JHU vision lab

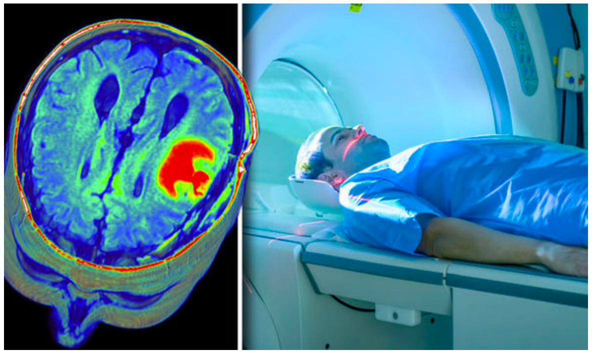
Provable Nonconvex Approaches for Low-rank Models

Zhihui Zhu
Mathematical Institute for Data Science
Johns Hopkins University

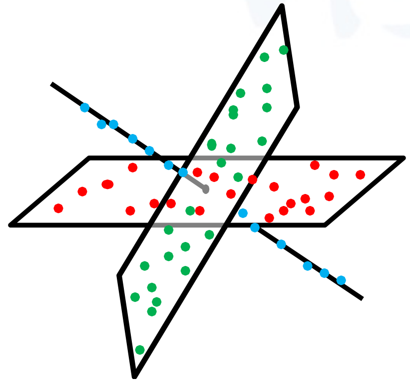
Workshop on Low-Rank Models and Applications
Mons, Belgium, September 12-13, 2019

JOHNS HOPKINS
MATHEMATICAL INSTITUTE *for* DATA SCIENCE

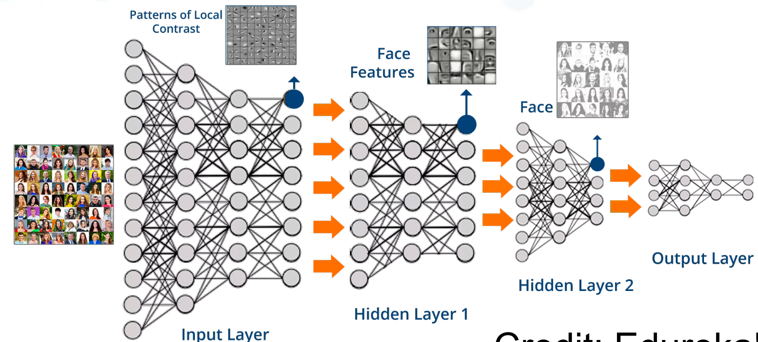
Optimization is a Core Pillar to Data Science



Computational imaging



Machine learning



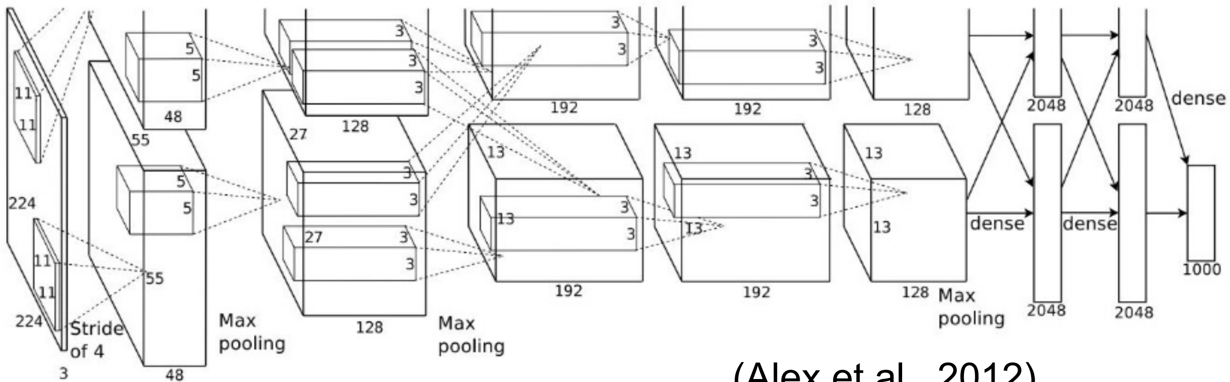
Credit: Edureka!

Artificial intelligence

- Standard formulation:
  - Observed data: $\boldsymbol{y}$
  - Unknown object (signal, image, information, parameter): $\boldsymbol{x}$
  - "loss" function

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{y})$$

What is Deep Learning



"cat"

(Alex et al., 2012)

- Network is trained by solving

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}; \boldsymbol{y})$$

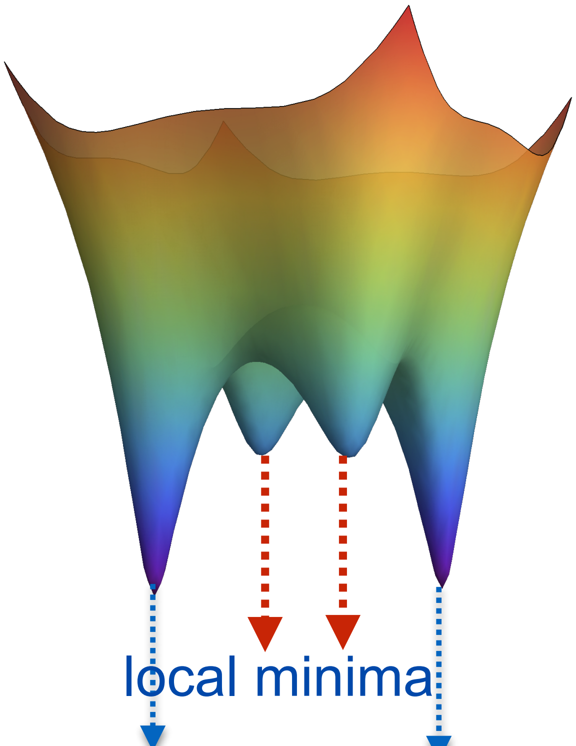
Network parameters

Training data

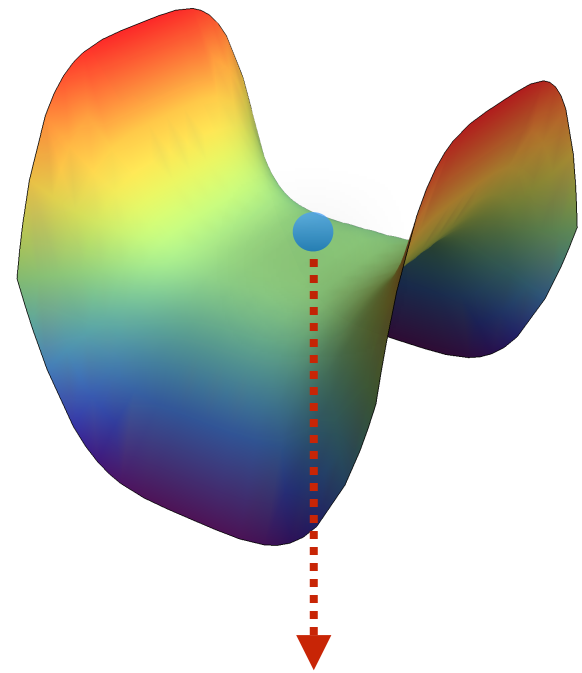
- High-dimensional
- Large-scale
- (Nonsmooth) Nonconvex

Why Nonconvexity is Often Scary?

“bad” local minima



“flat” saddle point



Computing a local minimizer of a general nonconvex problem could be NP-hard!

[1] Murty & Kabadi, Some NP-complete problems in quadratic and nonlinear programming, Mathematical programming, 1987.

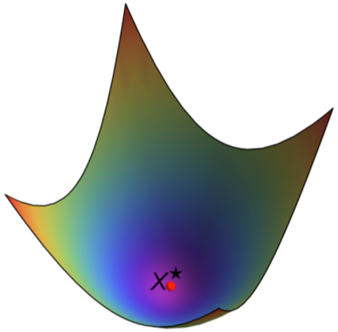
Why Nonsmoothness Makes Scarier?

Credit: Scott T Leutenegger

Not differentiable

Convex vs Nonconvex Optimization

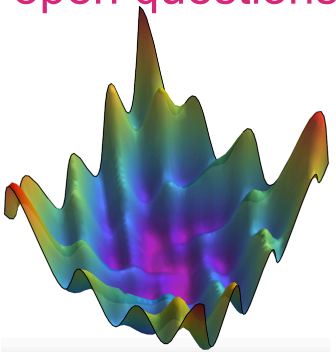
- Convex optimization:
  - Guaranteed convergence/optimality



Phase retrieval
Blind deconvolution
Matrix sensing
Matrix completion
Robust PCA
Clustering

  - Large problem size
  - Non-scalable

- Nonconvex optimization
  - Many open questions



  - Small problem size
  - Scalable

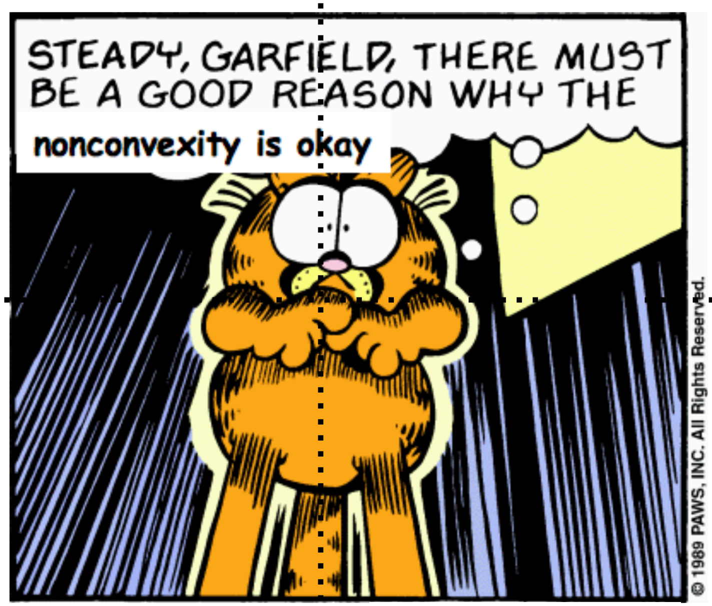
- Many problems are fundamentally nonconvex:

PCA Dictionary learning NMF Deep learning

- How to deal with nonconvexity 

Optimality Guarantee for Nonconvex Problems

Non-smooth



Credit: Chi & Chen

Smooth

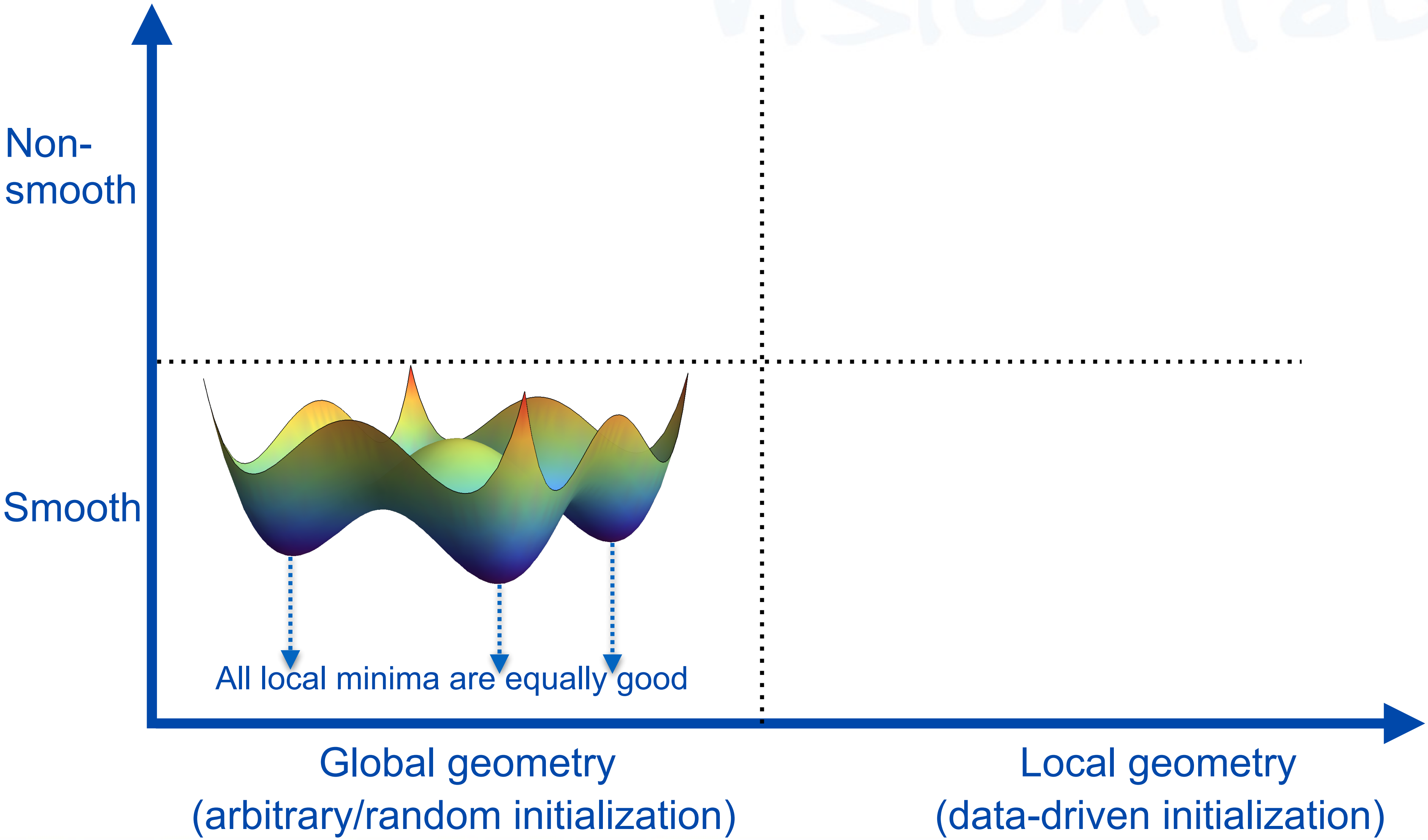
Global geometry
(arbitrary/random initialization)

