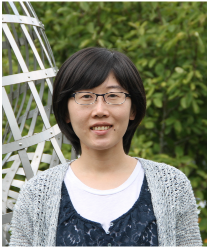
Implicit Regularization in Nonconvex Statistical Estimation

Yuxin Chen

Electrical Engineering, Princeton University

Cong Ma
Princeton ORFE

Kaizheng Wang
Princeton ORFE

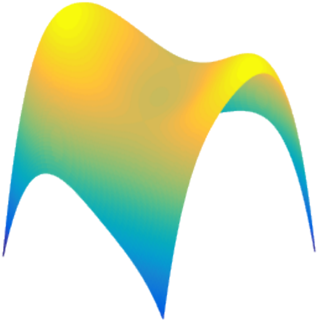


Yuejie Chi
CMU ECE

Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & \ell(\boldsymbol{x}; \boldsymbol{y}) \quad \rightarrow \quad \text{may be nonconvex} \\ \text{subj. to} \quad & \boldsymbol{x} \in \mathcal{S} \quad \rightarrow \quad \text{may be nonconvex} \end{aligned}$$

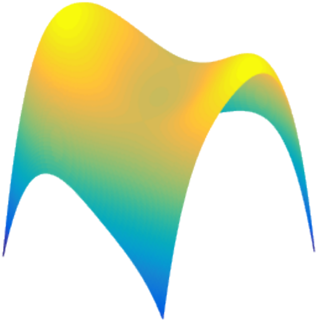


Nonconvex estimation problems are everywhere

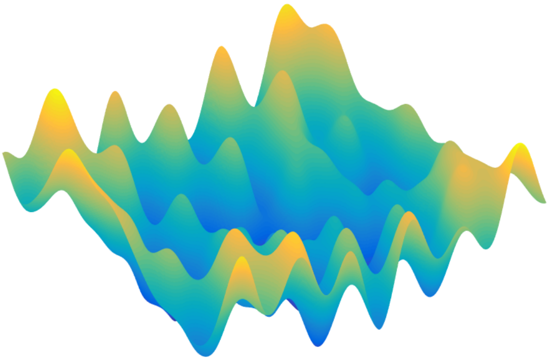
Empirical risk minimization is usually nonconvex

$$\begin{array}{rcl} \text{minimize}_{\boldsymbol{x}} & \ell(\boldsymbol{x}; \boldsymbol{y}) & \rightarrow \quad \text{may be nonconvex} \\ \text{subj. to} & \boldsymbol{x} \in \mathcal{S} & \rightarrow \quad \text{may be nonconvex} \end{array}$$

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



Nonconvex optimization may be super scary



There may be bumps everywhere and exponentially many local optima

e.g. 1-layer neural net (Auer, Herbster, Warmuth '96; Vu '98)

Nonconvex optimization may be super scary

There may be bumps everywhere and exponentially many local optima

e.g. 1-layer neural net (Auer, Herbster, Warmuth '96; Vu '98)

... but is sometimes much nicer than we think

Under certain statistical models,
we see benign global geometry: no spurious local optima

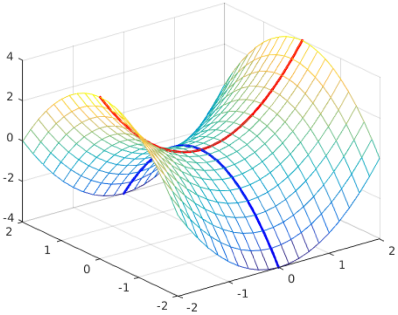
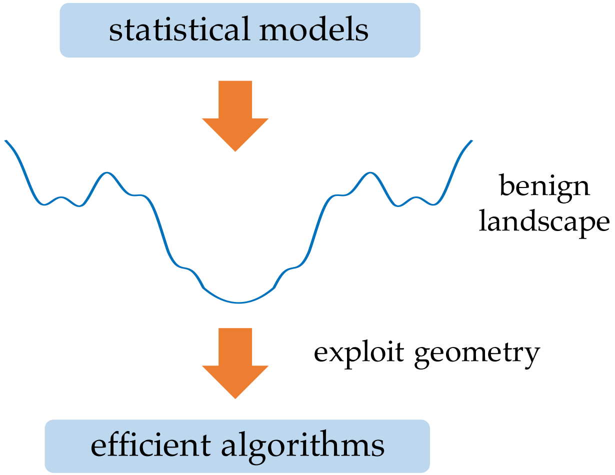
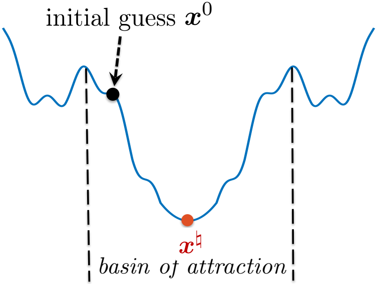


