Escaping Saddle Points or Not?

a case study of robust matrix sensing

Jianhao Ma
University of Michigan

FAI Seminar

April 7, 2023

Content

Content

Nonconvex Optimization

- A generic optimization problem aims to solve $\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$.
- Nonconvex optimization is much more difficult than its convex counterpart.

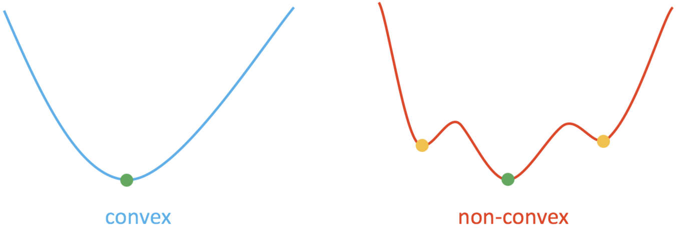


Figure: convex v.s. nonconvex optimizations[1]

[1] source: `https://stanford.edu/~pilanci/papers/TALK_Sketching.pdf`

Classification of Stationary Points

- First-order stationary point: $\nabla f(\mathbf{x}) = 0$.
- Second-order stationary point: $\nabla f(\mathbf{x}) = 0$, and $\nabla^2 f(\mathbf{x}) \succeq 0$.
- Approximate second-order stationary point: $\|\nabla f(\mathbf{x})\| \leq \varepsilon_g$, $\lambda_{\min}\left(\nabla^2 f(\mathbf{x})\right) \geq -\varepsilon_H$.

Figure: stationary points[2]

[2]source:
https://pythoninchemistry.org/ch40208/comp_chem_methods/geometry_optimisation.html

Nonconvex Optimization in ML: ERM Framework

- Provided with the training dataset $S = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \overset{\text{iid}}{\sim} \mathcal{D}$, the empirical risk minimization (ERM) aims to solve

$$\boxed{\min_{\theta \in \Theta} \mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \lambda \mathcal{R}(\theta).}$$

- Let $\Theta^\star = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta) = \mathbb{E}\left[\mathcal{L}_n(\theta)\right]$ be the set of all the ground truths. The challenging question is:

  For a **finite** *sample size* n*, can we find a* $\hat{\theta}_n$ *from solving the* **nonconvex** *optimization* $\min_{\theta \in \Theta} \mathcal{L}_n(\theta)$ *such that* $\hat{\theta}_n$ *is close to* $\Theta^\star$*?*

Difficulty of ERM

The fundamental question is:

> *For a **finite** sample size n, can we find a $\hat{\theta}_n$ from solving the **nonconvex** optimization $\min_{\theta \in \Theta} \mathcal{L}_n(\theta)$ such that $\hat{\theta}_n$ is close to $\Theta^\star$?*

Three difficulties:

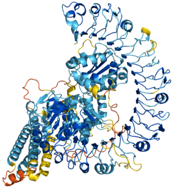
- **Optimization:** The optimization problem $\min_{\theta \in \Theta} \mathcal{L}_n(\theta)$ is highly nonconvex and can be nonsmooth.
- **Generalization:** A good solution of $\mathcal{L}_n(\theta)$ can be far away from $\Theta^\star$ provided with limited data (overfitting).
- **Sample complexity:** How many samples are sufficient to find a solution close to the ground truth (e.g., d v.s. d^{100})?

Empirical Success of Nonconvex Optimization in DL

Observation

Deep learning has achieved empirical success in many fields, such as computer vision, natural language processing, and robotics. The underlying mechanism is to solve a highly **nonconvex optimization** via first-order methods, like Adam, SGD.

(a) object detection



(b) AlphaFold

(c) GAN

General Nonconvex Optimization is Hard

However...

Fact

Solving nonconvex optimization is hard in the **worst** case. Sepcifically, finding a global solution is **NP-hard**.

- For local search algorithms, only local guarantees are available, i.e., converging to first/second-order stationary points [JGN^{+}17].
- GD can take exponential time to escape saddle points [DJL^{+}17].
- SGD can converge to local maxima [ZLSU21].
- Simple GD-like algorithm has a worse convergence rate for nonsmooth functions [B^{+}15].

Structured Nonconvex Optimization

Question: How to close this gap between theory and practice?

Hypothesis

If the function class has some special structures like convexity, then the optimization should be easy!

- Weak convexity.
- Polyak-Łojasiewicz (PL) condition.
- Restricted (strong) convexity.
- **Benign landscape**, i.e., no spurious local minima, strict saddle property.
- ...

