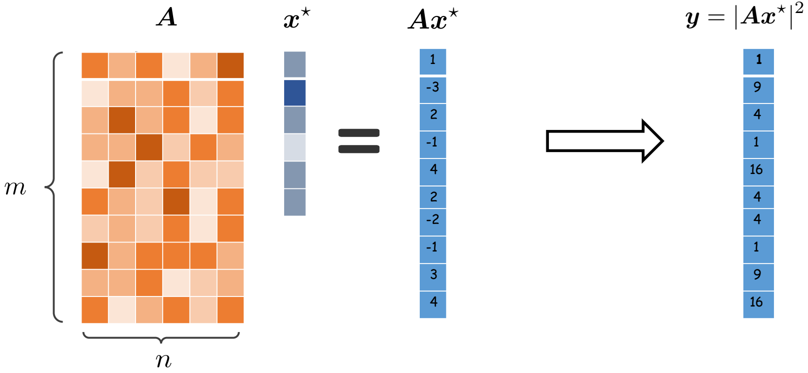
Nonconvex Optimization for High-Dimensional Estimation (Part 2)

Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

A case study: solving quadratic systems of equations

Solving quadratic systems of equations



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

$$y_k \quad = \quad (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \qquad k = 1, \ldots, m$$

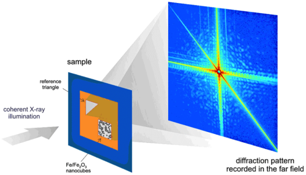
assume w.l.o.g. $\|\boldsymbol{x}^\star\|_2 = 1$

Motivation: phase retrieval

Detectors record intensities of diffracted rays

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

figure credit: Stanford SLAC



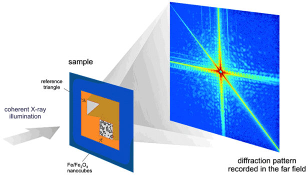
intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Motivation: phase retrieval

Detectors record intensities of diffracted rays

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

figure credit: Stanford SLAC

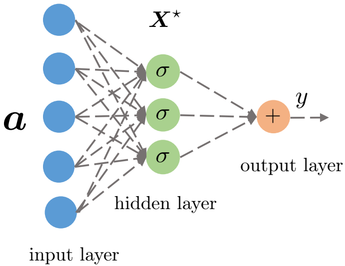


intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

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

Motivation: learning neural nets with quadratic activation

— *Soltanolkotabi, Javanmard, Lee '17, Li, Ma, Zhang '17*

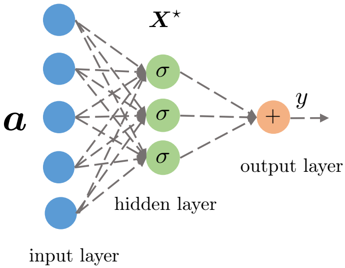


input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

output: $$y = \sum_{i=1}^{r} \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i^\star)$$

Motivation: learning neural nets with quadratic activation

— *Soltanolkotabi, Javanmard, Lee '17, Li, Ma, Zhang '17*

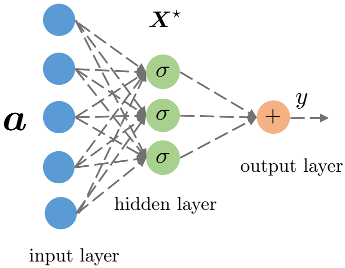


input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

$$\text{output:} \quad y = \sum_{i=1}^{r} \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i^\star) \overset{\sigma(z)=z^2}{:=} \sum_{i=1}^{r} (\boldsymbol{a}^\top \boldsymbol{x}_i^\star)^2$$

Motivation: learning neural nets with quadratic activation

— *Soltanolkotabi, Javanmard, Lee '17, Li, Ma, Zhang '17*



input features: a; weights: $X^\star = [x_1^\star, \cdots, x_r^\star]$

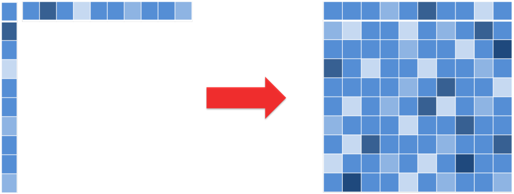
$$\text{output:} \quad y = \sum_{i=1}^{r} \sigma(a^\top x_i^\star) \overset{\sigma(z)=z^2}{:=} \sum_{i=1}^{r} (a^\top x_i^\star)^2$$

We consider simplest model when $r = 1$ (higher r is similar)

An equivalent view: low-rank factorization

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

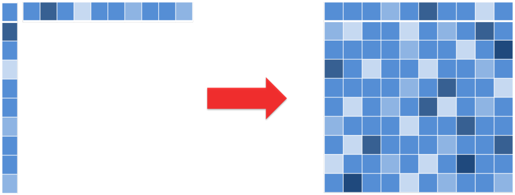
$$y_k = (\boldsymbol{a}_k^\top \boldsymbol{x})^2 = \boldsymbol{a}_k^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} \qquad \Longrightarrow \qquad y_k = \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k$$



An equivalent view: low-rank factorization

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

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

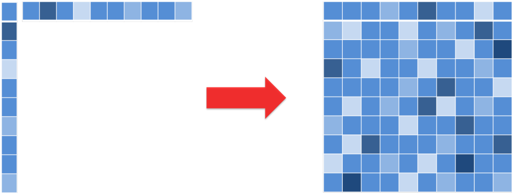


$$\begin{array}{lll} \text{find} & \boldsymbol{X} & \\ \text{s.t.} & y_k = \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k, & k = 1, \cdots, m \\ & \text{rank}(\boldsymbol{X}) = 1 & \end{array}$$

An equivalent view: low-rank factorization

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

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



$$\begin{array}{lll} \text{find} & \boldsymbol{X} & \\ \text{s.t.} & y_k = \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k, & k = 1, \cdots, m \\ & \operatorname{rank}(\boldsymbol{X}) = 1 & \end{array}$$

Solving quadratic systems is essentially low-rank matrix completion

A natural least-squares formulation

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

$$\Downarrow$$

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

A natural least-squares formulation

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

$$\Downarrow$$

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

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

A natural least-squares formulation

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

$$\Downarrow$$

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

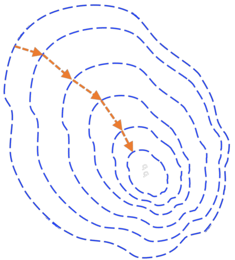
- **pros:** often exact as long as sample size is sufficiently large
- **cons:** $f(\cdot)$ is highly nonconvex
  $\longrightarrow$ *computationally challenging!*

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)

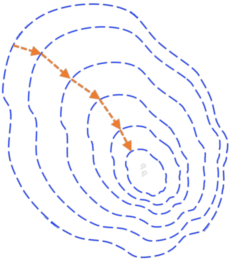
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[\left(\boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$



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

Wirtinger flow (Candès, Li, Soltanolkotabi ’14)

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



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

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

Spectral initialization

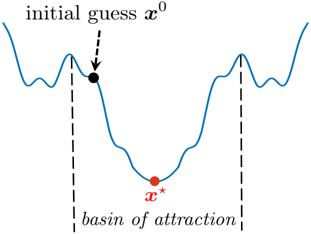
$\boldsymbol{x}^0 \longleftarrow$ leading eigenvector of

