Refined analysis of local convergence: implicit regularization

Cong Ma

University of Chicago, Winter 2024

Revisit phase retrieval

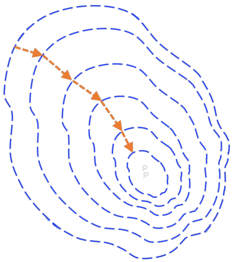
$$\text{given:} \qquad y_k \;=\; (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \quad 1 \le k \le m$$

$$\Downarrow$$

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

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
- **gradient descent:**

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

First theory of WF

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\star\|_2\}$$

Theorem 9.1 (Candès, Li, Soltanolkotabi '14)

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

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

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

- Iteration complexity: $O(n \log \frac{1}{\epsilon})$
- Sample complexity: $O(n \log n)$
- Derived based on (worst-case) local geometry

What does optimization theory say about WF?

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

What does optimization theory say about WF?

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

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

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

What does optimization theory say about WF?

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

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

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

What does 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 ($m \asymp n \log n$)

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

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

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

$\Uparrow$

Step size taken to be $\eta = O(1/n)$

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

$$\Uparrow$$

Step size taken to be $\eta = O(1/n)$

$$\Uparrow$$

This choice is suggested by worst-case optimization theory

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

$\Uparrow$

Step size taken to be $\eta = O(1/n)$

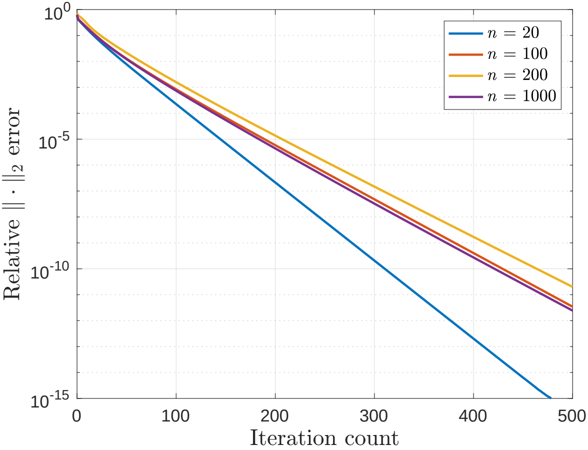
$\Uparrow$

This choice is suggested by worst-case optimization theory

$\Uparrow$

Does it capture what really happens?

Numerical efficiency with $\eta_t = 0.1$



Vanilla GD (WF) converges fast for a constant step size!

Improved theory of WF

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\star\|_2\}$$

Theorem 9.2 (Ma, Wang, Chi, Chen '17)

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

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

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

- Iteration complexity: $O\big(n \log \frac{1}{\epsilon}\big) \searrow \quad O\big(\log n \log \frac{1}{\epsilon}\big)$
- Sample complexity: $O(n \log n)$
- Derived based on finer analysis of GD trajectory

A second look at gradient descent theory

Which local region enjoys both strong convexity and smoothness?

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

A second look at gradient descent theory

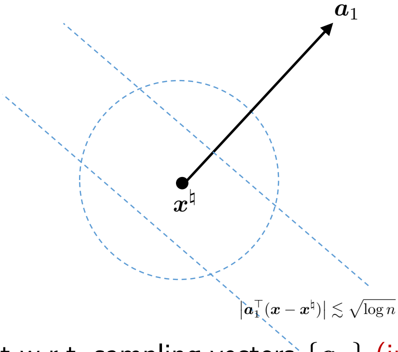
Which local region enjoys both strong convexity and smoothness?

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

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

A second look at gradient descent theory

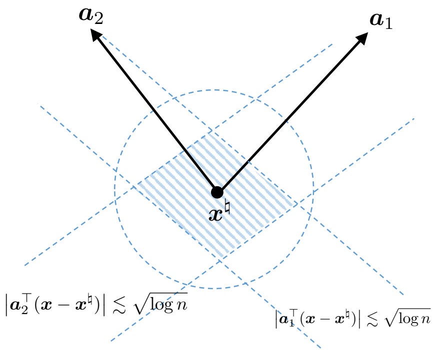
Which local region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

A second look at gradient descent theory

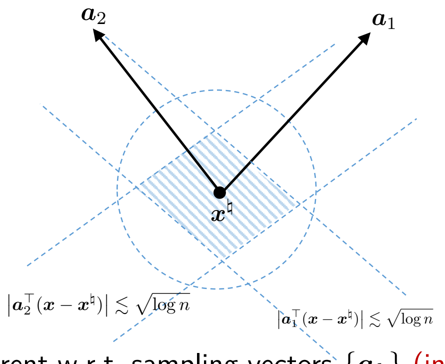
Which local region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

A second look at gradient descent theory

Which local region enjoys both strong convexity and smoothness?