Landscape Analysis

Hypothesis

Though it is nonconvex, the global landscape might enjoy some **benign** landscape properties so that local-search algorithms find ground truth.

Strict Saddle (Optimization, [JGN+17])

For any $\theta \in \Theta$, at least one of following holds

- $\|\nabla \mathcal{L}_n(\theta)\| \geq \varepsilon_g$;
- $\lambda_{\min}(\nabla^2 \mathcal{L}_n(\theta)) \leq -\varepsilon_H$;
- θ is ε-close to $\Theta^\star$ — the set of local minima.

Implication: first-order algorithms escape saddle points and converge to local minima.

How to Escape Saddle Points Efficiently? [JGN$^+$17]

- When gradient norm is large, i.e., $\|\nabla \mathcal{L}_n(\theta_t)\| \geq \varepsilon_g$, we apply gradient descent lemma

$$\mathcal{L}_n(\theta_{t+1}) - \mathcal{L}_n(\theta_t) \leq -\frac{\eta}{2} \|\nabla \mathcal{L}_n(\theta_t)\|^2 \leq -\frac{\eta}{2}\varepsilon_g^2. \tag{1}$$

- When θ_t is close to a saddle point, i.e., $\|\nabla \mathcal{L}_n(\theta_t)\| \leq \varepsilon_g$, running perturbed GD is similar to one-step Hessian update $\theta_{t+1} = \theta_t - \eta_H v$ where v is the eigenvector of $\lambda_{\min}(\nabla^2 \mathcal{L}_n(\theta_t))$. We have

$$\begin{aligned}
\mathcal{L}_n(\theta_{t+1}) - \mathcal{L}_n(\theta_t) &\leq \eta_H \langle \nabla \mathcal{L}_n(\theta_t), v\rangle + \frac{1}{2}\eta_H^2 v^\top \nabla^2 \mathcal{L}_n(\theta_t) v + \mathcal{O}(\eta_H^3) \\
&\leq \eta_H \varepsilon_g - \frac{1}{2}\eta_H^2 \gamma + \mathcal{O}(\eta_H^3) \qquad (2) \\
&\leq -\frac{1}{4}\eta_H^2 \gamma.
\end{aligned}$$

Benign Landscape (cont'd)

No Spurious Local Minimum (Statistics) [GLM16]

All the local minima are global.

Implication: local-search algorithms find the global minimum.

Identifiability [MF23]

The ground truth $\theta^\star \in \Theta^\star$ is identifiable if it is a stationary point of $\mathcal{L}_n(\theta)$.

Implication: local-search algorithms **could** find ground truth. Furthermore, if it corresponds to a global minimum, then local-search algorithms find ground truth.

Examples of Benign Landscape

Example:

- Matrix sensing [PKCS17].
- Matrix completion [GLM16, GJZ17].
- Deep linear neural network [Kaw16].
- Two hidden unit ReLU network [WLL18].
- ...

Implication: local-search algorithms work.

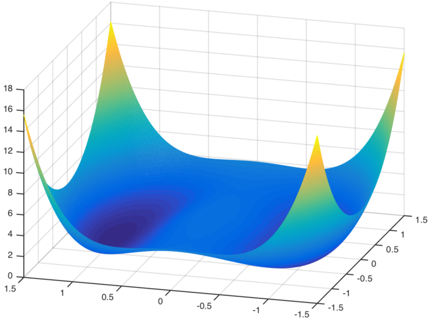


Figure: landscape of phase retrieval [SQW15]

Trajectory Analysis

Hypothesis

Even if the global landscape can be bad, the landscape of the **solution trajectory** might enjoy good properties, and algorithms have **implicit biases** towards the ground truth.

- **Pros:** This is nearly the minimal requirement for a local-search algorithm to succeed!
- **Cons:** How to prove it?

Examples of Trajectory Analysis

Example:

- Overparameterized sparse recovery [VKR19].
- Overparameterized matrix factorization [LMZ18, SS21].
- Deep linear neural network [ACGH18].
- ...

