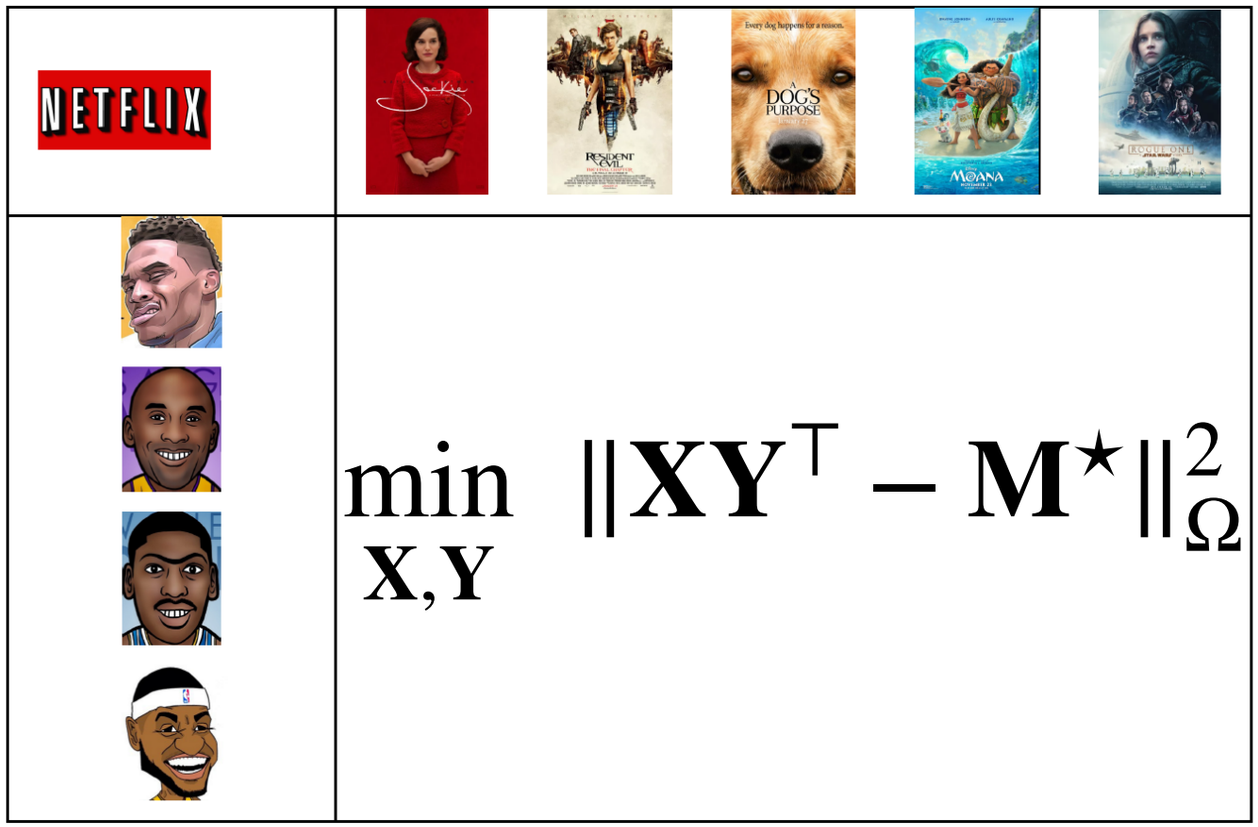
Alternating Minimizations Converge to Second-order Optimal Solutions

Qiuwei Li[1]

Joint work with Zhihui Zhu[2] and Gongguo Tang[1]

[1] Colorado School of Mines
[2] Johns Hopkins University

Why is Alternating Minimization so popular?



$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \; f(\mathbf{x}, \mathbf{y})$$

Many optimization problems have variables with natural partitions

Nonnegative MF	Matrix sensing/completion	Games	Dictionary learning
Blind deconvolution	Tensor decomposition	EM algorithm	.....

Why is Alternating Minimization so popular?

$$\mathbf{y}_k = \operatorname{argmin}_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y})$$

$$\mathbf{x}_k = \operatorname{argmin}_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_k)$$

Advantages

- Simple to implement : No stepsize tuning
- Good empirical performance

Disadvantages

- No global optimality guarantee for general problems
- Only exists 1st-order convergence

Our Approach

Provide the 2nd-order convergence to partially solve the issue of “no global optimality guarantee”.