Fig credit: Sun, Qu & Wright

… but is sometimes much nicer than we think

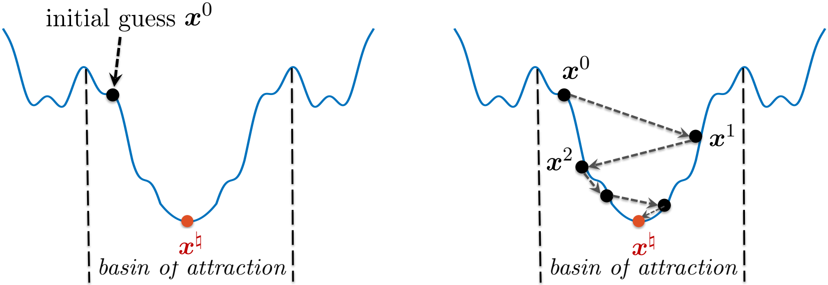


Optimization-based methods: two-stage approach



- Start from an appropriate initial point

Optimization-based methods: two-stage approach



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

Roles of regularization

- Prevents overfitting and improves generalization
  - e.g. ℓ_1 penalization, SCAD, nuclear norm penalization, ...

Roles of regularization

- Prevents overfitting and improves generalization
  - e.g. ℓ_1 penalization, SCAD, nuclear norm penalization, ...

- Improves computation by stabilizing search directions
  - e.g. trimming, projection, regularized loss

Roles of regularization

- Prevents overfitting and improves generalization
  - e.g. ℓ_1 penalization, SCAD, nuclear norm penalization, ...

- Improves computation by stabilizing search directions
  $\Longrightarrow$ focus of this talk
  - e.g. trimming, projection, regularized loss

