# Implicit Regularization in Nonconvex Statistical Estimation


Yuejie Chi

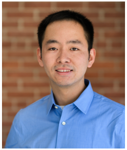
Department of Electrical and Computer Engineering

Carnegie Mellon University

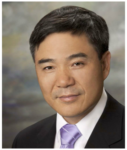

May 2018

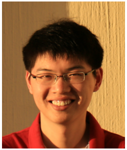
# Acknowledgements

Thanks to my collaborators:











Y. Chen
Princeton

J. Fan
Princeton

C. Ma
Princeton

K. Wang
Princeton

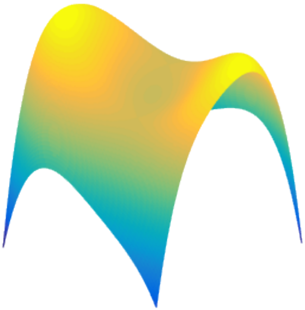
Y. Li
CMU

This research is supported by NSF, ONR and AFOSR.

# Nonconvex problems are abundant

Empirical risk minimization is usually nonconvex

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

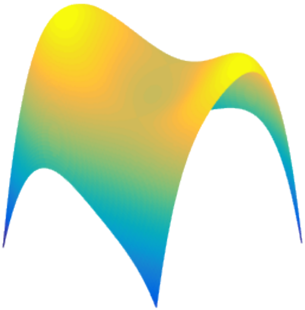


# Nonconvex problems are abundant

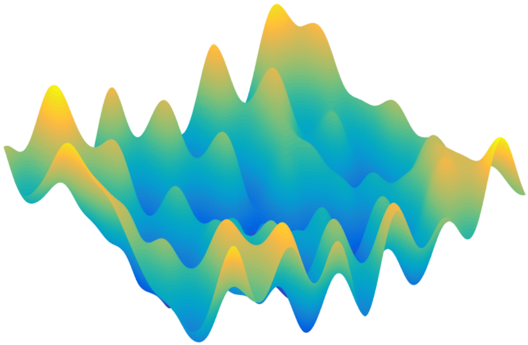
Empirical risk minimization is usually nonconvex

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

- low-rank matrix completion
- phase retrieval
- dictionary learning
- blind deconvolution
- mixture models
- deep learning
- ...



# Nonconvex optimization is daunting in theory



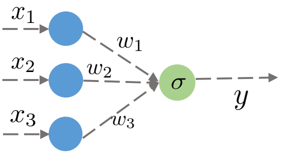
There may be exponentially many local optima

e.g. a single neuron model (Auer, Herbster, Warmuth '96)

# Exponentially many local minima for perceptron

Given training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$,

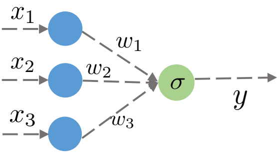
$$\text{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} \quad \ell_n(\boldsymbol{w}) := \frac{1}{2n}\sum_{i=1}^n \Big(y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\Big)^2$$



# Exponentially many local minima for perceptron

Given training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$,

$$\text{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} \quad \ell_n(\boldsymbol{w}) := \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right)^2$$

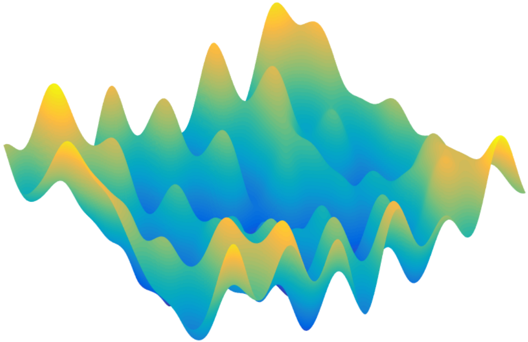



**Theorem (Auer et al., 1995)**

*Let $\sigma(\cdot)$ be sigmoid and $\ell(\cdot)$ be the quadratic loss function. There exists a sequence of training samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ such that $\ell_n(\boldsymbol{w})$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.*

No. of local minima grows exponentially with the dimension $d$!

# Nonconvex optimization is daunting in theory



There may be exponentially many local optima

e.g. a single neuron model (Auer, Herbster, Warmuth '96)

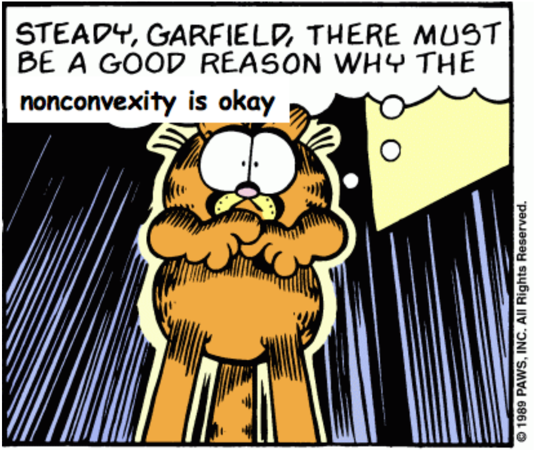
# Nonconvex optimization is daunting in theory

There may be exponentially many local optima

e.g. a single neuron model (Auer, Herbster, Warmuth '96)

# But they're solved on a daily basis in practice

Using simple algorithms such as gradient descent, e.g., "back propagation" for training deep neural networks...



# Statistical models come to rescue

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

$$\text{minimize}_{\boldsymbol{x}} \ f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i; \boldsymbol{x})$$

# Statistical models come to rescue

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

$$\text{minimize}_{\boldsymbol{x}}\ f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i;\boldsymbol{x}) \quad \overset{m\to\infty}{\Longrightarrow} \quad \mathbb{E}[\ell(y;\boldsymbol{x})]$$

# Statistical models come to rescue

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

$$\text{minimize}_{\boldsymbol{x}}\ f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i;\boldsymbol{x}) \quad \overset{m\to\infty}{\Longrightarrow} \quad \mathbb{E}[\ell(y;\boldsymbol{x})]$$

empirical risk $\approx$ population risk (often nice!)

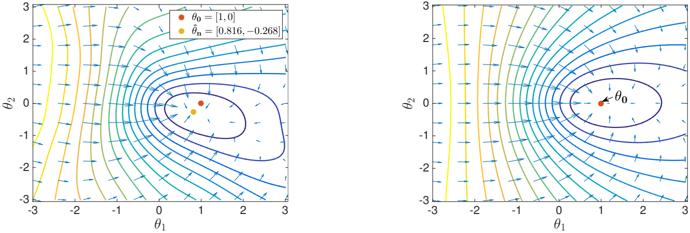
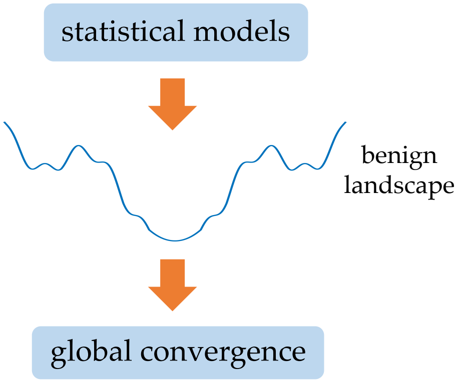
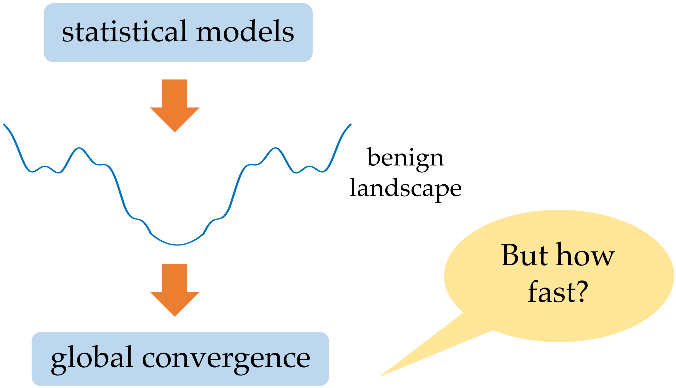



