Geometry and Regularization in Nonconvex Low-Rank Estimation

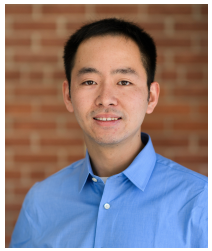
Yuejie Chi

Carnegie Mellon University

University of Texas, Austin
Feb. 2019

Acknowledgements

Thanks to my collaborators:



Y. Chen
Princeton

C. Ma
Princeton

K. Wang
Princeton

Y. Li
CMU

This research is supported by NSF, ONR, AFOSR and ARO.

Empirical risk minimization

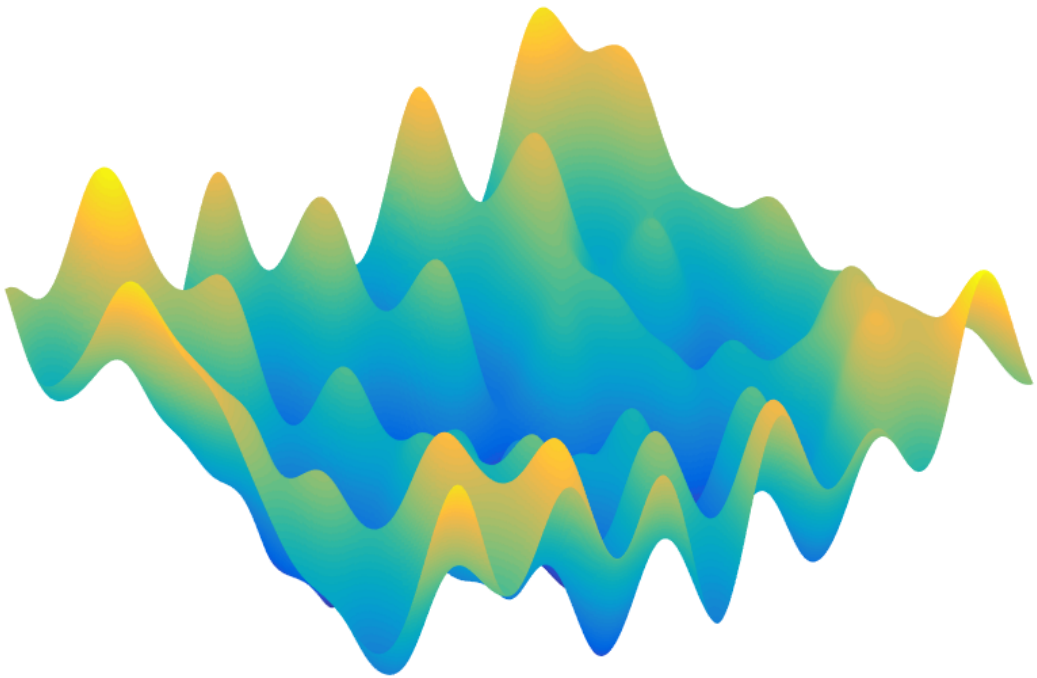
Given data $\boldsymbol{z}$, estimate parameters $\boldsymbol{x} \in \mathbb{R}^n$:

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

Empirical risk minimization

Given data $\boldsymbol{z}$, estimate parameters $\boldsymbol{x} \in \mathbb{R}^n$:

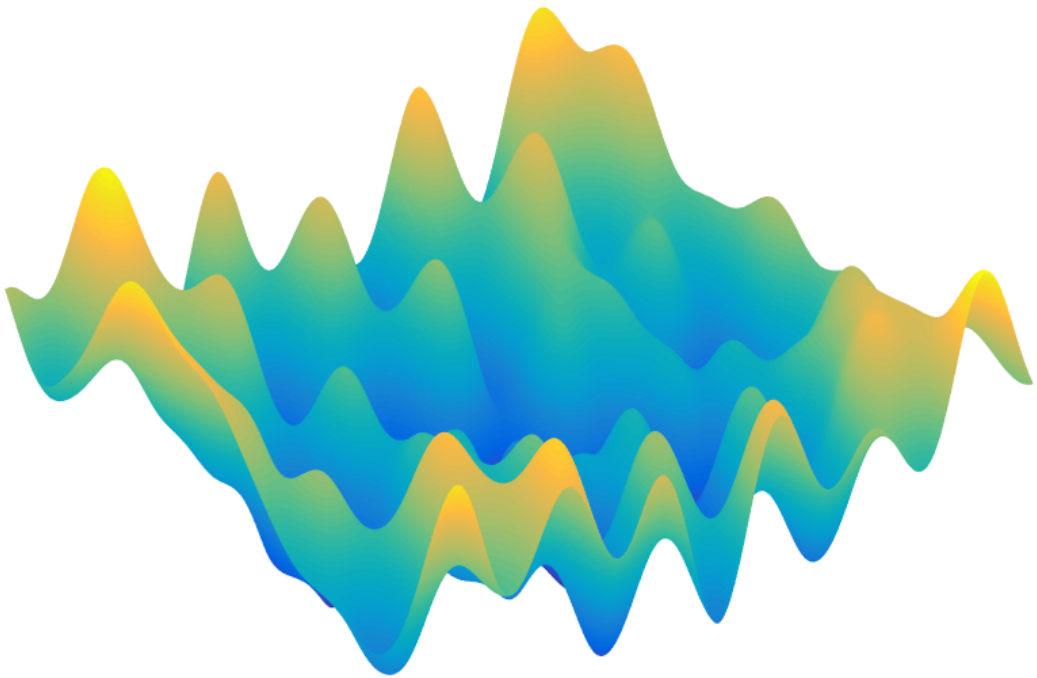
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) := \frac{1}{m} \sum_{i=1}^{m} \ell(z_i; \boldsymbol{x})$$



Empirical risk minimization

Given data $\boldsymbol{z}$, estimate parameters $\boldsymbol{x} \in \mathbb{R}^n$:

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

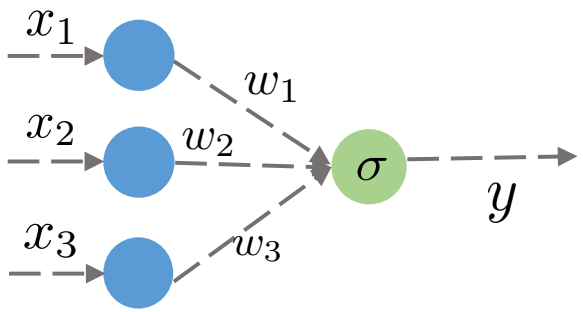


There may be exponentially many local optima

Exponentially many local minima

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

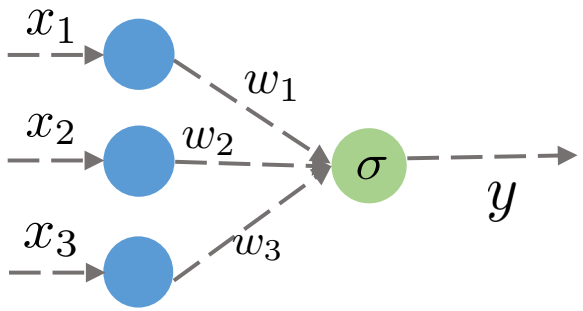
$$\text{minimize}_{\boldsymbol{w}\in\mathbb{R}^n} \quad \ell_m(\boldsymbol{w}) := \frac{1}{2m}\sum_{i=1}^{m}\Big(y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\Big)^2$$



Exponentially many local minima

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

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



Theorem (Auer et al., 1996)

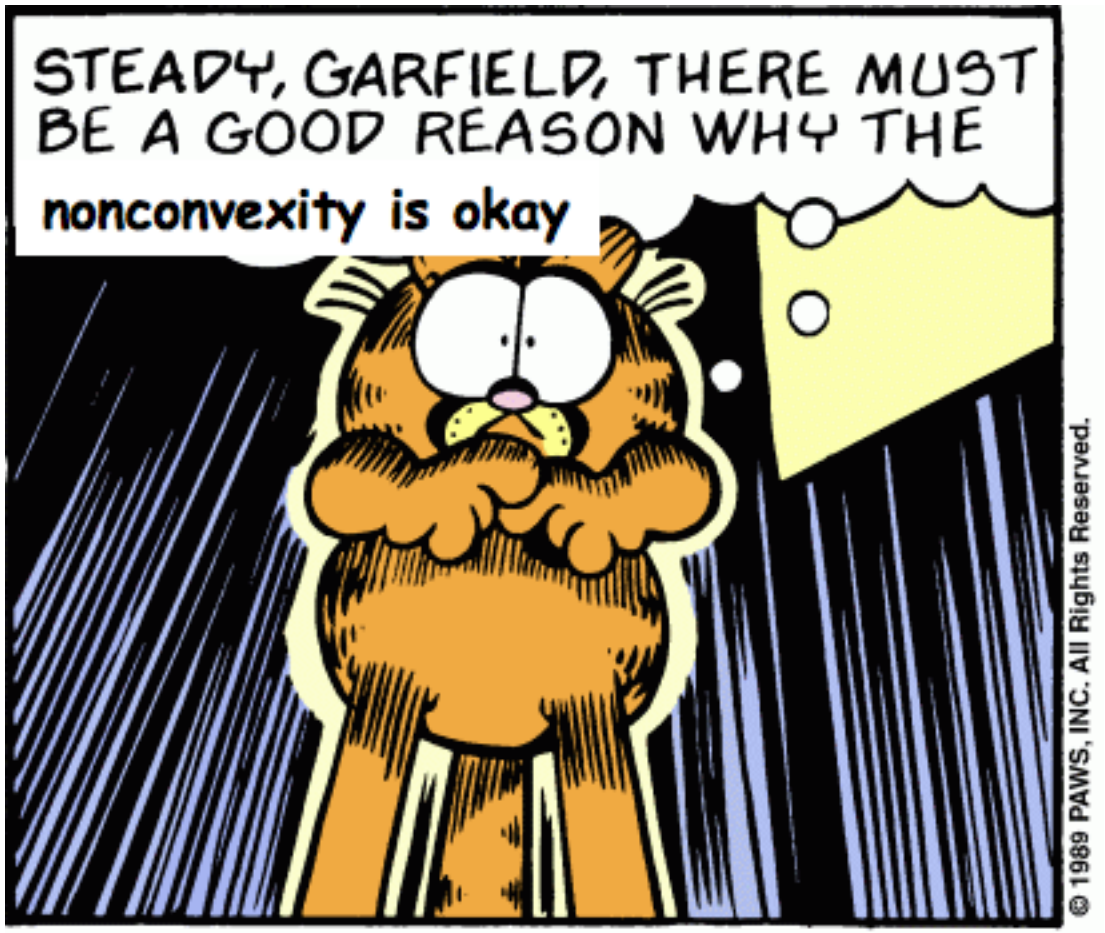
Let $\sigma(\cdot)$ be sigmoid and $\ell(\cdot)$ be the quadratic loss function. There exists a sequence of training samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$ such that $\ell_m(\boldsymbol{w})$ has $\lfloor \frac{m}{n} \rfloor^n$ distinct local minima.

Auer et al, "Exponentially many local minima for single neurons," NIPS 1996.

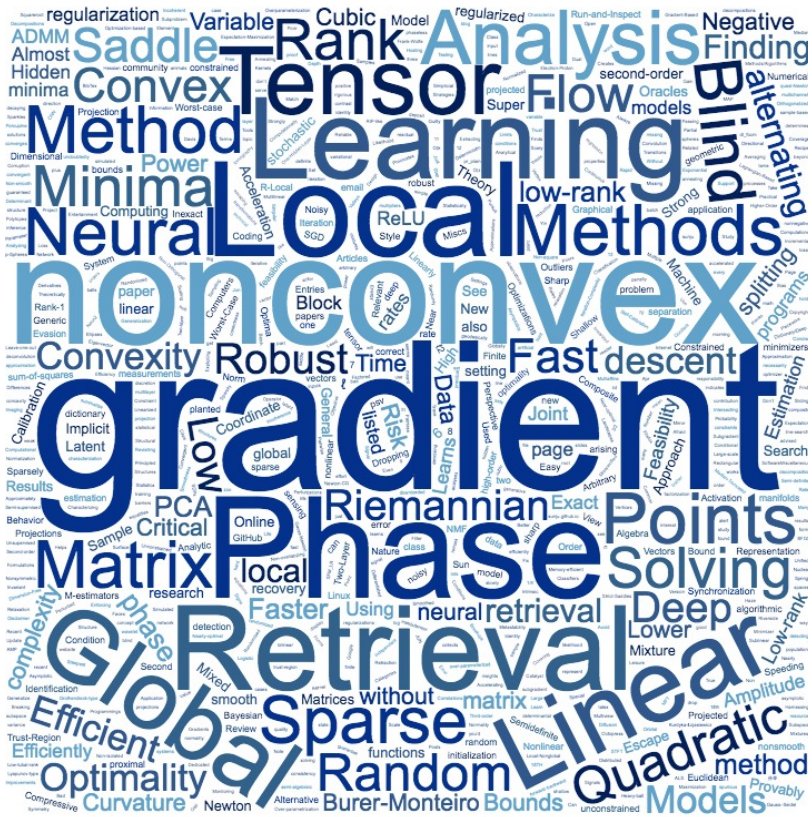
Nonconvex problems are hard!