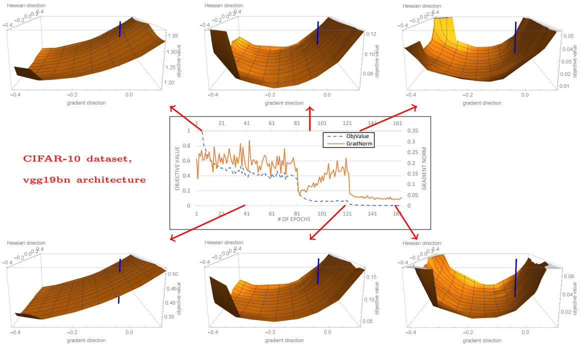


Figure: adapted from [AZLS19]

Content

Takeaway Message

There exists an instance of statistical learning problems (robust matrix sensing) such that with high probability:

1. GD can find a ground truth $\theta^\star_{\mathsf{GD}} \in \Theta^\star$ [MF22].
2. All the elements in $\Theta^\star$ are saddle points of $\mathcal{L}_n(\theta)$ [MF23].

Discussion

- Saddle-avoiding algorithms fail in this case.
- Landscape analysis has fundamental limits, so we must develop sophisticated trajectory analysis in the general case!

Robust Matrix Sensing

Problem (Robust Matrix Sensing)

The robust matrix sensing problem aims to

$$\text{find} \quad X^\star \qquad \text{subject to:} \quad \mathbf{y} = \mathcal{A}(X^\star) + \mathbf{s}, \quad \operatorname{rank}(X^\star) = r^\star.$$

- **Low-rank ground truth:** $X^\star \in \mathbb{R}^{d\times d}$ is PSD and $r^\star \ll d$.
- **Gaussian measurement matrices:** $\mathcal{A}(\cdot) = [\langle A_1, \cdot\rangle, \cdots, \langle A_n, \cdot\rangle]^\top$ where $A_1, \cdots, A_n \in \mathbb{R}^{d\times d}$ are i.i.d. standard Gaussian matrices.
- **Huber's contamination model:** $\lceil pn \rceil$ of measurements are corrupted by outliers $\mathbf{s}$ i.i.d. drawn from some unknown distribution $\mathcal{D}_{\text{outlier}}$ with $0 < p < 1$.

Nonconvex Optimization Formulation

Optimization

We solve the following optimization problem

$$\min_{U \in \mathbb{R}^{d \times r'}} \quad \mathcal{L}_n(U) := \frac{1}{n} \left\| \mathbf{y} - \mathcal{A}\left(UU^\top\right) \right\|_1 = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \left\langle A_i, UU^\top \right\rangle \right|. \tag{3}$$

Here $r^\star \leq r' \leq d$ is the search rank.

- **Why nonconvex optimization?** Traditional convex relaxation methods do not scale well.
- **Why ℓ_1-loss?** To promote robustness.
- **Why overparameterized model?** In practice, it is nontrivial to estimate the true rank $r^\star$. $UU^\top$ enforces PSD naturally.

Ground Truths are Saddles

Theorem (Informal)

Suppose the sample size $n \lesssim dr'$, the corruption probability $0 < p < 1$, and the radius $\gamma \leq 1/\operatorname{poly}(d)$. Then, with high probability, for any $U^\star$ such that $U^\star U^{\star\top} = X^\star$, we have

$$\min_{\|\Delta U\|_F \leq \gamma} \mathcal{L}_n(U^\star + \Delta U) - \mathcal{L}_n(U^\star) = -\Theta(\gamma^2). \tag{4}$$

- **Information lower bound:** $n = \Theta(dr^\star)$.
- **First-order stationary point:** Let the radius $\gamma \to 0$, we have

$$\lim_{\gamma \to 0} \sup_{\|\Delta U\|_F \leq \gamma} \frac{|\mathcal{L}_n(U^\star + \Delta U) - \mathcal{L}_n(U^\star)|}{\|\Delta U\|_F} = \lim_{\gamma \to 0} \Theta(\gamma) = 0.$$

GD Finds Ground Truth Efficiently

Theorem (informal)

Suppose the sample size $n \gtrsim dr^{\star 2}$ and the corruption probability $0 < p < 1$. For arbitrary accuracy $\varepsilon > 0$, we use GD with a proper learning rate regime and initialization. Then, with high probability, we have

$$\left\| U_T U_T^\top - X^\star \right\|_F \le \varepsilon, \tag{5}$$

after $T = \mathcal{O}\left(\kappa^2 \log^3(d/\varepsilon)\right)$ iterations.

- **Exact recovery:** We can set ε sufficiently small.
- **Near optimal sample complexity:** $dr^{\star 2}$ v.s. $dr^\star$ where $r^\star \ll d$.
- **Near linear convergence:** Our iteration complexity has polylog dependence with ε.