Why is Alternating Minimization so popular?

$$\mathbf{y}_k = \operatorname{argmin}_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y})$$

$$\mathbf{x}_k = \operatorname{argmin}_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_k)$$

Advantages

- Simple to implement : No stepsize tuning
- Good empirical performance

Disadvantages

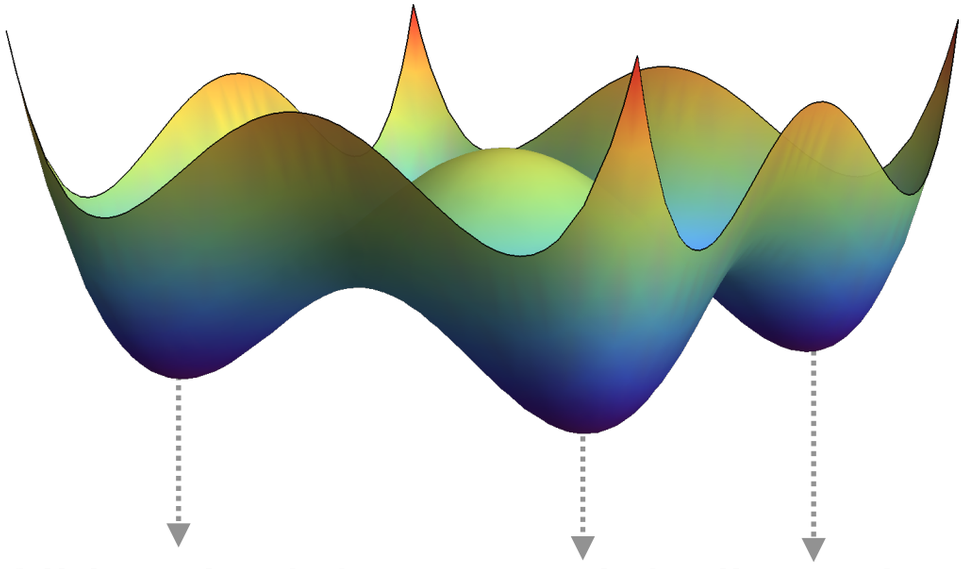
- No global optimality guarantee for general problems
- Only exists 1st-order convergence

Theorem 1

Assume f is strongly bi-convex with a full-rank cross Hessian at all strict saddles. Then **AltMin** almost surely converges to a **2nd-order stationary point** from random initialization.

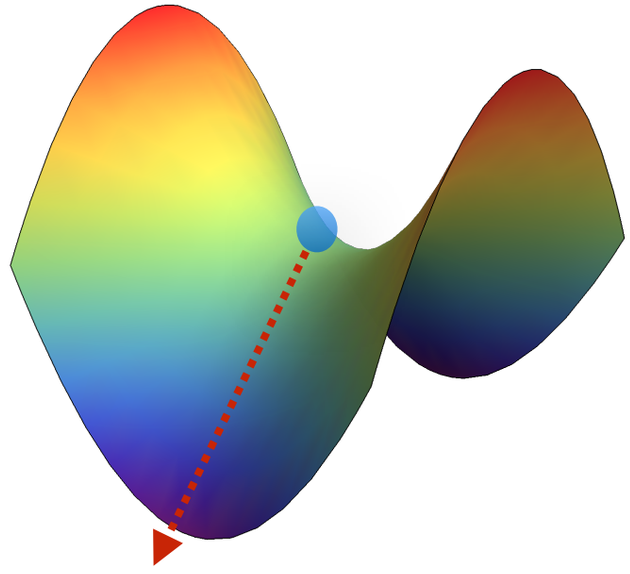
Why second-order convergence is enough?