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...



Recent developments: provable nonconvex optimization



Phase retrieval: Gerchberg-Saxton '72, Netrapalli et al. '13, Candès, Li, Soltanolkotabi '14, Chen, Candès '15, Cai, Li, Ma '15, Zhang et al. 16, Wang et al. '16, Sun et al. '16, Ma et al. '17, Chen et al. '18, ...

Matrix completion: Keshavan et al. '09, Jain et al. '09, Hardt '13, Sun, Luo '15, Chen, Wainwright '15, Zheng, Lafferty '16, Ge et al. '16, Jin et al. '16, Ma et al. '17, ...

Matrix sensing: Jain et al. '13, Tu et al. '15, Zheng, Lafferty '15, Bhojanapalli et al. 16, Li, Zhu, Tang '18, ...

Blind deconvolution / demixing: Li et al. '16, Lee et al. '16, Ling, Strohmer '16, Huang, Hand '16, Ma et al. '17, Zhang et al. '18, Li, Bresler '18, Dong, Shi '18, ...

Dictionary learning: Arora et al. '14, Sun et al. '15, Chatterji, Bartlett '17, ...

Robust principal component analysis: Netrapalli et al. '14, Yi et al. '16, Gu et al. '16, Ge et al. '17, Cherapanamjeri et al. '17, ...

Statistical thinking in nonconvex optimization

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 thinking in nonconvex optimization

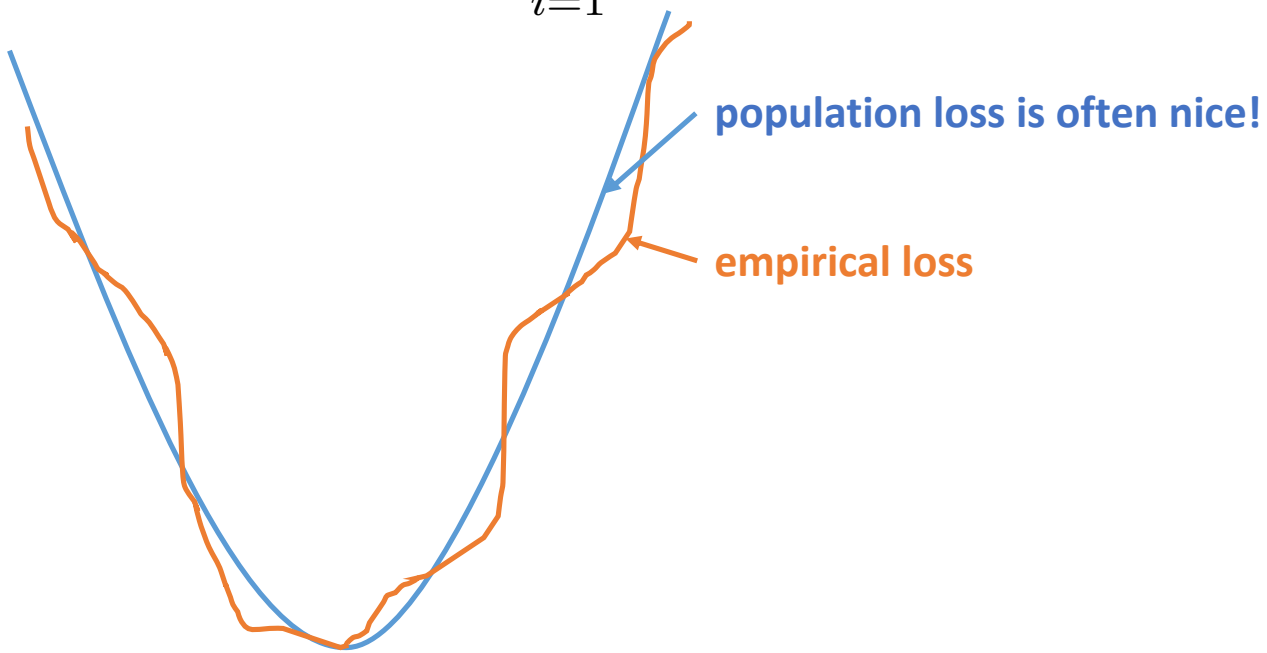
Data/measurements follow certain **statistical models** and hence are not worst-case instances.

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

Statistical thinking in nonconvex optimization

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

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



When $m \to \infty$, empirical risk $\approx$ population risk!

From population risk to empirical risk

Sample-starved / finite-sample regime:

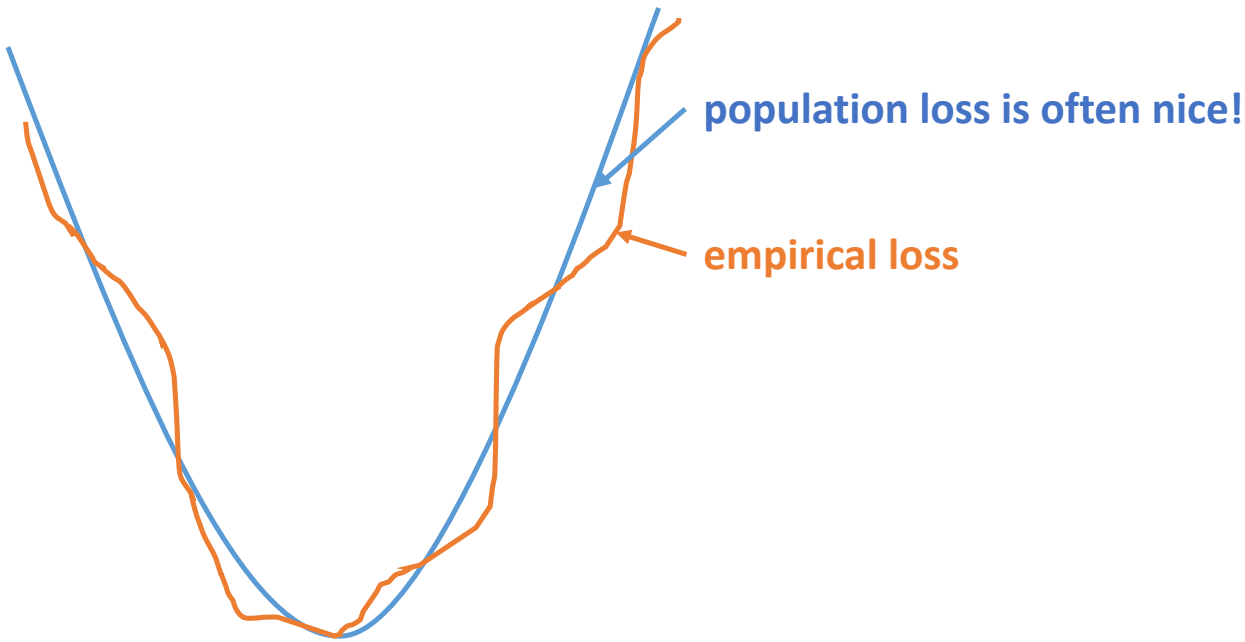
sample size $m \approx \tilde{O}(n)$, the number of parameters.

From population risk to empirical risk

Sample-starved / finite-sample regime:

sample size $m \approx \tilde{O}(n)$, the number of parameters.

Even when $\mathbb{E}[f(\boldsymbol{x})]$ is (locally) strongly convex and smooth, the empirical loss $f(\boldsymbol{x})$ may not be when $m \approx \tilde{O}(n)$.

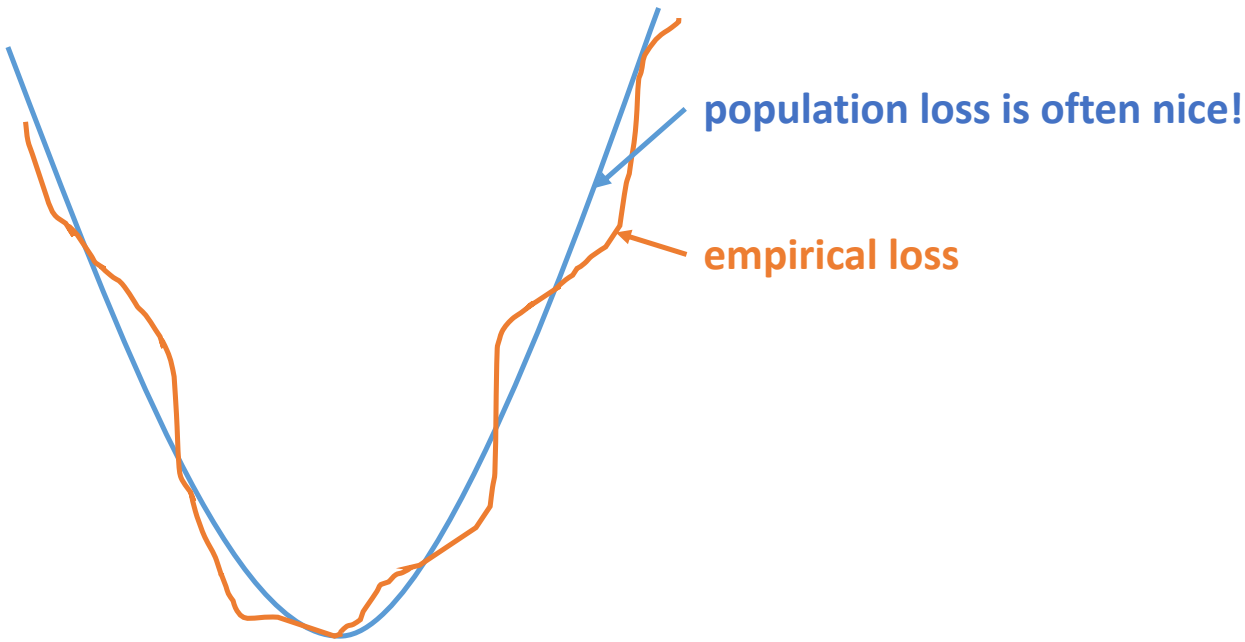


From population risk to empirical risk

Sample-starved / finite-sample regime:

sample size $m \approx \tilde{O}(n)$, the number of parameters.

Even when $\mathbb{E}[f(\boldsymbol{x})]$ is (locally) strongly convex and smooth, the empirical loss $f(\boldsymbol{x})$ may not be when $m \approx \tilde{O}(n)$.



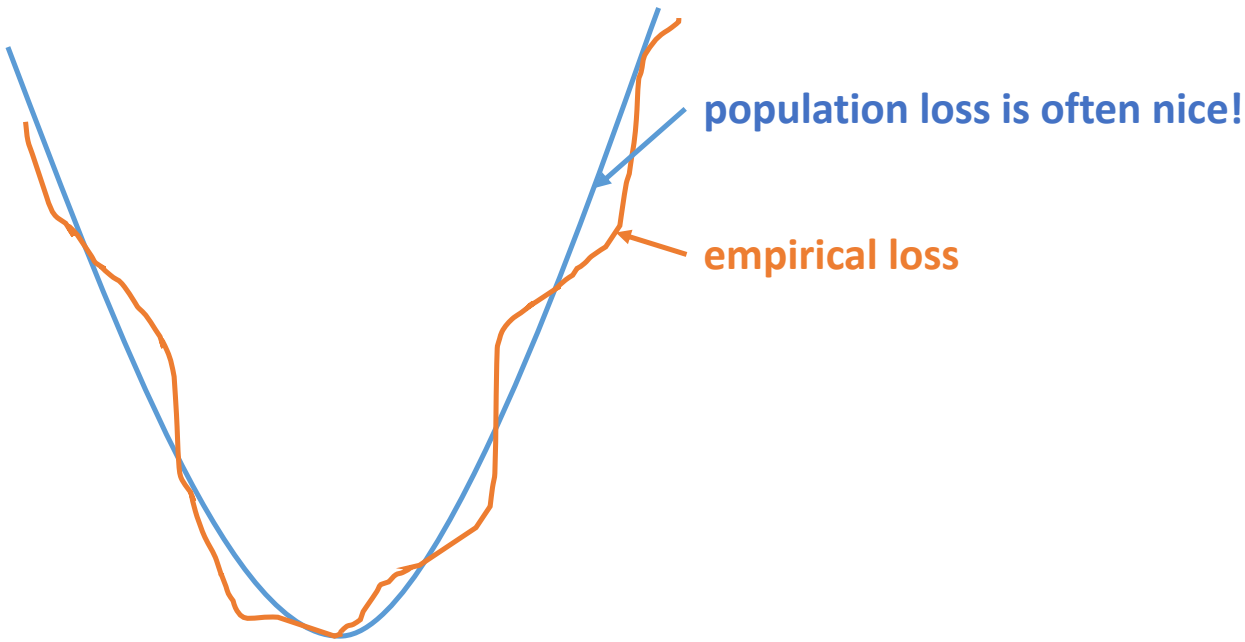
$f(\boldsymbol{x})$ may lack curvatures in certain regions/directions!

