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \boldsymbol{x}^\natural \right| + \max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural \right) \right| \\ &\lesssim \sqrt{\log n} + \sqrt{\log n} \asymp \sqrt{\log n} \end{aligned}$$

Re-examine the Hessian in incoherence region

The term $\mathbf{\Lambda}_1$ is okay now:

$$\|\mathbf{\Lambda}_1\| \leq \left\| \frac{3}{m} \sum_{j=1}^{m} \left|\boldsymbol{a}_j^\top (\boldsymbol{x} - \boldsymbol{x}^\natural)\right| \left|\boldsymbol{a}_j^\top (\boldsymbol{x} + \boldsymbol{x}^\natural)\right| \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\|.$$

- In the local neighborhood and incoherence region, we have

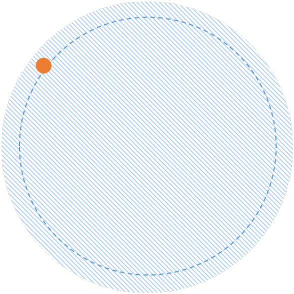
$$\begin{aligned}
\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural\right)\right| &\lesssim \sqrt{\log n} \quad \text{by Cauchy-Schwartz} \\
\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} + \boldsymbol{x}^\natural\right)\right| &\leq 2 \max_{1\leq j\leq m} \left|\boldsymbol{a}_j^\top \boldsymbol{x}^\natural\right| + \max_{1\leq j\leq m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x} - \boldsymbol{x}^\natural\right)\right| \\
&\lesssim \sqrt{\log n} + \sqrt{\log n} \asymp \sqrt{\log n}
\end{aligned}$$

$\Longrightarrow$

$$\|\mathbf{\Lambda}_1\| \lesssim \log n \cdot \left\| \frac{1}{m} \sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{a}_j^\top \right\| \asymp \log n,$$

A second look at gradient descent theory

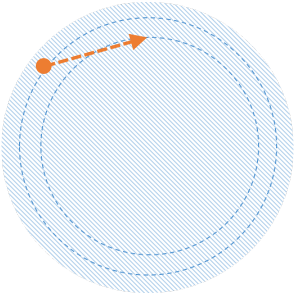
region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

A second look at gradient descent theory

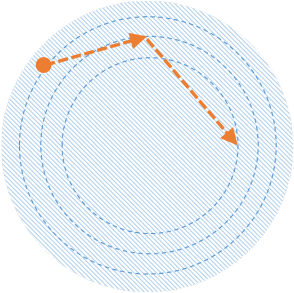
region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

A second look at gradient descent theory

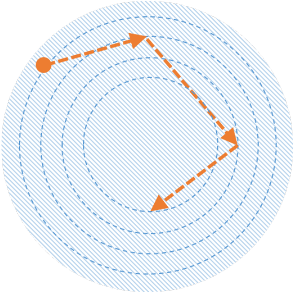
region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

A second look at gradient descent theory

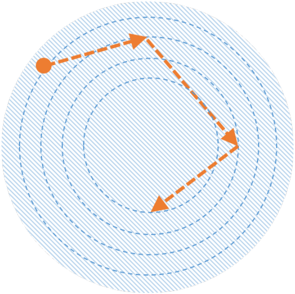
region of local strong convexity + smoothness

