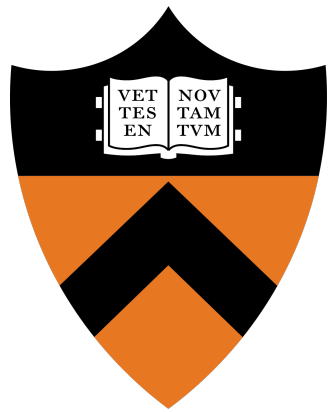
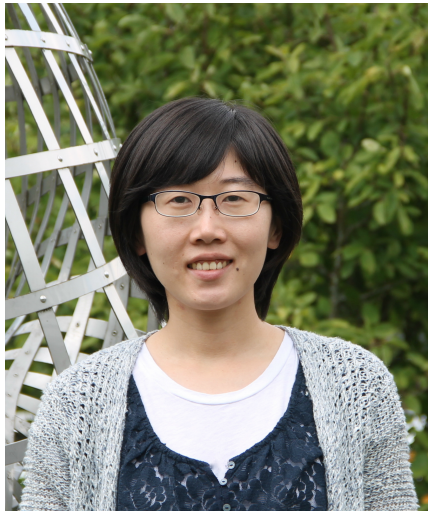
# Nonconvex Matrix Completion without Regularization



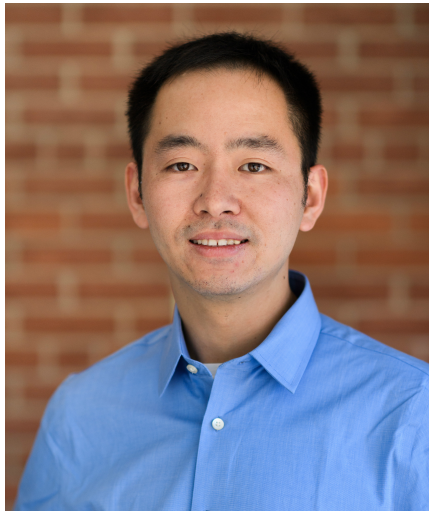

Cong Ma

ORFE, Princeton University


Kaizheng Wang
Princeton ORFE



Yuejie Chi
CMU ECE

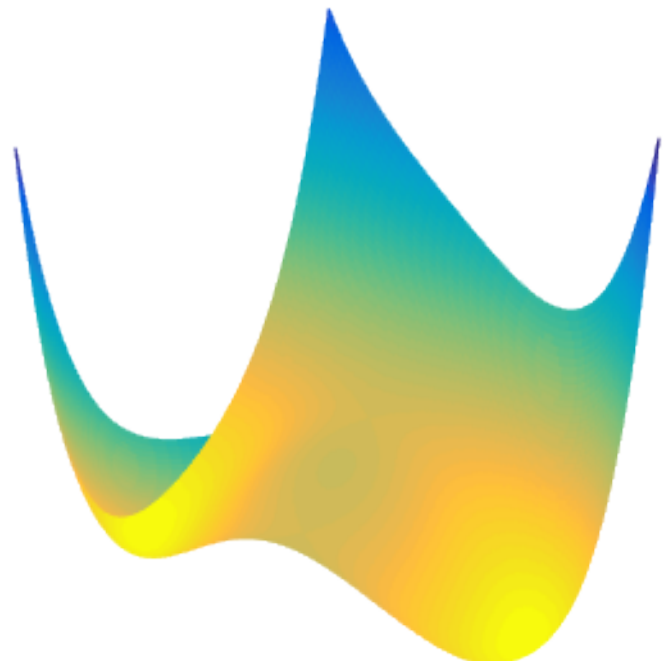


Yuxin Chen
Princeton EE

# Nonconvex problems are everywhere

Empirical risk minimization is usually nonconvex

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

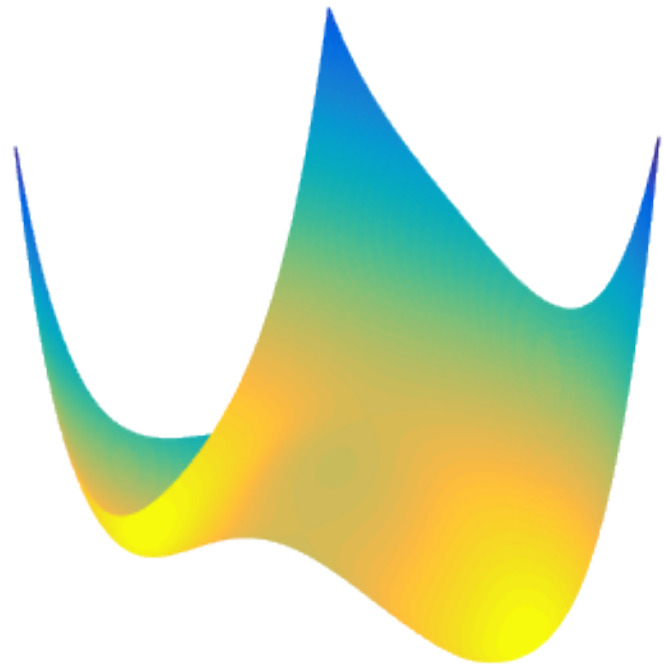


# Nonconvex problems are everywhere

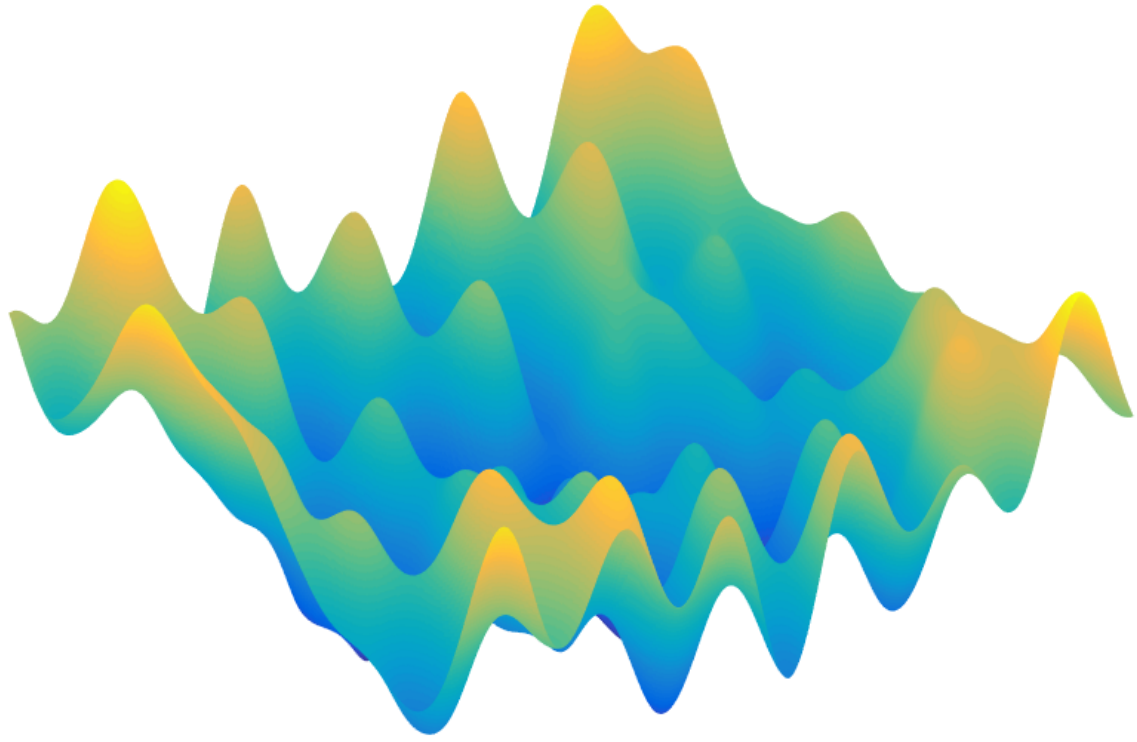
Empirical risk minimization is usually nonconvex

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

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



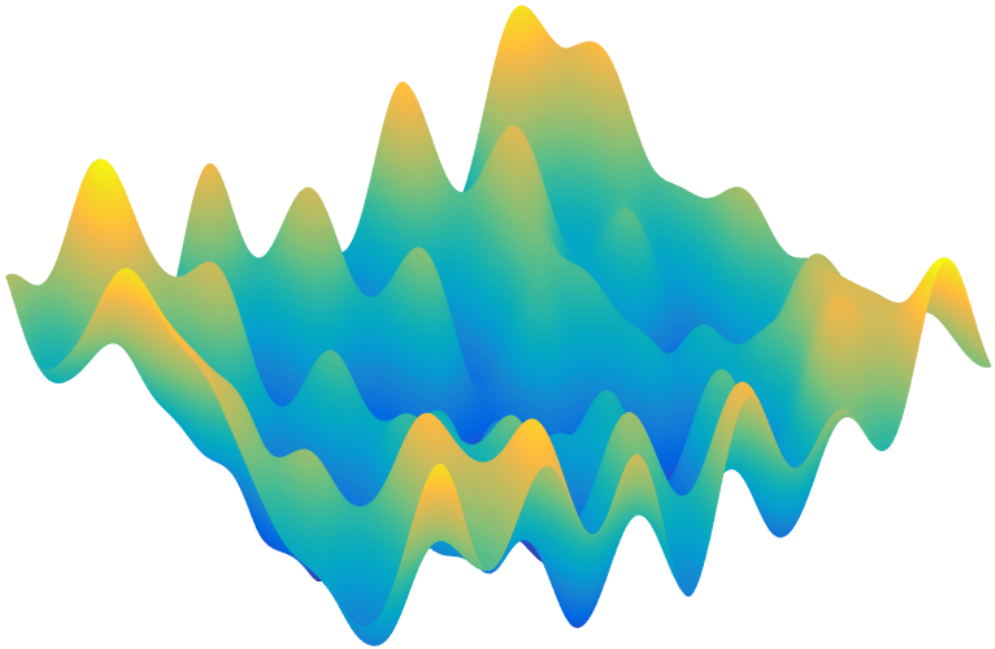
# Nonconvex optimization may be super scary

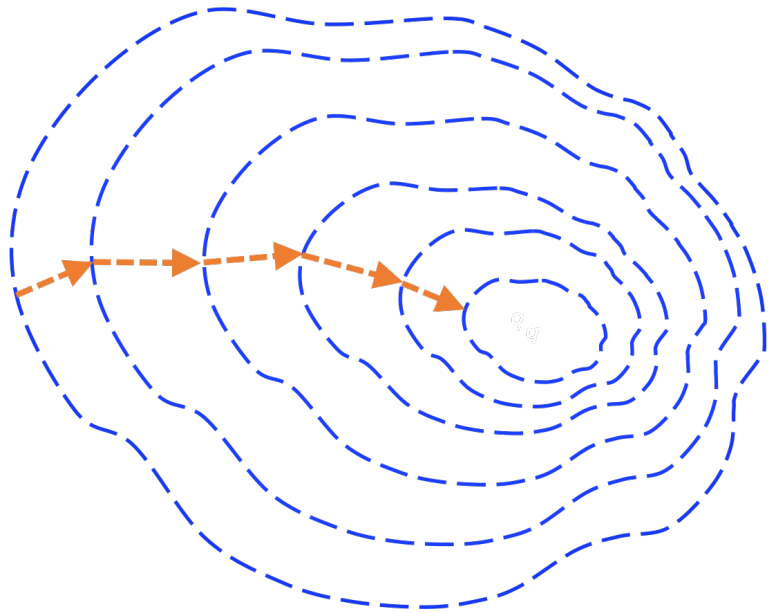


There may be bumps everywhere and exponentially many local optima

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

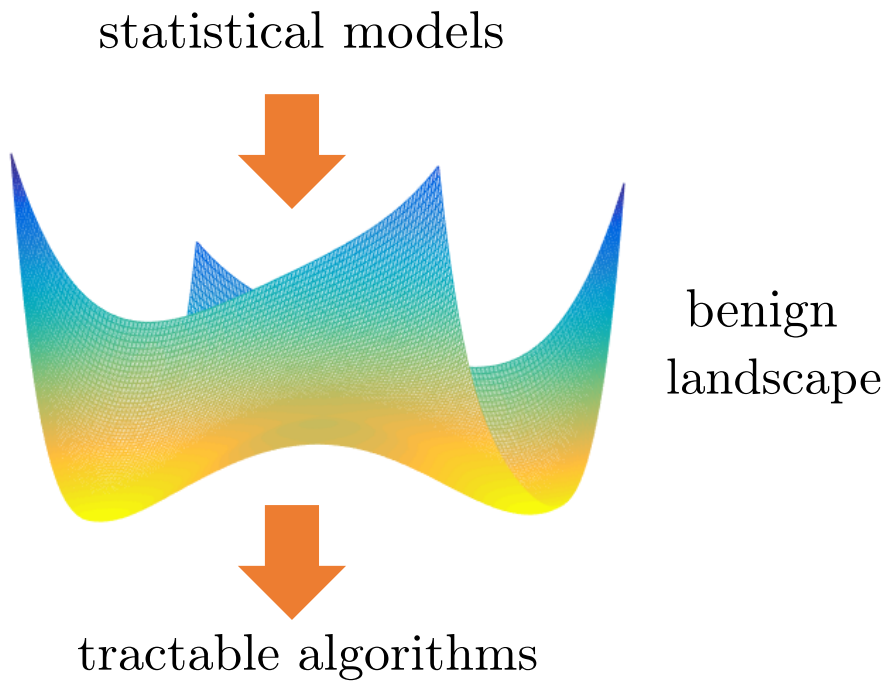
# Nonconvex optimization may be super scary





