Symmetry, Saddle Points, and Global Geometry of Nonconvex Matrix Factorization

Xingguo Li

Joint work with Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao

University of Minnesota
Johns Hopkins University
Princeton University
Georgia Institute of Technology

Background

Consider a low-rank matrix estimation problem:

$$\min_{M} f(M) \quad \text{subject to } \operatorname{rank}(M) \leq r,$$

where $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ is convex and smooth

- Fit **Wide** class of problems; **NP-hard** in general

Background

Consider a low-rank matrix estimation problem:

$$\min_{M} f(M) \quad \text{subject to } \operatorname{rank}(M) \leq r,$$

where $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is convex and smooth

- Fit **Wide** class of problems; **NP-hard** in general

➔ Convex relaxation:

$$\min_{M} f(M) \quad \text{subject to } ||M||_{*} \leq \tau,$$

- **Easy** to analyze; Computationally **Expensive**, e.g., SVD

Background

Consider a low-rank matrix estimation problem:

$$\min_{M} f(M) \quad \text{subject to } \text{rank}(M) \leq r,$$

where $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is convex and smooth

- Fit **Wide** class of problems; **NP-hard** in general

→ Convex relaxation:

$$\min_{M} f(M) \quad \text{subject to } ||M||_* \leq \tau,$$

- **Easy** to analyze; Computationally **Expensive**, e.g., SVD

→ Nonconvex formulation:

$$\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} f(XY^\top),$$

- **Good** empirical performance; **Challenging** for analysis

Background

Challenges in $\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} f(XY^\top)$:

- Infinitely many nonisolated saddle points
  Example: (X, Y) is a saddle $\rightarrow$ $(X\Phi, Y\Phi)$ is also a saddle $\forall\ \Phi$
- Nonconvex on X, Y, even $f(\cdot)$ is convex

Background

Challenges in $\min_{X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{m \times r}} f(XY^\top)$:

- Infinitely many nonisolated saddle points
  Example: (X, Y) is a saddle $\rightarrow$ $(X\Phi, Y\Phi)$ is also a saddle $\forall\ \Phi$
- Nonconvex on X, Y, even $f(\cdot)$ is convex

Existing approach:

- Generalization of convexity: Local regularity condition (Candes et al., 2015)
- Geometric characterization: Local properties vs. Global properties

(Ge et al., 2016; Sun et al., 2016)

Background

Challenges in $\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} f(XY^\top)$:

- Infinitely many nonisolated saddle points
  Example: (X, Y) is a saddle $\to$ $(X\Phi, Y\Phi)$ is also a saddle $\forall\ \Phi$
- Nonconvex on X, Y, even $f(\cdot)$ is convex

Existing approach:

- Generalization of convexity: Local regularity condition (Candes et al., 2015)
- Geometric characterization: Local properties vs. Global properties (Ge et al., 2016; Sun et al., 2016)

Our approach:

- A **novel theory** characterizing stationary points
- A **full geometric characterization** of low-rank matrix factorization
- An extension to **constrained problems**

Different Types of Stationary Points

Definition

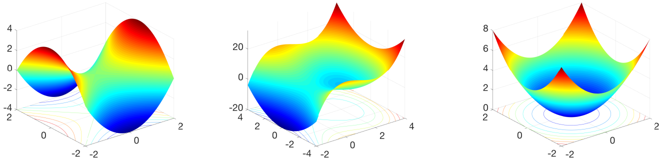
Given a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, a point $x \in \mathbb{R}^n$ is called:

(i) a **stationary point**, if $\nabla f(x) = 0$;

(ii) a **local minimum**, if x is a stationary and $\exists$ a neighborhood $\mathcal{B} \subseteq \mathbb{R}^n$ of x such that $f(x) \leq f(y)$ for any $y \in \mathcal{B}$;

(iii) a **global minimum**, if x is a stationary and $f(x) \leq f(y)$, $\forall y \in \mathbb{R}^n$;

(iv) a **strict saddle point**, if x is a stationary and $\forall$ neighborhood $\mathcal{B} \subseteq \mathbb{R}^n$ of x, $\exists y, z \in \mathcal{B}$ s.t. $f(z) \leq f(x) \leq f(y)$ & $\lambda_{\min}(\nabla^2 f(x)) < 0$.



(a) strict saddle (b) local minimum (c) global minimum

A Generic Theory for Stationary Points

- Invariant group $\mathcal{G}$ of f: A subgroup of a special linear group, if $f(x) = f(g(x))$ for all $x \in \mathbb{R}^m$ and $g \in \mathcal{G}$.
- Fixed point $x_{\mathcal{G}}$ of a group $\mathcal{G}$: if $g(x_{\mathcal{G}}) = x_{\mathcal{G}}$ for all $g \in \mathcal{G}$.