*Figure credit: Mei, Bai and Montanari*

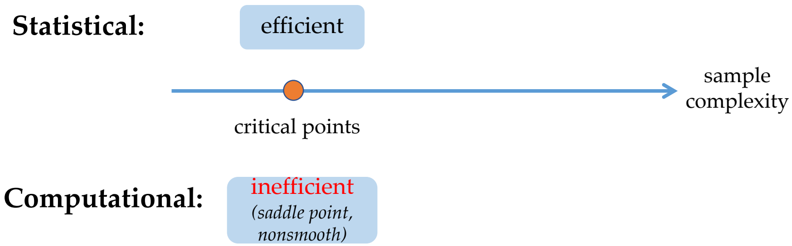
# Putting together...



# Computational efficiency?

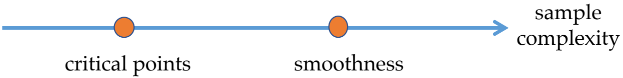


# What we know in theory



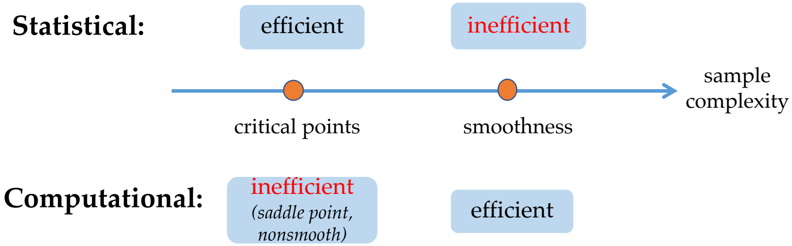
# What we know in theory

**Statistical:** efficient

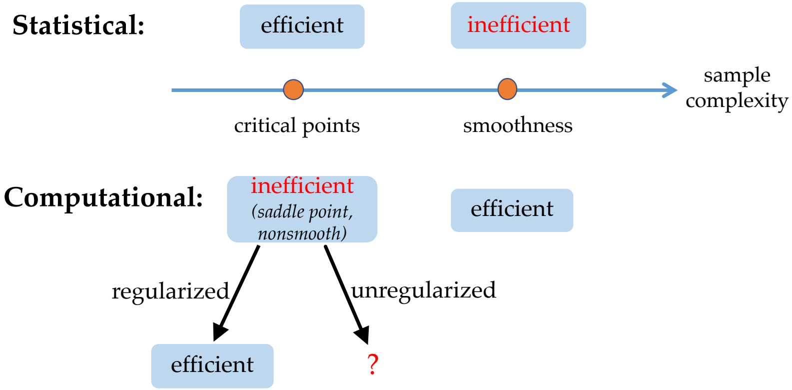



**Computational:** inefficient *(saddle point, nonsmooth)*

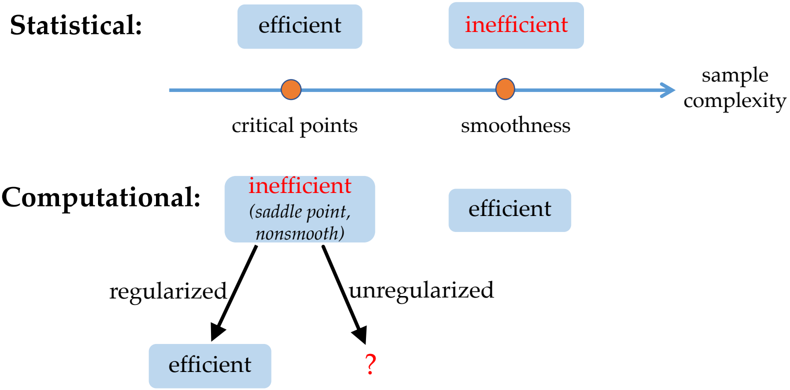
# What we know in theory



# What we know in theory



# What we know in theory




*Can we simultaneously achieve statistical and computational efficiency using unregularized methods?*

# Three problems I care about

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

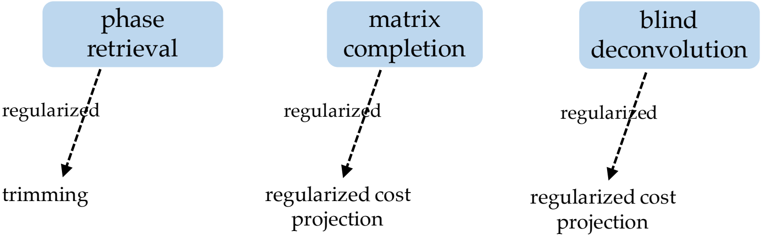
phase retrieval

matrix completion

blind deconvolution

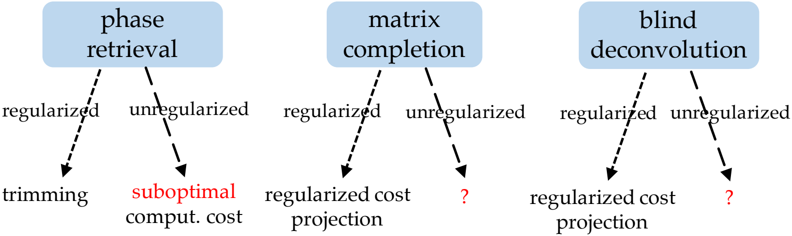
# Regularized gradient descent

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



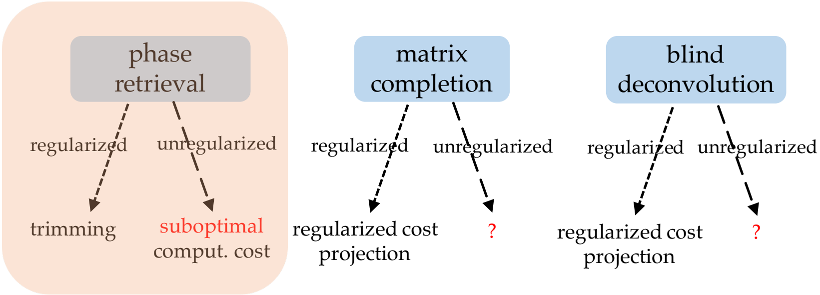
# Regularized vs. unregularized gradient descent