But they are solved on a daily basis via simple algorithms like *(stochastic) gradient descent*

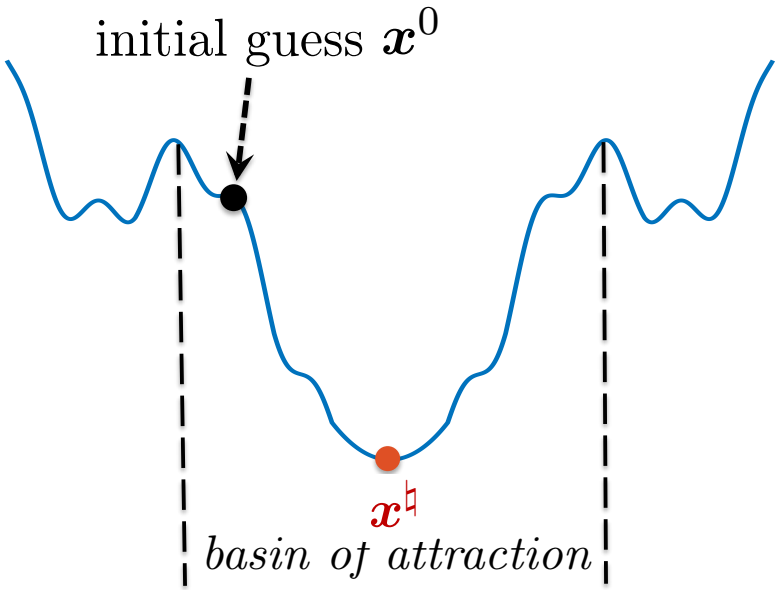
# Statistical models come to rescue




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

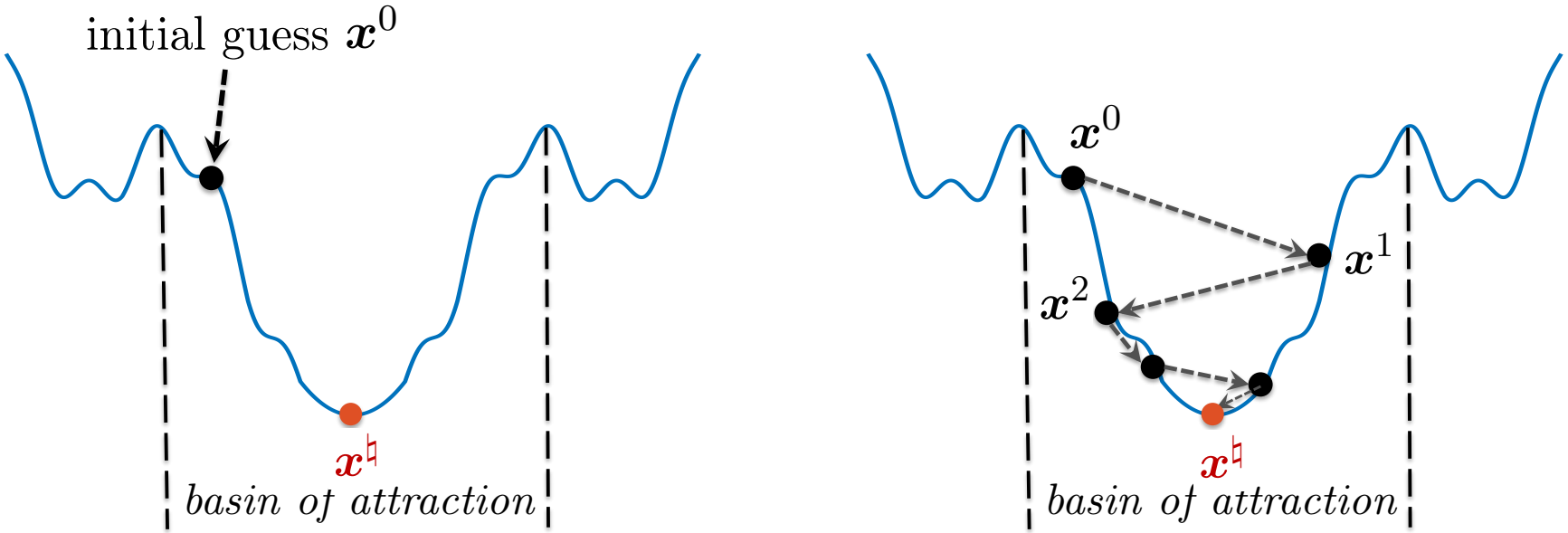
— *Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview*
*Chi, Lu, Chen '18*

# A popular two-stage approach




1. initialize within $\underbrace{\text{local basin sufficiently close to } x^\natural}_{\text{(restricted) strongly convex and smooth}}$

# A popular two-stage approach




1. initialize within $\underbrace{\text{local basin sufficiently close to } \boldsymbol{x}^\natural}_{\text{(restricted) strongly convex and smooth}}$

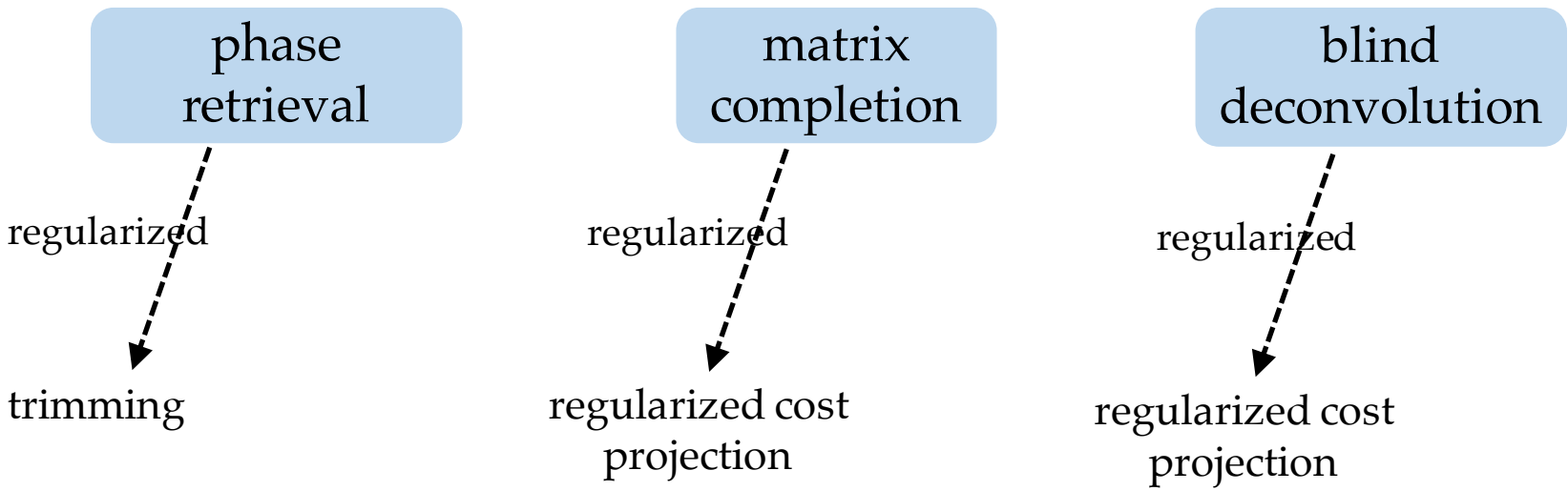
2. iterative refinement

# Proper regularization is *often* recommended

Improves computation by stabilizing search directions

# Proper regularization is *often* recommended

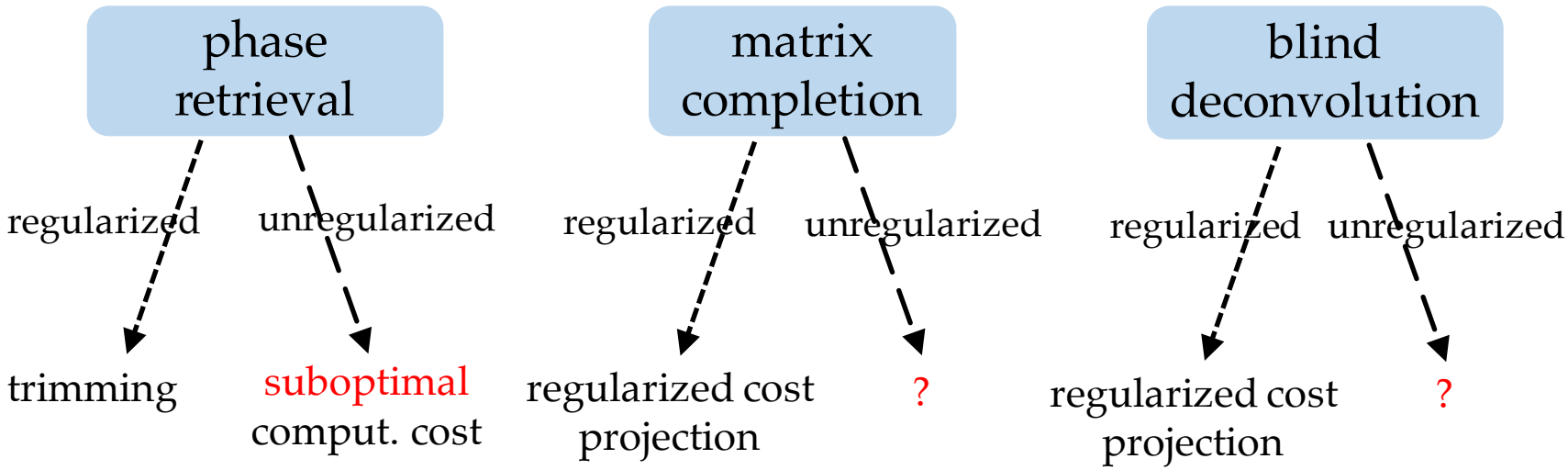
Improves computation by stabilizing search directions




- trimming: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \mathcal{T}\left(\nabla f\left(\boldsymbol{x}^t\right)\right)$
- regularized cost: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \left(\nabla f\left(\boldsymbol{x}^t\right) + \nabla R\left(\boldsymbol{x}^t\right)\right)$
- projection: $\boldsymbol{x}^{t+1} = \mathcal{P}\left(\boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right)\right)$