Local geometry
(data-driven initialization)

See http://sunju.org/research/nonconvex/ for a detailed list of references

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Optimality Guarantee for Nonconvex Problems



See http://sunju.org/research/nonconvex/ for a detailed list of references

Optimality Guarantee for Nonconvex Problems

Non-smooth

Smooth

1st/2nd-order methods [Lee et al., 2017]
Alternating minimization [Li et al. 2019]
Shallow linear neural network (NN) [Kawaguchi 2016; Zhu et al., 2018]
Low-rank matrix opt [Zhu et al., 2017]
Matrix completion [Ma et al., 2016]
Dictionary learning [Sun et al., 2016]
Phase retrieval [Sun et al., 2016]

Global geometry (arbitrary/random initialization)

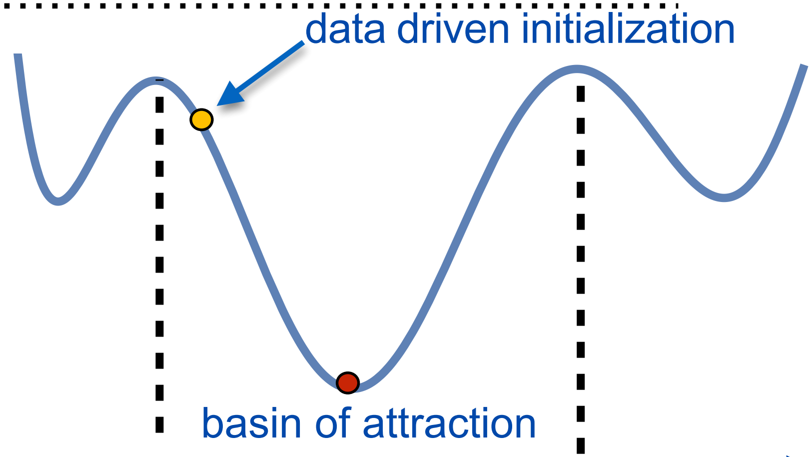
Local geometry (data-driven initialization)

Optimality Guarantee for Nonconvex Problems

Non-smooth

Smooth

1st/2nd-order methods [Lee et al., 2017]
Alternating minimization [Li et al. 2019]
Shallow linear neural network (NN) [Kawaguchi 2016; Zhu et al., 2018]
Low-rank matrix opt [Zhu et al., 2017]
Matrix completion [Ma et al., 2016]
Dictionary learning [Sun et al., 2016]
Phase retrieval [Sun et al., 2016]



Global geometry (arbitrary/random initialization)

Local geometry (data-driven initialization)

Optimality Guarantee for Nonconvex Problems

Non-smooth

Smooth

Global geometry (arbitrary/random initialization)

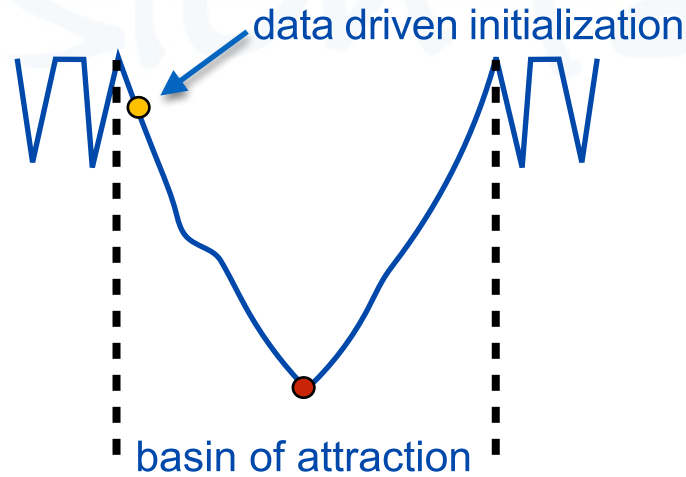
- 1st/2nd-order methods [Lee et al., 2017]
- Alternating minimization [Li et al. 2019]
- Shallow linear neural network (NN) [Kawaguchi 2016; Zhu et al., 2018]
- Low-rank matrix opt [Zhu et al., 2017]
- Matrix completion [Ma et al., 2016]
- Dictionary learning [Sun et al., 2016]
- Phase retrieval [Sun et al., 2016]

Local geometry (data-driven initialization)

- first-order methods [Candes et al., 2014]
- 2nd-order methods [Yue et al, 2018]
- Tensor recovery [Jain et al., 2013]
- Matrix completion [Sun et al, 2014, Verdun 2019]
- Matrix sensing [Tu et al, 2016]
- Blind phase retrieval [Qu et al., 2014]
- Blind deconvolution [Li et al., 2016; Qu et al. 2019]
- Phase retrieval [Candes et al., 2014]

See http://sunju.org/research/nonconvex/ for a detailed list of references

Optimality Guarantee for Nonconvex Problems



Non-smooth

Smooth

1st/2nd-order methods [Lee et al., 2017]
Alternating minimization [Li et al. 2019]
Shallow linear neural network (NN) [Kawaguchi 2016; Zhu et al., 2018]
Low-rank matrix opt [Zhu et al., 2017]
Matrix completion [Ma et al., 2016]
Dictionary learning [Sun et al., 2016]
Phase retrieval [Sun et al., 2016]

first-order methods [Candes et al., 2014]
2nd-order methods [Yue et al, 2018]
Tensor recovery [Jain et al., 2013]
Matrix completion [Sun et al, 2014, Verdun 2019]
Matrix sensing [Tu et al, 2016]
Blind phase retrieval [Qu et al., 2014]
Blind deconvolution [Li et al., 2016; Qu et al. 2019]
Phase retrieval [Candes et al., 2014]

Global geometry (arbitrary/random initialization)

Local geometry (data-driven initialization)

See http://sunju.org/research/nonconvex/ for a detailed list of references

Optimality Guarantee for Nonconvex Problems

Non-smooth

Local geometry (data-driven initialization):

Riemannian subgradient
- robust subspace learning [Zhu et al., 2018]
- dictionary learning [Bai et al. , 2019; Zhu et al., 2019]

Subgradient method
- overparameterized NN [Du et al., 2019]
- robust low-rank recovery[Li et al., 2018]
- robust phase retrieval [Davis et al., 2017]

Smooth

Global geometry (arbitrary/random initialization):

- 1st/2nd-order methods [Lee et al., 2017]
- Alternating minimization [Li et al. 2019]
- Shallow linear neural network (NN) [Kawaguchi 2016; Zhu et al., 2018]
- Low-rank matrix opt [Zhu et al., 2017]
- Matrix completion [Ma et al., 2016]
- Dictionary learning [Sun et al., 2016]
- Phase retrieval [Sun et al., 2016]

Local geometry (data-driven initialization):

- first-order methods [Candes et al., 2014]
- 2nd-order methods [Yue et al, 2018]
- Tensor recovery [Jain et al., 2013]
- Matrix completion [Sun et al, 2014, Verdun 2019]
- Matrix sensing [Tu et al, 2016]
- Blind phase retrieval [Qu et al., 2014]
- Blind deconvolution [Li et al., 2016; Qu et al. 2019]
- Phase retrieval [Candes et al., 2014]

Global geometry (arbitrary/random initialization)

Local geometry (data-driven initialization)

See http://sunju.org/research/nonconvex/ for a detailed list of references

Optimality Guarantee for Nonconvex Problems

Non-smooth

Riemannian subgradient
robust subspace learning [Zhu et al., 2018]
dictionary learning
[Bai et al. , 2019; Zhu et al., 2019]

Subgradient method
overparameterized NN [Du et al., 2019]
robust low-rank recovery[Li et al., 2018]
robust phase retrieval [Davis et al., 2017]

Smooth

1st/2nd-order methods [Lee et al., 2017]
Alternating minimization [Li et al. 2019]
Shallow linear neural network (NN)
[Kawaguchi 2016; Zhu et al., 2018]
Low-rank matrix opt [Zhu et al., 2018]
Matrix completion [Ma et al., 2016]
Dictionary learning [Sun et al., 2016]
Phase retrieval [Sun et al., 2016]

first-order methods [Candes et al., 2014]
2nd-order methods [Yue et al, 2018]
Tensor recovery [Jain et al., 2013]
Matrix completion [Sun et al, 2014, Verdun 2019]
Matrix sensing [Tu et al, 2016]
Blind phase retrieval [Qu et al., 2014]
Blind deconvolution [Li et al., 2016; Qu et al. 2019]
Phase retrieval [Candes et al., 2014]

Global geometry
(arbitrary/random initialization)

Local geometry
(data-driven initialization)

See http://sunju.org/research/nonconvex/ for a detailed list of references

Outline

- Global geometric analysis for nonconvex optimization
  - Low-rank matrix optimization
- From global to local geometric analysis: building practical optimization methods
  - Robust low-rank matrix recovery
- Incremental methods for weakly convex optimization
- Conclusion and future work

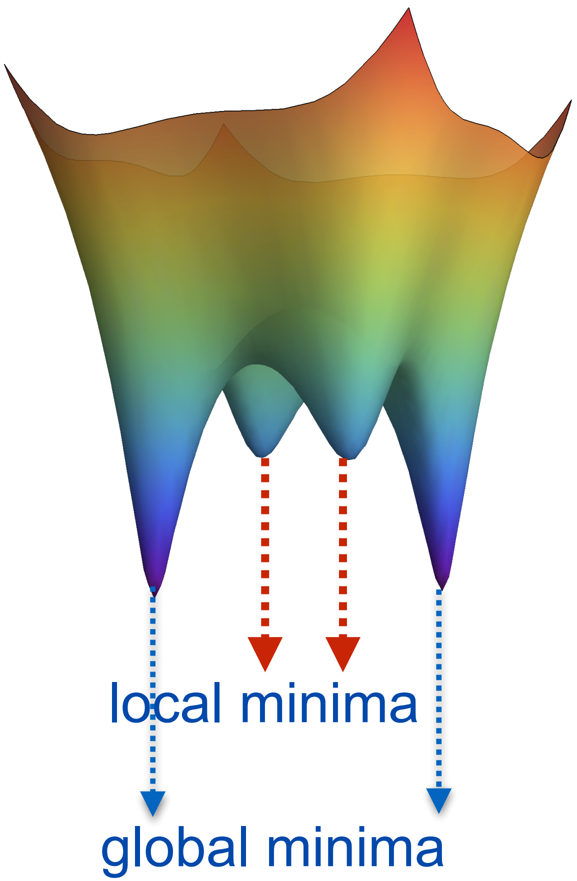
JHU vision lab

Global Geometric Analysis for Nonconvex Optimization

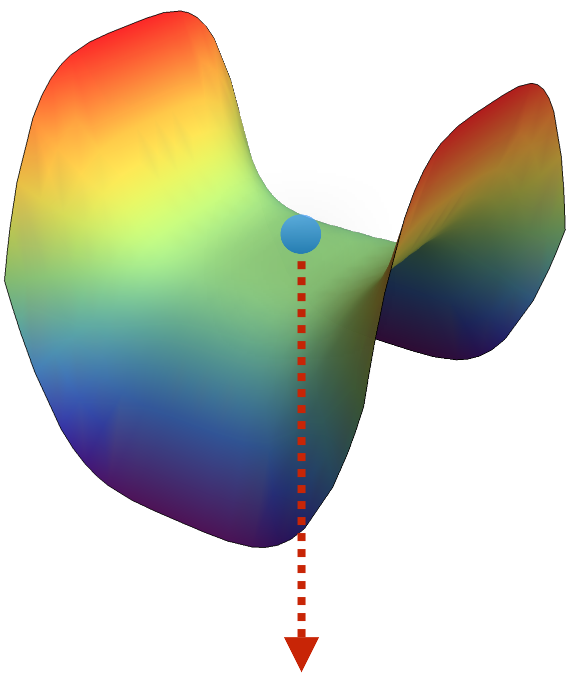
JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Recap: Why Nonconvexity is Often Scary?