Theorem (Stationary Fixed Point)

Suppose f has an invariant group $\mathcal{G}$ and $\mathcal{G}$ has a fixed point $x_{\mathcal{G}}$. If we have

$$\mathcal{G}(\mathbb{R}^m) \triangleq \mathrm{Span}\{g(x) - x \mid g \in \mathcal{G}, x \in \mathbb{R}^m\} = \mathbb{R}^m,$$

then $x_{\mathcal{G}}$ is a stationary point of f.

A Generic Theory for Stationary Points

- Invariant group $\mathcal{G}$ of f: A subgroup of a special linear group, if $f(x) = f(g(x))$ for all $x \in \mathbb{R}^m$ and $g \in \mathcal{G}$.
- Fixed point $x_{\mathcal{G}}$ of a group $\mathcal{G}$: if $g(x_{\mathcal{G}}) = x_{\mathcal{G}}$ for all $g \in \mathcal{G}$.

Theorem (Stationary Fixed Point)

Suppose f has an invariant group $\mathcal{G}$ and $\mathcal{G}$ has a fixed point $x_{\mathcal{G}}$. If we have

$$\mathcal{G}(\mathbb{R}^m) \triangleq \mathrm{Span}\{g(x) - x \mid g \in \mathcal{G}, x \in \mathbb{R}^m\} = \mathbb{R}^m,$$

then $x_{\mathcal{G}}$ is a stationary point of f.

Corollary

If $y_{\mathcal{G}_{\mathcal{Y}}}$ is a fixed point of $\mathcal{G}_{\mathcal{Y}}$, an induced subgroup of $\mathcal{G}$, and

$$z^*(y_{\mathcal{G}_{\mathcal{Y}}}) \in \arg\,\underset{z}{\mathrm{zero}}\, \nabla_z f(y_{\mathcal{G}_{\mathcal{Y}}} \oplus z),$$

then $g(y_{\mathcal{G}_{\mathcal{Y}}} \oplus z^*)$ is a stationary point for all $g \in \mathcal{G}$.

Examples

- Low-rank Matrix Factorization:

  $\min_X f(X) = \frac{1}{4}\|XX^\top - M^*\|_F^2$, where $M^* = UU^\top$

  - Invariant group:$\mathfrak{O}_r = \{\Psi \in \mathbb{R}^{r\times r} \mid \Psi\Psi^\top = \Psi^\top\Psi = I_r\}$; Fixed point:0
  - $\mathcal{Y} = \mathcal{L}_{U_{r-s}} \subseteq \mathcal{L}_U$ and $\mathcal{Z} = \mathcal{L}_{U_s} \subseteq \mathcal{L}_U$

    $\Rightarrow$ $U_s\Psi_r$ is stationary, where $\Psi_r \in \mathfrak{O}_r$, $U_s = \Phi\Sigma S\Theta^\top$,$U = \Phi\Sigma\Theta^\top$ (SVD), and S is a diagonal matrix w/ s entries 1 and 0 o.w. $\forall s \in [r]$

Examples

- Low-rank Matrix Factorization:

  $\min_X f(X) = \frac{1}{4}\|XX^\top - M^*\|_F^2$, where $M^* = UU^\top$

  - Invariant group: $\mathfrak{O}_r = \{\Psi \in \mathbb{R}^{r\times r} \mid \Psi\Psi^\top = \Psi^\top\Psi = I_r\}$; Fixed point:0
  - $\mathcal{Y} = \mathcal{L}_{U_{r-s}} \subseteq \mathcal{L}_U$ and $\mathcal{Z} = \mathcal{L}_{U_s} \subseteq \mathcal{L}_U$

    $\Rightarrow$ $U_s\Psi_r$ is stationary, where $\Psi_r \in \mathfrak{O}_r$, $U_s = \Phi\Sigma S\Theta^\top, U = \Phi\Sigma\Theta^\top$ (SVD), and S is a diagonal matrix w/ s entries 1 and 0 o.w. $\forall s \in [r]$

- Phase Retrieval: $\min_x h(x) = \frac{1}{2m}\sum_{i=1}^m \left(y_i^2 - |a_i^H x|^2\right)^2$

  Expected objective: $f(x) = \mathbb{E}(h(x)) = \|x\|_2^4 + \|u\|_2^4 - \|x\|_2^2\|u\|_2^2 - |x^H u|^2$

  - Invariant group: $\mathcal{G} = \left\{e^{i\theta} \mid \theta \in [0, 2\pi)\right\}$; Fixed point:0
  - $\mathcal{Y} = \{y_i = 0, \forall i \in \mathcal{C}\}$ and $\mathcal{Z} = \{z_i = 0, \forall i \in [n]\backslash\mathcal{C}\}$, $\mathcal{C} \subseteq [n]$, $|\mathcal{C}| \leq n$