From population risk to empirical risk

Sample-starved / finite-sample regime:

sample size $m \approx \tilde{O}(n)$, the number of parameters.

Even when $\mathbb{E}[f(\boldsymbol{x})]$ is (locally) strongly convex and smooth, the empirical loss $f(\boldsymbol{x})$ may not be when $m \approx \tilde{O}(n)$.

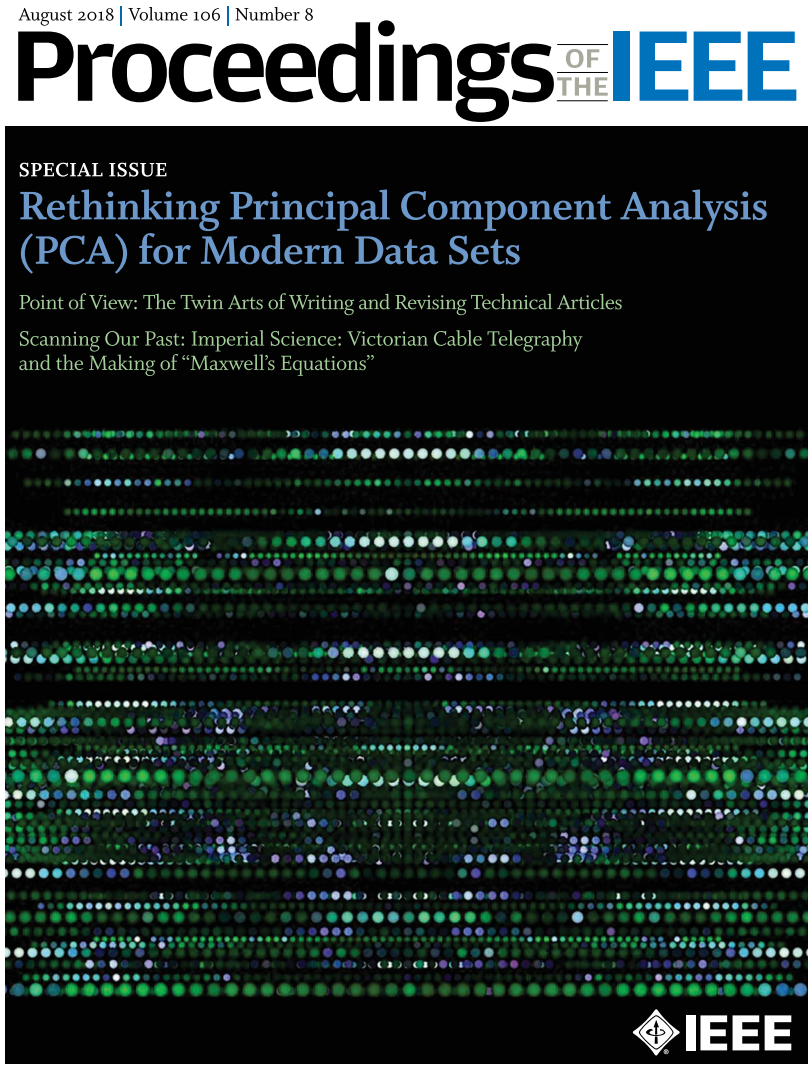


$f(\boldsymbol{x})$ may lack curvatures in certain regions/directions!

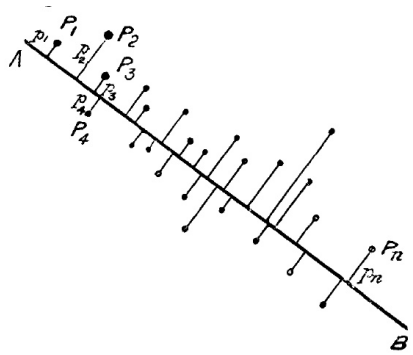
Will this be problematic?

a case study with low-rank matrix estimation

Rethinking PCA for modern datasets

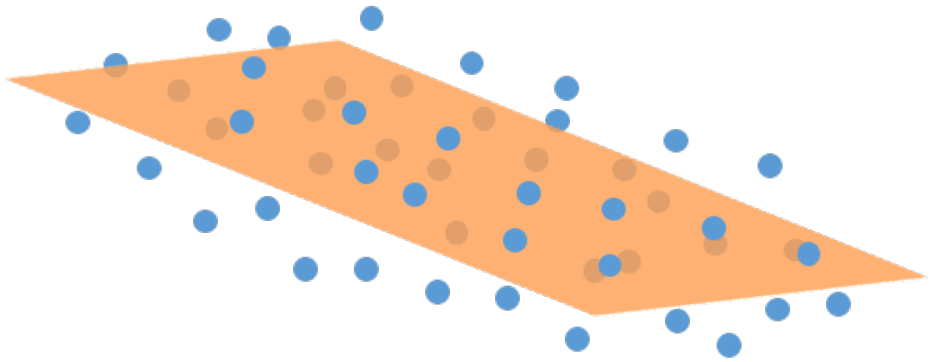


Classical PCA (Pearson 1901):



Modern PCA (Big, Messy Data):

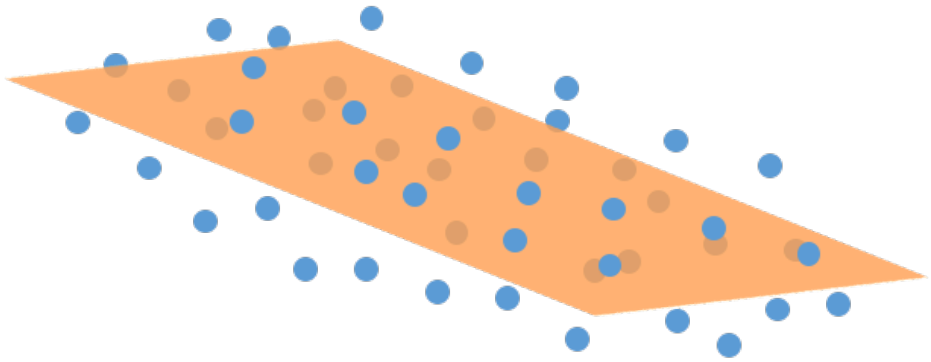
Revisiting PCA: in search of low-rank representation



Given $\boldsymbol{M} \succeq 0 \in \mathbb{R}^{n \times n}$ (e.g. sample covariance matrix), find its best rank-r approximation:

$$\underbrace{\widehat{\boldsymbol{M}} = \mathsf{argmin}_{\boldsymbol{Z}} \, \|\boldsymbol{Z} - \boldsymbol{M}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \mathsf{rank}(\boldsymbol{Z}) \le r}_{\text{nonconvex optimization!}}$$

Revisiting PCA: in search of low-rank representation



This problem admits a closed-form solution:

- let $\boldsymbol{M} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be eigen-decomposition of $\boldsymbol{M}$ ($\lambda_1 \geq \cdots \geq \lambda_n$), then

$$\widehat{\boldsymbol{M}} = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$$

— *nonconvex, but tractable*

An optimization viewpoint

Burer-Monteiro factorization: if we factorize $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, then it leads to a nonconvex problem:

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

An optimization viewpoint

Burer-Monteiro factorization: if we factorize $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, then it leads to a nonconvex problem:

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

- **Pro:** reduce parameter space from $O(n^2)$ to $O(n)$;
- **Con:** nonconvex and susceptible to local optima.

An optimization viewpoint

Burer-Monteiro factorization: if we factorize $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, then it leads to a nonconvex problem:

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

- **Pro:** reduce parameter space from $O(n^2)$ to $O(n)$;
- **Con:** nonconvex and susceptible to local optima.

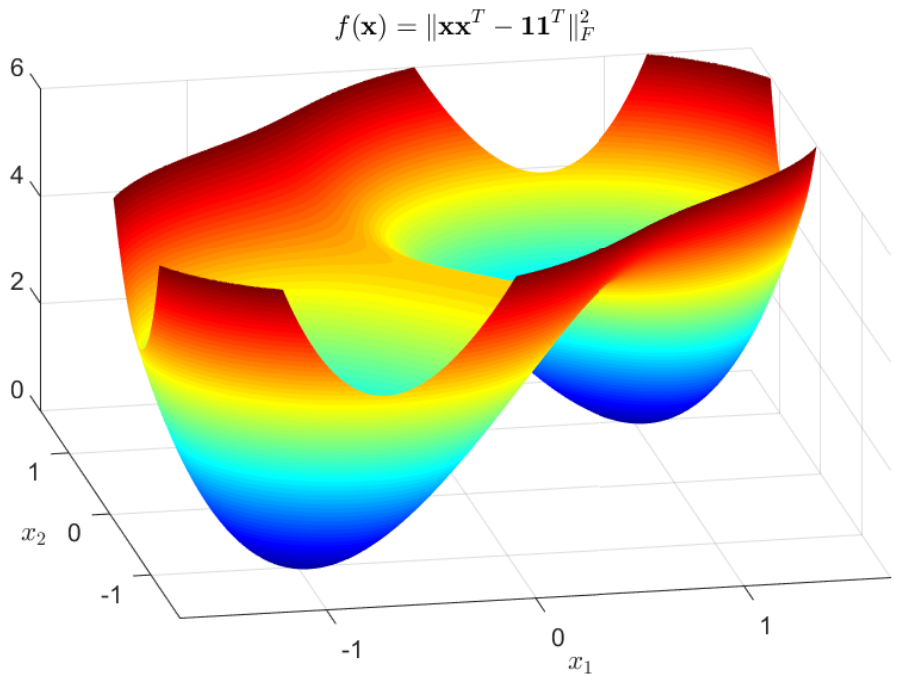
Theorem (PCA doesn't have spurious local minima, Baldi and Hornik, 1989)

Suppose $\boldsymbol{M}$ has a strict eigen-gap between λ_r and λ_{r+1}, the critical points of $f(\boldsymbol{X})$ can be categorized into

- *global minima;*
- *strict saddle points, from which there exist directions to strictly decrease $f(\boldsymbol{X})$.*

Benign landscape of PCA

For example, for 2-dimensional case $f(\boldsymbol{x}) = \left\| \boldsymbol{x}\boldsymbol{x}^\top - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{F}}^2$

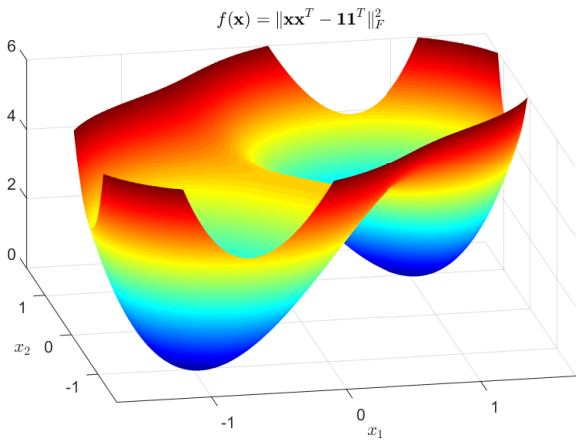


global minima: $\boldsymbol{x} = \pm \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; strict saddles: $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\pm \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

— No "spurious" local minima!

Parameter recovery via gradient descent

a two-step recovery strategy:



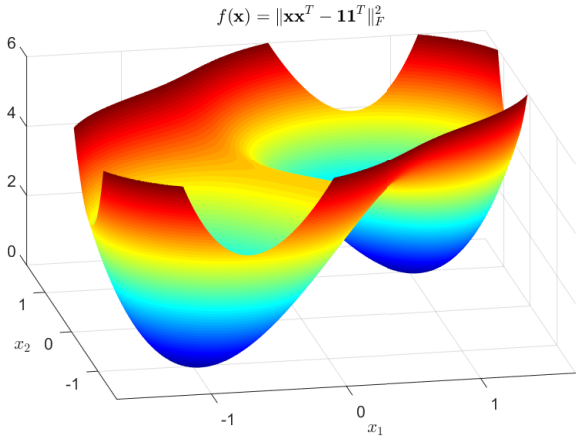
- **Initialization by spectral method**
- **Gradient iterations:**

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

for $t = 0, 1, \ldots$

Parameter recovery via gradient descent

a two-step recovery strategy:



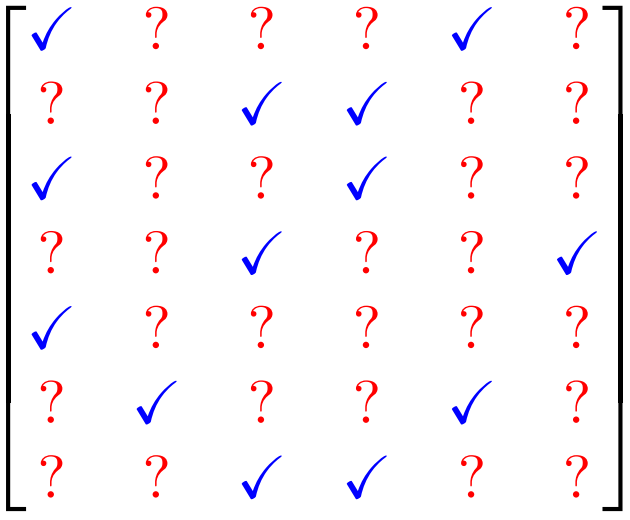
- **Initialization by spectral method**
- **Gradient iterations:**

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

for $t = 0, 1, \ldots$

- The initial point falls into a "basin of attraction";
- Low-complexity local refinements via gradient descent.

