Euclidean Distance Matrix Completion has No Spurious Local Minimum

Sudeep Raja, Arun Rajkumar

Abstract

The Euclidean Distance Matrix(EDM) of a set of points is the matrix whose entries are the squared Euclidean distances between pairs of points. EDMs find applications in areas such as Machine Learning, Sensor Networks, Psychometrics, Molecular Conformations and many more. In this paper, we address the problem of recovering a point configuration in a specified dimension, given only a random subset of the (noisy) distance measurements between points. This is known as the EDM Completion Problem(EDMCP). We consider minimizing the position based non-convex objective function for EDMCP. We prove that 1) all local minima for this objective are also globally optimal; 2) no high order saddle points exists. Hence local search methods like (stochastic) Gradient Descent can find a point configuration, even with random initialization.

1 Introduction

EDMs find applications in a variety of areas. Some of these include localization of a wireless sensor networks, determining conformation of molecules in chemistry and proteins in bioinformatics, and nonlinear dimensionality reduction in statistics and machine learning. In several of these applications, it is cumbersome to measure the distances between all pairs of points. Given the distances between random pairs of points, the goal of EDMCP is to determine the missing distances and additionally determine the set of locations for the points in a given dimension.

Over the years, several approaches for solving the EDMCP have been developed. It is typically posed as an optimization problem. We focus on the non-convex position based formulation. In recent years, non-convex optimization has emerged as a very powerful tool in Machine Learning. Several inherently non-convex problems such as tensor decomposition, dictionary learning, matrix sensing, matrix completion and robust PCA have been shown to possess a well behaved optimization landscape: all local optima are also globally optimal. Such problems can be efficiently solved by basic optimization algorithms such as stochastic gradient descent. We show that the non-convex objective function of EDMCP also has a similar property.

2 Preliminaries

2.1 Notation

For a vector $\mathbf{v}$, $\|\mathbf{v}\|$ denotes its ℓ_2 norm. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|$ denotes its spectral norm and $\|\mathbf{M}\|_F$ denotes its Frobenius norm. $\langle \mathbf{u}, \mathbf{v} \rangle$ denotes the inner-product of vectors and for matrices $\langle \mathbf{M}, \mathbf{N} \rangle = \text{tr}(\mathbf{M}\mathbf{N}^\top) = \sum_{i,j} \mathbf{M}_{ij}\mathbf{N}_{ij}$. Let $\mathbf{M} : \mathcal{H} : \mathbf{N} = \langle \mathbf{M}, \mathcal{H} \circ \mathbf{N} \rangle = \sum_{i,j} \mathbf{M}_{ij}\mathcal{H}_{ij}\mathbf{N}_{ij}$.

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \in \mathbb{R}^{n \times k}$ be a set of n points in k dimensions. The squared Euclidean distance between the two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is d_{ij}

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \mathbf{x}_i^\top \mathbf{x}_i + \mathbf{x}_j^\top \mathbf{x}_j - 2\mathbf{x}_i^\top \mathbf{x}_j$$

Let $\mathbf{X}$ be the $n \times k$ matrix where $\mathbf{x}_i$ is the ith row. Let $\mathbf{D}$ be the Euclidean Distance Matrix of these points. The entry of $\mathbf{D}$ at location (i, j) is $d_{i,j}$. $\mathbf{D}$ can be computed using the following equation.

$$\mathbf{D} = (\text{diag}(\mathbf{X}\mathbf{X}^\top))\mathbf{e}^\top + \mathbf{e}(\text{diag}(\mathbf{X}\mathbf{X}^\top))^\top - 2\mathbf{X}\mathbf{X}^\top$$

Here $\operatorname{diag}(\mathbf{A})$ is the diagonal vector of $\mathbf{A}$ and $\mathbf{e}$ is the vector of size n consisting of all ones. Let $\mathcal{K}(\mathbf{A}) = (\operatorname{diag}(\mathbf{A}))\mathbf{e}^\top + \mathbf{e}(\operatorname{diag}(\mathbf{A}))^\top - \mathbf{A} - \mathbf{A}^\top$. So $\mathbf{D} = \mathcal{K}(\mathbf{X}\mathbf{X}^\top)$.

Let $\mathbf{D}^\star$ be the EDM of unknown point matrix $\mathbf{X}^\star \in \mathbb{R}^{n\times k}$, i.e $\mathbf{D}^\star = \mathcal{K}(\mathbf{X}^\star\mathbf{X}^{\star\top})$. Let $\Omega \subseteq [n]\times[n]$ be the entries of $\mathbf{D}^\star$ which are observed. Since $\mathbf{D}^\star$ is a zero diagonal symmetric matrix, $(i,i) \notin \Omega$ and $(i,j) \in \Omega \iff (j,i) \in \Omega$ for all $i \neq j$ and $i,j \in [n]$. For any matrix $\mathbf{M}$, let $\mathbf{M}_\Omega$ be the matrix whose entries outside of Ω are set to 0. Let $\mathcal{H} = (\mathbf{e}\mathbf{e}^\top)_\Omega$.