    $\Rightarrow$ x is stationary, if $x^H u = 0$, $x_{\mathcal{Y}} = 0$, $\|x\|_2 = \|u\|_2/\sqrt{2}$

Examples

- Low-rank Matrix Factorization:
  $\min_X f(X) = \frac{1}{4}\|XX^\top - M^*\|_F^2$, where $M^* = UU^\top$
  - Invariant group:$\mathfrak{O}_r = \{\Psi \in \mathbb{R}^{r\times r} \mid \Psi\Psi^\top = \Psi^\top\Psi = I_r\}$; Fixed point:0
  - $\mathcal{Y} = \mathcal{L}_{U_{r-s}} \subseteq \mathcal{L}_U$ and $\mathcal{Z} = \mathcal{L}_{U_s} \subseteq \mathcal{L}_U$
    $\Rightarrow$ $U_s\Psi_r$ is stationary, where $\Psi_r \in \mathfrak{O}_r$, $U_s = \Phi\Sigma S\Theta^\top, U = \Phi\Sigma\Theta^\top$ (SVD), and S is a diagonal matrix w/ s entries 1 and 0 o.w. $\forall s \in [r]$

- Phase Retrieval: $\min_x h(x) = \frac{1}{2m}\sum_{i=1}^m \left(y_i^2 - |a_i^{\mathrm{H}} x|^2\right)^2$
  Expected objective: $f(x) = \mathbb{E}(h(x)) = \|x\|_2^4 + \|u\|_2^4 - \|x\|_2^2\|u\|_2^2 - |x^{\mathrm{H}}u|^2$
  - Invariant group:$\mathcal{G} = \left\{e^{i\theta} \mid \theta \in [0, 2\pi)\right\}$; Fixed point:0
  - $\mathcal{Y} = \{y_i = 0, \forall i \in \mathcal{C}\}$ and $\mathcal{Z} = \{z_i = 0, \forall i \in [n]\backslash\mathcal{C}\}$, $\mathcal{C} \subseteq [n]$, $|\mathcal{C}| \leq n$
    $\Rightarrow$ x is stationary, if $x^{\mathrm{H}}u = 0$, $x_{\mathcal{Y}} = 0$, $\|x\|_2 = \|u\|_2/\sqrt{2}$

- Deep Linear Neural Networks ...

Null Space of Hessian Matrix at Stationary Points

Definition (Tangent Space)

Let $\mathcal{M} \subset \mathbb{R}^m$ be a smooth k-dimensional manifold. Given $x \in \mathcal{M}$, we call $v \in \mathbb{R}^m$ as a **tangent vector** of $\mathcal{M}$ at x if there exists a smooth curve $\gamma : \mathbb{R} \to \mathcal{M}$ with $\gamma(0) = x$ and $v = \gamma'(0)$. The set of tangent vectors of $\mathcal{M}$ at x is called the **tangent space** of $\mathcal{M}$ at x, denoted as

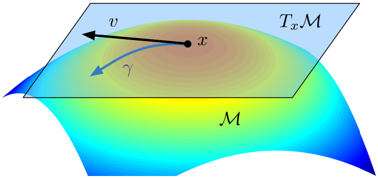
$$T_x\mathcal{M} = \{\gamma'(0) \mid \gamma : \mathbb{R} \to \mathcal{M} \text{ is smooth }, \; \gamma(0) = x\} .$$

Null Space of Hessian Matrix at Stationary Points

Definition (Tangent Space)

Let $\mathcal{M} \subset \mathbb{R}^m$ be a smooth k-dimensional manifold. Given $x \in \mathcal{M}$, we call $v \in \mathbb{R}^m$ as a **tangent vector** of $\mathcal{M}$ at x if there exists a smooth curve $\gamma : \mathbb{R} \to \mathcal{M}$ with $\gamma(0) = x$ and $v = \gamma'(0)$. The set of tangent vectors of $\mathcal{M}$ at x is called the **tangent space** of $\mathcal{M}$ at x, denoted as

$$T_x\mathcal{M} = \{\gamma'(0) \mid \gamma : \mathbb{R} \to \mathcal{M} \text{ is smooth }, \ \gamma(0) = x\} .$$



Null Space of Hessian Matrix at Stationary Points

Definition (Tangent Space)

Let $\mathcal{M} \subset \mathbb{R}^m$ be a smooth k-dimensional manifold. Given $x \in \mathcal{M}$, we call $v \in \mathbb{R}^m$ as a **tangent vector** of $\mathcal{M}$ at x if there exists a smooth curve $\gamma : \mathbb{R} \to \mathcal{M}$ with $\gamma(0) = x$ and $v = \gamma'(0)$. The set of tangent vectors of $\mathcal{M}$ at x is called the **tangent space** of $\mathcal{M}$ at x, denoted as

$$T_x\mathcal{M} = \{\gamma'(0) \mid \gamma : \mathbb{R} \to \mathcal{M} \text{ is smooth },\ \gamma(0) = x\}.$$