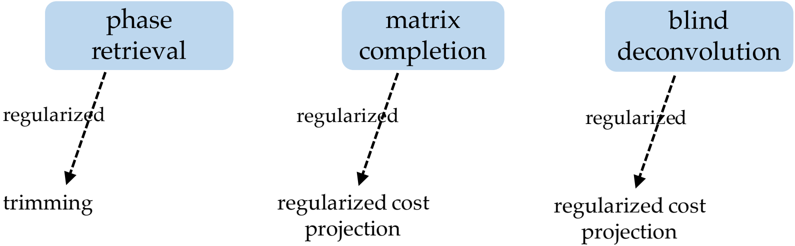
3 representative nonconvex problems

phase retrieval

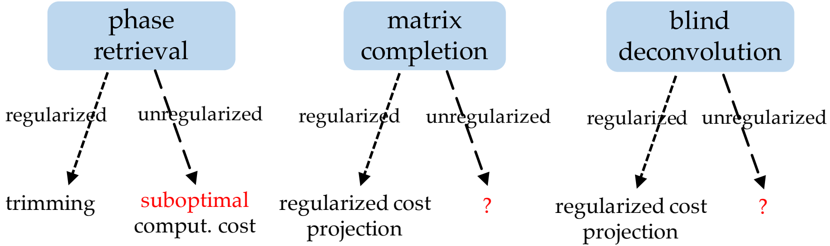
matrix completion

blind deconvolution

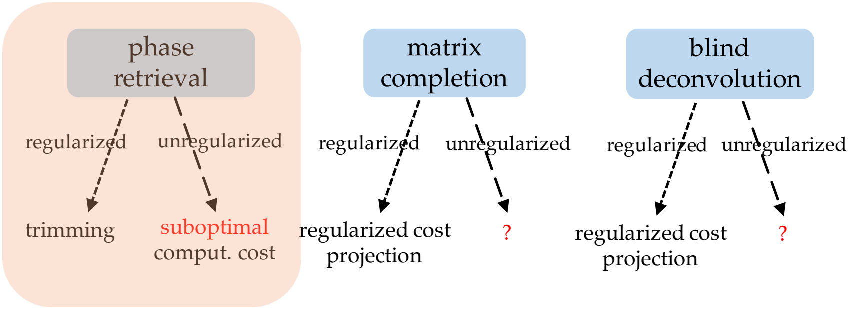
Regularized methods



Regularized vs. unregularized methods



Regularized vs. unregularized methods



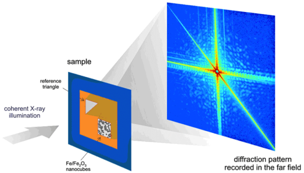
Are unregularized methods suboptimal for nonconvex estimation?

Missing phase problem

Detectors record intensities of diffracted rays

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

Fig credit: Stanford SLAC



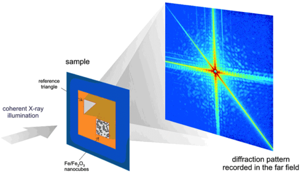
intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Missing phase problem

Detectors record intensities of diffracted rays

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

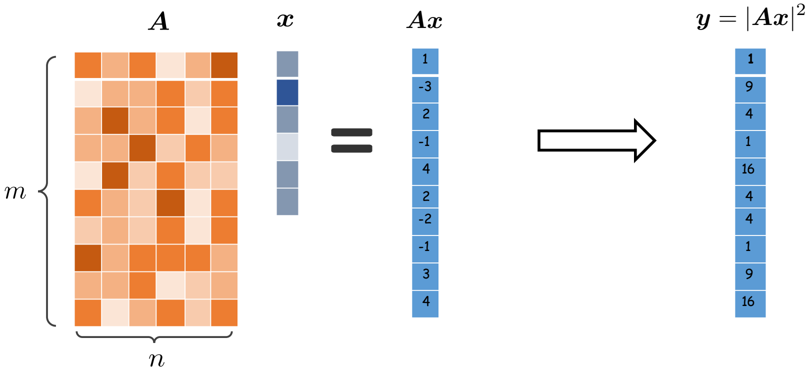
Fig credit: Stanford SLAC



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 $\left|\widehat{x}(f_1, f_2)\right|^2$

Solving quadratic systems of equations



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

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

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

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

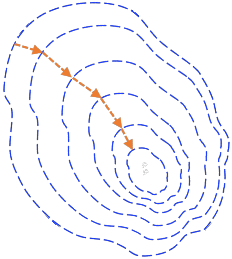
Empirical risk minimization

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

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

Empirical risk minimization

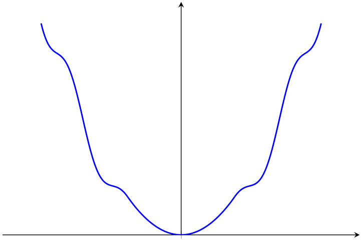
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[\left(\boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$



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

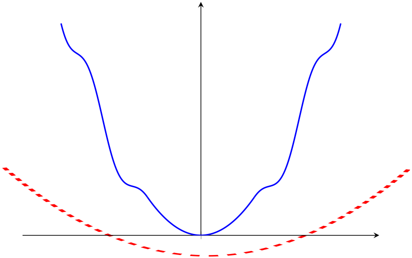
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$$

Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

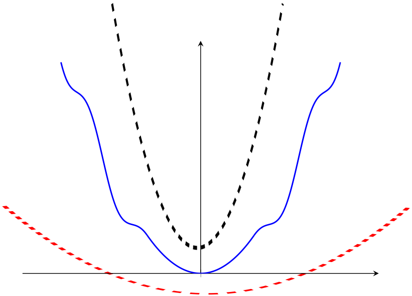
Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

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

Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

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

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

Gradient descent theory revisited

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

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

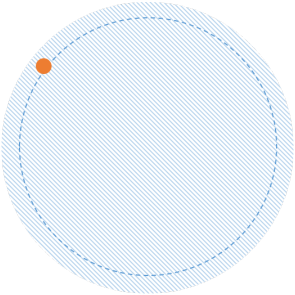
ℓ_2 **error contraction:** GD with $\eta = 1/\beta$ obeys