The objective is to find $\mathbf{X} \in \mathbb{R}^{n\times k}$ such that the following function is minimized.

$$f(\mathbf{X}) = \frac{1}{2}\|(\mathcal{K}(\mathbf{X}\mathbf{X}^\top) - \mathbf{D}^\star)_\Omega\|_F^2 = \frac{1}{2}(\mathcal{K}(\mathbf{X}\mathbf{X}^\top) - \mathbf{D}^\star) : \mathcal{H} : (\mathcal{K}(\mathbf{X}\mathbf{X}^\top) - \mathbf{D}^\star) \tag{1}$$

Let $\mathbf{E} = (\mathcal{K}(\mathbf{X}\mathbf{X}^\top) - \mathbf{D})_\Omega$ and $\mathbf{1_{n\times k}}$ be a matrix of all ones.

We propose the following gradient descent procedure.

Algorithm 1 Gradient Descent for EDM completion

```
Initialize $\mathbf{X}_0$ randomly
for Each iteration $t$ do
    $\mathbf{X}_t = \mathbf{X}_{t-1} - \eta((\mathbf{E}\mathbf{1_{n\times k}}) \circ \mathbf{X}_{t-1} - \mathbf{E}\mathbf{X}_{t-1})$
end for
```

We show that every local minima is a also a global minima and no higher order saddle points exist, by following the Proof Technique of [1] or [2].

3 Proofs

3.1 First and Second order optimality

$$\begin{aligned}
f(\mathbf{X}) &= \frac{1}{2}\|(\mathcal{K}(\mathbf{X}\mathbf{X}^\top) - \mathbf{D}^\star)_\Omega\|_F^2 \\
f(\mathbf{X}+\mathbf{Z}) &= \frac{1}{2}\|(\mathcal{K}((\mathbf{X}+\mathbf{Z})(\mathbf{X}+\mathbf{Z})^\top) - \mathbf{D}^\star)_\Omega\|_F^2 \\
&= \frac{1}{2}\|(\mathcal{K}(\mathbf{X}\mathbf{X}^\top + \mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top + \mathbf{Z}\mathbf{Z}^\top) - \mathbf{D}^\star)_\Omega\|_F^2 \\
&= \frac{1}{2}\|(\mathbf{D} - \mathbf{D}^\star)_\Omega + (\mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top) + \mathcal{K}(\mathbf{Z}\mathbf{Z}^\top))_\Omega\|_F^2 \\
&= \frac{1}{2}\|(\mathbf{D} - \mathbf{D}^\star)_\Omega\|_F^2 \\
&\quad + \langle(\mathbf{D} - \mathbf{D}^\star)_\Omega, \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top)\rangle \\
&\quad + \langle(\mathbf{D} - \mathbf{D}^\star)_\Omega, \mathcal{K}(\mathbf{Z}\mathbf{Z}^\top)\rangle + \frac{1}{2}\|(\mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top))_\Omega\|_F^2 + o(\|\mathbf{Z}\|^2)
\end{aligned}$$

By Taylor's expansion,

$$f(\mathbf{X}+\mathbf{Z}) = f(\mathbf{X}) + \langle\nabla f(\mathbf{X}), \mathbf{Z}\rangle + \frac{1}{2}(\mathbf{Z} : \nabla^2 f(\mathbf{X}) : \mathbf{Z}) + o(\|\mathbf{Z}\|^2)$$

Comparing the corresponding parts of same degree, we get:

$$\begin{aligned}
\langle\nabla f(\mathbf{X}), \mathbf{Z}\rangle &= \langle(\mathbf{D} - \mathbf{D}^\star)_\Omega, \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top)\rangle \\
&= (\mathbf{D} - \mathbf{D}^\star) : \mathcal{H} : \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top) \\
\mathbf{Z} : \nabla^2 f(\mathbf{X}) : \mathbf{Z} &= 2\langle(\mathbf{D} - \mathbf{D}^\star)_\Omega, \mathcal{K}(\mathbf{Z}\mathbf{Z}^\top)\rangle + \|(\mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top))_\Omega\|_F^2 \\
&= 2(\mathbf{D} - \mathbf{D}^\star) : \mathcal{H} : \mathcal{K}(\mathbf{Z}\mathbf{Z}^\top) + \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top) : \mathcal{H} : \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top)
\end{aligned}$$