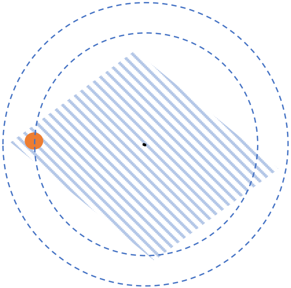


- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

A second look at gradient descent theory

 region of local strong convexity + smoothness

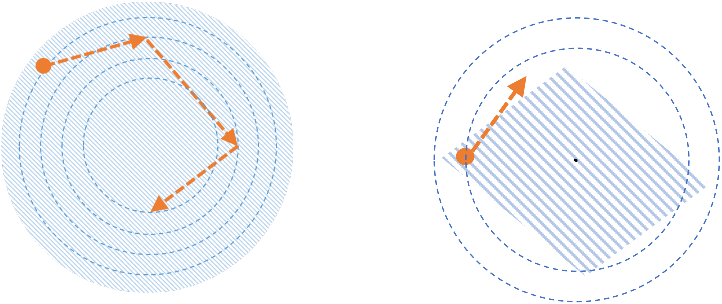




- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

A second look at gradient descent theory

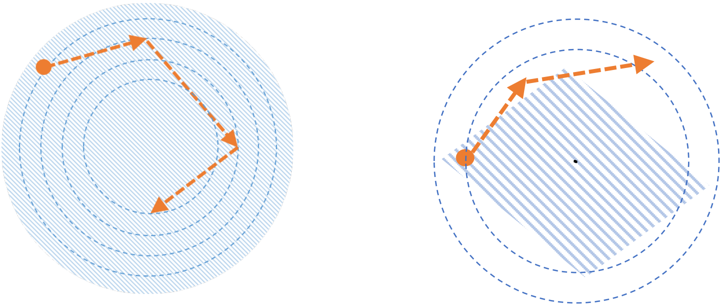
 region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

A second look at gradient descent theory

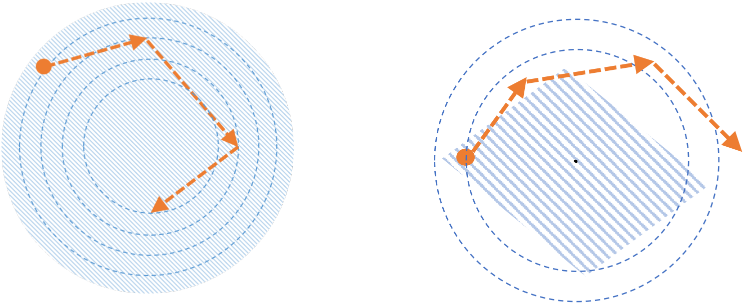
 region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

A second look at gradient descent theory

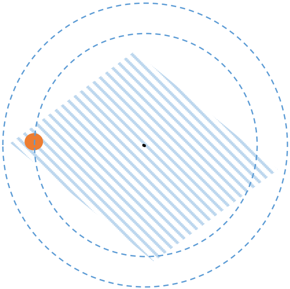
region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

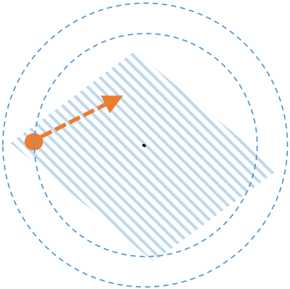
Surprising message: GD is implicitly regularized

region of local strong convexity + smoothness



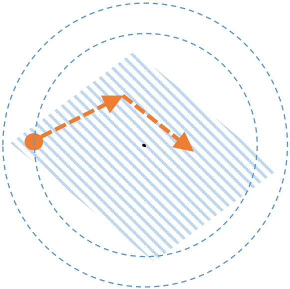
Surprising message: GD is implicitly regularized

region of local strong convexity + smoothness



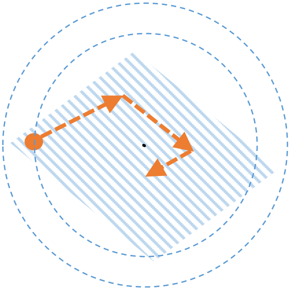
Surprising message: GD is implicitly regularized

