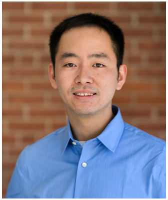
Random Initialization and Implicit Regularization in Nonconvex Statistical Estimation

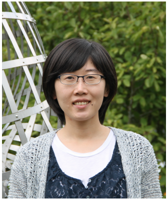
Cong Ma

ORFE, Princeton University

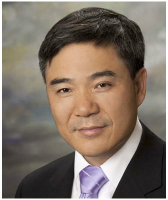


Yuxin Chen
Princeton EE

Kaizheng Wang
Princeton ORFE



Yuejie Chi
CMU ECE

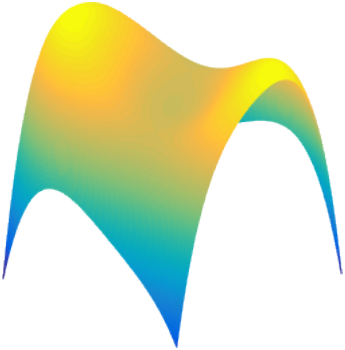


Jianqing Fan
Princeton ORFE

Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x};\boldsymbol{y}) \quad \rightarrow \quad \text{loss function may be nonconvex}$$

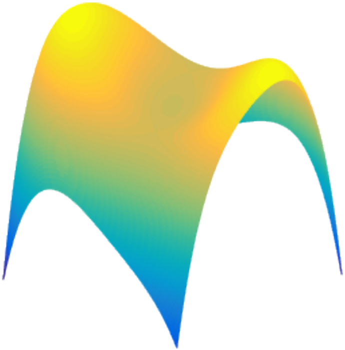


Nonconvex estimation problems are everywhere

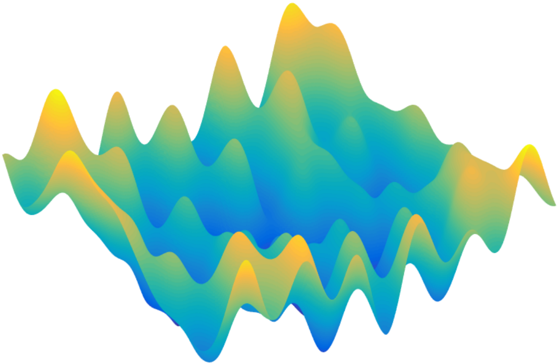
Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}; \boldsymbol{y}) \quad \rightarrow \quad \text{loss function may be nonconvex}$$

- low-rank matrix completion
- blind deconvolution
- 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)

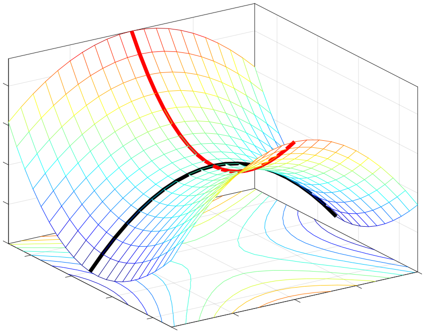
Nonconvex optimization may be super scary

There may be bumps everywhere and exponentially many local optima

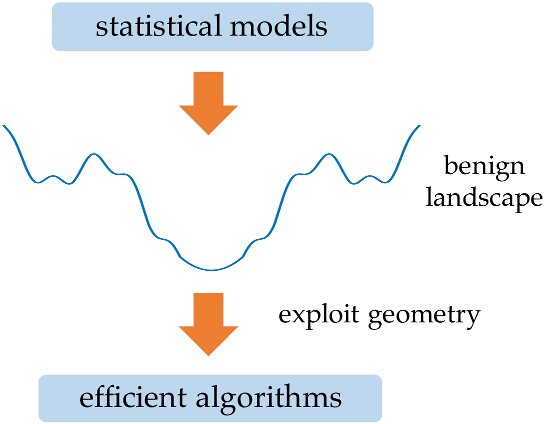
e.g. 1-layer neural net (Auer, Herbster, Warmuth '96; Vu '98)

… but is sometimes much nicer than we think

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



… but is sometimes much nicer than we think



Even simplest possible nonconvex methods
might be remarkably efficient under suitable statistical models

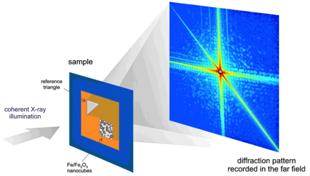
This talk: a case study — phase retrieval

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



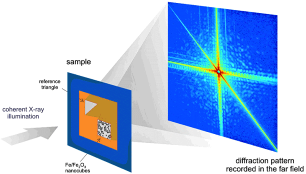
intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Missing phase problem

Detectors record intensities of diffracted rays

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

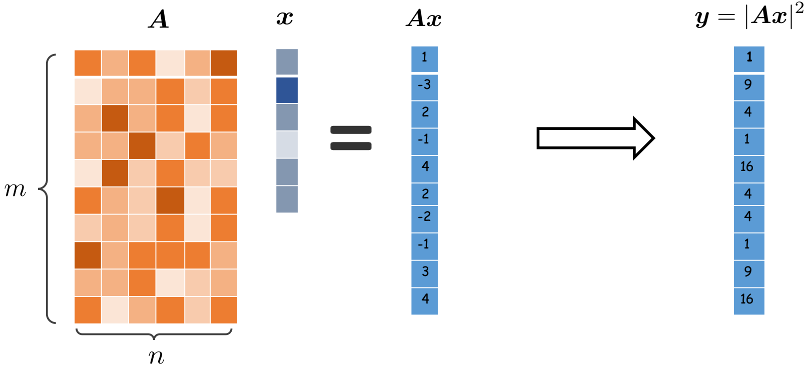
Fig credit: Stanford SLAC



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

$$y_k = |a_k^\top x^\natural|^2, \qquad k = 1, \ldots, m$$

assume w.l.o.g. $\|x^\natural\|_2 = 1$

A natural least squares formulation

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

$$\Downarrow$$

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

A natural least squares formulation

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

$$\Downarrow$$

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

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

A natural least squares formulation

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

$$\Downarrow$$

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

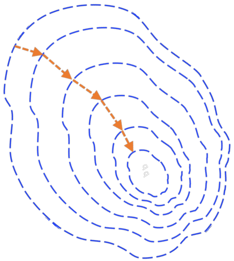
- **pros:** often exact as long as sample size is sufficiently large
- **cons:** $f(\cdot)$ is highly nonconvex
  $\longrightarrow$ *computationally challenging!*

