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
Consider a low-rank matrix estimation problem:

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

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

- Fit **Wide** class of problems; **NP-hard** in general
Consider a low-rank matrix estimation problem:

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

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:

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

- **Easy** to analyze; Computationally **Expensive**, e.g., SVD
Consider a low-rank matrix estimation problem:

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

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

- 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)

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$, 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 $B \subseteq \mathbb{R}^n$ of $x$ such that $f(x) \leq f(y)$ for any $y \in B$;

(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 $B \subseteq \mathbb{R}^n$ of $x$, $\exists y, z \in B$ s.t. $f(z) \leq f(x) \leq f(y)$ & $\lambda_{\text{min}}(\nabla^2 f(x)) < 0$.

(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

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

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

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

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

**Corollary**

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

$$z^*(y_{\mathcal{G} Y}) \in \text{arg min}_z \nabla_z f(y_{\mathcal{G} Y} \oplus z),$$

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

→ Low-rank Matrix Factorization:

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

• Invariant group: \( \mathcal{O}_r = \{ \Psi \in \mathbb{R}^{r \times r} \mid \Psi \Psi^\top = \Psi^\top \Psi = I_r \} \); Fixed point: 0

• \( \mathcal{Y} = \mathcal{L}_{U_{r-s}} \subseteq \mathcal{L}_U \) and \( \mathcal{Z} = \mathcal{L}_{U_s} \subseteq \mathcal{L}_U \)

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

→ Low-rank Matrix Factorization:
\[
\min_X f(X) = \frac{1}{4} \|XX^\top - M^*\|_F^2, \text{ where } M^* = UU^\top
\]

- Invariant group: \( \mathcal{O}_r = \{ \Psi \in \mathbb{R}^{r \times r} \mid \Psi \Psi^\top = \Psi^\top \Psi = I_r \} \); 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 \( \Psi_r \in \mathcal{O}_r \), \( U_s = \Phi \Sigma S \Theta^\top \), \( U = \Phi \Sigma \Theta^\top \) (SVD), and \( S \) is a diagonal matrix w/ \( s \) entries 1 and 0 o.w. \( \forall s \in [r] \)

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

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

- Invariant group: \( \mathcal{G} = \{ e^{i\theta} \mid \theta \in [0, 2\pi) \} \); Fixed point: 0
- \( \mathcal{Y} = \{ y_i = 0, \forall i \in C \} \) and \( \mathcal{Z} = \{ z_i = 0, \forall i \in [n] \setminus C \} \), \( C \subseteq [n] \), \( |C| \leq n \)
  \( \Rightarrow x \) is stationary, if \( x^H u = 0 \), \( x_{\mathcal{Y}} = 0 \), \( \|x\|_2 = \|u\|_2 / \sqrt{2} \)
Examples

→ Low-rank Matrix Factorization:
\[
\min_X f(X) = \frac{1}{4} \| XX^\top - M^* \|_F^2, \text{ where } M^* = UU^\top
\]

- Invariant group: \( \mathcal{O}_r = \{ \Psi \in \mathbb{R}^{r \times r} \mid \Psi \Psi^\top = \Psi^\top \Psi = I_r \} \); Fixed point: 0

- \( \mathcal{Y} = \mathcal{L}_{U_{r-s}} \subseteq \mathcal{L}_U \) and \( \mathcal{Z} = \mathcal{L}_{U_s} \subseteq \mathcal{L}_U \)
  \[\Rightarrow U_s \Psi_r \text{ is stationary, where } \Psi_r \in \mathcal{O}_r, \ U_s = \Phi \Sigma S \Theta^\top, \ U = \Phi \Sigma \Theta^\top \] (SVD), and \( S \) is a diagonal matrix w/ \( s \) entries 1 and 0 o.w. \( \forall s \in [r] \)

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

  Expected objective: \( f(x) = \mathbb{E}(h(x)) = \|x\|_4^4 + \|u\|_2^4 - \|x\|_2^2 \|u\|_2^2 - |x^H u|^2 \)

- Invariant group: \( \mathcal{G} = \{ e^{i\theta} \mid \theta \in [0, 2\pi) \} \); Fixed point: 0

- \( \mathcal{Y} = \{ y_i = 0, \forall i \in C \} \) and \( \mathcal{Z} = \{ z_i = 0, \forall i \in [n] \setminus C \}, \ C \subseteq [n], \ |C| \leq n \)
  \[\Rightarrow x \text{ is stationary, if } x^H u = 0, \ x_{\mathcal{Y}} = 0, \ \|x\|_2 = \|u\|_2/\sqrt{2} \]

