Analysis of global convergence: random initialization

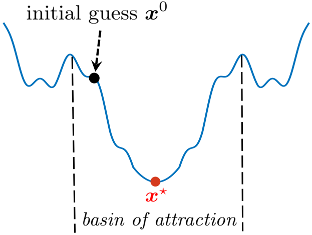
Cong Ma

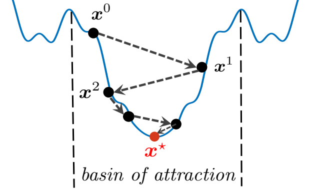
University of Chicago, Winter 2024

Outline

- Strict saddle property
- Global landscape analysis: matrix sensing
- Gradient descent with random initialization: phase retrieval
- Generic saddle-escaping algorithms

Rationale of two-stage approach



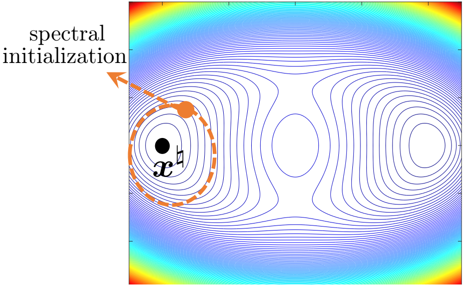


1. initialize within $\underbrace{\text{local basin sufficiently close to } \boldsymbol{x}^\natural}_{\text{(restricted) strongly convex}}$

2. iterative refinement

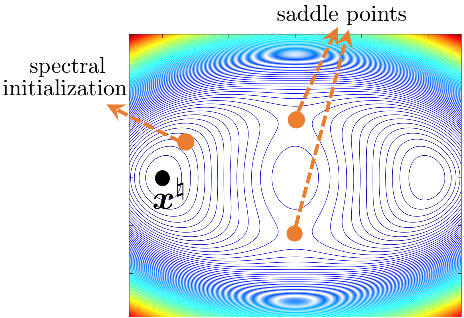
Is careful initialization necessary for fast convergence?

Initialization



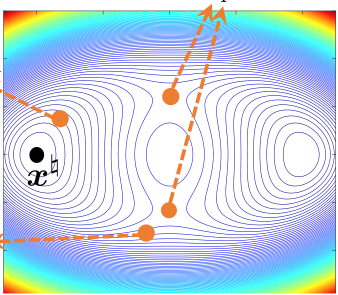
- spectral initialization gets us to (restricted) strongly cvx region

Initialization



- spectral initialization gets us to (restricted) strongly cvx region
- cannot initialize GD anywhere, e.g., might get stuck at saddles

Initialization



- spectral initialization gets us to (restricted) strongly cvx region
- cannot initialize GD anywhere, e.g., might get stuck at saddles

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

Generic saddle-escaping algorithms

Strict saddle property: qualitative version

All critical points can be classified into two categories

- local minimizers
- strict saddle points: Hessian has a strictly negative eigenvalue

Let $\boldsymbol{x}$ be a critical point. Taylor expansion yields

$$f(\boldsymbol{x}+\boldsymbol{\Delta}) \approx f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\Delta}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{\Delta}$$

GD converges to local minimizers

Theorem 10.1

Consider any twice continuously differentiable function f that satisfies the strict saddle property. If $\eta < 1/\beta$ with β the smoothness parameter, then GD with a random initialization converges to a local minimizer or $-\infty$ almost surely.

- This also holds for other optimization algorithms
- Exponential time for GD to converge in the worst case

An example: low-rank matrix sensing

Low-rank matrix sensing

- Groundtruth: rank-r psd matrix $\boldsymbol{M}^\star = \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \in \mathbb{R}^{n\times n}$
- Observations:

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{M}^\star \rangle, \qquad \text{for } 1 \le i \le m$$

- Goal: recover $\boldsymbol{M}^\star$ based on linear measurements $\{\boldsymbol{A}_i, y_i\}_{1\le i\le m}$

Restricted isometry property (RIP)

Define linear operator $\mathcal{A}: \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^m$ to be

$$\mathcal{A}(\boldsymbol{M}) = [m^{-1/2} \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle]_{1 \le i \le m}$$

Definition 10.2

The operator $\mathcal{A}$ is said to satisfy r-RIP with RIP constant $\delta_r < 1$ if

$$(1 - \delta_r) \|\boldsymbol{M}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{M})\|_2^2 \le (1 + \delta_r) \|\boldsymbol{M}\|_{\mathsf{F}}^2$$

holds simultaneously for all $\boldsymbol{M}$ of rank at most r.