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

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

Wirtinger flow (Candès, Li, Soltanolkotabi ’14)

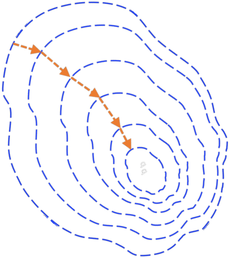
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



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

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

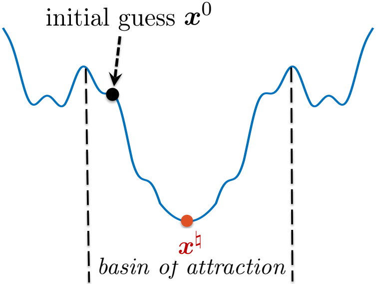
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



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

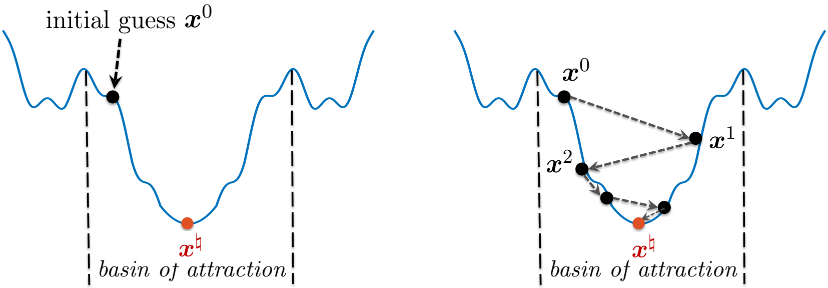
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta\,\nabla f(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

Rationale of two-stage approach



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

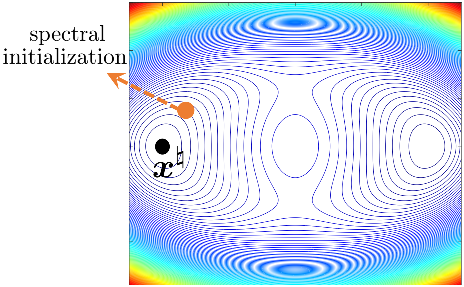
Rationale of two-stage approach



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

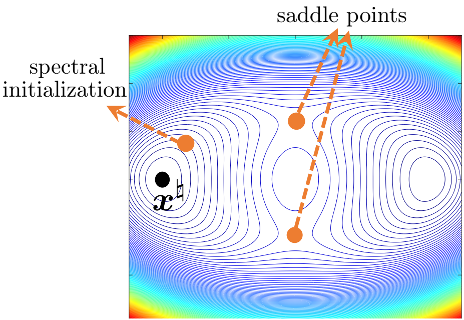
Is carefully-designed initialization necessary for fast convergence?

Initialization



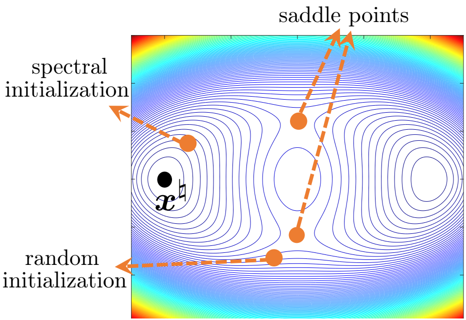
- spectral initialization gets us reasonably close to truth

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

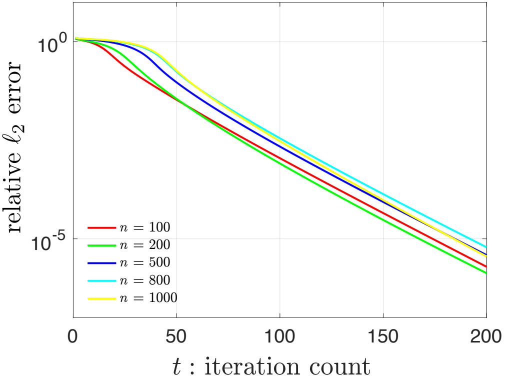


- spectral initialization gets us reasonably close to truth
- cannot initialize GD from anywhere, e.g. it might get stucked at local stationary points (e.g. saddle points)

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

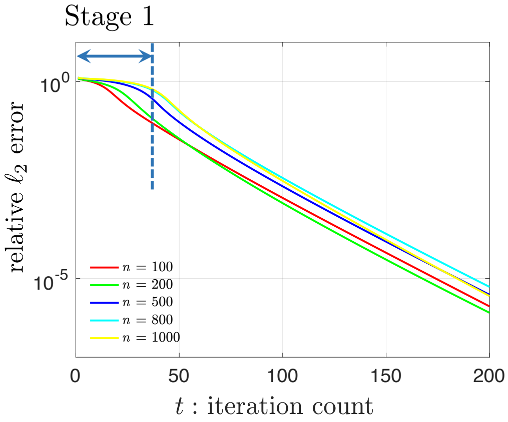
Numerical efficiency of randomly initialized GD

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



Numerical efficiency of randomly initialized GD

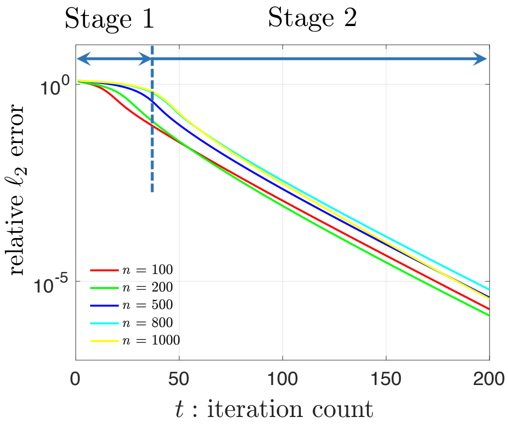
$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



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

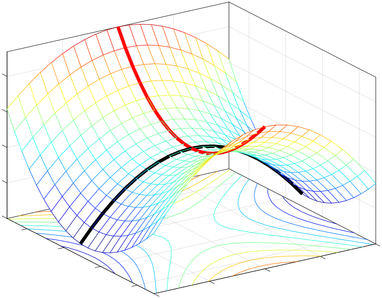
Numerical efficiency of randomly initialized GD

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



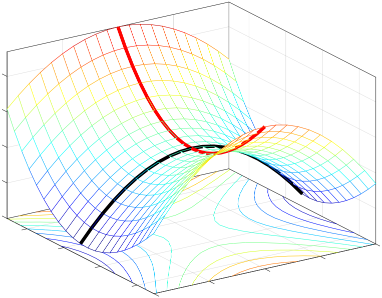
Randomly initialized GD enters local basin within **a few iterations**

What does prior theory say?



