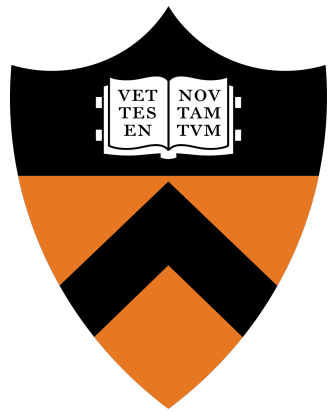
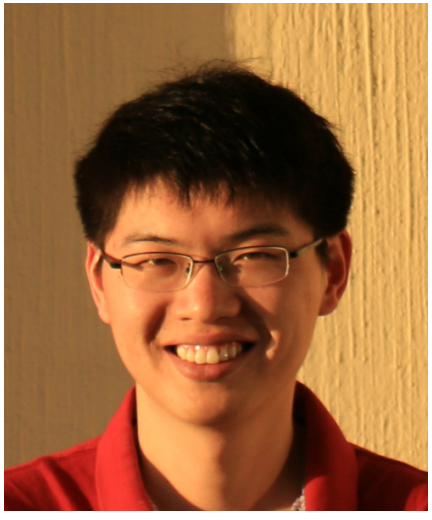
# Random Initialization in Nonconvex Phase Retrieval

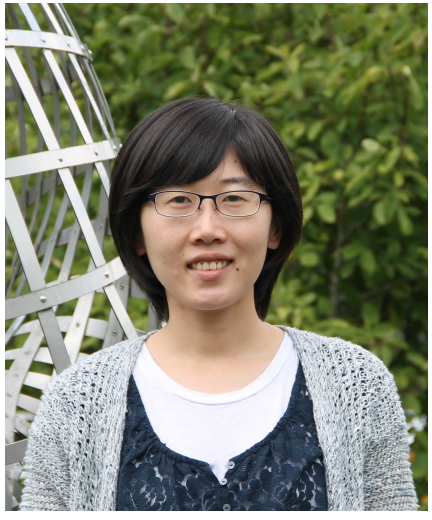



Yuxin Chen

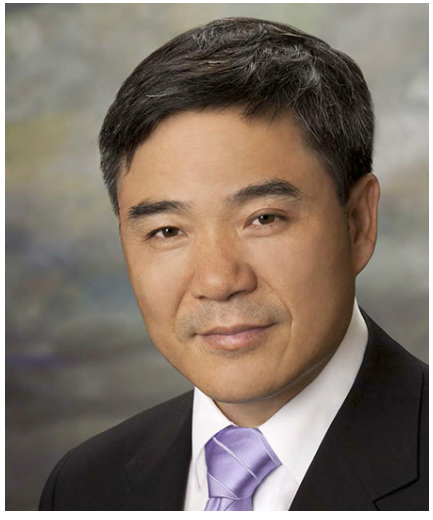
Electrical Engineering, Princeton University



Cong Ma
Princeton ORFE



Yuejie Chi
CMU ECE

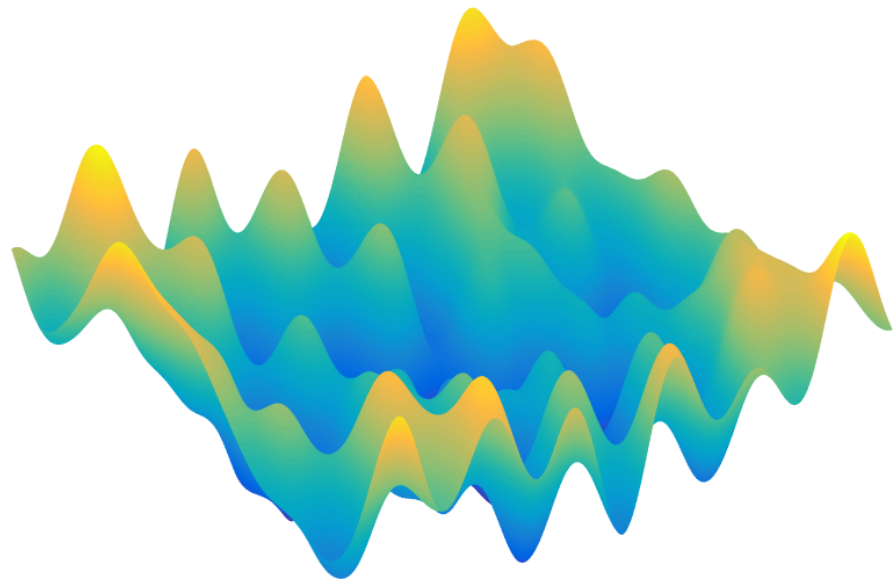


Jianqing Fan
Princeton ORFE

# Nonconvex problems are everywhere

Empirical risk minimization is usually nonconvex

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

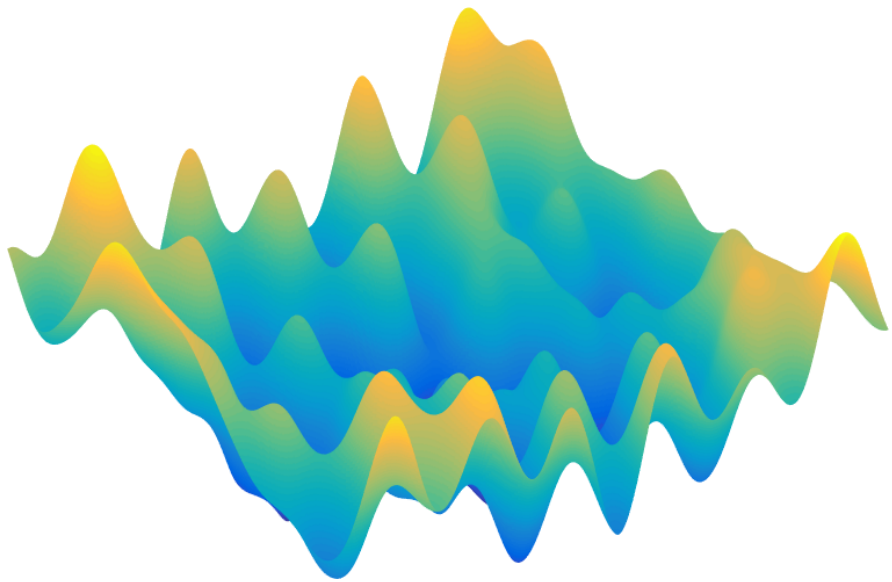


# Nonconvex problems are everywhere

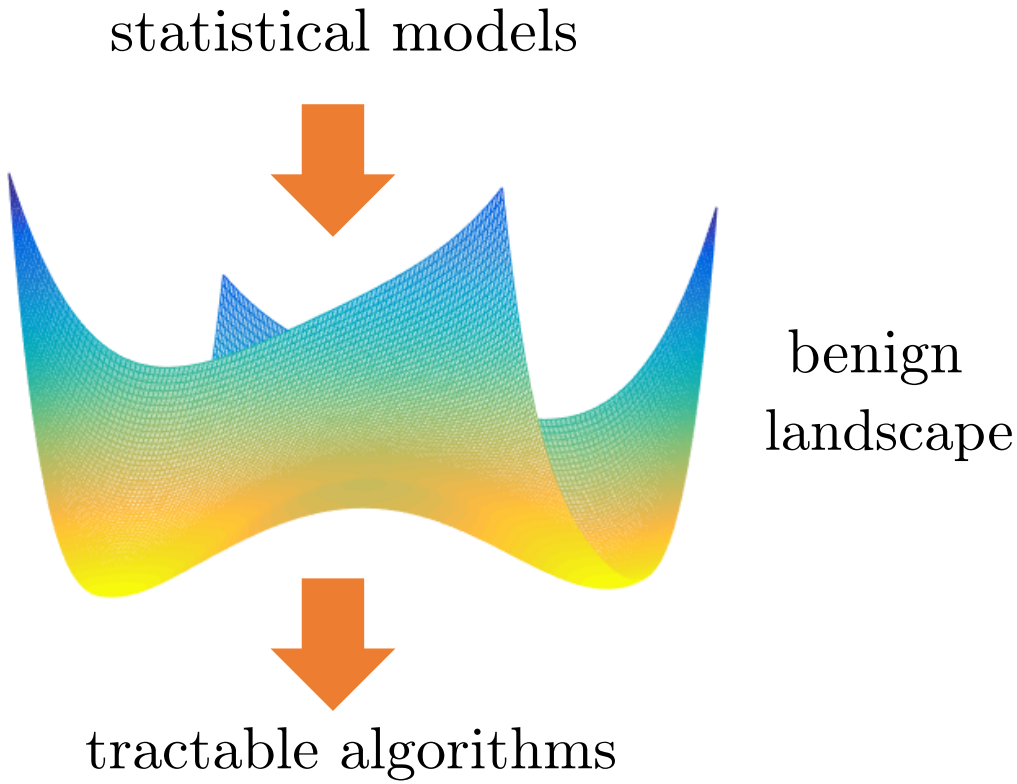
Empirical risk minimization is usually nonconvex