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



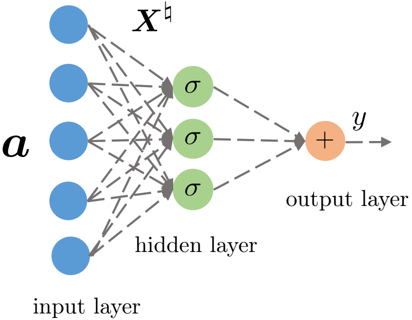
# Regularized vs. unregularized gradient descent

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




*This talk: vanilla gradient descent runs as fast as regularized ones!*

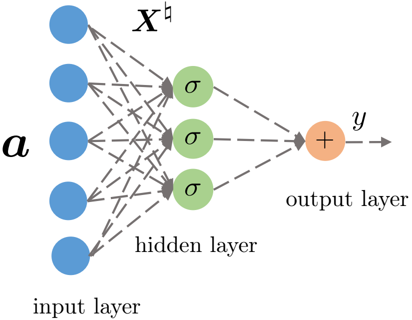
# Shallow neural network




Set $\boldsymbol{X}^\natural = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r]$, then

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

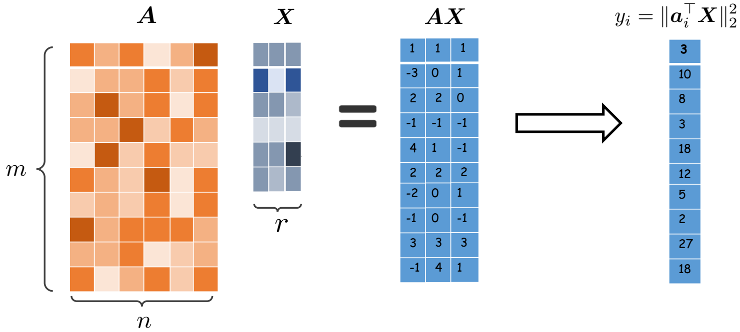
# Shallow neural network with quadratic activation




Set $\boldsymbol{X}^\natural = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r]$, then

$$y = \sum_{i=1}^{r} \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i) \overset{\sigma(z)=z^2}{:=} \sum_{i=1}^{r} (\boldsymbol{a}^\top \boldsymbol{x}_i)^2 = \left\| \boldsymbol{a}^\top \boldsymbol{X}^\natural \right\|_2^2 .$$

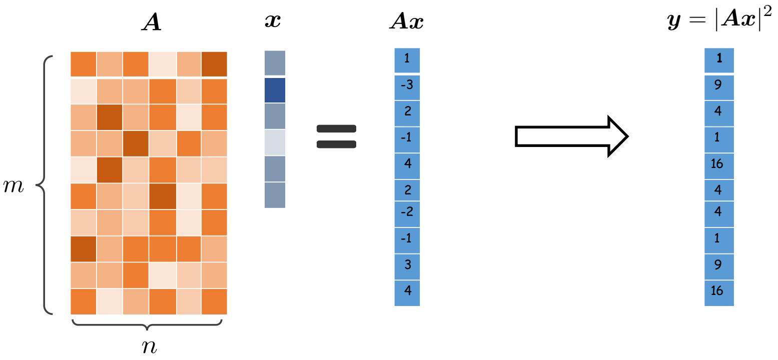
# Generalized phase retrieval




Recover $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$ from $m$ "random" quadratic measurements

$$y_i = \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2, \qquad i = 1, \ldots, m$$

# Single neuron with quadratic activation




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

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

where $m$ is about as large as $n$. *Assume w.l.o.g.* $\|x^\natural\|_2 = 1$

# A natural least squares formulation

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

$$\Downarrow$$

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

# A natural least squares formulation

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

$$\Downarrow$$

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

- **pros:** global minimizers are the truth as long as sample size is sufficiently large

# A natural least squares formulation

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

$$\Downarrow$$

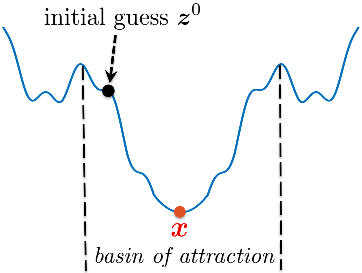
$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left|\boldsymbol{a}_k^\top \boldsymbol{x}\right|^2 - y_k\right]^2$$

- **pros:** global minimizers are the truth as long as sample size is sufficiently large
- **cons:** $f(\cdot)$ is nonconvex

  $\longrightarrow$ *computationally challenging!*

# Two-step nonconvex procedure

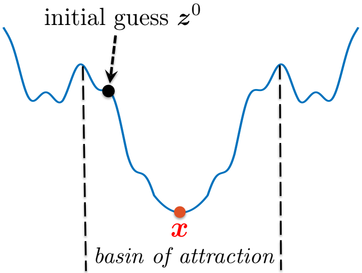
$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, f(\boldsymbol{x}) := \frac{1}{m}\sum_{i=1}^{m} \ell(y_i; \boldsymbol{x})$$

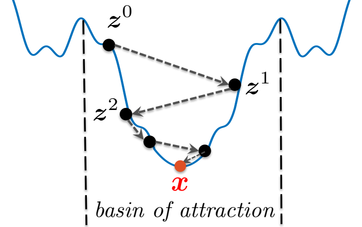



- Initialize $\boldsymbol{x}^0$ via *spectral* methods properly;
- Update using *simple* iterative methods, e.g. gradient descent.

# Two-step nonconvex procedure

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\ f(\boldsymbol{x}) := \frac{1}{m}\sum_{i=1}^{m} \ell(y_i; \boldsymbol{x})$$




- Initialize $\boldsymbol{x}^0$ via *spectral* methods properly;
- Update using *simple* iterative methods, e.g. gradient descent.

# Wirtinger flow (Candès, Li, Soltanolkotabi '14)

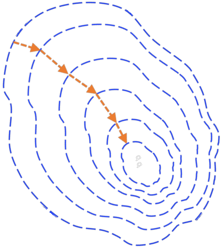
Empirical risk minimization

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

# Wirtinger flow (Candès, Li, Soltanolkotabi '14)

Empirical risk minimization

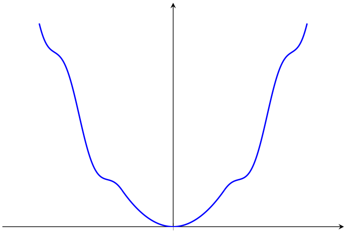
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[ \left( \boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$



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

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

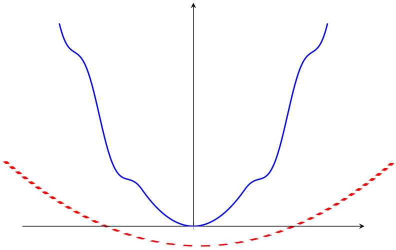
# Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

at least along certain descent directions.

# Gradient descent theory revisited

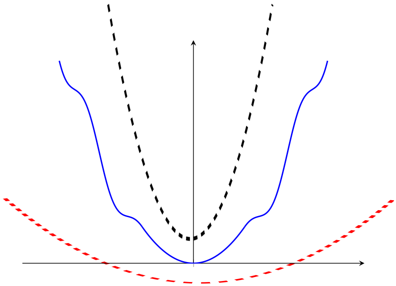


Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity

at least along certain descent directions.

# Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity
- (local) smoothness

at least along certain descent directions.

# Gradient descent theory revisited

$f$ is said to be $\alpha$-strongly convex and $\beta$-smooth if

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

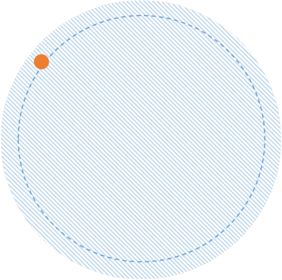
$\ell_2$ **error contraction:** GD with $\eta = 1/\beta$ obeys

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

# Gradient descent theory revisited

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

region of local strong convexity + smoothness

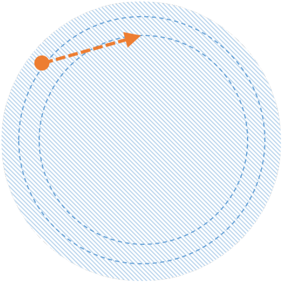


# Gradient descent theory revisited

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



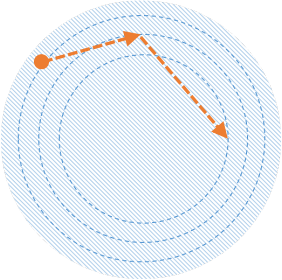
region of local strong convexity + smoothness



# Gradient descent theory revisited

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

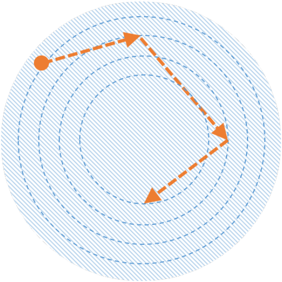
region of local strong convexity + smoothness



# Gradient descent theory revisited

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

region of local strong convexity + smoothness



# Gradient descent theory revisited

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

$\ell_2$ **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 $\frac{\beta}{\alpha}$ determines rate of convergence

# Gradient descent theory revisited

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

$\ell_2$ **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 $\frac{\beta}{\alpha}$ determines rate of convergence
- Attains $\varepsilon$-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)**

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

$$\boldsymbol{I}_n \preceq \mathbb{E}\left[\nabla^2 f(\boldsymbol{x})\right] \preceq 10\boldsymbol{I}_n$$

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

# What does this optimization theory say about WF?

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

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

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

# 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 ($m \asymp n \log n$)**

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

$$\frac{1}{2}\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)\boldsymbol{I}_n$$

# What does this optimization theory say about WF?

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

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

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

$$\frac{1}{2}\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)\boldsymbol{I}_n$$

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

# What does this optimization theory say about WF?

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

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

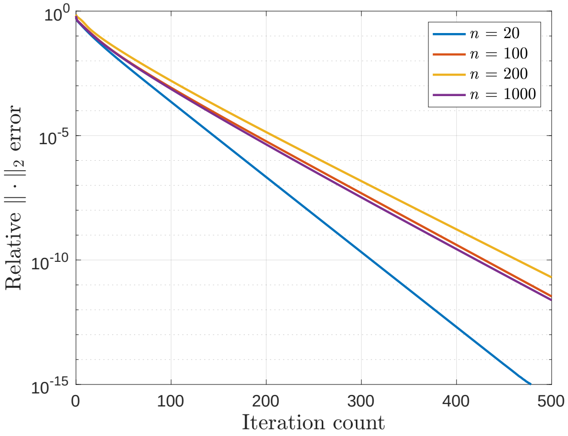
$$\nabla^2 f(\boldsymbol{x}) \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp\, n} \text{ (even locally)}$$

$$\frac{1}{2}\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)\boldsymbol{I}_n$$

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

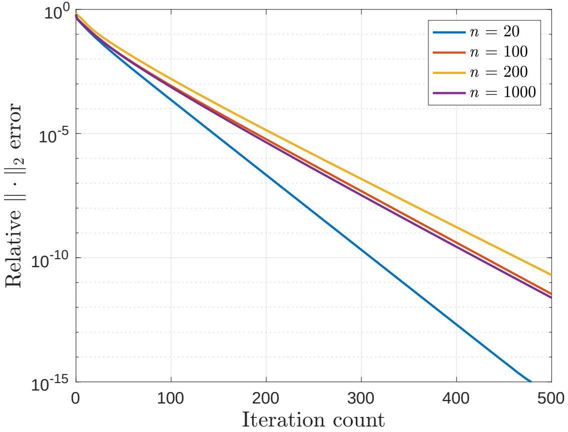
*Too slow ...*

# Numerical experiment with $\eta_t = 0.1$




Vanilla GD (WF) can proceed much more aggressively!

# Numerical experiment with $\eta_t = 0.1$




Generic optimization theory is too pessimistic!

# 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

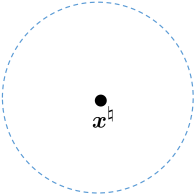
Which region enjoys both strong convexity and smoothness?

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

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

# A second look at gradient descent theory

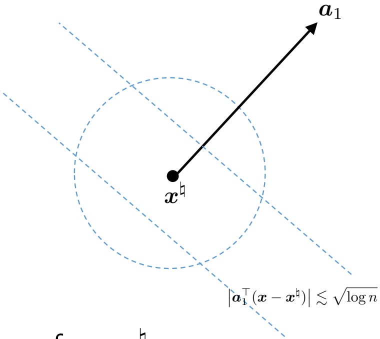
Which region enjoys both strong convexity and smoothness?




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

# A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

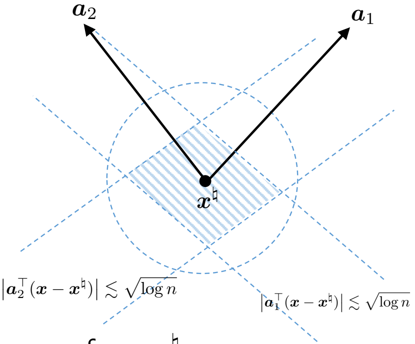



- $\boldsymbol{x}$ is not far away from $\boldsymbol{x}^\natural$
- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors (incoherence region)

$$(1/2) \cdot \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log n) \cdot \boldsymbol{I}_n$$

# A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

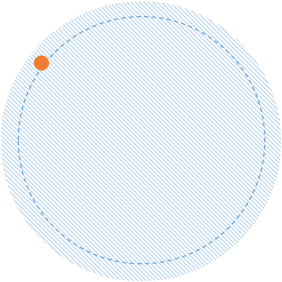



- $\boldsymbol{x}$ is not far away from $\boldsymbol{x}^\natural$
- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors (incoherence region)

$$(1/2) \cdot \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log n) \cdot \boldsymbol{I}_n$$

# A second look at gradient descent theory

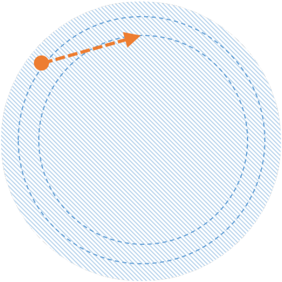
region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