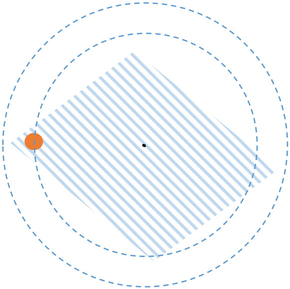


- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

Prior works suggest enforcing regularization (e.g. truncation, projection, regularized loss) to promote incoherence

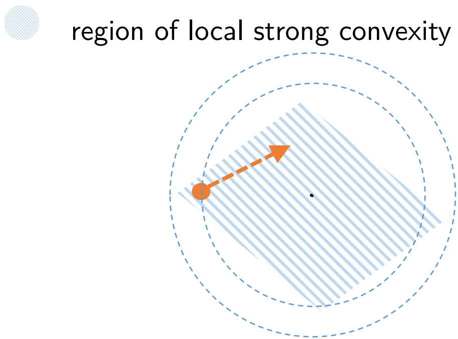
Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness

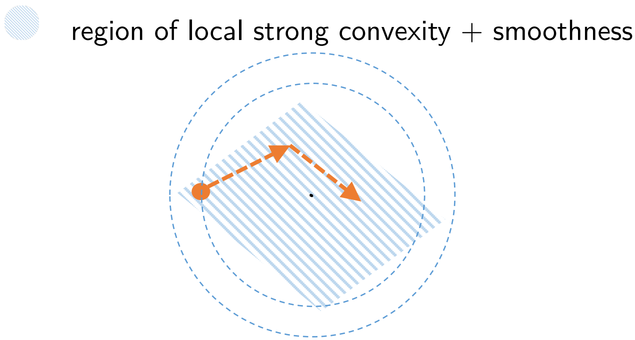


Encouraging message: GD is implicitly regularized

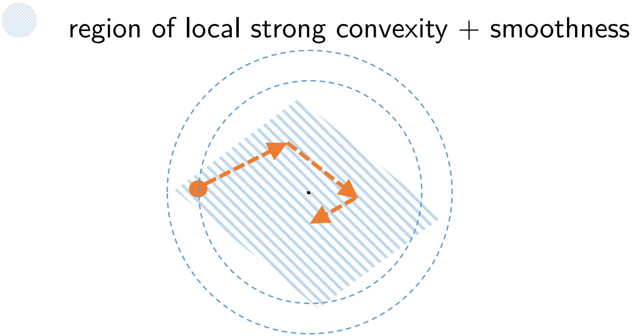
region of local strong convexity + smoothness



Encouraging message: GD is implicitly regularized

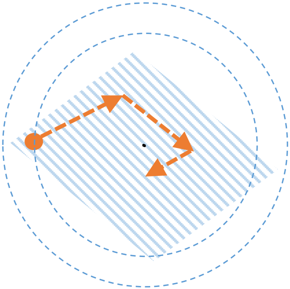


Encouraging message: GD is implicitly regularized



Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent with $\{\boldsymbol{a}_k\}$

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

— cannot be derived from generic optimization theory; relies on finer statistical analysis for entire trajectory of GD

Theoretical guarantees for local refinement stage

Theorem 9.3 (Ma, Wang, Chi, Chen '17)

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

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

Theoretical guarantees for local refinement stage

Theorem 9.3 (Ma, Wang, Chi, Chen '17)