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

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

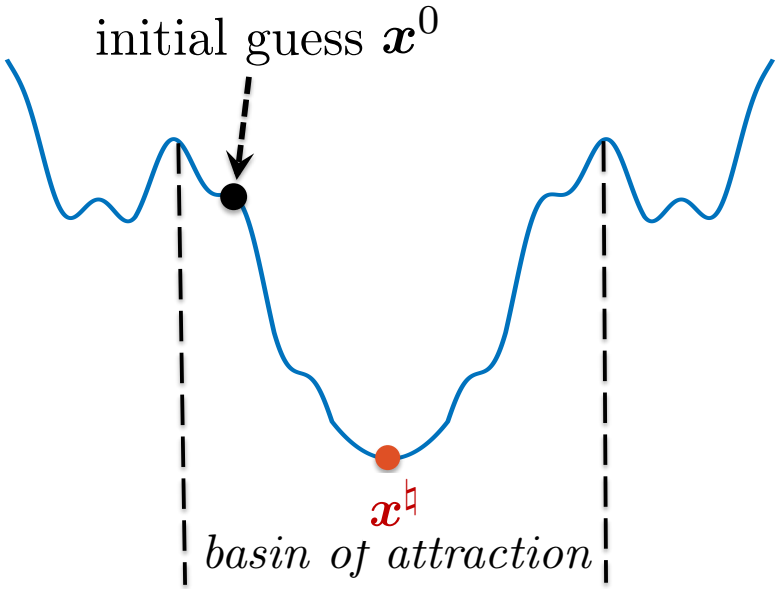


# Statistical models come to rescue



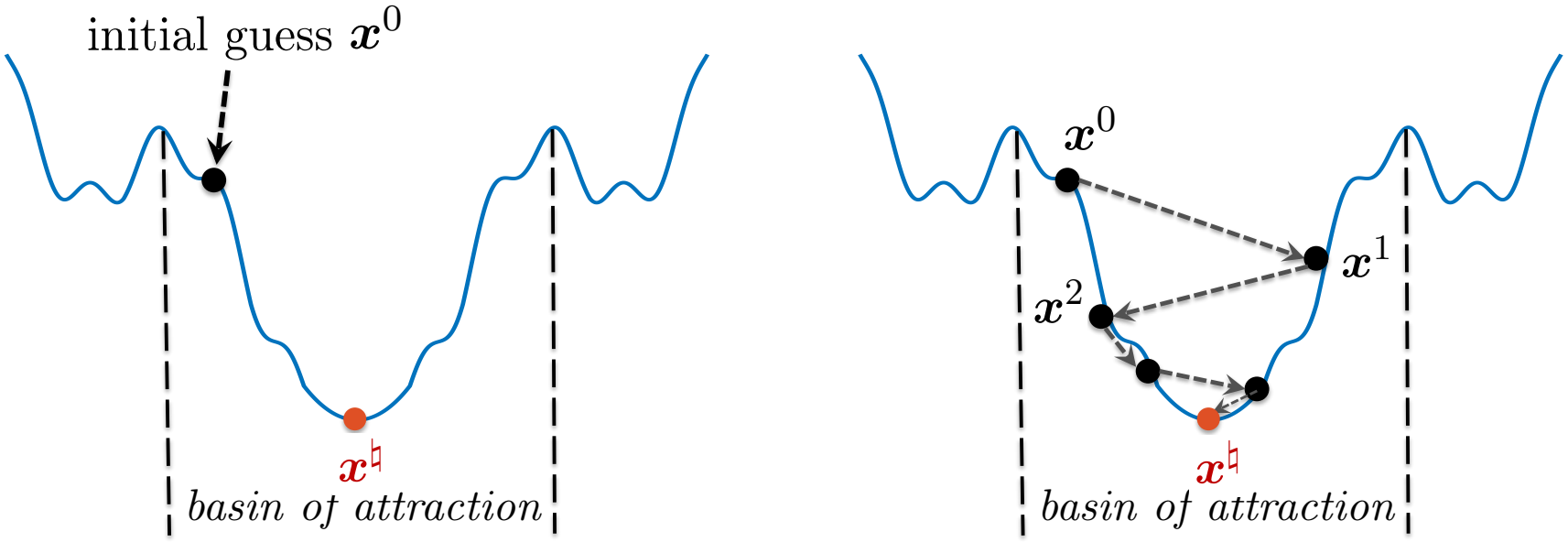

When data are generated by certain statistical models, problems are often much nicer than worst-case instances

# A popular two-stage approach




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

# A popular two-stage approach




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

2. iterative refinement

# A highly incomplete list of two-stage methods

## phase retrieval:

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

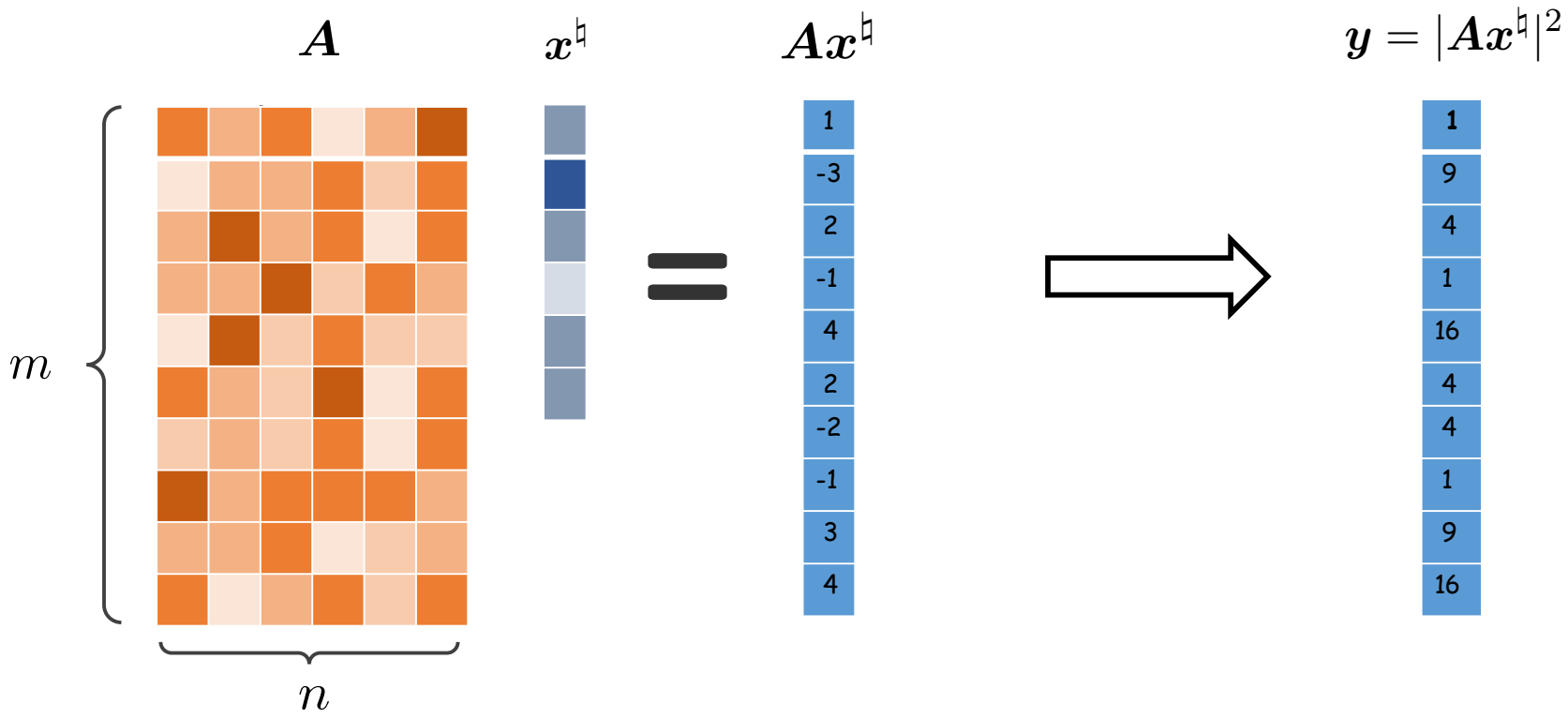
## other problems:

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

*Is carefully-designed initialization necessary for fast convergence?*

**A case study: phase retrieval**

# Phase retrieval / solving quadratic systems of equations




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

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

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

# A natural least squares formulation

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

# A natural least squares formulation

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

- **issue:** $f(\cdot)$ is highly nonconvex

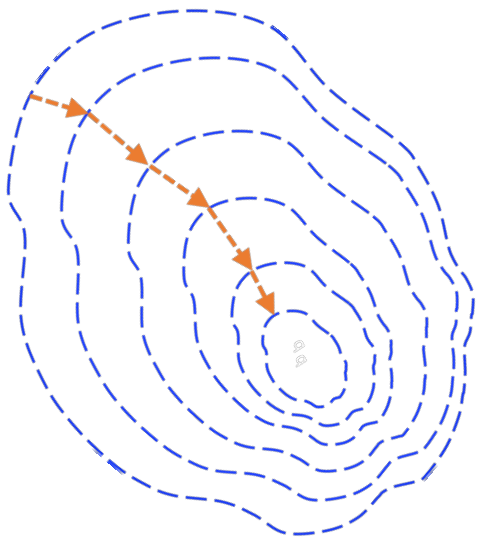
  $\longrightarrow$ *computationally challenging!*

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

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

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

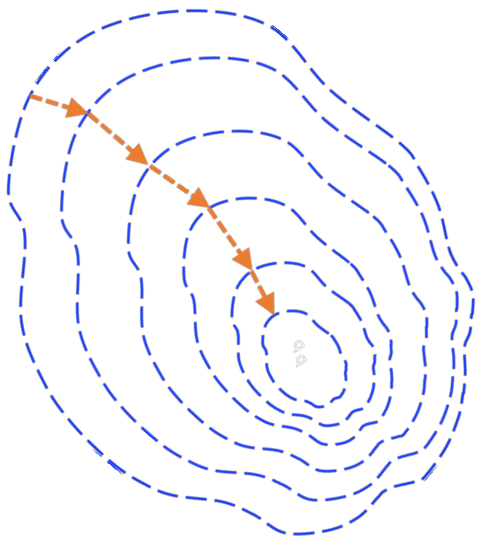
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[ \left( \boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$