3.2 Projection

Consider the $n \times n$ symmetric matrix J given by

$$\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{e}\mathbf{e}^\top$$

$\mathbf{J}$ is known as the centering matrix. The product $\mathbf{JX}$ centers the set of points around the origin. This does not change the distances between the points, so $\mathcal{K}(\mathbf{X}\mathbf{X}^\top) = \mathcal{K}(\mathbf{J}\mathbf{X}\mathbf{X}^\top\mathbf{J})$. Let $\mathbf{M} = \mathbf{J}\mathbf{X}\mathbf{X}^\top\mathbf{J}$ and $\mathbf{M}^\star = \mathbf{J}\mathbf{X}^\star\mathbf{X}^{\star\top}\mathbf{J}$ be the centered gram matrices of $\mathbf{X}$ and $\mathbf{X}^\star$. A few crucial relationships are:

$$\mathcal{K}(\mathbf{M}) = (\text{diag}(\mathbf{M}))\mathbf{e}^\top + \mathbf{e}(\text{diag}(\mathbf{M}))^\top - 2\mathbf{M} = D$$

$$\mathcal{T}(\mathbf{D}) = -\frac{1}{2}\mathbf{J}\mathbf{D}\mathbf{J} = \mathbf{M}$$

Definition 1. Given the matrices $\mathbf{X}, \mathbf{X}^\star$, define their difference as $\Delta = \mathbf{JX} - \mathbf{J}\mathbf{X}^\star\mathbf{R}_\mathbf{X}$, where $\mathbf{R}_\mathbf{X} \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{R}_\mathbf{X} = \underset{\mathbf{R}\mathbf{R}^\top = \mathbf{R}^\top\mathbf{R} = \mathbf{I}}{\text{argmin}} \|\mathbf{JX} - \mathbf{J}\mathbf{X}^\star\mathbf{R}\|_F^2$$

This definition centers the matrices $\mathbf{X}$ and $\mathbf{X}^\star$, and tries to align them before taking their difference. As long as $\mathcal{K}(\mathbf{X}\mathbf{X}^\top)$ is close to $\mathcal{K}(\mathbf{X}^\star\mathbf{X}^{\star\top})$, we have a small Δ.

Note that $\mathbf{JJ} = \mathbf{J}$. So, $\mathbf{J}\Delta = \Delta$.

3.3 What needs to be proved

We say function $f(\cdot)$ is (θ, γ, ζ)-**strict saddle**. That is, for any x, at least one of followings holds:

1. $\|\nabla f(x)\| \geq \theta$.
2. $\lambda_{\min}(\nabla^2 f(x)) \leq -\gamma$.
3. x is ζ-close to $\mathcal{X}^\star$ – the set of local minima.

We need to show that $f(X)$ has the strict saddle property. This can be done by showing that when X is not close to $X^\star$ (i.e. $\|\Delta\|_F^2 \geq \zeta$), we have $\Delta : \nabla^2 f(X) : \Delta \leq -\gamma\|\Delta\|_F^2$. This proves $f(X)$ is a $(\epsilon, \gamma, \zeta)$-strict saddle. Take $\epsilon = 0$. We know all stationary points with $\|\Delta\|_F \neq 0$ are saddle points. This means all local minima are global minima.

References

[1] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[2] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

A Auxiliary Inequalities

SAME as Appendix F of [2]. In this section, we provide some frequently used lemmas regarding matrices. Our first two lemmas lower bound $\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2$ by $\|(\mathbf{U}-\mathbf{Y})(\mathbf{U}-\mathbf{Y})^\top\|_F^2$ and $\|\mathbf{U}-\mathbf{Y}\|_\text{F}^2$.

Lemma 1. *Let* $\mathbf{U}$ *and* $\mathbf{Y}$ *be two* $n \times k$ *matrices. Further let* $\mathbf{U}^\top\mathbf{Y} = \mathbf{Y}^\top\mathbf{U}$ *be a PSD matrix. Then,*

$$\|(\mathbf{U}-\mathbf{Y})(\mathbf{U}-\mathbf{Y})^\top\|_F^2 \leq 2\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2$$

Proof. To prove this, we let $\Delta = \mathbf{U} - \mathbf{Y}$, and expand:

$$\begin{aligned}
\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2 =& \|\mathbf{U}\Delta^\top + \Delta\mathbf{U}^\top - \Delta\Delta^\top\|_F^2 \\
=& \text{tr}(2\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta + (\Delta^\top\Delta)^2 + 2(\mathbf{U}^\top\Delta)^2 - 4\mathbf{U}^\top\Delta\Delta^\top\Delta) \\
=& \text{tr}(2\mathbf{U}^\top(\mathbf{U} - \Delta)\Delta^\top\Delta + (\frac{1}{\sqrt{2}}\Delta^\top\Delta - \sqrt{2}\mathbf{U}^\top\Delta)^2 + \frac{1}{2}(\Delta^\top\Delta)^2) \\
\geq& \text{tr}(2\mathbf{U}^\top\mathbf{Y}\Delta^\top\Delta + \frac{1}{2}(\Delta^\top\Delta)^2) \geq \frac{1}{2}\|\Delta\Delta^\top\|_F^2
\end{aligned}$$