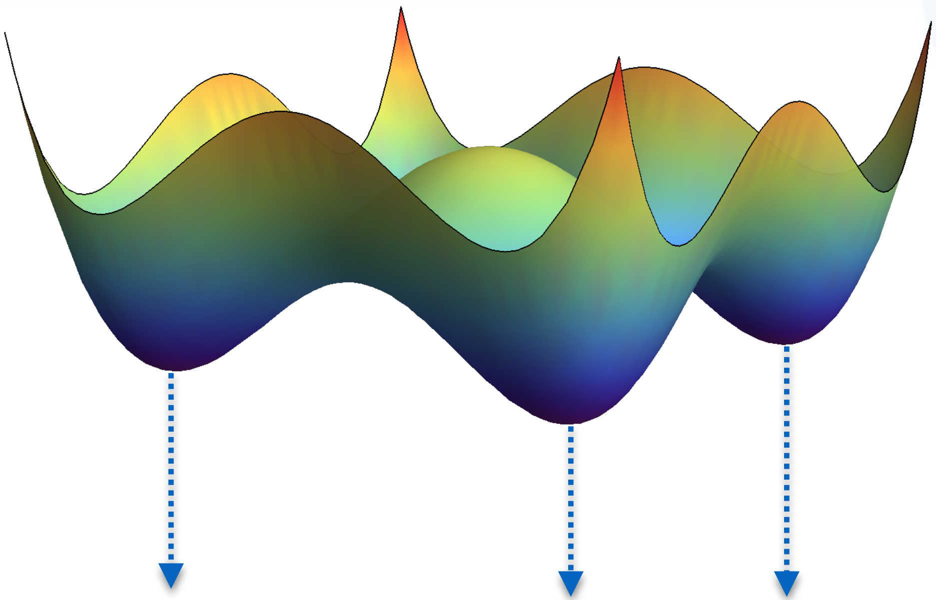
“bad” local minima



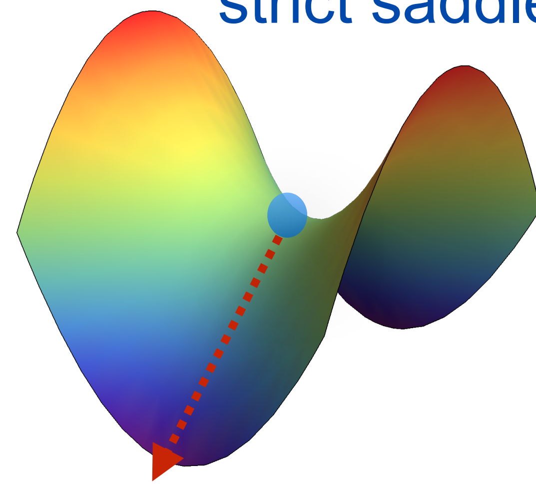
“flat” saddle point



When Nonconvexity is Not Scary?



All local minima are equally good



strict saddle

Negative curvature

- No spurious local minima: all local minima are global

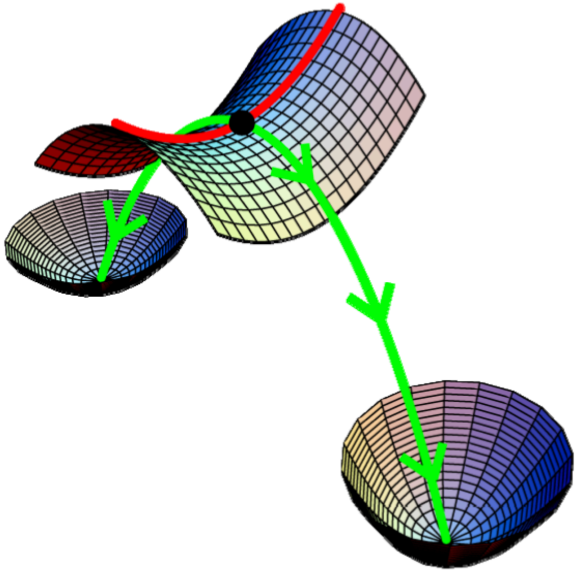
- Strict saddle function: any critical point that is not a local minimum is a strict saddle

Benign Geometry Leads to Algorithm Design

- Benign geometric structures allow simple, efficient iterative methods to obtain global optima.

- First-order algorithms:
  - Gradient decent with random initialization
  - Perturbed gradient descent

- Second-order algorithms:
  - Cubic regularization of Newton method
  - Trust region algorithm
  - Gradient + Negative Curvature Direction
- Alternating minimization



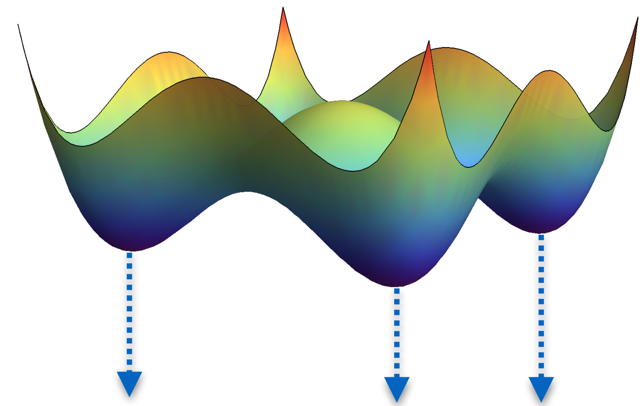
Credit: Turnhout et al.

[1] Lee, Simchowitz, Jordan & Recht, B, Gradient descent converges to minimizers, COLT, 2016.
[2] Jin, Ge, Netrapalli, Kakade, Jordan, How to escape saddle points efficiently, ICML, 2017.
[3] Nesterov & Polyak, Cubic regularization of newton method and its global performance, MP, 2006.
[4] Curtis, Robinson & Samadi, A trust region algorithm with a worst-case iteration complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization, MP, 2014.
[5] Curtis & Robinson, Exploiting negative curvature in deterministic and stochastic optimization, 2017
[6] Li*, **Z.*** & G. Tang, Alternating Minimizations Converge to Second-Order Optimal Solutions, ICML, 2019

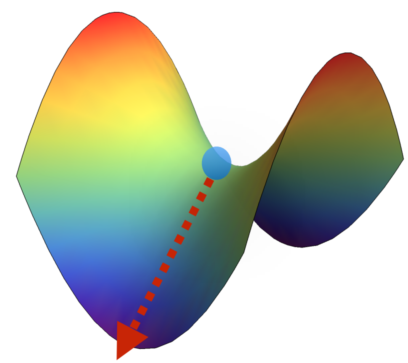
JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Examples that Obey Benign Geometry

- **Low-rank matrix optimization** [this talk]
- Shallow linear neural network
- Orthogonal tensor decomposition
- Phase retrieval
- Dictionary learning
- Phase synchronization/Community detection
- ……



All local minima are equally good



Negative curvature

[1] Kawaguchi, Deep learning without poor local minima, NIPS, 2016.
[2] **Z.**, Soudry, Eldar & Wakin, The global optimization geometry of shallow linear neural networks, 2018.
[3]Ge et al., Escaping from saddle points—online stochastic gradient for tensor decomposition, COLT, 2015.
[4] Sun, Qu & Wright, Complete dictionary recovery over the sphere, TIT, 2018.
[5] Sun, Qu & Wright, A geometric analysis of phase retrieval, FOCS, 2018.
[6] Ling, Xu & Bandeira, On the Landscape of Synchronization Networks: A Perspective from Nonconvex Optimization, 2018.
[7] Uschmajew and Vandereycken, On critical points of quadratic low-rank matrix optimization problems, 2018

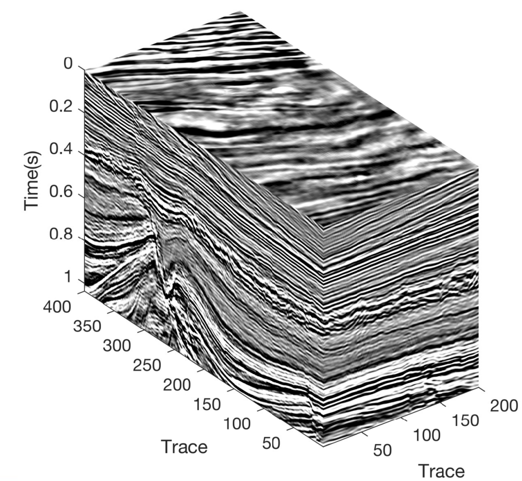
Outline

- Global geometric analysis for nonconvex optimization
  - **Low-rank matrix optimization**
- From global to local geometric analysis: building practical optimization methods
  - Robust low-rank matrix recovery
- Incremental methods for weakly convex optimization
- Conclusion and future work

[1] **Z.**, Li, Tang & Wakin, Global optimality in low-rank matrix optimization, TSP, 2018.
[2] Li, **Z.**, Tang & Wakin, The non-convex geometry of low-rank matrix optimization, Information and Inference, 2018.

Low-rank Modeling

- Low-rank modeling appears widely
  - Nonnegative matrix factorization
  - Subspace clustering
  - Matrix completion
  - Seismic denoising
  - Quantum state tomography
  - Video surveillance
  - ……



Low-rank Modeling: Convex Approach

- Nuclear norm optimization

nuclear norm $\sum_i \sigma_i(\boldsymbol{X})$

$$\min_{\boldsymbol{X} \in \mathbb{R}^{m \times n}} f(\boldsymbol{X}; \boldsymbol{Y}) + \lambda \|\boldsymbol{X}\|_*$$

- Matrix completion

$$\min_{\boldsymbol{X} \in \mathbb{R}^{m \times n}} \sum_{\text{observed } (i,j)} (\boldsymbol{X}_{ij} - \boldsymbol{Y}_{ij})^2 + \lambda \|\boldsymbol{X}\|_*$$

- Matrix sensing

$$\min_{\boldsymbol{X} \in \mathbb{R}^{m \times n}} \|\mathcal{A}(\boldsymbol{X}) - \mathcal{A}(\boldsymbol{Y})\|_2^2 + \lambda \|\boldsymbol{X}\|_*$$

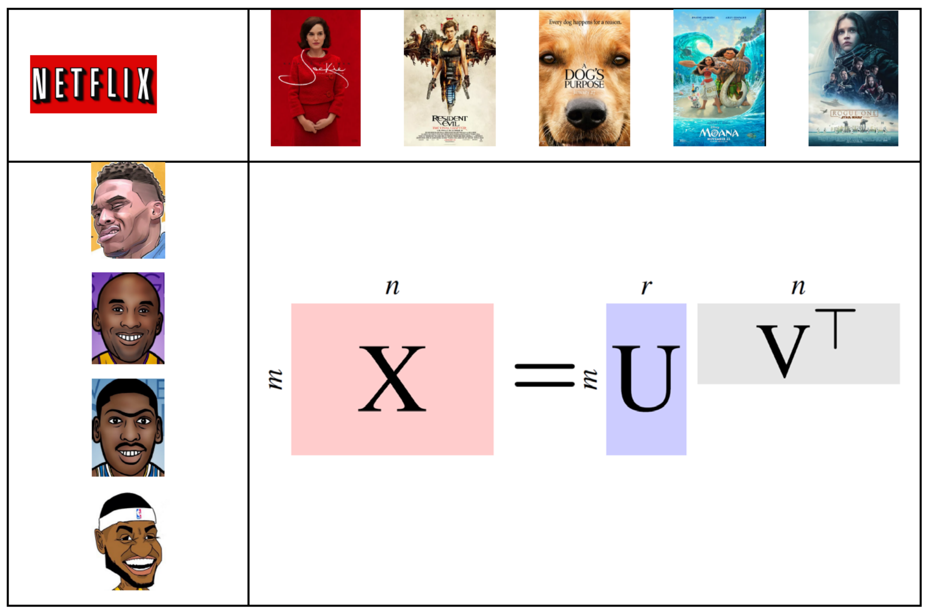
[1] Candès & Recht. "Exact matrix completion via convex optimization." FOCM, 2009
[2] Recht, Fazel&Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." SIAM review, 2011.