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

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

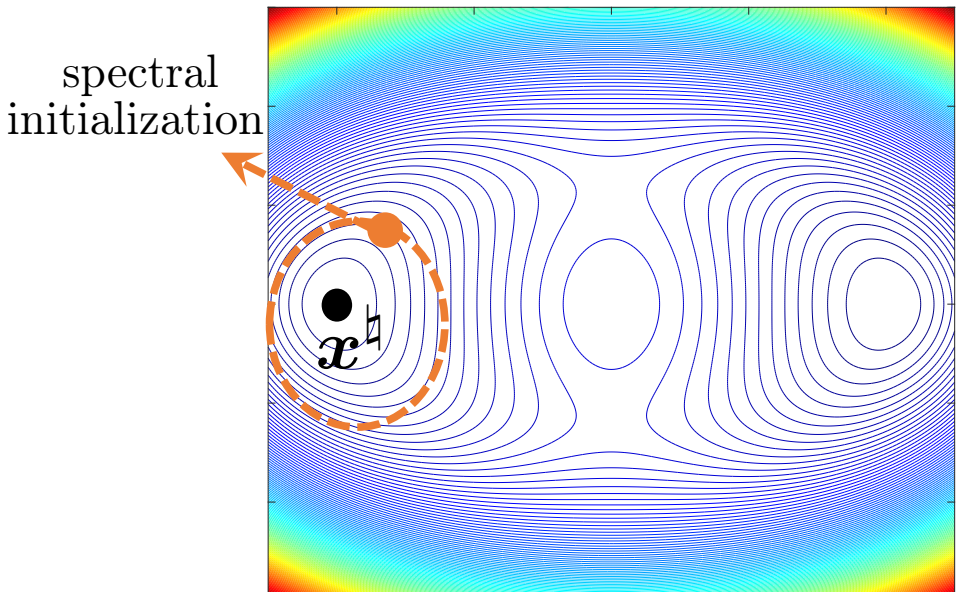
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



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

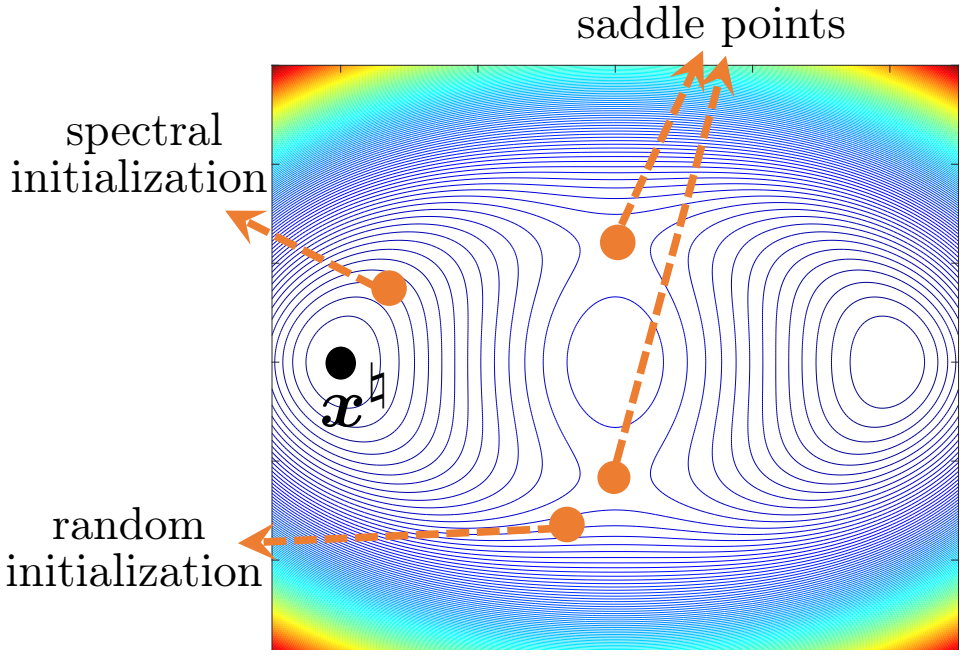
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

# Initialization




- spectral initialization gets us to (restricted) strongly cvx region

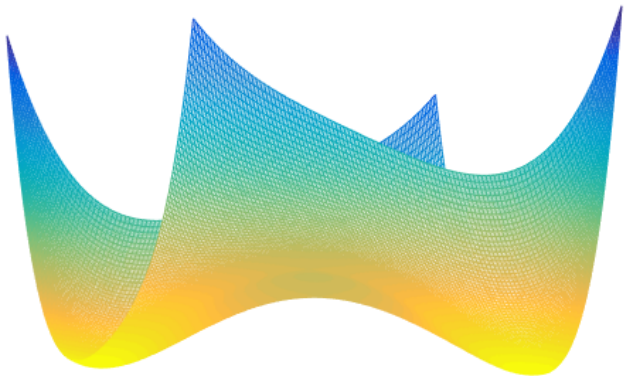
# Initialization




- spectral initialization gets us to (restricted) strongly cvx region

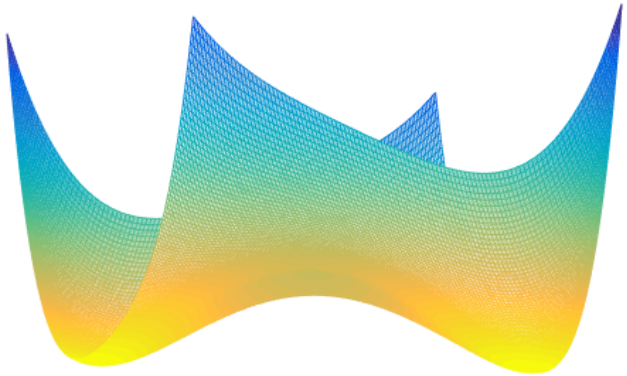
Can we initialize GD randomly, which is simpler and model-agnostic?

# What does prior theory say?



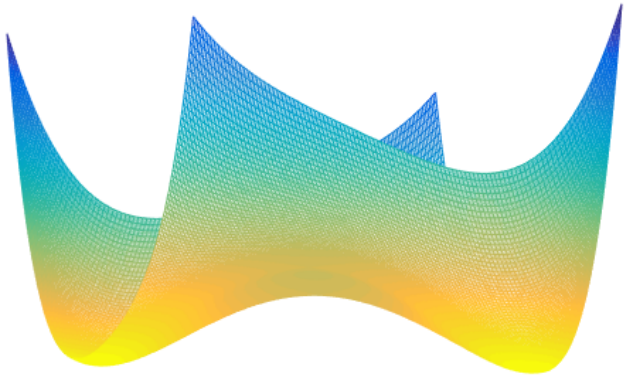
- **landscape:** no spurious local mins (Sun et al. '16)

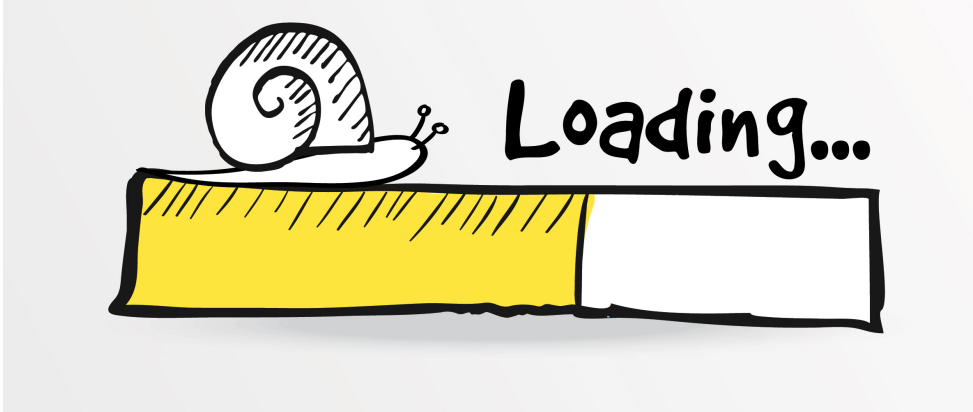
# What does prior theory say?



- **landscape:** no spurious local mins (Sun et al. '16)
- randomly initialized GD converges almost surely (Lee et al. '16)

# What does prior theory say?



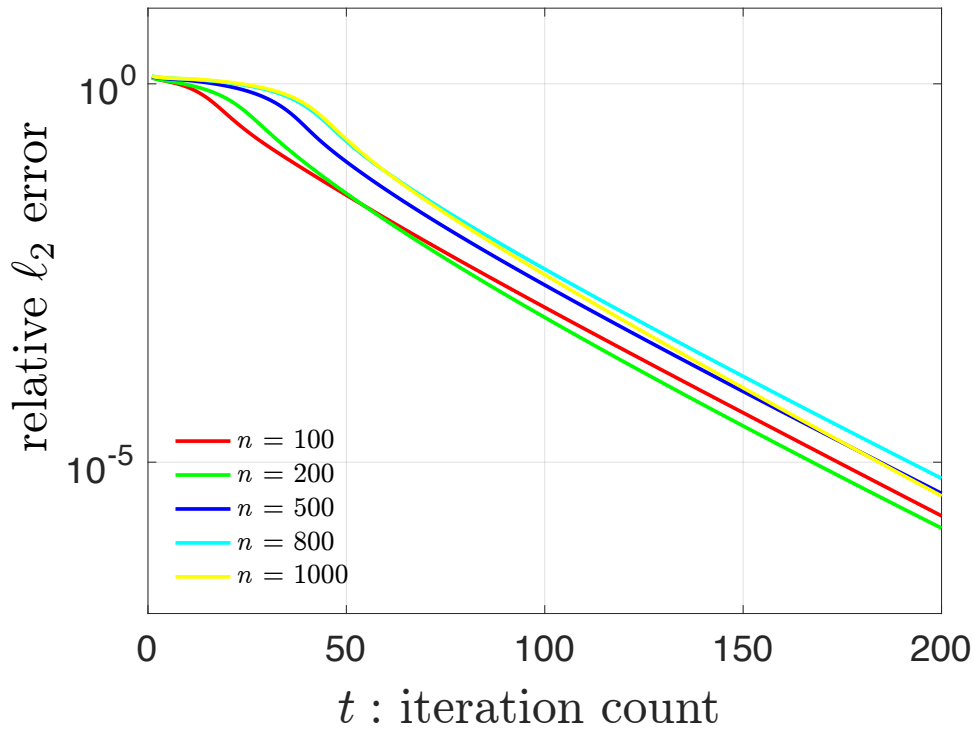



- **landscape:** no spurious local mins (Sun et al. '16)
- randomly initialized GD converges almost surely (Lee et al. '16)