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

Rationale: under random Gaussian design $\boldsymbol{a}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$,

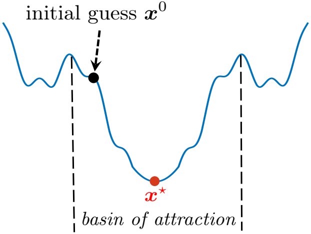
$$\mathbb{E}[\boldsymbol{Y}] := \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{y}_k \boldsymbol{a}_k \boldsymbol{a}_k^\top\right] = \underbrace{\|\boldsymbol{x}^\star\|_2^2\, \boldsymbol{I} + 2\boldsymbol{x}^\star \boldsymbol{x}^{\star\top}}_{\text{leading eigenvector: } \pm \boldsymbol{x}^\star}$$

Rationale of two-stage approach



1. initialize within $\underbrace{\text{local basin sufficiently close to } x^\star}_{\text{(restricted) strongly convex; no saddles / spurious local mins}}$

Rationale of two-stage approach



1. initialize within $\underbrace{\text{local basin sufficiently close to } \boldsymbol{x}^\star}_{\text{(restricted) strongly convex; no saddles / spurious local mins}}$

2. iterative refinement

A highly incomplete list of two-stage methods

phase retrieval:

- Netrapalli, Jain, Sanghavi '13
- Candès, Li, Soltanolkotabi '14
- Chen, Candès '15
- Cai, Li, Ma '15
- Wang, Giannakis, Eldar '16
- Zhang, Zhou, Liang, Chi '16
- Kolte, Ozgur '16
- Zhang, Chi, Liang '16
- Soltanolkotabi '17
- Vaswani, Nayer, Eldar '16
- Chi, Lu '16
- Wang, Zhang, Giannakis, Akcakaya, Chen '16
- Tan, Vershynin '17
- Ma, Wang, Chi, Chen '17
- Duchi, Ruan '17
- Jeong, Gunturk '17
- Yang, Yang, Fang, Zhao, Wang, Neykov '17
- Qu, Zhang, Wright '17
- Goldstein, Studer '16
- Bahmani, Romberg '16
- Hand, Voroninski '16
- Wang, Giannakis, Saad, Chen '17
- Barmherzig, Sun '17
- ...

other problems:

- Keshavan, Montanari, Oh '09
- Sun, Luo '14
- Chen, Wainwright '15
- Tu, Boczar, Simchowitz, Soltanolkotabi, Recht '15
- Zheng, Lafferty '15
- Balakrishnan, Wainwright, Yu '14
- Chen, Suh '15
- Chen, Candès '16
- Li, Ling, Strohmer, Wei '16
- Yi, Park, Chen, Caramanis '16
- Jin, Kakade, Netrapalli '16
- Huang, Kakade, Kong, Valiant '16
- Ling, Strohmer '17
- Li, Ma, Chen, Chi '18
- Aghasi, Ahmed, Hand '17
- Lee, Tian, Romberg '17
- Li, Chi, Zhang, Liang '17
- Cai, Wang, Wei '17
- Abbe, Bandeira, Hall '14
- Chen, Kamath, Suh, Tse '16
- Zhang, Zhou '17
- Boumal '16
- Zhong, Boumal '17
- ...

First theory of WF

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

Theorem 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:* $m \gtrsim n \log n$.

First theory of WF

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

Theorem 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: $O\big(n \log \frac{1}{\epsilon}\big)$

First theory of WF

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

Theorem 1 (Candès, Li, Soltanolkotabi '14)

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

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

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

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

First theory of WF

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

Theorem 1 (Candès, Li, Soltanolkotabi '14)

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

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

with high prob., provided that step size and sample size: .

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

Improved theory of WF

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

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

What does optimization theory say about WF?

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

What does optimization theory say about WF?

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

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

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

What does optimization theory say about WF?

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

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

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

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 $O(n \log \frac{1}{\varepsilon})$ 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 $\eta = O(1/n)$

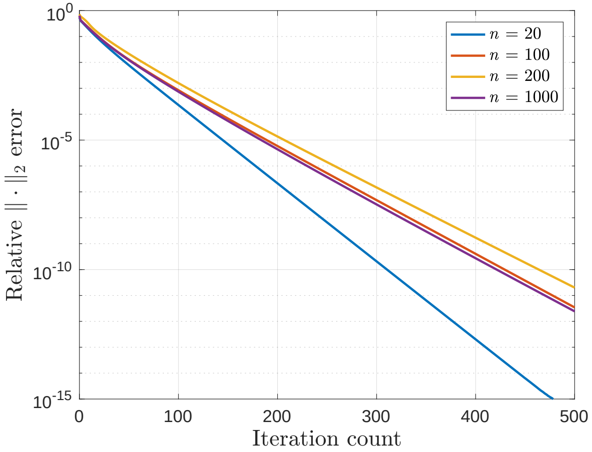
$\Uparrow$

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!

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

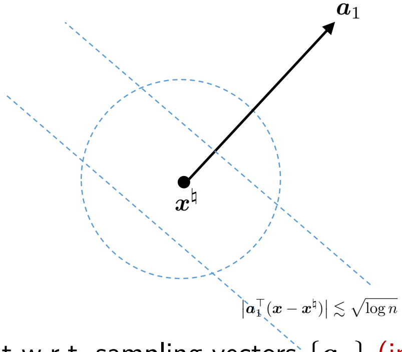
Which local region enjoys both strong convexity and smoothness?

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

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

A second look at gradient descent theory

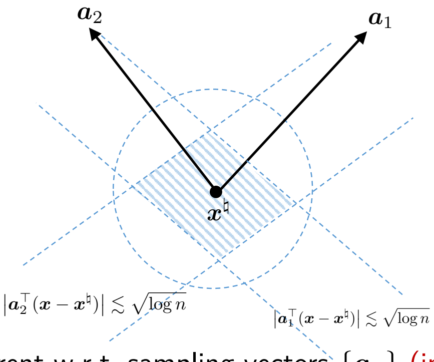
Which local region enjoys both strong convexity and smoothness?



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

A second look at gradient descent theory

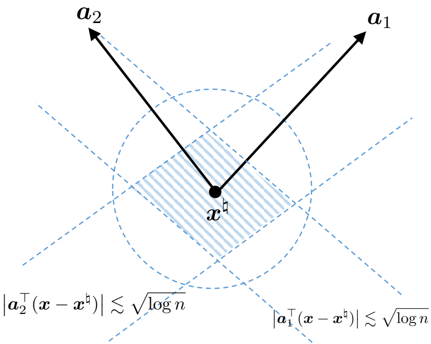
Which local region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ 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?