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

Gradient descent theory revisited

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

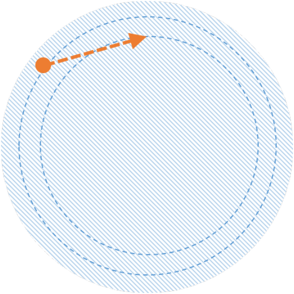
region of local strong convexity + smoothness



Gradient descent theory revisited

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

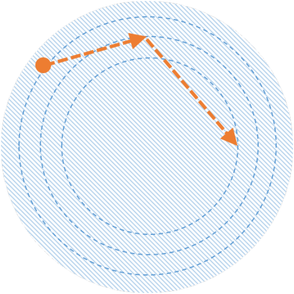
region of local strong convexity + smoothness



Gradient descent theory revisited

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

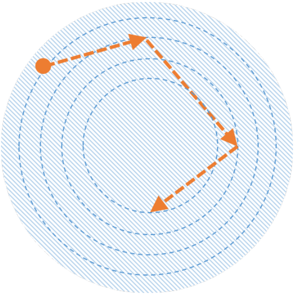
 region of local strong convexity + smoothness



Gradient descent theory revisited

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

 region of local strong convexity + smoothness



Gradient descent theory revisited

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

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

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

- Condition number β/α determines rate of convergence

Gradient descent theory revisited

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

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

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

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

What does this optimization theory say about WF?

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

What does this optimization theory say about WF?

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

Population level (infinite samples)

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

Consequence: WF converges within $O\left(\log\frac{1}{\varepsilon}\right)$ iterations if $m \to \infty$

What does this optimization theory say about WF?

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

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

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

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

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

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 $O(n \log \frac{1}{\varepsilon})$ iterations if $m \asymp n \log n$

What does this optimization theory say about WF?

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

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

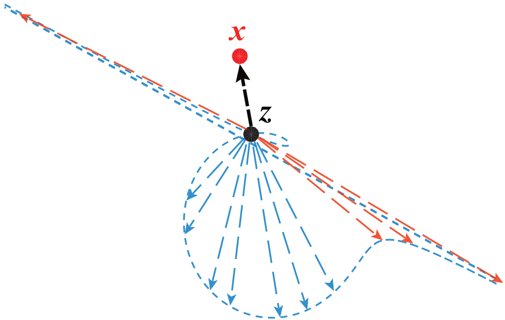
$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp n} \text{ (even locally)}$$

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 $\eta_t = O(1/n)$

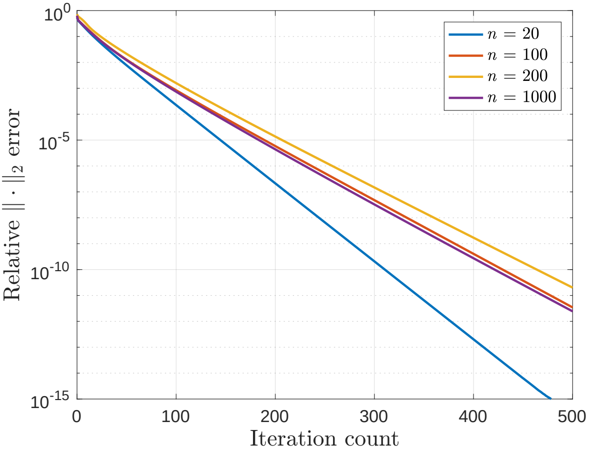
$\Uparrow$

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?

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

A second look at gradient descent theory

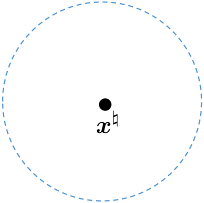
Which region enjoys both strong convexity and smoothness?