"almost surely" might mean "takes forever"

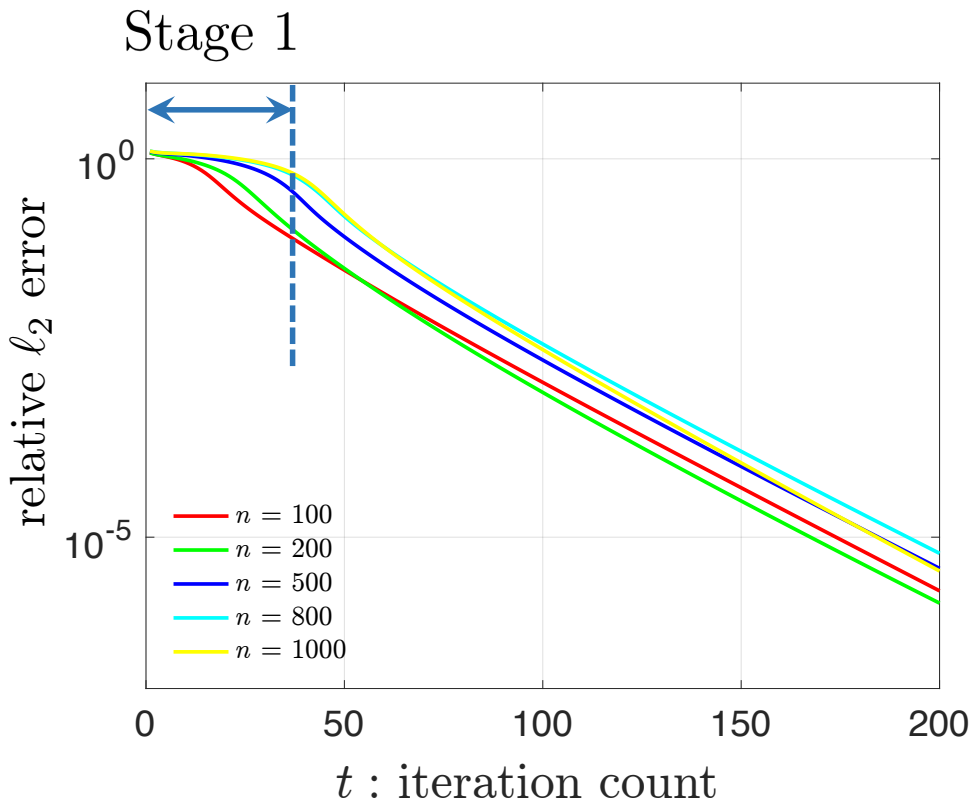
# Numerical efficiency of randomly initialized GD

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



# Numerical efficiency of randomly initialized GD

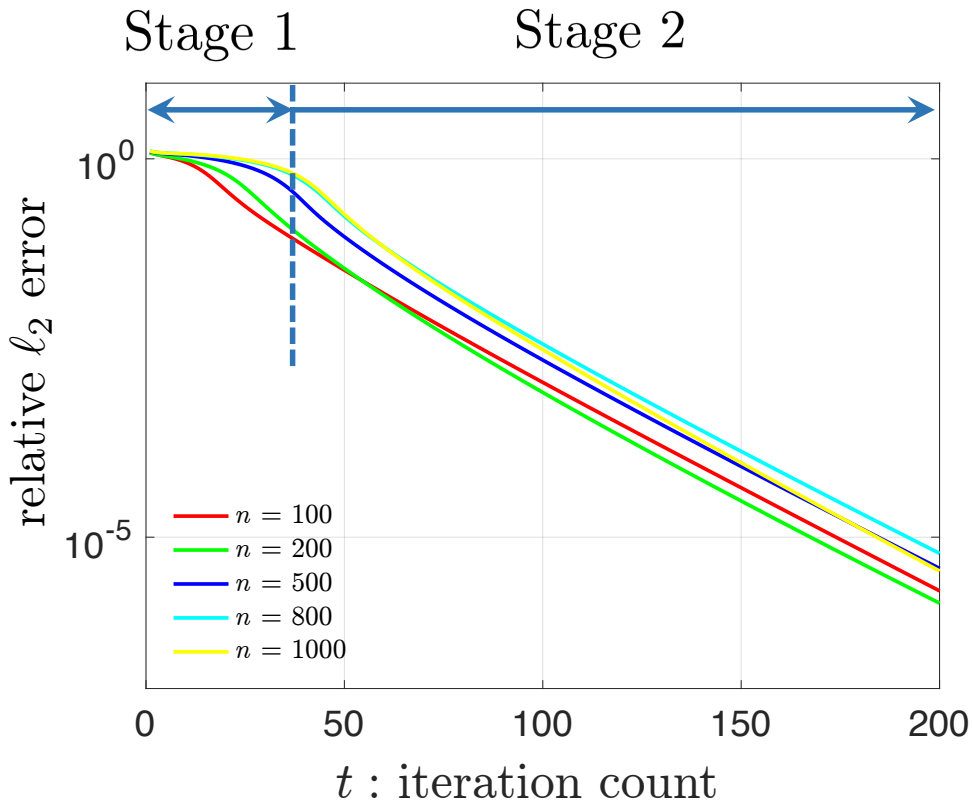
$$\eta = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$




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

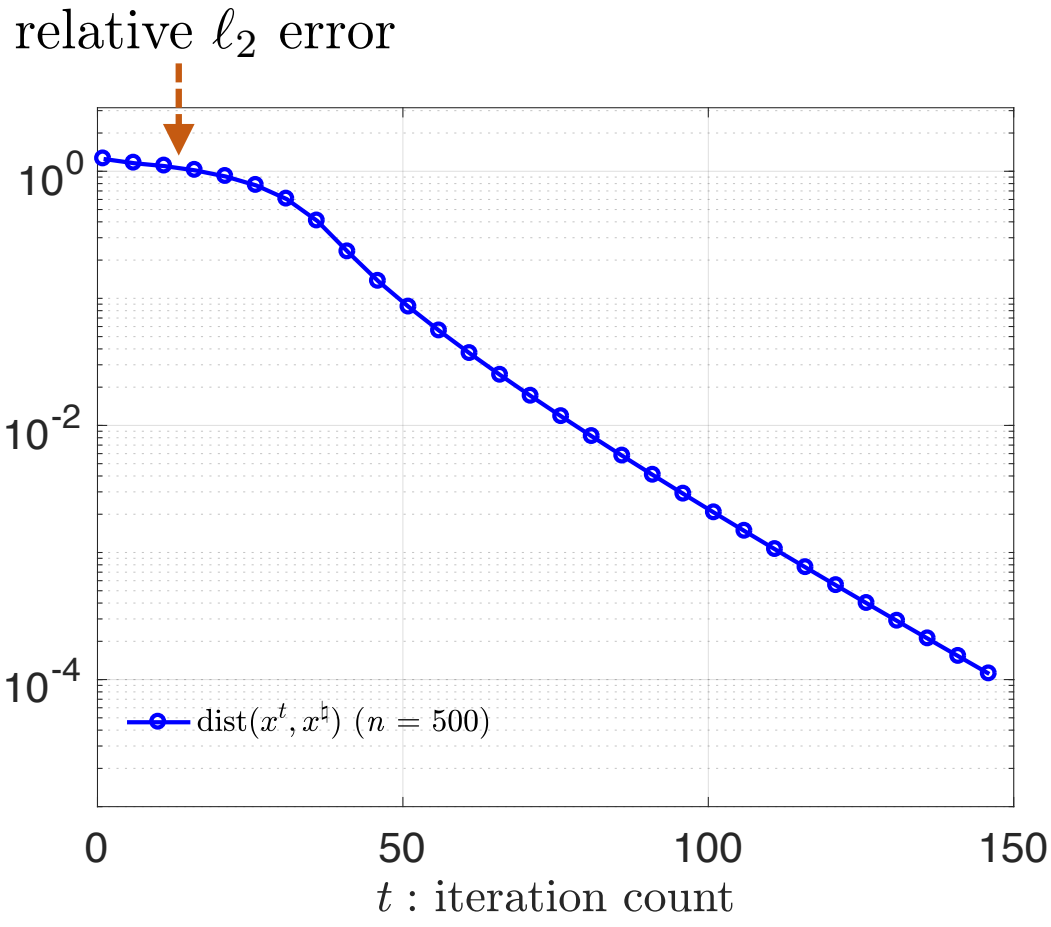
# Numerical efficiency of randomly initialized GD