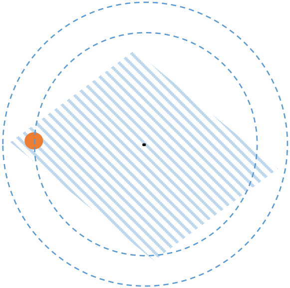


- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

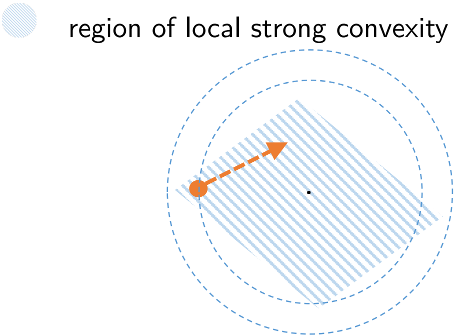
Prior works suggest enforcing regularization (e.g. truncation, projection, regularized loss) to promote incoherence

Encouraging message: GD is implicitly regularized

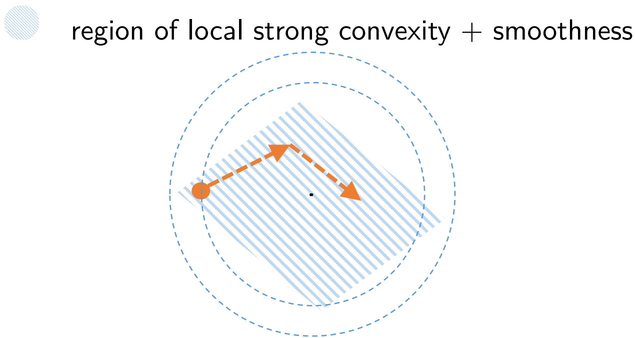
region of local strong convexity + smoothness



Encouraging message: GD is implicitly regularized

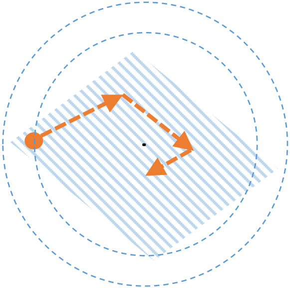


Encouraging message: GD is implicitly regularized

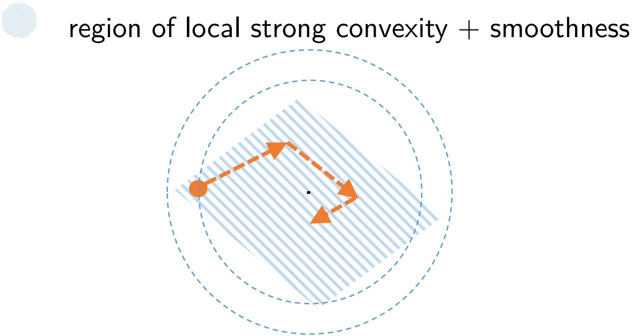


Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness



Encouraging message: GD is implicitly regularized



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 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 3 (Ma, Wang, Chi, Chen '17)

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

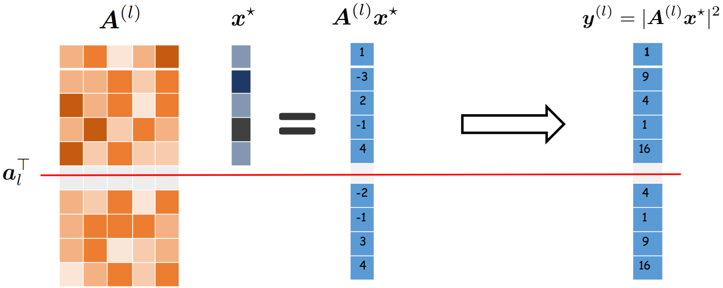
- $\max_k \left|\boldsymbol{a}_k^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\star\|_2$ *(incoherence)*
- $\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\star\|_2$ *(linear convergence)*

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

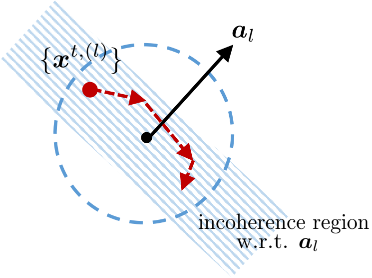
- Attains ε accuracy within $O\big(\log n \, \log \frac{1}{\varepsilon}\big)$ iterations

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

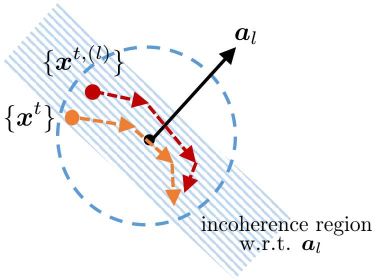


Key proof idea: leave-one-out analysis



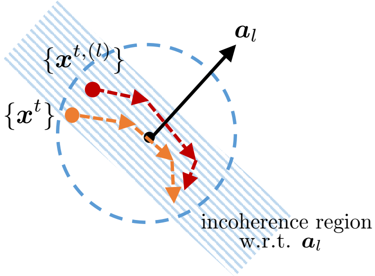
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$

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

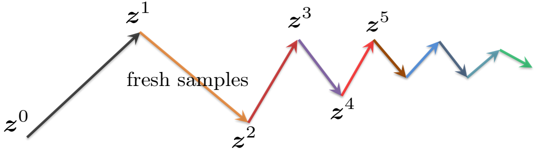


- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)} \approx$ true iterate $\boldsymbol{x}^t$

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

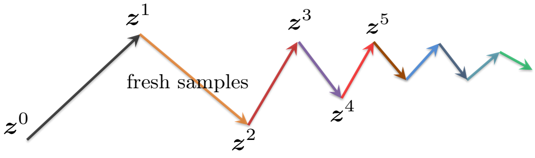
No need of sample splitting

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

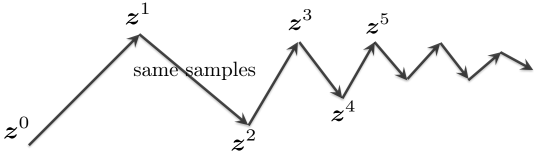


No need of sample splitting

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



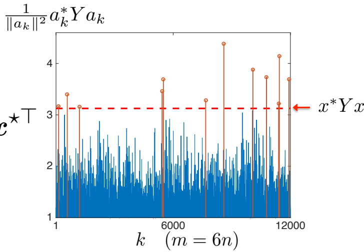
- **This tutorial:** reuses all samples in all iterations



Can we further improve sample complexity (to $O(n)$)?

Truncated spectral initialization

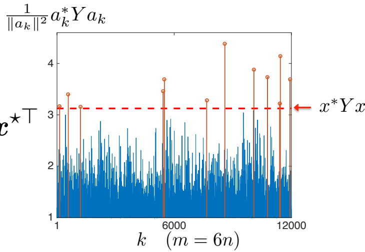
$$\mathbb{E}[\boldsymbol{Y}] := \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{y}_k \boldsymbol{a}_k \boldsymbol{a}_k^\top\right] = \|\boldsymbol{x}^\star\|_2^2\,\boldsymbol{I} + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}$$



problem: unless $m \gg n$, dangerous to use empirical average because large observations $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2$ bear too much influence

Truncated spectral initialization

$$\mathbb{E}[\boldsymbol{Y}] := \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{y}_k \boldsymbol{a}_k \boldsymbol{a}_k^\top\right] = \|\boldsymbol{x}^\star\|_2^2\, \boldsymbol{I} + 2\boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$$

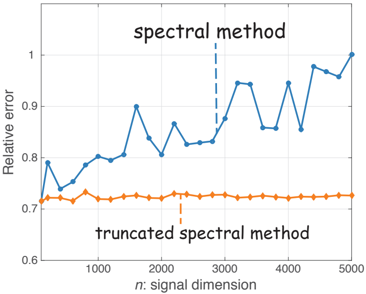


problem: unless $m \gg n$, dangerous to use empirical average because large observations $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2$ bear too much influence