# A second look at gradient descent theory

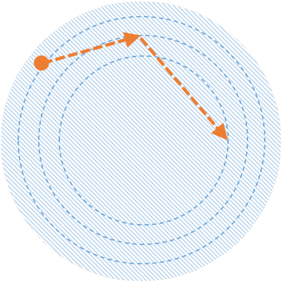
region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

# A second look at gradient descent theory

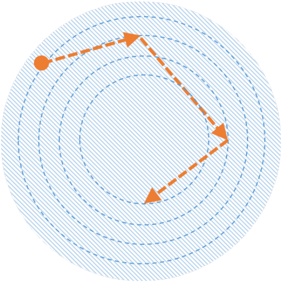
region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

# A second look at gradient descent theory

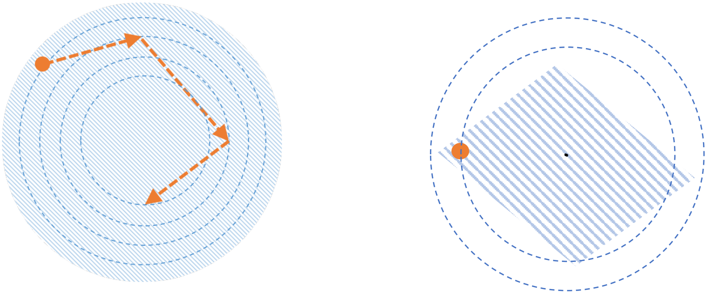
 region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

# A second look at gradient descent theory

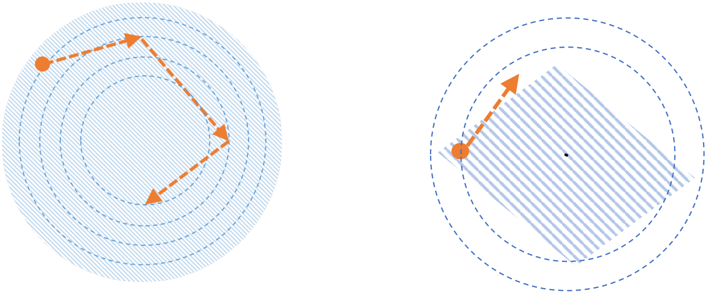
 region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

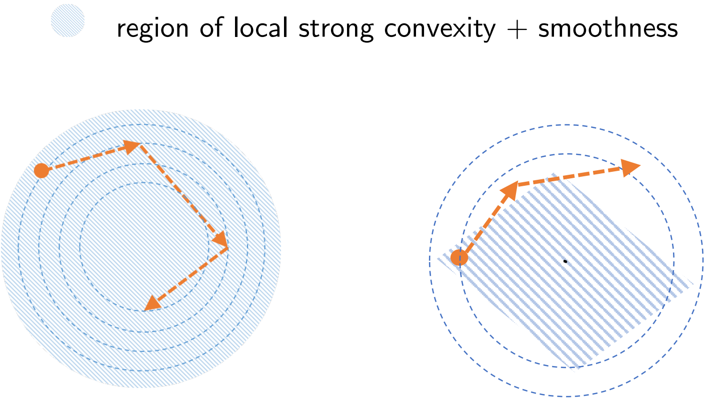
# A second look at gradient descent theory

region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

# A second look at gradient descent theory

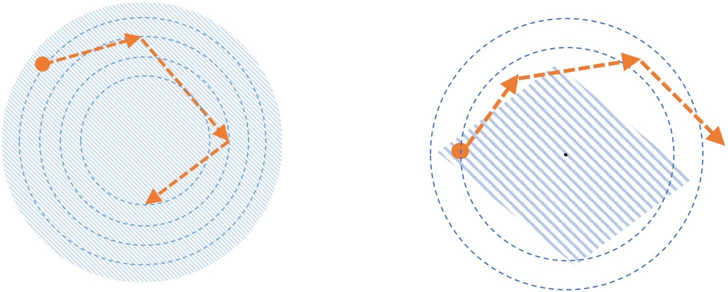



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

# A second look at gradient descent theory

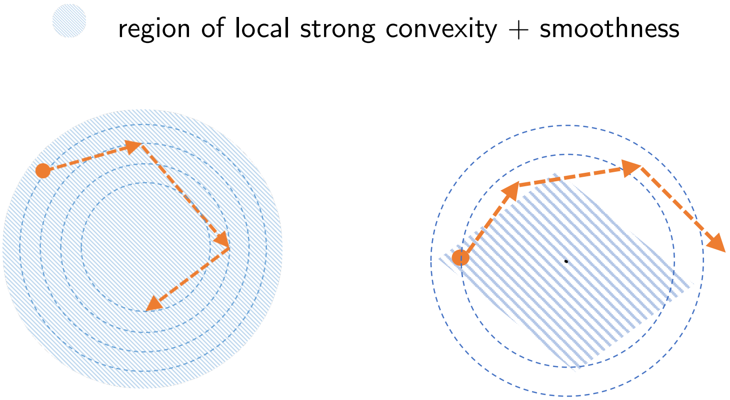


region of local strong convexity + smoothness



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

# A second look at gradient descent theory

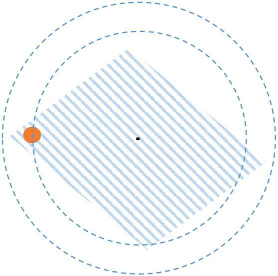



- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
- *Existing algorithms enforce regularization, or apply sample splitting to promote incoherence*

# Our findings: GD is implicitly regularized



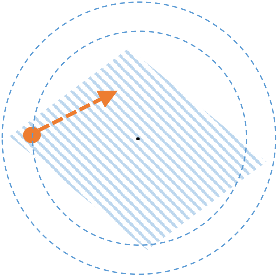
region of local strong convexity + smoothness



# Our findings: GD is implicitly regularized

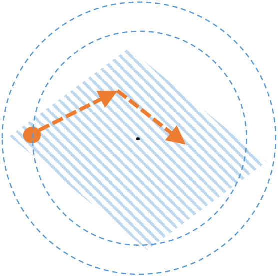


region of local strong convexity + smoothness

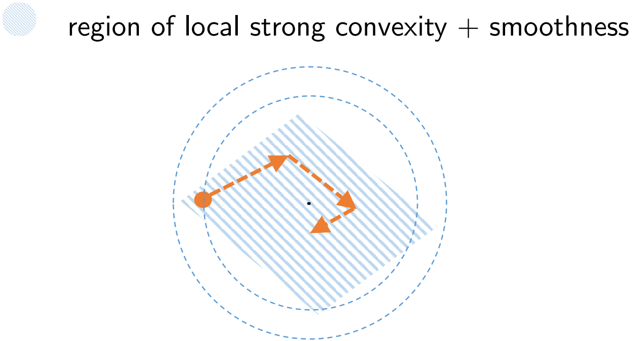


# Our findings: GD is implicitly regularized

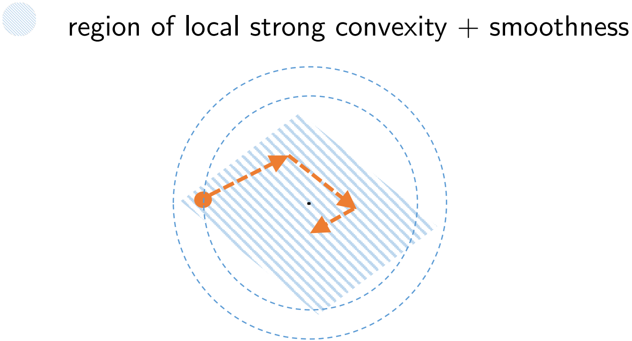
region of local strong convexity + smoothness