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

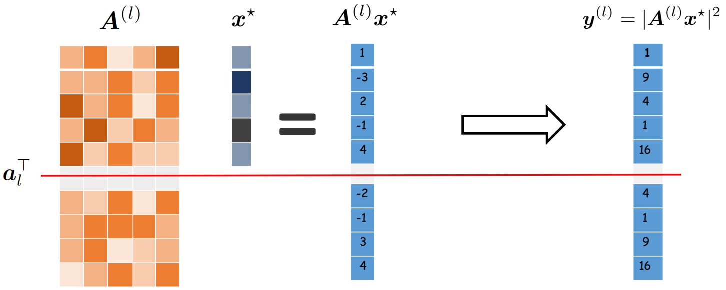
- $\max_k \left|\boldsymbol{a}_k^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\star\|_2$ *(incoherence)*
- $\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\star\|_2$ *(linear convergence)*

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

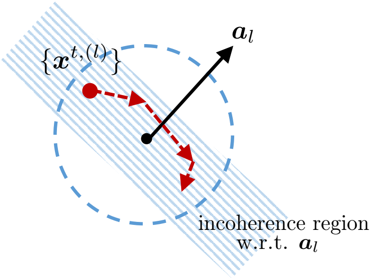
- Attains ε accuracy within $O\big(\log n \, \log \frac{1}{\varepsilon}\big)$ iterations

Key proof idea: leave-one-out analysis

For each $1 \leq l \leq m$, introduce leave-one-out iterates $\boldsymbol{x}^{t,(l)}$ by dropping lth measurement

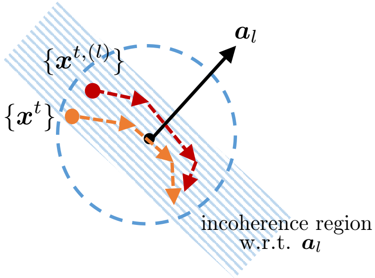


Key proof idea: leave-one-out analysis



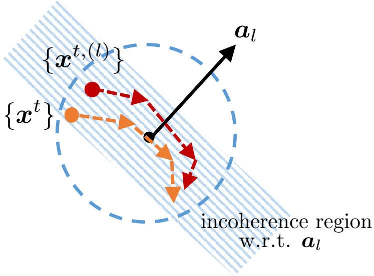
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$

Key proof idea: leave-one-out analysis



- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)} \approx$ true iterate $\boldsymbol{x}^t$

Key proof idea: leave-one-out analysis

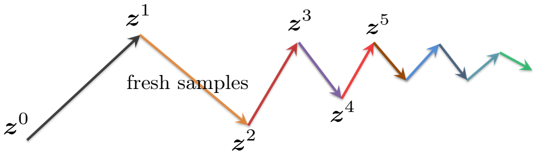


- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)} \approx$ true iterate $\boldsymbol{x}^t$

$$\Longrightarrow \quad \boldsymbol{x}^t \text{ is } \underbrace{\text{nearly independent of}}_{\text{nearly orthogonal to}} \boldsymbol{a}_l$$

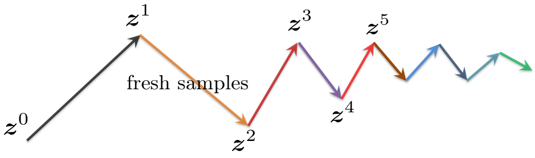
No need of sample splitting

- Several prior works use sample-splitting: require fresh samples at each iteration; not practical but helps analysis

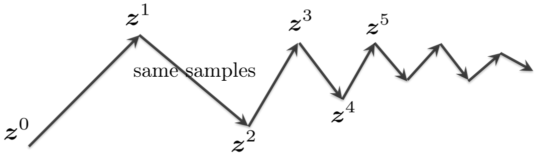


No need of sample splitting

- Several prior works use sample-splitting: require fresh samples at each iteration; not practical but helps analysis



- In contrast, we reuse all samples in all iterations



Architecture of the proof

Local geometry

Lemma 9.4

Suppose $m \geq c_0 n \log n$ *for some sufficiently large constant* $c_0 > 0$. *With high probability,*

$$\nabla^2 f(\boldsymbol{x}) \succeq (1/2) \cdot \boldsymbol{I}_n$$

holds simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n$ *satisfying* $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq 2C_1$; *and*

$$\nabla^2 f(\boldsymbol{x}) \preceq (5C_2(10 + C_2)\log n) \cdot \boldsymbol{I}_n$$

holds simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n$ *obeying*

$$\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq 2C_1, \tag{9.1a}$$

$$\max_{1 \leq j \leq m} \left| \boldsymbol{a}_j^\top (\boldsymbol{x} - \boldsymbol{x}^\star) \right| \leq C_2 \sqrt{\log n}. \tag{9.1b}$$

