Mathematical Methods for Machine Learning and Signal Processing SS 2019

Lecture 19–20: Matrix Completion

V. Morgenshtern

Agenda:

1. The Netflix problem
2. Which matrices can we recover?
3. Recovery algorithm
4. Coherence
5. Recovery via nuclear norm minimization
6. Proof strategy

In this lecture we will not prove all the results. For detailed treatment please see [1–5]

1 The Netflix problem

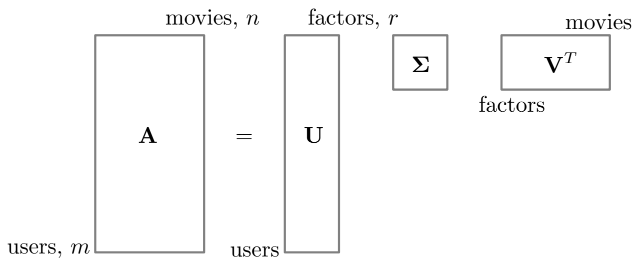
Netflix database consists of about $m \approx 10^6$ users and about $n \approx 25000$ movies. Users rate movies; the ratings are recorded into matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$. Of course most of the users have only seen a small fractions of the movies, and therefore only a small subset of entries of $\mathbf{A}$ have been observed:

						...
Alice	1			4		
Bob		2	5			
Carol			4	5		
Dave	5				4	
⋮						

The goal is to predict which movies a particular user might like. Mathematically this means that we would like to complete matrix $\mathbf{A}$ based on the partial observations of some of its entries.

Clearly, without additional assumptions we cannot recover the entries of $\mathbf{A}$ we have not seen. In the Netflix problem we can assume that matrix $\mathbf{A}$ is low-rank: $r = \text{rank}(\mathbf{A}) \ll n$.

This modeling assumption can be justified as follows. It is reasonable to assume that there is a small number, r, of hidden factors, which are common to all users, so that the preference of each user is largely determined by his/her (linear) response to these factors. The hidden factors might be the genre of the movie or if Leonardo DiCaprio is playing in the movie. The assumption directly leads to the (thin) SVD decomposition of $\mathbf{A}$, in which the inner dimension, r, is much smaller than the outer dimensions m, n:



Originally, $\mathbf{A}$ had mn degrees of freedom (unknown entries). If the low rank model holds, to specify $\mathbf{A}$ we just need to specify $\mathbf{U}, \Sigma, \mathbf{V}$, and so we only need to specify $(m + n - r)r \ll mn$ parameters (remember the orthogonality constraints in the SVD). Therefore, we might hope that it will be possible to recover $\mathbf{A}$ from $\mathcal{O}(r \max(m, n))$ entries.

2 Which matrices can we recover?

In the remainder of the lecture, let's assume, to simplify writing, that $m = n$. We will also assume that $\mathbf{A}$ is real.

Which matrices can we hope to recover?

Let $\mathbf{e}_1, \dots, \mathbf{e}_n$ denote the standard basis in $\mathbb{R}^n$, i.e. $\mathbf{e}_k = [0, \dots, 0, 1, 0, \dots, 0]^\mathsf{T}$. Consider

$$\mathbf{A} = \mathbf{e}_1 \mathbf{e}_n^\mathsf{T} = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Clearly, this matrix cannot be recovered unless we observe $\mathbf{A}_{1,n}$.

Consider rank-1 matrix formed by two arbitrary vectors $\mathbf{x}, \mathbf{y}$: $\mathbf{A} = \mathbf{x}\mathbf{y}^\mathsf{T}$. So that $\mathbf{A}_{i,j} = x_i y_j$. Clearly, if a single row or a single column of $\mathbf{A}$ is not sampled, then recovery is not possible.

We will assume that the sampling set (the set of element of $\mathbf{A}$ that we observe) is random. We will denote this set by Ω.

3 Recovery algorithm

We hope that there exist only one low-rank matrix that is consistent with the sampled entries. Therefore, we might try the following recovery algorithm:

$$\begin{aligned}&\text{minimize}_{\hat{\mathbf{A}}}\ \text{rank}(\hat{\mathbf{A}})\\&\text{subject to } \hat{\mathbf{A}}_{i,j} = \mathbf{A}_{i,j} \text{ for } (i,j) \in \Omega.\end{aligned}$$

Similar to P0, this algorithm is NP-hard. Convex relaxation of the rank minimization problem is

$$\begin{aligned}&\text{minimize}_{\hat{\mathbf{A}}}\ \|\hat{\mathbf{A}}\|_* \qquad (1)\\&\text{subject to } \hat{\mathbf{A}}_{i,j} = \mathbf{A}_{i,j} \text{ for } (i,j) \in \Omega.\end{aligned}$$

Above, $\|\mathbf{A}\|_*$ denotes the nuclear norm of $\mathbf{A}$ defined as

$$\|\mathbf{A}\|_* = \text{tr}\left(\sqrt{\mathbf{A}^\mathsf{T}\mathbf{A}}\right) = \sum_{i=1}^{n} \sigma_i(\mathbf{A}),$$

where $\sigma_i(\mathbf{A})$ are the singular values of $\mathbf{A}$.

