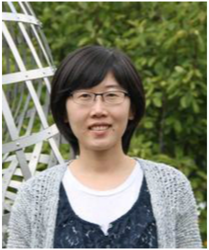
Implicit Regularization in Nonconvex Statistical Estimation

Yuxin Chen

Electrical Engineering, Princeton University

Cong Ma
Princeton ORFE

Kaizheng Wang
Princeton ORFE

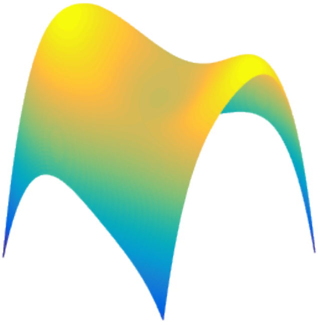


Yuejie Chi
CMU ECE / OSU ECE

Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad \ell(\boldsymbol{x}; \boldsymbol{y})$$

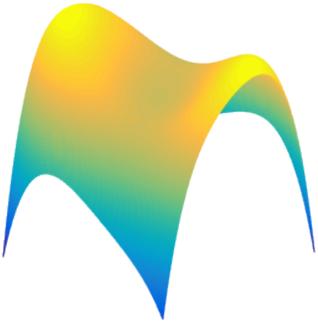


Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

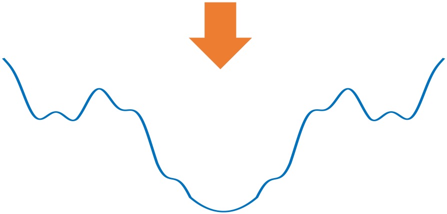
$$\text{minimize}_{\boldsymbol{x}} \quad \ell(\boldsymbol{x}; \boldsymbol{y})$$

- low-rank matrix completion
- graph clustering
- dictionary learning
- mixture models
- deep learning
- ...



Blessing of randomness

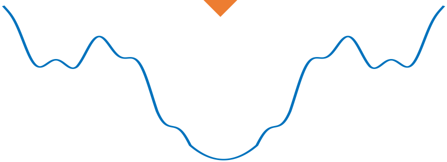
statistical models



benign landscape

Blessing of randomness

statistical models

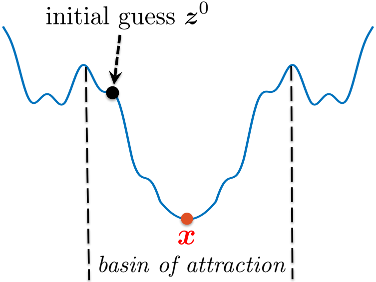


benign landscape

exploit geometry

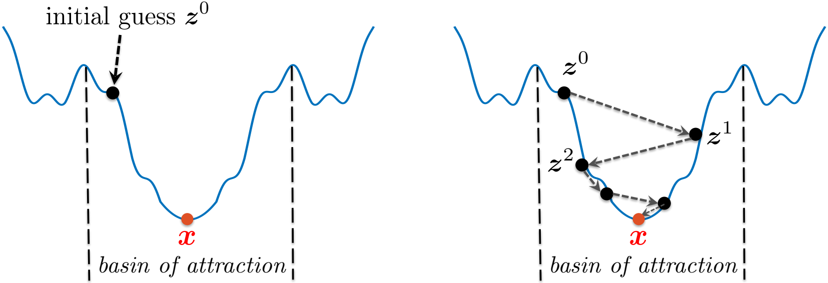
efficient algorithms

Optimization-based methods: two-stage approach



- Start from an appropriate initial point

Optimization-based methods: two-stage approach



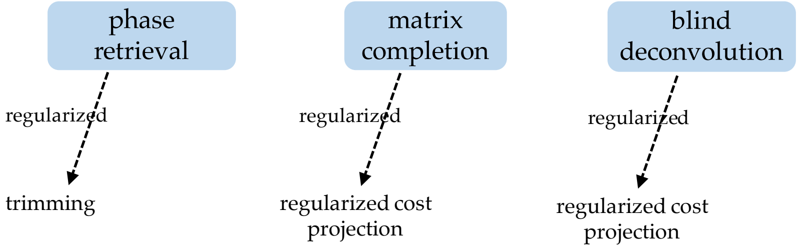
- Start from an appropriate initial point
- Proceed via some iterative optimization algorithms