Error contraction

Lemma 9.5

If $\boldsymbol{x}^t$ *obeys the conditions* (9.1), *whp. one has*

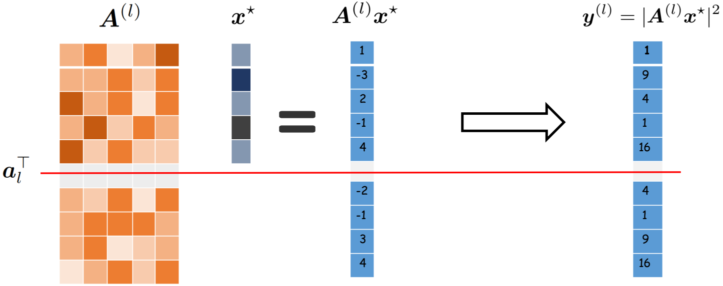
$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{\star}\right\|_2 \leq (1-\eta/2)\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{\star}\right\|_2 \tag{9.2}$$

provided that the step size satisfies $0<\eta\leq 1/\left[5C_2\left(10+C_2\right)\log n\right]$.

— *how to insure incoherence?*

Key proof idea: leave-one-out analysis

For each $1 \leq l \leq m$, introduce leave-one-out iterates $\boldsymbol{x}^{t,(l)}$ by dropping lth measurement



Induction hypotheses

We aim at proving the following claims using induction

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \leq C_1, \tag{9.3a}$$

$$\max_{1\leq l\leq m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \leq C_3\sqrt{\frac{\log n}{n}} \tag{9.3b}$$

$$\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \leq C_2\sqrt{\log n}. \tag{9.3c}$$

Proximity between x^t and $x^{t,(l)}$

Lemma 9.6

Suppose that the sample size obeys $m \geq Cn\log n$ for some sufficiently large constant $C > 0$ and that the stepsize obeys $0 < \eta < 1/[5C_2(10 + C_2)\log n]$. Then whp., one has

$$\max_{1\leq l\leq m} \left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)} \right\|_2 \leq C_3 \sqrt{\frac{\log n}{n}}. \tag{9.4}$$

Incoherence of leave-one-out iterates

By construction, $\boldsymbol{x}^{t+1,(l)}$ is statistically independent of the sampling vector $\boldsymbol{a}_l$. One thus has

$$
\begin{aligned}
\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right| &\leq 5\sqrt{\log n}\left\|\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^\star\right\|_2 \\
&\overset{\text{(i)}}{\leq} 5\sqrt{\log n}\left(\left\|\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1}\right\|_2+\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\right\|_2\right) \\
&\overset{\text{(ii)}}{\leq} 5\sqrt{\log n}\left(C_3\sqrt{\frac{\log n}{n}}+C_1\right) \\
&\leq C_4\sqrt{\log n} \qquad (9.5)
\end{aligned}
$$

holds for some constant $C_4 \geq 6C_1 > 0$ and n sufficiently large. Here, (i) comes from the triangle inequality, and (ii) arises from the proximity bound (9.4) and the conclusion (9.2).

Combining the bounds

$$
\begin{aligned}
\max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\right)\right| &\le \max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\big(\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\big)\right| \\
&\quad + \max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\big(\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^\star\big)\right| \\
&\overset{\text{(i)}}{\le} \max_{1\le l\le m}\|\boldsymbol{a}_l\|_2\big\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\big\|_2 + C_4\sqrt{\log n} \\
&\overset{\text{(ii)}}{\le} \sqrt{6n}\cdot C_3\sqrt{\frac{\log n}{n}} + C_4\sqrt{\log n} \le C_2\sqrt{\log n}
\end{aligned}
$$