- no spurious local mins (Sun et al. '16)

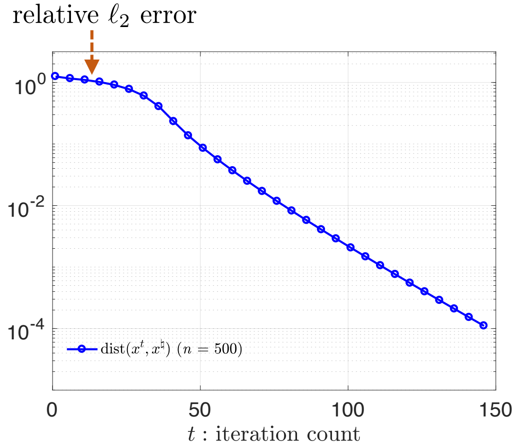
What does prior theory say?



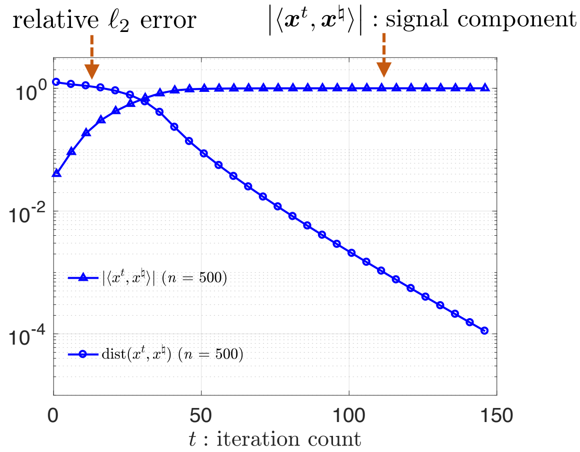
- no spurious local mins (Sun et al. '16)
- GD with random initialization converges to global min almost surely (Lee et al. '16)

No convergence rate guarantees for vanilla GD!

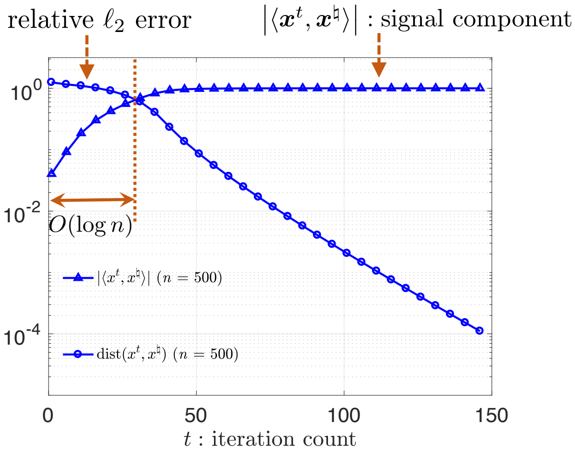
Exponential growth of signal strength in Stage 1



Exponential growth of signal strength in Stage 1

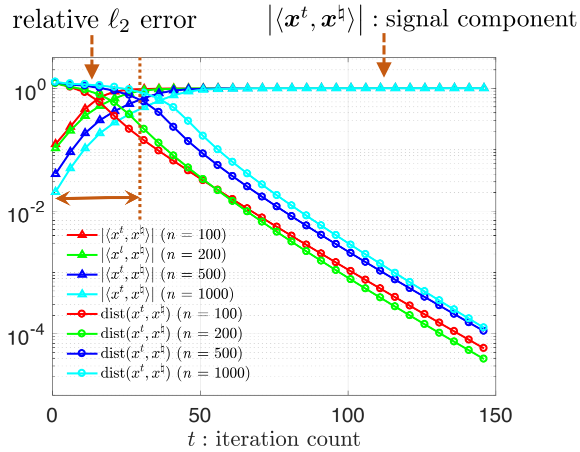


Exponential growth of signal strength in Stage 1



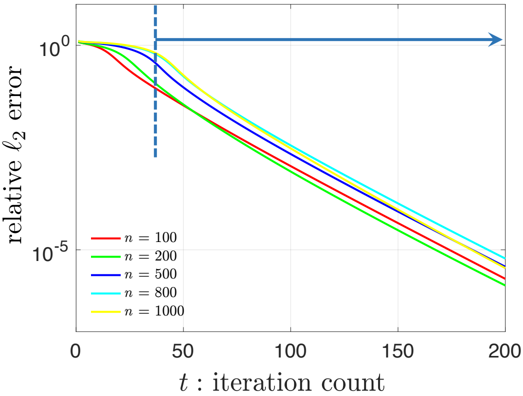
Numerically, $O(\log n)$ iterations are enough to enter local region

Exponential growth of signal strength in Stage 1

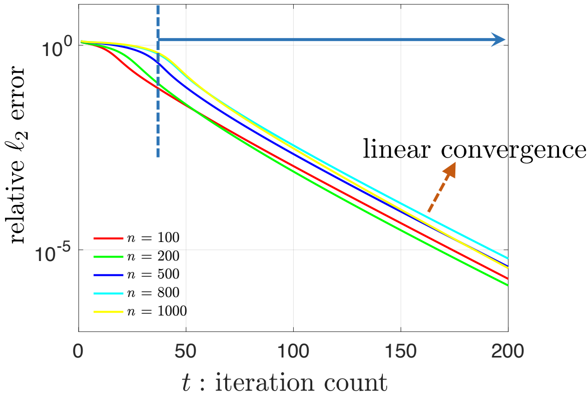


Numerically, $O(\log n)$ iterations are enough to enter local region

Linear / geometric convergence in Stage 2



Linear / geometric convergence in Stage 2



Numerically, GD converges linearly within local region

Experiments on images

- coded diffraction patterns
- $\boldsymbol{x}^{\natural} \in \mathbb{R}^{256 \times 256}$
- $m/n = 12$