Theorem

If f has an invariant group $\mathcal{G}$ and H_x is the Hessian matrix at a stationary point x, then we have

$$T_x\mathcal{G}(x) \subseteq Null(H_x).$$

Example

➔ Low-rank Matrix Factorization: Let $\gamma : \mathbb{R} \to \mathfrak{O}_r(X)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}$, $\exists \Psi_r \in \mathfrak{O}_r$ s.t. $\gamma(t) = g_t(X) = X\Psi_r$ and $\gamma(0) = g_0(X) = X$

$\Rightarrow$ $\gamma(t)\gamma(t)^T = XX^T$

$\Rightarrow$ $\gamma'(0)X^T + X\gamma'(0)^T = 0$ by differentiation

$\Rightarrow$ $T_X\mathfrak{O}_r(X) = \{XE \mid E \in \mathbb{R}^{r\times r}, E = -E^T\}$, e.g., $U_s\Psi_r E \in \mathrm{Null}(H_{U_s\Psi_r})$

Example

- Low-rank Matrix Factorization: Let $\gamma : \mathbb{R} \to \mathfrak{O}_r(X)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}$, $\exists \Psi_r \in \mathfrak{O}_r$ s.t. $\gamma(t) = g_t(X) = X\Psi_r$ and $\gamma(0) = g_0(X) = X$
  - $\Rightarrow$ $\gamma(t)\gamma(t)^T = XX^T$
  - $\Rightarrow$ $\gamma'(0)X^T + X\gamma'(0)^T = 0$ by differentiation
  - $\Rightarrow$ $T_X\mathfrak{O}_r(X) = \{XE \mid E \in \mathbb{R}^{r\times r}, E = -E^T\}$, e.g., $U_s\Psi_r E \in \mathrm{Null}(H_{U_s\Psi_r})$
- Phase Retrieval: Let $\gamma : \mathbb{R} \to \mathcal{G}(x)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}$, $\exists \theta \in [0, 2\pi)$ s.t. $\gamma(t) = x\mathrm{e}^{i\theta}$ and $\gamma(0) = x$
  - $\Rightarrow$ $\|\gamma(t)\|_2^2 = \|x\|_2^2$
  - $\Rightarrow$ $\gamma'(0)^{\mathrm{H}}x = -x^{\mathrm{H}}\gamma'(0)$ by differentiation w.r.t. t
  - $\Rightarrow$ $T_x\mathcal{G}(x) = ix$, e.g., $iu\mathrm{e}^{i\theta} \in \mathrm{Null}(H_{u\mathrm{e}^{i\theta}})$

Example

- Low-rank Matrix Factorization: Let $\gamma : \mathbb{R} \to \mathfrak{O}_r(X)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}$, $\exists \Psi_r \in \mathfrak{O}_r$ s.t. $\gamma(t) = g_t(X) = X\Psi_r$ and $\gamma(0) = g_0(X) = X$
  - $\Rightarrow \gamma(t)\gamma(t)^T = XX^T$
  - $\Rightarrow \gamma'(0)X^T + X\gamma'(0)^T = 0$ by differentiation
  - $\Rightarrow T_X\mathfrak{O}_r(X) = \{XE \mid E \in \mathbb{R}^{r\times r}, E = -E^T\}$, e.g., $U_s\Psi_r E \in \mathrm{Null}(H_{U_s\Psi_r})$
- Phase Retrieval: Let $\gamma : \mathbb{R} \to \mathcal{G}(x)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}$, $\exists\theta \in [0, 2\pi)$ s.t. $\gamma(t) = xe^{i\theta}$ and $\gamma(0) = x$
  - $\Rightarrow \|\gamma(t)\|_2^2 = \|x\|_2^2$
  - $\Rightarrow \gamma'(0)^{\mathrm{H}}x = -x^{\mathrm{H}}\gamma'(0)$ by differentiation w.r.t. t
  - $\Rightarrow T_x\mathcal{G}(x) = ix$, e.g., $iue^{i\theta} \in \mathrm{Null}(H_{ue^{i\theta}})$
- Deep Linear Neural Networks ...

A Geometric Analysis of Low-Rank Matrix Factorization

Given an objective $\mathcal{F}(X)$, our analysis consists of the following major arguments:

- Identify all stationary points, i.e., the solutions of $\nabla \mathcal{F}(X) = 0$
- Identify the strict saddle point and their neighborhood such that $\lambda_{\min}(\nabla^2 \mathcal{F}(X)) < 0$, denoted as $\mathcal{R}_1$
- Identify the global minimum, their neighborhood, and the directions such that $\lambda_{\min}(\nabla^2 \mathcal{F}(X)) > 0$, denoted as $\mathcal{R}_2$
- Verify that the gradient has a sufficiently large norm outside the regions described in (p2) and (p3), denoted as $\mathcal{R}_3$

$\Longrightarrow$ Iterative algorithms **DO NOT** converge to saddle point, e.g. first order methods (Ge et al., 2015) and second order methods (Sun et al., 2016).