Low-rank Modeling: Nonconvex Approach

- Matrix factorization approach:

$$\min_{\boldsymbol{U},\boldsymbol{V}} f(\boldsymbol{U}\boldsymbol{V}^\top;\boldsymbol{Y}) + \frac{\lambda}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2)$$

- It is nonconvex due to the bilinearity $(\boldsymbol{U},\boldsymbol{V}) \to \boldsymbol{U}\boldsymbol{V}^\top$

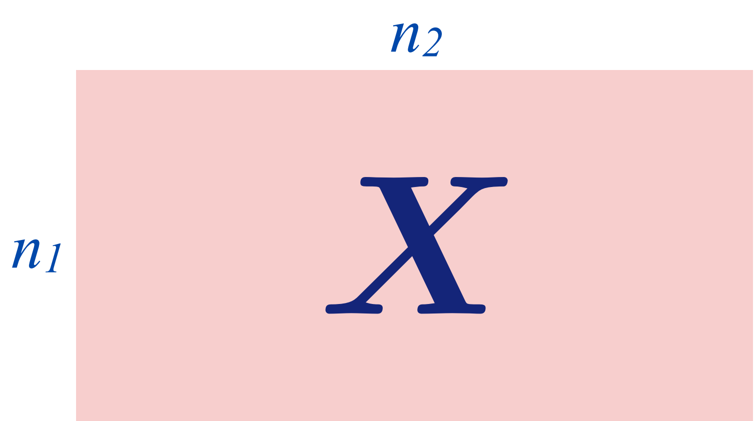


[1] Burer & Monteiro, A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization, Mathematical Programming, 2003.

Low-rank Modeling: Convex VS Nonconvex

- Convex formulation:

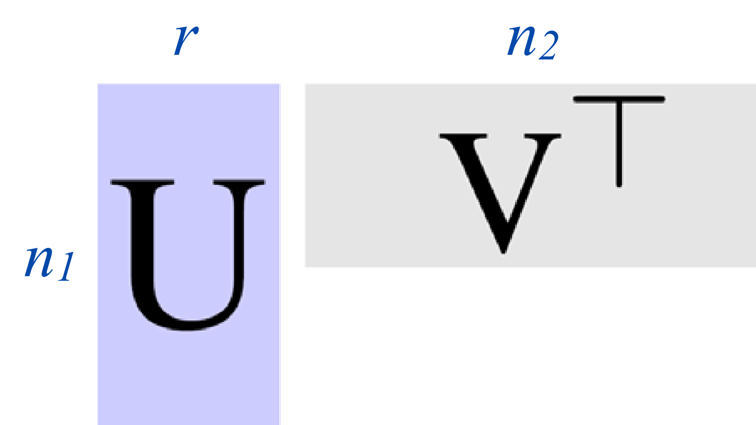
$$\min_{\boldsymbol{X} \in \mathbb{R}^{m \times n}} f(\boldsymbol{X}; \boldsymbol{Y}) + \lambda \|\boldsymbol{X}\|_*$$



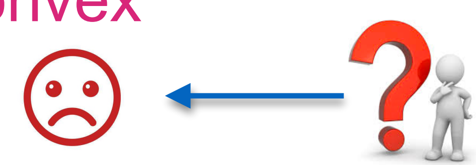
- Convex
  - Global optimality
  - Information-theoretical guarantee
- Problem size: $n_1 n_2$
- First-order solvers require SVD

- Factorization formulation:

$$\min_{\boldsymbol{U}, \boldsymbol{V}} f(\boldsymbol{U}\boldsymbol{V}^\top; \boldsymbol{Y}) + \frac{\lambda}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2)$$



- Nonconvex



- Problem size: $(n_1+n_2)r$
- No SVD required

[1] Chi, Lu, Chen, Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview, TSP, 2019.

Prior Guarantees for Matrix Factorization

- Prior guarantees for Matrix Factorization:
  - [1]: matrix sensing & global geometry & exact-parameterization & no regularizer
  - [2]: matrix completion & global geometry & exact-parameterization & no regularizer
  - [3]: general function & local geometry & no regularizer
  - [4]: general function & condition when a local minimum is global & nuclear norm
  - [5]: general function & condition when a local minimum is global & more regularizers
  - [6]: quadratic function & global geometry & exact-parameterization
  - ……
- Our work
  - general function & global geometry & nuclear norm
  - allows both exact-parameterization $r = \operatorname{rank}(\boldsymbol{X}^\star)$ and over-parameterization $r > \operatorname{rank}(\boldsymbol{X}^\star)$

[1] Bhojanapalli, Neyshabur & Srebro, Global optimality of local search for low rank matrix recovery, NIPS, 2016
[2] Ge, Lee & Ma, Matrix completion has no spurious local minimum, NIPS, 2006.
[3] Park et al., Finding low-rank solutions via non-convex matrix factorization, efficiently and provably, SIAMIS, 2018.
[4] Cabral et al., Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition, ICCV, 2013.
[5]Haeffele&Vidal, Structured low-rank matrix factorization: Global optimality, algorithms, and applications, 2017
[6] Uschmajew and Vandereycken, On critical points of quadratic low-rank matrix optimization problems, 2018

Benign Geometric Structure of Factorization Formulation

Convex formulation:

$$\boldsymbol{X}^\star \in \arg\min_{\boldsymbol{X}} f(\boldsymbol{X};\boldsymbol{Y}) + \lambda\|\boldsymbol{X}\|_*$$

Factorization formulation:

$$g(\boldsymbol{U},\boldsymbol{V}) = f(\boldsymbol{U}\boldsymbol{V}^\top;\boldsymbol{Y}) + \frac{\lambda}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2)$$

- **Restricted strongly convex and smooth:**

  for any $\operatorname{rank}(\boldsymbol{X}) \le 2r$, we have

$$\alpha\mathbf{I} \preceq \nabla^2_{\boldsymbol{X}\boldsymbol{X}} f(\boldsymbol{X};\boldsymbol{Y}) \preceq \beta\mathbf{I}, \quad \beta/\alpha \le 1.5$$

Satisfied in:
Matrix factorization
Matrix sensing
1-bit matrix recovery

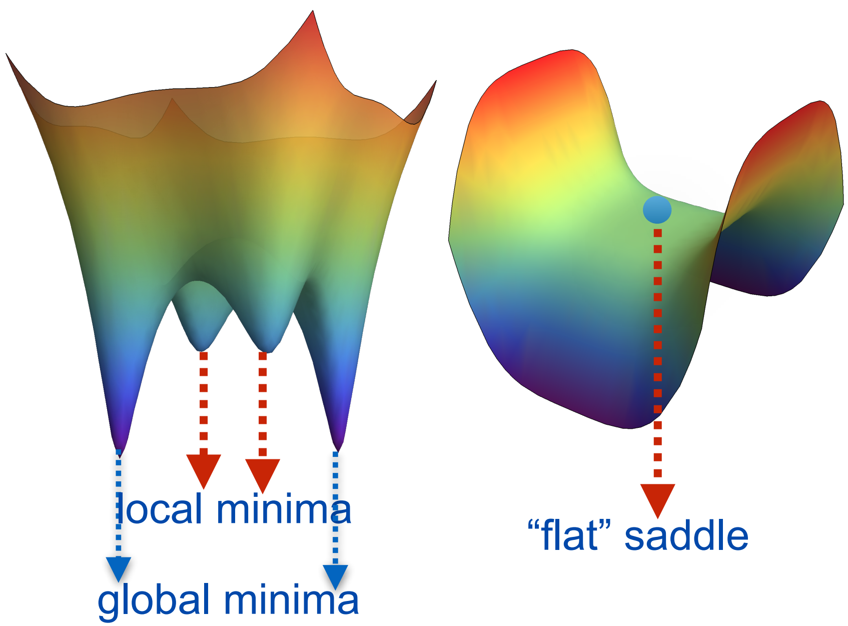
- **Theorem:** Let $\lambda \ge 0$. Suppose *f* satisfies above condition and set $r \ge \operatorname{rank}(\boldsymbol{X}^\star)$. Then any critical point (***U,V***) of the function *g* is either a global minimum ($\boldsymbol{U}\boldsymbol{V}^\top = \boldsymbol{X}^\star$), or a strict saddle.

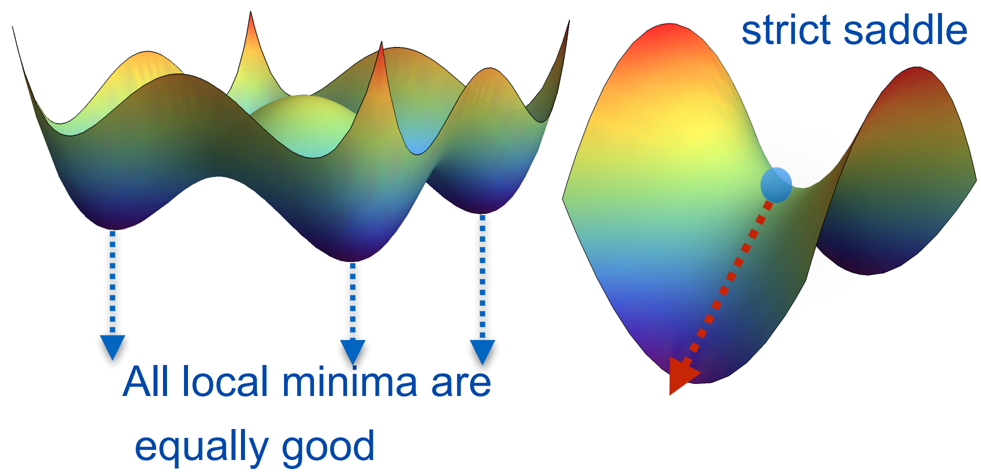
[1] **Z.**, Li, Tang & Wakin, Global optimality in low-rank matrix optimization, TSP, 2018.
[2] Li, **Z.**, Tang & Wakin, The non-convex geometry of low-rank matrix optimization, Information and Inference, 2018.

Benign Geometric Structure of Factorization Formulation

Factorization formulation:

$$g(\boldsymbol{U}, \boldsymbol{V}) = f(\boldsymbol{U}\boldsymbol{V}^\top; \boldsymbol{Y}) + \frac{\lambda}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2)$$





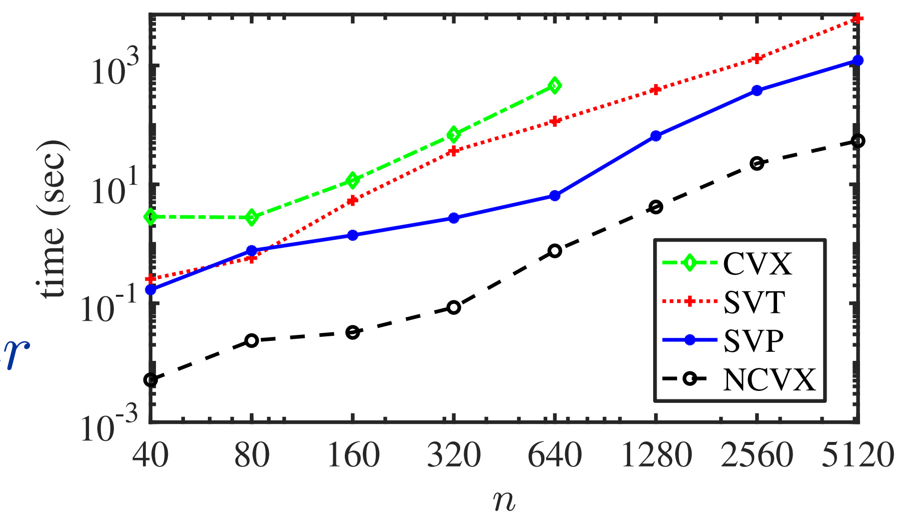
- Critical points of nonconvex function
- Critical points of factorization formulation

[1] **Z.**, Li, Tang & Wakin, Global optimality in low-rank matrix optimization, TSP, 2018.
[2] Li, **Z.**, Tang & Wakin, The non-convex geometry of low-rank matrix optimization, Information and Inference, 2018.