GD with random initialization

$\boldsymbol{x}^t$	$\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural$	$\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural$
GD iterate	signal component	perpendicular component

use Adobe to view the animation

Exponential growth of "signal-to-noise" ratio

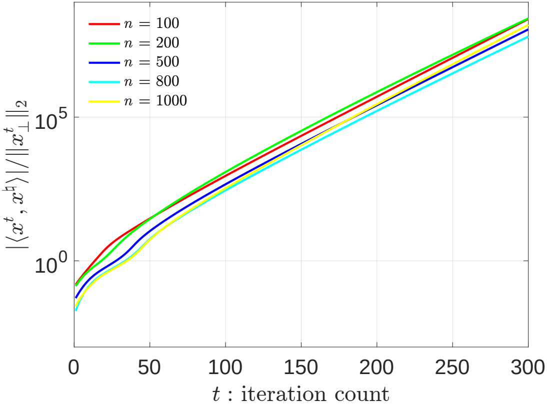
$$\frac{|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle|}{} \quad \rightarrow \quad \text{signal component}$$

Exponential growth of “signal-to-noise” ratio

$$\frac{\left|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle\right|}{\big\|\underbrace{\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural}_{:=\boldsymbol{x}_\perp^t}\big\|_2} \quad \begin{array}{l} \rightarrow \quad \text{signal component} \\ \rightarrow \quad \text{residual component} \end{array}$$

Exponential growth of "signal-to-noise" ratio

$$\frac{\left|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural\rangle\right|}{\big\|\underbrace{\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural\rangle \boldsymbol{x}^\natural}_{:=\boldsymbol{x}^t_\perp}\big\|_2} \quad \begin{array}{l} \rightarrow \quad \text{signal component} \\ \rightarrow \quad \text{residual component} \end{array}$$



Theoretical guarantees

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

Theorem 1 (Chen, Chi, Fan, Ma '18)

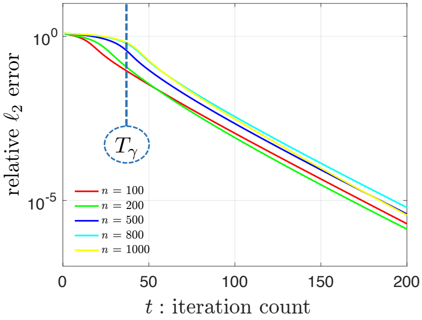
Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$ achieves

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

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

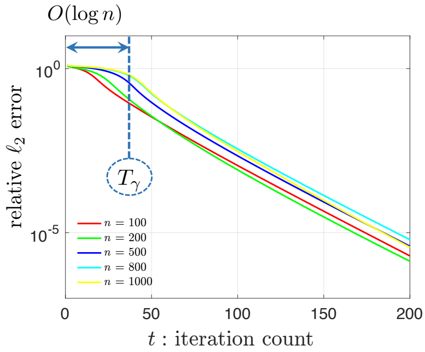
Theoretical guarantees

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



Theoretical guarantees

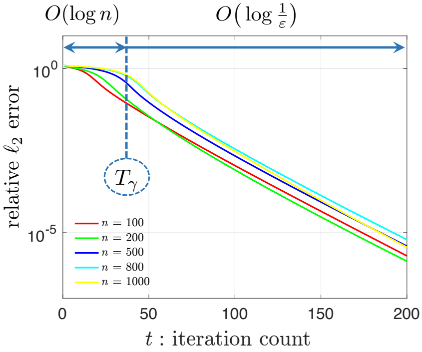
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \ge T_\gamma \asymp \log n$$



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

Theoretical guarantees

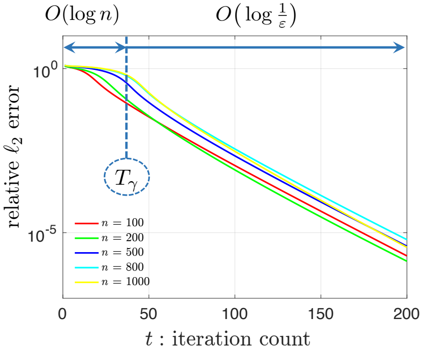
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma (1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^\natural\|_2, \quad t \ge T_\gamma \asymp \log n$$



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

Theoretical guarantees

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$

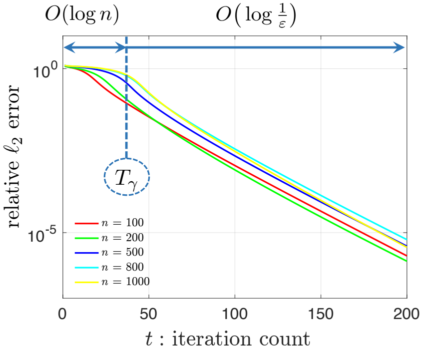


- *near-optimal compututational cost*:

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

Theoretical guarantees

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



- *near-optimal compututational cost*:
  — $O\left(\log n + \log \frac{1}{\varepsilon}\right)$ iterations to yield ε accuracy
- *near-optimal sample size*: $m \gtrsim n\mathrm{poly}\log m$

Comparison with prior theory

Iteration complexity:

	prior theory	**our theory**
Stage 1: random init $\rightarrow$ local region	almost surely (Lee et al. '16)	$O(\log n)$
Stage 2: local refinement		

Comparison with prior theory

Iteration complexity:

	prior theory	**our theory**
Stage 1: random init $\rightarrow$ local region	almost surely (Lee et al. '16)	$O(\log n)$
Stage 2: local refinement	$O\big(n \log \frac{1}{\varepsilon}\big)$ (Candes et al. '14)	$O\big(\log \frac{1}{\varepsilon}\big)$

Stage 1: random initialization $\rightarrow$ local region

What if we have infinite samples?