Proper regularization is *often* recommended

Improves computation by stabilizing search directions

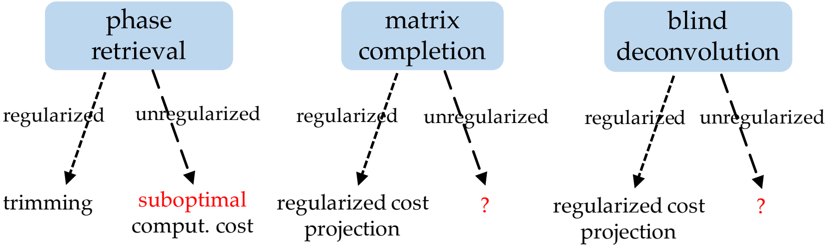
Proper regularization is *often* recommended

Improves computation by stabilizing search directions



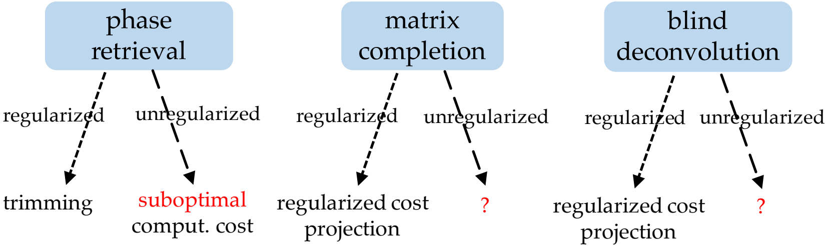
How about unregularized gradient methods?

Improves computation by stabilizing search directions



How about **unregularized** gradient methods?

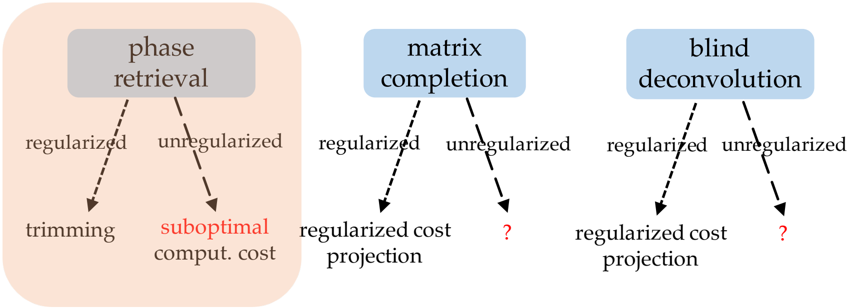
Improves computation by stabilizing search directions



Are unregularized methods suboptimal for nonconvex estimation?

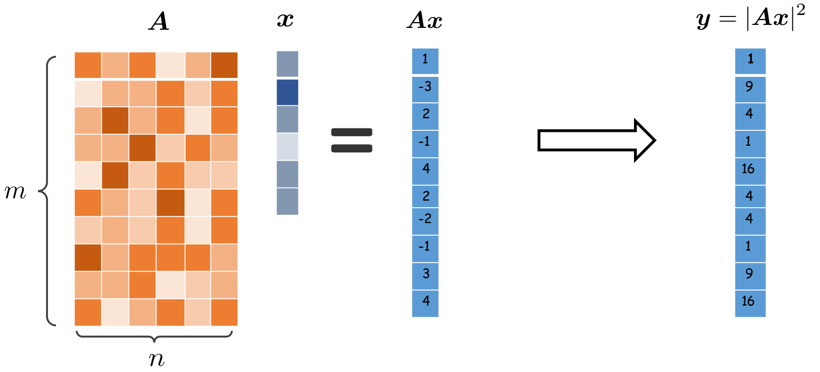
How about **unregularized** gradient methods?