region of local strong convexity + smoothness

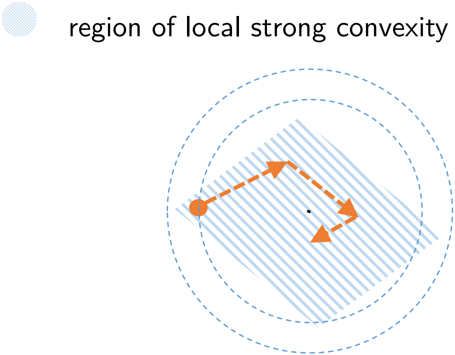


Surprising message: GD is implicitly regularized

region of local strong convexity + smoothness

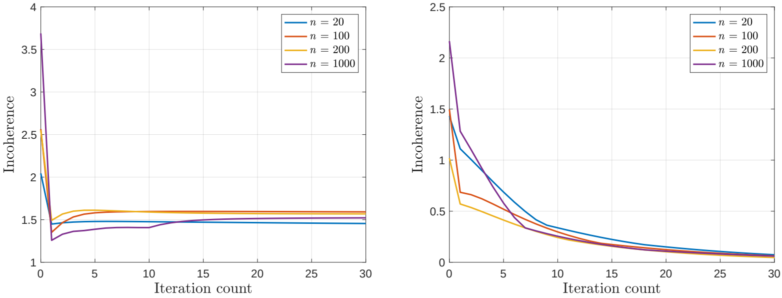


Surprising message: GD is implicitly regularized



GD implicitly forces iterates to remain incoherent

Implicit Regularization



(a) $\dfrac{\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top \boldsymbol{x}^t\right|}{\sqrt{\log n}\left\|\boldsymbol{x}^\natural\right\|_2}$ (b) $\dfrac{\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top (\boldsymbol{x}^t-\boldsymbol{x}^\natural)\right|}{\sqrt{\log n}\left\|\boldsymbol{x}^\natural\right\|_2}$

Figure 10.1: The incoherence measure vs. iteration count. The results are shown for $n \in \{20, 100, 200, 1000\}$ and $m = 10n$, with the step size taken to be $\eta_t = 0.1$.

Theoretical guarantees

Theorem 10.7 (Ma, Wang, Chi, Chen '17)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

- $\max_k \left|\boldsymbol{a}_k^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\natural\|_2$ *(incoherence)*

Theoretical guarantees

Theorem 10.7 (Ma, Wang, Chi, Chen '17)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

- $\max_k \left|\boldsymbol{a}_k^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\natural\|_2$ *(incoherence)*
- $\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$ *(linear convergence)*

provided that step size $\eta \asymp 1/\log n$ *and sample size* $m \gtrsim n \log n$.

Key ingredient: leave-one-out analysis

How to establish $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\,\|\boldsymbol{x}^\natural\|_2$?

Key ingredient: leave-one-out analysis

How to establish $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\,\|\boldsymbol{x}^\natural\|_2$?

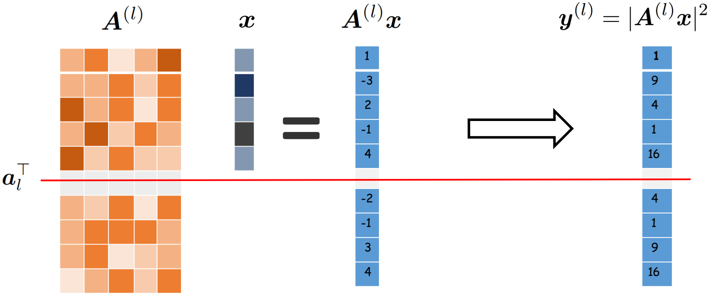
Technical difficulty: $\boldsymbol{x}^t$ is statistically dependent with $\{\boldsymbol{a}_l\}$;

Key ingredient: leave-one-out analysis

How to establish $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\,\|\boldsymbol{x}^\natural\|_2$?

Technical difficulty: $\boldsymbol{x}^t$ is statistically dependent with $\{\boldsymbol{a}_l\}$;

Leave-one-out trick: For each $1 \le l \le m$, introduce leave-one-out iterates $\boldsymbol{x}^{t,(l)}$ by dropping lth sample



Leave-one-out trick

- For each $1 \le l \le m$, we define the leave-one-out empirical loss function as

$$f^{(l)}(\boldsymbol{x}) := \frac{1}{4m} \sum_{j:j \neq l} \left[\left(\boldsymbol{a}_j^\top \boldsymbol{x} \right)^2 - y_j \right]^2,$$

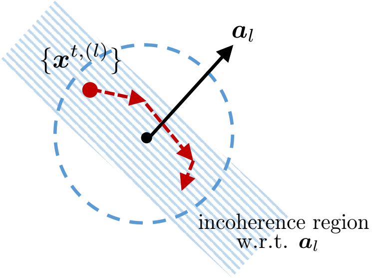
  and the auxiliary trajectory $\left\{ \boldsymbol{x}^{t,(l)} \right\}_{t \ge 0}$ is constructed by running WF w.r.t. $f^{(l)}(\boldsymbol{x})$.

- The initialization $\boldsymbol{x}^{0,(l)}$ is computed based on

$$\boldsymbol{Y}^{(l)} := \frac{1}{m} \sum_{j:j \neq l} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top .$$

- Clearly, the entire sequence $\left\{ \boldsymbol{x}^{t,(l)} \right\}_{t \ge 0}$ is independent of the lth sampling vector $\boldsymbol{a}_l$.