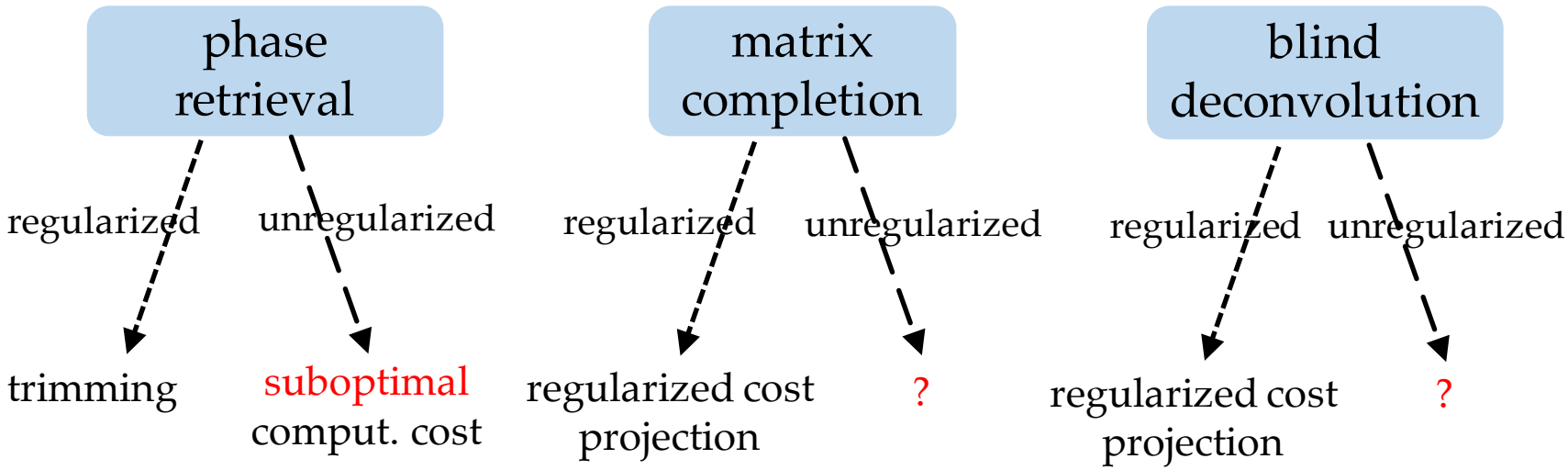
# How about unregularized gradient methods?

Improves computation by stabilizing search directions



# How about unregularized gradient methods?

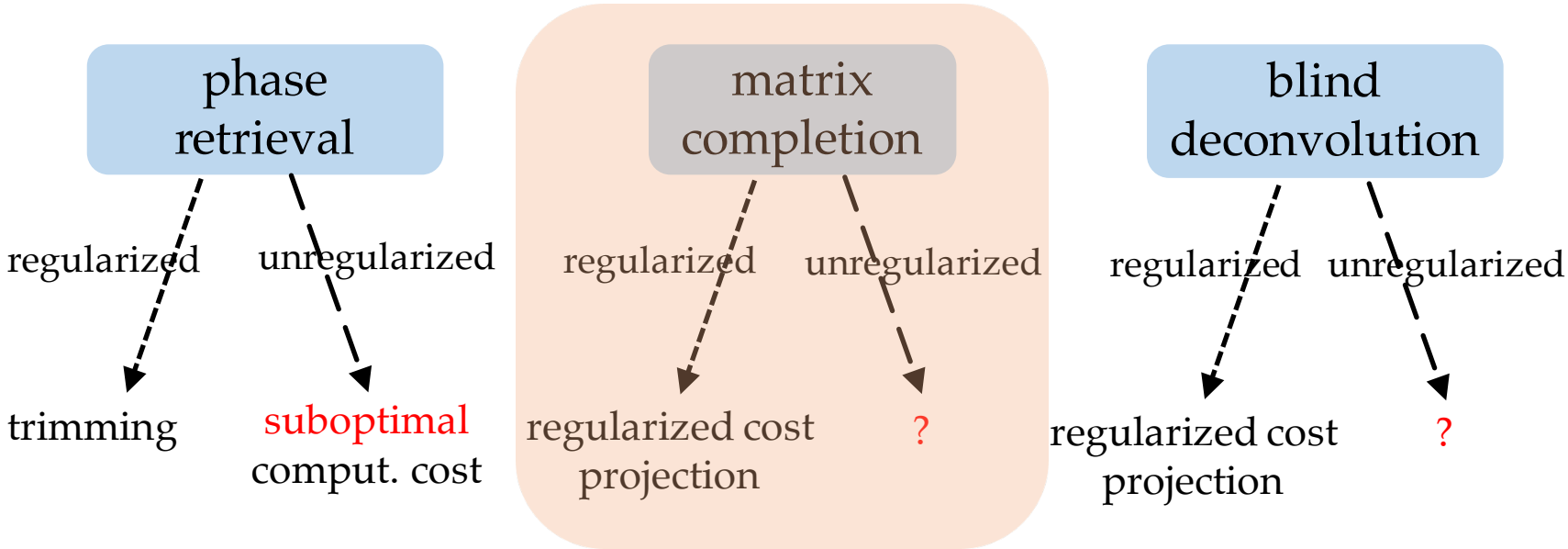
Improves computation by stabilizing search directions




*Are unregularized methods suboptimal for nonconvex estimation?*

# How about unregularized gradient methods?

Improves computation by stabilizing search directions




*Are unregularized methods suboptimal for nonconvex estimation?*

# Low-rank matrix completion

*figure credit: Emmanuel Candès*

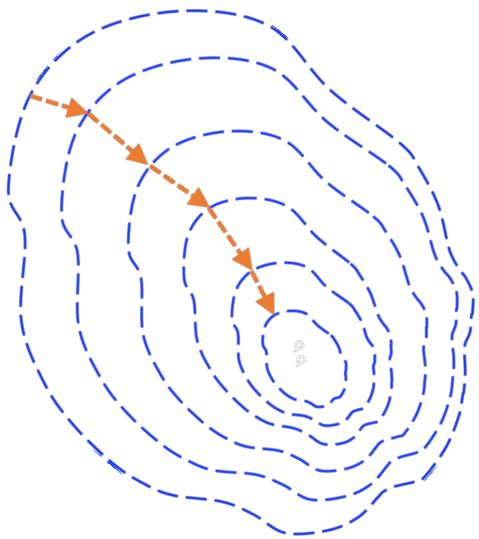
Given partial samples $\Omega$ of a *low-rank* matrix $M^\natural$, fill in missing entries

# A natural least squares formulation

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

# A natural least squares formulation

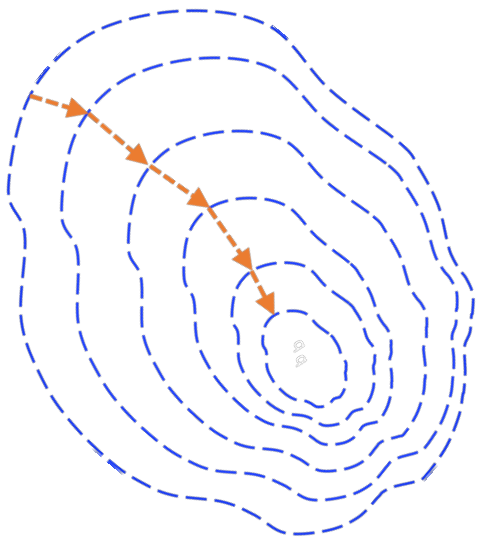
$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - M_{j,k}^\natural\right)^2$$



- **spectral initialization:** $\boldsymbol{X}^0 \leftarrow$ leading eigenvectors of data matrix

# A natural least squares formulation

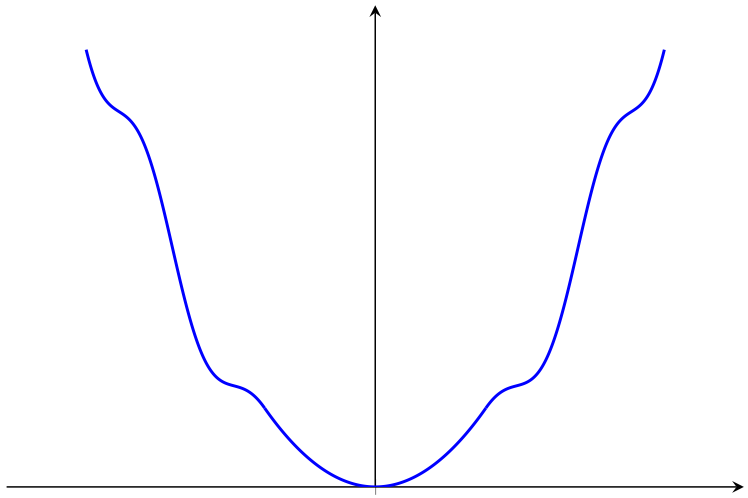
$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - M_{j,k}^\natural\right)^2$$



- **spectral initialization:** $\boldsymbol{X}^0 \leftarrow$ leading eigenvectors of data matrix
- **gradient descent:**

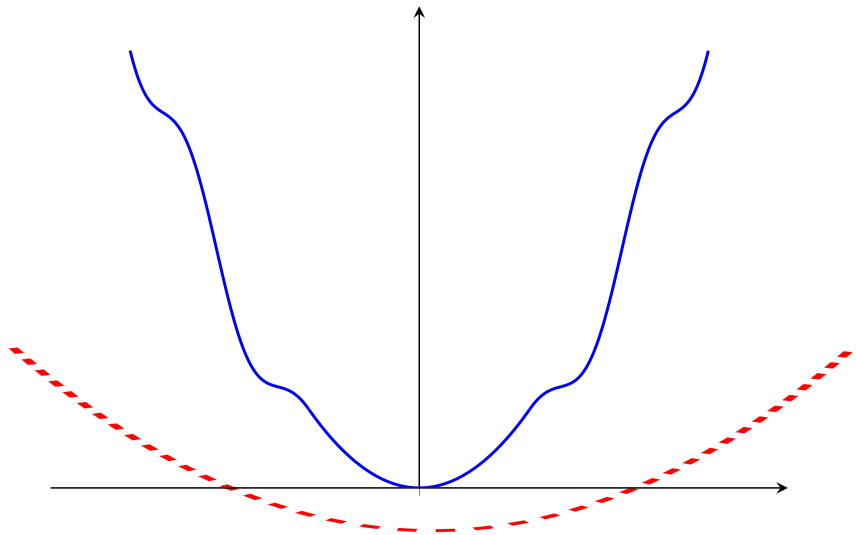
$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla f(\boldsymbol{X}^t), \quad t = 0, 1, \cdots$$

# Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

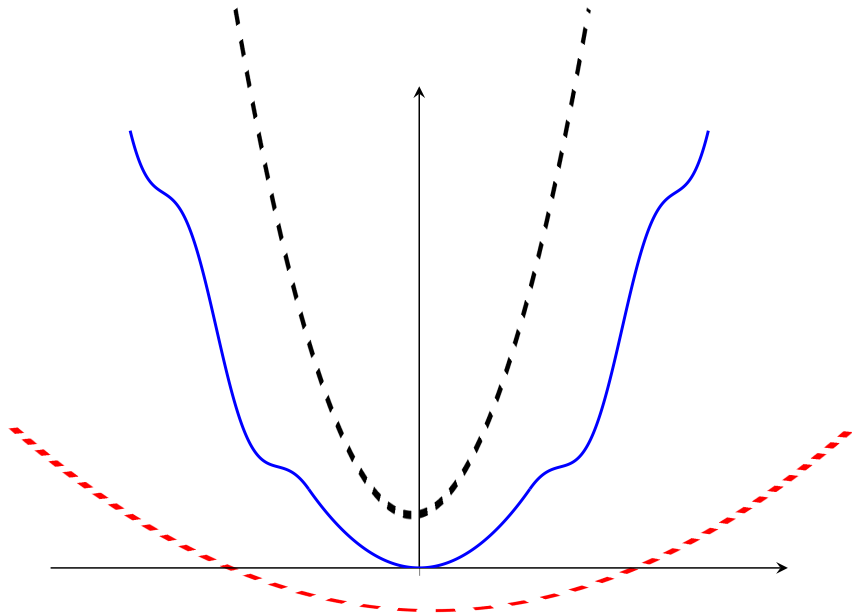
# Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity

# Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

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

# Gradient descent theory revisited

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

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

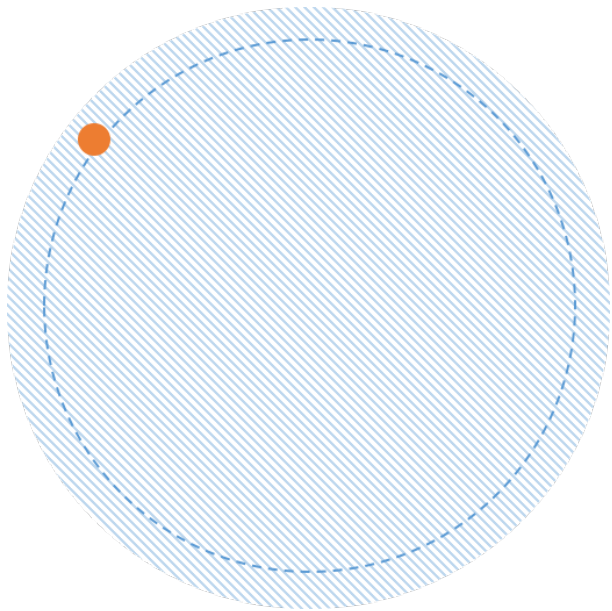
$\ell_2$ **error contraction:** GD with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\natural\|_{\mathrm{F}} \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{X}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}}$$

# Gradient descent theory revisited

$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\natural\|_{\mathrm{F}} \leq (1 - \alpha/\beta)\,\|\boldsymbol{X}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}}$$

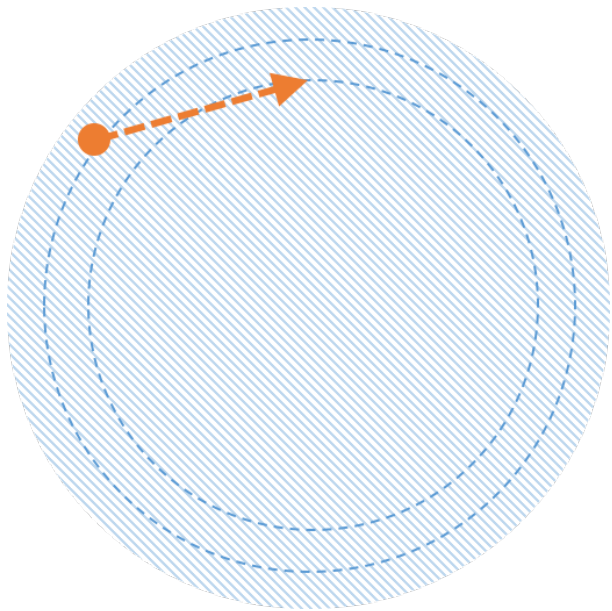
region of local strong convexity + smoothness



# Gradient descent theory revisited

$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\natural\|_{\mathrm{F}} \leq (1 - \alpha/\beta)\,\|\boldsymbol{X}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}}$$

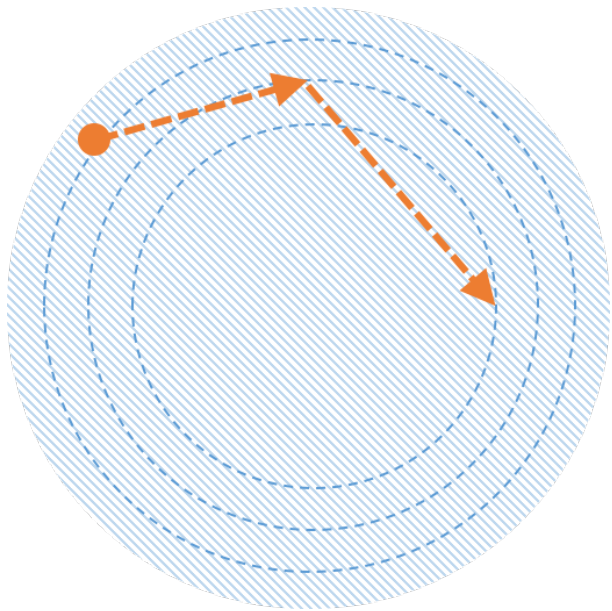
region of local strong convexity + smoothness



# Gradient descent theory revisited

$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\natural\|_{\mathrm{F}} \leq (1 - \alpha/\beta)\,\|\boldsymbol{X}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}}$$

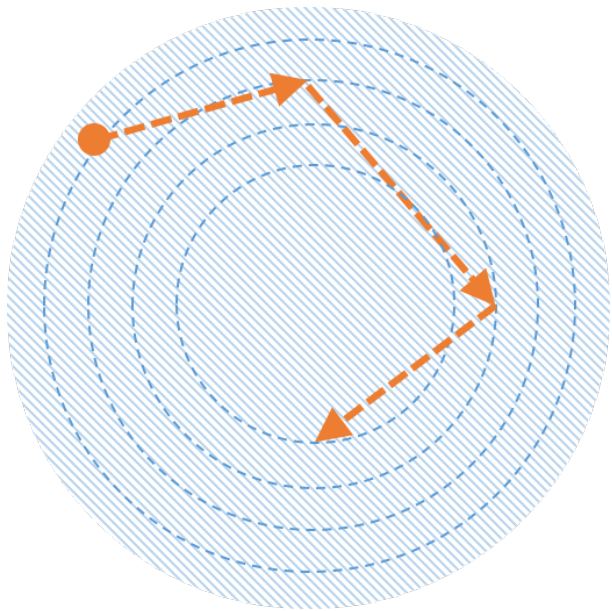
region of local strong convexity + smoothness



# Gradient descent theory revisited

$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\natural\|_{\mathrm{F}} \leq (1 - \alpha/\beta)\,\|\boldsymbol{X}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}}$$

region of local strong convexity + smoothness



# What does this optimization theory say about GD?

Independent Bernoulli sampling: $\boldsymbol{M}^{\natural}_{j,k}$ is observed with probability $p$

**Population level ($p = 1$)**

$f(\boldsymbol{X})$ is locally restricted strongly convex and smooth

$$\|\boldsymbol{V}\|_{\mathrm{F}}^2 \lesssim \mathrm{vec}\left(\boldsymbol{V}\right)^\top \mathbb{E}\left[\nabla^2 f\left(\boldsymbol{X}\right)\right] \mathrm{vec}\left(\boldsymbol{V}\right) \lesssim \|\boldsymbol{V}\|_{\mathrm{F}}^2$$

for all $\boldsymbol{X}$ and $\boldsymbol{V}$ such that

# What does this optimization theory say about GD?

Independent Bernoulli sampling: $\boldsymbol{M}^{\natural}_{j,k}$ is observed with probability $p$

**Population level ($p = 1$)**

$f(\boldsymbol{X})$ is locally restricted strongly convex and smooth

$$\|\boldsymbol{V}\|_{\mathrm{F}}^2 \lesssim \mathrm{vec}\left(\boldsymbol{V}\right)^\top \mathbb{E}\left[\nabla^2 f\left(\boldsymbol{X}\right)\right] \mathrm{vec}\left(\boldsymbol{V}\right) \lesssim \|\boldsymbol{V}\|_{\mathrm{F}}^2$$

for all $\boldsymbol{X}$ and $\boldsymbol{V}$ such that

- $\boldsymbol{X}$ is not far away from $\boldsymbol{X}^{\natural}$ in $\ell_{\mathrm{F}}$ sense.
- $\boldsymbol{V}$ points towards $\boldsymbol{X}^{\natural}$

# What does this optimization theory say about GD?

Independent Bernoulli sampling: $\boldsymbol{M}_{j,k}^{\natural}$ is observed with probability $p$

**Finite-sample level** ($p \asymp \text{poly}\log n/n$)

$f(\boldsymbol{X})$ is locally restricted strongly convex and smooth

$$\|\boldsymbol{V}\|_{\mathrm{F}}^2 \lesssim \mathrm{vec}\left(\boldsymbol{V}\right)^{\top} \mathbb{E}\left[\nabla^2 f\left(\boldsymbol{X}\right)\right] \mathrm{vec}\left(\boldsymbol{V}\right) \lesssim \|\boldsymbol{V}\|_{\mathrm{F}}^2$$

for all $\boldsymbol{X}$ and $\boldsymbol{V}$ such that

- $\boldsymbol{X}$ is not far away from $\boldsymbol{X}^{\natural}$ in $\ell_{\mathrm{F}}$ sense.
- $\boldsymbol{V}$ points towards $\boldsymbol{X}^{\natural}$

# What does this optimization theory say about GD?

Independent Bernoulli sampling: $\boldsymbol{M}^{\natural}_{j,k}$ is observed with probability $p$

**Finite-sample level** ($p \asymp \text{poly}\log n/n$)

$f(\boldsymbol{X})$ is locally restricted strongly convex and smooth

$$\|\boldsymbol{V}\|_{\mathrm{F}}^2 \lesssim \mathrm{vec}\left(\boldsymbol{V}\right)^\top \mathbb{E}\left[\nabla^2 f\left(\boldsymbol{X}\right)\right] \mathrm{vec}\left(\boldsymbol{V}\right) \lesssim \|\boldsymbol{V}\|_{\mathrm{F}}^2$$

for all $\boldsymbol{X}$ and $\boldsymbol{V}$ such that

- $\boldsymbol{X}$ is not far away from $\boldsymbol{X}^{\natural}$ in $\ell_{\mathrm{F}}$ sense.
- $\boldsymbol{V}$ points towards $\boldsymbol{X}^{\natural}$
- $\boldsymbol{X}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\}$

# Incoherence

**Definition 1 (Incoherence for matrix completion)**

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

$$\max_i \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2 = \|\boldsymbol{U}\|_{2,\infty} \le \sqrt{\frac{\mu}{n}}\,\|\boldsymbol{U}\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}.$$

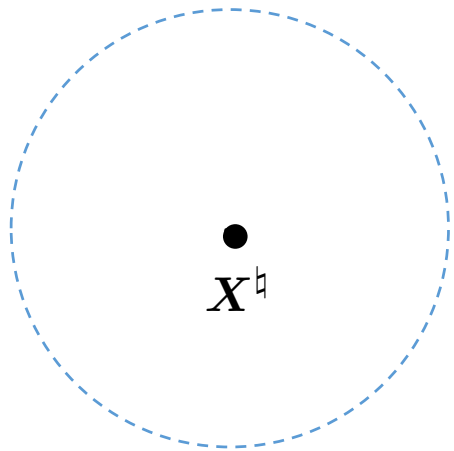
$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\text{hard } \mu=n} \quad \text{vs.} \quad \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots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

# Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

# Incoherence region

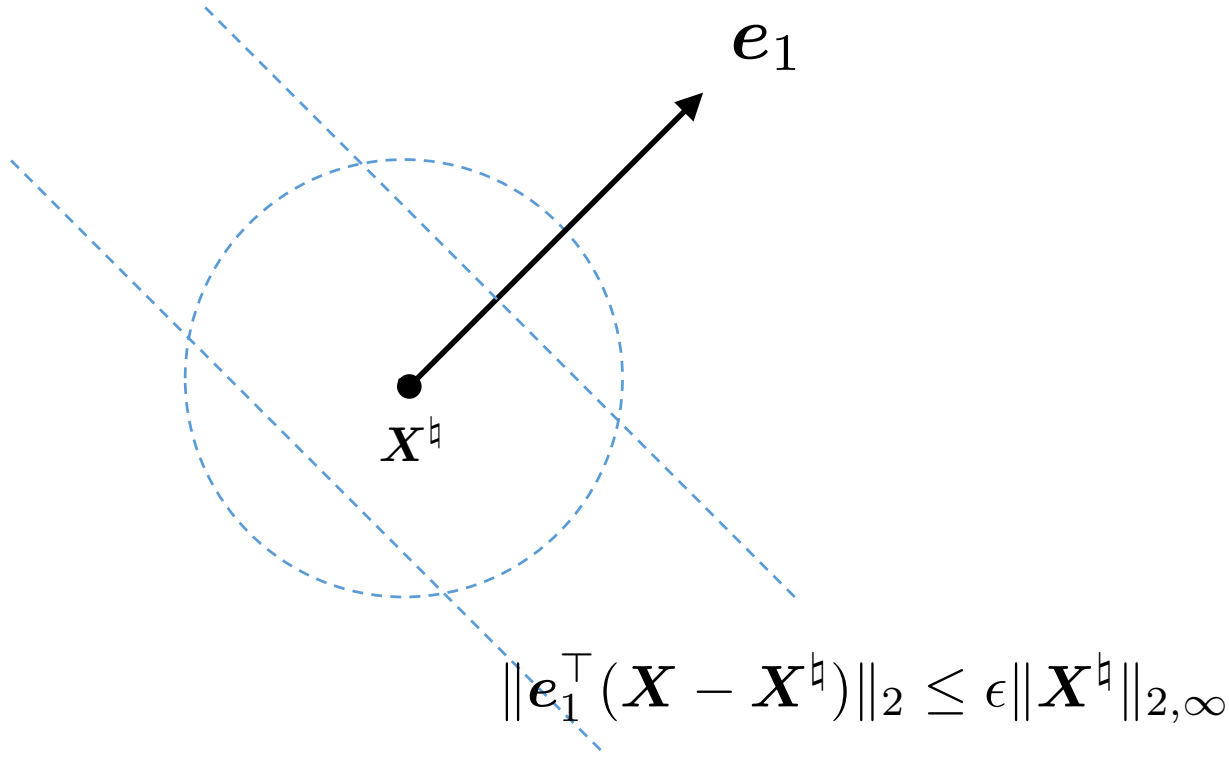
Which region enjoys both restricted strong convexity and smoothness?




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

# Incoherence region

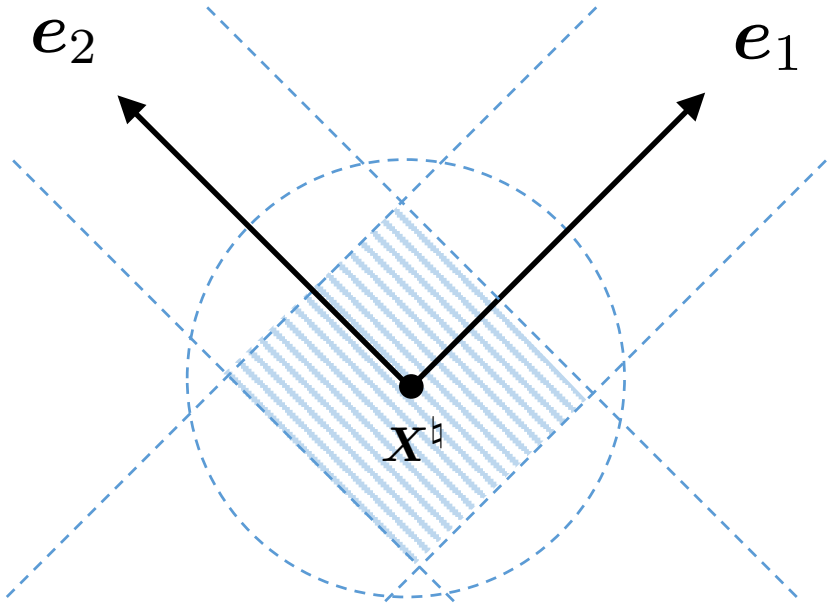
Which region enjoys both restricted strong convexity and smoothness?




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

# Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

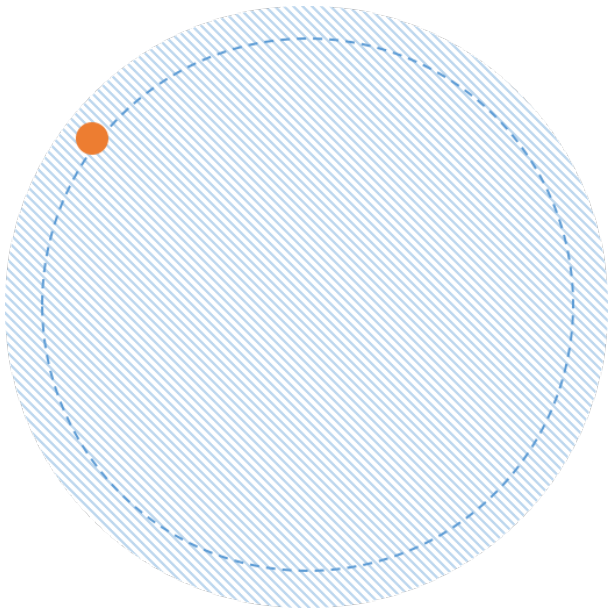



$$\|\boldsymbol{e}_2^\top(\boldsymbol{X}-\boldsymbol{X}^\natural)\|_2 \le \epsilon\|\boldsymbol{X}^\natural\|_{2,\infty} \qquad \|\boldsymbol{e}_1^\top(\boldsymbol{X}-\boldsymbol{X}^\natural)\|_2 \le \epsilon\|\boldsymbol{X}^\natural\|_{2,\infty}$$

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

# "Failure" of gradient descent?

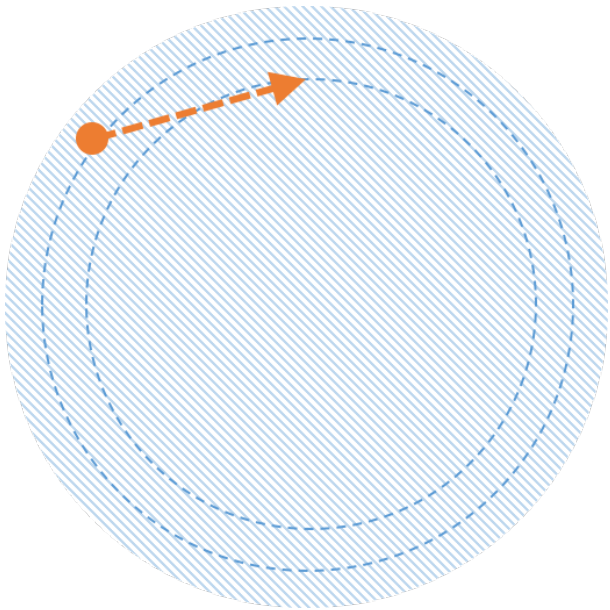
region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# "Failure" of gradient descent?

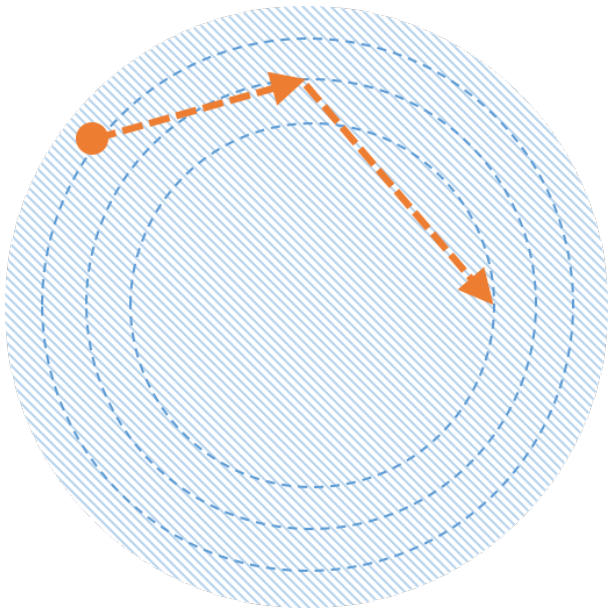
region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# "Failure" of gradient descent?

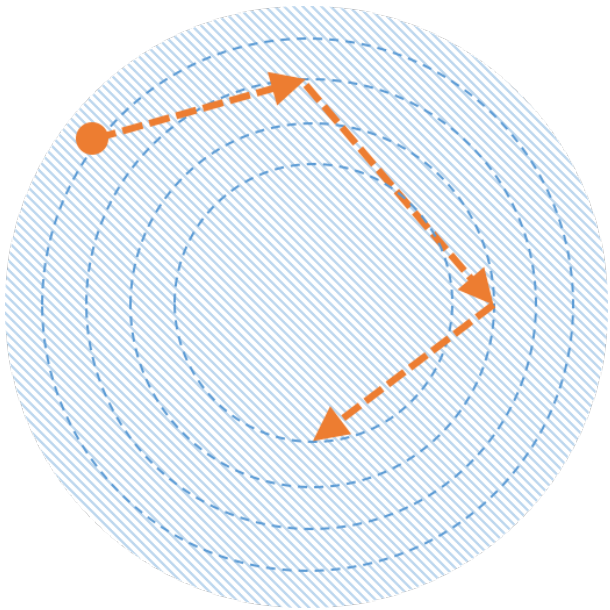
region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# "Failure" of gradient descent?

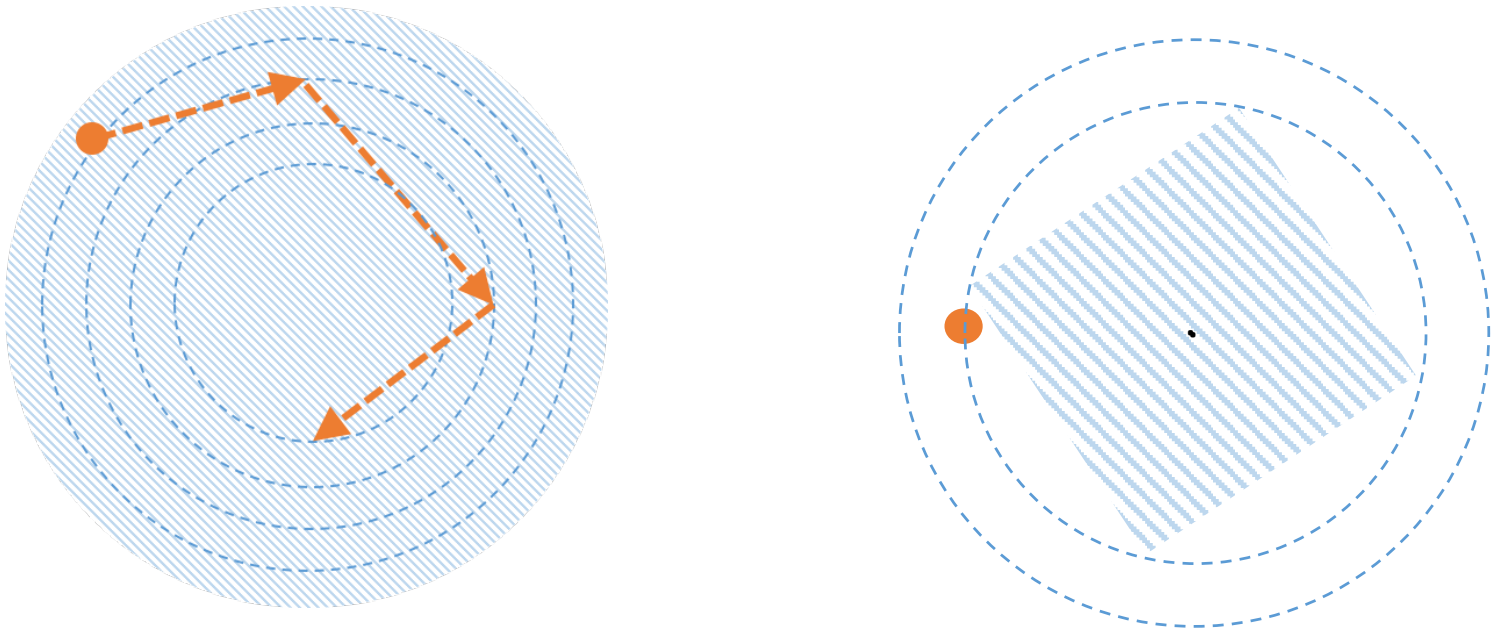
region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# "Failure" of gradient descent?