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

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

A second look at gradient descent theory

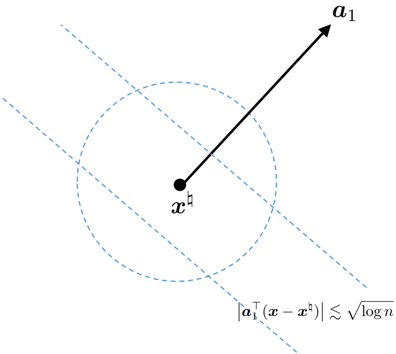
Which region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ is not far away from $\boldsymbol{x}^\natural$

A second look at gradient descent theory

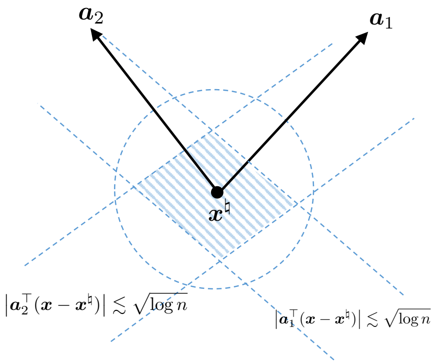
Which region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ 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

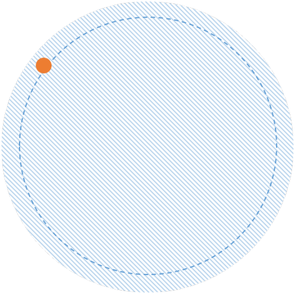
Which region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ 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

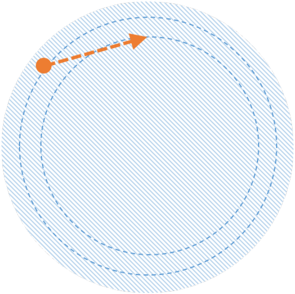
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

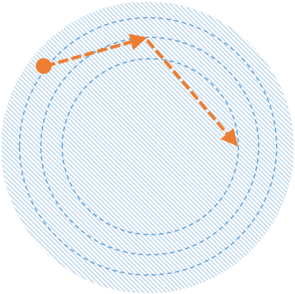
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

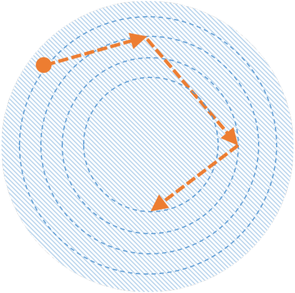
region of local strong convexity + smoothness



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

A second look at gradient descent theory

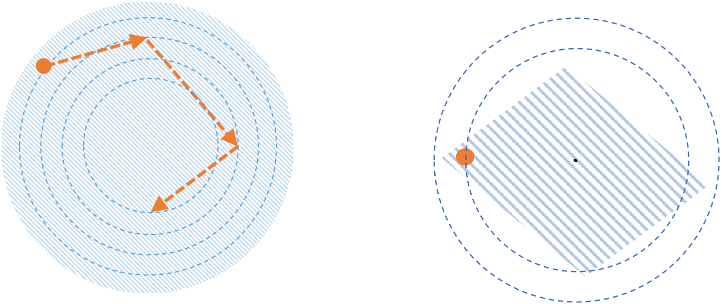
region of local strong convexity + smoothness



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

A second look at gradient descent theory

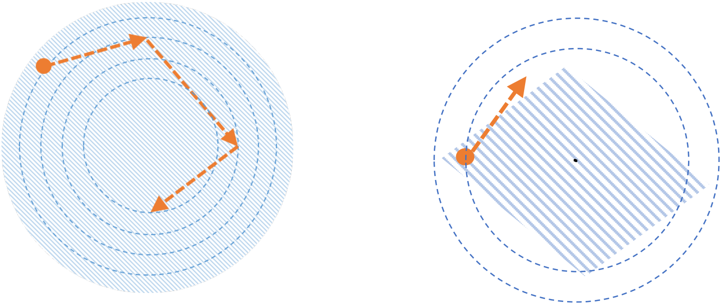
region of local strong convexity + smoothness



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

A second look at gradient descent theory

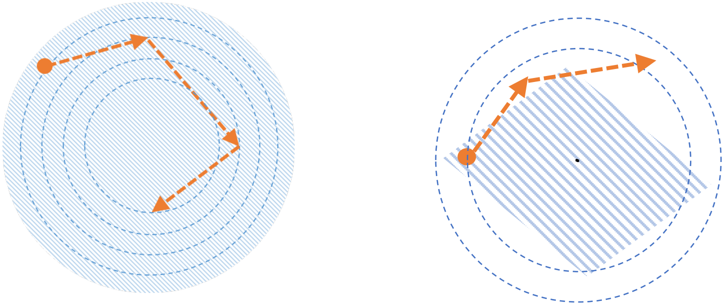
region of local strong convexity + smoothness