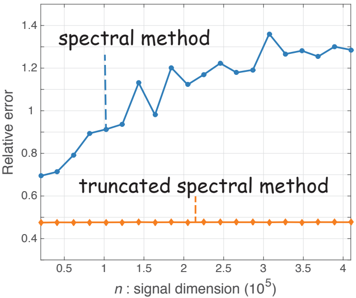
solution: discard high leverage samples and compute leading eigenvector of truncated sum

$$\frac{1}{m}\sum_{k=1}^{m} y_k \boldsymbol{a}_k \boldsymbol{a}_k^\top \cdot \mathbf{1}_{\{|y_k| \le \alpha^2 \,\mathrm{Avg}(|y_j|)\}}$$

Importance of truncated spectral initialization



real Gaussian $m = 6n$

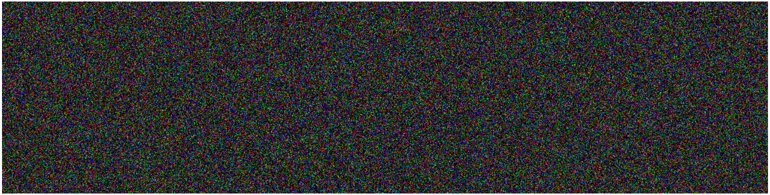


complex CDP $m = 12n$

Importance of truncated spectral initialization

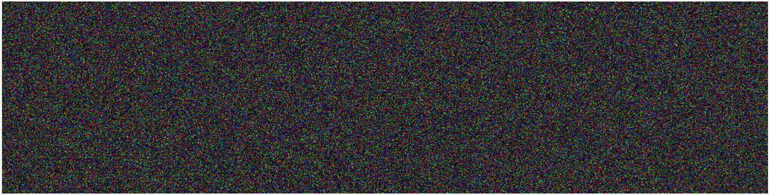
Original image

Importance of truncated spectral initialization



Spectral initialization

Importance of truncated spectral initialization



Spectral initialization

Truncated spectral initialization

Precise asymptotic characterization (Lu, Li '17)

- $m/n \asymp 1$
- i.i.d. Gaussian design

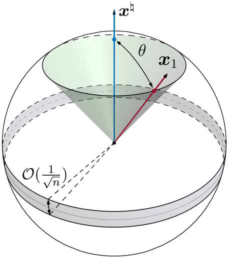


Fig. credit: Lu, Li '17

Theorem 4 (Lu, Li '17, Mondelli, Montanari '17)

There exist analytical formulas $\rho(\cdot)$ *and constants* $\alpha_{\min}$ *and* $\alpha_{\max}$ *s.t.*

$$\underbrace{\frac{\left(\boldsymbol{x}^{\star\top}\boldsymbol{x}^0\right)^2}{\|\boldsymbol{x}^\star\|_2^2\|\boldsymbol{x}^0\|_2^2}}_{\text{cosine similarity}} \longrightarrow \begin{cases} 0, & \text{if } m/n < \alpha_{\min} \\ \rho(m/n), & \text{if } m/n > \alpha_{\max} \end{cases}$$

Theoretical prediction vs. simulations

image size: 64×64

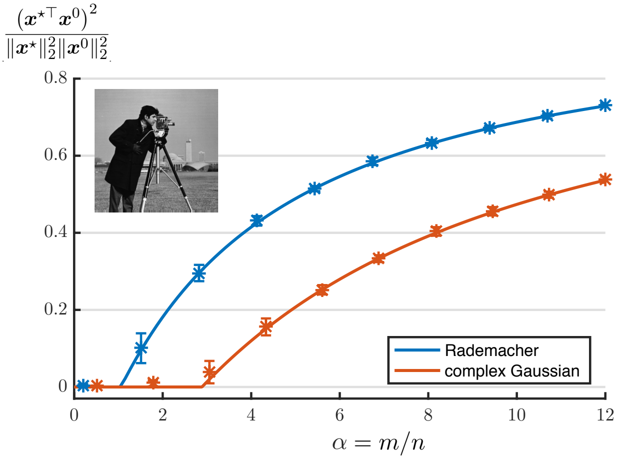


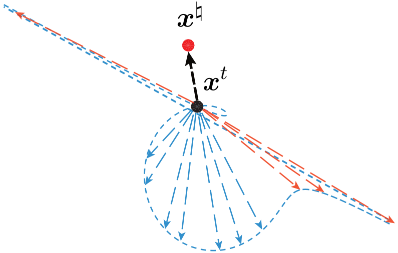
Fig. credit: Lu, Li '17

Improving search directions

$$\text{WF (GD):} \quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_k \nabla f_k(\boldsymbol{x}^t)$$

Improving search directions

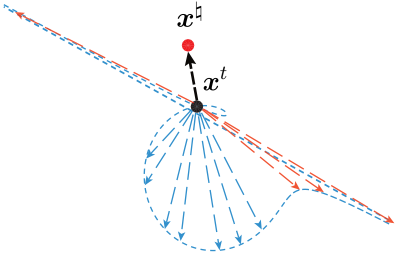
$$\text{WF (GD):} \quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_k \nabla f_k(\boldsymbol{x}^t)$$



locus of $\{\nabla f_k(\boldsymbol{z})\}$

Improving search directions

$$\text{WF (GD):} \quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m}\sum_k \nabla f_k(\boldsymbol{x}^t)$$

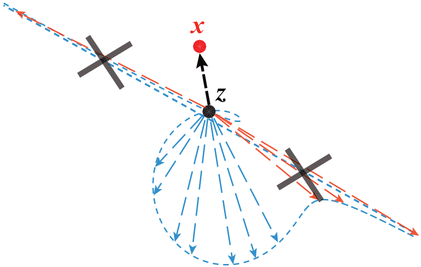


locus of $\{\nabla f_k(\boldsymbol{z})\}$

Problem: descent direction might have large variability

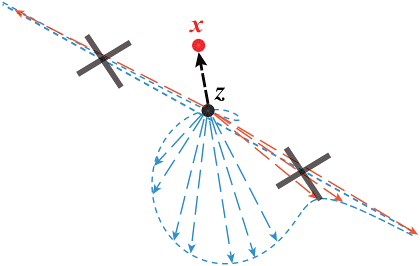
Solution: variance reduction via trimming

More adaptive rule: $\quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_{k \in \mathcal{T}_t} \nabla f_k(\boldsymbol{x}^t)$



Solution: variance reduction via trimming

More adaptive rule: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m}\sum_{k \in \mathcal{T}_t} \nabla f_k(\boldsymbol{x}^t)$



- $\mathcal{T}_t$ trims away excessively large grad components

$$\mathcal{T}_t := \left\{ k : \quad \left\|\nabla f_k(\boldsymbol{x}^t)\right\|_2 \lesssim \text{typical-size}\left\{\left\|\nabla f_l(\boldsymbol{x}^t)\right\|_2\right\}_{1 \le l \le m} \right\}$$

Slight bias $\quad + \quad$ much reduced variance

Summary: truncated Wirtinger flow

(1) **Regularized spectral initialization:** $\boldsymbol{x}^0 \leftarrow$ principal component of