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

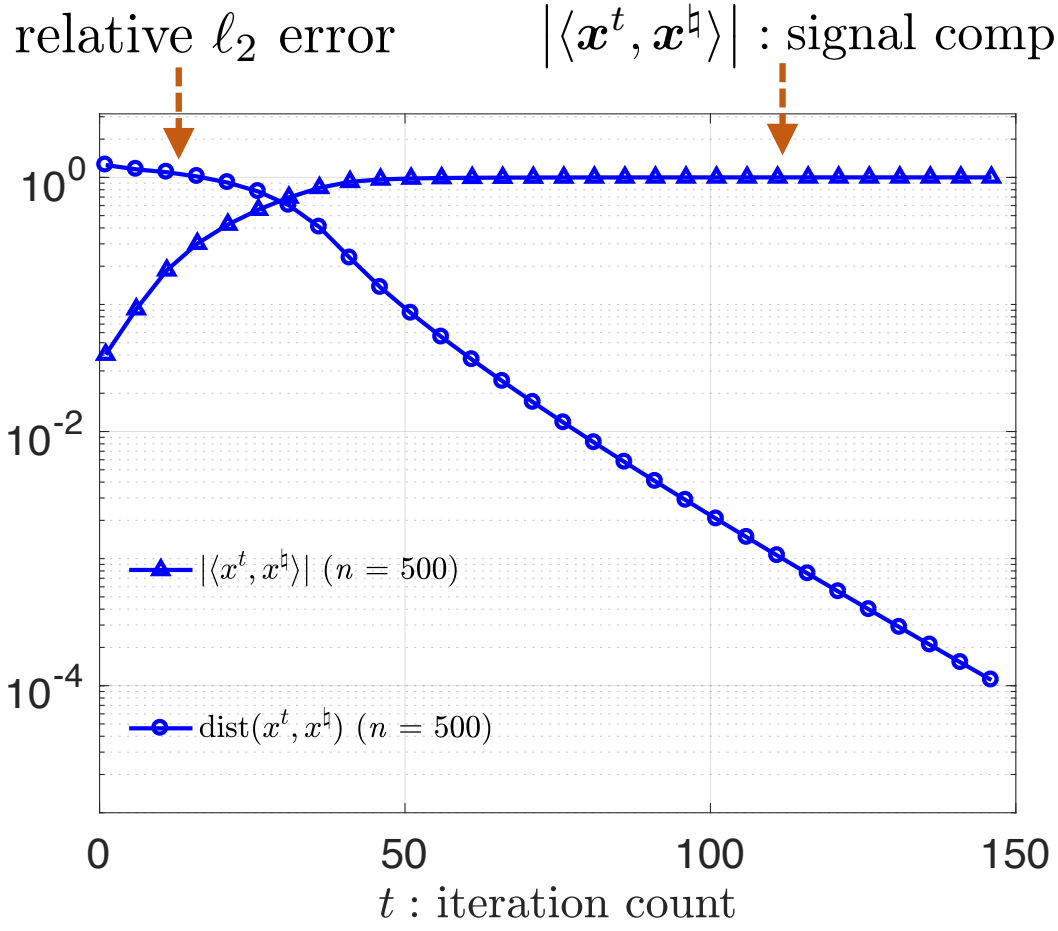



Randomly initialized GD enters local basin within **a few iterations**

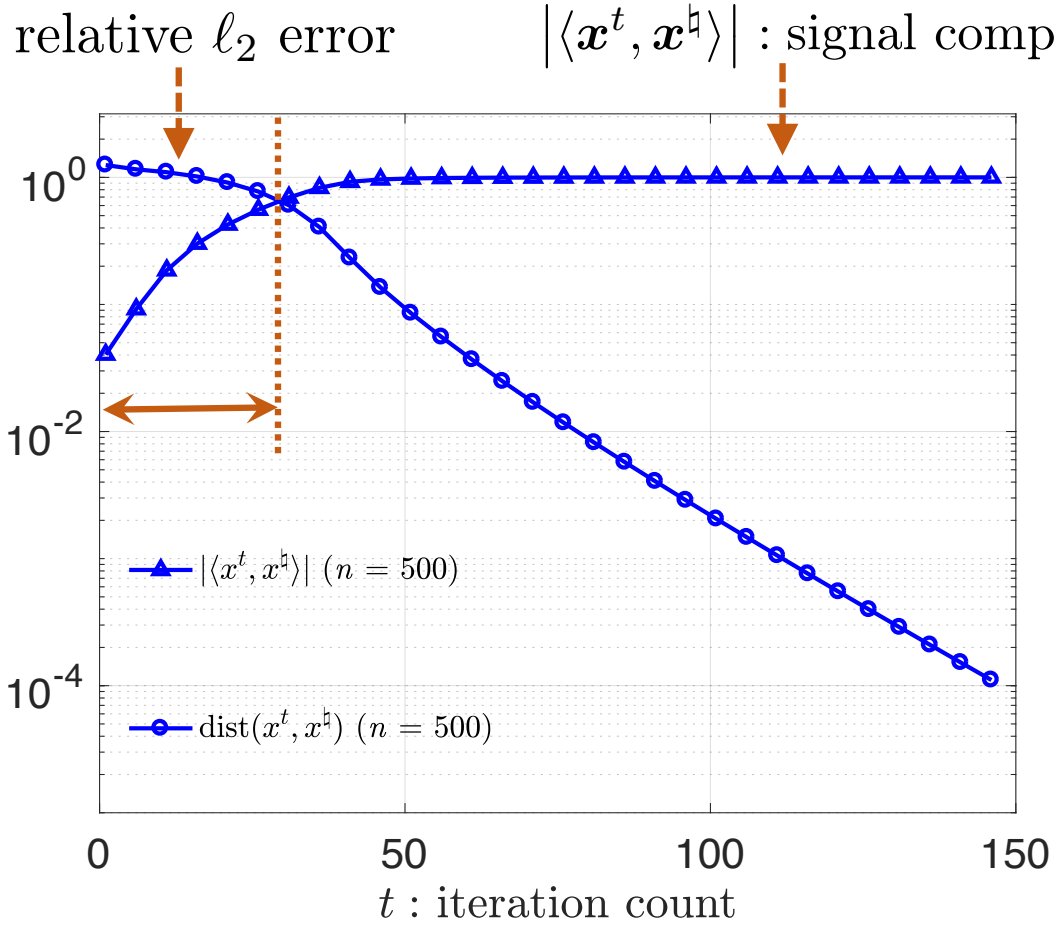
# Exponential growth of signal strength in Stage 1



# Exponential growth of signal strength in Stage 1



# Exponential growth of signal strength in Stage 1




Numerically, a few iterations suffice for entering local region

# Our theory

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

# Our theory

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

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

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

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

# Our theory

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

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

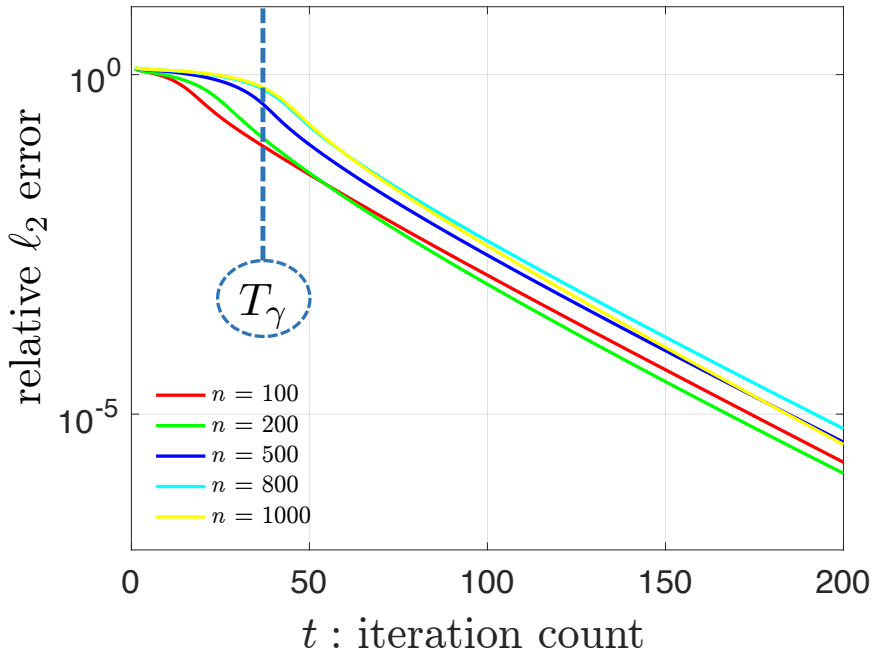
**Theorem 1 (Chen, Chi, Fan, Ma '18)**

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

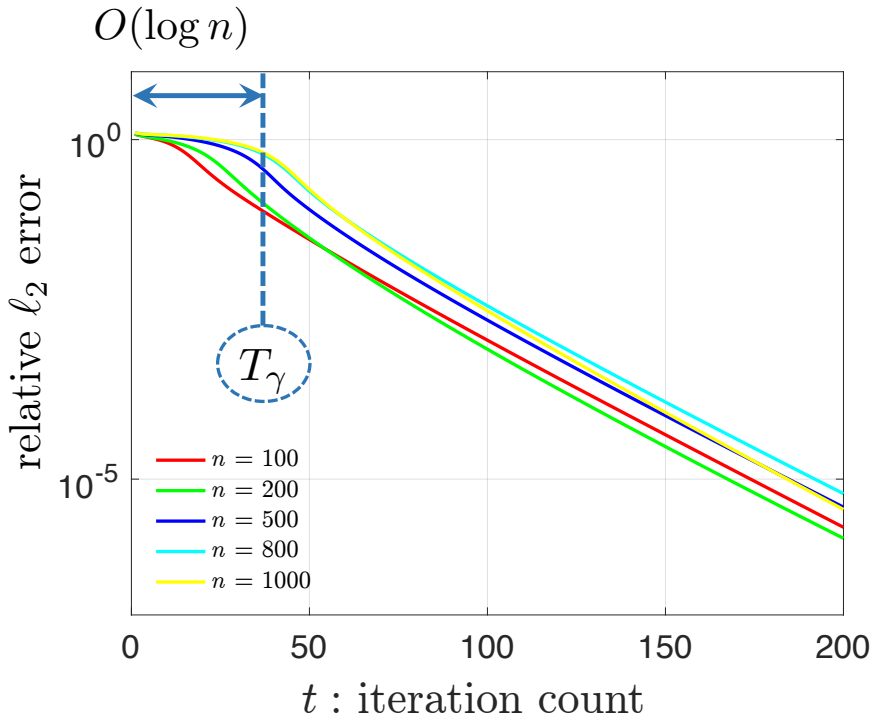
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \qquad t \ge T_\gamma$$