Content

Local Perturbation Analysis

Theorem

Suppose the sample size $n \lesssim dr'$, the corruption probability $0 < p < 1$, and the radius $\gamma \leq 1/\operatorname{poly}(d)$. Then, with high probability, for any $U^\star$ such that $U^\star U^{\star\top} = X^\star$, we have

$$\min_{\|\Delta U\|_F \leq \gamma} \mathcal{L}_n(U^\star + \Delta U) - \mathcal{L}_n(U^\star) = -\Theta(\gamma^2).$$

Comments:

- For simplicity, we only focus on the upper bound and set $r' = d$.
- Not so easy...
  - Cannot use first-order approximation.
  - Cannot use common concentration bounds for independent random variables.

Local Perturbation Analysis (cont'd)

The difference can be written as

$$\boxed{\mathcal{L}_n(U^\star + \Delta U) - \mathcal{L}_n(U^\star) = \frac{1}{n}\sum_{i\in I} |\langle A_i, \Delta X\rangle| + \frac{1}{n}\sum_{i\in O} \left(|\langle A_i, \Delta X\rangle - s_i| - |s_i|\right).} \tag{6}$$

Notations.

- Suppose $U^\star$ satisfies $U^\star U^{\star\top} = X^\star$.
- Denote $\Delta X = (U^\star + \Delta U)(U^\star + \Delta U)^\top - U^\star U^{\star\top}$.
- $y_i - \left\langle A_i, UU^\top \right\rangle = \left\langle A_i, X^\star - UU^\top \right\rangle + s_i$.
- $[n] = I \cup O$ (inliers and outliers) such that $s_i = 0, \forall i \in I$.

Main Idea: Construct a specific perturbation ΔU to minimize the above difference.

Structured Perturbation

Observation:

$$\Delta X = \underbrace{\Delta U U^{\star\top} + U^{\star}\Delta U^{\top}}_{\text{first-order term, rank-}r} + \underbrace{\Delta U \Delta U^{\top}}_{\text{second-order term}} . \tag{7}$$

Step 1. Cancel out the first-order term by restricting the perturbation such that $\Delta U U^{\star\top} = 0$. Hence,

$$\mathcal{L}_n(U^\star + \Delta U) - \mathcal{L}_n(U^\star) = \frac{1}{n}\sum_{i\in I}\left|\left\langle A_i, \Delta U \Delta U^\top \right\rangle\right| + \frac{1}{n}\sum_{i\in O}\left(\left|\left\langle A_i, \Delta U \Delta U^\top \right\rangle - s_i\right| - |s_i|\right). \tag{8}$$

Dimension of perturbation space: $d^2 \to d(d - r^\star) = \Omega(d^2)$.

Structured Perturbation (cont'd)

Step 2. For sufficiently small perturbation radius γ, with high probability, we have $\mathrm{Sign}(s_i - \langle A_i, \Delta U \Delta U^\top \rangle) = \mathrm{Sign}(s_i)$. Hence, we further have

$$\mathcal{L}_n(U^\star + \Delta U) - \mathcal{L}_n(U^\star) = \frac{1}{n}\sum_{i \in I} \left| \left\langle A_i, \Delta U \Delta U^\top \right\rangle \right| + \underbrace{\frac{1}{n}\sum_{i \in O} \mathrm{Sign}(s_i) \left\langle A_i, \Delta U \Delta U^\top \right\rangle}_{\text{Gaussian process}}.$$

Step 3. Choose $\Delta U = \arg\min_{\|\Delta U\|_F \le \gamma} \frac{1}{n} \sum_{i \in O} \mathrm{Sign}(s_i) \left\langle A_i, \Delta U \Delta U^\top \right\rangle$.

Structured Perturbation (cont'd)

- Our choice of ΔU is independent of $A_i, \forall i \in I$ so that

$$\begin{aligned}\frac{1}{n}\sum_{i\in I}\left|\left\langle A_i, \Delta U \Delta U^\top \right\rangle\right| &\approx (1-p)\mathbb{E}\left[\left|\left\langle A_i, \Delta U \Delta U^\top \right\rangle\right|\right] \\ &= \sqrt{2/\pi}(1-p)\left\|\Delta U \Delta U^\top\right\|_F .\end{aligned} \tag{9}$$

- Applying Sudakov inequality, we have

$$\begin{aligned}\frac{1}{n}\sum_{i\in O}\mathrm{Sign}(s_i)\left\langle A_i, \Delta U \Delta U^\top \right\rangle &\lesssim -\sqrt{\frac{p\dim(\Delta U)}{n}}\left\|\Delta U \Delta U^\top\right\|_F \\ &\lesssim -\sqrt{\frac{pd^2}{n}}\left\|\Delta U \Delta U^\top\right\|_F .\end{aligned} \tag{10}$$

Supremum of Gaussian Process[3]

Definition (Gaussian process)

A random process $\{X_t\}_{t\in T}$ is called a Gaussian process if, for any finite subset $T_0 \subset \mathcal{T}$, the random vector $\{X_t\}_{t\in T_0}$ has a normal distribution.