→ Deep Linear Neural Networks ...
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 $\gamma : \mathbb{R} \rightarrow \mathcal{M}$ 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

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

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

![Diagram of a tangent space](image)
**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 $G$ and $H_x$ is the Hessian matrix at a stationary point $x$, then we have*

$$T_x G(x) \subseteq \text{Null}(H_x).$$
→ **Low-rank Matrix Factorization:** Let $\gamma : \mathbb{R} \rightarrow \mathcal{O}_r(X)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}, \exists \Psi_r \in \mathcal{O}_r$ 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

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

→ Low-rank Matrix Factorization: Let $\gamma : \mathbb{R} \to \mathcal{O}_r(X)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}, \exists \Psi_r \in \mathcal{O}_r$ 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

$\Rightarrow T_x\mathcal{O}_r(X) = \{XE | E \in \mathbb{R}^{r \times r}, E = -E^T\}$, e.g., $U_s\Psi_rE \in \text{Null}(H_{U_s\Psi_r})$

→ 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 $\gamma(0) = x$

$\Rightarrow \|\gamma(t)\|_2^2 = \|x\|_2^2$

$\Rightarrow \gamma'(0)^Hx = -x^H\gamma'(0)$ by differentiation w.r.t. $t$

$\Rightarrow T_x\mathcal{G}(x) = ix$, e.g., $iue^{i\theta} \in \text{Null}(H_{ue^{i\theta}})$
→ Low-rank Matrix Factorization: Let $\gamma : \mathbb{R} \to \mathbb{O}_r(X)$ be a smooth curve, i.e., $\forall t \in \mathbb{R}, \exists \Psi_r \in \mathbb{O}_r$ 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

$\Rightarrow T_X\mathbb{O}_r(X) = \{XE | E \in \mathbb{R}^{r \times r}, E = -E^T\}$, e.g., $U^s\Psi_r E \in \text{Null}(H_{U^s\Psi_r})$

→ Phase Retrieval: Let $\gamma : \mathbb{R} \to \mathbb{G}(x)$ 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 $\gamma(0) = x$

$\Rightarrow \|\gamma(t)\|_2^2 = \|x\|_2^2$

$\Rightarrow \gamma'(0)^Hx = -x^H\gamma'(0)$ by differentiation w.r.t. $t$

$\Rightarrow T_x\mathbb{G}(x) = ix$, e.g., $iue^{i\theta} \in \text{Null}(H_{ue^{i\theta}})$

→ 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 $\lambda_{\text{min}}(\nabla^2 \mathcal{F}(X)) < 0$, denoted as $\mathcal{R}_1$

• Identify the global minimum, their neighborhood, and the directions such that $\lambda_{\text{min}}(\nabla^2 \mathcal{F}(X)) > 0$, denoted as $\mathcal{R}_2$

• Verify that the gradient has a sufficiently large norm outside the regions described in (p2) and (p3), denoted as $\mathcal{R}_3$

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

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

$$\mathcal{R}_2 \triangleq \{y \in \mathbb{R}^n | ||y - u||_2 \leq \frac{1}{8} ||u||_2\},$$

and

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

Then the following properties hold.

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

Low-Rank Matrix Factorization: Rank-r Case

Introduce two sets:

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

**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

\[ \mathcal{R}_1 \triangleq \{ Y \in \mathbb{R}^{n \times r} \mid \sigma_r(Y) \leq \frac{1}{2} \sigma_r(U), \| YY^\top \|_F \leq 4 \| M^* \|_F \} , \]
\[ \mathcal{R}_2 \triangleq \{ Y \in \mathbb{R}^{n \times r} \mid \min_{\Psi \in \mathcal{O}_r} \| Y - U \Psi \|_2 \leq \frac{\sigma_r^2(U)}{8 \sigma_1(U)} \} , \]
\[ \mathcal{R}_3' \triangleq \{ Y \in \mathbb{R}^{n \times r} \mid \sigma_r(Y) > \frac{1}{2} \sigma_r(U), \min_{\Psi \in \mathcal{O}_r} \| Y - U \Psi \|_2 > \frac{\sigma_r^2(U)}{8 \sigma_1(U)}, \| YY^\top \|_F \leq 4 \| M^* \|_F \} , \text{ and} \]
\[ \mathcal{R}_3'' \triangleq \{ Y \in \mathbb{R}^{n \times r} \mid \| YY^\top \|_F > 4 \| M^* \|_F \} . \]
Then the following properties hold.