Low-Rank Matrix Factorization: Rank-1 Case

Theorem

Consider $\min_{x\in\mathbb{R}^n} \mathcal{F}(x)$, *where* $\mathcal{F}(x) = \frac{1}{4}||M^* - xx^\top||_F^2$. *Define*

$$\mathcal{R}_1 \triangleq \left\{y \in \mathbb{R}^n \mid ||y||_2 \le \frac{1}{2}||u||_2\right\},$$

$$\mathcal{R}_2 \triangleq \left\{y \in \mathbb{R}^n \mid ||y - u||_2 \le \frac{1}{8}||u||_2\right\}, \text{ and}$$

$$\mathcal{R}_3 \triangleq \left\{y \in \mathbb{R}^d \mid ||y||_2 > \frac{1}{2}||u||_2,\ ||y - u||_2 > \frac{1}{8}||u||_2\right\}.$$

Then the following properties hold.

- $x = 0$, u *and* $-u$ *are the only stationary points of* $\mathcal{F}(x)$.
- $x = 0$ *is a strict saddle point with* $\lambda_{\min}(\nabla^2\mathcal{F}(0)) = -||u||_2^2$. *Moreover, for any* $x \in \mathcal{R}_1$, $\lambda_{\min}(\nabla^2\mathcal{F}(x)) \le -\frac{1}{2}||u||_2^2$.
- *For* $x = \pm u$, x *is a global minimum with* $\lambda_{\min}(\mathcal{F}(x)) = ||u||_2^2$. *Moreover, for any* $x \in \mathcal{R}_2$, $\lambda_{\min}(\nabla^2\mathcal{F}(x)) \ge \frac{1}{5}||u||_2^2$.
- *For any* $x \in \mathcal{R}_3$, *we have* $||\nabla\mathcal{F}(x)||_2 > \frac{||u||_2^3}{8}$.

Low-Rank Matrix Factorization: Rank-r Case

Introduce two sets:

$$\mathcal{X} = \left\{X = \Phi\Sigma_2\Theta_2 \mid U = \Phi\Sigma_1\Theta_1(\text{SVD}),\ (\Sigma_2^2 - \Sigma_1^2)\Sigma_2 = 0, \Theta_2 \in \mathfrak{O}_r\right\},$$
$$\mathcal{U} = \left\{X \in \mathcal{X} \mid \Sigma_2 = \Sigma_1\right\}.$$

Theorem

Consider $\min_{X\in\mathbb{R}^{n\times r}} \mathcal{F}(X)$, *where* $\mathcal{F}(X) = \frac{1}{4}\|M^* - XX^\top\|_F^2$ *for* $r \geq 1$. *Define*

$$\mathcal{R}_1 \triangleq \left\{Y \in \mathbb{R}^{n\times r} \mid \sigma_r(Y) \leq \tfrac{1}{2}\sigma_r(U),\ \|YY^\top\|_F \leq 4\|M^*\|_F\right\},$$
$$\mathcal{R}_2 \triangleq \left\{Y \in \mathbb{R}^{n\times r} \mid \min_{\Psi\in\mathfrak{O}_r}\|Y - U\Psi\|_2 \leq \frac{\sigma_r^2(U)}{8\sigma_1(U)}\right\},$$
$$\mathcal{R}_3' \triangleq \left\{Y \in \mathbb{R}^{n\times r} \mid \sigma_r(Y) > \tfrac{1}{2}\sigma_r(U),\ \min_{\Psi\in\mathfrak{O}_r}\|Y - U\Psi\|_2 > \frac{\sigma_r^2(U)}{8\sigma_1(U)},\ \|YY^\top\|_F \leq 4\|M^*\|_F\right\},\ \textit{and}$$
$$\mathcal{R}_3'' \triangleq \left\{Y \in \mathbb{R}^{n\times r} \mid \|YY^\top\|_F > 4\|M^*\|_F\right\}.$$

Low-Rank Matrix Factorization: Rank-r Case

Theorem (Continued...)

Then the following properties hold.