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

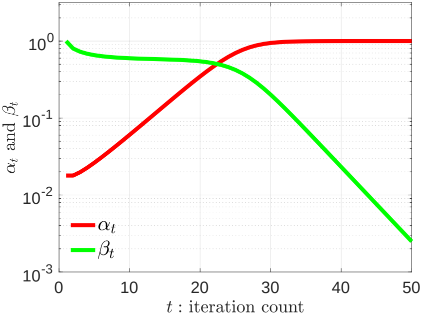
Population level (infinite samples)

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

where

$$\nabla F(\boldsymbol{x}) := \mathbb{E}[\nabla f(\boldsymbol{x})] = (3\|\boldsymbol{x}\|_2^2 - 1)\boldsymbol{x} - 2(\boldsymbol{x}^{\natural\top}\boldsymbol{x})\boldsymbol{x}^{\natural}$$

Population-level state evolution



Let $\alpha_t := \underbrace{|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle|}_{\text{signal strength}}$ and $\beta_t = \underbrace{\|\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural\|_2}_{\text{size of residual component}}$, then

$$\alpha_{t+1} = \{1 + 3\eta[1 - (\alpha_t^2 + \beta_t^2)]\}\alpha_t$$
$$\beta_{t+1} = \{1 + \eta[1 - 3(\alpha_t^2 + \beta_t^2)]\}\beta_t$$

2-parameter dynamics

Back to finite-sample analysis

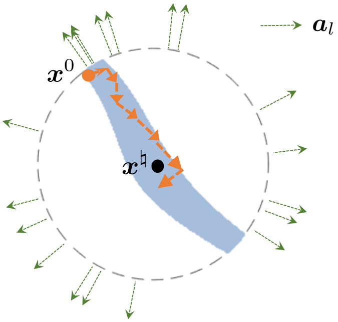
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$$

Back to finite-sample analysis

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t) - \eta \underbrace{\big(\nabla f(\boldsymbol{x}^t) - \nabla F(\boldsymbol{x}^t)\big)}_{:=\boldsymbol{r}(\boldsymbol{x}^t)}$$

Back to finite-sample analysis

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t) - \eta \underbrace{\big(\nabla f(\boldsymbol{x}^t) - \nabla F(\boldsymbol{x}^t)\big)}_{:= \boldsymbol{r}(\boldsymbol{x}^t)}$$

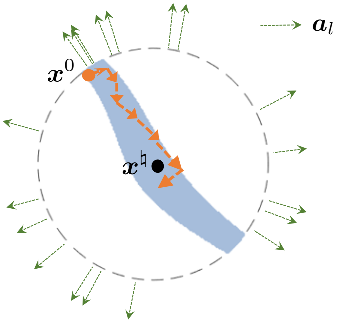


a region with well-controlled $\boldsymbol{r}(\boldsymbol{x})$

- population-level analysis holds *approximately* if $\boldsymbol{r}(\boldsymbol{x}^t) \ll \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t)$

Back to finite-sample analysis

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t) - \eta \underbrace{\big(\nabla f(\boldsymbol{x}^t) - \nabla F(\boldsymbol{x}^t)\big)}_{:=\boldsymbol{r}(\boldsymbol{x}^t)}$$

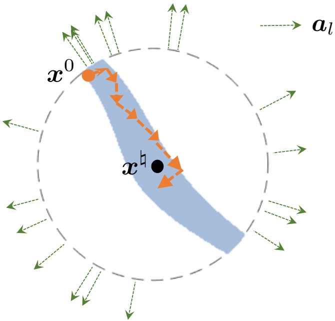


a region with well-controlled $\boldsymbol{r}(\boldsymbol{x})$

- population-level analysis holds *approximately* if $\boldsymbol{r}(\boldsymbol{x}^t) \ll \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t)$
- $\boldsymbol{r}(\boldsymbol{x}^t)$ is well-controlled if $\boldsymbol{x}^t$ is independent of $\{\boldsymbol{a}_k\}$

Back to finite-sample analysis

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta \nabla f(\boldsymbol{x}^{t}) = \boldsymbol{x}^{t} - \eta \nabla F(\boldsymbol{x}^{t}) - \eta \underbrace{\left(\nabla f(\boldsymbol{x}^{t}) - \nabla F(\boldsymbol{x}^{t})\right)}_{:=\boldsymbol{r}(\boldsymbol{x}^{t})}$$

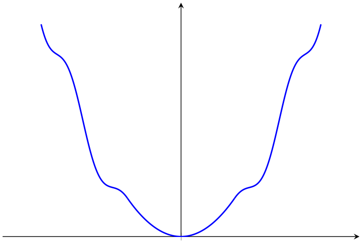


a region with well-controlled $\boldsymbol{r}(\boldsymbol{x})$

- population-level analysis holds *approximately* if $\boldsymbol{r}(\boldsymbol{x}^t) \ll \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t)$
- $\boldsymbol{r}(\boldsymbol{x}^t)$ is well-controlled if $\boldsymbol{x}^t$ is independent of $\{\boldsymbol{a}_k\}$
- **key analysis ingredient:** show $\boldsymbol{x}^t$ is "nearly-independent" of each $\boldsymbol{a}_k$

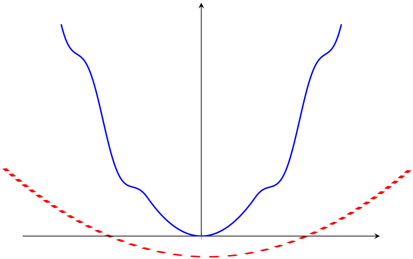
Stage 2: local refinement

Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

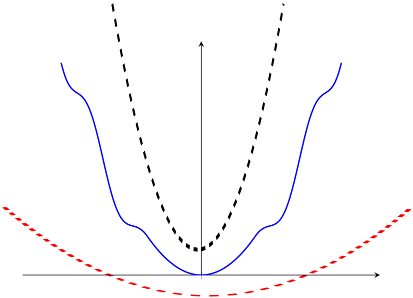
Gradient descent theory revisited



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

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

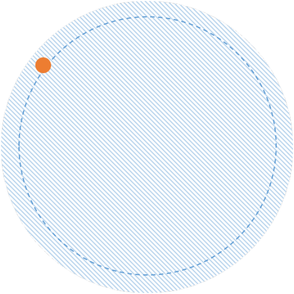
ℓ_2 **error contraction:** GD with $\eta = 1/\beta$ obeys

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

Gradient descent theory revisited

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

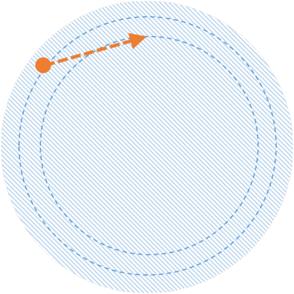
 region of local strong convexity + smoothness