Lemma (Sudakov's minoration inequality, informal)

For a centered Gaussian process $\{X_t\}_{t\in T}$ *with variance proxy* $\sigma^2 = \inf_{t\in T}\mathbb{E}[X_t^2]$, *we have*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \gtrsim \sigma\sqrt{\dim(T)}.$$

[3] Reference: `https://www.bilibili.com/video/BV1CU4y1h7Ao/?spm_id_from=333.999.0.0&vd_source=e0e131166191f0238a27cd7bf5ad57a3`

Content

Algorithm

- **Optimization formulation:**

$$\min_{U \in \mathbb{R}^{d \times r'}} \quad \mathcal{L}(U) := \frac{1}{n} \left\| \mathbf{y} - \mathcal{A}\left(UU^\top\right) \right\|_1, \tag{11}$$

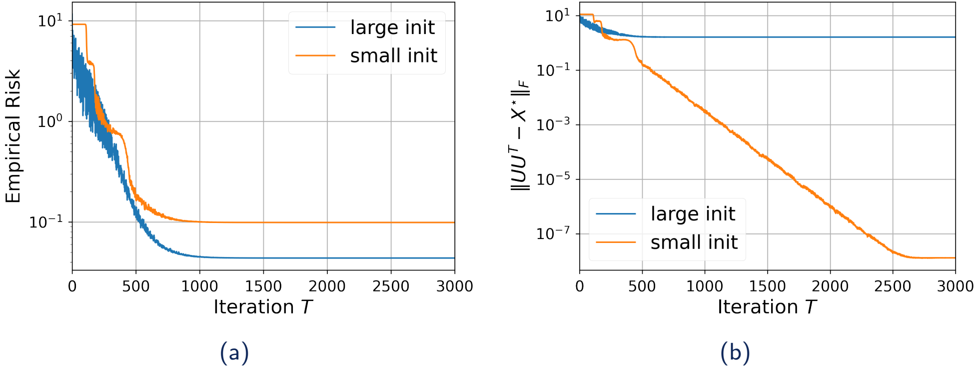
- **Algorithm:** GD with geometric stepsize $U_{t+1} = U_t - \eta \rho^t D_t$ where

$$D_t \in \partial \mathcal{L}(U_t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Sign}\left(\langle A_i, U_t U_t^\top - X^\star \rangle\right) \left(A_i + A_i^\top\right) U_t.$$

- **Initialization:** small (spectral) initialization $U_0 = \alpha B$, where $BB^\top$ is the robust analog of spectral initialization, satisfying $BB^\top \approx X^\star$.

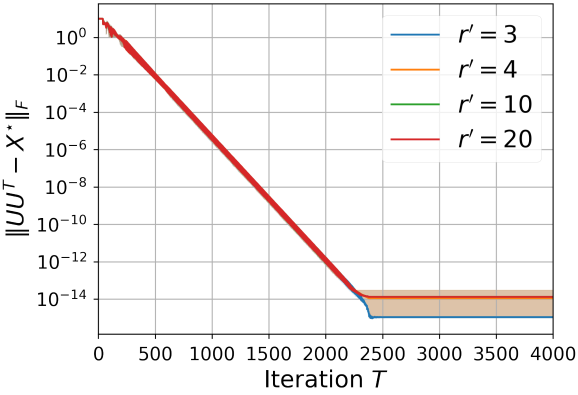
Emergence of "spurious" global minima

- Better objective value $\not\Rightarrow$ better generalization error.
- Plain landscape analysis fails! $\Rightarrow$ trajectory analysis!



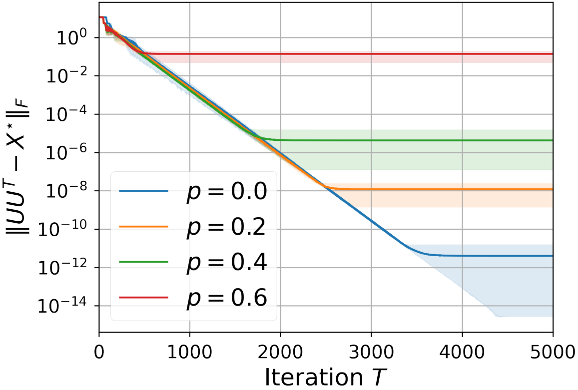
GD is Agnostic to Over-parameterization

- The dimension $d = 20$, the rank $r^\star = 3$, sample size $n = 300$, and the corruption probability $p = 0.1$. For each choice, we run 5 independent trials.