The last inequality is due to $\mathbf{U}^\top\mathbf{Y}$ is a PSD matrix. □

Lemma 2. *Let $\mathbf{U}$ and $\mathbf{Y}$ be two $n \times k$ matrices. Further let $\mathbf{U}^\top\mathbf{Y} = \mathbf{Y}^\top\mathbf{U}$ be a PSD matrix. Then,*

$$\sigma_{\min}(\mathbf{U}^\top\mathbf{U})\|\mathbf{U} - \mathbf{Y}\|_F^2 \leq \|(\mathbf{U} - \mathbf{Y})\mathbf{U}^\top\|_F^2 \leq \frac{1}{2(\sqrt{2}-1)}\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2$$

Proof. The left inequality is basic, we only need to prove right inequality. To prove this, we let $\Delta = \mathbf{U} - \mathbf{Y}$, and expand:

$$\begin{aligned}
\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2 =& \|\mathbf{U}\Delta^\top + \Delta\mathbf{U}^\top - \Delta\Delta^\top\|_F^2 \\
=& \text{tr}(2\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta + (\Delta^\top\Delta)^2 + 2(\mathbf{U}^\top\Delta)^2 - 4\mathbf{U}^\top\Delta\Delta^\top\Delta) \\
=& \text{tr}((4 - 2\sqrt{2})\mathbf{U}^\top(\mathbf{U} - \Delta)\Delta^\top\Delta + (\Delta^\top\Delta - \sqrt{2}\mathbf{U}^\top\Delta)^2 + 2(\sqrt{2} - 1)\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta) \\
\geq& \text{tr}((4 - 2\sqrt{2})\mathbf{U}^\top\mathbf{Y}\Delta^\top\Delta + 2(\sqrt{2} - 1)\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta) \geq 2(\sqrt{2} - 1)\|\mathbf{U}\Delta^\top\|_{\mathrm{F}}^2
\end{aligned}$$

The last inequality is due to $\mathbf{U}^\top\mathbf{Y}$ is a PSD matrix. □

Lemma 3. *Let $\mathbf{A}$ be a $n \times n$ transform matrix and $\mathbf{x}$ be a n dimensional vector. Then*

$$\|\mathbf{A}\mathbf{x}\|_2^2 \leq \sigma_{\max}(\mathbf{A}^\top\mathbf{A})\|\mathbf{x}\|_2^2 \tag{2}$$

Proof. Let the rank of $\mathbf{A}$ be r and its SVD be $\mathbf{U}\mathbf{S}\mathbf{V}^T$. Here $\mathbf{U}$ and $\mathbf{V}$ are $n \times r$ unitary matrices, and $\mathbf{S}$ is the $r \times r$ diagonal matrix containing the singular values of $\mathbf{A}$. We have, $\mathbf{A}\mathbf{x} = \sum_{i=1}^r S_{ii}\mathbf{u}_i\mathbf{v}_i^\top\mathbf{x}$. So,

$$\|\sum_{i=1}^r S_{ii}\mathbf{u}_i\mathbf{v}_i^\top\mathbf{x}\|_2^2 \leq \max_i(S_{ii})^2\|x\|_2^2$$

$$\|\mathbf{A}\mathbf{x}\|_2^2 \leq \sigma_{\max}(\mathbf{A}^\top\mathbf{A})\|\mathbf{x}\|_2^2$$

□

Lemma 4. *Let $\mathbf{M}$ be a $n \times n$ matrix. Then*

$$\|\mathcal{K}(\mathbf{M})\|_F^2 \leq 4n\|\mathbf{M}\|_F^2 \tag{3}$$

Proof. Let $\mathbf{B}$ be the $n^2 \times n^2$ transform matrix corresponding to the linear function $\mathcal{K}$. The rank of B is $n(n-1)/2$ and its non-zero singular values are 2, $\sqrt{2n}$, $\sqrt{4n}$. We have $\|\text{vec}(M)\|_2^2 = \|M\|_F^2$. Using the above lemma, we have,

$$\|\mathcal{K}(\mathbf{M})\|_F^2 = \|\mathbf{B}\text{vec}(M)\|_2^2 \leq 4n\|\text{vec}(\mathbf{M})\|_2^2 = 4n\|\mathbf{M}\|_F^2$$

□

Lemma 5. *Let $\mathbf{M}$ be a $n \times n$ PSD matrix. Then*

$$\|\mathcal{K}(\mathbf{M})\|_F^2 \leq \frac{2n^2}{n-1}\|\mathbf{M}\|_F^2 \tag{4}$$