Experiments on Matrix Completion

	Scalability	**Optimality**
Convex formulation	✗	✔
Factorization formulation	✔	✔

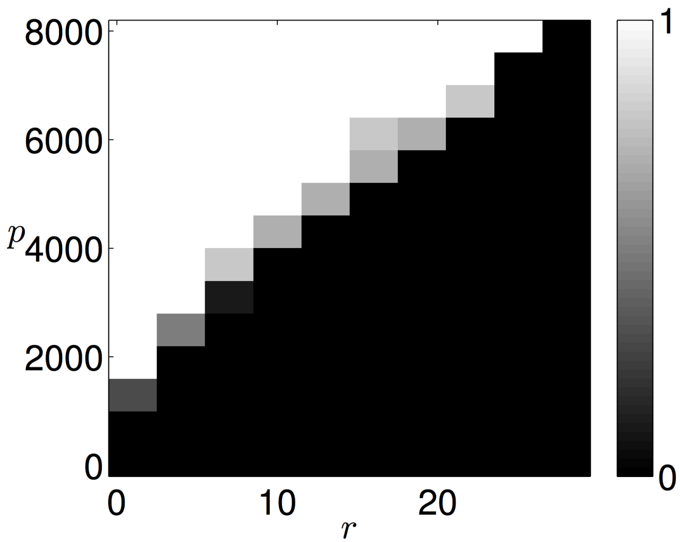
$$n_1 = n_2 = n$$
$$\text{rank} = 5$$
$$\#\text{samples} = 6nr$$



Experiments on Matrix Completion

	Scalability	**Optimality**
Convex formulation	✗	✔
Factorization formulation	✔	✔

$n_1 = n_2 = 100$



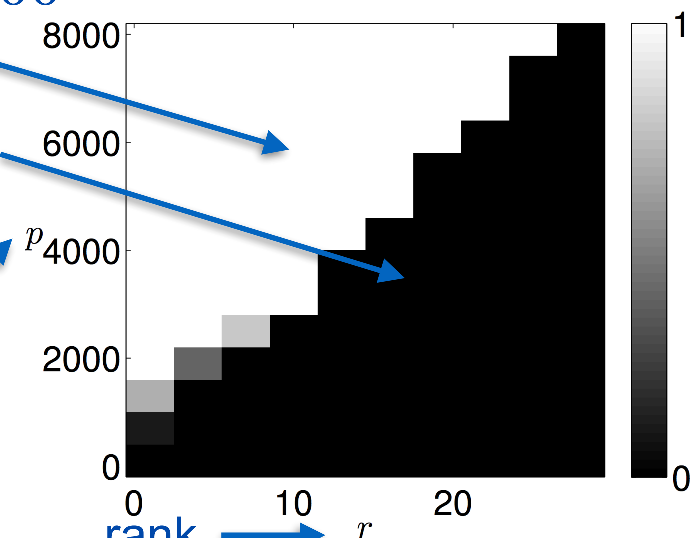
CVX: nuclear norm minimization

success

failure

samples

rank



NCVX: matrix factorization by gradient descent

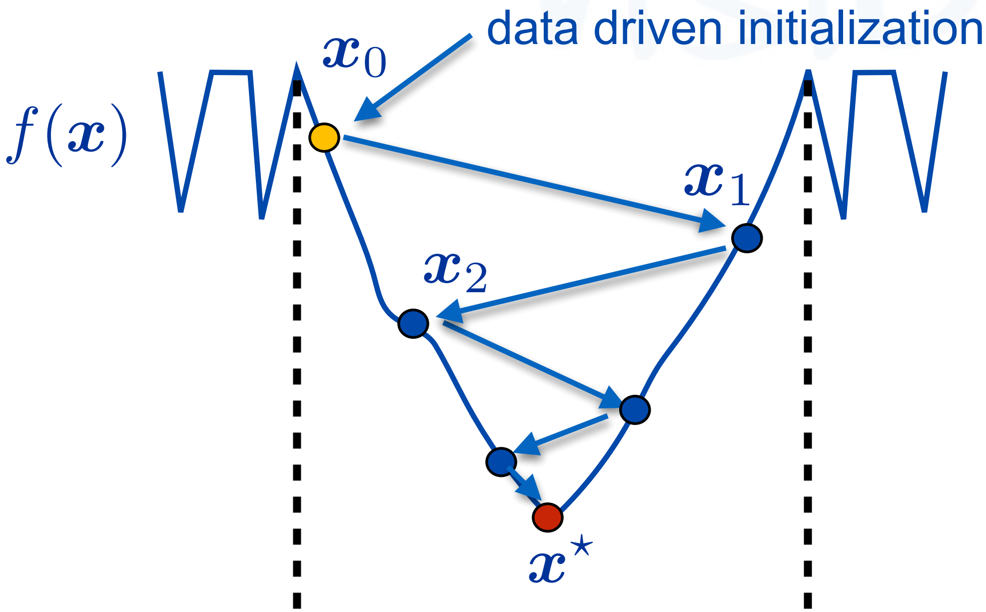
Michael, Boyd & Ye, CVX: Matlab software for disciplined convex programming, 2008.

JHU vision lab

From Global to Local Geometric Analysis: Building Practical Optimization Methods

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

What if Geometry is not Globally Benign?



basin of attraction
(no saddles/spurious local minima)

- Two-stage approach:
  - Initialize within a local basis of attraction sufficiently close to $\boldsymbol{x}^\star$
  - Local refinement (e.g., gradient descent) that exploits certain property in the basis of attraction

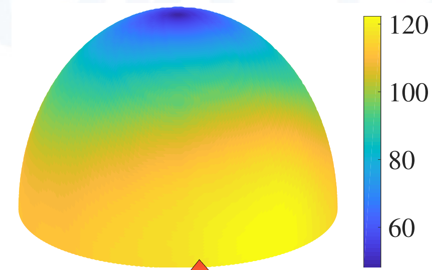
Two-stage Methods

Non-smooth

Riemannian subgradient

Robust subspace learning [Li et al. 2018]

Dictionary learning [Bai et al. , 2019; Zhu et al., 2019]

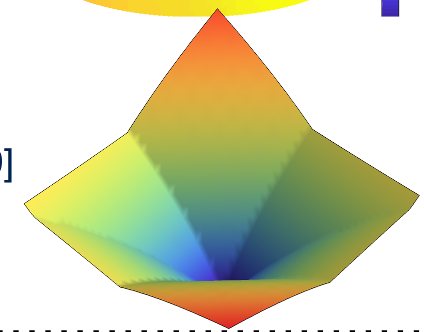


Subgradient method

Overparameterized Neural Network [Du et al., 2019]

Robust low-rank recovery [Zhu et al., 2018]

Robust phase retrieval [Davis et al., 2017]



Smooth

1st/2nd-order methods, alternating min. [Candes et al., 2014, Yue et al., 2018]

Tensor recovery [Jain et al., 2013]

Matrix completion [Sun et al, 2014, Verdun et al., 2019]

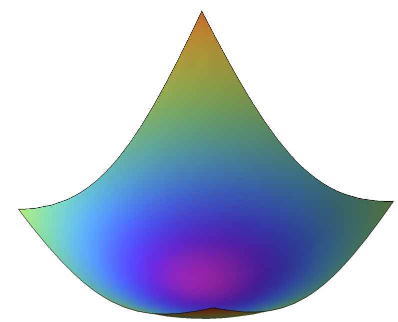
Matrix sensing [Tu et al, 2016]

Blind phase retrieval [Qu et al., 2014]

Blind deconvolution [Li et al., 2016; Qu et al. 2019]

Phase retrieval [Candes et al., 2014]

Dictionary learning [Agarwal et al., 2014]



Outline

- Global geometric analysis for nonconvex optimization
  - Low-rank matrix optimization
- From global to local geometric analysis: building practical optimization methods
  - **Robust low-rank matrix recovery**
- Incremental methods for weakly convex optimization
- Conclusion and future work

[1] Li*, **Z.***, So & Vidal, Nonconvex robust low-rank matrix recovery, 2018

Sparse-Corruption Measurement Model

- Sparse-corruption measurement model:

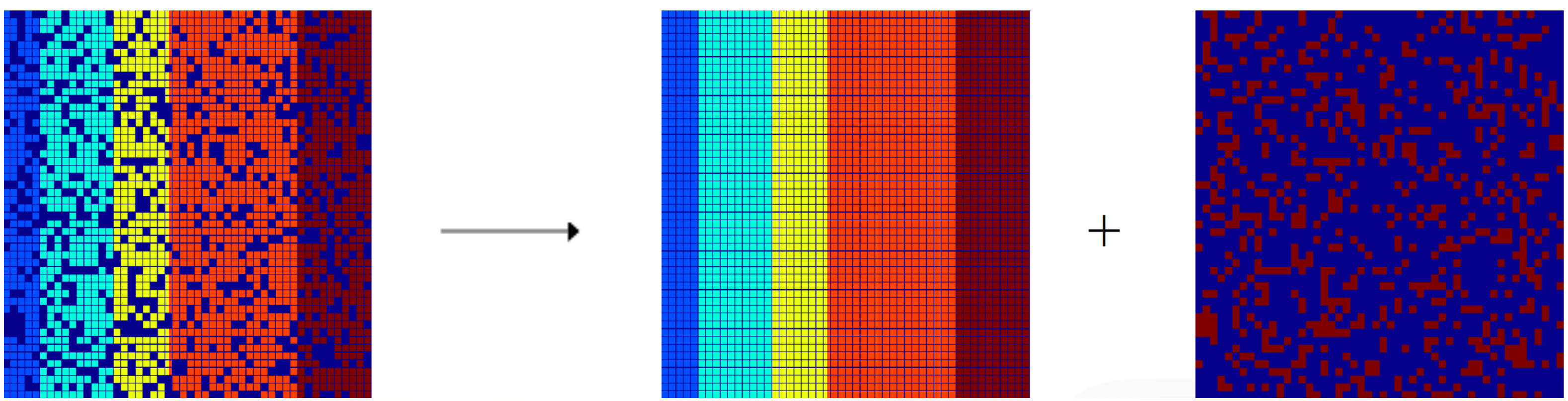
$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star) + \boldsymbol{s}$$

linear operator: $\mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$

sparse corruption

$$p = \frac{\#\{s_i > 0\}}{m}$$

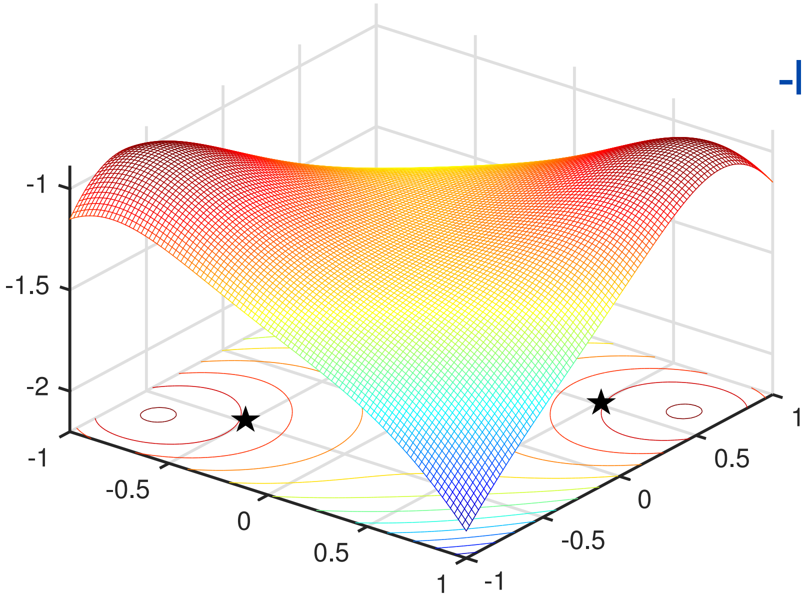
- Outliers are prevalent in the context of (1) gross errors, or (2) applications where the outliers are the important observations
- Robust matrix sensing: $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^\star \rangle + s_i$
- Robust PCA:

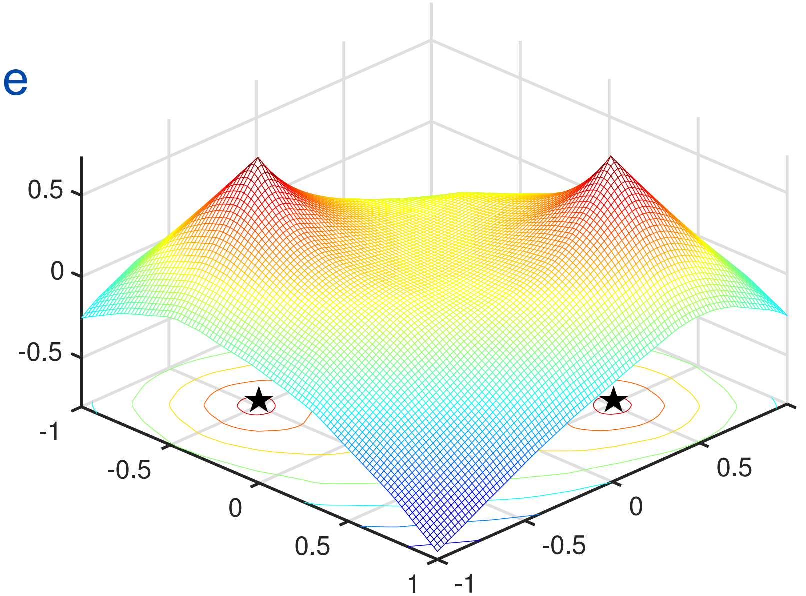


[1] Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.
[2] Chandrasekaran, Sanghavi, Parrilo, Willsky, Rank-sparsity incoherence for matrix decomposition, SIOPT,201
[3] Hosseini, and Uschmajew. A gradient sampling method on algebraic varieties and application to nonsmooth low-rank optimization, 2017.