# Our findings: GD is implicitly regularized



# Our findings: GD is implicitly regularized




GD implicitly forces iterates to remain incoherent
even without regularization

# Theoretical guarantees

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

**Big computational saving**: WF attains $\varepsilon$-accuracy within $O\big(\log n \log \frac{1}{\varepsilon}\big)$ iterations with $\eta \asymp 1/\log n$ if $m \asymp n \log n$

# Key ingredient: leave-one-out analysis

How to establish $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\,\|\boldsymbol{x}^\natural\|_2$?

# Key ingredient: leave-one-out analysis

How to establish $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\,\|\boldsymbol{x}^\natural\|_2$?

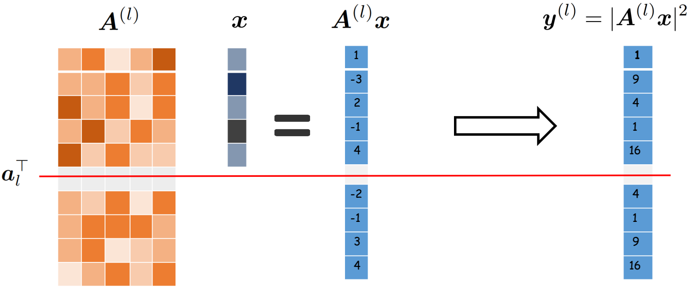
Technical difficulty: $\boldsymbol{x}^t$ is statistically dependent with $\{\boldsymbol{a}_l\}$;

# Key ingredient: leave-one-out analysis

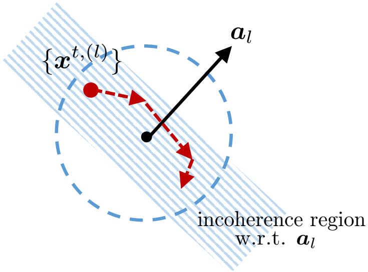
How to establish $\left|\boldsymbol{a}_l^\top(\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\,\|\boldsymbol{x}^\natural\|_2$?

Technical difficulty: $\boldsymbol{x}^t$ is statistically dependent with $\{\boldsymbol{a}_l\}$;

Leave-one-out trick: For each $1 \leq l \leq m$, introduce leave-one-out iterates $\boldsymbol{x}^{t,(l)}$ by dropping $l$th sample

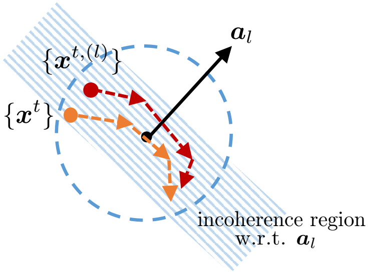


# Key ingredient: leave-one-out analysis



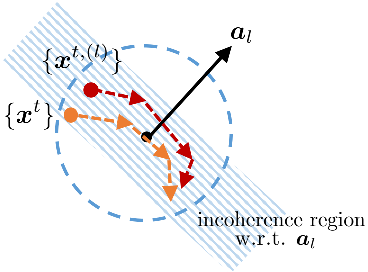

- Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.

## 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 $\boldsymbol{x}^t$

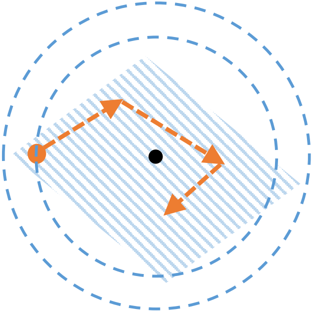
# Key ingredient: leave-one-out analysis



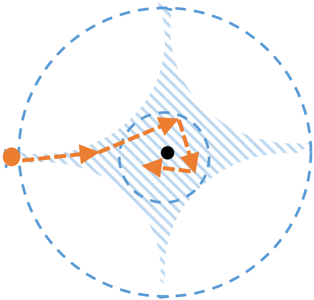

- Leave-one-out iterates $\{\boldsymbol{x}^{t,(l)}\}$ are independent of $\boldsymbol{a}_l$, and are hence **incoherent** w.r.t. $\boldsymbol{a}_l$ with high prob.
- Leave-one-out iterates $\boldsymbol{x}^{t,(l)} \approx$ true iterates $\boldsymbol{x}^t$
- Finish by triangle inequality

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

# Incoherence region in high dimensions



2-dimensional

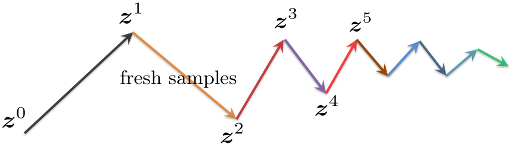


high-dimensional

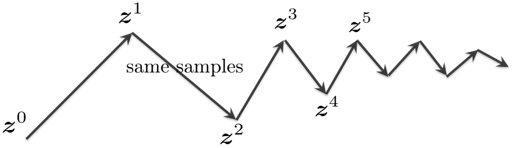
incoherence region is vanishingly small

# No 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



*This recipe is quite general*

# Low-rank matrix completion

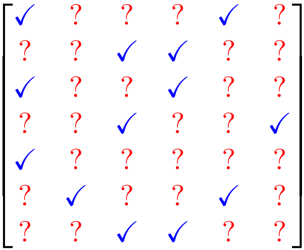


*Fig. credit: Candès*

Given partial samples of a *low-rank* matrix $\boldsymbol{M}$ in an index set $\Omega$, fill in missing entries.