Improves computation by stabilizing search directions



Are unregularized methods suboptimal for nonconvex estimation?

Phase retrieval / solving quadratic systems



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

$$y_k = |\boldsymbol{a}_k^\top \boldsymbol{x}^\natural|^2, \qquad k = 1, \ldots, m$$

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

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

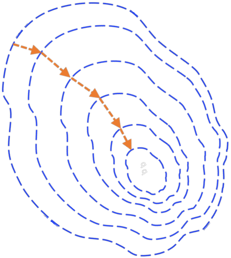
Empirical loss minimization

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

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

Empirical loss minimization

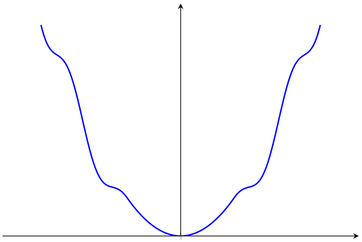
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



- **Initialization by spectral method**
- **Gradient iterations:** for $t = 0, 1, \ldots$

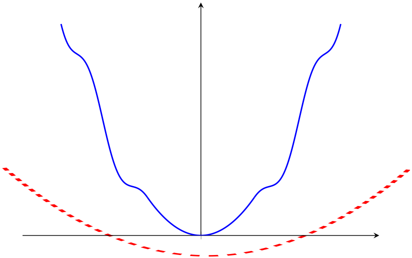
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)$$

Gradient descent theory revisited



Two standard conditions that enable linear convergence of GD

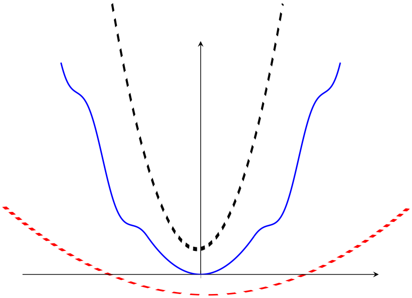
Gradient descent theory revisited



Two standard conditions that enable linear convergence of GD

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

Gradient descent theory revisited



Two standard conditions that enable linear convergence of GD

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

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

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

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

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

- Attains ε-accuracy within $O\left(\frac{\beta}{\alpha} \log \frac{1}{\varepsilon}\right)$ 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?

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

Population level (infinite samples)

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

Consequence: WF converges within logarithmic iterations if $m \to \infty$

What does this optimization theory say about WF?

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

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

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

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 ($m \asymp n \log n$)

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

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 ($m \asymp n \log n$)

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

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

What does this optimization theory say about WF?

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

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

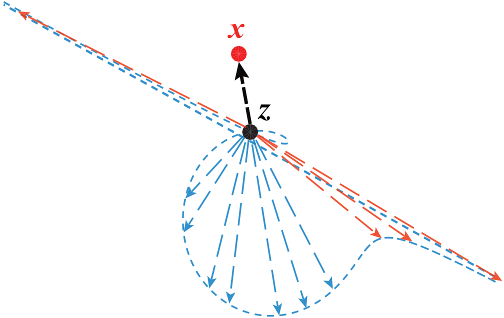
$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp\, n} \text{ (even locally)}$$

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

Too slow ... can we accelerate it?

One solution: truncated WF (Chen, Candès ’15)

Regularize / trim gradient components to accelerate convergence



But wait a minute ...

WF converges in $O(n)$ iterations

But wait a minute ...

WF converges in $O(n)$ iterations

$\Uparrow$

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

But wait a minute ...

WF converges in $O(n)$ iterations

$\Uparrow$

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

$\Uparrow$

This choice is suggested by generic optimization theory

But wait a minute ...

WF converges in $O(n)$ iterations

$\Uparrow$

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

$\Uparrow$

This choice is suggested by worst-case optimization theory

But wait a minute ...