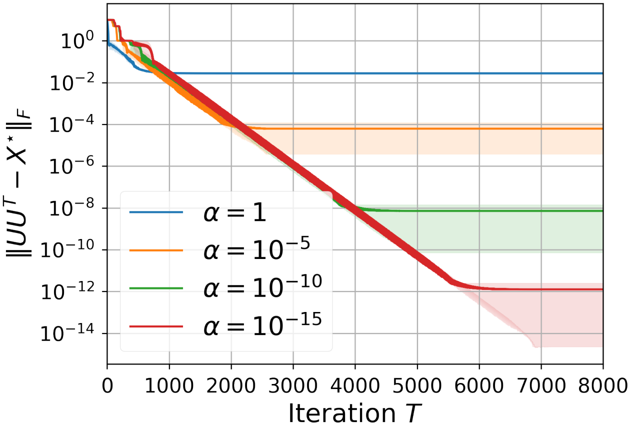
GD is Robust against Outlier

- The same setting as the last slide. The search rank is set to be $r' = 20$.



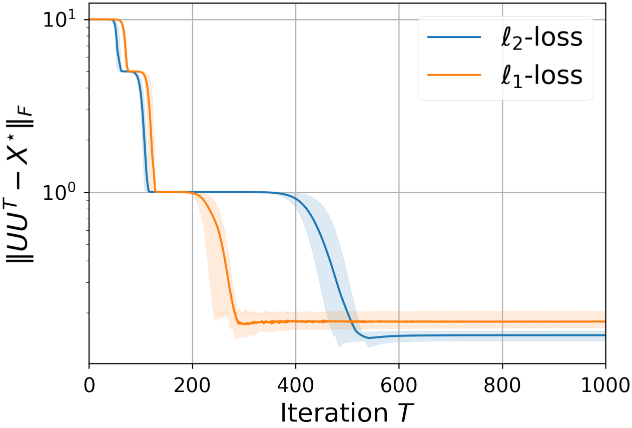
Effect of Small Initialization

- The error is proportional to the initialization scale: $\left\|UU^\top - X^\star\right\|_F \propto \alpha^\gamma$.
- In practice, we can choose $\alpha = \varepsilon^{1/\gamma}$ to make $\left\|UU^\top - X^\star\right\|_F \lesssim \varepsilon$.



Successful in Gaussian Noise Model

- The performance of ℓ_1-loss is comparable with ℓ_2-loss, which is minimax optimal.
- We have information lower bound $\left\| UU^\top - X^\star \right\|_F \gtrsim \sqrt{\frac{dr^\star}{n}}$.



Proof Sketch

Matrix Factorization ($n \to \infty$)

We start with the **population** loss ($n \to \infty$) and **noiseless** setting ($p = 0$).

Overparameterized Matrix Factorization (ℓ_1-loss)

Suppose the measurement matrices are standard Gaussian, and the noise vector is zero. When the measurement number $n \to \infty$, the objective function becomes

$$\min_{U \in \mathbb{R}^{d \times r'}} \bar{\mathcal{L}}(U) := \sqrt{\frac{2}{\pi}} \left\| UU^\top - X^\star \right\|_F .$$

Equivalence between ℓ_1- and ℓ_2-loss

Observation

Using GD to solve

$$\min_{U \in \mathbb{R}^{d \times r'}} \bar{\mathcal{L}}(U) := \frac{1}{2} \left\| UU^\top - X^\star \right\|_F \quad \text{with stepsize } \bar{\eta}_t = \eta_0 \left\| U_t U_t^\top - X^\star \right\|_F$$

is equivalent to using GD to solve

$$\min_{U \in \mathbb{R}^{d \times r'}} \frac{1}{4} \left\| UU^\top - X^\star \right\|_F^2 \quad \text{with stepsize } \bar{\eta}_t = \eta_0.$$

Intuition: solving ℓ_2-loss via constant stepsize GD might be easy to analyze [LMZ18, ZKHC21, SS21].