It can be shown that the nuclear norm 1-ball is the convex hull of rank-1 matrices obeying $\|\mathbf{x}\mathbf{y}^\mathsf{T}\|_* \leq 1$.

Nuclear norm minimization is a convex optimization problem and, therefore, can be solved efficiently. It can be reduced to semi-definite convex program (SDP).

4 Coherence

In this section we motivate coherence of a matrix, a measure that will be important for matrix recovery. Consider examples.

Let

$$\mathbf{A} = \sum_{k=1}^{2} \sigma_k \mathbf{u}_k \mathbf{u}_k^\mathsf{T}$$

where $\mathbf{u}_1 = (\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}$ and $\mathbf{u}_2 = (\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}$, i.e.

$$\mathbf{A} = \begin{bmatrix} * & * & 0 & \dots & 0 \\ * & * & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

Clearly, $\mathbf{A}$ cannot be recovered from a small set of entries.

Let

$$\mathbf{A} = \mathbf{e}_1\mathbf{x}^\mathsf{T} = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_{n-1} & x_n \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Clearly, $\mathbf{A}$ cannot be recovered from a small set of entries. The intuition here is that column and row spaces cannot be aligned with standard basis vectors.

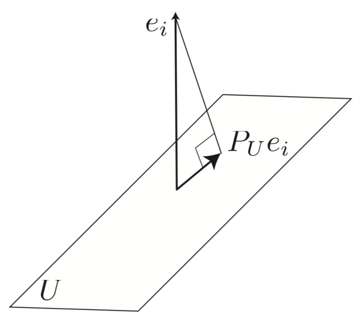
Definition 1. Assume that $\text{rank}(\mathbf{A}) = r$ so that the SVD of $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\mathsf{T} = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^*.$$

with $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{n\times r}$. Coherence of $\mathbf{A}$ with respect to standard basis, $\mathbf{e}_i$, is the smallest scalar μ obeying

$$\max_{1\leq k\leq n} \|\mathbf{U}^\mathsf{T}\mathbf{e}_k\|^2 \leq \mu\frac{r}{n}, \quad \max_{1\leq k\leq n} \|\mathbf{V}\mathbf{e}_k\|^2 \leq \mu\frac{r}{n}.$$

The coherence quantifies how close are the elements of the standard basis, $\mathbf{e}_k$, are to the subspace spanned by the columns of $\mathbf{U}$ (and $\mathbf{V}$) by measuring the length of projection of $\mathbf{e}_k$ onto these subspaces:



To gain intuition, observe that for an $n \times r$ matrix $\mathbf{U}$ with orthonormal columns, the smallest μ that can satisfy

$$\max_{1\leq k\leq n} \|\mathbf{U}^\mathsf{T}\mathbf{e}_k\|^2 \leq \mu\frac{r}{n} \tag{2}$$

is $\mu = 1$. This happens when $\mathbf{U}$ consists of vectors whose entries all have magnitude $\frac{1}{\sqrt{n}}$. This is the situation in which the eigenvectors of $\mathbf{A}$ have their energy uniformly spread-out across all n dimensions. Coherence is small. We expect that recovery will be possible in this case.

On the other hand, when $\mathbf{e}_k \in \text{span}(\mathbf{u}_1, \ldots, \mathbf{u}_r)$, the smallest μ that satisfies (2) is $\mu = \frac{n}{r}$. Coherence is large. It is impossible to recover $\mathbf{A}$ in this case.

It turns out that for many ensembles of random matrices, $\mu \sim \mathcal{O}(1)$ if r is sufficiently small.

The role of the coherence in this theory is also very natural, and can be understood when thinking about the prediction of movie ratings. Here, we can imagine that the complete matrix of ratings is (approximately) low rank because users' preferences are correlated. Now the reason why matrix completion is possible under incoherence is that we can exploit correlations and infer how a specific user is going to like a movie she has not yet seen, by examining her ratings and learning about her general preferences, and inferring how other users with such preferences have rated this particular item. Whenever we have users or small groups of users that are very singular in the sense that their ratings are orthogonal to those of all other users, it is not possible to correctly predict their missing entries. Such matrices have large coherence.