WF converges in $O(n)$ iterations

$\Uparrow$

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

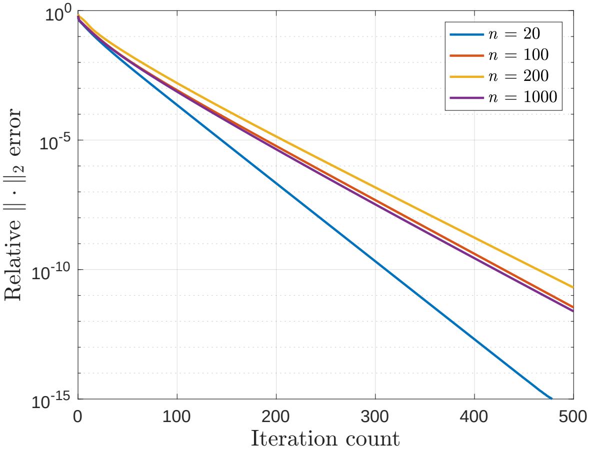
$\Uparrow$

This choice is suggested by worst-case optimization theory

$\Uparrow$

Does it capture what really happens?

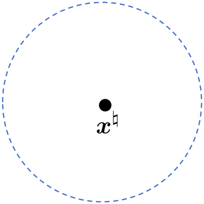
Numerical surprise with $\eta_t = 0.1$



Vanilla GD (WF) can proceed much more aggressively!

A second look at gradient descent theory

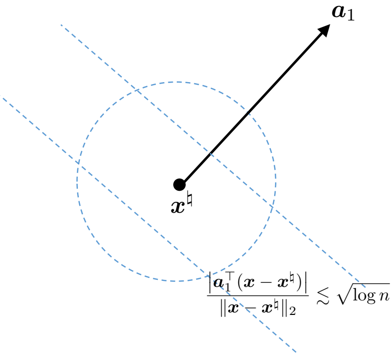
Which region enjoys both strong convexity and smoothness?



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

A second look at gradient descent theory

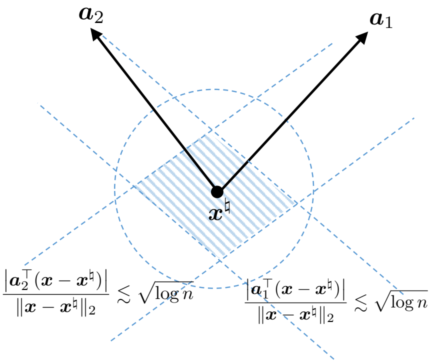
Which region enjoys both strong convexity and smoothness?



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

A second look at gradient descent theory

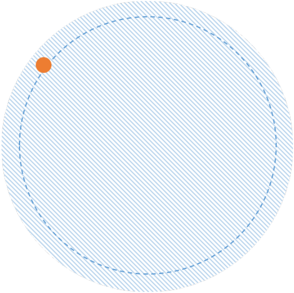
Which region enjoys both strong convexity and smoothness?



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

A second look at gradient descent theory

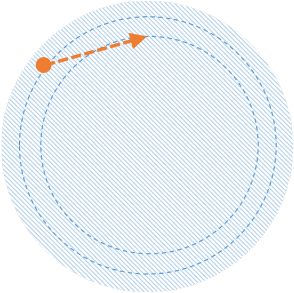
region of local strong convexity + smoothness



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

A second look at gradient descent theory

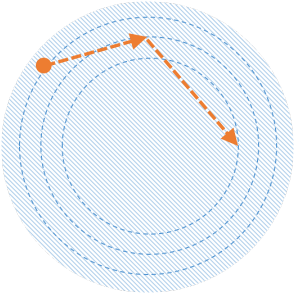
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

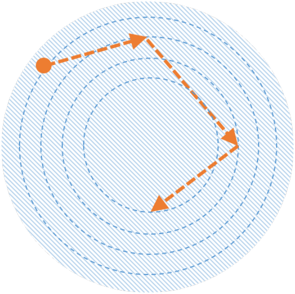
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