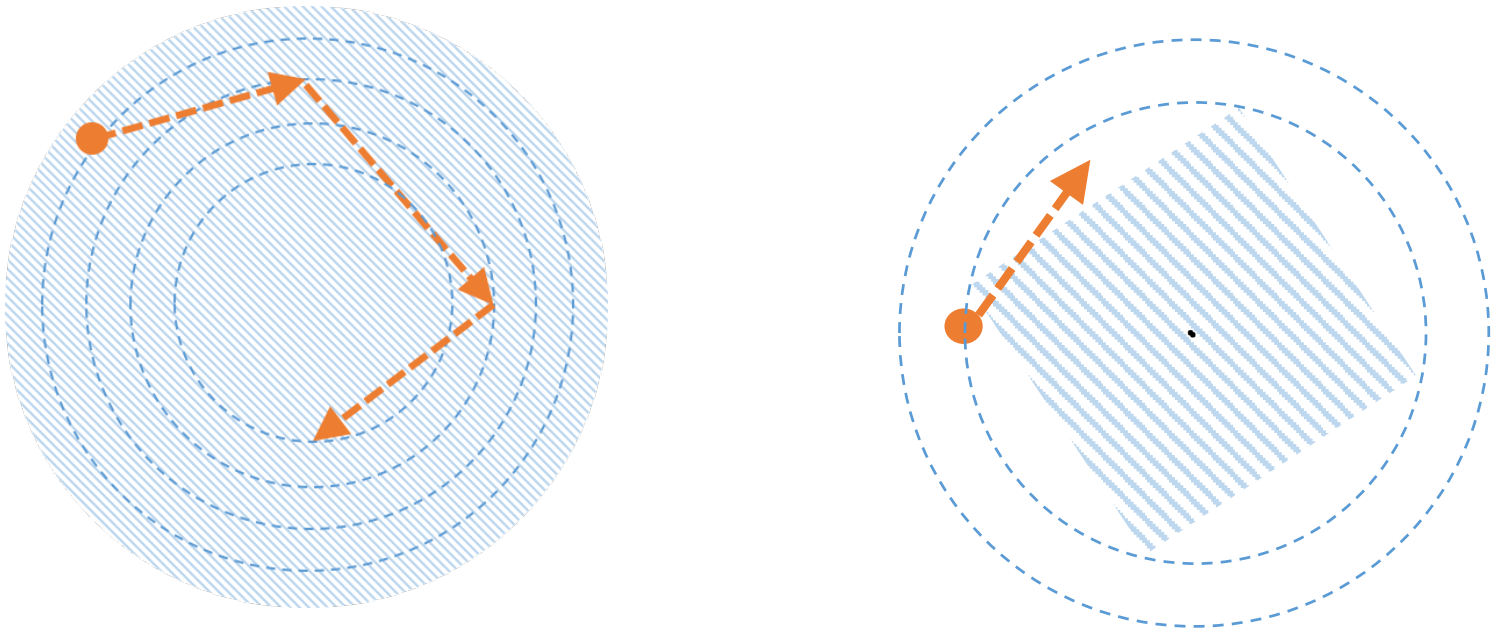
region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# "Failure" of gradient descent?

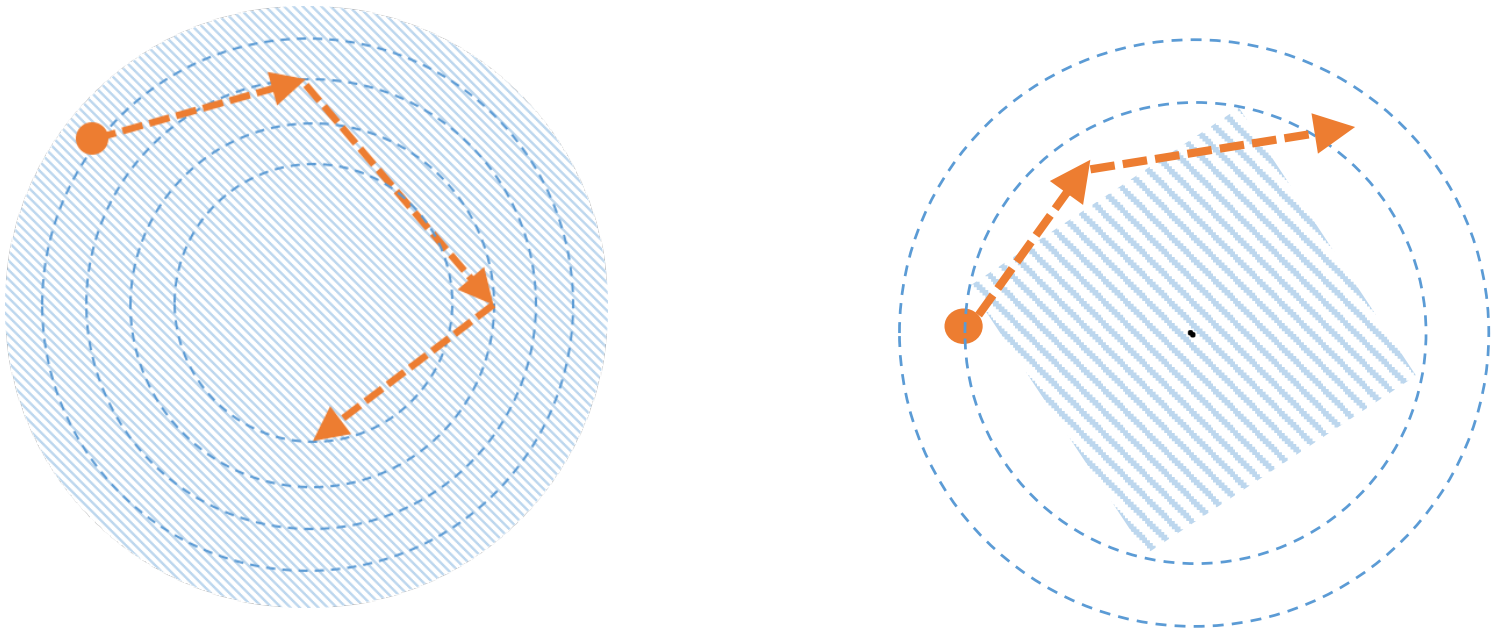
region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# "Failure" of gradient descent?

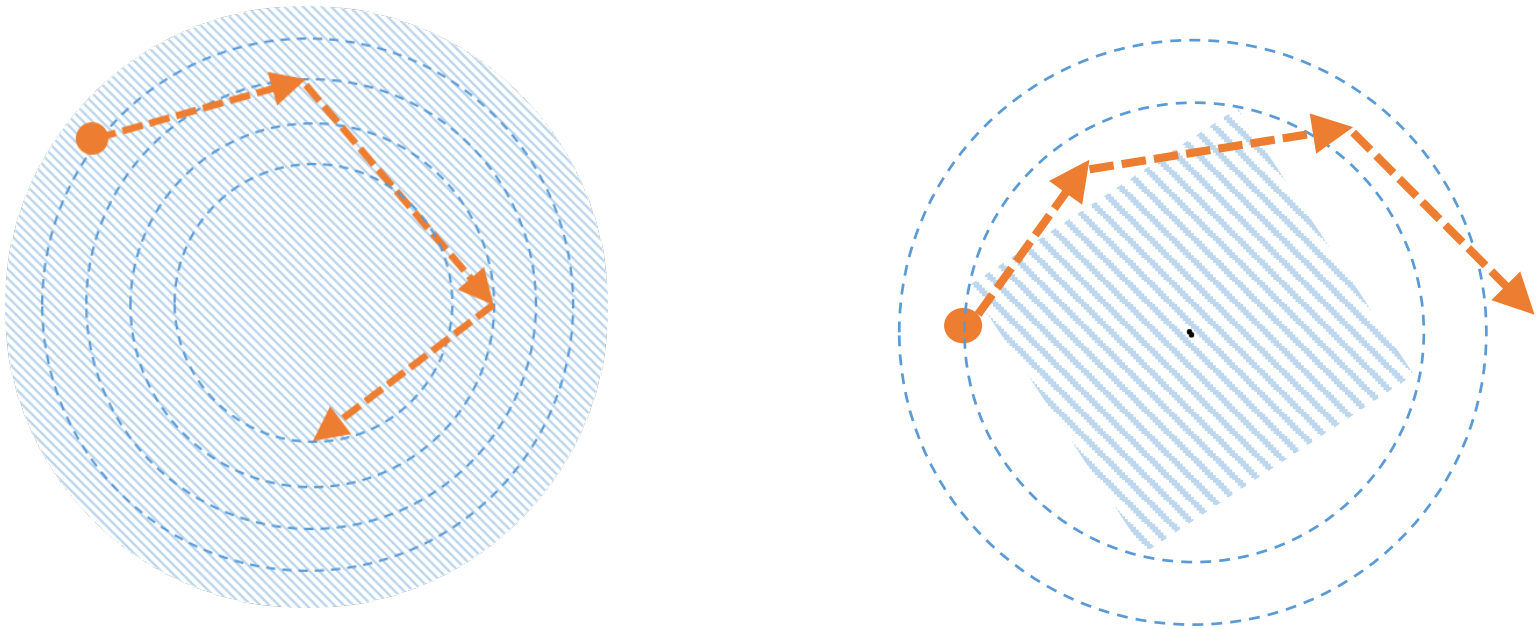
region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# "Failure" of gradient descent?

region of local strong convexity + smoothness



- Prior theory only ensures that iterates remain in $\ell_{\mathrm{F}}$ ball but not incoherence region

# Existing solutions

- regularized loss (solve minimize$_{\boldsymbol{X}}$ $f(\boldsymbol{X}) + R(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16, Chen, Li '17

# Existing solutions

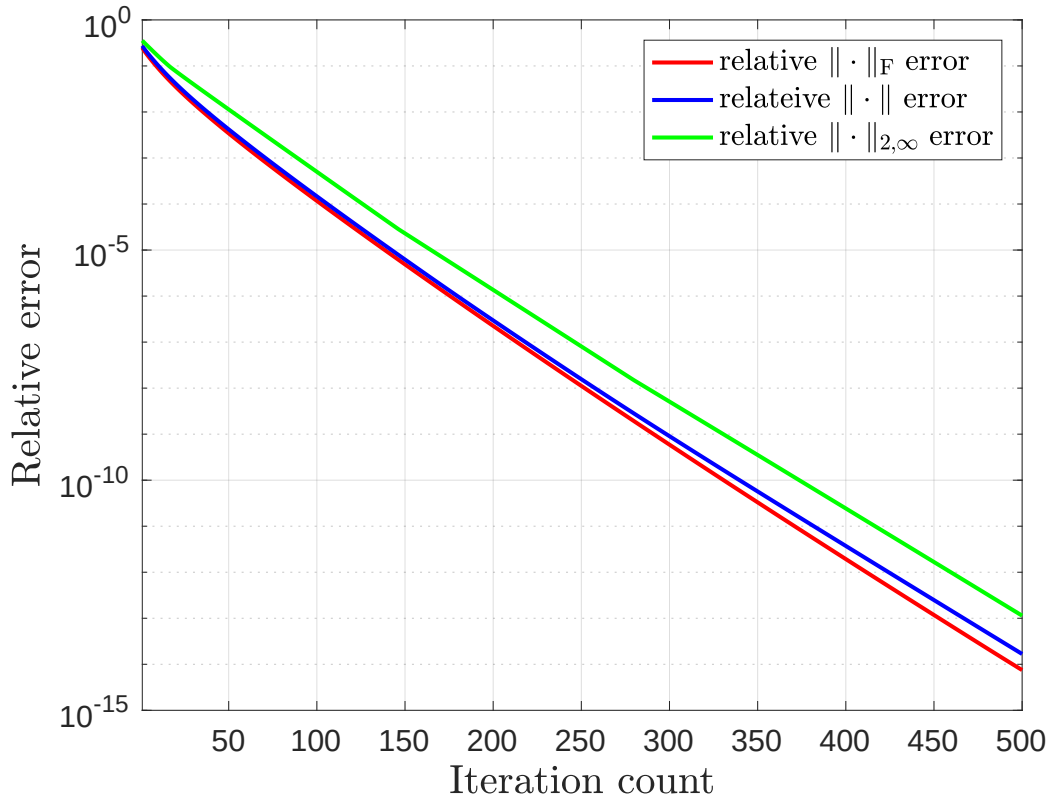
- regularized loss (solve minimize$_{\boldsymbol{X}}$ $f(\boldsymbol{X}) + R(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16, Chen, Li '17