To convince oneself, consider situations where one user enter ratings based on the outcome of coin tosses $(t_1, t_2, \ldots)$:

$$\mathbf{A} = \underbrace{\frac{1}{\sqrt{n}} \frac{1}{\sqrt{(m-1)}} \begin{bmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{u}_1} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} + \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\mathbf{u}_2} \begin{bmatrix} t_1 & t_2 & \cdots & t_4 \end{bmatrix}$$

The rank of $\mathbf{A}$ is 2. However we still cannot recover $\mathbf{A}$ perfectly from a few measurements. This is because if even one elements of the $t_1, t_2, \ldots$ is not observed, there is no way to recover it based on other elements. The coherence of the matrix is high, because $\mathbf{u}_2 = \mathbf{e}_1$.

In the following we will assume that the sampling process follows the Bernoulli model, where each entry of $\mathbf{A}$ is observed independently and identically with probability p. Therefore, the total number of observations is about pn^2.

5 Recovery via nuclear norm minimization

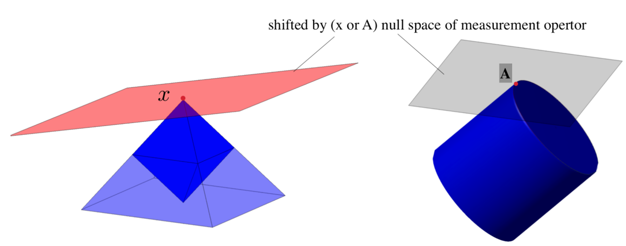
We are now ready to state the main result of this lecture.

Theorem 1 ([6], [3])**.** Suppose $\mathbf{A}$ is fixed but otherwise arbitrary $n \times n$ matrix with coherence μ and rank r. There exist universal constants $c_0, c_1, c_2 > 0$ such that if

$$p \geq c_0 \frac{\mu r \log^2(n)}{n}$$

then $\mathbf{A}$ is the unique solution to (1) with probability at least $1 - c_1 n^{-c_2}$.

Similar to l1-minimization, the success of (1) in recovering the matrix $\mathbf{A}$ can be intuitively understood geometrically and formalized via presenting a dual certificate. Here is the geometry of the nuclear norm minimization (on the right) along with the already familiar geometry of l1 minimization on the left, for reference:



(a) ℓ_1 ball **(b)** Nuclear ball

In grey we depict the null space of the measurement operator shifted by $\mathbf{A}$: this is a linear space of all $n \times n$ matrices that agree with $\mathbf{A}$ on Ω and arbitrary everywhere else. $\mathbf{A}$ is the unique solution of (1) if and only if there is no way to move inside the shifted null space while making the nuclear norm smaller. Equivalently, $\mathbf{A}$ is the unique solution of (1) if and only if the cone of descent of the nuclear norm at $\mathbf{A}$ does not intersect the shifted null space. The nuclear ball is depicted in dark blue in the case when $\mathbf{A}$ is a 2×2 symmetric matrix, and therefore depends on 3 parameters.

Looking at the figure, we can intuitively understand why minimizing the l1 and nuclear norms recovers sparse and low-rank objects: indeed, as the figure suggests, the cone of descent (light blue in the left figure) to the l1 norm is 'narrow' at sparse vectors and, therefore, even though the null space is of small codimension (equal to the number of measurements), it is likely that if the number of measurements is large enough, it will miss the tangent cone. A similar observation applies to the nuclear ball, which also appears pinched at low-rank objects.

6 Proof strategy

Similar to results on l1 minimization, the favorable geometry can be guaranteed by constructing a dual certificate. We need the following definitions:

Definition 2. Let $\mathcal{T}$ denote the span of all matrices with row space *or* column space included in that of $\mathbf{A}$. The orthogonal projection onto $\mathcal{T}$ is therefore given by

$$\mathcal{P}_{\mathcal{T}}(\mathbf{Z}) = \mathbf{U}\mathbf{U}^\top\mathbf{Z} + \mathbf{Z}\mathbf{V}\mathbf{V}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{Z}\mathbf{V}\mathbf{V}^\top.$$

and the projection onto the orthogonal complement is $\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z}) = \mathbf{Z} - \mathcal{P}_{\mathcal{T}}(\mathbf{Z})$.

To understand the definition, consider the following cases. The full proof is left as an exercise to the reader.