Lemma 6. *Let* $\mathbf{D}$ *be a* $n \times n$ *matrix. Then*

$$\|\mathcal{T}(\mathbf{D})\|_F^2 \le \frac{1}{2}\|\mathbf{D}\|_F^2 \tag{5}$$

Proof. Let $\mathbf{B}$ be the $n^2 \times n^2$ transform matrix corresponding to the linear function $\mathcal{T}$. The rank of B is $n(n-1)$ and its non-zero singular values are $1/2, 1/\sqrt{2}$. We have $\|vec(D)\|_2^2 = \|D\|_F^2$. Using the above lemma, we have,

$$\|\mathcal{T}(\mathbf{D})\|_F^2 = \|\mathbf{B}\text{vec}(D)\|_2^2 \le \frac{1}{2}\|\text{vec}(\mathbf{D})\|_F^2 = \frac{1}{2}\|\mathbf{D}\|_F^2$$

□

Lemma 7. *Let* $\mathbf{D}$ *be a* $n \times n$ *matrix. Then*

$$\|\mathcal{T}(\mathbf{D})\|_F^2 \le \frac{1}{4}\|\mathbf{D}\|_F^2 \tag{6}$$

Proof. We know that $\|J\|_2 = 1$. So,

$$\begin{aligned}\|\mathcal{T}(\mathbf{D})\|_F^2 &= \frac{1}{4}\|\mathbf{J}\mathbf{D}\mathbf{J}\|_F^2 \le \frac{1}{4}\|\mathbf{J}\|_2^4\|\mathbf{D}\|_F^2 \\ &\le \frac{1}{4}\|\mathbf{D}\|_F^2\end{aligned}$$

□

B Symmetric PSD matrix completion

Lemma 8. *Given matrices* $\mathbf{U}, \mathbf{U}^\star \in \mathbb{R}^{d\times r}$*, let* $\mathbf{M} = \mathbf{U}\mathbf{U}^\top$*,* $\mathbf{M}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\top$ *and* $\sigma_r^\star$ *is the smallest singular value of* $\mathbf{M}^*$*, and let* Δ *be defined as in Definition 1, then we have* $\|\Delta\Delta^\top\|_F^2 \le 2\|\mathbf{M}-\mathbf{M}^\star\|_F^2$*, and* $\sigma_r^\star\|\Delta\|_F^2 \le \frac{1}{2(\sqrt{2}-1)}\|\mathbf{M}-\mathbf{M}^\star\|_F^2$.

Lemma 9 (Main)**.** *For the objective* (1)*, let* Δ *be defined as in Definition 1 and* $\mathbf{M} = \mathbf{U}\mathbf{U}^\top$*. Then, for any* $\mathbf{U} \in \mathbb{R}^{d\times r}$*, we have*

$$\begin{aligned}\Delta : \nabla^2 f(\mathbf{U}) : \Delta =& \Delta\Delta^\top : \mathcal{H} : \Delta\Delta^\top - 3(\mathbf{M}-\mathbf{M}^\star) : \mathcal{H} : (\mathbf{M}-\mathbf{M}^\star) \\ &+ 4\langle\nabla f(\mathbf{U}), \Delta\rangle + [\Delta : \nabla^2 Q(\mathbf{U}) : \Delta - 4\langle\nabla Q(\mathbf{U}), \Delta\rangle]\end{aligned}$$

Lemma 10. *When* $p = 1$*, we have* 1) *all local minima satisfy* $\mathbf{U}\mathbf{U}^\top = M$ *and* 2) *the function is a* $(\epsilon, (\sqrt{2}-1)\sigma_r^\star, \frac{4\epsilon}{(\sqrt{2}-1)\sigma_r^\star})$*-strict saddle.*

Proof.

$$\begin{aligned}\Delta : \nabla^2 f(\mathbf{U}) : \Delta &= \Delta\Delta^\top : \mathcal{H} : \Delta\Delta^\top - 3(\mathbf{M}-\mathbf{M}^\star) : \mathcal{H} : (\mathbf{M}-\mathbf{M}^\star) + 4\langle\nabla f(\mathbf{U}), \Delta\rangle \\ &= \|\Delta\Delta^\top\|_F^2 - 3\|\mathbf{M}-\mathbf{M}^\star\|_F^2 + 4\langle\nabla f(\mathbf{U}), \Delta\rangle \\ &\le 2\|\mathbf{M}-\mathbf{M}^\star\|_F^2 - 3\|\mathbf{M}-\mathbf{M}^\star\|_F^2 + 4\langle\nabla f(\mathbf{U}), \Delta\rangle \\ &\le -1\|\mathbf{M}-\mathbf{M}^\star\|_F^2 + 4\langle\nabla f(\mathbf{U}), \Delta\rangle \\ &\le -2(\sqrt{2}-1)\sigma_r^\star\|\Delta\|_F^2 + 4\epsilon\|\Delta\|_F\end{aligned}$$