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

# Our theory

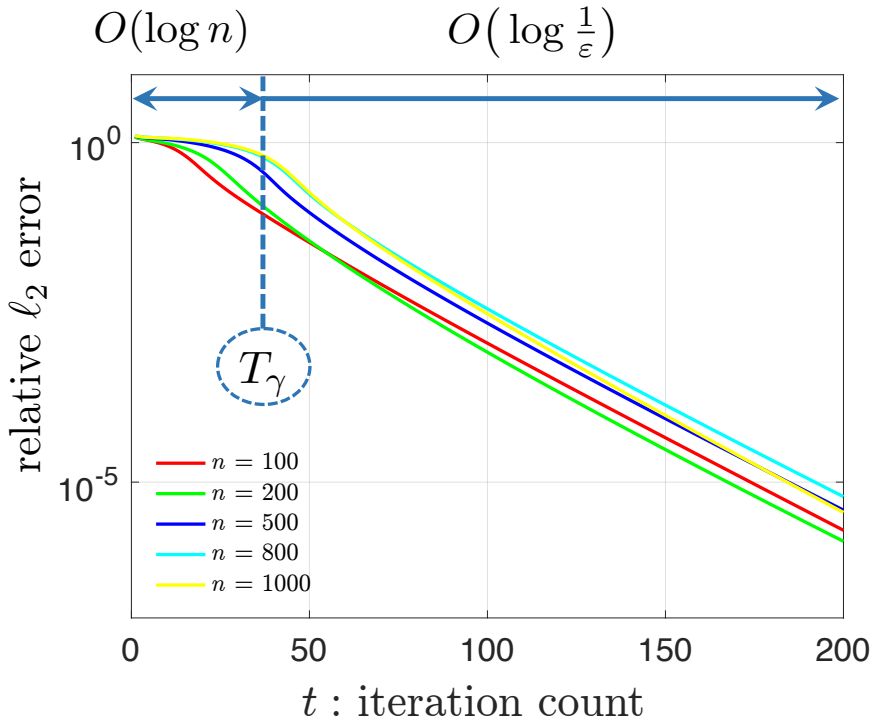

relative $\ell_2$ error
$10^0$
$10^{-5}$
$T_\gamma$
$n = 100$
$n = 200$
$n = 500$
$n = 800$
$n = 1000$
0
50
100
150
200
$t$ : iteration count

# Our theory



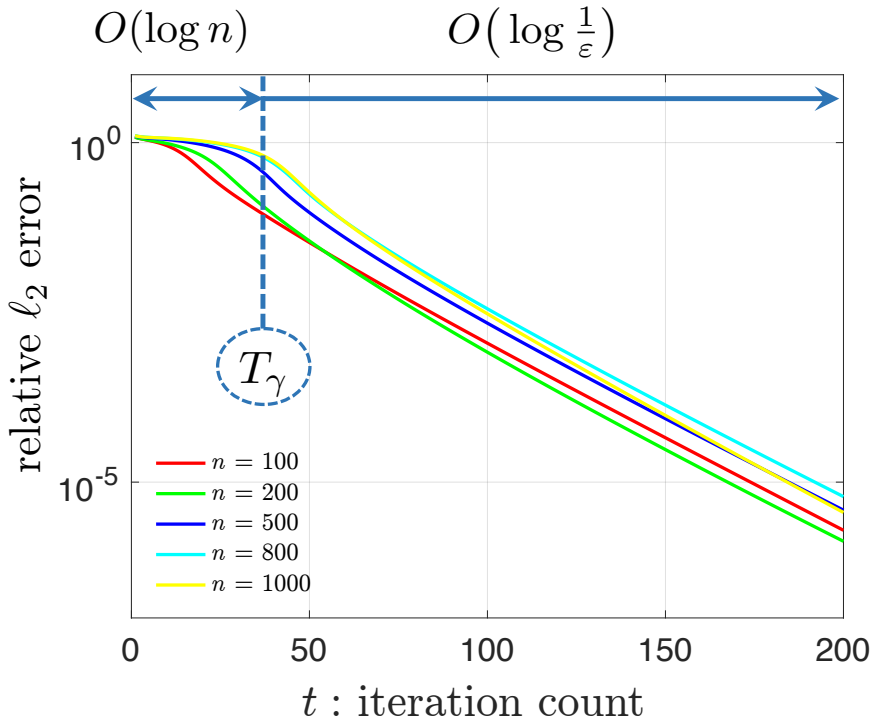

- *Stage 1:* takes $O(\log n)$ iterations to reach $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \underbrace{\gamma}_{\text{small}}$

# Our theory




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

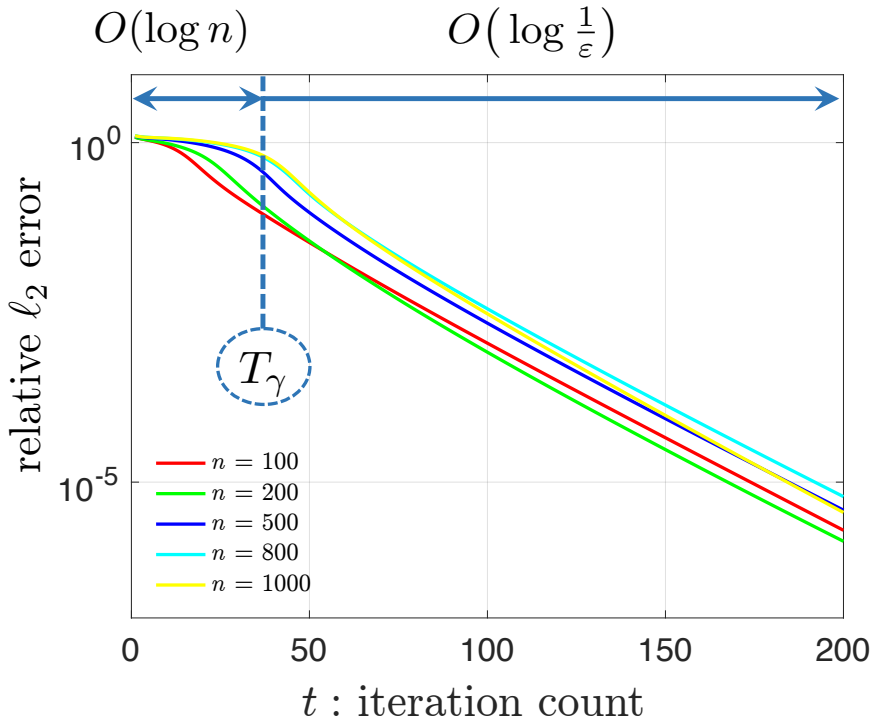
# Our theory




- *near-optimal computational cost*:

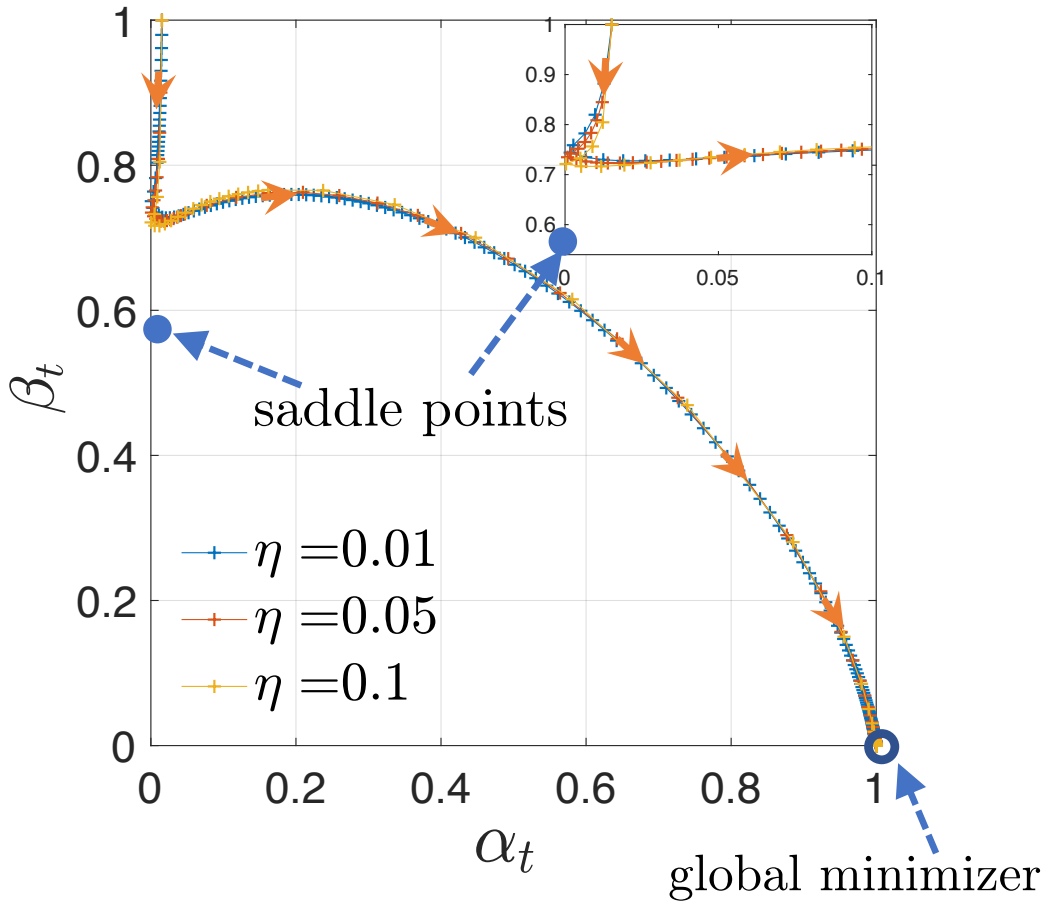
  — $O\big(\log n + \log \frac{1}{\varepsilon}\big)$ iterations to yield $\varepsilon$ accuracy