Robust Low-Rank Matrix Factorization

-log scale





$$\frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top)\|_2^2$$

$$\frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top)\|_1$$

$$\boldsymbol{X}^\star = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}$$

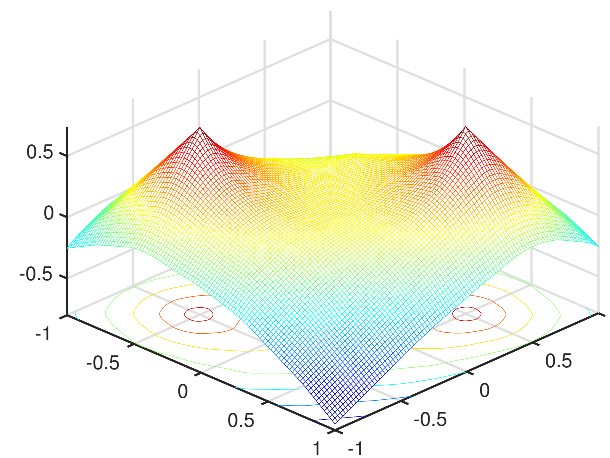
10% sparse corruptions

Contributions

Nonconvex approach

Robust estimator

$$f(\boldsymbol{U}) = \frac{1}{m}\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top) - \boldsymbol{y}\|_1$$



- When the global minimum "is" $\boldsymbol{X}^\star$?
  - The objective function is sharp
- How to efficiently find it?
  - The objective function is weakly convex
  - Weak convexity + sharpness → large basin of attraction
  - Subgradient method converges in a linear rate

[1] Tukey, A survey of sampling from contaminated distributions, 1960.
[2] Huber et al., "Robust regression: Asymptotics, conjectures and monte carlo," Ann. Stat. , 1973.
[3] Duchi and Ruan, Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval, 2018.
[4] Davis, Drusvyatskiy, Paquette, The nonsmooth landscape of phase retrieval, IMAJNA, 2019
[5] Li*, **Z***, So & Vidal, Nonconvex robust low-rank matrix recovery, 2018

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

ℓ_1/ℓ_2-RIP

- We say $\mathcal{A}$ satisfies the ℓ_1/ℓ_2-RIP if there exist c_1, c_2 such that

$$c_1 \|\boldsymbol{X}\|_F \leq \frac{1}{m} \|\mathcal{A}(\boldsymbol{X})\|_1 \leq c_2 \|\boldsymbol{X}\|_F$$

holds for any rank-2r matrix $\boldsymbol{X}$.

- Gaussian measurement ensemble $\{\boldsymbol{A}_i, i = 1, \ldots, m\}$ satisfies with

$$c_1 = \sqrt{\frac{2}{\pi}} - \delta, c_2 = \sqrt{\frac{2}{\pi}} + \delta$$

w.h.p when $m = O(\max(n_1, n_2))$.

- Subgaussian measurement ensemble also satisfies w.h.p.

[1] Chen, Chi & Goldsmith, Exact and stable covariance estimation from quadratic sampling via convex programming, TIT, 2014.

Sharpness

$$f(\boldsymbol{U}) = \frac{1}{m}\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top) - \boldsymbol{y}\|_1$$

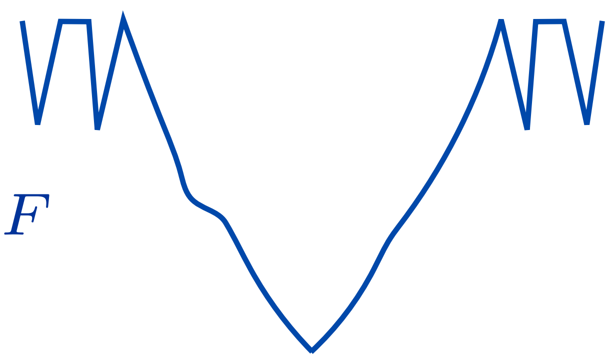
- **Theorem** (sharpness): Suppose $\mathcal{A}$ satisfies the ℓ_1/ℓ_2-RIP. Then for any p<1/2, there exists a>0 (which depends on c_1, c_2, p) s.t.,

$$f(\boldsymbol{U}) - f(\boldsymbol{U}^\star) \geq a\,\mathrm{dist}(\boldsymbol{U}, \boldsymbol{U}^\star)$$

- $\boldsymbol{U}^\star$ is a global minimum of f

$$\mathrm{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) = \min_{\boldsymbol{R}\in\mathbb{O}_r} \|\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{R}\|_F$$

orthogonal Procrustes problem



$\boldsymbol{U}^\star$

[1] Burke, Ferris. Weak Sharp Minima in Mathematical Programming. SIAM J Control and Optimization,1993.
[2] Higham and Papadimitriou. Matrix procrustes problems. Rapport technique, University of Manchester,1995

Weak Convexity

- We say *g* is b-weakly convex if $\boldsymbol{x} \mapsto g(\boldsymbol{x}) + \frac{b}{2}\|\boldsymbol{x}\|_2^2$ is convex
- Any gradient Lipschitz smooth function is weakly convex
- Any composite function $h(F(x))$ is weakly convex

convex and Lipschitz

smooth map with Lipschitz gradient

- f is $2c_2$-weakly convex

$$f(\boldsymbol{U}) = \frac{1}{m}\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top) - \boldsymbol{y}\|_1$$

[1] Vial. Strong and weak convexity of sets and functions. Mathematics of Operations Research, 1983.

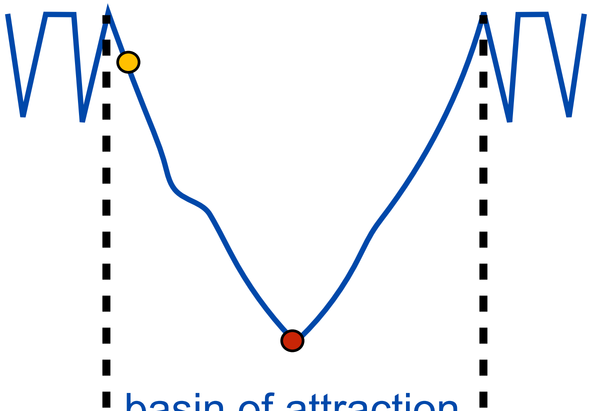
Consequence of Sharpness and Weak Convexity

- b-weakly convex: $\forall \boldsymbol{D} \in \partial f(\boldsymbol{U})$

$$f(\boldsymbol{U}^\star) - f(\boldsymbol{U}) \geq \langle \boldsymbol{D}, \boldsymbol{U}^\star - \boldsymbol{U} \rangle - \frac{b}{2} \|\boldsymbol{U} - \boldsymbol{U}^\star\|_F^2$$

- a-sharp:

$$f(\boldsymbol{U}) - f(\boldsymbol{U}^\star) \geq a \operatorname{dist}(\boldsymbol{U}, \boldsymbol{U}^\star)$$



Regularity condition

$$\langle \boldsymbol{D}, \boldsymbol{U} - \boldsymbol{U}^\star \rangle \geq a \operatorname{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) \underbrace{- \frac{b}{2} \operatorname{dist}^2(\boldsymbol{U}, \boldsymbol{U}^\star)}_{\text{Positive when dist} \leq 2a/b}$$

[1] Davis, Drusvyatskiy, MacPhee & Paquette, Subgradient methods for sharp weakly convex functions, JOTA, 2018.
[2] Li*, **Z.***, So & Vidal, Nonconvex robust low-rank matrix recovery, 2018

SubGradient Method (SGM)

- **Geometrically diminishing step size**:

$$\mu_k = \mu_0 \beta^k, \ \beta < 1$$

- **SGM update**:

$$\boldsymbol{U}_{k+1} = \boldsymbol{U}_k - \mu_k \boldsymbol{D}_k, \ \boldsymbol{D}_k \in \partial f(\boldsymbol{U}_k)$$

- **Theorem**: If $\operatorname{dist}(\boldsymbol{U}_0, \boldsymbol{U}^\star) \le \frac{2a}{b}$, SGM converges in a linear rate

$$\operatorname{dist}(\boldsymbol{U}_k, \boldsymbol{U}^\star) \lesssim \beta^k$$

- A truncated spectral initialization satisfies this w.h.p.

[1] Li*, **Z.***, So & Vidal, Nonconvex robust low-rank matrix recovery, 2018
[2] Davis, Drusvyatskiy, MacPhee & Paquette, Subgradient methods for sharp weakly convex functions, JOTA, 2018.
[3] Li, Chi, Zhang & Liang, Nonconvex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent, 2017.

Extension to General Case

$$f(\boldsymbol{U},\boldsymbol{V}) = \|\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\top) - \boldsymbol{y}\|_1 + \lambda\|\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\|_F$$

- The regularizer is to avoid scaling ambiguity $(t\boldsymbol{U})(t^{-1}\boldsymbol{V}^\top) = \boldsymbol{U}\boldsymbol{V}^\top$
- Using $\|\cdot\|_F$ instead of $\|\cdot\|_F^2$ is to preserve sharpness

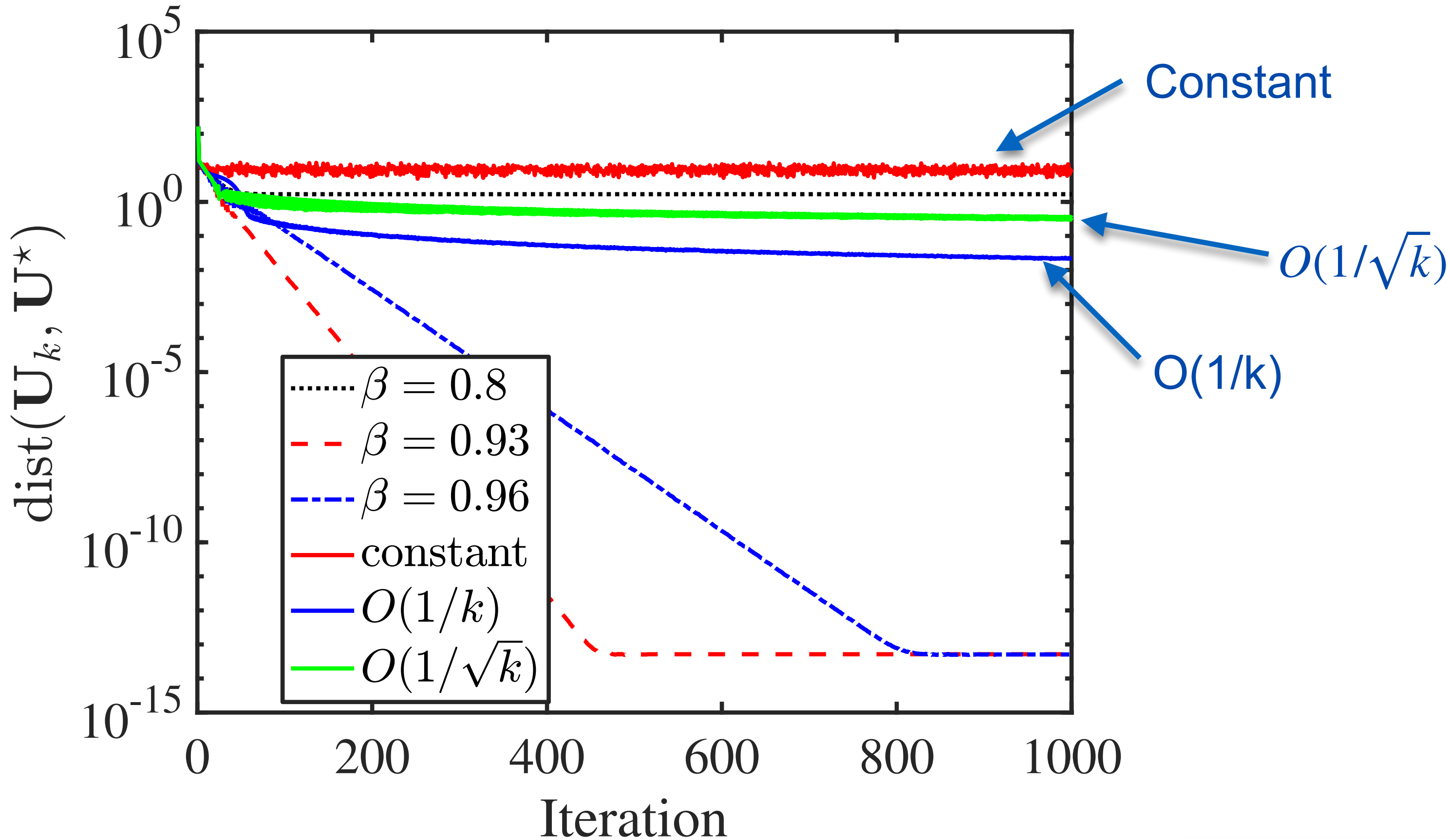
- **Theorem** (informal): Suppose $\mathcal{A}$ satisfies the ℓ_1/ℓ_2-RIP. If the ratio of outliers is less than a half, then f is sharp and weakly convex, and subgradient method finds the target solution in a linear rate.