- projection onto set of incoherent matrices
  - e.g. Chen, Wainwright '15, Zheng, Lafferty '16

*Is regularization necessary for nonconvex matrix completion?*

# Numerical surprise with unregularized GD

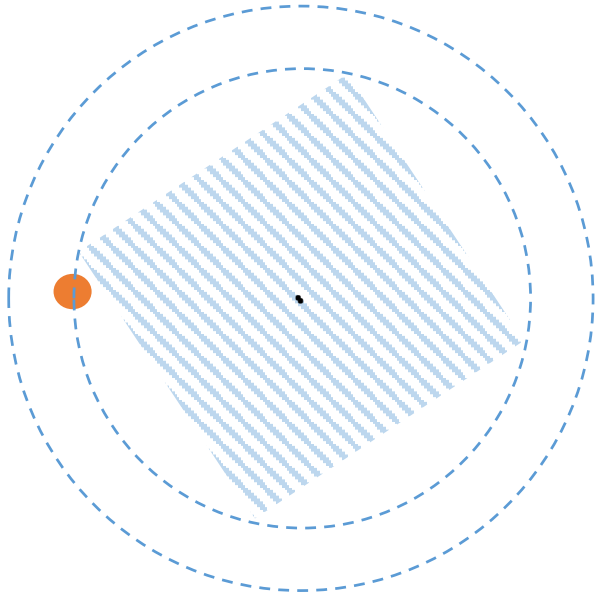
$$n = 1000, r = 10, p = 0.1, \eta = 0.2$$




Vanilla GD without regularization converges fast for MC!

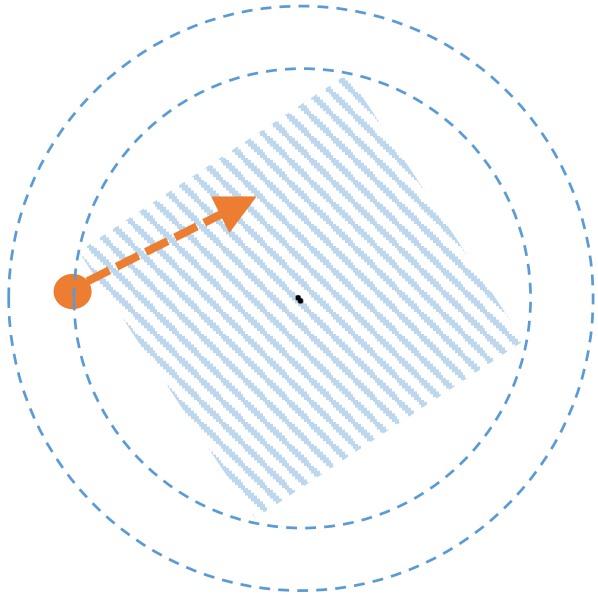
# Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



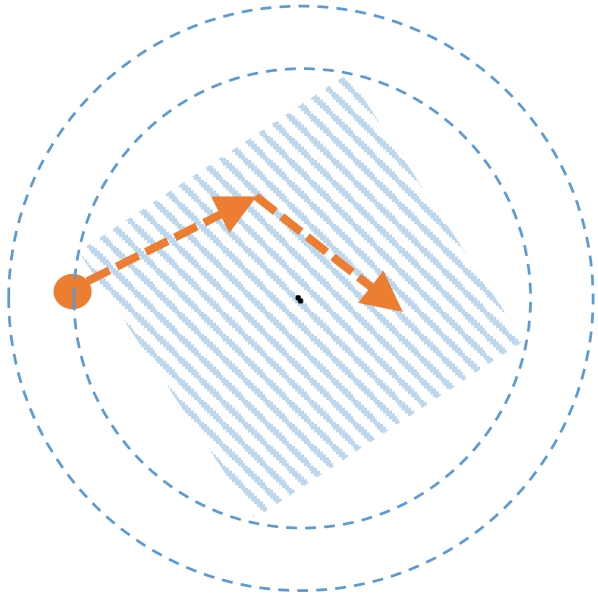
# Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



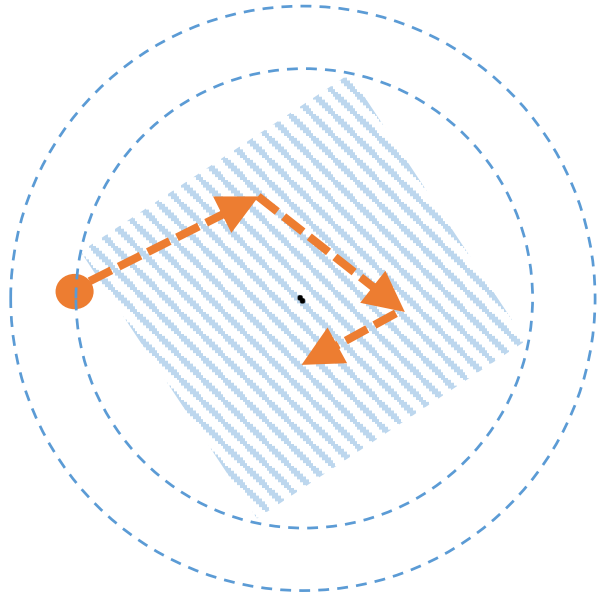
# Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



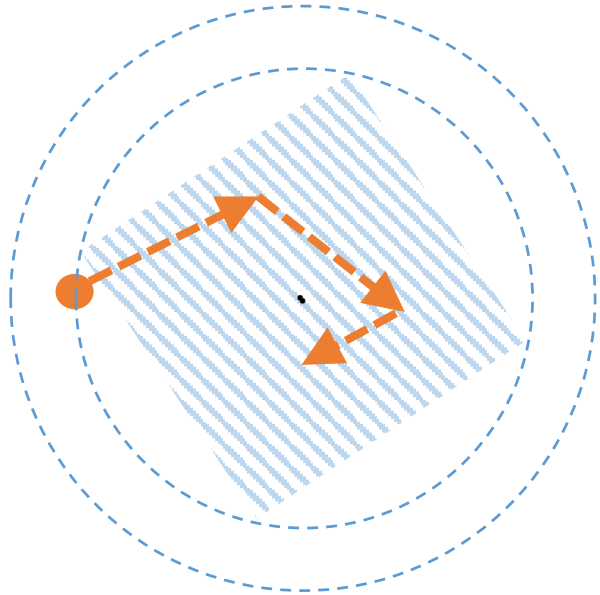
# Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



# Our findings: GD is implicitly regularized

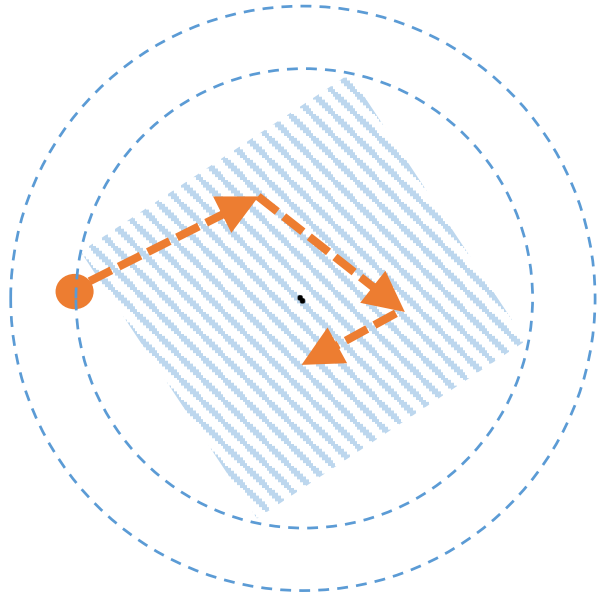
region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent

# Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent

— cannot be derived from generic optimization theory; relies on finer statistical analysis for entire trajectory of GD

# Theoretical guarantees

**Theorem 2 (Matrix completion)**

*Suppose $\boldsymbol{M}$ is rank-$r$, incoherent and well-conditioned. Vanilla gradient descent (with spectral initialization) achieves $\varepsilon$ accuracy*

- *in $O\big(\log \frac{1}{\varepsilon}\big)$ iterations*

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

# Theoretical guarantees

**Theorem 2 (Matrix completion)**

*Suppose* $\boldsymbol{M}$ *is rank-*$r$*, incoherent and well-conditioned. Vanilla gradient descent (with spectral initialization) achieves* $\varepsilon$ *accuracy*

- *in* $O\big(\log \frac{1}{\varepsilon}\big)$ *iterations w.r.t.* $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|$, *and* $\underbrace{\|\cdot\|_{2,\infty}}_{\text{incoherence}}$

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

# Theoretical guarantees

**Theorem 2 (Matrix completion)**

*Suppose* $\boldsymbol{M}$ *is rank-*$r$*, incoherent and well-conditioned. Vanilla gradient descent (with spectral initialization) achieves* $\varepsilon$ *accuracy*