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

A second look at gradient descent theory

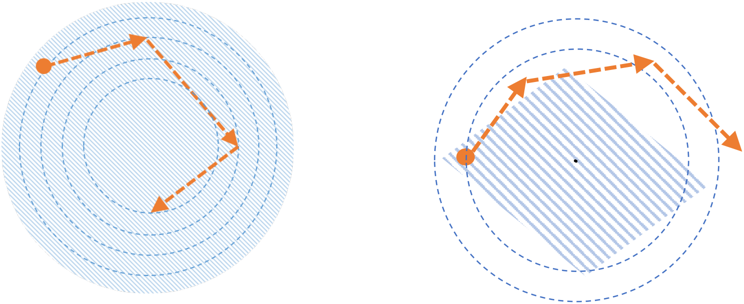
region of local strong convexity + smoothness



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

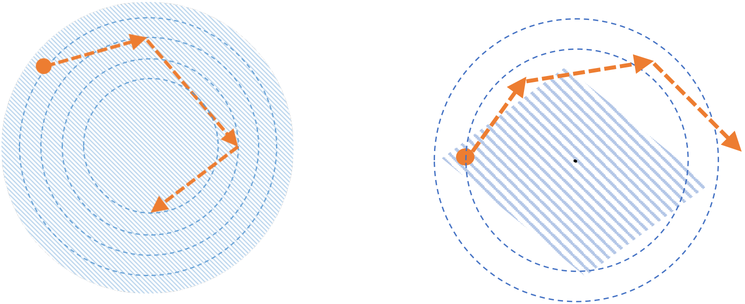
A second look at gradient descent theory

 region of local strong convexity + smoothness



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

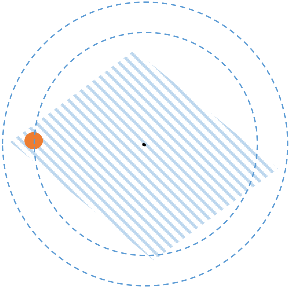
A second look at gradient descent theory



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

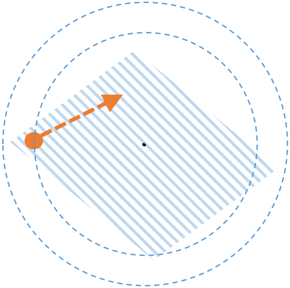
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



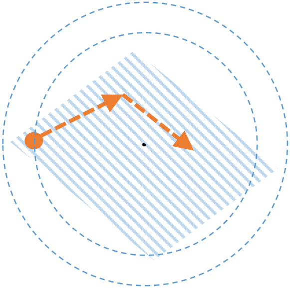
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



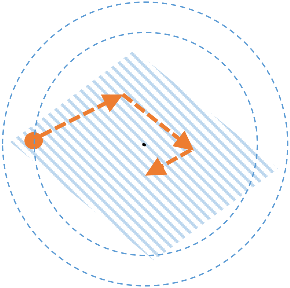
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness

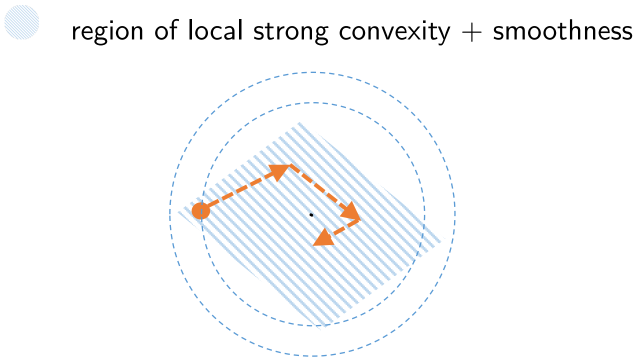


Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



Our findings: GD is implicitly regularized



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* $m \gtrsim n \log n$.