- $\forall X \in \mathcal{X}$, *X is a stationary point of* $\mathcal{F}(X)$.
- $\forall X \in \mathcal{X} \backslash \mathcal{U}$, *X is a strict saddle point with* $\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \leq -\lambda^2_{\max}(\Sigma_1 - \Sigma_2)$. *Moreover, for any* $X \in \mathcal{R}_1$, $\nabla^2 \mathcal{F}(X)$, $\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \leq -\frac{\sigma_r^2(U)}{4}$.
- $\forall X \in \mathcal{U}$, *X is a global minimum of* $\mathcal{F}(X)$ *with nonzero* $\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \geq \sigma_r^2(U)$ *($r(r-1)/2$ zero eigenvalues). Moreover,* $\forall X \in \mathcal{R}_2$, $z^\top \nabla^2 \mathcal{F}(X) z \geq \frac{1}{5}\sigma_r^2(U)||z||_2^2$, $\forall z \perp \mathcal{E}$, *where* $\mathcal{E} \subseteq \mathbb{R}^{n \times r}$ *is a subspace spanned by eigenvectors of* $\nabla^2 \mathcal{F}(K_E)$ *with negative eigenvalues,* $E = X - U\Psi_X$, *and* $K_E \triangleq \begin{bmatrix} E_{(*,1)}E_{(*,1)}^\top & E_{(*,2)}E_{(*,1)}^\top & \cdots & E_{(*,r)}E_{(*,1)}^\top \\ E_{(*,1)}E_{(*,2)}^\top & E_{(*,2)}E_{(*,2)}^\top & \cdots & E_{(*,r)}E_{(*,2)}^\top \\ \vdots & \vdots & \ddots & \vdots \\ E_{(*,1)}E_{(*,r)}^\top & E_{(*,2)}E_{(*,r)}^\top & \cdots & E_{(*,r)}E_{(*,r)}^\top \end{bmatrix}$.
- $\forall X \in \mathcal{R}_3'$, $||\nabla \mathcal{F}(X)||_F > \frac{\sigma_r^4(U)}{9\sigma_1(U)}$ *and* $\forall X \in \mathcal{R}_3''$, $\|\nabla \mathcal{F}(X)\|_F > \frac{3}{4}\sigma_1^3(X)$.

Geometric Interpretation

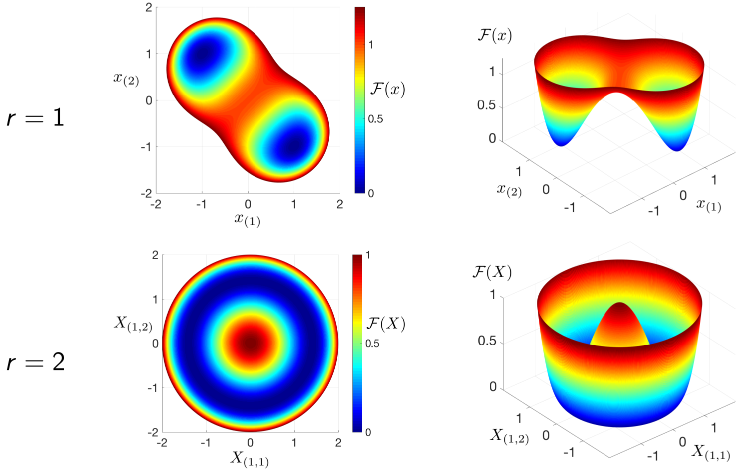


Figure: In the case $r = 1$, the true model is $u = [1 \ \ -1]^\top$. In the case $r = 2$, the true model is $U = [1 \ \ -1]$.

Extensions

- General Rectangular Matrix: we have $M^* = UV^\top$ and solve

$$\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} \mathcal{F}_\lambda(X, Y) = \frac{1}{8}||XY^\top - M^*||_F^2 + \frac{\lambda}{4}||X^\top X - Y^\top Y||_F^2$$

Extensions

➔ General Rectangular Matrix: we have $M^* = UV^\top$ and solve

$$\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} \mathcal{F}_\lambda(X,Y) = \frac{1}{8}\|XY^\top - M^*\|_F^2 + \frac{\lambda}{4}\|X^\top X - Y^\top Y\|_F^2$$

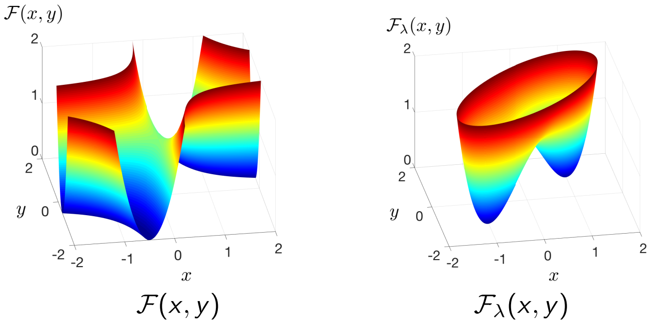


Figure: $r = 1$, the true model is $u = v = 1$.