Low-rank matrix completion: dealing with missing data



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

$$\text{find low-rank } \widehat{\boldsymbol{M}} \quad \text{s.t.} \quad \mathcal{P}_{\Omega}(\widehat{\boldsymbol{M}}) = \mathcal{P}_{\Omega}(\boldsymbol{M})$$

Applications: recommendation systems, ...

A natural least-squares formulation

$$\text{given:} \quad \mathcal{P}_\Omega(\boldsymbol{M})$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad f(\boldsymbol{X}) = \left\| \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}) \right\|_{\mathrm{F}}^2$$

A natural least-squares formulation

$$\text{given:} \quad \mathcal{P}_\Omega(\boldsymbol{M})$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad f(\boldsymbol{X}) = \left\| \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}) \right\|_{\mathrm{F}}^2$$

- **Bernoulli sampling:** *Assume every entry is observed i.i.d. with* $0 < p \leq 1$*:*

$$\mathbb{E}[f(\boldsymbol{X})] = p \left\| \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M} \right\|_{\mathrm{F}}^2.$$

does not imply optimization efficiency!

Incoherence

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}$$

vs.

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}$$

Incoherence

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

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 μ-incoherent if

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

Note: $\|\boldsymbol{U}\|_{2,\infty} = \max_i \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2$.

Lower bound [Candès and Tao]: $p \gtrsim \mu r \log n / n$.

Incoherence

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

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 μ-incoherent if

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

Note: $\|\boldsymbol{U}\|_{2,\infty} = \max_i \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2$.

Lower bound [Candès and Tao]: $p \gtrsim \mu r \log n / n$.

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 μ-incoherent if

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

Note: $\|\boldsymbol{U}\|_{2,\infty} = \max_i \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2$.

Lower bound [Candès and Tao]: $p \gtrsim \mu r \log n / n$.

What does the population level look like?

Assume every entry is observed i.i.d. with probability $0 < p \leq 1$.

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

What does the population level look like?

Assume every entry is observed i.i.d. with probability $0 < p \leq 1$.

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

Population level ($p = 1$): this is PCA.

$$\begin{aligned}&\operatorname{vec}(\boldsymbol{V})^\top \mathbb{E}\left[\nabla^2 f(\boldsymbol{X})\right] \operatorname{vec}(\boldsymbol{V}) \\ &= \underbrace{\frac{1}{2}\left\|\boldsymbol{V}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{V}^\top\right\|_{\mathrm{F}}^2 + \left\langle \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}, \boldsymbol{V}\boldsymbol{V}^\top \right\rangle}_{\textit{locally}\text{ restricted strongly convex and smooth}}\end{aligned}$$

along descent direction $\boldsymbol{V}$ when $\boldsymbol{X}$ is close to $\boldsymbol{X}^\natural$.

Consequence: GD converges within $O\left(\log \frac{1}{\varepsilon}\right)$ iterations if $p = 1$.

What does the finite-sample level look like?

Assume every entry is observed i.i.d. with probability $0 < p \leq 1$.

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

What does the finite-sample level look like?

Assume every entry is observed i.i.d. with probability $0 < p \leq 1$.

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

Finite-sample level ($p \asymp \frac{\mathsf{polylog} n}{n}$)

$\nabla^2 f(\boldsymbol{X})$ strongly convex and smooth

along descent direction $\boldsymbol{V}$ only when $\boldsymbol{X}$ is incoherent:

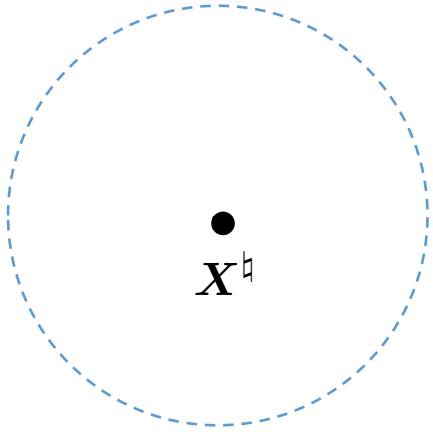
$$\|\boldsymbol{X} - \boldsymbol{X}^\natural\|_{2,\infty} \ll \|\boldsymbol{X}^\natural\|_{2,\infty}$$

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

Incoherence region

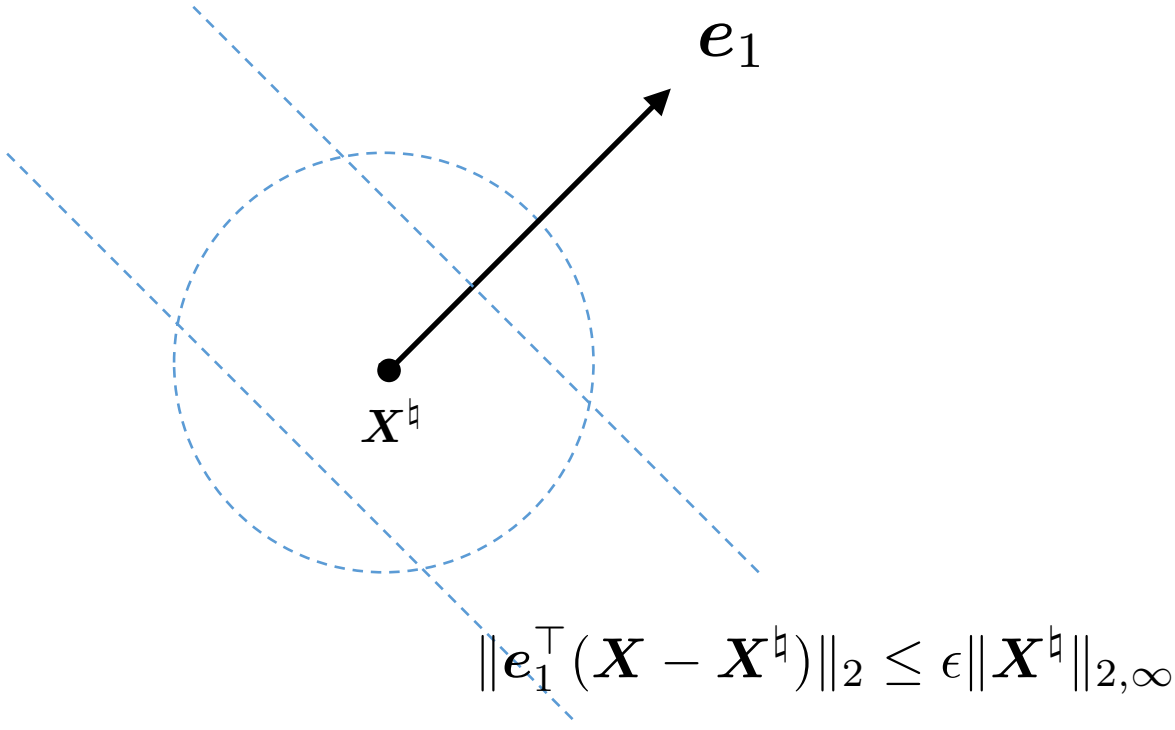
Which region enjoys both restricted strong convexity and smoothness?



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

Incoherence region

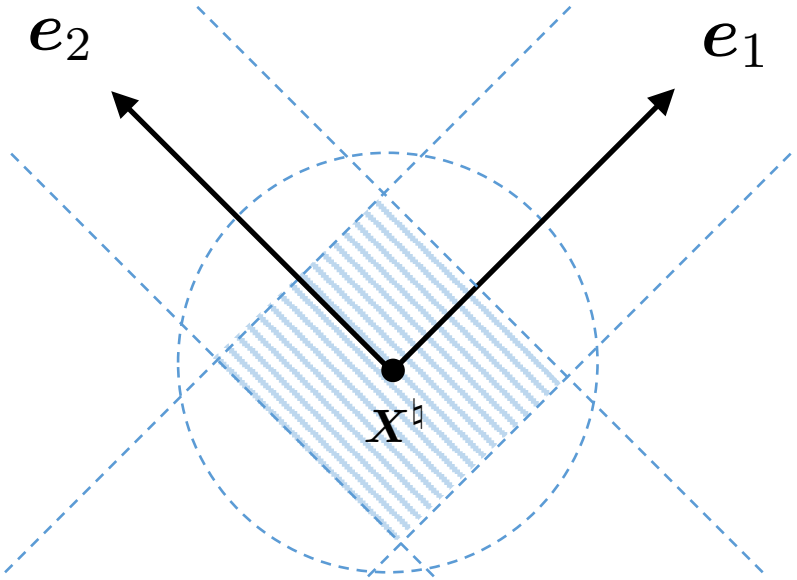
Which region enjoys both restricted strong convexity and smoothness?



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

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

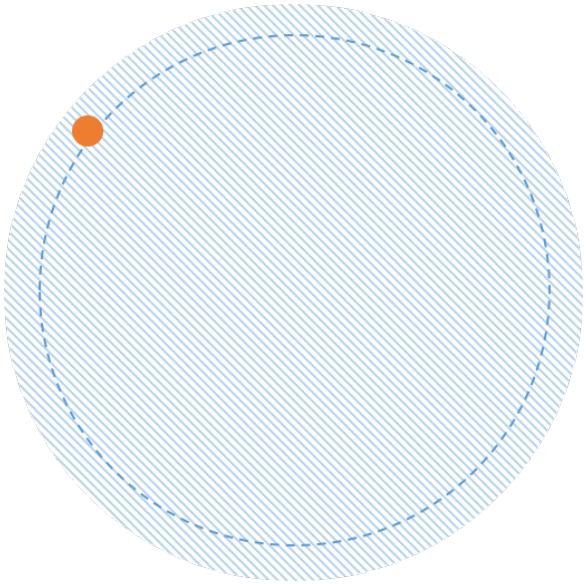


$$\|\boldsymbol{e}_2^\top(\boldsymbol{X}-\boldsymbol{X}^\natural)\|_2 \le \epsilon\|\boldsymbol{X}^\natural\|_{2,\infty} \qquad \|\boldsymbol{e}_1^\top(\boldsymbol{X}-\boldsymbol{X}^\natural)\|_2 \le \epsilon\|\boldsymbol{X}^\natural\|_{2,\infty}$$

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

Vanilla gradient descent is at risk

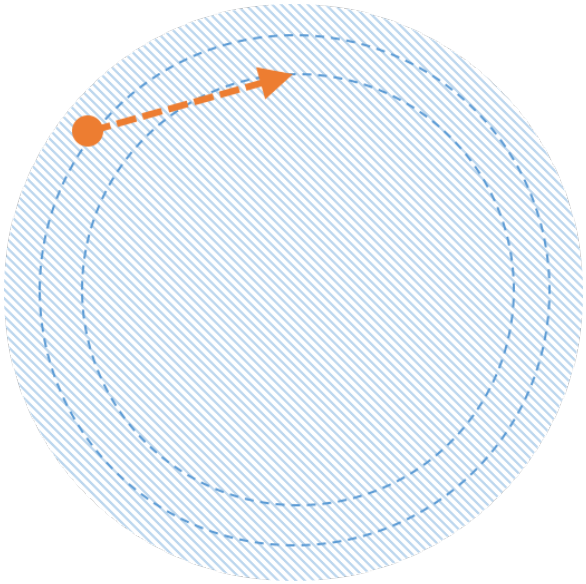
region of local strong convexity + smoothness



GD on the pop. loss

Vanilla gradient descent is at risk

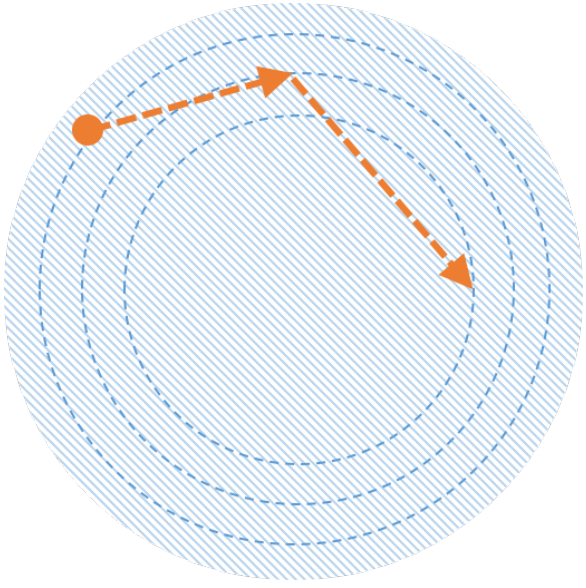
region of local strong convexity + smoothness