Gradient descent theory revisited

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

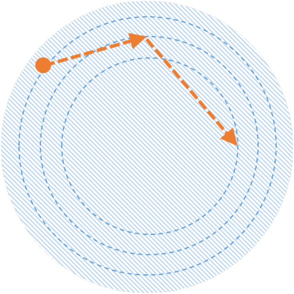
 region of local strong convexity + smoothness



Gradient descent theory revisited

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

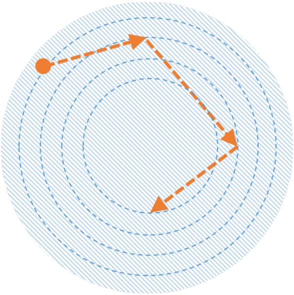
 region of local strong convexity + smoothness



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

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

- 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?

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

Population level (infinite samples)

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

Consequence: Given good initialization, WF converges within $O\big(\log \frac{1}{\varepsilon}\big)$ 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?

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

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

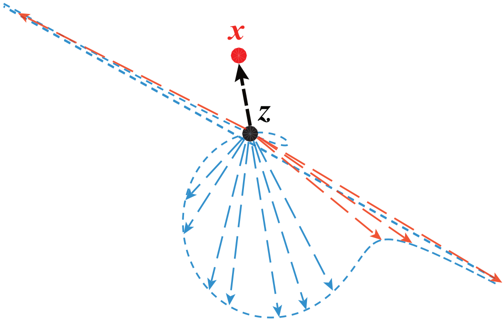
$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp n} \text{ (even locally)}$$

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

Too slow ... can we accelerate it?

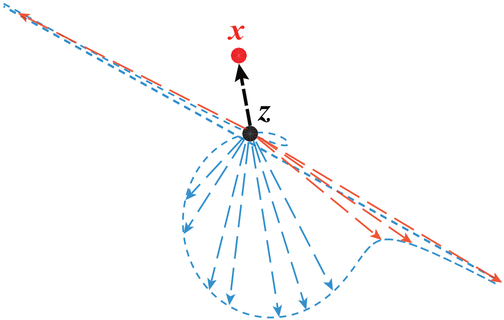
Improvement: truncated WF (Chen, Candès '15)

Regularize / trim gradient components to accelerate convergence



Improvement: truncated WF (Chen, Candès ’15)

Regularize / trim gradient components to accelerate convergence



But it still needs certain spectral initialization ...

Recall

WF converges in $O(n)$ iterations

Recall

WF converges in $O(n)$ iterations

$$\Uparrow$$

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

Recall

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

Recall

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

Recall

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?

A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

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

A second look at gradient descent theory

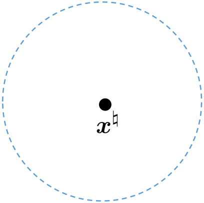
Which region enjoys both strong convexity and smoothness?

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

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

A second look at gradient descent theory

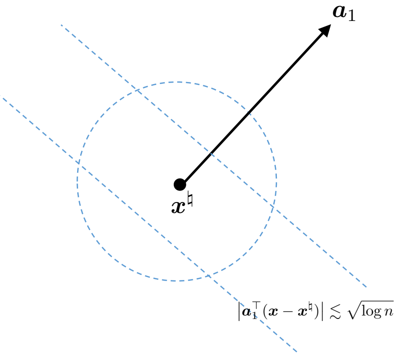
Which region enjoys both strong convexity and smoothness?



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

A second look at gradient descent theory

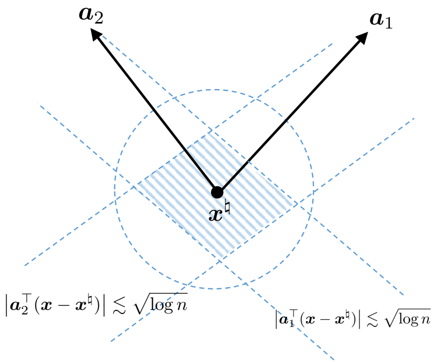
Which region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ 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

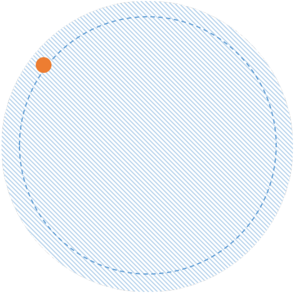
Which region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ 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

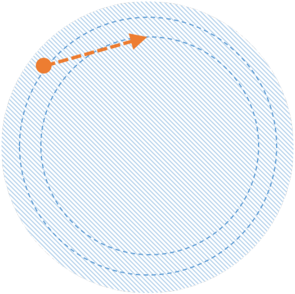
region of local strong convexity + smoothness



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

A second look at gradient descent theory

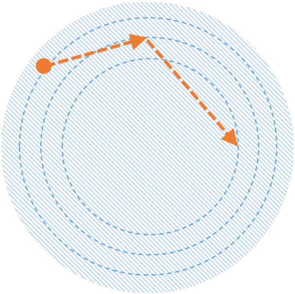
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

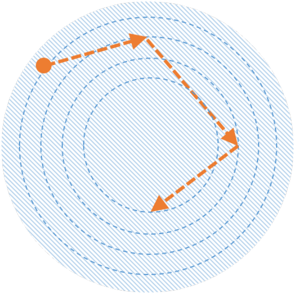
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

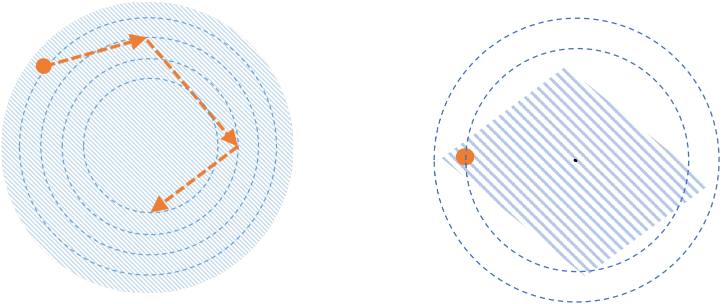
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