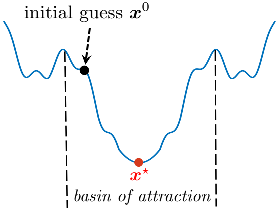
$$\frac{1}{m}\sum\nolimits_{k\in\mathcal{T}_0} y_k\,\boldsymbol{a}_k\boldsymbol{a}_k^\top$$

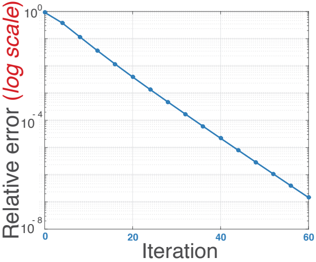
(2) Follow **adaptive gradient descent**

$$\boldsymbol{x}^t = \boldsymbol{x}^t - \frac{\eta_t}{m}\sum\nolimits_{k\in\mathcal{T}_t} \nabla f_k(\boldsymbol{x}^t)$$

Adaptive and iteration-varying rules: discard high-leverage data $\{y_k : k \notin \mathcal{T}_t\}$

Theoretical guarantees (noiseless data)





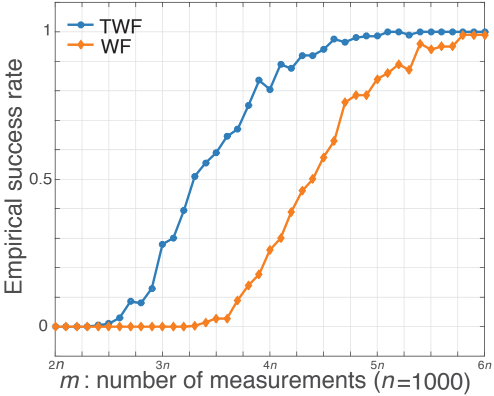
Theorem 5 (Chen, Candès '15)

Suppose $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ *and sample size* $m \gtrsim n$*. With high prob.,*

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) := \min \|\boldsymbol{x}^t \pm \boldsymbol{x}^\star\|_2 \ \leq \ \nu\,(1-\rho)^t\,\|\boldsymbol{x}^\star\|_2$$

where $0 < \nu, \rho < 1$ *are universal constants*

Empirical success rate (noiseless data)



Empirical success rate vs. sample size

Stability vis a vis noise?

Stability under noisy data

- Noisy data: $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2 + \eta_k$
- Signal-to-noise ratio:

$$\mathsf{SNR} := \frac{\sum_k (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^4}{\sum_k \eta_k^2} \approx \frac{3m\|\boldsymbol{x}^\star\|_2^4}{\|\boldsymbol{\eta}\|_2^2}$$

- i.i.d. Gaussian design $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

Stability under noisy data

- Noisy data: $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2 + \eta_k$
- Signal-to-noise ratio:

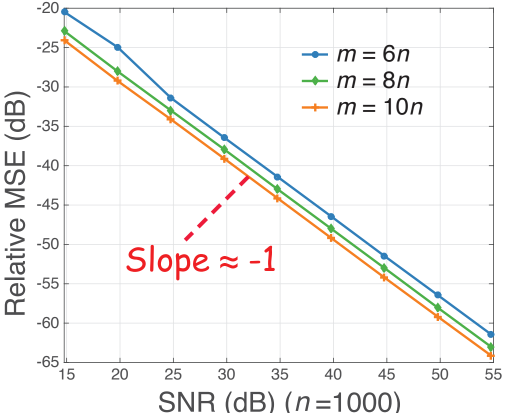
$$\mathsf{SNR} := \frac{\sum_k (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^4}{\sum_k \eta_k^2} \approx \frac{3m\|\boldsymbol{x}^\star\|_2^4}{\|\boldsymbol{\eta}\|_2^2}$$

- i.i.d. Gaussian design $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

Theorem 6 (Chen, Candès '15)

Relative error of TWF converges to $O(\frac{1}{\sqrt{\mathsf{SNR}}})$

Relative MSE vs. SNR (Poisson data)

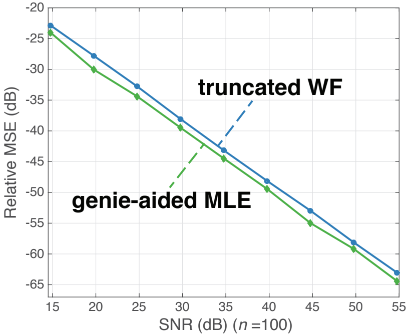


Empirical evidence: relative MSE scales inversely with SNR

This accuracy is nearly un-improvable (empirically)

Comparison with <u>ideal</u> MLE (with phase info. revealed)

ideal knowledge: $y_k \sim \text{Poisson}\left(\left|\boldsymbol{a}_k^\top \boldsymbol{x}^\star\right|^2 \right)$ and $\varepsilon_k = \text{sign}(\boldsymbol{a}_k^\top \boldsymbol{x}^\star)$



Little loss due to missing phases!

This accuracy is nearly un-improvable (theoretically)

- Poisson data: $y_k \overset{\text{ind.}}{\sim} \text{Poisson}(\,|\boldsymbol{a}_k^\top \boldsymbol{x}^\star|^2\,)$
- Signal-to-noise ratio:

$$\text{SNR} \;\approx\; \frac{\sum_k |\boldsymbol{a}_k^\top \boldsymbol{x}^\star|^4}{\sum_k \text{Var}(y_k)} \;\approx\; 3\|\boldsymbol{x}^\star\|_2^2$$

This accuracy is nearly un-improvable (theoretically)

- Poisson data: $y_k \overset{\text{ind.}}{\sim} \text{Poisson}(\,|\boldsymbol{a}_k^\top \boldsymbol{x}^\star|^2\,)$
- Signal-to-noise ratio:

$$\text{SNR} \;\approx\; \frac{\sum_k |\boldsymbol{a}_k^\top \boldsymbol{x}^\star|^4}{\sum_k \text{Var}(y_k)} \;\approx\; 3\|\boldsymbol{x}^\star\|_2^2$$

Theorem 7 (Chen, Candès '15)

Under i.i.d. Gaussian design, for any estimator $\widehat{\boldsymbol{x}}$*,*

$$\inf_{\widehat{\boldsymbol{x}}} \; \sup_{\boldsymbol{x}^\star:\; \|\boldsymbol{x}^\star\|_2 \geq \log^{1.5} m} \frac{\mathbb{E}\left[\text{dist}\left(\widehat{\boldsymbol{x}}, \boldsymbol{x}^\star\right) \mid \{\boldsymbol{a}_k\}\right]}{\|\boldsymbol{x}^\star\|_2} \;\gtrsim\; \frac{1}{\sqrt{\text{SNR}}},$$

provided that sample size $m \asymp n$

Other examples: low-rank matrix estimation

Low-rank matrix completion

Problem: complete a rank-r matrix $\boldsymbol{M}$ from partial entries: $M_{i,j}$, $(i,j) \in \Omega$

- *random sampling:* (i,j) is included in Ω independently with prob. p

$$\text{find low-rank } \widehat{\boldsymbol{M}} \quad \text{s.t.} \quad \mathcal{P}_{\Omega}(\widehat{\boldsymbol{M}}) = \mathcal{P}_{\Omega}(\boldsymbol{M})$$