GD on the pop. loss

Vanilla gradient descent is at risk

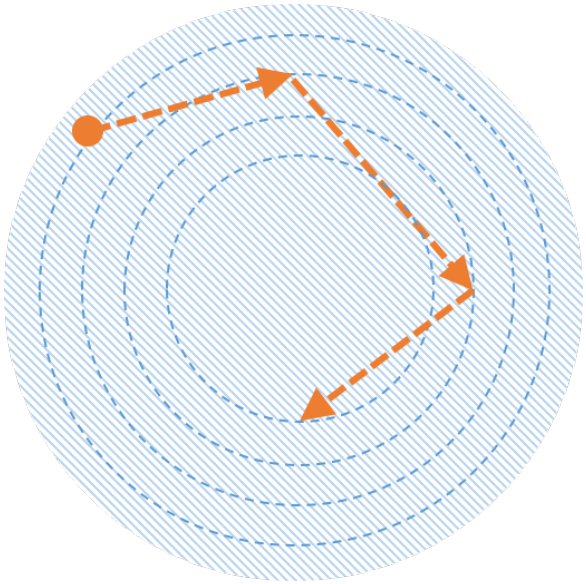
region of local strong convexity + smoothness



GD on the pop. loss

Vanilla gradient descent is at risk

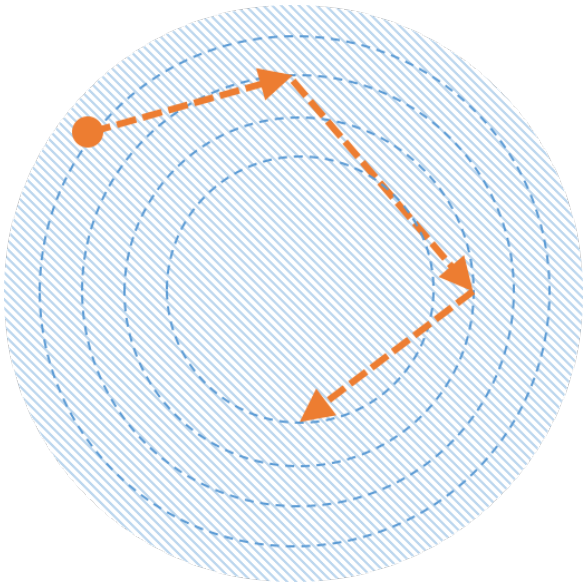
region of local strong convexity + smoothness

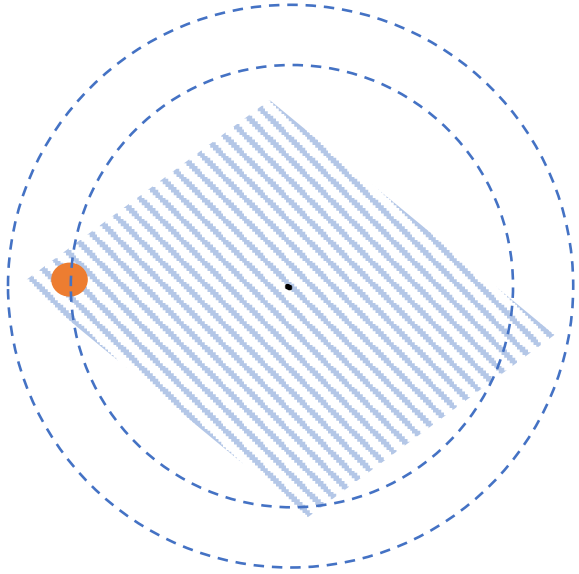


GD on the pop. loss

Vanilla gradient descent is at risk

region of local strong convexity + smoothness



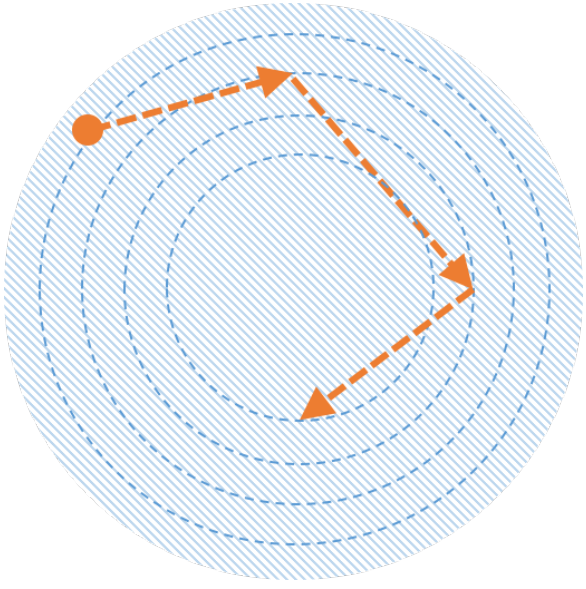


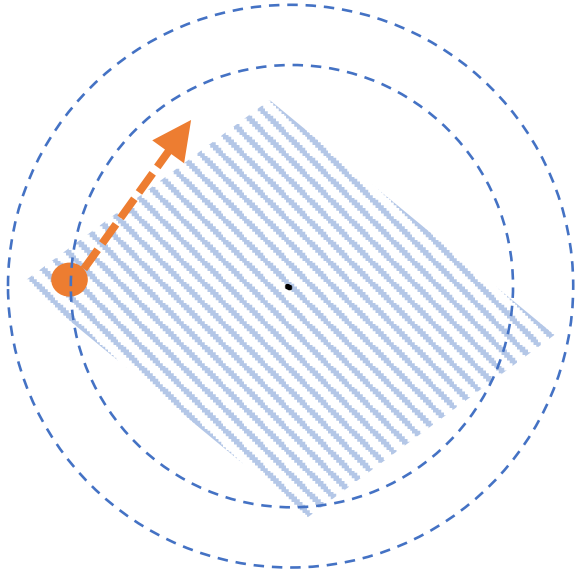
GD on the pop. loss *GD on the emp. loss*

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

Vanilla gradient descent is at risk

region of local strong convexity + smoothness



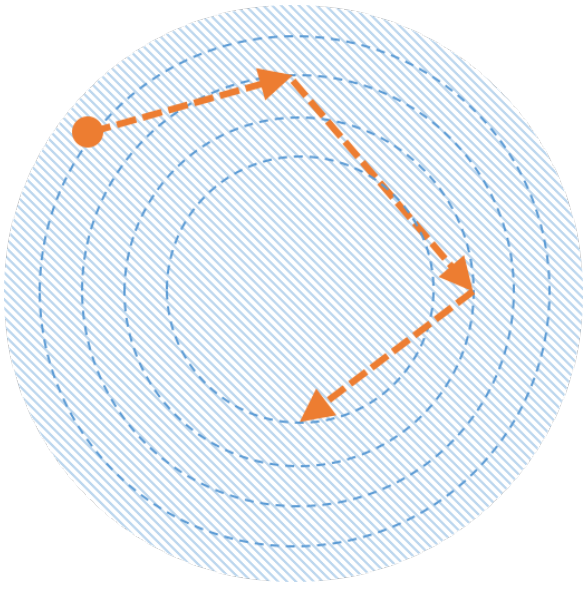


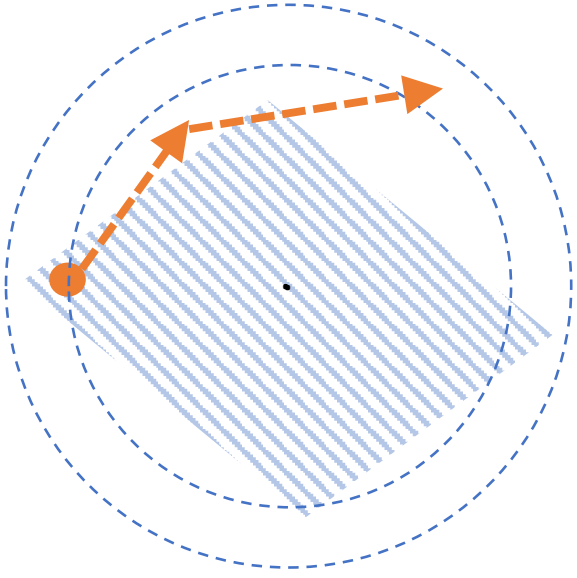
GD on the pop. loss *GD on the emp. loss*

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

Vanilla gradient descent is at risk

region of local strong convexity + smoothness



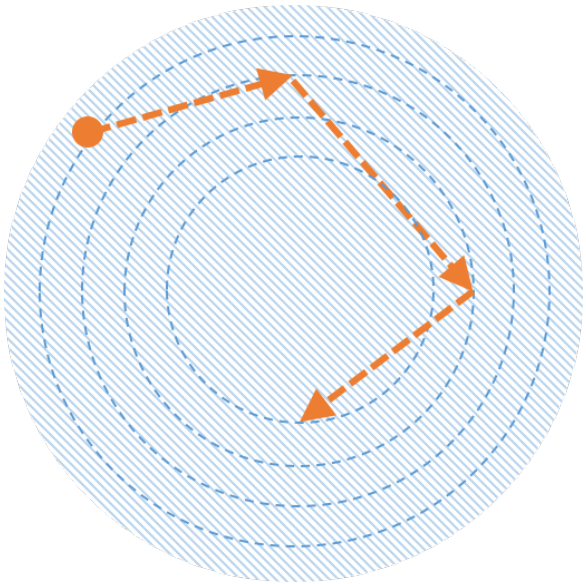


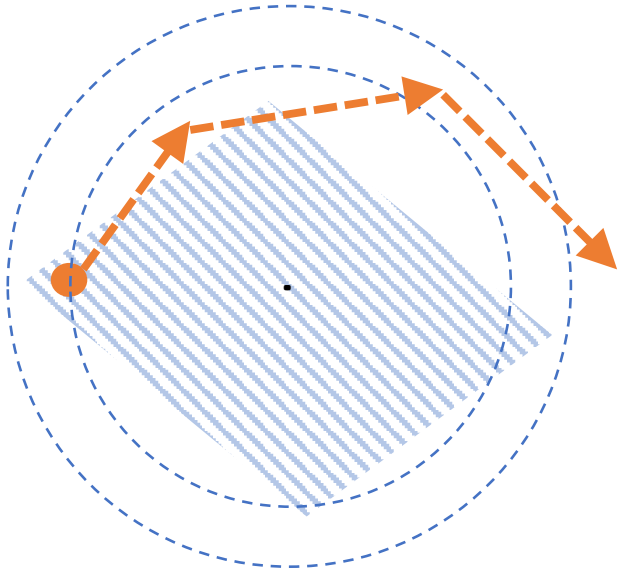
GD on the pop. loss *GD on the emp. loss*

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

Vanilla gradient descent is at risk

region of local strong convexity + smoothness



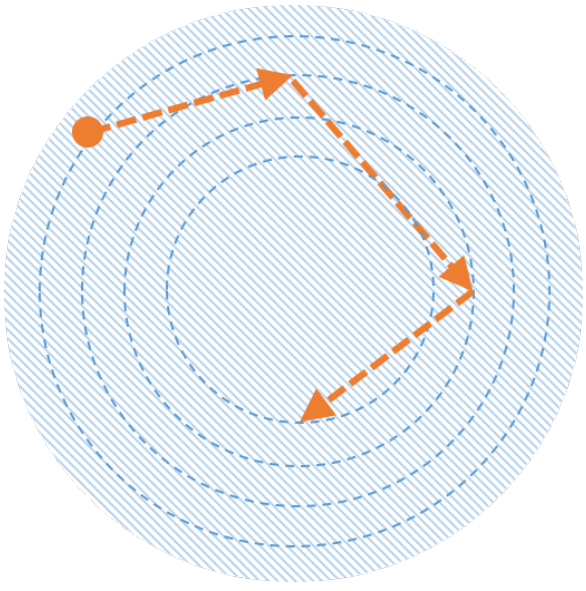


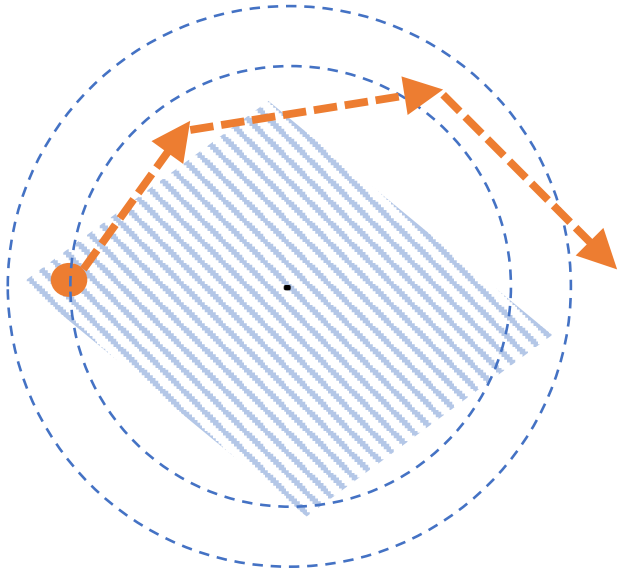
GD on the pop. loss *GD on the emp. loss*