When $\mathbf{U}$ is not close to $\mathbf{U}^\star$ (i.e. if $\|\Delta\| \ge \frac{4\epsilon}{(\sqrt{2}-1)\sigma_r^\star}$), then $\Delta : \nabla^2 f(\mathbf{U}) : \Delta \le -(\sqrt{2}-1)\sigma_r^\star\|\Delta\|_F^2$. This proves $(\epsilon, (\sqrt{2}-1)\sigma_r^\star, \frac{4\epsilon}{(\sqrt{2}-1)\sigma_r^\star})$-strict saddle property. Take $\epsilon = 0$, we know all stationary points with $\|\Delta\|_F \ne 0$ are saddle points. This means, all local minima are global minima satisfying $\mathbf{U}\mathbf{U}^\top = \mathbf{M}^\star$. □

C Proofs for EDMC

First we prove a lemma which connects difference in the matrix $\mathbf{X}$ and the difference in the matrix $\mathcal{G}$.

Lemma 11. *Given matrices $\mathbf{X}, \mathbf{X}^\star \in \mathbb{R}^{n\times k}$, let $\mathbf{M} = \mathbf{J}\mathbf{X}\mathbf{X}^\top\mathbf{J}$, $\mathbf{M}^\star = \mathbf{J}\mathbf{X}^\star(\mathbf{X}^\star)^\top\mathbf{J}$ and $\sigma_r^\star$ is the smallest singular value of $\mathbf{M}^*$, and let Δ be defined as in Definition 1, then we have $\|\Delta\Delta^\top\|_F^2 \le 2\|\mathbf{M}-\mathbf{M}^\star\|_F^2$, and $\sigma_r^\star\|\Delta\|_F^2 \le \frac{1}{2(\sqrt{2}-1)}\|\mathbf{M}-\mathbf{M}^\star\|_F^2$.*

Proof. Recall in Definition 1, $\Delta = \mathbf{J}\mathbf{X} - \mathbf{J}\mathbf{X}^\star\mathbf{R}_\mathbf{X}$ where

$$\mathbf{R}_\mathbf{X} = \underset{\mathbf{R}^\top\mathbf{R}=\mathbf{R}\mathbf{R}^\top=\mathbf{I}}{\operatorname{argmin}} \|\mathbf{J}\mathbf{X} - \mathbf{J}\mathbf{X}^\star\mathbf{R}\|_\mathrm{F}^2.$$

We first prove following claim, which will be used in many places across this proof:

$$\mathbf{X}^\top\mathbf{J}\mathbf{J}\mathbf{X}^\star\mathbf{R}_\mathbf{X} \text{ is a symmetric PSD matrix.} \tag{7}$$

This because by expanding the Frobenius norm, and letting the SVD of $\mathbf{X}^{\star\top}\mathbf{J}\mathbf{J}\mathbf{X}$ be $\mathbf{U}\mathbf{S}\mathbf{V}^\top$, we have:

$$\begin{aligned}\underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\operatorname{argmin}} \|\mathbf{J}\mathbf{X} - \mathbf{J}\mathbf{X}^\star\mathbf{R}\|_\mathrm{F}^2 &= \underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\operatorname{argmin}} -\langle \mathbf{J}\mathbf{X}, \mathbf{J}\mathbf{X}^\star\mathbf{R}\rangle \\ = \underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\operatorname{argmin}} -\mathrm{tr}(\mathbf{X}^\top\mathbf{J}\mathbf{J}\mathbf{X}^\star\mathbf{R}) &= \underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\operatorname{argmin}} -\mathrm{tr}(\mathbf{S}\mathbf{U}^\top\mathbf{R}\mathbf{V})\end{aligned}$$

Since $\mathbf{U}, \mathbf{V}, \mathbf{R} \in \mathbb{R}^{k\times k}$ are all orthonormal matrix, we know $\mathbf{U}^\top\mathbf{R}\mathbf{V}$ is also orthonormal matrix. Moreover for any orthonormal matrix $\mathbf{T}$, we have:

$$\mathrm{tr}(\mathbf{S}\mathbf{T}) = \sum_i \mathbf{S}_{ii}\mathbf{T}_{ii} \le \sum_i \mathbf{S}_{ii}$$

The last inequality is because $\mathbf{S}_{ii}$ is singular value thus non-negative, and $\mathbf{T}$ is orthonormal, thus $\mathbf{T}_{ii} \le 1$. This means the maximum of $\mathrm{tr}(\mathbf{S}\mathbf{T})$ is achieved when $\mathbf{T} = \mathbf{I}$, i.e., the minimum of $-\mathrm{tr}(\mathbf{S}\mathbf{U}^\top\mathbf{R}\mathbf{V})$ is achieved when $\mathbf{R} = \mathbf{U}\mathbf{V}^\top$. Therefore, $\mathbf{X}^\top\mathbf{J}\mathbf{J}\mathbf{X}^\star\mathbf{R}_\mathbf{X} = \mathbf{V}\mathbf{S}\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top = \mathbf{V}\mathbf{S}\mathbf{V}^\top$ is symmetric PSD matrix.