No spurious local minima



All local minima are globally optimal

All saddles are strict



Negative curvature

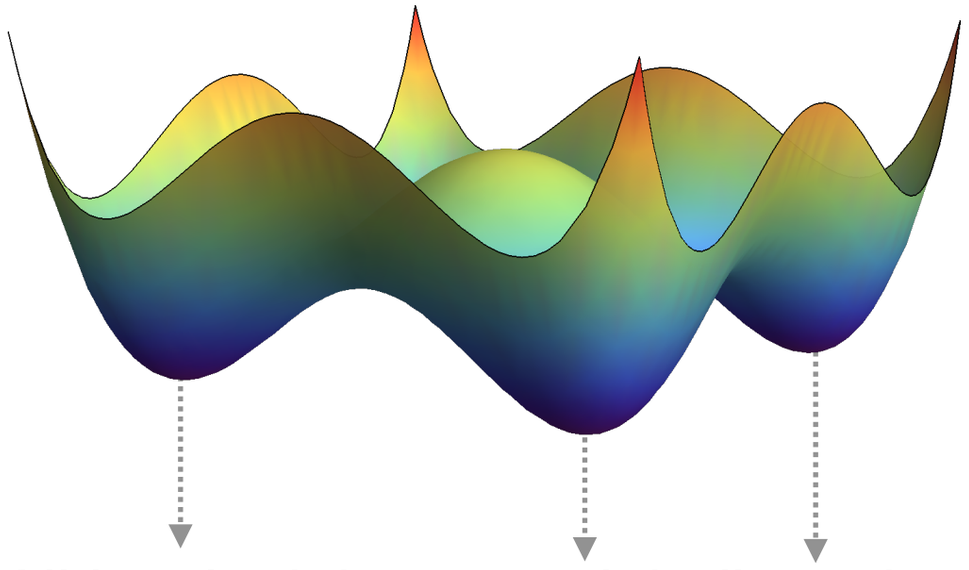
2nd-order optimal solution = globally optimal solution

Matrix factorization [1]	Matrix sensing [2]	Matrix completion [3]
Dictionary learning [4]	Blind deconvolution [5]	Tensor decomposition [6]

[1] Jain et al. Global Convergence of Non-Convex Gradient Descent for Computing Matrix Squareroot
[2] Bhojanapalli et al. Global Optimality of Local Search for Low Rank Matrix Recovery
[3] Ge et al. Matrix Completion Has No Spurious Local Minimum
[4] Sun et al. Complete Dictionary Recovery over The Sphere
[5] Zhang et al. On the Global Geometry of Sphere-Constrained Sparse Blind Deconvolution
[6] Ge et al. Online Stochastic Gradient for Tensor Decomposition

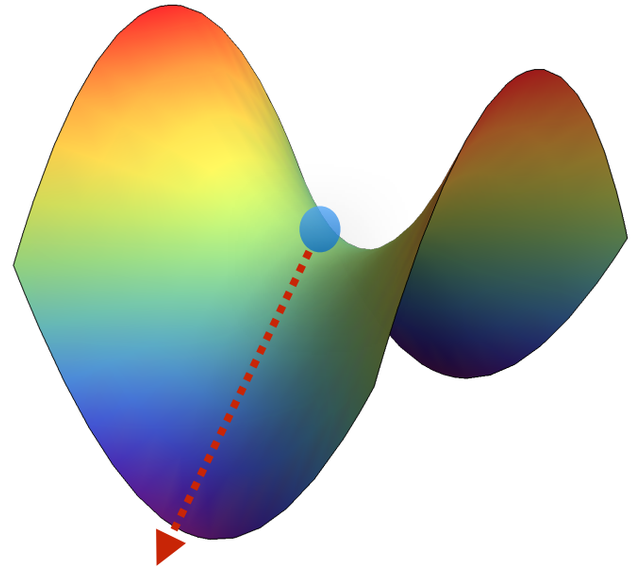
Why second-order convergence is enough?

No spurious local minima



All local minima are globally optimal

All saddles are strict



Negative curvature

2nd-order optimal solution = globally optimal solution

Matrix factorization [1]
Matrix sensing [2]
Matrix completion [3]
Dictionary learning [4]
Blind deconvolution [5]
Tensor decomposition [6]

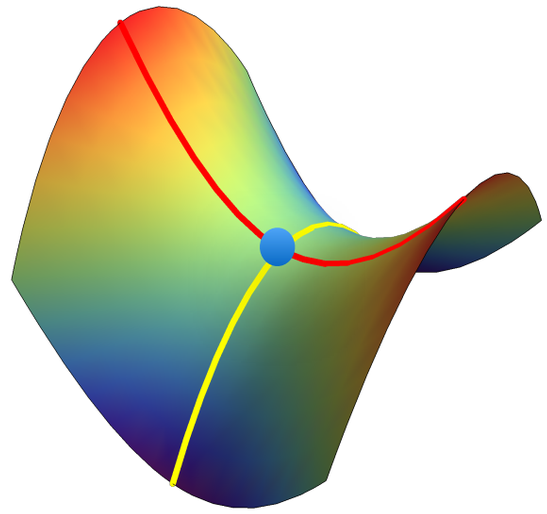
1st-order convergence + avoid strict saddles = 2nd-order convergence

It suffices to show alternating minimization **avoids strict saddles!**

How to show avoiding strict saddles?