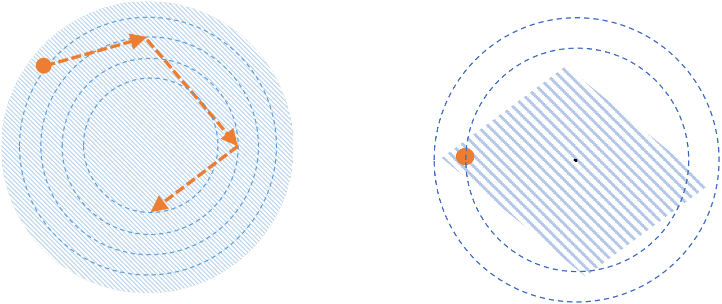
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

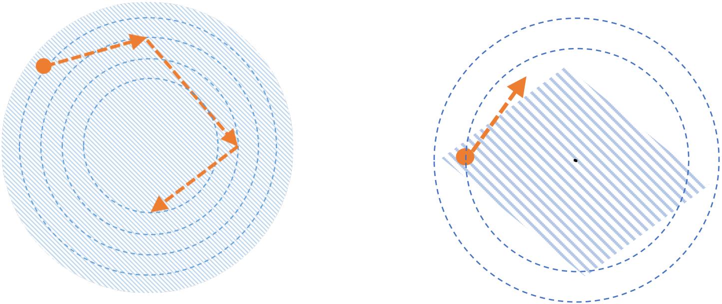
 region of local strong convexity + smoothness



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

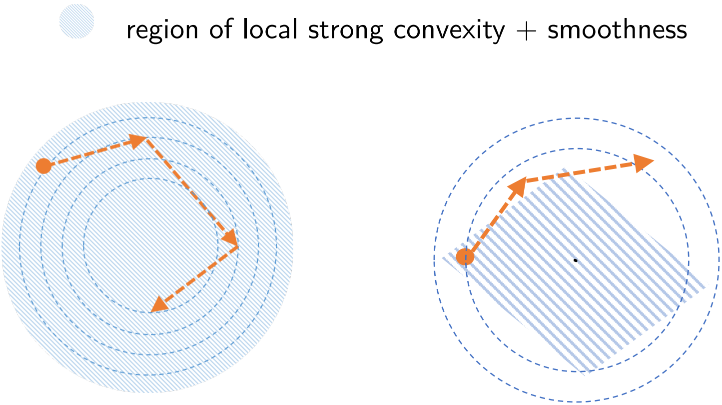
A second look at gradient descent theory

 region of local strong convexity + smoothness



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

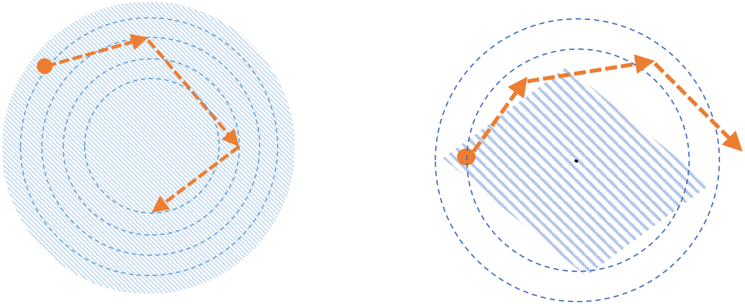
A second look at gradient descent theory



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

A second look at gradient descent theory

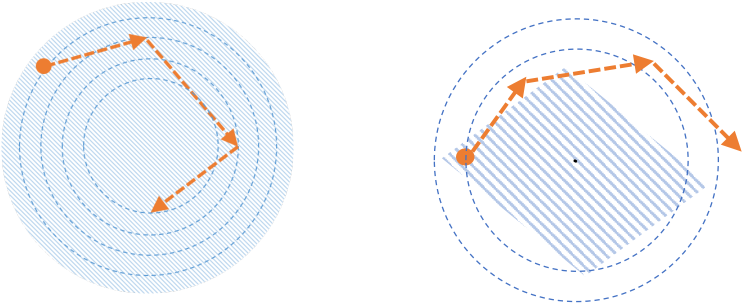
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