Signal-Residual Decomposition

- $X^\star = V\Sigma V^\top$, where $\Sigma = \mathrm{Diag}\{\sigma_1, \cdots, \sigma_{r^\star}\}$.
- **Signal-Residual Decomposition:** We project the matrix U_t onto the column space of V, and its orthogonal complement $V^\perp$ (recall that $X^\star = V\Sigma V^\top$)

$$U_t = VS_t + V_\perp E_t, \quad \text{where } \underbrace{S_t = V^\top U_t}_{\text{rank-}r^\star}, \quad \underbrace{E_t = V_\perp^\top U_t}_{\text{dense, but small???}} .$$

Lemma (Signal-Residual Decomposition)

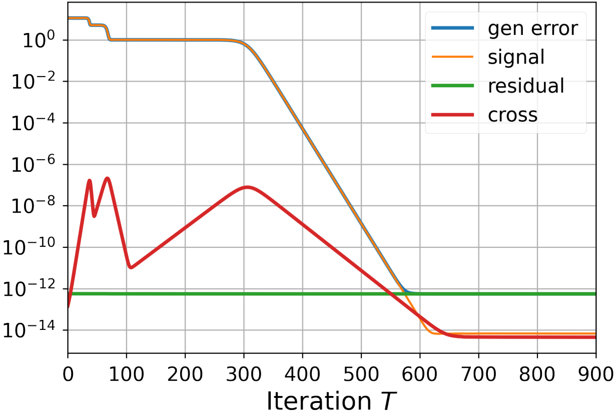
The generalization error can be decomposed as

$$U_t U_t^\top - X^\star = \underbrace{V\left(S_t S_t^\top - \Sigma\right)V^\top + VS_t E_t^\top V_\perp^\top + V_\perp E_t S_t^\top V^\top}_{\text{rank-}3r^\star} + \underbrace{V_\perp E_t E_t^\top V_\perp^\top}_{\text{small???}},$$

$$\left\|U_t U_t^\top - X^\star\right\| \leq \underbrace{\left\|\Sigma - S_t S_t^\top\right\|}_{\text{signal}} + 2\underbrace{\left\|S_t E_t^\top\right\|}_{\text{cross}} + \underbrace{\left\|E_t E_t^\top\right\|}_{\text{residual}}.$$

Signal-Residual Decomposition (cont'd)

- The projected signal term $S_t S_t^\top$ satisfies a local regularity condition (an analog of strong convexity) so that it converges linearly to the projected ground truth.



Robust Matrix Sensing: Noiseless Case

- Recall that in the population case, we can choose the stepsize $\bar{\eta}_t = \eta_0 \left\| UU^\top - X^\star \right\|_F$.
- In the noiseless case, we can choose $\eta_t = \eta_0 \frac{1}{n} \sum_{i=1}^n \left| y_i - \left\langle A_i, U_t U_t^\top \right\rangle \right|$, which is a good approximation of $\bar{\eta}_t$ up to some constant, i.e., $\eta_t \asymp \bar{\eta}_t$.
- **Question 1:** Is the sub-differential $\partial \mathcal{L}(U)$ close to $\partial \bar{\mathcal{L}}(U)$?
- **Question 2:** Does the trajectory $U_0, \cdots, U_T$ have similar behavior as that corresponds to the population loss provided an affirmative answer to Question 1?

Uniform Convergence of Sub-differential

Theorem (Uniform convergence of sub-differential)

For standard Gaussian matrices, suppose the measurement number $n = \tilde{\Omega}(dr^\star)$. Then with high probability, for arbitrary ε-approximate rank-$\mathcal{O}(r^\star)$ matrix U, we have

$$\left\|\partial\mathcal{L}(U) - \partial\bar{\mathcal{L}}(U)\right\| \lesssim \|U\|\,\delta.$$

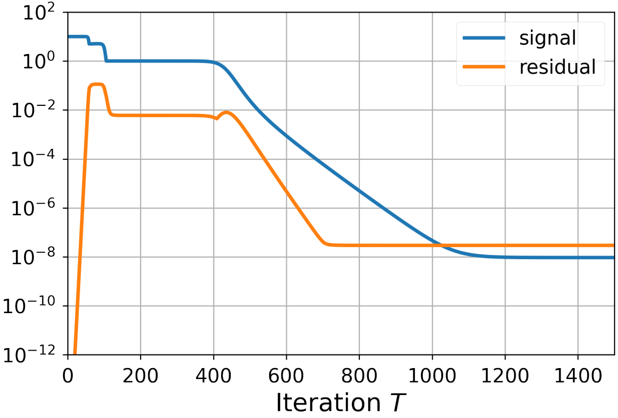
Here ε, δ are small numbers depending only on $m, d, r^\star$.