With Eq.(7), the remaining of proof directly follows from the results by substituting $(\mathbf{U}, \mathbf{Y})$ in Lemma 1 and 2 with $(\mathbf{J}\mathbf{X}^\star\mathbf{R}_\mathbf{X}, \mathbf{J}\mathbf{X})$. □

Lemma 12 (Main). *For the objective* (1)*, let Δ be defined as in Definition 1, $\mathbf{M} = \mathbf{J}\mathbf{X}\mathbf{X}^\top\mathbf{J}$ and $\mathbf{D} = \mathcal{K}(\mathbf{X}\mathbf{X}^\top)$. Then, for any $\mathbf{X} \in \mathbb{R}^{n\times k}$, we have*

$$\begin{aligned}\Delta : \nabla^2 f(\mathbf{X}) : \Delta =& \mathcal{K}(\Delta\Delta^\top) : \mathcal{H} : \mathcal{K}(\Delta\Delta^\top) - 3(\mathbf{D}-\mathbf{D}^\star) : \mathcal{H} : (\mathbf{D}-\mathbf{D}^\star) \\ &+ 4\langle \nabla f(\mathbf{X}), \Delta\rangle + [\Delta : \nabla^2 Q(\mathbf{X}) : \Delta - 4\langle \nabla Q(\mathbf{X}), \Delta\rangle]\end{aligned}$$

Proof. Recall the objective function is:

$$f(\mathbf{X}) = \frac{1}{2}(\mathcal{K}(\mathbf{X}\mathbf{X}^\top) - \mathbf{D}^\star) : \mathcal{H} : (\mathcal{K}(\mathbf{X}\mathbf{X}^\top) - \mathbf{D}^\star) + Q(\mathbf{X})$$

Calculating gradient and Hessian, we have for any $\mathbf{Z} \in \mathbb{R}^{d\times r}$:

$$\begin{aligned}\langle \nabla f(\mathbf{X}), \mathbf{Z}\rangle =& (\mathbf{D}-\mathbf{D}^\star) : \mathcal{H} : \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top) + \langle \nabla Q(\mathbf{X}), \mathbf{Z}\rangle \qquad (8)\\ \mathbf{Z} : \nabla^2 f(\mathbf{X}) : \mathbf{Z} =& 2(\mathbf{D}-\mathbf{D}^\star) : \mathcal{H} : \mathcal{K}(\mathbf{Z}\mathbf{Z}^\top) + \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top) : \mathcal{H} : \mathcal{K}(\mathbf{X}\mathbf{Z}^\top + \mathbf{Z}\mathbf{X}^\top) + \mathbf{Z} : \nabla^2 Q(\mathbf{X}) : \mathbf{Z}\end{aligned}$$

Let $\mathbf{Z} = \Delta = \mathbf{J}\mathbf{X} - \mathbf{J}\mathbf{X}^\star\mathbf{R}_\mathbf{X}$ as in Definition 1 and note $\mathbf{M} - \mathbf{M}^\star + \Delta\Delta^\top = \mathbf{J}(\mathbf{X}\Delta^\top + \Delta\mathbf{X}^\top)\mathbf{J}$. This implies that

$\mathcal{K}(\mathbf{M}-\mathbf{M}^\star+\Delta\Delta^\top)=\mathcal{K}(\mathbf{X}\Delta^\top+\Delta\mathbf{X}^\top)$. Then