# Our theory




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

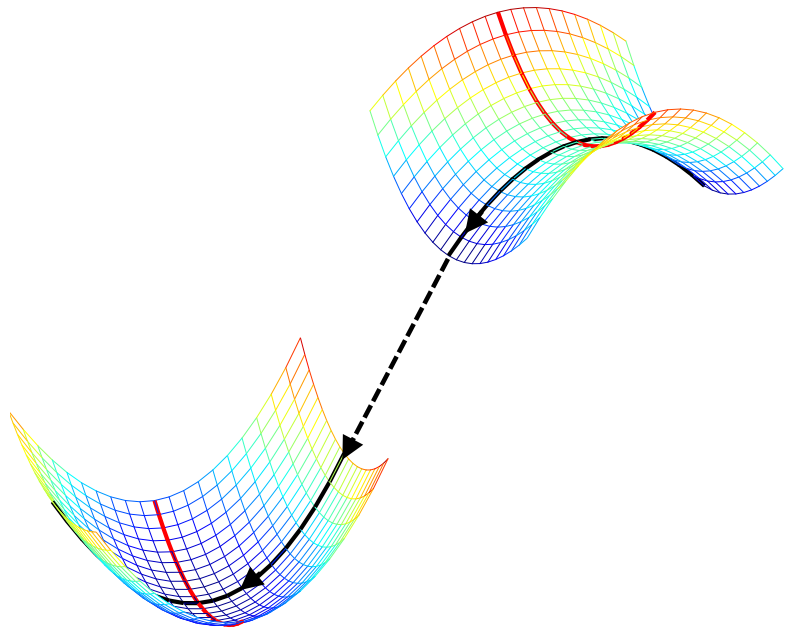
# Saddle-escaping schemes?




Randomly initialized GD never hits saddle points!

# Other saddle-escaping schemes based on generic landscape analysis

| | iteration complexity |
|---|---|
| **trust-region** (Sun et al. '16) | $n^7 + \log\log\frac{1}{\varepsilon}$ |
| **perturbed GD** (Jin et al. '17) | $n^3 + n\log\frac{1}{\varepsilon}$ |
| **perturbed accelerated GD** (Jin et al. '17) | $n^{2.5} + \sqrt{n}\log\frac{1}{\varepsilon}$ |
| **GD (ours)** (Chen et al. '18) | $\log n + \log\frac{1}{\varepsilon}$ |



Generic optimization theory yields highly suboptimal convergence guarantees

# A little analysis

# What if we have infinite samples?

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

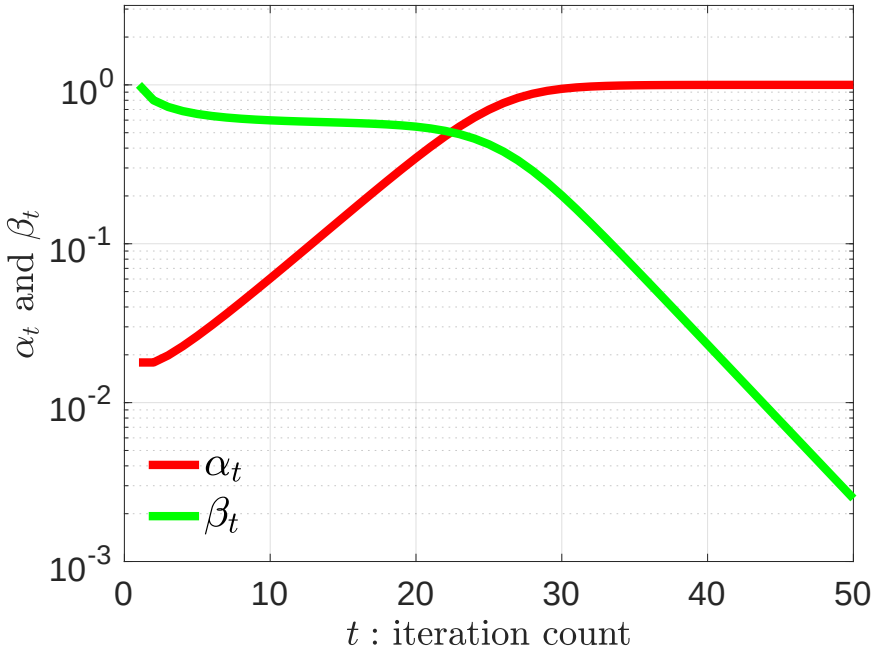
**Population level (infinite samples)**

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

where

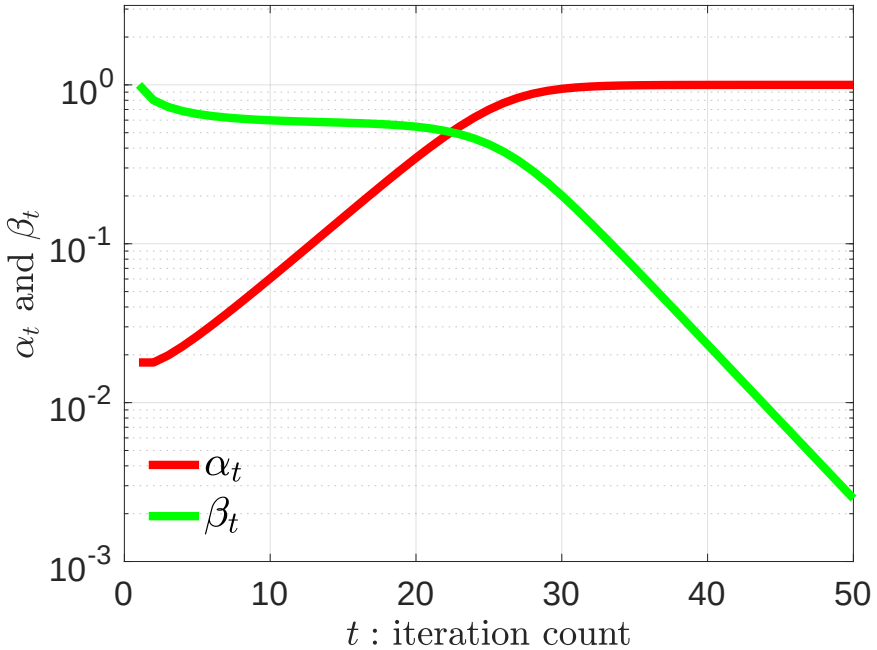
$$\nabla F(\boldsymbol{x}) := \mathbb{E}[\nabla f(\boldsymbol{x})] = (3\|\boldsymbol{x}\|_2^2 - 1)\boldsymbol{x} - 2(\boldsymbol{x}^{\natural\top}\boldsymbol{x})\boldsymbol{x}^{\natural}$$

# 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

# 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

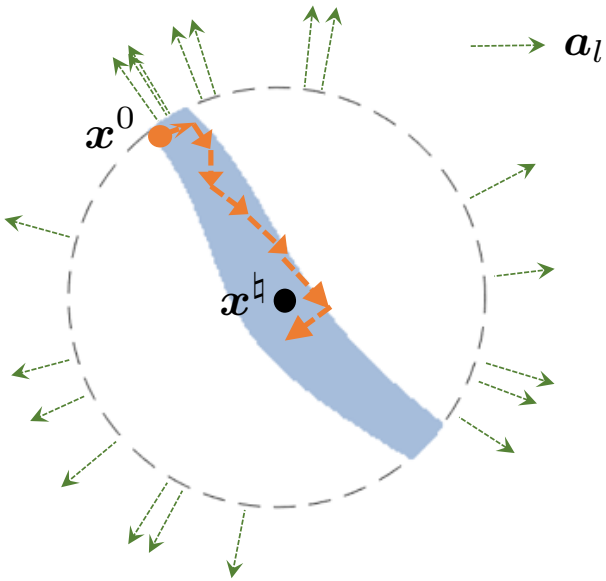
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$$

# Back to finite-sample analysis

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

# Back to finite-sample analysis

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

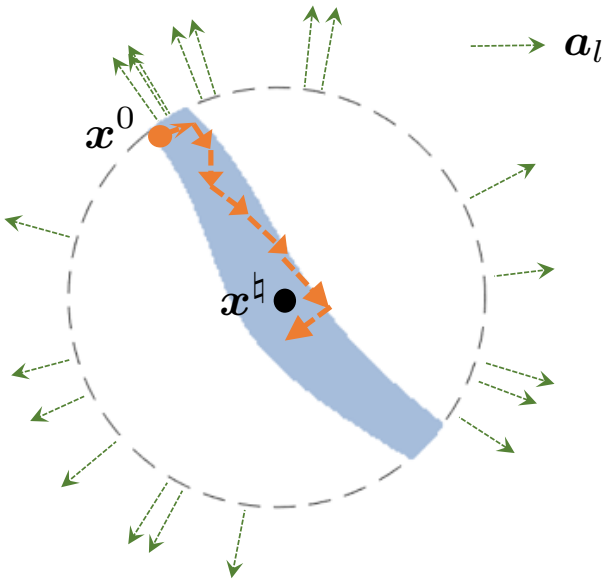



a region with
well-controlled residual

- population-level analysis holds *approximately* if $\boldsymbol{x}^t$ is independent of $\{\boldsymbol{a}_l\}$