*Applications: recommendation systems, ...*

# Incoherence

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

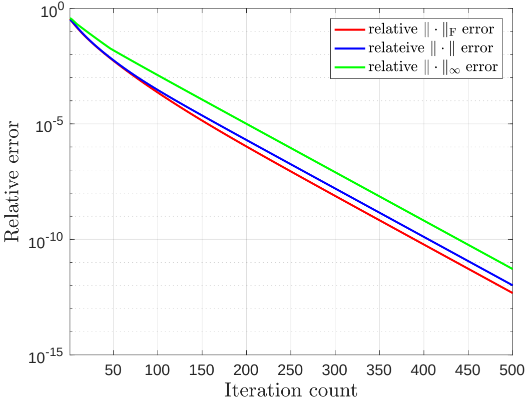
**Definition (Incoherence for matrix completion)**

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

$$\left\| \boldsymbol{U}^\natural \right\|_{2,\infty} \leq \sqrt{\frac{\mu}{n}} \left\| \boldsymbol{U}^\natural \right\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}.$$

# Matrix completion via vanilla GD

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



# Prior theory

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

Existing theory promotes incoherence explicitly:

# Prior theory

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

Existing theory promotes incoherence explicitly:

- 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 theory

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

Existing theory promotes incoherence explicitly:

- 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
- no theory on vanilla / unregularized gradient descent

# Our theory

**Theorem (Matrix completion)**

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

- $\max_i \|\boldsymbol{e}_i^\top(\boldsymbol{X}^t - \boldsymbol{X}^{\natural})\|_2 \ll \|\boldsymbol{X}^{\natural}\|_{2,\infty}$ *(incoherence)*
- *in* $O\big(\log\frac{1}{\varepsilon}\big)$ *iterations*

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

# Our theory

**Theorem (Matrix completion)**

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

- $\max_i \|e_i^\top (X^t - X^\natural)\|_2 \ll \|X^\natural\|_{2,\infty}$ *(incoherence)*
- *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* $\eta \lesssim 1/\sigma_{\max}(M)$ *and sample size* $\gtrsim nr^3 \log^3 n$

- **near-optimal entrywise error control** $\big\|X^t X^{t\top} - M^\natural\big\|_\infty$.
- $O(\log 1/\varepsilon)$ iteration complexity.
- First result on vanilla gradient descent for matrix completion.

# Noiseless matrix completion via Vanilla GD

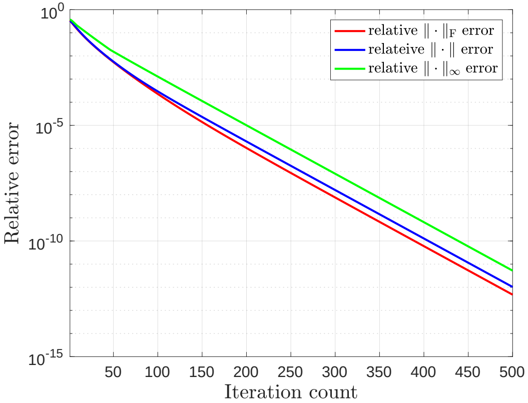



Figure: Relative error of $\boldsymbol{X}^t \boldsymbol{X}^{t\top}$ (measured by $\|\cdot\|_F, \|\cdot\|, \|\cdot\|_\infty$) vs. iteration count for matrix completion, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$.

# Noisy matrix completion via Vanilla GD

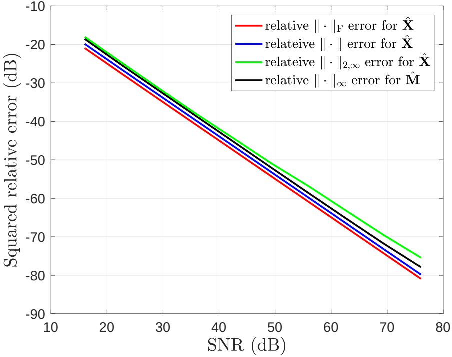
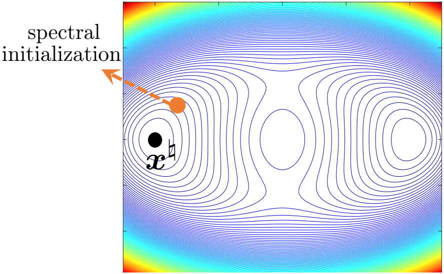



Figure: Squared relative error of the estimate $\hat{\boldsymbol{X}}$ (measured by $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_{2,\infty}$) and $\hat{\boldsymbol{M}} = \hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^{\top}$ (measured by $\|\cdot\|_{\infty}$) vs. SNR, where $n = 500$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. Here, SNR $:= \frac{\|\boldsymbol{M}^{\natural}\|_{\mathrm{F}}^2}{n^2\sigma^2}$.

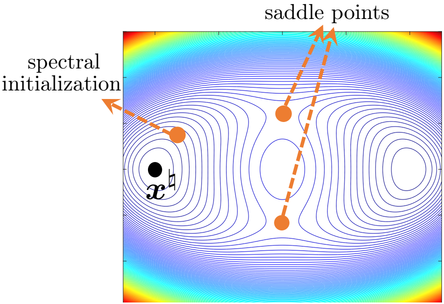
*What about random initialization?*

# Initialization



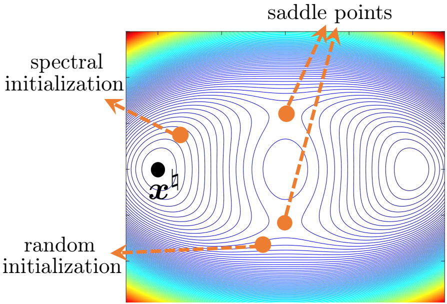

- spectral initialization gets us reasonably close to truth

# Initialization




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

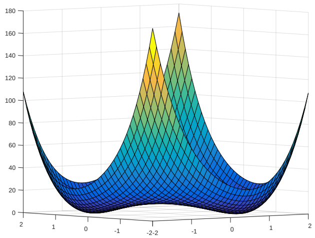
# Initialization




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

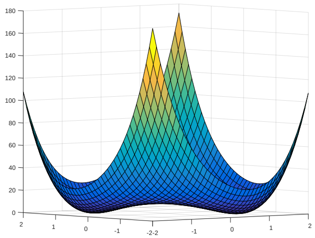
Can we initialize GD randomly?

# What does prior theory say?




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

# What does prior theory say?

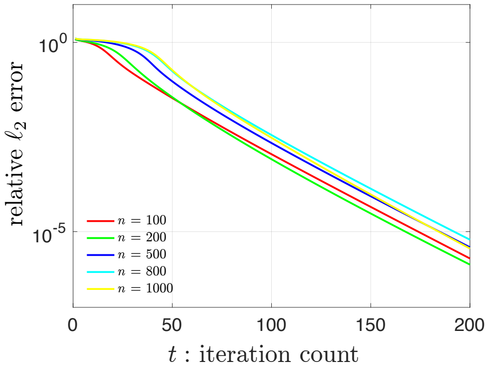



- no spurious local mins (Sun et al. '16)
- Vanilla GD with random initialization converges to global min almost surely (Lee et al. '16)

No convergence rate guarantees for vanilla GD!