Low-rank matrix completion

Problem: complete a rank-r matrix $\boldsymbol{M}$ from partial entries: $M_{i,j}$, $(i,j) \in \Omega$

- *random sampling:* (i,j) is included in Ω independently with prob. p

$$\text{find low-rank } \widehat{\boldsymbol{M}} \quad \text{s.t.} \quad \mathcal{P}_{\Omega}(\widehat{\boldsymbol{M}}) = \mathcal{P}_{\Omega}(\boldsymbol{M})$$

Strong convexity and smoothness do not hold in general

$\rightarrow$ need to regularize loss function by promoting **incoherent** solutions

Incoherence for matrix completion

Definition 8 (Incoherence for matrix completion)

A rank-r matrix $\boldsymbol{M}$ with eigendecomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ is said to be μ-incoherent if

$$\|\boldsymbol{U}\|_{2,\infty} \le \sqrt{\frac{\mu}{n}}\|\boldsymbol{U}\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}$$

e.g., $\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\text{hard } \mu=n}$ vs. $\underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}_{\text{easy } \mu=1}$

Gradient descent for matrix completion

Let $\boldsymbol{M} = \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}$. Observe

$$Y_{i,j} = M_{i,j} + E_{i,j}, \quad (i,j) \in \Omega$$

where $(i,j) \in \Omega$ independently with prob. p, and $E_{i,j} \sim \mathcal{N}(0, \sigma^2)$[1]

$$\text{minimize} \left\| \mathcal{P}_\Omega(\widehat{\boldsymbol{M}} - \boldsymbol{Y}) \right\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \text{rank}(\widehat{\boldsymbol{M}}) \le r$$

[1] can be relaxed to sub-Gaussian noise and the asymmetric case

Gradient descent for matrix completion

Let $\boldsymbol{M} = \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}$. Observe

$$Y_{i,j} = M_{i,j} + E_{i,j}, \quad (i,j) \in \Omega$$

where $(i,j) \in \Omega$ independently with prob. p, and $E_{i,j} \sim \mathcal{N}(0, \sigma^2)$[1]

$$\text{minimize} \left\| \mathcal{P}_\Omega(\widehat{\boldsymbol{M}} - \boldsymbol{Y}) \right\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \text{rank}(\widehat{\boldsymbol{M}}) \le r$$

$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad \underbrace{f(\boldsymbol{X}) = \sum_{(j,k) \in \Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k} \right)^2}_{\text{unregularized least-squares loss}}$$

[1] can be relaxed to sub-Gaussian noise and the asymmetric case

Gradient descent for matrix completion

1. **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 $\boldsymbol{X}^0 = \boldsymbol{U}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$

2. **gradient descent updates:**
$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla f\big(\boldsymbol{X}^t\big), \qquad t = 0, 1, \cdots$$

Gradient descent for matrix completion

Define the optimal rotation from the tth iterate $\boldsymbol{X}^t$ to $\boldsymbol{X}^\star$ as

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

where $\mathcal{O}^{r\times r}$ is the set of $r \times r$ orthonormal matrices

- orthogonal Procrustes problem

Gradient descent for matrix completion

Theorem 9 (Noiseless MC, Ma, Wang, Chi, Chen '17)

Suppose $\boldsymbol{M} = \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}$ *is rank-*r*, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves*

- $\|\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\star\|_{\mathrm{F}} \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\star\|_{\mathrm{F}}$,
- $\|\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\star\| \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\star\|$, *(spectral)*
- $\|\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\star\|_{2,\infty} \lesssim \rho^t \mu r \sqrt{\frac{\log n}{np}} \|\boldsymbol{X}^\star\|_{2,\infty}$, *(incoherence)*

where $0 < \rho < 1$*, if the step size* $\eta \asymp 1/\sigma_{max}$ *and the sample complexity* $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$

Gradient descent for matrix completion

Theorem 9 (Noiseless MC, Ma, Wang, Chi, Chen '17)

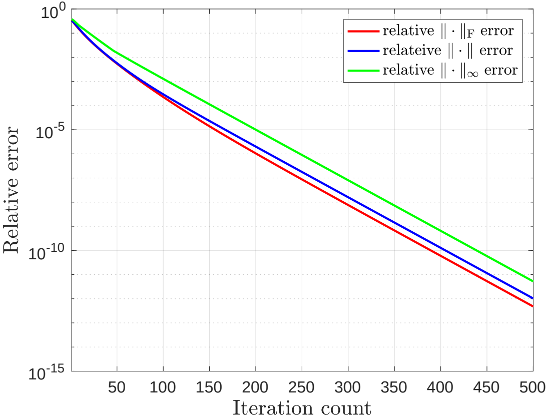
Suppose $M = X^\star X^{\star\top}$ is rank-r, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves

- $\|X^t Q^t - X^\star\|_{\mathrm{F}} \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|X^\star\|_{\mathrm{F}}$,
- $\|X^t Q^t - X^\star\| \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|X^\star\|$, *(spectral)*
- $\|X^t Q^t - X^\star\|_{2,\infty} \lesssim \rho^t \mu r \sqrt{\frac{\log n}{np}} \|X^\star\|_{2,\infty}$, *(incoherence)*

where $0 < \rho < 1$, if the step size $\eta \asymp 1/\sigma_{max}$ and the sample complexity $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$

- vanilla GD converges linearly for matrix completion!

Numerical evidence for noiseless data



Relative error of $\boldsymbol{X}^t\boldsymbol{X}^{t\top}$ (measured by $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_{\infty}$) vs. iteration count for matrix completion, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$

Related theory

$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k) \in \Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k} \right)^2$$

Related theory promotes incoherence explicitly:

Related theory

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k}\right)^2$$

Related theory promotes incoherence explicitly:

- regularized loss (solve $\min_{\boldsymbol{X}} \; f(\boldsymbol{X}) + Q(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16

Related theory

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k}\right)^2$$

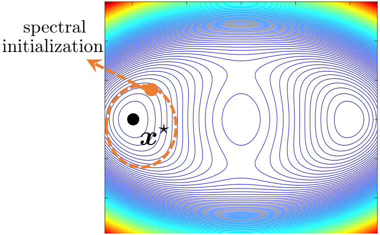
Related theory promotes incoherence explicitly:

- regularized loss (solve $\min_{\boldsymbol{X}} \; f(\boldsymbol{X}) + Q(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16
- projection onto set of incoherent matrices
  - e.g. Chen, Wainwright '15, Zheng, Lafferty '16

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

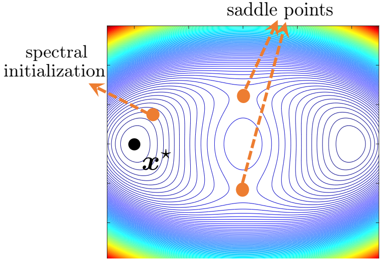
Are carefully-designed initialization or saddle-point escaping schemes necessary for fast convergence?