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

Vanilla gradient descent is at risk

region of local strong convexity + smoothness



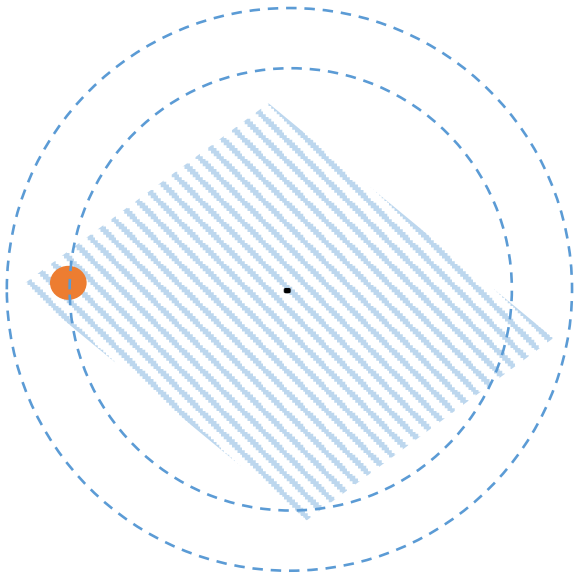


GD on the pop. loss *GD on the emp. loss*

- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region
- *Existing algorithms enforce regularization, or apply sample splitting to promote incoherence*

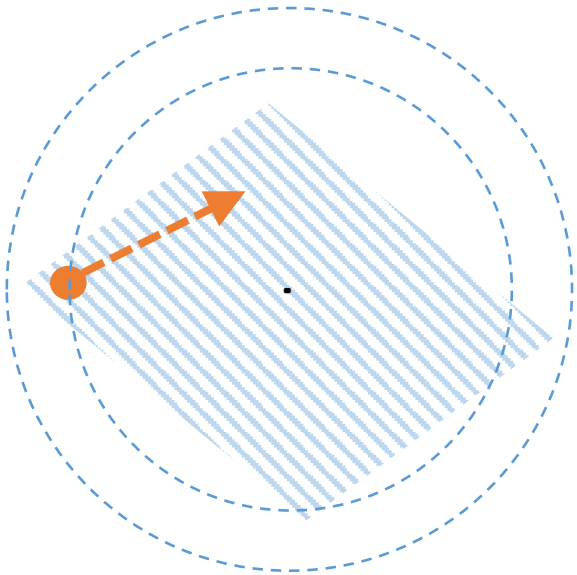
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



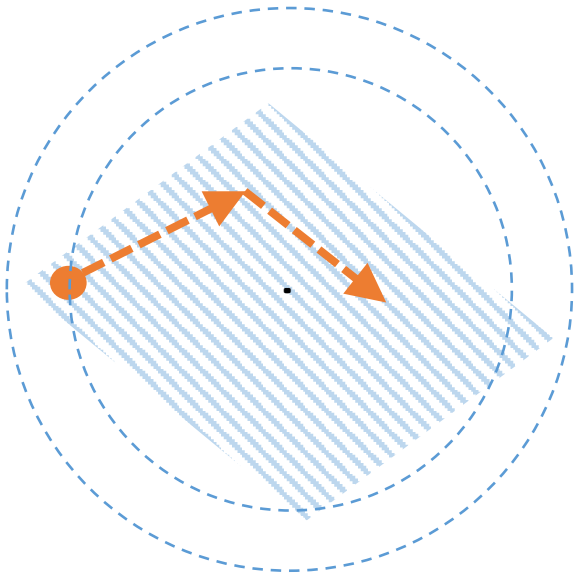
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



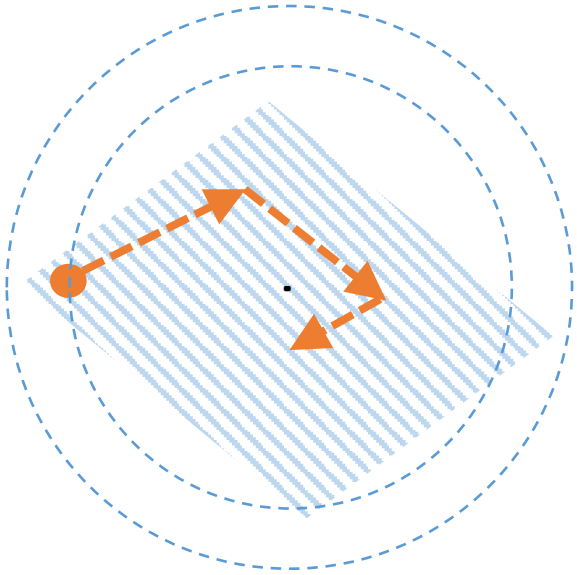
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



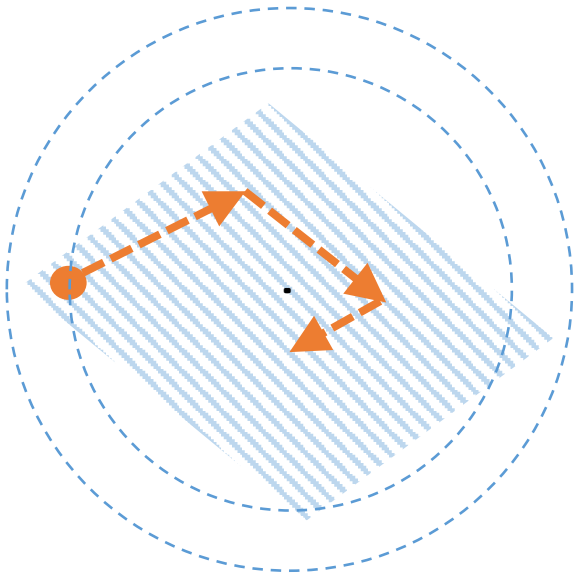
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



Our findings: GD is implicitly regularized

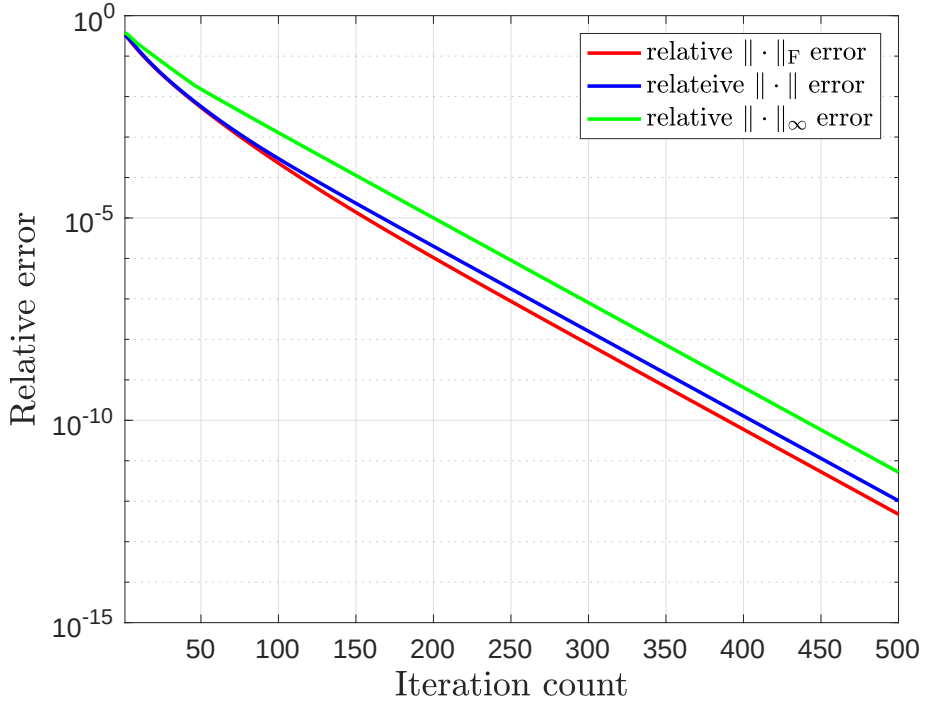
region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent even without regularization

Matrix completion via vanilla GD

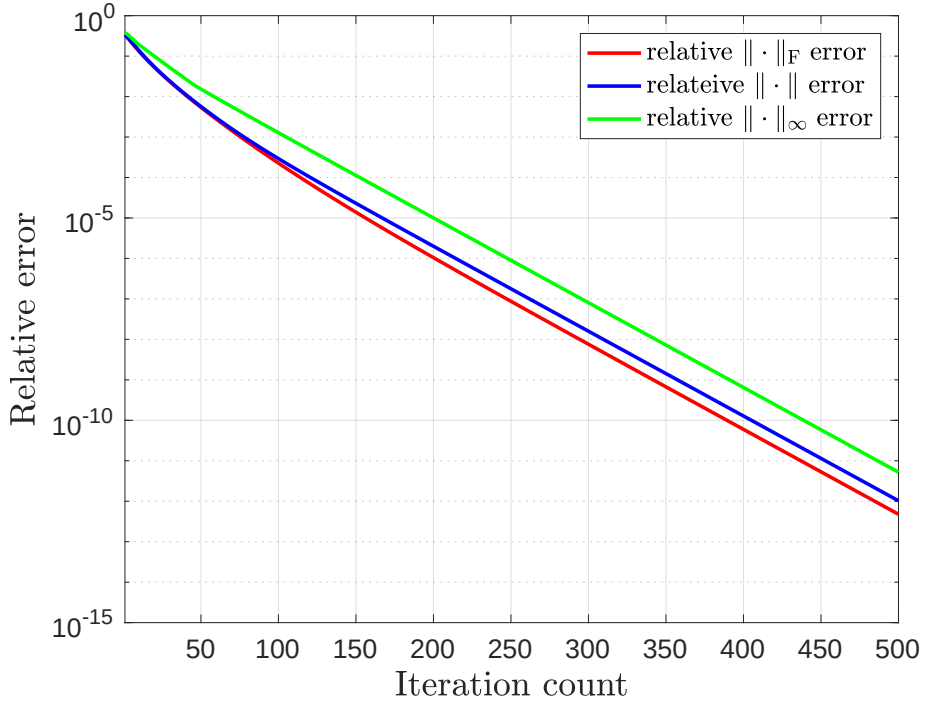
$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - M_{j,k}\right)^2$$



Vanilla GD converges fast without regularization!

Matrix completion via vanilla GD

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



Our finding: GD is implicitly regularized to stay incoherent!

Theoretical guarantees - noise-free case

Theorem (Ma, Wang, Chi, Chen)

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

- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}} \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|_{\mathrm{F}}$,
- $\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\| \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|$, *(spectral)*
- $\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\|_{2,\infty} \lesssim \rho^t \mu r \sqrt{\frac{\log n}{np}} \|\boldsymbol{X}^\natural\|_{2,\infty}$, *(incoherence)*

where $\rho = 1 - \frac{\sigma_{\min}\eta}{5} < 1$, *if step size* $\eta \asymp 1/\sigma_{\max}$ *and sample complexity* $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$.

Theoretical guarantees - noise-free case

Theorem (Ma, Wang, Chi, Chen)

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

- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}} \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|_{\mathrm{F}}$,
- $\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\| \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|$, *(spectral)*
- $\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\|_{2,\infty} \lesssim \rho^t \mu r \sqrt{\frac{\log n}{np}} \|\boldsymbol{X}^\natural\|_{2,\infty}$, *(incoherence)*

where $\rho = 1 - \frac{\sigma_{\min}\eta}{5} < 1$, *if step size* $\eta \asymp 1/\sigma_{\max}$ *and sample complexity* $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$.

- A recent follow-up by Xiaodong Li improves the sample complexity to $O(\mu^2 n r^2 \log n)$.