A Key Result

Lee et al [1,2] use **Stable Manifold Theorem [3]** to show that iterations defined by a **global diffeom** avoids **unstable fixed points.**



An Improved Version (Zero-Property Theorem [4] + Max-Rank Theorem [5])

This work relaxes the global diffeom condition to show that a **local diffeom (at all unstable fixed points)** can avoid **unstable fixed points.**

General Recipe

(1) Construct algorithm mapping g and show it is a local diffeom (i.e., Show Dg is nonsingular);

(2) Show all strict saddles of f are unstable fixed points of g;

[1] Lee et al. Gradient Descent Converges to Minimizers.
[2] Lee et al. First-order Methods Almost Always Avoid Saddle Points
[3] Shub. Global Stability of Dynamical Systems
[4] Ponomarev et al. Submersions and Preimages of Sets of Measure Zero
[5] Bamber and Van. How Many Parameters Can A Model Have and still Be Testable

A Proof Sketch

Construct the mapping

$$\begin{cases} \mathbf{y}_k = \phi(\mathbf{x}_{k-1}) = \operatorname{argmin}_{\mathbf{y}} \ f(\mathbf{x}_{k-1}, \mathbf{y}) \\ \mathbf{x}_k = \psi(\mathbf{y}_k) = \operatorname{argmin}_{\mathbf{x}} \ f(\mathbf{x}, \mathbf{y}_k) \end{cases} \implies \mathbf{x}_k = g(\mathbf{x}_{k-1}) \doteq \psi(\phi(\mathbf{x}_{k-1}))$$

Compute the Jacobian (use Implicit function theorem and chain rule)

$$Dg(\mathbf{x}^\star) \sim \underbrace{\left(\nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{-\frac{1}{2}} \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{-\frac{1}{2}} \right) \left(\nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{-\frac{1}{2}} \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{-\frac{1}{2}} \right)^\top}_{\mathbf{L}\mathbf{L}^\top}$$

Show all strict saddles are "unstable" (Connect Dg with "Schur complement" of the Hessian)

$$\nabla^2 f(\mathbf{x}^\star, \mathbf{y}^\star) = \begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{1/2} & \\ & \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{1/2} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{I}_n & \mathbf{L} \\ \mathbf{L}^\top & \mathbf{I}_m \end{bmatrix}}_{\mathbf{\Phi}} \begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{1/2} & \\ & \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{1/2} \end{bmatrix}$$

Finally, by using a Schur complement theorem:

$$\nabla^2 f(\mathbf{x}^\star, \mathbf{y}^\star) \not\succeq 0 \iff \mathbf{\Phi} \not\succeq 0 \iff \mathbf{\Phi/I} \doteq \mathbf{I} - \mathbf{L}\mathbf{L}^\top \not\succeq 0 \iff \|\mathbf{L}\| > 1 \iff \rho(Dg(\mathbf{x}^\star)) > 1. \square$$

Proximal Alternating Minimization

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \;\; f(\mathbf{x}, \mathbf{y})$$

Proximal Alternating Minimization

$$\mathbf{x}_k = \text{argmin}_{\mathbf{x}} \; f(\mathbf{x}, \mathbf{y}_{k-1}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|_2^2$$

$$\mathbf{y}_k = \text{argmin}_{\mathbf{y}} \; f(\mathbf{x}_k, \mathbf{y}) + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{y}_{k-1}\|_2^2$$

Key Assumption (Lipschitz bi-smoothness)

$$\max\{\|\nabla_{\mathbf{x}}^2 f(\mathbf{x}, \mathbf{y})\|, \|\nabla_{\mathbf{y}}^2 f(\mathbf{x}, \mathbf{y})\|\} \leq L, \; \forall \mathbf{x}, \mathbf{y}$$

Theorem 2

Assume f is L-Lipschitz bi-smooth and $\lambda > L$. Then Proximal AltMin almost surely converges to a **2nd-order stationary point** from random initialization.

110