Another example: Low-rank matrix completion

Low-rank matrix completion

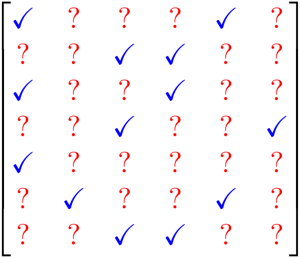
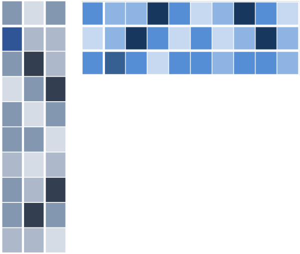


figure credit: Candès

- consider a low-rank matrix $\boldsymbol{M}^{\star} = \boldsymbol{U}^{\star}\boldsymbol{\Sigma}^{\star}\boldsymbol{U}^{\star\top}$
- each entry $M^{\star}_{i,j}$ is observed independently with prob. p
- **Goal:** estimate $\boldsymbol{M}^{\star}$

A natural least-squares loss

Represent low-rank matrix by $\boldsymbol{X}\boldsymbol{X}^\top$ with $\underbrace{\boldsymbol{X} \in \mathbb{R}^{n\times r}}_{\text{low-rank factor}}$



$$\underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X}) = \sum_{(i,j)\in\Omega} \left[\left(\boldsymbol{X}\boldsymbol{X}^\top\right)_{i,j} - M^\star_{i,j} \right]^2$$

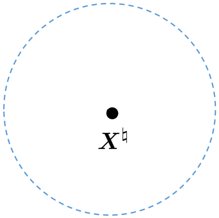
—how does local geometry look like?

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

Incoherence region

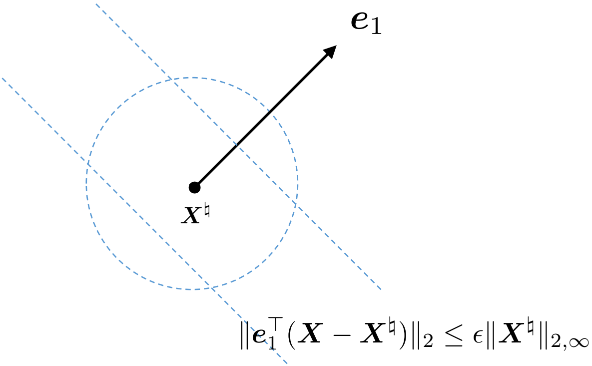
Which region enjoys both restricted strong convexity and smoothness?



- X is not far away from $X^\natural$ in Euclidean metric

Incoherence region

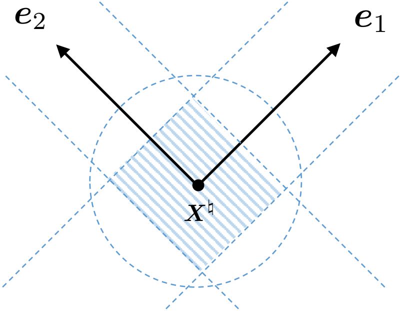
Which region enjoys both restricted strong convexity and smoothness?



- X is not far away from $X^\natural$ in Euclidean metric
- X is incoherent w.r.t. sampling basis (incoherence region)

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?



$\|e_2^\top(X - X^\natural)\|_2 \le \epsilon \|X^\natural\|_{2,\infty}$ $\|e_1^\top(X - X^\natural)\|_2 \le \epsilon \|X^\natural\|_{2,\infty}$

- X is not far away from $X^\natural$ in Euclidean metric
- X is incoherent w.r.t. sampling basis (incoherence region)

Local geometry of $f(\cdot)$

Lemma 9.7

Suppose that $n^2 p \geq C\kappa^2 \mu r n \log n$ *for some sufficiently large constant* $C > 0$. *Then with high probability, the Hessian* $\nabla^2 f(\boldsymbol{X})$ *obeys*

$$\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f\left(\boldsymbol{X}\right) \mathrm{vec}\left(\boldsymbol{V}\right) \geq \frac{\sigma_{\min}}{2} \left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2$$

$$\left\|\nabla^2 f\left(\boldsymbol{X}\right)\right\| \leq \frac{5}{2}\sigma_{\max}$$

for all $\boldsymbol{X}$, $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{H}_Y - \boldsymbol{X}^\star$ *s.t.*
$\boldsymbol{H}_Y := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{Y}\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}$,

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_{2,\infty} \leq \epsilon \left\|\boldsymbol{X}^\star\right\|_{2,\infty},$$

where $\epsilon \ll 1/\sqrt{\kappa^3 \mu r \log^2 n}$.