- ∀\(X \in \mathcal{X}\), \(X\) is a stationary point of \(\mathcal{F}(X)\).
- ∀\(X \in \mathcal{X}\setminus \mathcal{U}\), \(X\) is a strict saddle point with \(\lambda_{\text{min}}(\nabla^2 \mathcal{F}(X)) \leq -\lambda_{\text{max}}^2(\Sigma_1 - \Sigma_2)\). Moreover, for any \(X \in \mathcal{R}_1\), \(\nabla^2 \mathcal{F}(X)\), \(\lambda_{\text{min}}(\nabla^2 \mathcal{F}(X)) \leq -\frac{\sigma^2_r(U)}{4}\).
- ∀\(X \in \mathcal{U}\), \(X\) is a global minimum of \(\mathcal{F}(X)\) with nonzero \(\lambda_{\text{min}}(\nabla^2 \mathcal{F}(X)) \geq \sigma^2_r(U)\) \((r(r-1)/2\) zero eigenvalues). Moreover, ∀\(X \in \mathcal{R}_2\), \(z^\top \nabla^2 \mathcal{F}(X)z \geq \frac{1}{5} \sigma^2_r(U)\|z\|^2_2\), ∀\(z \perp \mathcal{E}\), 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 \(K_E = \begin{bmatrix} E_{(1,1)}^\top E_{(1,1)}^\top & E_{(1,2)}^\top E_{(1,2)}^\top & \cdots & E_{(r,1)}^\top E_{(r,1)}^\top \\ E_{(1,2)}^\top E_{(1,2)}^\top & E_{(2,2)}^\top E_{(2,2)}^\top & \cdots & E_{(r,2)}^\top E_{(r,2)}^\top \\ \vdots & \vdots & \ddots & \vdots \\ E_{(1,r)}^\top E_{(1,r)}^\top & E_{(2,r)}^\top E_{(2,r)}^\top & \cdots & E_{(r,r)}^\top E_{(r,r)}^\top \end{bmatrix}\).
- ∀\(X \in \mathcal{R}_3'\), \(\|\nabla \mathcal{F}(X)\|_F > \frac{\sigma^4_r(U)}{9\sigma_1(U)}\) and ∀\(X \in \mathcal{R}_3''\), \(\|\nabla \mathcal{F}(X)\|_F > \frac{3}{4} \sigma^3_1(X)\).
Geometric Interpretation

$r = 1$

$r = 2$

**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

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

- 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 ...

⇒ 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)
Consider the generalized eigenvalue decomposition (GEV) problem:

\[
\min_{X \in \mathbb{R}^{d \times r}} \mathcal{F}(X) = -\text{tr}(X^\top AX) \quad \text{subject to} \quad X^\top BX = I_r
\]

- Apply the method of Lagrange multipliers,

\[
\min_X \max_Y \mathcal{L}(X, Y) = -\text{tr}(X^\top AX) + \langle Y, X^\top BX - I_r \rangle
\]

- The gradient of Lagrangian function:

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

- At a stationary point, the dual variable satisfies

\[
Y = \mathcal{D}(X) \triangleq X^\top AX
\]
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 $B \subseteq \mathbb{R}^{d \times r}$ of $X$, there exist $X_1, X_2 \in B$ such that

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

and $\lambda_{\min} \left( \nabla^2_X \mathcal{L}(X, Y)|_{Y = D(X)} \right) \leq 0$

- 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.

\[
(X, Y) = \arg \min_{\tilde{X}} \max_{\tilde{Y}} \mathcal{L}(\tilde{X}, \tilde{Y}).
\]
Consider nonsingular $B$:

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

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

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

**Theorem (Symmetry Property)**

Suppose that $A$ and $B$ are symmetric and $B$ is nonsingular. Then $(X, D(X))$ 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 $\tilde{X} \in \mathcal{G}_{\mathcal{U}_S}(V)$ with any $V \in \mathcal{V}_{\tilde{S}}$, where $\mathcal{G}_{\mathcal{U}_S}(V) = \{ g_U : g_{\mathcal{U}_S}(V) = g(U \oplus V), g \in \mathcal{G}, U \in \mathcal{U}_S \}$.
The GEV problem reduces to

$$
\tilde{X}^* = \arg\min_{\tilde{X} \in \mathbb{R}^{d \times r}} \quad -\text{tr}(\tilde{X}^T \tilde{A} \tilde{X}) \quad \text{s.t.} \quad \tilde{X}^T \tilde{X} = I_r,
$$

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

**Lemma**

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

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

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

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!