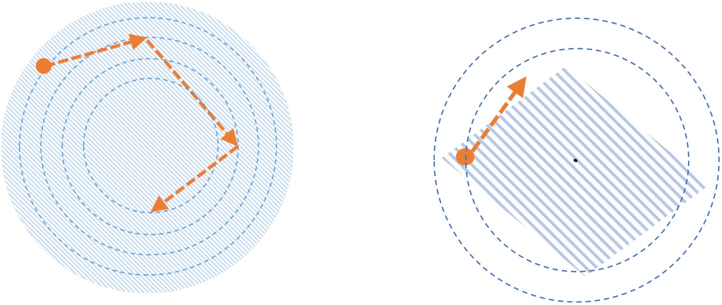
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

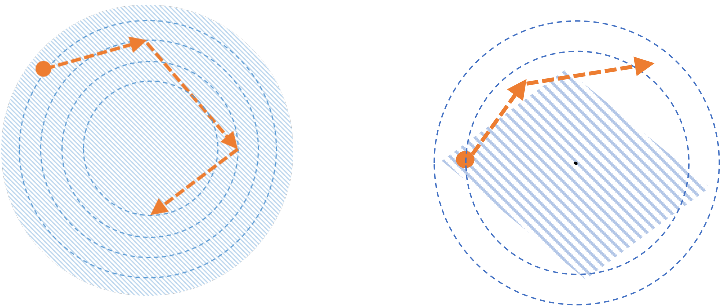
region of local strong convexity + smoothness



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

A second look at gradient descent theory

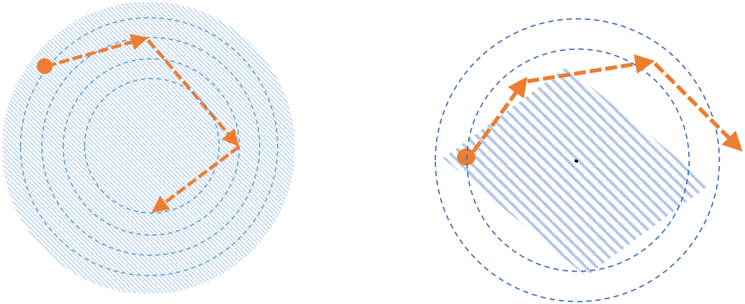
region of local strong convexity + smoothness



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

A second look at gradient descent theory

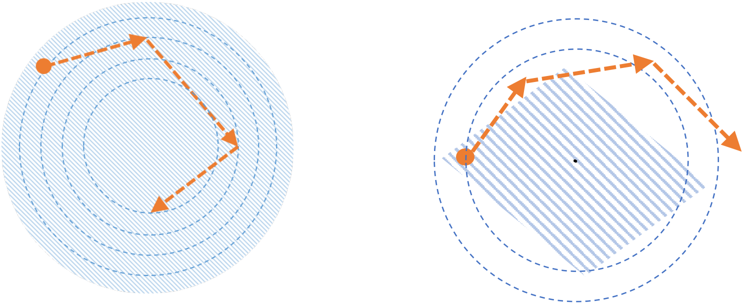
 region of local strong convexity + smoothness



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

A second look at gradient descent theory

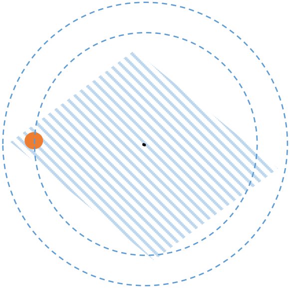
region of local strong convexity + smoothness



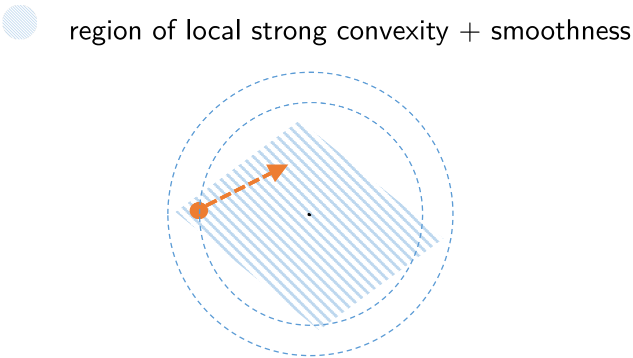
- Prior theory only ensures that iterates remain in ℓ_2 ball but not incoherence region
- *Prior theory enforces regularization to promote incoherence*

Our findings: GD is implicitly regularized

region of local strong convexity + smoothness

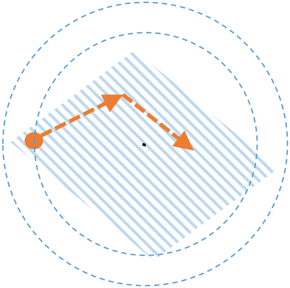


Our findings: GD is implicitly regularized



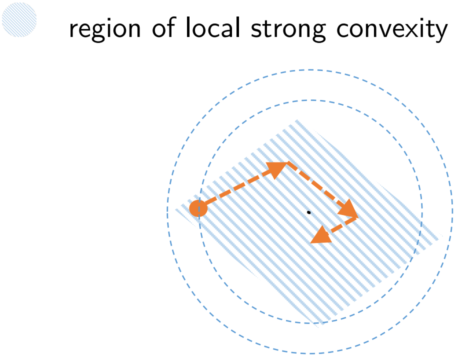
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



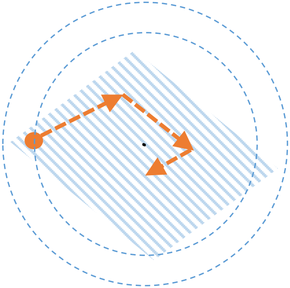
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent

Theoretical guarantees for Stage 2

Theorem 2 (Phase retrieval)

Under i.i.d. Gaussian design, GD with random initialization achieves for $t \geq T_\gamma + 1$

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

Theoretical guarantees for Stage 2

Theorem 2 (Phase retrieval)

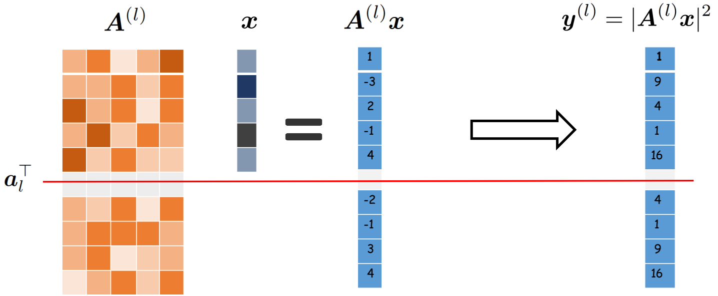
Under i.i.d. Gaussian design, GD with random initialization achieves for $t \geq T_\gamma + 1$

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

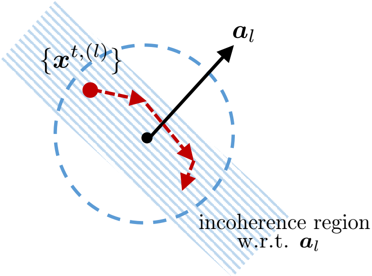
provided that step size $\eta \asymp c$ *and sample size* $m \gtrsim n\, \mathrm{poly} \log m$.