[1] Tu et al., Low-rank solutions of linear matrix equations via procrustes flow, ICML, 2015.
[2] Bhojanapalli, Neyshabur & Srebro, Global optimality of local search for low rank matrix recovery, NIPS, 2016
[3] Li*, **Z.***, So & Vidal, Nonconvex robust low-rank matrix recovery, 2018

Experiments on Robust Matrix Sensing

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^\star \rangle + s_i \longrightarrow \text{nonzero } \mathcal{N}(0,100)$$

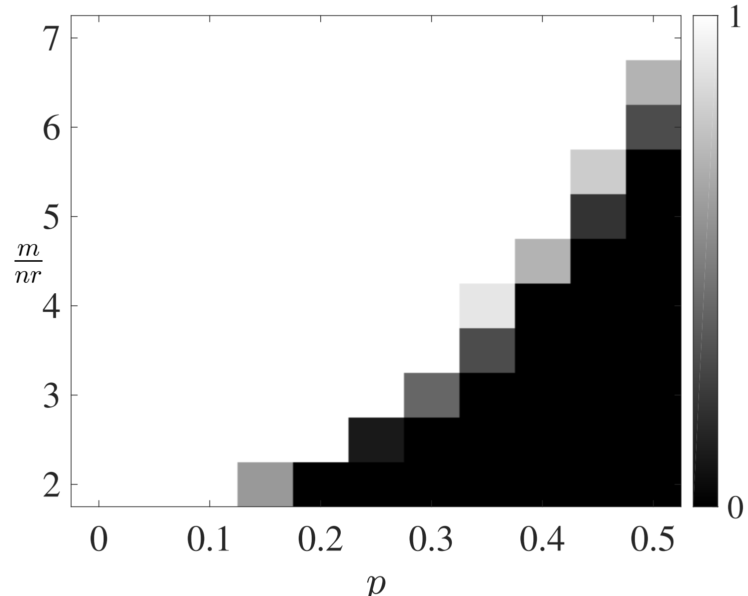
n= 50, r = 7 m= 5nr p=30% sparse corruptions



Experiments on Robust Matrix Sensing

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^\star \rangle + s_i \longrightarrow \text{nonzero } \mathcal{N}(0,100)$$

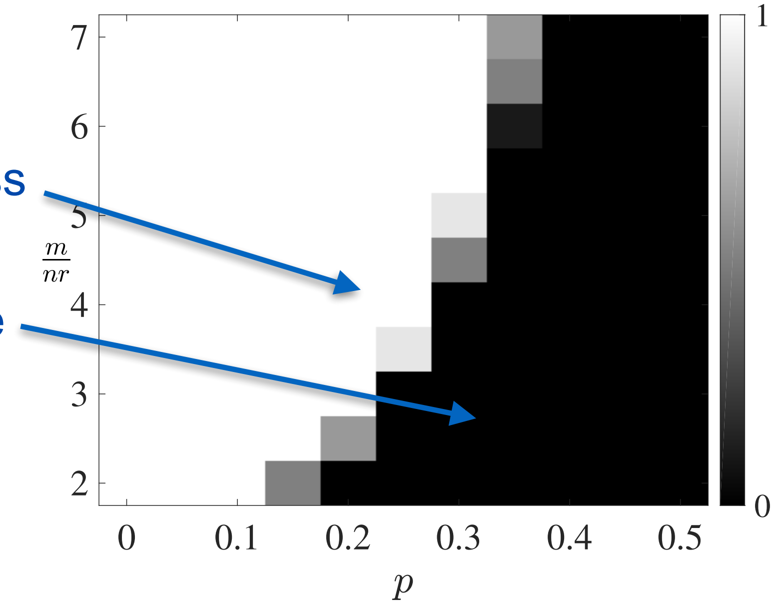
n= 50, r = 7



SubGradient Method

success

failure



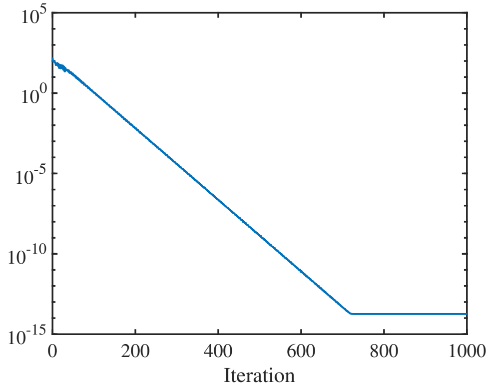
Median-truncated gradient descent

[1] Li, Chi, Zhang & Liang, Nonconvex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent, 2017.

Experiments on Robust PCA

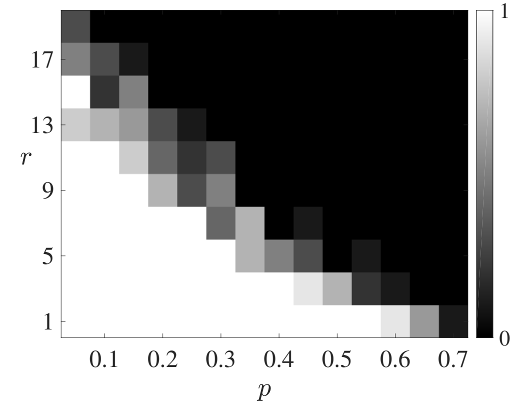
$$\boldsymbol{Y} = \boldsymbol{X}^{\star} + \boldsymbol{S}$$

dist

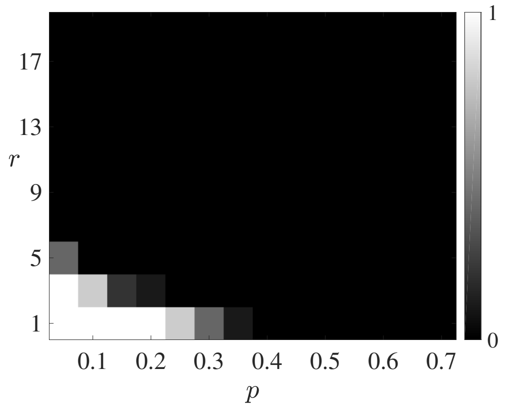


r= 10
p= 0.1

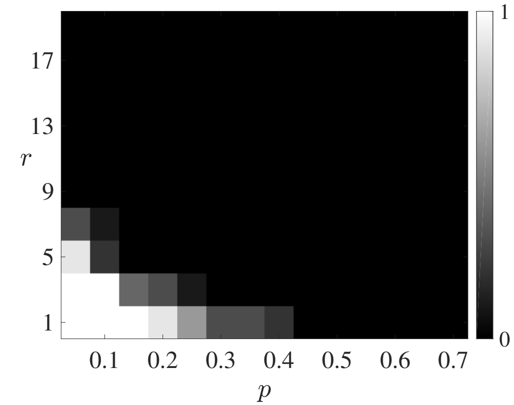
n= 50



SubGradient Method



ALM [Lin et al. 2010]



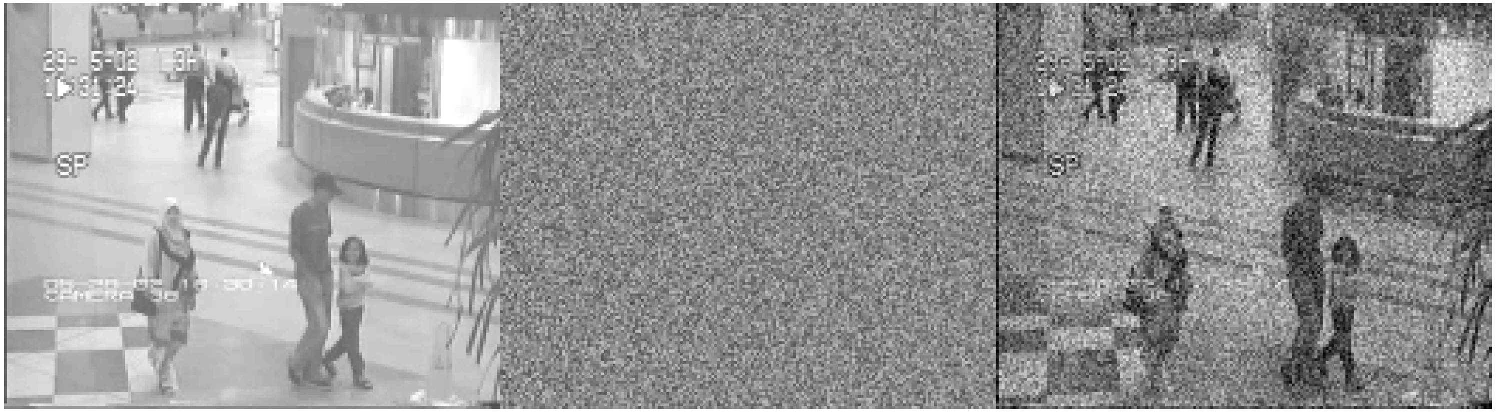
PPGD [Chen et al. 2010]

[1] Lin, Chen, & Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, 2010.
[2]Chen, Ganesh, Lin, Ma, Wright & Wu, Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix, 2009

Experiments on Background Subtraction

original frame

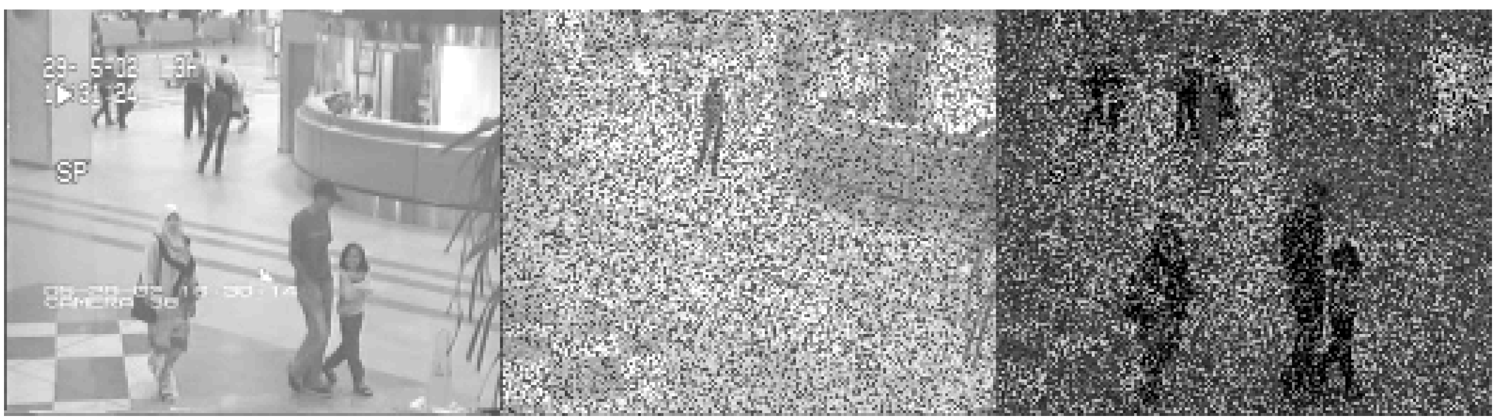
initialization



Experiments on Background Subtraction

original frame

10^{th} iteration



Experiments on Background Subtraction

original frame

20th iteration

Experiments on Background Subtraction

original frame

30th iteration

Experiments on Background Subtraction

original frame

40th iteration

Experiments on Background Subtraction

original frame

50th iteration

Incremental Methods for Weakly Convex Optimization

[1] Li, **Z.**, So, and Lee, Incremental Methods for Weakly Convex Optimization, 2019.