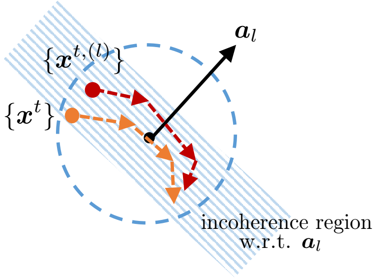
Key ingredient: leave-one-out analysis



- Step 1: Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.

$$\max_{1 \le l \le m} \left| \boldsymbol{a}_l^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\natural \right) \right| \lesssim \sqrt{\log n}.$$

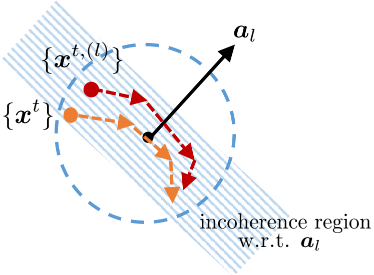
Key ingredient: leave-one-out analysis



- Step 1: Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.

$$\max_{1\le l\le m} \left|\boldsymbol{a}_l^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\natural\right)\right| \lesssim \sqrt{\log n}.$$

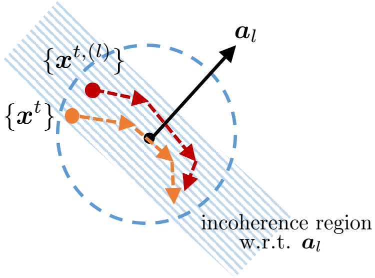
Key ingredient: leave-one-out analysis



- Step 2: Leave-one-out iterates $\boldsymbol{x}^{t,(l)} \approx$ true iterates $\boldsymbol{x}^t$

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

Key ingredient: leave-one-out analysis



- Step 3: Finish by triangle inequality

$$\begin{aligned}
\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| &\le \left|\boldsymbol{a}_l^\top(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\natural)\right| + \left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)})\right| \\
&\le \left|\boldsymbol{a}_l^\top(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\natural)\right| + \|\boldsymbol{a}_l\| \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\| \\
&\lesssim \sqrt{\log n} + \sqrt{n}\sqrt{\frac{\log n}{n}} \asymp \sqrt{\log n}.
\end{aligned}$$

Proximity of leave-one-out iterates

$$
\begin{aligned}
&\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)} \\
&= \boldsymbol{x}^{t} - \eta \nabla f\left(\boldsymbol{x}^{t}\right) - \left[\boldsymbol{x}^{t,(l)} - \eta \nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)\right] \\
&= \boldsymbol{x}^{t} - \eta \nabla f\left(\boldsymbol{x}^{t}\right) - \left[\boldsymbol{x}^{t,(l)} - \eta \nabla f\big(\boldsymbol{x}^{t,(l)}\big)\right] - \eta\left[\nabla f\big(\boldsymbol{x}^{t,(l)}\big) - \nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)\right] \\
&= \underbrace{\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} - \eta\left[\nabla f\left(\boldsymbol{x}^{t}\right) - \nabla f\big(\boldsymbol{x}^{t,(l)}\big)\right]}_{:=\boldsymbol{\nu}_1^{(l)}} - \underbrace{\frac{\eta}{m}\left[\big(\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}\big)^2 - \big(\boldsymbol{a}_l^\top \boldsymbol{x}^{\natural}\big)^2\right](\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)})\boldsymbol{a}_l}_{:=\boldsymbol{\nu}_2^{(l)}},
\end{aligned}
$$

- By incoherence:

$$
\begin{aligned}
\|\boldsymbol{\nu}_2^{(l)}\|_2 &\leq \eta \frac{\|\boldsymbol{a}_l\|_2}{m}\left|\big(\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}\big)^2 - \big(\boldsymbol{a}_l^\top \boldsymbol{x}^{\natural}\big)^2\right|\left|\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}\right| \\
&\lesssim \eta \frac{\sqrt{n\log n}}{m}\log n \lesssim \eta\sqrt{\frac{\log n}{n}}
\end{aligned}
$$

where the last line follows from $m \gtrsim n\log n$.