An optimization-based method

Then least-squares estimation yields

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}} \qquad f(\boldsymbol{X}) = \frac{1}{4m}\sum_{i=1}^{m}\Big(\langle \boldsymbol{A}_i, \boldsymbol{X}\boldsymbol{X}^\top\rangle - y_i\Big)^2$$

Global landscape

Theorem 10.3

Assume that the measurement operator $\mathcal{A}$ satisfies $2r$-RIP with RIP constant $\delta_{2r} \leq 1/10$. Then for the matrix sensing objective, one has

- *For any critical point $\boldsymbol{U}$ that is not a local minimum, one has $\lambda_{\min}(\nabla^2 f(\boldsymbol{U})) \leq -2/5\sigma_r(\boldsymbol{M}^\star)$;*
- *All local minimizers are global.*

- Matrix sensing obeys strict saddle property
- In addition, all local minimizers are global — GD converges to global minimizer

Strict saddle property: quantitative version

Definition 10.4

A function $f(\cdot)$ is said to satisfy the $(\varepsilon, \gamma, \xi)$-strict saddle property for some positive ε, γ, ξ, if for each $\boldsymbol{x}$, at least one of the following is true

- (**strong gradient**) $\|\nabla f(x)\|_2 \geq \varepsilon$;
- (**negative curvature**) $\lambda_{\min}(\nabla^2 f(\boldsymbol{x})) \leq -\gamma$;
- (**local minimum**) there exists a local minimum $\boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 \leq \xi$.

Saddle-escaping algorithms

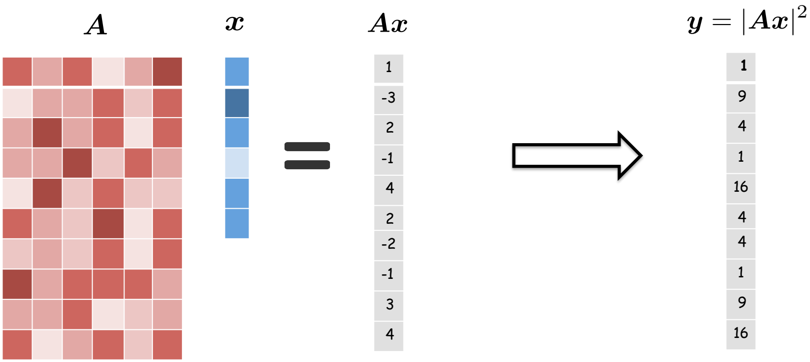
$$f(\boldsymbol{x}+\boldsymbol{\Delta}) \approx f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\Delta}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{\Delta}$$

A rough categorization:

- Hessian-based algorithms
- Gradient-based algorithms

Another example: phase retrieval

Solving quadratic systems of equations



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

$$y_k = \left(\boldsymbol{a}_k^\top \boldsymbol{x}^\natural\right)^2 + \text{noise}, \qquad k = 1, \ldots, m$$

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

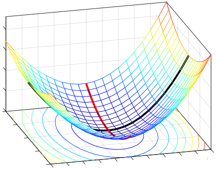
A natural least-squares formulation

given: $\quad y_k \;=\; (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \quad 1 \le k \le m$

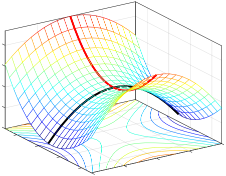
$$\Downarrow$$

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

What does prior theory say?



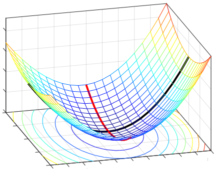
global minimum



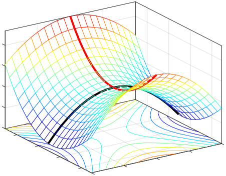
saddle point

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

What does prior theory say?



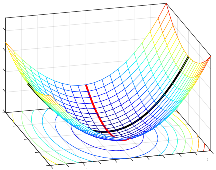
global minimum



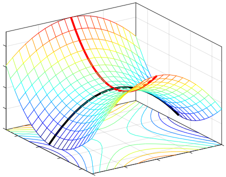
saddle point

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

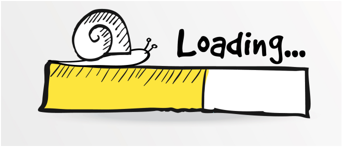
What does prior theory say?



global minimum



saddle point

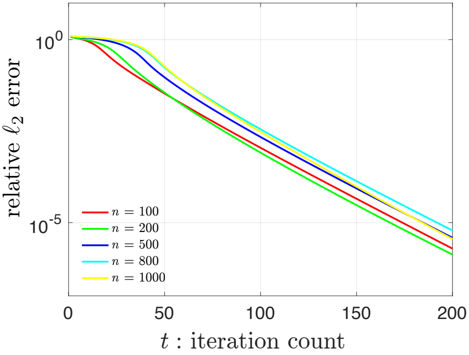


- **landscape:** no spurious local mins (Sun et al. '16)
- randomly initialized GD converges almost surely (Lee et al. '16)

"almost surely" might mean "takes forever"

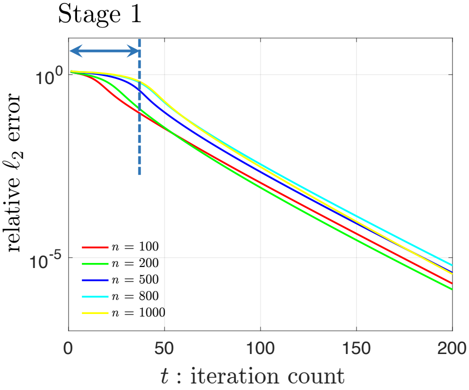
Numerical efficiency of randomly initialized GD