Restricted local strong convexity

- Due to rotation ambiguity, $f(\cdot)$ cannot be strongly convex along every direction; it is strongly convex along specific directions $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{H}_Y - \boldsymbol{X}^\star$
- Instead of ℓ_{F} ball, $f(X)$ is strongly convex in a local $\ell_{2,\infty}$ ball; $\boldsymbol{X}$ needs to be incoherent in the sense that

$$\|\boldsymbol{X}\|_{2,\infty} \lesssim \sqrt{\frac{\mu r}{n}}\|\boldsymbol{X}^\star\|$$

Revisit Incoherence

Definition 9.8

Fix an orthonormal matrix $\boldsymbol{U}^\star \in \mathbb{R}^{n \times r}$. Define its incoherence to be

$$\mu(\boldsymbol{U}^\star) := \frac{n\|\boldsymbol{U}^\star\|_{2,\infty}^2}{r}$$

—*recover incoherence of eigenvector when* $r = 1$

- For $\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{U}^{\star\top}$, define $\mu(\boldsymbol{M}^\star) := \mu(\boldsymbol{U}^\star)$

Existing solutions to guarantee incoherence

- regularized loss (solve $\text{minimize}_{\boldsymbol{X}} \ f(\boldsymbol{X}) + R(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16, Chen, Li '17

Existing solutions to guarantee incoherence

- regularized loss (solve $\text{minimize}_{\boldsymbol{X}} \; f(\boldsymbol{X}) + R(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16, Chen, Li '17

- projection onto set of incoherent matrices
  - e.g. Chen, Wainwright '15, Zheng, Lafferty '16

Projected gradient descent for matrix completion

(1) **Projected spectral initialization**: let $\boldsymbol{U}^0 \boldsymbol{\Sigma}^0 \boldsymbol{U}^{0\top}$ be rank-r eigendecomposition of

$$\frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{Y}).$$

and set $\boldsymbol{Z}^0 = \boldsymbol{U}^0 \left(\boldsymbol{\Sigma}^0\right)^{1/2}$, and incoherence set

$$\mathcal{C} := \{\boldsymbol{X} \mid \|\boldsymbol{X}\|_{2,\infty} \le \sqrt{\frac{2\mu r}{n}}\|\boldsymbol{Z}^0\|\}$$

let $\boldsymbol{X}^0 = \mathcal{P}_{\mathcal{C}}(\boldsymbol{Z}^0)$

(2) **Projected gradient descent updates:**

$$\boldsymbol{X}^{t+1} = \mathcal{P}_{\mathcal{C}}(\boldsymbol{X}^t - \eta_t \nabla f(\boldsymbol{X}^t)), \qquad t = 0, 1, \cdots$$

Projection operator

Projection onto can be implemented via a row-wise "clipping operation"

$$[\mathcal{P}_{\mathcal{C}}(\boldsymbol{X})]_{i,\cdot} = \min\left\{1, \sqrt{\frac{2\mu r}{n}}\frac{\|\boldsymbol{Z}^0\|}{\|\boldsymbol{X}_{i,\cdot}\|_2}\right\} \cdot \boldsymbol{X}_{i,\cdot}$$

Performance guarantees

Theorem 9.9

Suppose that $n^2 p \geq c_0 \mu^2 r^2 \kappa^2 n \log n$ *for some large constant* $c_0 > 0$. *With high probability, one has for all* $t \geq 0$

$$\|\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\star\|_{\mathrm{F}}^2 \leq \left(1 - \frac{c_1}{\mu^2 r^2 \kappa^2}\right)^t \sigma_r(\boldsymbol{M}^\star),$$

provided that step size is chosen as $\eta \asymp \frac{1}{\mu^2 r^2 \kappa \sigma_1(\boldsymbol{M}^\star)}$