Initialization



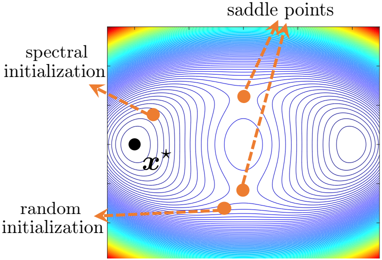
- Spectral initialization gets us reasonably close to truth

Initialization



- Spectral initialization gets us reasonably close to truth
- Cannot initialize GD from anywhere, e.g. it might get stucked at local stationary points (e.g. saddle points)

Initialization

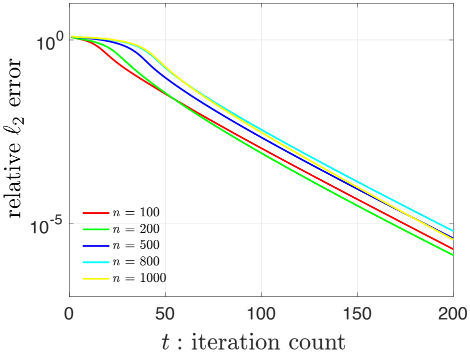


- Spectral initialization gets us reasonably close to truth
- Cannot initialize GD from anywhere, e.g. it might get stucked at local stationary points (e.g. saddle points)

Can we initialize GD randomly, which is simpler and model-agnostic?

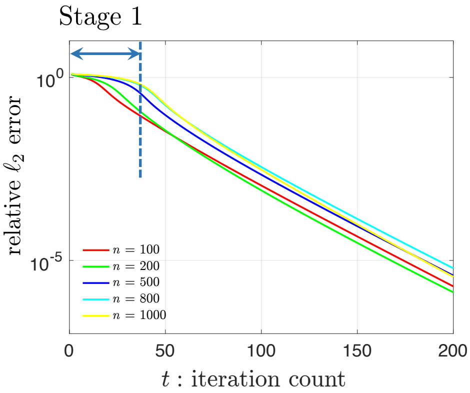
Numerical efficiency of randomly initialized GD

$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



Numerical efficiency of randomly initialized GD

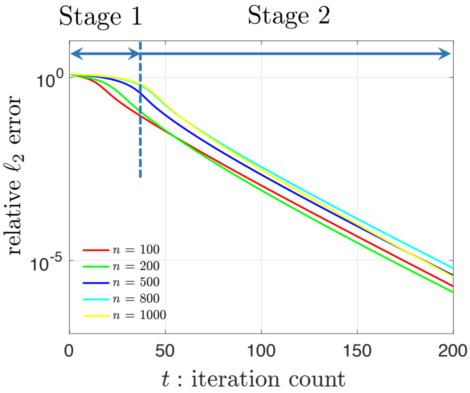
$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



Randomly initialized GD enters local basin within **a few iterations**

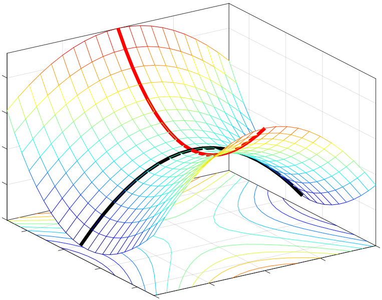
Numerical efficiency of randomly initialized GD

$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$$



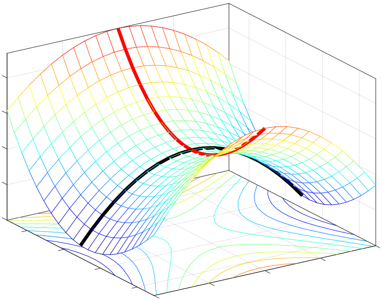
Randomly initialized GD enters local basin within **a few iterations**

A geometric analysis



- if $m \gtrsim n \log^3 n$, then (Sun et al. '16)
  - there is no spurious local mins
  - all saddle points are strict (i.e. associated Hessian matrices have at least one sufficiently negative eigenvalue)

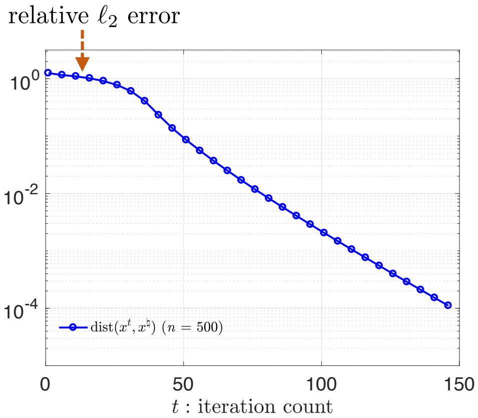
A geometric analysis



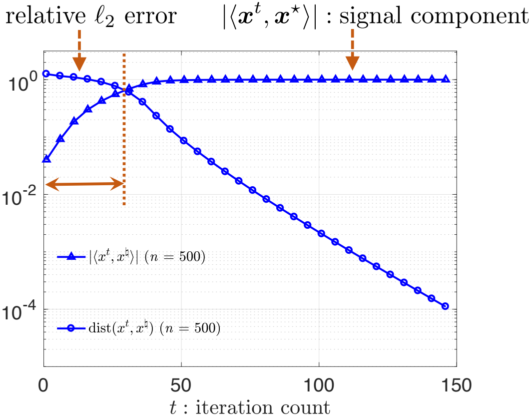
- With such benign landscape, GD with random initialization converges to global min almost surely (Lee et al. '16)

No convergence rate guarantees for vanilla GD!

Exponential growth of signal strength in Stage 1

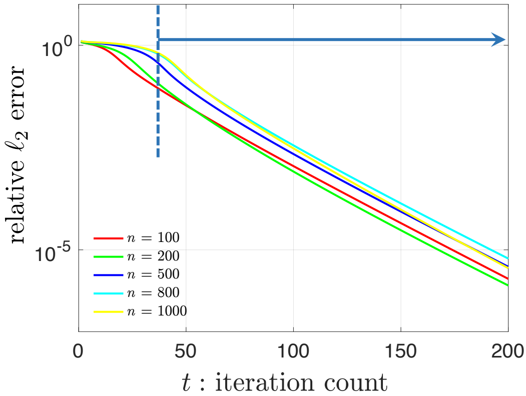


Exponential growth of signal strength in Stage 1

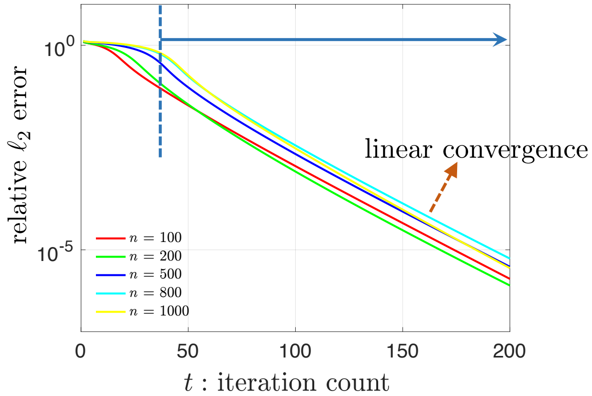


Numerically, $O(\log n)$ iterations are enough to enter local region

Linear / geometric convergence in Stage 2



Linear / geometric convergence in Stage 2



Numerically, GD converges linearly within local region

Theoretical guarantees for randomly initialized GD

These numerical findings can be formalized when $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$:

Theorem 10 (Chen, Chi, Fan, Ma '18)