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



Numerical efficiency of randomly initialized GD

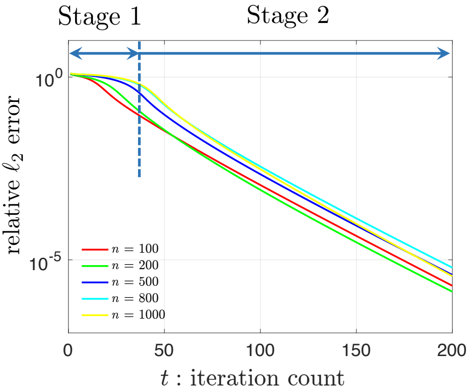
$$\eta = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



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

Numerical efficiency of randomly initialized GD

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



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

Our theory: noiseless case

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

Our theory: noiseless case

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

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

Theorem 10.5 (Chen, Chi, Fan, Ma '18)

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

Our theory: noiseless case

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

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

Theorem 10.5 (Chen, Chi, Fan, Ma '18)

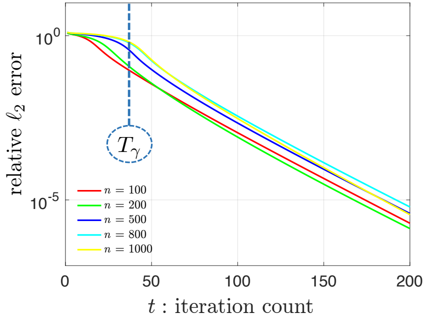
Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$ achieves

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

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

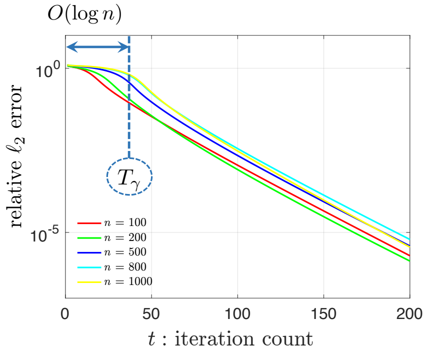
Our theory

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$



Our theory

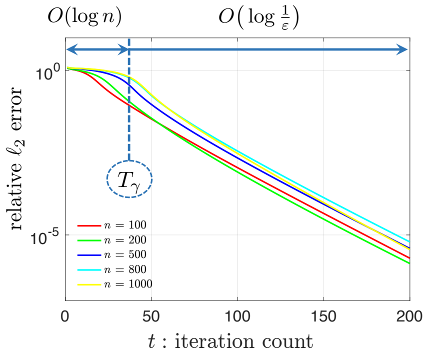
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \ge T_\gamma \asymp \log n$$



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

Our theory

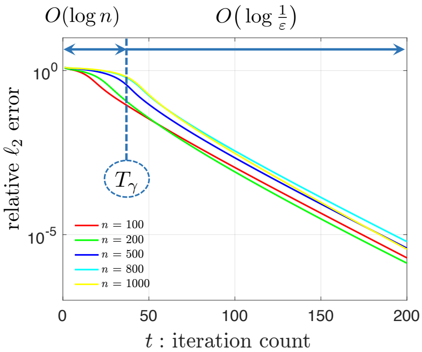
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \ge T_\gamma \asymp \log n$$



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

Our theory

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

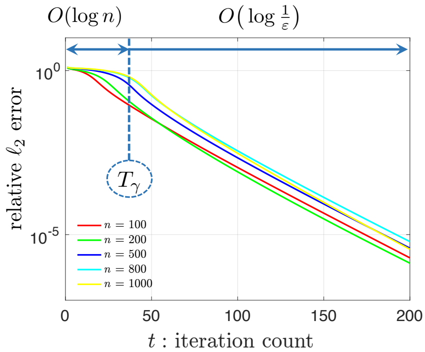


- *near-optimal computational cost*:

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

Our theory

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$



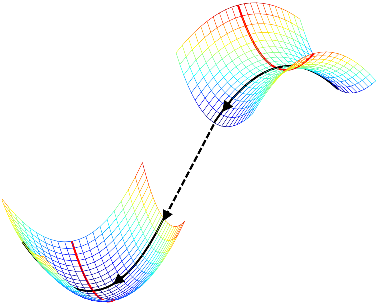
- *near-optimal computational cost*:

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

- *near-optimal sample size*: $m \gtrsim n\text{poly}\log m$

Generic algorithm design and analysis

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



Generic optimization theory yields highly suboptimal convergence guarantees

What we have not discussed so far

- A lot of interesting problems that nonconvex optimization could work, e.g., robust PCA, tensor estimation, mixture models, etc.
- A lot of algorithms, e.g., expectation maximization, alternating minimization, scaledGD, etc.
- Inference for nonconvex estimators
- Connections between nonconvex and convex estimators