Here $\boldsymbol{Q}^t$ is the optimal alignment matrix between $\boldsymbol{X}^t$ and $\boldsymbol{X}^\star$

$$\boldsymbol{Q}^t := \mathrm{argmin}_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \left\| \boldsymbol{X}^t \boldsymbol{R} - \boldsymbol{X}^\star \right\|_{\mathrm{F}}$$

Regularity condition

Key to prove convergence is the following regularity condition

Lemma 9.10

Suppose that $n^2 p \geq \mu^2 r^2 \kappa^2 n \log n$*. Then with high probability, for all* $\boldsymbol{X} \in \mathcal{C}$*, and* $\|\boldsymbol{X} - \boldsymbol{X}^\star \boldsymbol{H}\|_{\mathrm{F}}^2 \leq \frac{1}{16}\sigma_r(\boldsymbol{M}^\star)$ f *obeys*

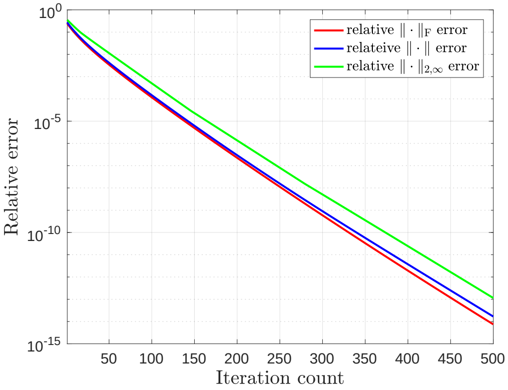
$$\langle \nabla f(\boldsymbol{X}), \boldsymbol{X} - \boldsymbol{X}^\star \boldsymbol{H} \rangle \geq \frac{99}{512}\sigma_r(\boldsymbol{M}^\star)\|\boldsymbol{X} - \boldsymbol{X}^\star \boldsymbol{H}\|_{\mathrm{F}}^2 + \frac{1}{13196\mu^2 r^2 \kappa \sigma_1(\boldsymbol{M}^\star)}\|\nabla f(\boldsymbol{X})\|_{\mathrm{F}}^2$$

Here $\boldsymbol{H}$ is optimal alignment matrix

Is regularization necessary for nonconvex matrix completion?

Numerical surprise with unregularized GD

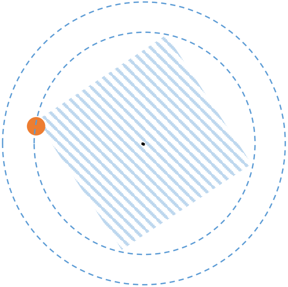
$$n = 1000, r = 10, p = 0.1, \eta = 0.2$$



Vanilla GD without regularization converges fast for MC!

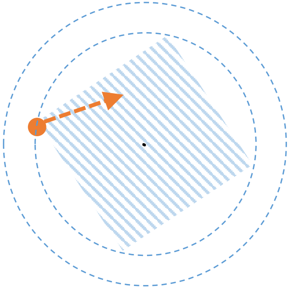
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



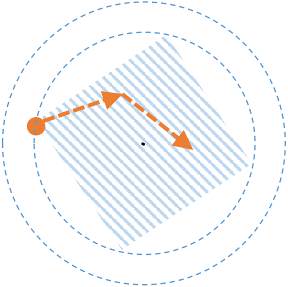
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



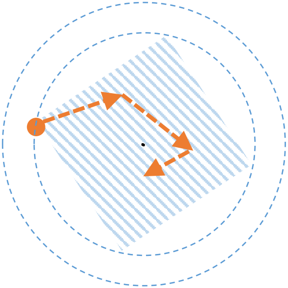
Our findings: GD is implicitly regularized

 region of local strong convexity + smoothness



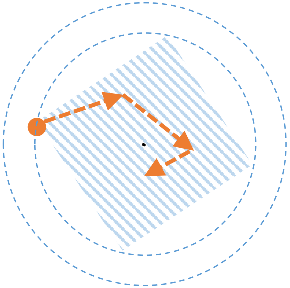
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