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

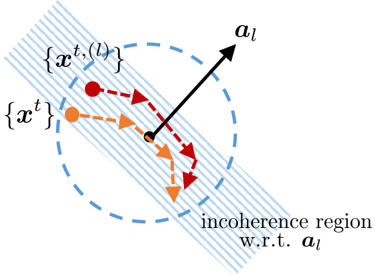


Key ingredient: leave-one-out analysis



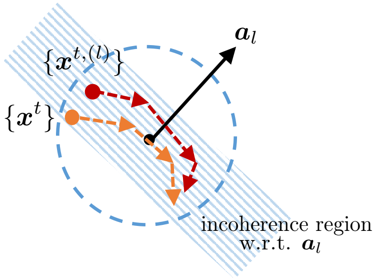
- Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.

Key ingredient: leave-one-out analysis



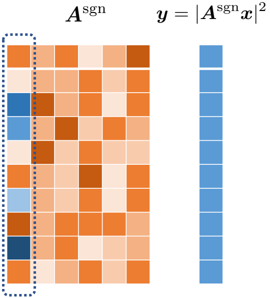
- Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.
- Leave-one-out iterates $\boldsymbol{x}^{t,(l)} \;\approx\;$ true iterates $\boldsymbol{x}^t$

Key ingredient: leave-one-out analysis

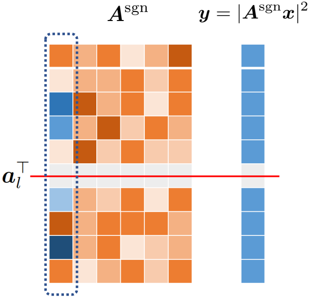


- Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.
- Leave-one-out iterates $\boldsymbol{x}^{t,(l)} \;\approx\;$ true iterates $\boldsymbol{x}^t$
- $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \leq \left|\boldsymbol{a}_l^\top(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\natural)\right| + \left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)})\right|$

Other leave-one-out sequences

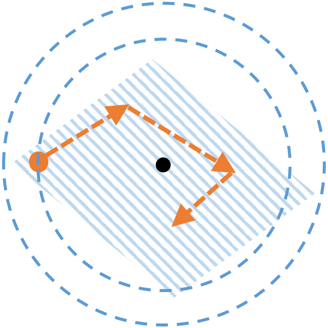


$\boldsymbol{x}^{t,\mathsf{sgn}}$: indep. of sign info of $\{a_{i,1}\}$

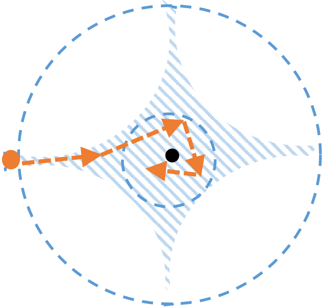


$\boldsymbol{x}^{t,\mathsf{sgn},(l)}$: indep. of both sign info of $\{a_{i,1}\}$ and $\boldsymbol{a}_l$

Incoherence region in high dimensions



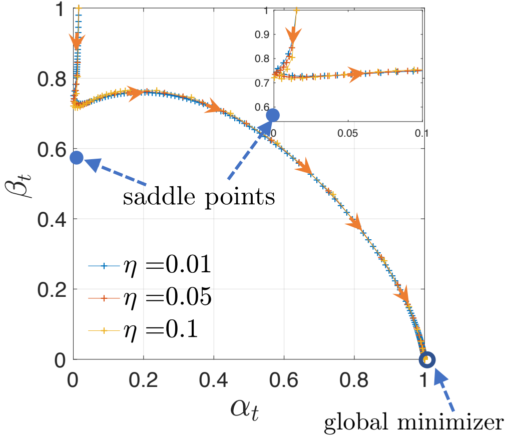
2-dimensional



high-dimensional (mental representation)

incoherence region is vanishingly small

Saddle-escaping schemes?



Randomly initialized GD never hits saddle points in phase retrieval!

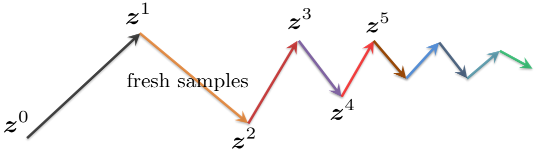
Other saddle-escaping schemes

	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 (ours) (Chen et al. '18)	$\log n + \log\frac{1}{\varepsilon}$	$\log n$	$\log\frac{1}{\varepsilon}$

Generic optimization theory yields highly suboptimal convergence guarantees

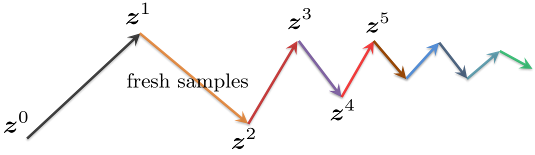
No need of sample splitting

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

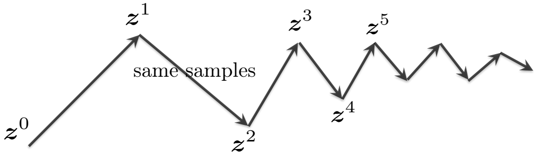


No need of sample splitting

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



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



Summary

- **Blessings of statistical models:** GD with random initialization converges fast
- **Implict regularization:** vanilla gradient descent automatically foces iterates to stay *incoherent*

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

"Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval", Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma arXiv:XXXXXX