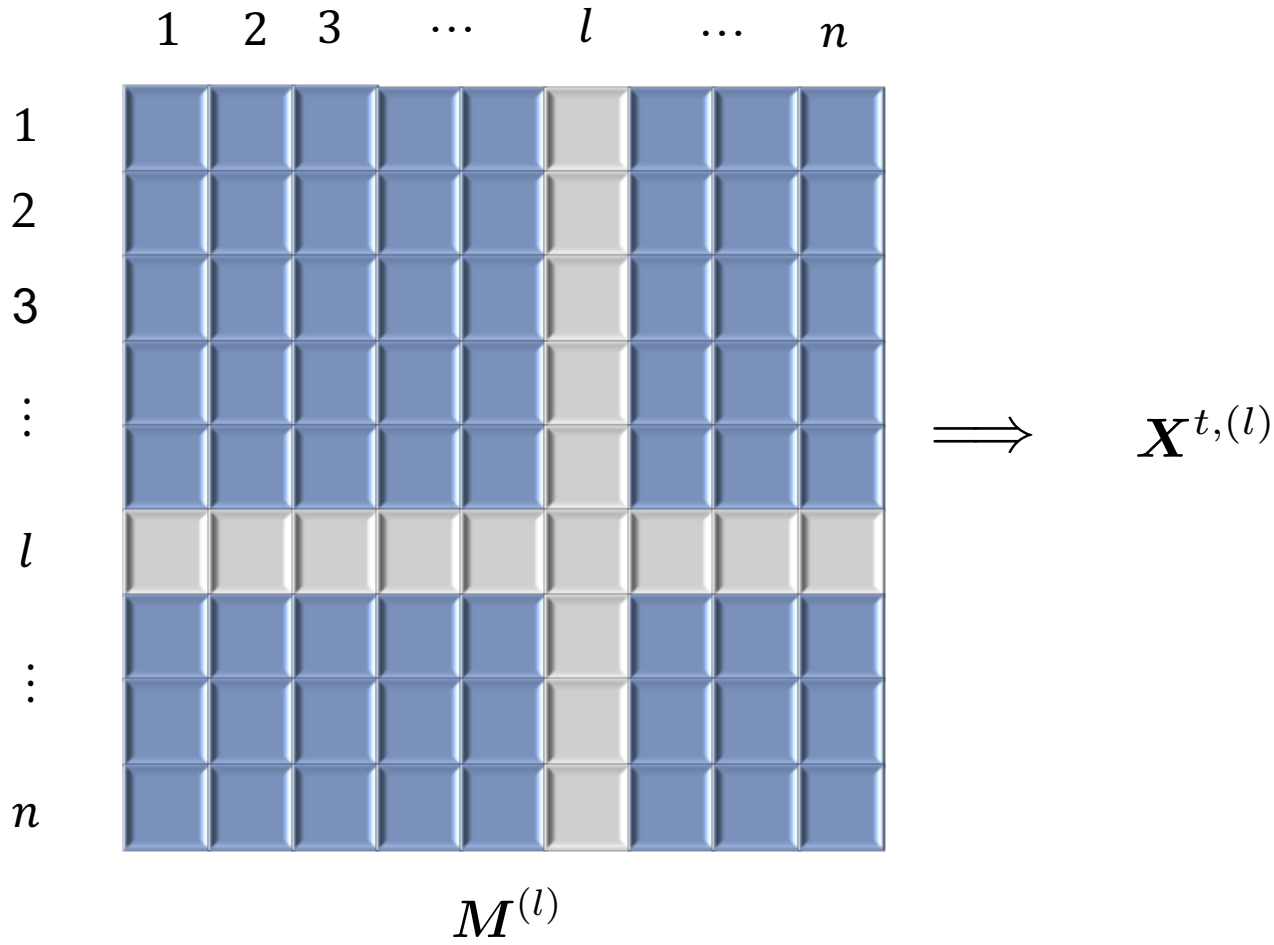
- *in* $O\big(\log \frac{1}{\varepsilon}\big)$ *iterations w.r.t.* $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|$, *and* $\underbrace{\|\cdot\|_{2,\infty}}_{\text{incoherence}}$

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

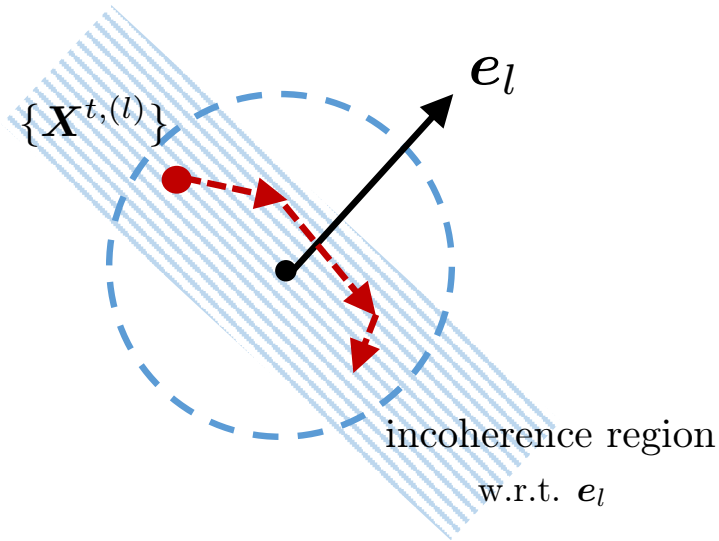
- *Byproduct: vanilla GD controls* ***entrywise error***
  *— errors are spread out across all entries*

# Key ingredient: leave-one-out analysis

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

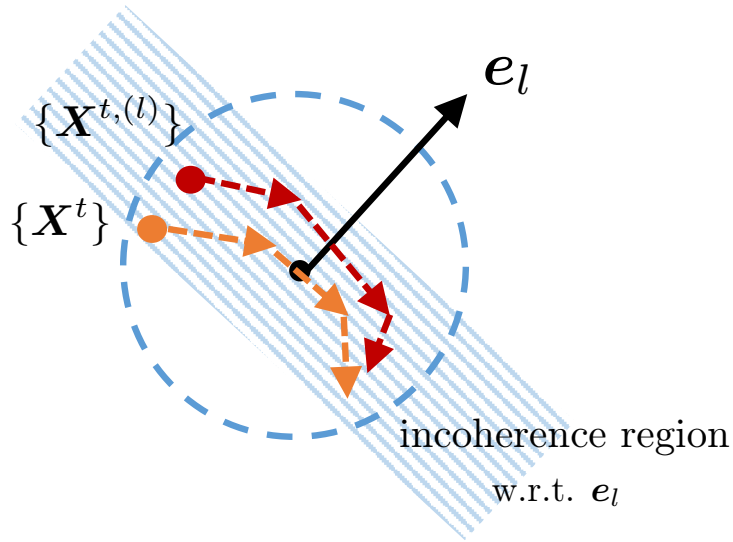


# Key ingredient: leave-one-out analysis



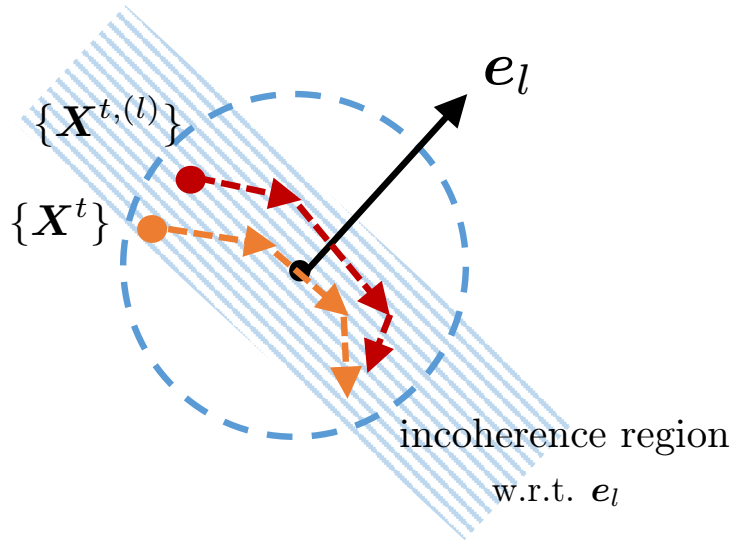

- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ contains more information of $l$th row of $\boldsymbol{X}^\natural$; indep. of randomness in $l$th row

# Key ingredient: leave-one-out analysis




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

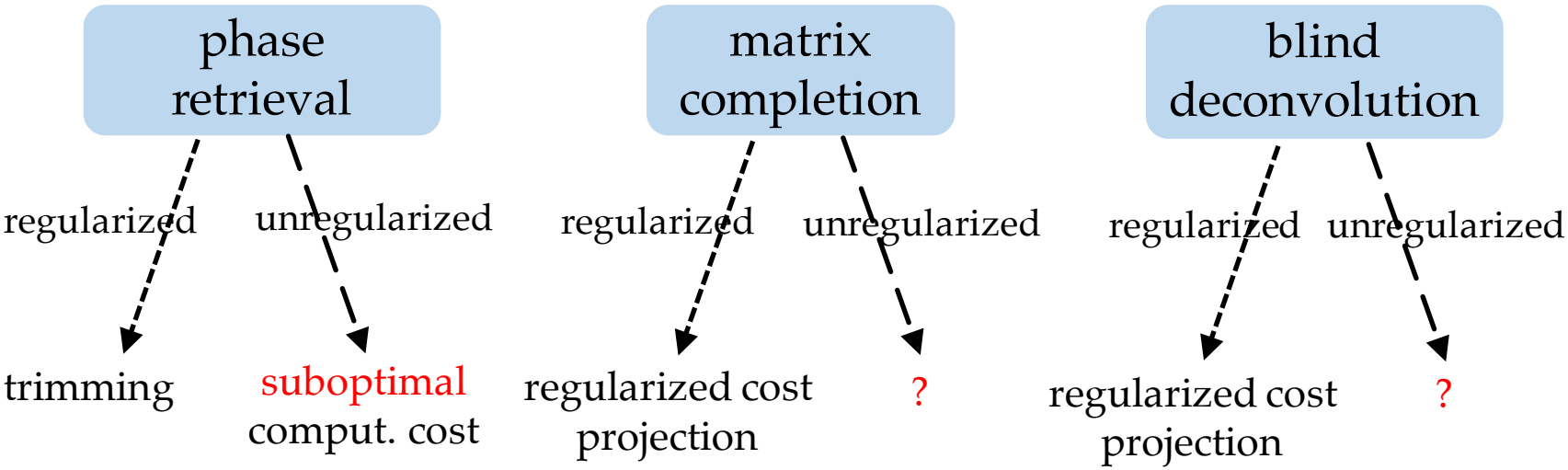
# Key ingredient: leave-one-out analysis




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

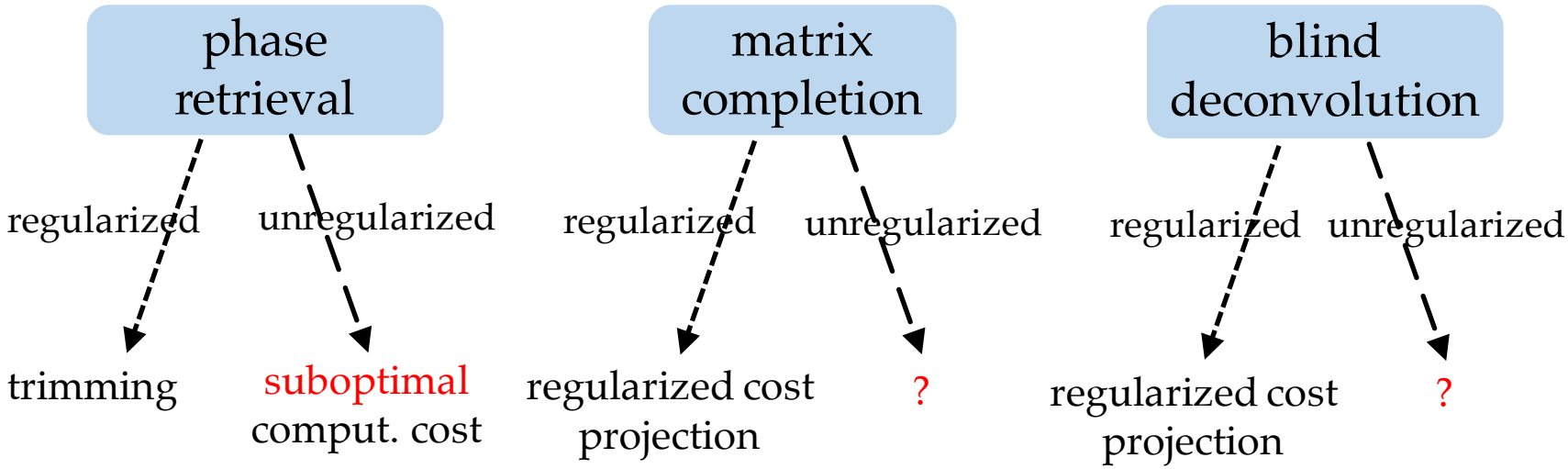
*This phenomenon is quite general*

# How about unregularized gradient methods?




*Are unregularized methods suboptimal for nonconvex estimation?*

# How about unregularized gradient methods?




*Are unregularized methods suboptimal for nonconvex estimation?*

No, for a variety of nonconvex statistical estimation problems.

# Summary

- **Implict regularization:** vanilla gradient descent automatically foces iterates to stay *incoherent*

# Summary

- **Implict regularization:** vanilla gradient descent automatically foces iterates to stay *incoherent*
- Enable error controls in a much stronger sense (e.g. *entrywise error control*)

**Paper**:

"Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution", Cong Ma, Kaizheng Wang, Yuejie Chi, Yuxin Chen, arXiv:1711.10467