Our findings: GD is implicitly regularized

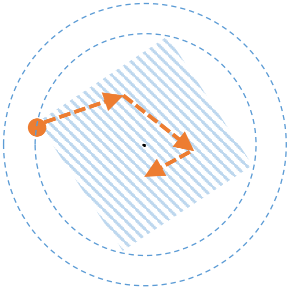
region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent

Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent

— cannot be derived from generic optimization theory; relies on finer statistical analysis for entire trajectory of GD

Theoretical guarantees

Theorem 9.11 (Matrix completion)

Suppose $\boldsymbol{M}$ is rank-r, incoherent and well-conditioned. Vanilla gradient descent (with spectral initialization) achieves ε accuracy

- *in $O\big(\log \frac{1}{\varepsilon}\big)$ iterations*

if step size $\eta \lesssim 1/\sigma_{\max}(\boldsymbol{M})$ and sample size $\gtrsim nr^3 \log^3 n$

Theoretical guarantees

Theorem 9.11 (Matrix completion)

Suppose $\boldsymbol{M}$ is rank-r, incoherent and well-conditioned. Vanilla gradient descent (with spectral initialization) achieves ε accuracy

- *in $O\big(\log\frac{1}{\varepsilon}\big)$ iterations w.r.t. $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|$, and $\underbrace{\|\cdot\|_{2,\infty}}_{\text{incoherence}}$*

if step size $\eta \lesssim 1/\sigma_{\max}(\boldsymbol{M})$ and sample size $\gtrsim nr^3 \log^3 n$

Theoretical guarantees

Theorem 9.11 (Matrix completion)

Suppose $\boldsymbol{M}$ is rank-r, incoherent and well-conditioned. Vanilla gradient descent (with spectral initialization) achieves ε accuracy

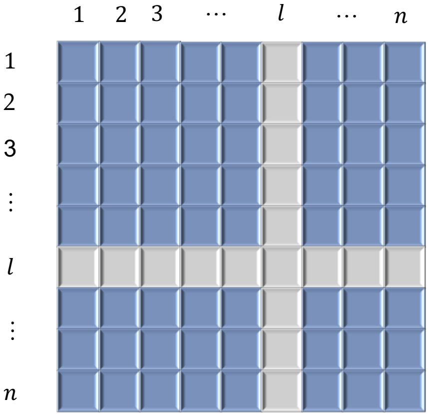
- *in $O\big(\log \frac{1}{\varepsilon}\big)$ iterations w.r.t. $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|$, and $\underbrace{\|\cdot\|_{2,\infty}}_{\textit{incoherence}}$*

if step size $\eta \lesssim 1/\sigma_{\max}(\boldsymbol{M})$ and sample size $\gtrsim nr^3 \log^3 n$

- *Byproduct: vanilla GD controls **entrywise error***
  — errors are spread out across all entries

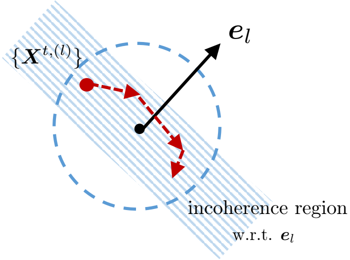
Key ingredient: leave-one-out analysis

For each $1 \le l \le n$, introduce leave-one-out iterates $\boldsymbol{X}^{t,(l)}$ by replacing lth row and column with true values



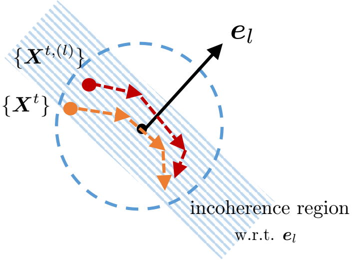
$\boldsymbol{M}^{(l)}$ $\Longrightarrow$ $\boldsymbol{X}^{t,(l)}$

Key ingredient: leave-one-out analysis



- $\boldsymbol{X}^{t,(l)}$ contains more information of lth row of $\boldsymbol{X}^{\natural}$; indep. of randomness in lth row

Key ingredient: leave-one-out analysis



- $\boldsymbol{X}^{t,(l)}$ contains more information of lth row of $\boldsymbol{X}^{\natural}$; indep. of randomness in lth row
- Leave-one-out iterates $\boldsymbol{X}^{t,(l)} \approx$ true iterates $\boldsymbol{X}^{t}$