- Step size: $\frac{1}{\log n}$ (vs. $\frac{1}{n}$)

Theoretical guarantees

Theorem 1 (Phase retrieval)

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

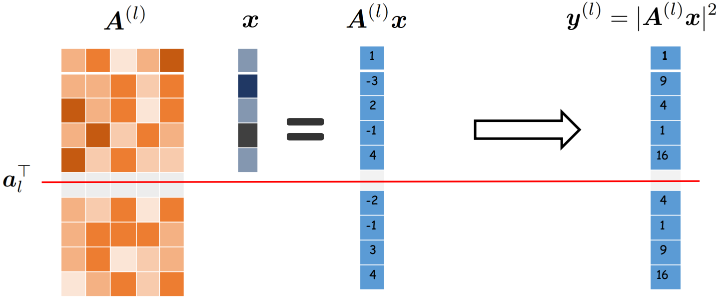
- $\max_k \left|\boldsymbol{a}_k^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\natural\|_2$ *(incoherence)*
- $\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2 \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$ *(near-linear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

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

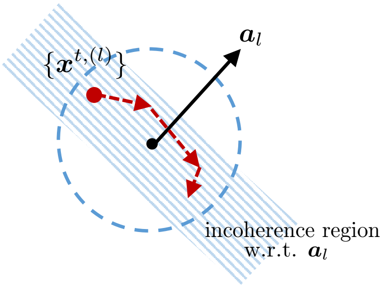
- Step size: $\frac{1}{\log n}$ (vs. $\frac{1}{n}$)
- Computational complexity: $\frac{n}{\log n}$ times faster than existing theory

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

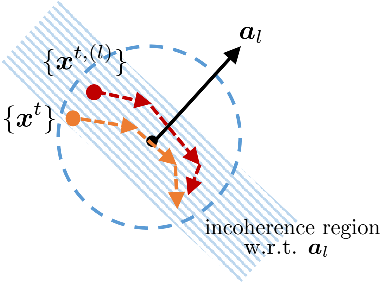


Key ingredient: leave-one-out analysis



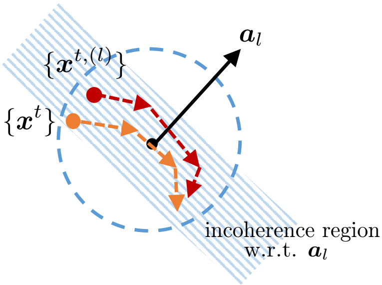
- Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.

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 $\boldsymbol{x}^t$

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 $\boldsymbol{x}^t$
- $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \leq \left|\boldsymbol{a}_l^\top(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\natural)\right| + \left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)})\right|$

This recipe is quite general

Low-rank matrix completion

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

Fig. credit: Candès

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

Prior art

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

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)
    - e.g. 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)
  - 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

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 $\underbrace{\|\cdot\|_{2,\infty}}_{\text{incoherence}}$*

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

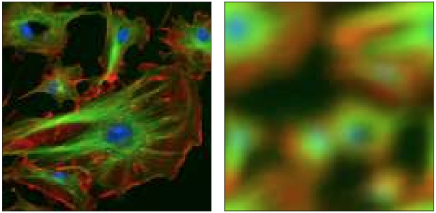


Fig. credit: Romberg

multipath in wireless comm

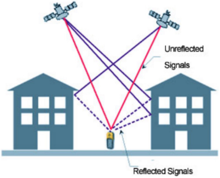


Fig. credit: EngineeringsALL

Reconstruct two signals from their convolution; equivalently,

$$\text{find} \quad \boldsymbol{h}, \boldsymbol{x} \in \mathbb{C}^n \qquad \text{s.t.} \quad \boldsymbol{b}_k^* \boldsymbol{h} \boldsymbol{x}^* \boldsymbol{a}_k = y_k, \quad 1 \leq k \leq m$$

Prior art

$$\text{minimize}_{\boldsymbol{x},\boldsymbol{h}} \quad f(\boldsymbol{x},\boldsymbol{h}) = \sum_{k=1}^{m} \left| \boldsymbol{b}_k^* \left(\boldsymbol{h}\boldsymbol{x}^* - \boldsymbol{h}^\natural \boldsymbol{x}^{\natural *} \right) \boldsymbol{a}_k \right|^2$$

$\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$ and $\{\boldsymbol{b}_k\}$: partial Fourier basis