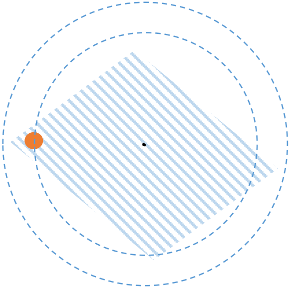
 region of local strong convexity + smoothness



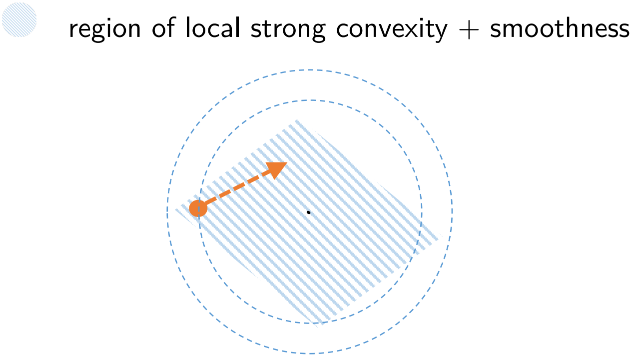
- Prior theory only ensures that iterates remain in ℓ_2 ball but not incoherence region
- *Prior works enforce explicit regularization to promote incoherence*

Our findings: GD is implicitly regularized

 region of local strong convexity + smoothness

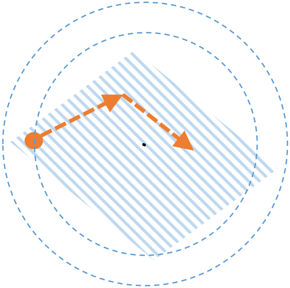


Our findings: GD is implicitly regularized

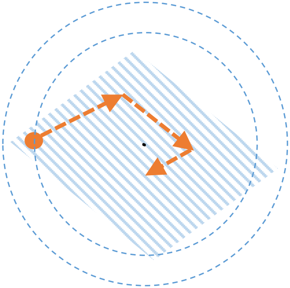


Our findings: GD is implicitly regularized

region of local strong convexity + smoothness

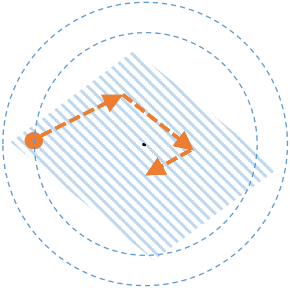


Our findings: GD is implicitly regularized



Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent

Theoretical guarantees

Theorem 1 (Phase retrieval)

Under i.i.d. Gaussian design, WF achieves

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

Theoretical guarantees

Theorem 1 (Phase retrieval)

Under i.i.d. Gaussian design, WF achieves

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

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

Theoretical guarantees

Theorem 1 (Phase retrieval)

Under i.i.d. Gaussian design, WF achieves

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

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

- Much more aggressive step size: $\frac{1}{\log n}$ (vs. $\frac{1}{n}$)

Theoretical guarantees

Theorem 1 (Phase retrieval)

Under i.i.d. Gaussian design, WF achieves

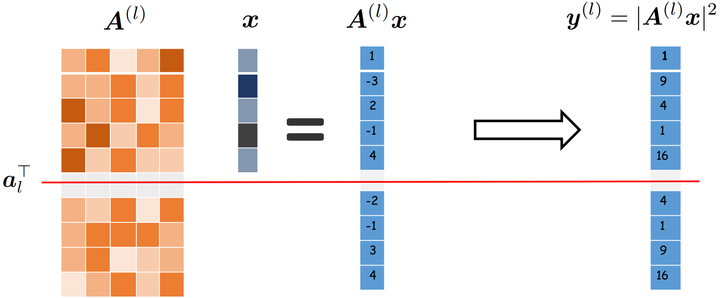
- $\max_k \left| \boldsymbol{a}_k^\top (\boldsymbol{x}^t - \boldsymbol{x}^\natural) \right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\natural\|_2$ *(incoherence)*
- $\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2 \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$ *(near-linear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