$$\begin{aligned}\Delta:\nabla^2 f(\mathbf{X}):\Delta =&\mathcal{K}(\mathbf{X}\Delta^\top+\Delta\mathbf{X}^\top):\mathcal{H}:\mathcal{K}(\mathbf{X}\Delta^\top+\Delta\mathbf{X}^\top)+2(\mathbf{D}-\mathbf{D}^\star):\mathcal{H}:\mathcal{K}(\Delta\Delta^\top)+\Delta:\nabla^2 Q(\mathbf{X}):\Delta\\
=&\mathcal{K}(\mathbf{M}-\mathbf{M}^\star+\Delta\Delta^\top):\mathcal{H}:\mathcal{K}(\mathbf{M}-\mathbf{M}^\star+\Delta\Delta^\top)+2(\mathbf{D}-\mathbf{D}^\star):\mathcal{H}:\mathcal{K}(\Delta\Delta^\top)\\
&+\Delta:\nabla^2 Q(\mathbf{X}):\Delta\\
=&\mathcal{K}(\Delta\Delta^\top):\mathcal{H}:\mathcal{K}(\Delta\Delta^\top)-3\mathcal{K}(\mathbf{M}-\mathbf{M}^\star):\mathcal{H}:\mathcal{K}(\mathbf{M}-\mathbf{M}^\star)\\
&+4\mathcal{K}(\mathbf{M}-\mathbf{M}^\star):\mathcal{H}:\mathcal{K}(\mathbf{M}-\mathbf{M}^\star+\Delta\Delta^\top)+\Delta:\nabla^2 Q(\mathbf{X}):\Delta\\
=&\mathcal{K}(\Delta\Delta^\top):\mathcal{H}:\mathcal{K}(\Delta\Delta^\top)-3(\mathbf{D}-\mathbf{D}^\star):\mathcal{H}:(\mathbf{D}-\mathbf{D}^\star)+4(\mathbf{D}-\mathbf{D}^\star):\mathcal{H}:\mathcal{K}(\mathbf{X}\Delta^\top+\Delta\mathbf{X}^\top)\\
&+\Delta:\nabla^2 Q(\mathbf{X}):\Delta\\
=&\mathcal{K}(\Delta\Delta^\top):\mathcal{H}:\mathcal{K}(\Delta\Delta^\top)-3(\mathbf{D}-\mathbf{D}^\star):\mathcal{H}:(\mathbf{D}-\mathbf{D}^\star)+4\langle\nabla f(\mathbf{X}),\Delta\rangle\\
&+[\Delta:\nabla^2 Q(\mathbf{X}):\Delta-4\langle\nabla Q(\mathbf{X}),\Delta\rangle]\end{aligned}$$

where in last line, we use the calculation of gradient $\nabla f(\mathbf{X})$ in Eq.8. This finishes the proof. □

Lemma 13. *Given matrices $\mathbf{X},\mathbf{X}^\star\in\mathbb{R}^{n\times k}$, let $\mathbf{M}=\mathbf{J}\mathbf{X}\mathbf{X}^\top\mathbf{J}$, $\mathbf{M}^\star=\mathbf{J}\mathbf{X}^\star(\mathbf{X}^\star)^\top\mathbf{J}$, $\mathbf{D}=\mathcal{K}(\mathbf{M})$ and $\mathbf{D}^\star=\mathcal{K}(\mathbf{M}^\star)$, and let Δ be defined as in Definition 1, then we have $\|\mathcal{K}(\Delta\Delta^\top)\|_F^2=\|\mathbf{D}-\mathbf{D}^\star\|_F^2-4tr(\mathcal{K}(\mathbf{X}\Delta^\top)\mathcal{K}(\mathbf{X}^\star\Delta^\top))$, and $\|\mathbf{M}-\mathbf{M}^\star\|_F^2\leq\frac{1}{4}\|\mathbf{D}-\mathbf{D}^\star\|_F^2$.*

Lemma 14. *Case where $p=1$, complete observability.*

Proof.

$$\begin{aligned}\Delta:\nabla^2 f(\mathbf{X}):\Delta =&\mathcal{K}(\Delta\Delta^\top):\mathcal{H}:\mathcal{K}(\Delta\Delta^\top)-3(\mathbf{D}-\mathbf{D}^\star):\mathcal{H}:(\mathbf{D}-\mathbf{D}^\star)+4\langle\nabla f(\mathbf{X}),\Delta\rangle\\
=&\|\mathcal{K}(\Delta\Delta^\top)\|_F^2-3\|\mathbf{D}-\mathbf{D}^\star\|_F^2+4\langle\nabla f(\mathbf{X}),\Delta\rangle\\
=&-2\|\mathbf{D}-\mathbf{D}^\star\|_F^2-4\text{tr}(\mathcal{K}(\mathbf{X}\Delta^\top)\mathcal{K}(\mathbf{X}^\star\Delta^\top))+4\langle\nabla f(\mathbf{X}),\Delta\rangle\\
=&-8\|\mathbf{M}-\mathbf{M}^\star\|_F^2-4\text{tr}(\mathcal{K}(\mathbf{X}\Delta^\top)\mathcal{K}(\mathbf{X}^\star\Delta^\top))+4\langle\nabla f(\mathbf{X}),\Delta\rangle\end{aligned}$$

So for this to work, $-4\text{tr}(\mathcal{K}(\mathbf{X}\Delta^\top)\mathcal{K}(\mathbf{X}^\star\Delta^\top))\leq 8\|\mathbf{M}-\mathbf{M}^\star\|_F^2$ must hold. I was able to generate instances randomly where this does not hold.

So one of two things: 1) There must be a condition weaker than strict saddle property, which is sufficient to imply that all local minimal are globally optimal. 2) GD does not always work for this problem. In my extended experiments however, it has always worked. □