Noisy matrix completion via vanilla GD

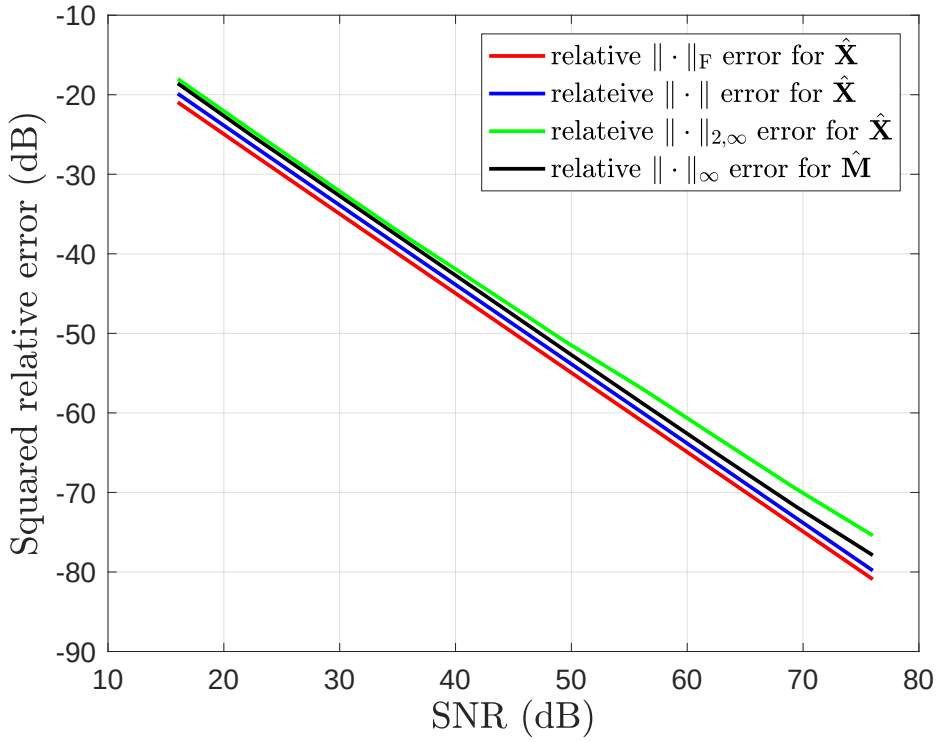


Figure: The relative error of $\hat{\boldsymbol{X}}$ in $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_{2,\infty}$ vs. SNR $:= \frac{\|\boldsymbol{M}^{\natural}\|_{\mathrm{F}}^2}{n^2\sigma^2}$.

Near-optimal entry-wise error control:

$$\left\|\boldsymbol{X}^t\boldsymbol{X}^{t\top} - \boldsymbol{M}^{\natural}\right\|_{\infty} \lesssim \left(\rho^t \mu r\sqrt{\frac{\log n}{np}} + \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{M}^{\natural}\right\|_{\infty}$$

Key ingredient: leave-one-out analysis

How to establish $\left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^\natural)\right\|_2 \ll \|\boldsymbol{X}^\natural\|_{2,\infty}$?

Key ingredient: leave-one-out analysis

How to establish $\left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^\natural)\right\|_2 \ll \|\boldsymbol{X}^\natural\|_{2,\infty}$?

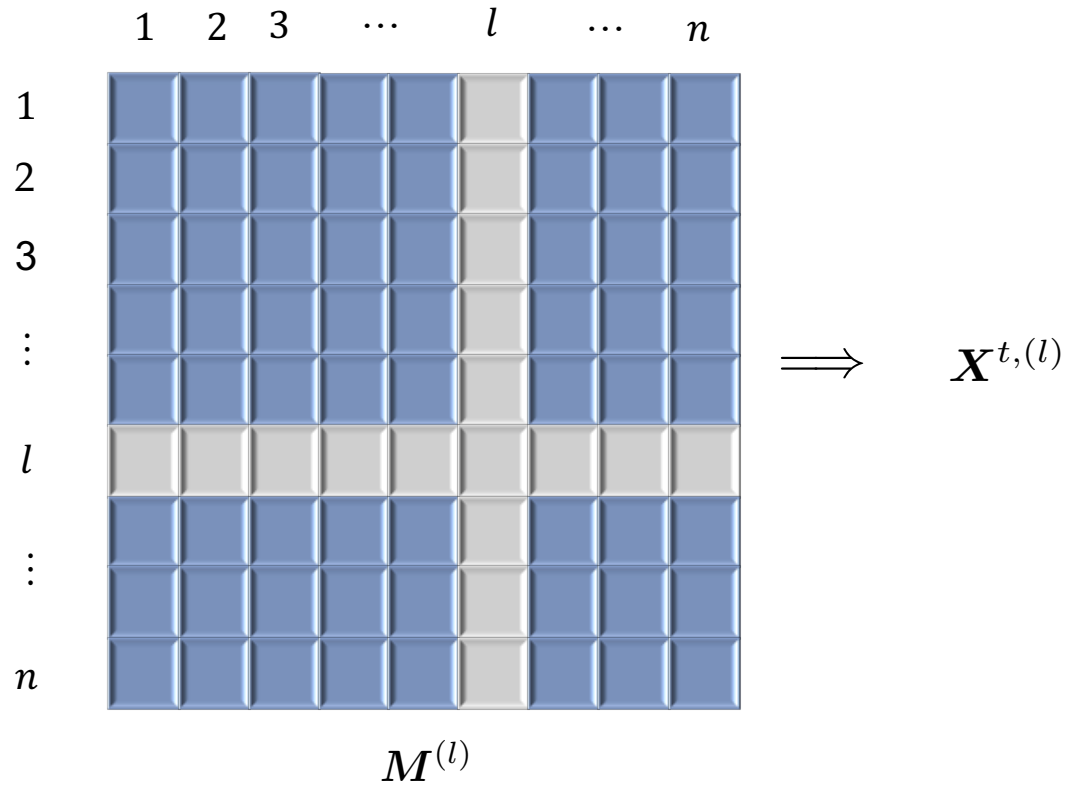
Technical difficulty: $\boldsymbol{X}^t$ is statistically dependent with Ω;

Key ingredient: leave-one-out analysis

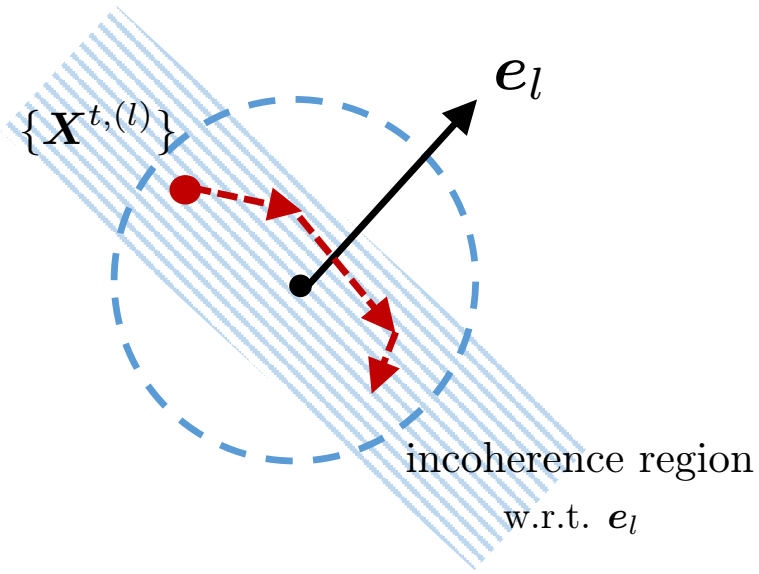
How to establish $\left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^\natural)\right\|_2 \ll \|\boldsymbol{X}^\natural\|_{2,\infty}$?

Technical difficulty: $\boldsymbol{X}^t$ is statistically dependent with Ω;

Leave-one-out trick: For each $1 \le l \le n$, introduce leave-one-out iterates $\boldsymbol{X}^{t,(l)}$ by replacing lth row and column with true values

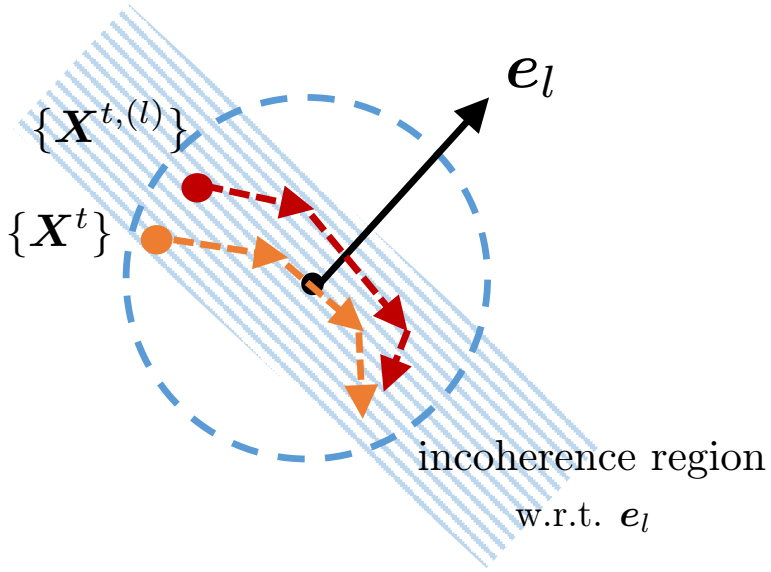


Key ingredient: leave-one-out analysis



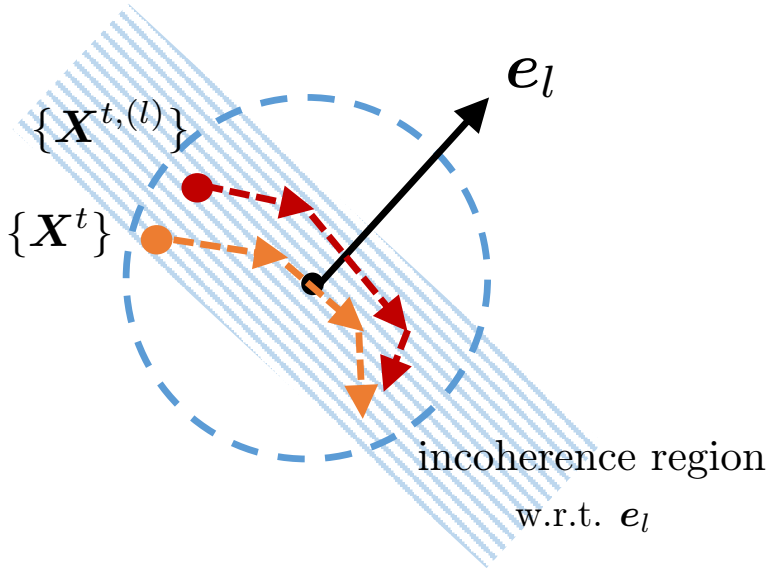
- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ contains more information of lth row of $\boldsymbol{X}^\natural$; indep. of randomness in lth row

Key ingredient: leave-one-out analysis



- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ contains more information of lth row of $\boldsymbol{X}^\natural$; indep. of randomness in lth row
- Leave-one-out iterates $\boldsymbol{X}^{t,(l)} \approx$ true iterates $\boldsymbol{X}^t$

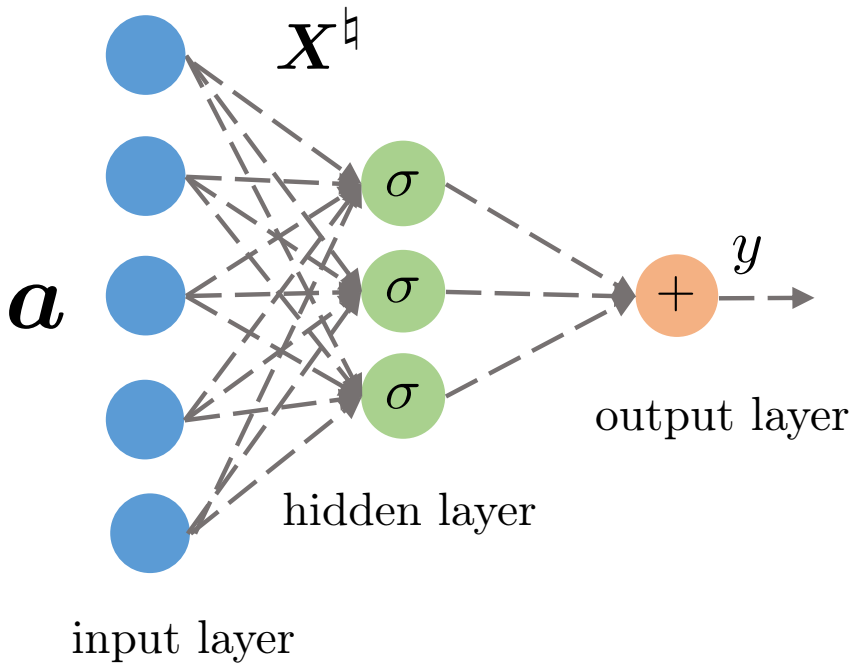
Key ingredient: leave-one-out analysis



- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ contains more information of lth row of $\boldsymbol{X}^\natural$; indep. of randomness in lth row
- Leave-one-out iterates $\boldsymbol{X}^{t,(l)} \approx$ true iterates $\boldsymbol{X}^t$
- $\left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^\natural)\right\|_2 \le \left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^{t,(l)} - \boldsymbol{X}^\natural)\right\|_2 + \left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^{t,(l)})\right\|_2$