Remark:

- Uniform result holds for all approximate low-rank matrices simultaneously.
- It holds for both outlier and Gaussian noise models.
- Highly nontrivial since sub-differential is discontinuous.

Decomposed Dynamics on Matrix Sensing

- Large $\left\| E_t E_t^\top \right\|$ $\Longrightarrow$ Further decomposition!



Decomposed Residual Dynamics

- Project the residual onto S_t and its orthogonal complement:

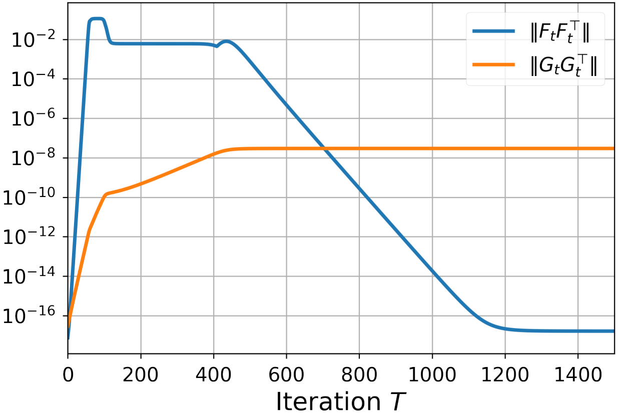
$$\underbrace{F_t := E_t \mathsf{P}_{S_t},}_{\text{rank-}r^\star} \qquad \underbrace{G_t := E_t \mathsf{P}_{S_t}^{\perp}}_{\text{dense, but small???}}, \text{where } \mathsf{P}_{S_t}, \mathsf{P}_{S_t}^{\perp} \text{ are projection operators.}$$

- We can decompose the generalization error as

$$\begin{aligned} & U_t U_t^\top - X^\star \\ & = \underbrace{V\left(S_t S_t^\top - \Sigma\right)V^\top + V S_t E_t^\top V_\perp^\top + V_\perp E_t S_t^\top V^\top + V_\perp F_t F_t^\top V_\perp^\top}_{\text{rank-}4r^\star} + \underbrace{V_\perp G_t G_t^\top V_\perp^\top}_{\text{small norm???}}. \end{aligned}$$

- We have $\left\|E_t E_t^\top\right\| \le \left\|F_t F_t^\top\right\| + \left\|G_t G_t^\top\right\|$.

Decomposed Residual Dynamics(cont'd)



Robust Matrix Sensing: Noisy Case

- In the existence of noise, $\eta_t = \eta_0 \frac{1}{n} \sum_{i=1}^{n} | \left\langle A_i, X^\star - U_t U_t^\top \right\rangle + s_i |$ is no longer a good approximation of $\bar{\eta}_t = \eta_0 \left\| U_t U_t^\top - X^\star \right\|_F$.

- Instead, we use **exponentially decayed** stepsize $\eta_t = \eta_0 \rho^t$.

- **Intuition:** If the algorithm works as expected, the error measure decreases linearly, i.e., $\left\| U_t U_t^\top - X^\star \right\|_F \asymp \rho^t$. Hence, $\eta_t = \eta_0 \rho^t \approx \eta_0 \left\| U_t U_t^\top - X^\star \right\|_F = \bar{\eta}_t$.

Content

Summary

There exists an instance of statistical learning problems (robust matrix sensing) such that with high probability:

1. GD can find a ground truth $\theta^\star_{\mathsf{GD}} \in \Theta^\star$.
2. All the elements in $\Theta^\star$ are saddle points of $\mathcal{L}_n(\theta)$.

Q & A

[ACGH18] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[B+15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[DJL+17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. *Advances in neural information processing systems*, 30, 2017.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

[JGN$^+$17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[Kaw16] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

[LMZ18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[MF22] Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *arXiv preprint arXiv:2202.08788*, 2022.

[MF23] Jianhao Ma and Salar Fattahi. On the optimization landscape of burer-monteiro factorization: When do global solutions correspond to ground truth? *arXiv preprint arXiv:2302.10963*, 2023.

[PKCS17] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74. PMLR, 2017.

[SQW15] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[SS21] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

[VKR19] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

[WLL18] Chenwei Wu, Jiajun Luo, and Jason D Lee. No spurious local minima in a two hidden unit relu network. 2018.

[ZKHC21] Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*, 2021.

[ZLSU21] Liu Ziyin, Botao Li, James B Simon, and Masahito Ueda. Sgd with a constant large learning rate can converge to local maxima. *arXiv preprint arXiv:2107.11774*, 2021.