Extensions

- General Rectangular Matrix: we have $M^* = UV^\top$ and solve

$$\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} \mathcal{F}_\lambda(X, Y) = \frac{1}{8}||XY^\top - M^*||_F^2 + \frac{\lambda}{4}||X^\top X - Y^\top Y||_F^2$$

- Matrix Sensing: we observe $y_{(i)} = \langle A_i, M^* \rangle + z_{(i)}$ for all $i \in [d]$, $\{z_{(i)}\}_{i=1}^d$ are noise, and solve

$$\min_X F(X) = \frac{1}{4d}\sum_{i=1}^d (y_i - \langle A_i, XX^\top \rangle)^2$$

Extensions

➔ General Rectangular Matrix: we have $M^* = UV^\top$ and solve

$$\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} \mathcal{F}_\lambda(X,Y) = \frac{1}{8}\|XY^\top - M^*\|_F^2 + \frac{\lambda}{4}\|X^\top X - Y^\top Y\|_F^2$$

➔ Matrix Sensing: we observe $y_{(i)} = \langle A_i, M^*\rangle + z_{(i)}$ for all $i \in [d]$, $\{z_{(i)}\}_{i=1}^d$ are noise, and solve

$$\min_X \; F(X) = \frac{1}{4d}\sum_{i=1}^d (y_i - \langle A_i, XX^\top\rangle)^2$$

➔ Matrix Completion ...

$\Longrightarrow$ Analogous geometric properties to those of low-rank matrix factorization.

Implication to Convergence Analysis

Direct result of convergence guarantees:

→ First order methods:

- Gradient descent: Asymptotic convergence guarantee of Q-linear convergence to a local minimum (Lee et al., 2016; Panageas and Piliouras, 2016)
- Noisy stochastic gradient descent: R-sublinear convergence to a local minimum (Ge et al., 2015)

Implication to Convergence Analysis

Direct result of convergence guarantees:

➔ First order methods:

- Gradient descent: Asymptotic convergence guarantee of Q-linear convergence to a local minimum (Lee et al., 2016; Panageas and Piliouras, 2016)
- Noisy stochastic gradient descent: R-sublinear convergence to a local minimum (Ge et al., 2015)

➔ Second order methods:

- Trust-region methods: R-quadratic convergence to a global minimum (Sun et al., 2016)
- Second-order majorization: Sublinear convergence guarantee (Carmon & Duchi, 2016)

Extension to Nonconvex Constrained Optimization

→ Consider the generalized eigenvalue decomposition (GEV) problem:

$$\min_{X \in \mathbb{R}^{d \times r}} \mathcal{F}(X) = -\operatorname{tr}(X^\top A X) \quad \text{subject to} \quad X^\top B X = I_r$$

- Apply the method of Lagrange multipliers,

$$\min_X \max_Y \mathcal{L}(X, Y) = -\operatorname{tr}(X^\top A X) + \langle Y, X^\top B X - I_r \rangle$$

- The gradient of Lagrangian function:

$$\nabla \mathcal{L} \triangleq \begin{bmatrix} \nabla_X \mathcal{L}(X, Y) \\ \nabla_Y \mathcal{L}(X, Y) \end{bmatrix} = \begin{bmatrix} 2BXY - 2AX \\ X^\top B X - I_r \end{bmatrix}.$$

- At a stationary point, the dual variable satisfies

$$Y = \mathcal{D}(X) \triangleq X^\top A X$$

Adaptation of Definition

Definition

Given the Lagrangian function $\mathcal{L}(X, Y)$, a pair of point (X, Y) is called:

- A **stationary point** of $\mathcal{L}(X, Y)$, if $\nabla\mathcal{L} = 0$
- An **unstable stationary point** of $\mathcal{L}(X, Y)$, if (X, Y) is a stationary point and for any neighborhood $\mathcal{B} \subseteq \mathbb{R}^{d\times r}$ of X, there exist $X_1, X_2 \in \mathcal{B}$ such that

$$\mathcal{L}(X_1, Y)|_{Y=\mathcal{D}(X_1)} \leq \mathcal{L}(X, Y)|_{Y=\mathcal{D}(X)} \leq \mathcal{L}(X_2, Y)|_{Y=\mathcal{D}(X_2)},$$

  and $\lambda_{\min}(\nabla_X^2\mathcal{L}(X, Y)|_{Y=\mathcal{D}(X)}) \leq 0$
- A **convex-concave saddle point**, or a **minimax point** of $\mathcal{L}(X, Y)$, if (X, Y) is a stationary point and (X, Y) is a global optimum, i.e.

$$(X, Y) = \arg\min_{\tilde{X}}\max_{\tilde{Y}} \mathcal{L}(\tilde{X}, \tilde{Y}).$$

Characterization of Stationary Point

- Consider nonsingular B:

Let the eigendecomposition be $B^{-1/2}AB^{-1/2} = O^{\dagger}\Lambda^{\dagger}(O^{\dagger})^{\top}$. Consider the following decomposition:

$$\mathcal{U}_{\mathcal{S}} = \left\{ U \in \mathbb{R}^{d\times s} : U = O^{\dagger}_{:,\mathcal{S}}, \mathcal{S} \subseteq [r] \text{ with } |\mathcal{S}| = s \leq r \right\},$$

$$\mathcal{V}_{\tilde{\mathcal{S}}} = \left\{ V \in \mathbb{R}^{d\times(r-s)} : V = O^{\dagger}_{:,\tilde{\mathcal{S}}}, \tilde{\mathcal{S}} \subseteq [d]\backslash[r] \text{ with } |\tilde{\mathcal{S}}| = r-s, |\mathcal{S}| = s \leq r \right\}.$$

Theorem (Symmetry Property)

Suppose that A and B are symmetric and B is nonsingular. Then $(X, \mathcal{D}(X))$ is a stationary is a stationary point of $\mathcal{L}(X, Y)$, i.e., $\nabla\mathcal{L} = 0$, if and only if $X = B^{-1/2}\tilde{X}$ for any $\tilde{X} \in \mathcal{G}_{\mathcal{U}_{\mathcal{S}}}(V)$ with any $V \in \mathcal{V}_{\tilde{\mathcal{S}}}$, where $\mathcal{G}_{\mathcal{U}_{\mathcal{S}}}(V) = \{g_{\mathcal{U}} : g_{\mathcal{U}_{\mathcal{S}}}(V) = g(U \oplus V), g \in \mathcal{G}, U \in \mathcal{U}_{\mathcal{S}}\}$.

Unstable Stationary vs. Saddle Point

The GEV problem reduces to

$$\tilde{X}^* = \underset{\tilde{X} \in \mathbb{R}^{d\times r}}{\mathrm{argmin}} \; -\mathrm{tr}(\tilde{X}^\top \tilde{A} \tilde{X}) \quad \text{s.t.} \quad \tilde{X}^\top \tilde{X} = I_r,$$

where $\tilde{X} = B^{1/2} X$ and $\tilde{A} = B^{-1/2} A B^{-1/2}$.

Lemma

Let $X = B^{-1/2}\tilde{X}$ for any $\tilde{X} \in \mathcal{G}_{\mathcal{U}_\mathcal{S}}(V)$ and any $V \in \mathcal{V}_{\tilde{\mathcal{S}}}$ with $\mathcal{S} \subseteq [r]$. If $\mathcal{S} = [r]$ and $\tilde{\mathcal{S}} = \emptyset$, then $(X, \mathcal{D}(X))$ is a saddle point of the min-max problem. Otherwise, if $\mathcal{S} \subset [r]$ and $\tilde{\mathcal{S}} \subseteq [d]\backslash[r]$, $\tilde{\mathcal{S}} \neq \emptyset$, with $|\mathcal{S}| + |\tilde{\mathcal{S}}| = r$, then $(X, \mathcal{D}(X))$ is an unstable stationary point with

$$\lambda_{\min}(H_X) \leq \frac{2\left(\lambda^\dagger_{\max \mathcal{S}\cup\tilde{\mathcal{S}}} - \lambda^\dagger_{\min \mathcal{S}^\perp\cap\tilde{\mathcal{S}}^\perp}\right)}{\|X_{:,\min \mathcal{S}^\perp\cap\tilde{\mathcal{S}}^\perp}\|_2^2} \quad \text{and} \quad \lambda_{\max}(H_X) \geq \frac{4\lambda^\dagger_{\min \mathcal{S}\cup\tilde{\mathcal{S}}}}{\|X_{:,\min \mathcal{S}\cup\tilde{\mathcal{S}}}\|_2^2},$$

where $\lambda^\dagger_{\max \mathcal{S}}$ ($\lambda^\dagger_{\min \mathcal{S}}$) is the smallest (largest) eigenvalue of $B^{-1/2}AB^{-1/2}$ indexed by a set $\mathcal{S}$.

Extension and Algorithm

- Extension to Singular B
  - Use generalized inverse, much more involved
- An asymptotic sublinear convergence of online optimization
  - Simple update: $X^{(k+1)} \leftarrow X^{(k)} - \eta \left(B^{(k)} X^{(k)} X^{(k)\top} - I_d \right) A^{(k)} X^{(k)}$
  - Characterization using stochastic differential equation (SDE)

Extension and Algorithm

- Extension to Singular B
  - Use generalized inverse, much more involved
- An asymptotic sublinear convergence of online optimization
  - Simple update: $X^{(k+1)} \leftarrow X^{(k)} - \eta \left(B^{(k)} X^{(k)} X^{(k)\top} - I_d \right) A^{(k)} X^{(k)}$
  - Characterization using stochastic differential equation (SDE)

Thank you !