The phenomenon is quite general

Shallow neural network with quadratic activation

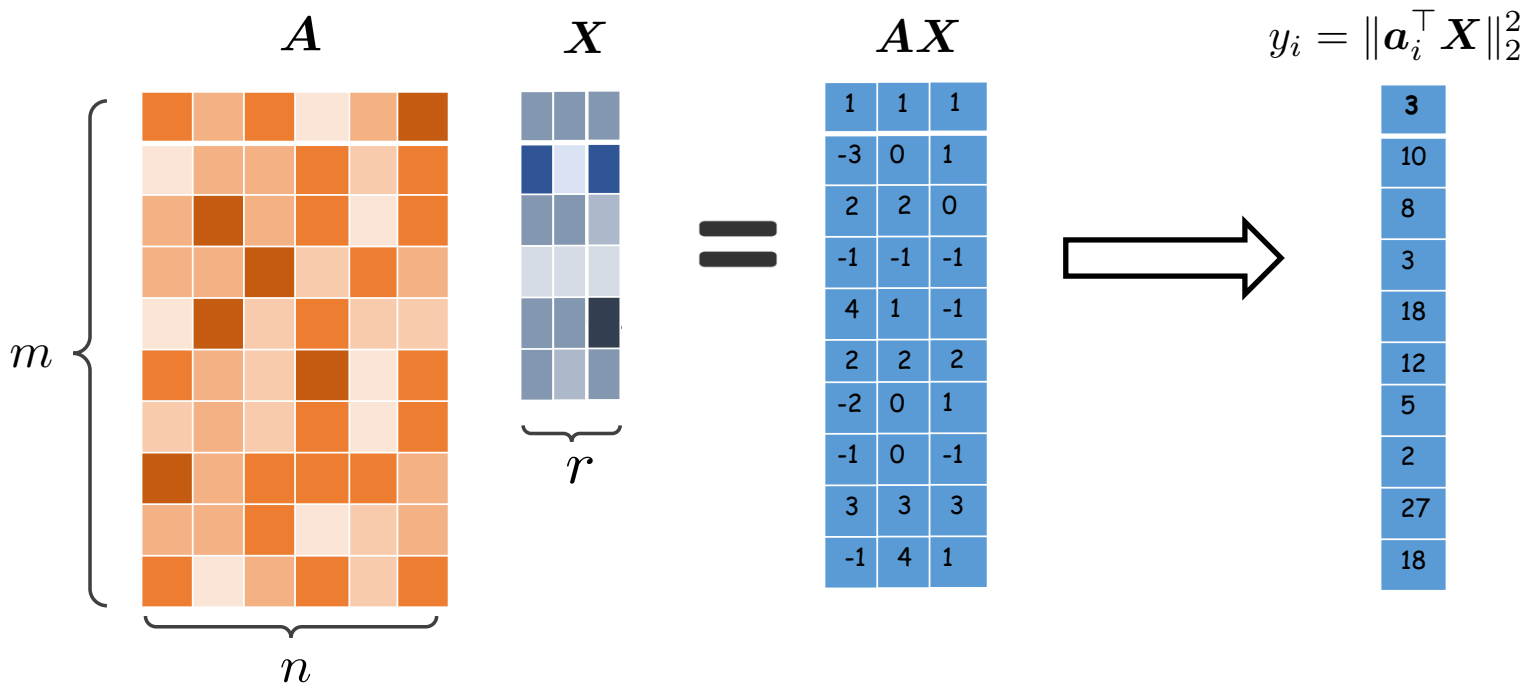


Set $\boldsymbol{X}^\natural = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r]$, then

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

Identifiability up to orthonormal transform of $\boldsymbol{X}^\natural$.

Generalized phase retrieval



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

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

Applications: optical imaging, phase space tomography ...

Implicit regularization for generalized phase retrieval

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

Implicit regularization for generalized phase retrieval

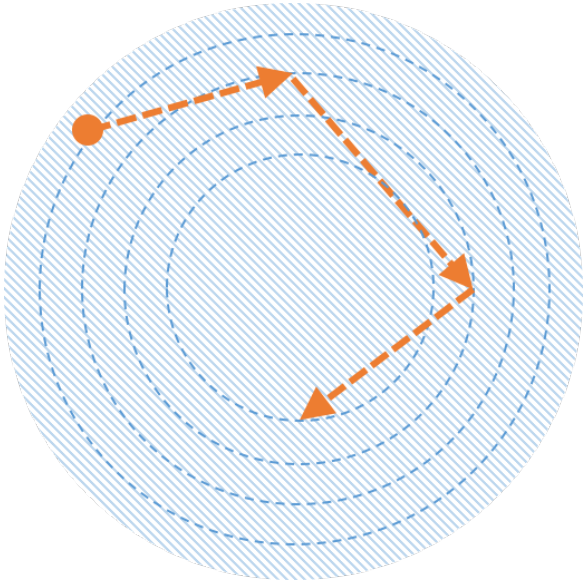
$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{1}{4m}\sum_{k=1}^{m}\left(\left\|\boldsymbol{a}_k^\top \boldsymbol{X}\right\|^2 - y_k\right)^2$$

- region of local strong convexity + smoothness

Implicit regularization for generalized phase retrieval

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

region of local strong convexity + smoothness

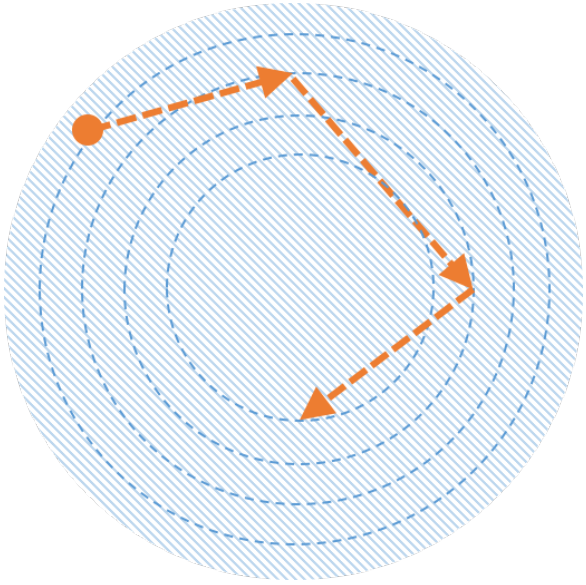


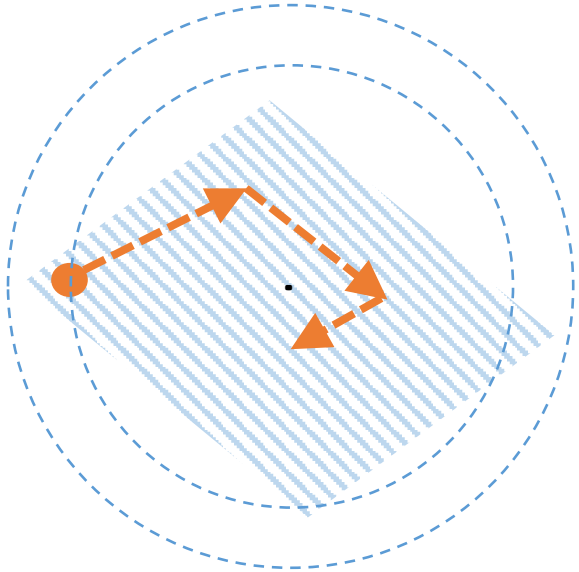
$$O(1) \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)$$

Implicit regularization for generalized phase retrieval

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

region of local strong convexity + smoothness





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

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

Theoretical guarantees

Theorem (Li, Ma, Chen, Chi)

Under i.i.d. Gaussian design, GD achieves linear convergence

- $\max_k \left\| \boldsymbol{a}_k^\top (\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural) \right\| \lesssim \sqrt{\log n}\, \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathrm{F}}}$ *(incoherence)*

Theoretical guarantees

Theorem (Li, Ma, Chen, Chi)

Under i.i.d. Gaussian design, GD achieves linear convergence

- $\max_k \left\|\boldsymbol{a}_k^\top (\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural)\right\| \lesssim \sqrt{\log n}\, \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathrm{F}}}$ *(incoherence)*
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}} \lesssim \left(1 - \frac{\sigma_r^2(\boldsymbol{X}^\natural)\eta}{2}\right)^t \|\boldsymbol{X}^\natural\|_{\mathrm{F}}$ *(linear convergence)*

provided that $\eta \asymp \frac{1}{(\log n \vee r)^2 \sigma_r^2(\boldsymbol{X}^\natural)}$ *and* $m \gtrsim nr^4 \log n$.

Theoretical guarantees

Theorem (Li, Ma, Chen, Chi)

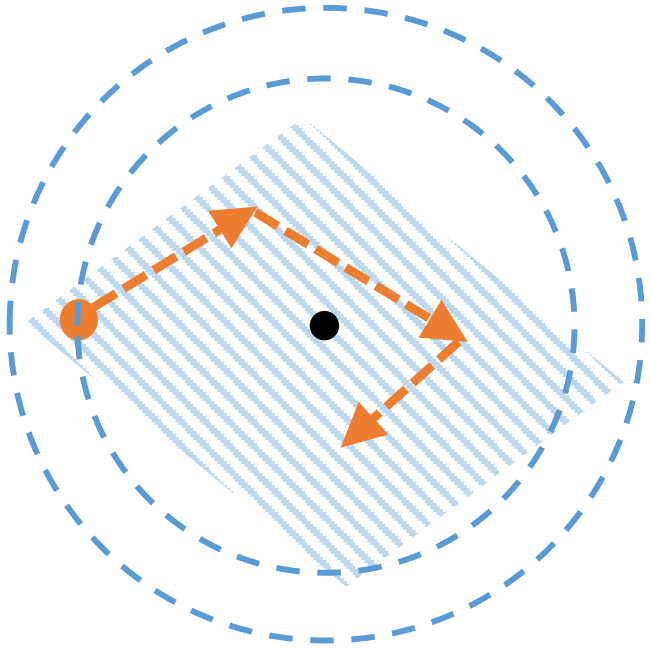
Under i.i.d. Gaussian design, GD achieves linear convergence

- $\max_k \left\|\boldsymbol{a}_k^\top (\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural)\right\| \lesssim \sqrt{\log n}\, \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathrm{F}}}$ *(incoherence)*
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}} \lesssim \left(1 - \frac{\sigma_r^2(\boldsymbol{X}^\natural)\eta}{2}\right)^t \|\boldsymbol{X}^\natural\|_{\mathrm{F}}$ *(linear convergence)*

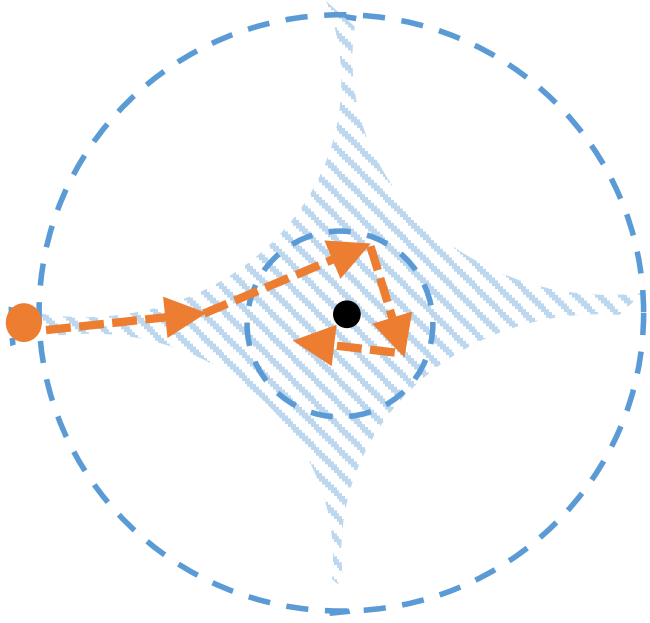
provided that $\eta \asymp \frac{1}{(\log n \vee r)^2 \sigma_r^2(\boldsymbol{X}^\natural)}$ *and* $m \gtrsim nr^4 \log n$.

Big computational saving: GD attains ε-accuracy within $O\big((\log n \vee r)^2 \log \frac{1}{\varepsilon}\big)$ iterations if $m \asymp nr^4 \log n$.

Incoherence region in high dimensions



2-dimensional



high-dimensional

incoherence region is vanishingly small

Conclusions

From population loss from empirical loss: vanilla gradient descent exploits local hidden convexity as if it almost runs on the population loss!

Computational:
near dimension-free
iteration complexity

Statistical:
near-optimal
sample complexity

Analytical toolkits: leave-one-out perturbation argument to establish near-independence. Useful in other problems!

References

1. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview, Y. Chi, Y. M. Lu and Y. Chen, arXiv:1809.09573.
2. Implicit Regularization for Nonconvex Statistical Estimation, C. Ma, K. Wang, Y. Chi and Y. Chen, arXiv:1711.10467.
3. Nonconvex Matrix Factorization from Rank-One Measurements, Y. Li, C. Ma, Y. Chen, and Y. Chi, arXiv:1802.06286.
4. Gradient Descent with *Random Initialization*: Fast Global Convergence for Nonconvex Phase Retrieval, Y. Chen, Y. Chi, J. Fan and C. Ma, arXiv:1803.07726.
5. Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation, Y. Chen and Y. Chi, survey article on IEEE Signal Processing Magazine, Jul. 2018.

`https://users.ece.cmu.edu/~yuejiec/`

Thank you!