Proximity of leave-one-out iterates

$$
\begin{aligned}
&\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)} \\
&= \boldsymbol{x}^{t} - \eta \nabla f\left(\boldsymbol{x}^{t}\right) - \left[\boldsymbol{x}^{t,(l)} - \eta \nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)\right] \\
&= \boldsymbol{x}^{t} - \eta \nabla f\left(\boldsymbol{x}^{t}\right) - \left[\boldsymbol{x}^{t,(l)} - \eta \nabla f\big(\boldsymbol{x}^{t,(l)}\big)\right] - \eta\left[\nabla f\big(\boldsymbol{x}^{t,(l)}\big) - \nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)\right] \\
&= \underbrace{\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} - \eta\left[\nabla f\left(\boldsymbol{x}^{t}\right) - \nabla f\big(\boldsymbol{x}^{t,(l)}\big)\right]}_{:=\boldsymbol{\nu}_1^{(l)}} - \underbrace{\frac{\eta}{m}\left[\big(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)}\big)^2 - \big(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{\natural}\big)^2\right](\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)})\boldsymbol{a}_l}_{:=\boldsymbol{\nu}_2^{(l)}},
\end{aligned}
$$

- By fundamental theorem of calculus:

$$
\boldsymbol{\nu}_1^{(l)} = \left[\boldsymbol{I}_n - \eta \int_0^1 \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) \mathrm{d}\tau\right]\big(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\big),
$$

where $\boldsymbol{x}\left(\tau\right) = \boldsymbol{x}^{t,(l)} + \tau(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)})$. As long as $\eta \asymp 1/\log n$ is small enough,

$$
\big\|\boldsymbol{\nu}_1^{(l)}\big\|_2 \le \left(1 - \eta/2\right)\big\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\big\|_2.
$$

Proximity of leave-one-out iterates

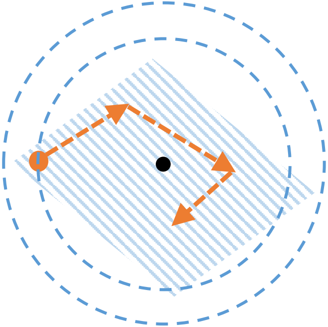
$$
\begin{aligned}
&\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)} \\
&= \boldsymbol{x}^{t} - \eta\nabla f\left(\boldsymbol{x}^{t}\right) - \left[\boldsymbol{x}^{t,(l)} - \eta\nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)\right] \\
&= \boldsymbol{x}^{t} - \eta\nabla f\left(\boldsymbol{x}^{t}\right) - \left[\boldsymbol{x}^{t,(l)} - \eta\nabla f\big(\boldsymbol{x}^{t,(l)}\big)\right] - \eta\left[\nabla f\big(\boldsymbol{x}^{t,(l)}\big) - \nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)\right] \\
&= \underbrace{\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} - \eta\left[\nabla f\left(\boldsymbol{x}^{t}\right) - \nabla f\big(\boldsymbol{x}^{t,(l)}\big)\right]}_{:=\boldsymbol{\nu}_1^{(l)}} \underbrace{- \frac{\eta}{m}\left[\big(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)}\big)^2 - \big(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{\natural}\big)^2\right](\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)})\boldsymbol{a}_l}_{:=\boldsymbol{\nu}_2^{(l)}},
\end{aligned}
$$

- Putting things together:

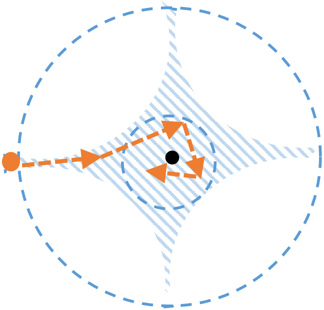
$$
\begin{aligned}
\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\right\|_2 &\leq (1-\eta/2)\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_2 + c\eta\sqrt{\frac{\log n}{n}} \\
&\lesssim \sqrt{\frac{\log n}{n}}
\end{aligned}
$$

by induction.

Incoherence region in high dimensions



2-dimensional



$\underbrace{\text{high-dimensional (mental representation)}}_{\color{blue}\text{incoherence region is vanishingly small}}$

Reference

[1] "*Phase retrieval via Wirtinger flow: Theory and algorithms*," E. Candes, X. Li, M. Soltanolkotabi, *IEEE Transactions on Information Theory*, 2015.

[2] "*Solving random quadratic systems of equations is nearly as easy as solving linear systems*," Y. Chen, E. Candes, Communications on Pure and Applied Mathematics, 2017.

[3] "*Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion and Blind Deconvolution*," C. Ma, K. Wang, Y. Chi and Y. Chen, *arXiv preprint arXiv:1711.10467*, 2017.