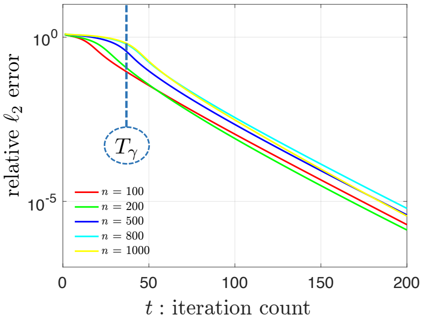
Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$ achieves

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \qquad t \geq T_\gamma$$

for $T_\gamma \lesssim \log n$ and some constants $\gamma, \rho > 0$, provided that step size $\eta \asymp 1$ and sample size $m \gtrsim n$ poly $\log m$

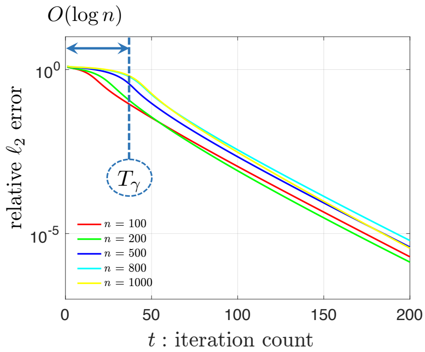
Theoretical guarantees for randomly initialized GD

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$



Theoretical guarantees for randomly initialized GD

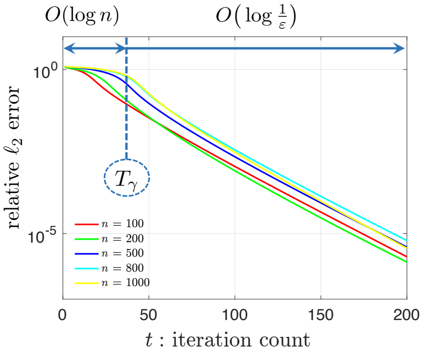
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$



- *Stage 1:* takes $O(\log n)$ iterations to reach $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma$

Theoretical guarantees for randomly initialized GD

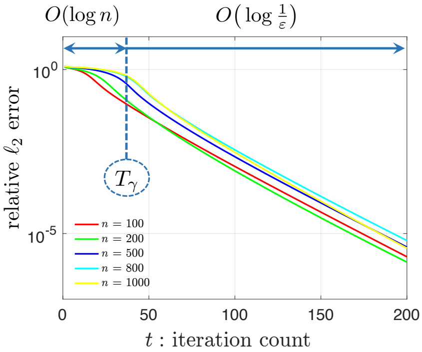
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma (1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$



- *Stage 1:* takes $O(\log n)$ iterations to reach $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma$
- *Stage 2:* linear convergence

Theoretical guarantees for randomly initialized GD

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma (1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$

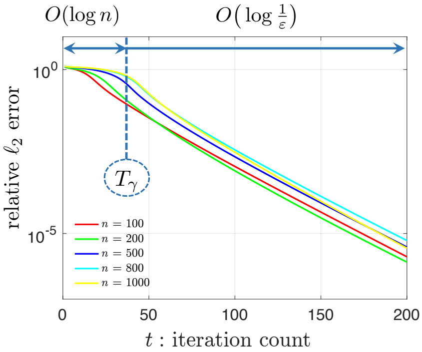


- *near-optimal compututational cost*:

  — $O\big(\log n + \log \frac{1}{\varepsilon}\big)$ iterations to yield ε accuracy

Theoretical guarantees for randomly initialized GD

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$

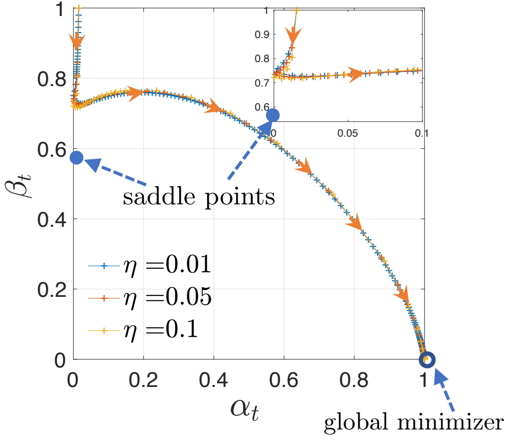


- *near-optimal compututational cost*:

  — $O\big(\log n + \log\frac{1}{\varepsilon}\big)$ iterations to yield ε accuracy

- *near-optimal sample size*: $m \gtrsim n\text{poly}\log m$

Saddle-escaping schemes?



Randomly initialized GD never hits saddle points in phase retrieval!

Other saddle-escaping schemes

	iteration complexity	num of iterations needed to escape saddles	local iteration complexity
Trust-region (Sun et al. '16)	$n^7 + \log\log\frac{1}{\varepsilon}$	n^7	$\log\log\frac{1}{\varepsilon}$
Perturbed GD (Jin et al. '17)	$n^3 + n\log\frac{1}{\varepsilon}$	n^3	$n\log\frac{1}{\varepsilon}$
Perturbed accelerated GD (Jin et al. '17)	$n^{2.5} + \sqrt{n}\log\frac{1}{\varepsilon}$	$n^{2.5}$	$\sqrt{n}\log\frac{1}{\varepsilon}$
GD (Chen et al. '18)	$\log n + \log\frac{1}{\varepsilon}$	$\log n$	$\log\frac{1}{\varepsilon}$

Generic optimization theory yields highly suboptimal convergence guarantees

Reference

- "*Nonconvex optimization meets low-rank matrix factorization: An overview*," Y. Chi, Y. M. Lu, Y. Chen, *IEEE Transactions on Signal Processing*, vol. 67, no. 20, 2019.
- "*Phase retrieval via Wirtinger flow: Theory and algorithms*," E. J. Candès, X. Li, M. Soltanolkotabi, *IEEE Transactions on Information Theory*, vol. 61, no. 4, 2015.
- "*Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution*," C. Ma, K. Wang, Y. Chi, Y. Chen, *Foundations of Computational Mathematics*, vol. 20, no. 3, 2020.
- "*Solving random quadratic systems of equations is nearly as easy as solving linear systems*," Y. Chen, E. J. Candès, *Communications on Pure and Applied Mathematics*, vol. 70, no. 5, 2017.

Reference

- "*Phase transitions of spectral initialization for high-dimensional non-convex estimation*," Y. M. Lu, G. Li, *Information and Inference*, vol. 9, no. 3, 2020.
- "*Fundamental limits of weak recovery with applications to phase retrieval*," M. Mondelli, A. Montanari, *Foundations of Computational Mathematics*, vol. 19, no. 3, 2019.
- "*Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval*," Y. Chen, J. Fan, Y. Chi, C. Ma, *Mathematical Programming*, vol. 176, no. 1, 2019.
- "*Matrix completion from a few entries*," R. H. Keshavan, A. Montanari, S. Oh, *IEEE Transactions on Information Theory*, vol. 56, no. 6, 2010.
- "*Guaranteed matrix completion via non-convex factorization*," R. Sun, Z. Q. Luo, *IEEE Transactions on Information Theory*, vol. 62, no. 11, 2016.

Reference

- "*Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees*," Y. Chen, M. Wainwright, 2015.
- "*Matrix completion has no spurious local minimum*," R. Ge, J. D. Lee, T. Ma, *NeurIPS*, 2016.