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

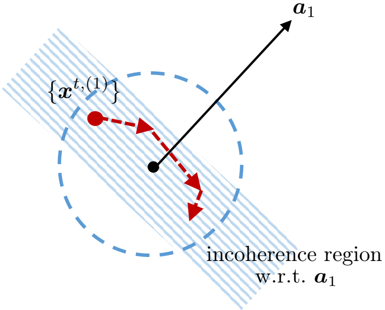
- Much more aggressive step size: $\frac{1}{\log n}$ (vs. $\frac{1}{n}$)
- Computational complexity: $n/\log n$ times faster than existing theory for WF

Key ingredient: leave-one-out analysis

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

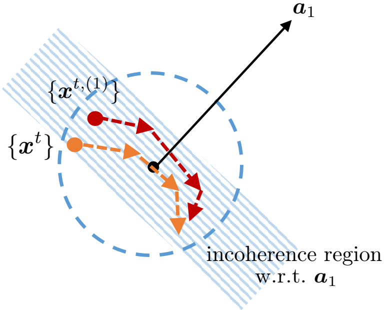


Key ingredient: leave-one-out analysis



- 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.

Key ingredient: leave-one-out analysis



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

This recipe is quite general

Low-rank matrix completion

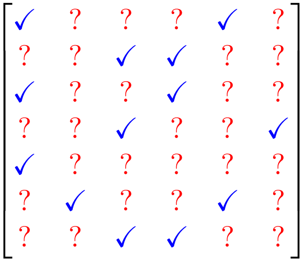


Fig. credit: Candès

Given partial samples Ω of a *low-rank* matrix $\boldsymbol{M}$, fill in missing entries

Prior art

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

Existing theory on gradient descent requires

Prior art

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

Existing theory on gradient descent requires

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

Prior art

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

Existing theory on gradient descent requires

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

Theoretical guarantees

Theorem 2 (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 2 (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 2 (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*

Blind deconvolution

image deblurring

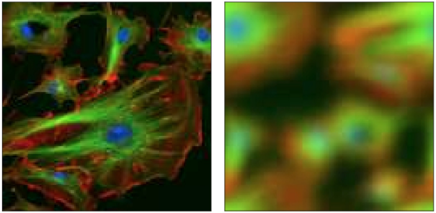


Fig. credit: Romberg

multipath in wireless comm

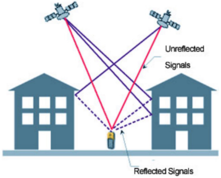
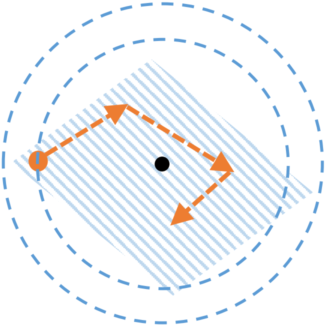


Fig. credit: EngineeringsALL

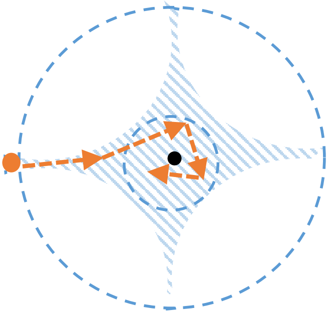
Reconstruct two signals from their convolution

Vanilla GD attains ε-accuracy within $O\big(\log\frac{1}{\varepsilon}\big)$ iterations

Incoherence region in high dimensions



2-dimensional



high-dimensional (mental representation)

incoherence region is vanishingly small

Summary

- **Implict regularization:** vanilla gradient descent automatically foces iterates to stay *incoherent*

Summary

- **Implict regularization:** vanilla gradient descent automatically foces iterates to stay *incoherent*
- Enable error controls in a much stronger sense (e.g. *entrywise error control*)

Paper:

"Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution", Cong Ma, Kaizheng Wang, Yuejie Chi, Yuxin Chen, arXiv:1711.10467