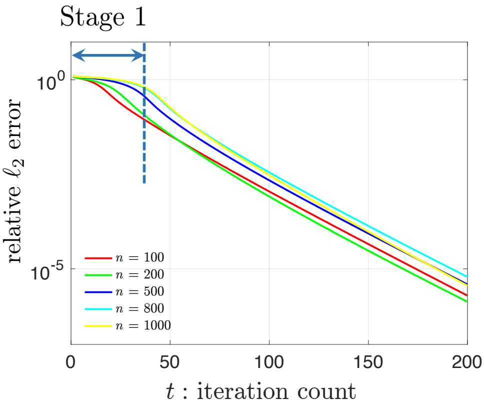
# Randomly initialized GD for phase retrieval

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



# Randomly initialized GD for phase retrieval

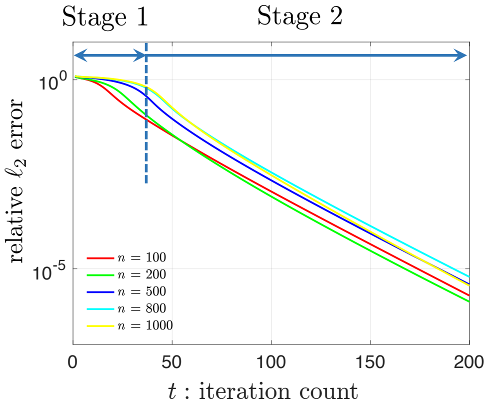
$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$




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

# Randomly initialized GD for phase retrieval

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




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

# 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 (Chen, Chi, Fan, Ma '18)**

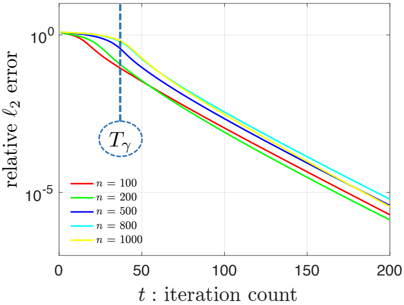
*Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$ achieves*

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

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

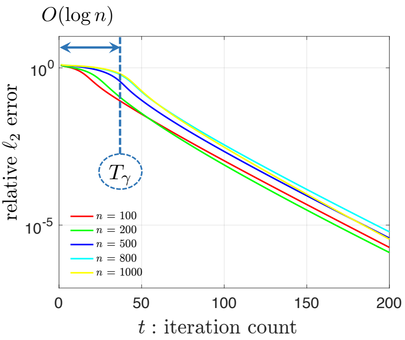
# Theoretical guarantees

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



# Theoretical guarantees

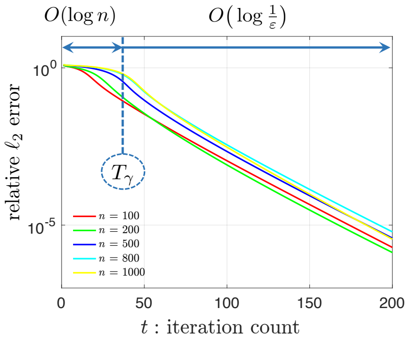
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma (1-\rho)^{t - T_\gamma} \|\boldsymbol{x}^\natural\|_2, \quad t \ge T_\gamma \asymp \log n$$




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

# Theoretical guarantees

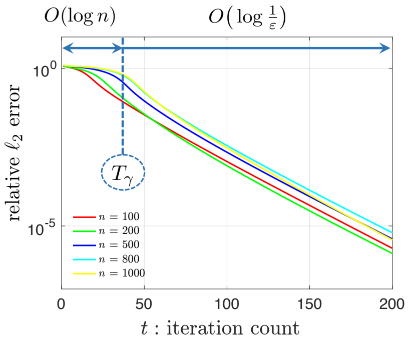
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma (1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$




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

# Theoretical guarantees

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



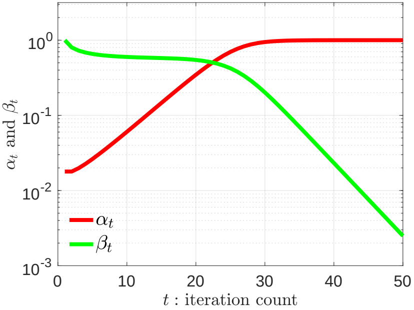

Randomly initialized WF attains $\varepsilon$-accuracy within $O\left(\log n + \log \frac{1}{\varepsilon}\right)$ iterations with $\eta \asymp 1$ if $m \asymp n\mathrm{poly}\log m$

# Population-level (infinite samples) state evolution

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \cdot \underbrace{\nabla F(\boldsymbol{x}^t)}_{\text{population gradient}}$$

Let $\alpha_t := \underbrace{\left|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle\right|}_{\text{signal strength}}$ ,

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

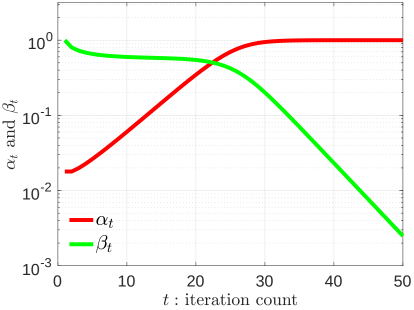


# Population-level (infinite samples) state evolution

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta \cdot \underbrace{\nabla F(\boldsymbol{x}^{t})}_{\text{population gradient}}$$

Let $\alpha_t := \underbrace{\left|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle\right|}_{\text{signal strength}}$ ,

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




2-parameter dynamics:

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

# Back to finite-sample analysis

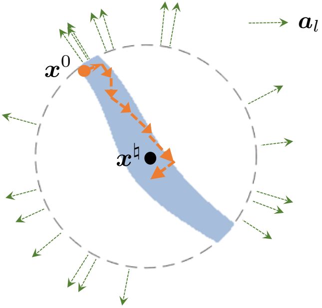
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$$

# Back to finite-sample analysis

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

# Back to finite-sample analysis

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

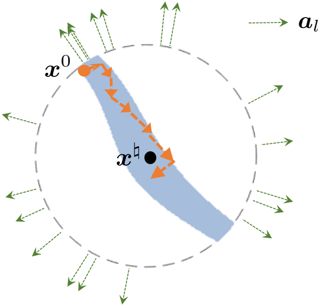



a region with well-controlled $\boldsymbol{r}(\boldsymbol{x})$

- 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

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

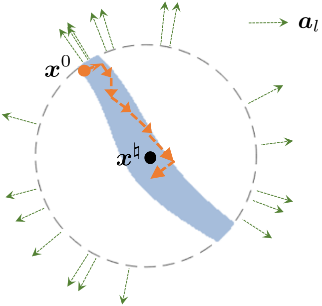



a region with well-controlled $\boldsymbol{r}(\boldsymbol{x})$

- 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 $\{\boldsymbol{a}_k\}$

# 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 $\{\boldsymbol{a}_k\}$
- **key analysis ingredient:** show $\boldsymbol{x}^t$ is "nearly-independent" of each $\boldsymbol{a}_k$ via leave-one-out analysis

# Conclusions

**optimization theory + statistical model:** vanilla gradient descent is "implicitly regularized" and runs fast!

**Computational:**
near dimension-free
iteration complexity

**Statistical:**
near-optimal
sample complexity

It will be interesting to study "implicit regularization" via the leave-one-out argument for other algorithms such as alternating minimization, and other problems.

# References

1. Implicit Regularization for Nonconvex Statistical Estimation, C. Ma, K. Wang, Y. Chi and Y. Chen, arXiv:1711.10467.
2. Nonconvex Matrix Factorization from Rank-One Measurements, Y. Li, C. Ma, Y. Chen, and Y. Chi, arXiv:1802.06286.
3. Gradient Descent with *Random Initialization*: Fast Global Convergence for Nonconvex Phase Retrieval, Y. Chen, Y. Chi, J. Fan and C. Ma, arXiv:1803.07726.
4. Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation, Y. Chen and Y. Chi, survey article on IEEE Signal Processing Magazine, to appear.

Thank you!