# Back to finite-sample analysis

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




a region with
well-controlled residual

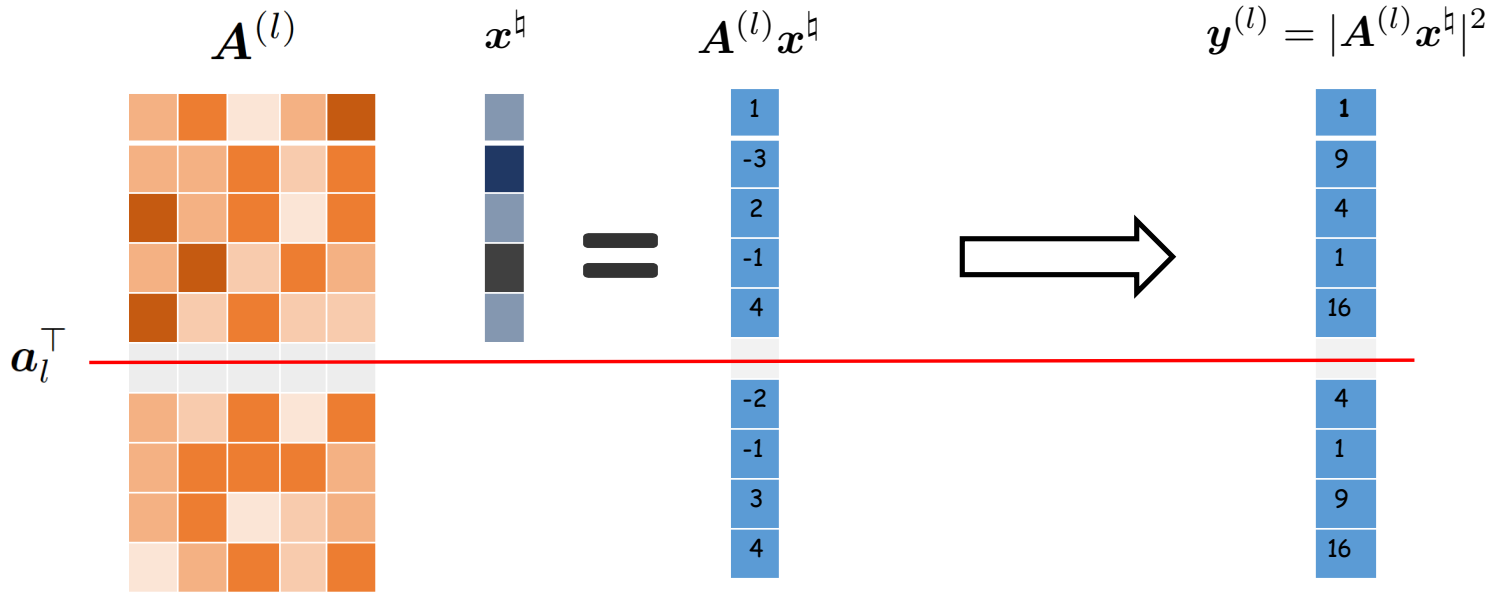
- population-level analysis holds *approximately* if $\boldsymbol{x}^t$ is independent of $\{\boldsymbol{a}_l\}$
- **key analysis ingredient:** show $\boldsymbol{x}^t$ is "nearly-independent" of each $\boldsymbol{a}_l$

# Key proof idea: leave-one-out analysis

Leave out a small amount of information from data and run GD

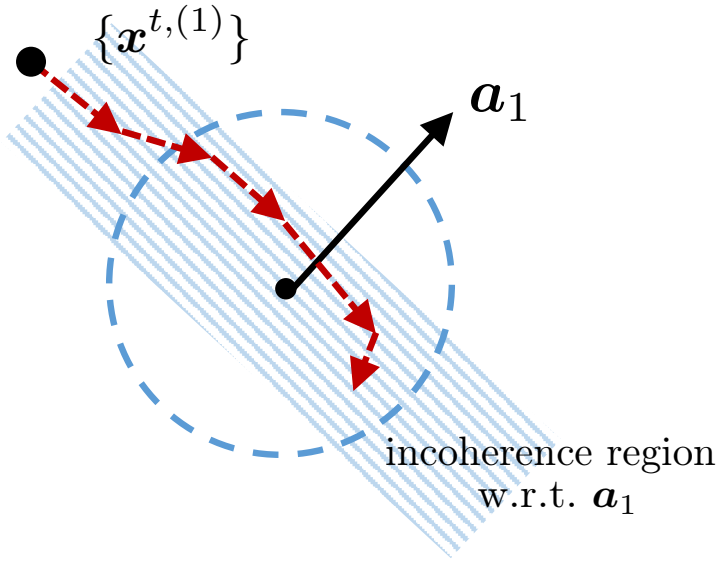
# Key proof idea: leave-one-out analysis

Leave out a small amount of information from data and run GD



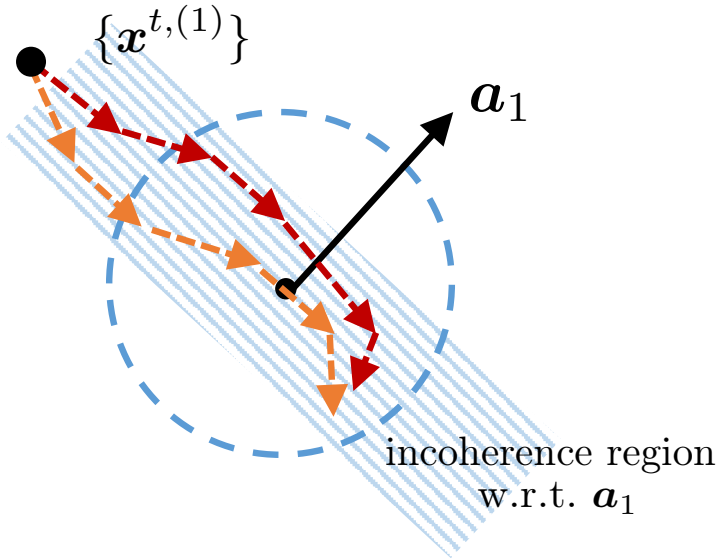

e.g. introduce leave-one-out iterates $\boldsymbol{x}^{t,(l)}$ by running GD without $l$th sample

# Key proof idea: leave-one-out analysis



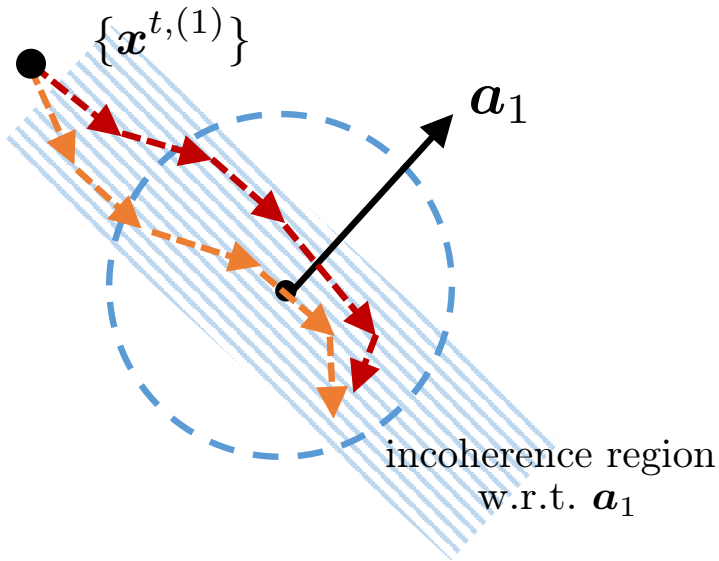

- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$

# Key proof idea: leave-one-out analysis




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

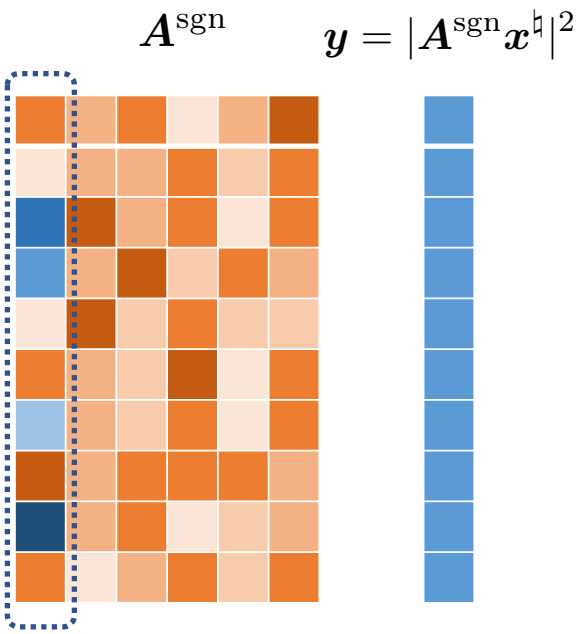
# Key proof idea: leave-one-out analysis



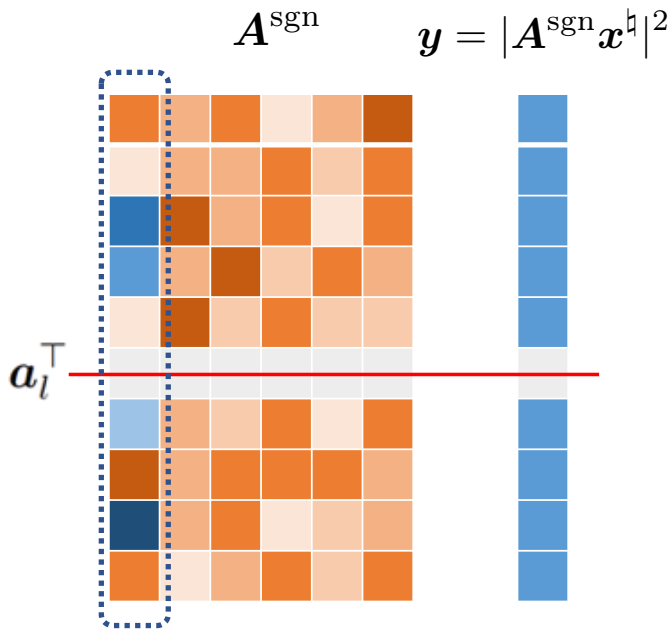

- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)} \approx$ true iterate $\boldsymbol{x}^t$

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

# 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 $\boldsymbol{a}_l$

# Concluding remarks

Even simplest nonconvex methods
are remarkably efficient under suitable statistical models

| smart initialization | sample splitting | saddle escaping |
|---|---|---|
|  |  |  |



"Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval", Y. Chen, Y. Chi, J. Fan, C. Ma, arXiv:1803.07726