Prior art

$$\text{minimize}_{\boldsymbol{x},\boldsymbol{h}} \quad f(\boldsymbol{x},\boldsymbol{h}) = \sum_{k=1}^{m} \left| \boldsymbol{b}_k^* \left(\boldsymbol{h}\boldsymbol{x}^* - \boldsymbol{h}^\natural \boldsymbol{x}^{\natural *} \right) \boldsymbol{a}_k \right|^2$$

$$\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \qquad \text{and} \qquad \{\boldsymbol{b}_k\} : \text{ partial Fourier basis}$$

Existing theory on gradient descent requires

- regularized loss + projection
  - e.g. Li, Ling, Strohmer, Wei '16, Huang, Hand '17, Ling, Strohmer '17

Prior art

$$\text{minimize}_{\boldsymbol{x},\boldsymbol{h}} \quad f(\boldsymbol{x},\boldsymbol{h}) = \sum_{k=1}^{m} \left| \boldsymbol{b}_k^* \left(\boldsymbol{h}\boldsymbol{x}^* - \boldsymbol{h}^\natural \boldsymbol{x}^{\natural *} \right) \boldsymbol{a}_k \right|^2$$

$$\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \qquad \text{and} \qquad \{\boldsymbol{b}_k\} : \text{ partial Fourier basis}$$

Existing theory on gradient descent requires

- regularized loss + projection
  - e.g. Li, Ling, Strohmer, Wei '16, Huang, Hand '17, Ling, Strohmer '17
  - requires m iterations even with regularization

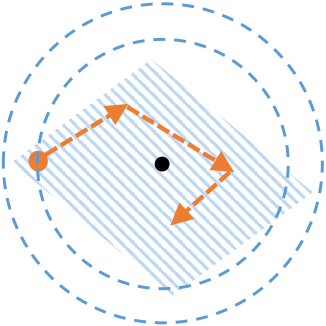
Theoretical guarantees

Theorem 3 (Blind deconvolution)

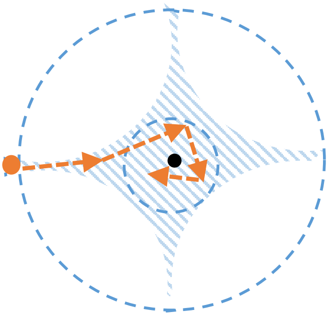
Suppose $\boldsymbol{h}^\natural$ is incoherent w.r.t. $\{\boldsymbol{b}_k\}$. Vanilla gradient descent (with spectral initialization) achieves ε accuracy in $O\big(\log\frac{1}{\varepsilon}\big)$ iterations, provided that step size $\eta \lesssim 1$ and sample size $m \gtrsim n \mathrm{poly}\log(m)$.

- Regularization-free
- Converges in $O\big(\log\frac{1}{\varepsilon}\big)$ iterations (vs. $O\big(m\log\frac{1}{\varepsilon}\big)$ iterations in prior theory)

Incoherence region in high dimensions



2-dimensional

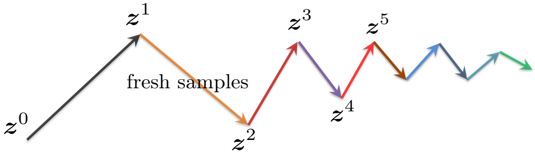


high-dimensional (mental representation)

incoherence region is vanishingly small

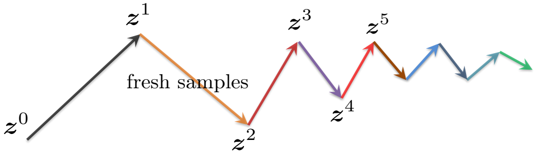
Complicated dependencies across iterations

- Several prior sample-splitting approaches: require fresh samples at each iteration; not what we actually run in practice

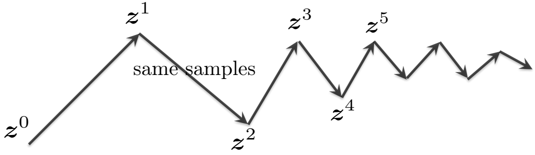


Complicated dependencies across iterations

- Several prior sample-splitting approaches: require fresh samples at each iteration; not what we actually run in practice



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



Summary

- **Implicit regularization:** vanilla gradient descent automatically forces iterates to stay *incoherent*

Summary

- **Implicit regularization:** vanilla gradient descent automatically forces 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