Incremental Methods for Weakly Convex Optimization

- Consider $\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$ $\qquad f(U) = \frac{1}{m} \|\mathcal{A}(UU^\top) - y\|_1$

  Weakly convex but could be nondifferentiable

- Widely appeared in ML&SP:
  - m could be huge: computing subgradients for all i is time consuming
  - A practical approach is to compute a subgradient for one i at each time

Stochastic method

1	2	3	……	m

$\uparrow i$

Incremental method

$i \longrightarrow$

1	2	3	……	m

Random shuffling

$i \longrightarrow$

3	m	1	……	4

- Easy to analyze: $\mathbb{E}[\partial f_i(x)] = \partial f(x)$
- Not often used in practice

TensorFlow

- Hard to analyze
- Widely used in practice

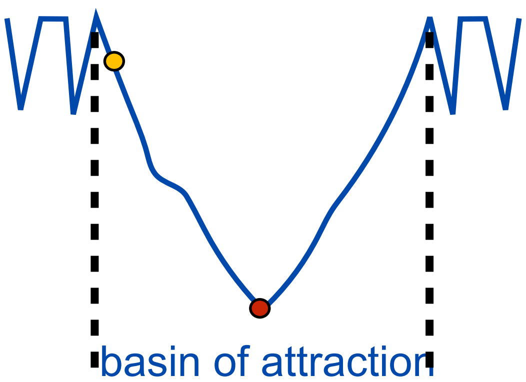
[1] Nedic and Bertsekas, Convergence rate of incremental subgradient algorithms, 2001.
[2] Bottou, Stochastic gradient descent tricks, in Neural networks: Tricks of the trade, 2012.
[3] Davis and Drusvyatskiy, Stochastic model-based minimization of weakly convex functions, SIOPT 2018.
[4] Duchi and Ruan, Stochastic methods for composite and weakly convex optimization problems, 2018
[5] Haochen and Sra, Random shuffling beats SGD after finite epochs, ICML 2019.

Incremental Methods for Weakly Convex Optimization

- Consider $\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$ $\quad f(U) = \frac{1}{m} \|\mathcal{A}(UU^\top) - y\|_1$

- **Theorem** (informal). Suppose each f_i is weakly convex and Lipschitz. Then incremental subgradient/proximal point/prxo-linear methods and their random shuffling versions all converge to a critical point at rate $\mathrm{O}(k^{-1/4})$.
  - Has the same rate as the stochastic one, (but with slightly worse bound on m)

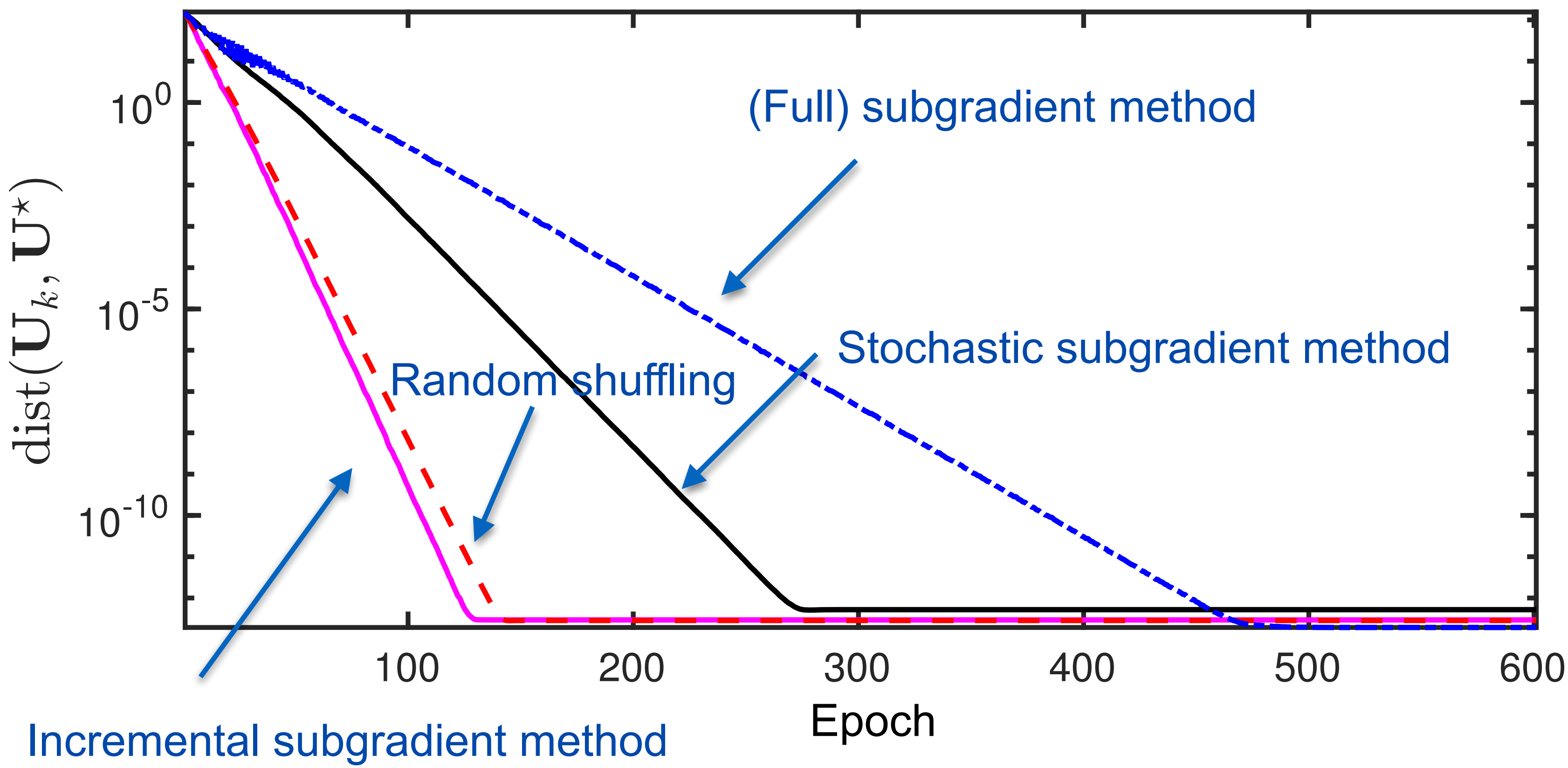
- **Theorem** (informal). If the function f is also sharp, then with a good initialization these incremental methods converge to a global minimum in a linear rate.
  - Iterates always within the basin of attraction
  - Has the same rate as the stochastic one (which only holds w.h.p)



[1] Li, **Z.**, So, and Lee, Incremental Methods for Weakly Convex Optimization, 2019.
[2] Davis and Drusvyatskiy, Stochastic model-based minimization of weakly convex functions, SIOPT 2018.
[3] Davis, Drusvyatskiy, and Charisopoulos, Stochastic algorithms with geometric step decay converge linearly on sharp functions, 2019.

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Experiments on Robust Low-Rank Matrix Recovery

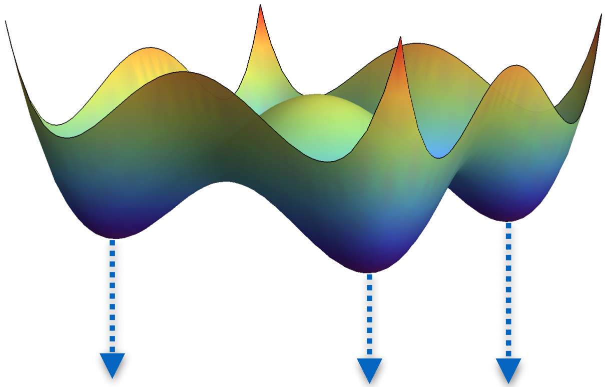


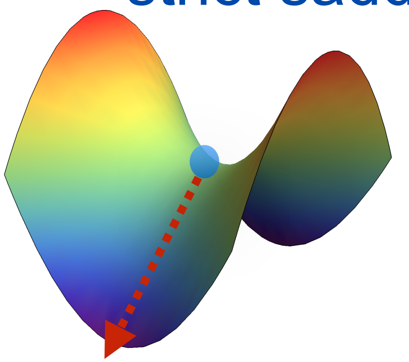
[1] Li, **Z.**, So, and Lee, Incremental Methods for Weakly Convex Optimization, 2019.

Conclusion and Future Work

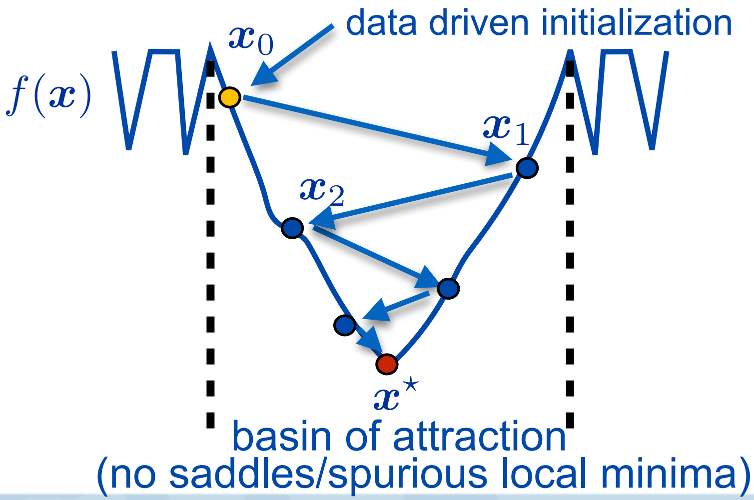
Recap: Taming Nonconvexity by Geometric Analysis

- Benign global geometric property





- Fast optimization methods based on local geometric property



Optimality Guarantee for Nonconvex Problems

Non-smooth

Riemannian subgradient
robust subspace learning [Z. et al., 2018]
dictionary learning
[Bai et al. , 2019; Z. et al., 2019]

Subgradient method
overparameterized NN [Du et al., 2019]
robust low-rank recovery[Li et al., 2018]
robust phase retrieval [Davis et al., 2017]

Smooth

1st/2nd-order methods [Lee et al., 2017]
Alternating minimization [Li et al. 2019]
Shallow linear neural network (NN)
[Kawaguchi 2016; Zhu et al., 2018]
Low-rank matrix opt [Z. et al., 2017]
Matrix completion [Ma et al., 2016]
Dictionary learning [Sun et al., 2016]
Phase retrieval [Sun et al., 2016]

first-order methods [Candes et al., 2014]
2nd-order methods [Yue et al, 2018]
Tensor recovery [Jain et al., 2013]
Matrix completion [Sun et al, 2014]
Matrix sensing [Tu et al, 2016]
Blind phase retrieval [Qu et al., 2014]
Blind deconvolution [Li et al., 2016; Qu et al. 2019]
Phase retrieval [Candes et al., 2014]

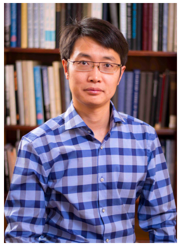
Global geometry
(arbitrary/random initialization)

Local geometry
(data-driven initialization)

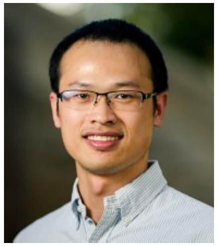
See http://sunju.org/research/nonconvex/ for a detailed list of references

Thanks to

M. Wakin
CSM



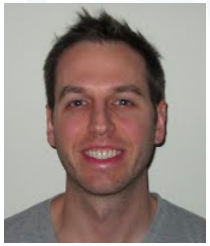
G. Tang
CSM



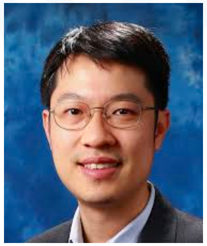
Q. Li
CSM

M. Tsakiris
ShanghaiTech

R. Vidal
JHU



D. Robinson
JHU



A. So
CHUK

X. Li
CHUK

J. Lee
Princeton

Credit: https://sites.google.com/site/nicolasgillis/projects/team-members

and **You!**