Case 1: $\mathbf{Z} = \mathbf{U}_1\Lambda\mathbf{V}_1^\top$ where $\mathbf{U}_1$ contains a subset of columns of $\mathbf{U}$ and $\mathbf{V}_1$ contains a subset of columns of $\mathbf{V}$. Then,

$$\begin{aligned}\mathcal{P}_{\mathcal{T}}(\mathbf{Z}) &= \mathbf{U}\mathbf{U}^\top\mathbf{U}_1\Lambda\mathbf{V}_1^\top + \mathbf{U}_1\Lambda\mathbf{V}_1^\top\mathbf{V}\mathbf{V}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{U}_1\Lambda\mathbf{V}_1^\top\mathbf{V}\mathbf{V}^\top \\ &= \mathbf{U}_1\Lambda\mathbf{V}_1^\top + \mathbf{U}_1\Lambda\mathbf{V}_1^\top - \mathbf{U}_1\Lambda\mathbf{V}_1^\top \\ &= \mathbf{Z}.\end{aligned}$$

Case 2: $\mathbf{Z} = \mathbf{U}_1\Lambda\tilde{\mathbf{V}}^\top$ where $\mathbf{U}_1$ contains a subset of columns of $\mathbf{U}$ and $\tilde{\mathbf{V}}$ does not contain columns that belong to the span of columns of $\mathbf{V}$. Then,

$$\begin{aligned}\mathcal{P}_{\mathcal{T}}(\mathbf{Z}) &= \mathbf{U}\mathbf{U}^\top\mathbf{U}_1\Lambda\tilde{\mathbf{V}}^\top + \mathbf{U}_1\Lambda\tilde{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{U}_1\Lambda\tilde{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top \\ &= \mathbf{U}_1\Lambda\mathbf{V}_1^\top + \mathbf{0} - \mathbf{0} \\ &= \mathbf{Z}.\end{aligned}$$

Case 3: $\mathbf{Z} = \tilde{\mathbf{U}}\Lambda\tilde{\mathbf{V}}^\top$ where $\tilde{\mathbf{U}}$ does not contain columns that belong to the span of columns of $\mathbf{U}$ and $\tilde{\mathbf{V}}$ does not contain columns that belong to the span of columns of $\mathbf{V}$. Then,

$$\begin{aligned}\mathcal{P}_{\mathcal{T}}(\mathbf{Z}) &= \mathbf{U}\mathbf{U}^\top\tilde{\mathbf{U}}\Lambda\tilde{\mathbf{V}}^\top + \tilde{\mathbf{U}}\Lambda\tilde{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top - \mathbf{U}\mathbf{U}^\top\tilde{\mathbf{U}}\Lambda\tilde{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top \\ &= \mathbf{0} + \mathbf{0} - \mathbf{0} \\ &= \mathbf{0}.\end{aligned}$$

Definition 3. Define the orthogonal projection operator onto the span of all matrices that vanish on $\Omega^{\perp}$: $\mathcal{P}_{\Omega}(\mathbf{Z})$ is the matrix with $\mathcal{P}_{\Omega}(\mathbf{Z}) = \mathbf{Z}_{i,j}$ if $(i,j) \in \Omega$ and zero otherwise. For convenience, $\mathcal{R}_{\Omega}(\mathbf{Z}) = \frac{1}{p}\mathcal{P}_{\Omega}(\mathbf{Z})$.

A sufficient condition for successful recovery can now be expressed in term of dual certificate:

Lemma 2 ([6])**.** Suppose $p \geq \frac{1}{n}$. The matrix $\mathbf{A}$ is the unique optimal solution to (1) if the following conditions hold:

1. $\|\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\|_{op} \leq \frac{1}{2}$.
2. There exists an approximate dual certificate $\mathbf{Y} \in \mathbb{R}^{n\times n}$ which satisfies $\mathcal{P}_{\Omega}(\mathbf{Y}) = \mathbf{Y}$ and
   - $\|\mathcal{P}_{\mathcal{T}}(\mathbf{Y}) - \mathbf{U}\mathbf{V}^{\mathsf{T}}\|_F \leq \frac{1}{4n}$
   - $\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{Y})\| \leq \frac{1}{2}$

Above $\|\mathbb{A}\|_{op} = \sup_{\mathbf{Z}\in\mathbb{R}^{n\times n}} \|\mathbb{A}(\mathbf{Z})\|_F / \|\mathbf{Z}\|_F$ is the operator norm and $\|\mathbf{Z}\|_F = \sqrt{\sum_i \sum_j \mathbf{Z}_{i,j}} = \sqrt{\operatorname{tr} \mathbf{Z}^{\mathsf{T}}\mathbf{Z}} = \sqrt{\sum_i \sigma^2(\mathbf{Z})}$ is the Frobenius norm and $\|\mathbf{A}\|$ is the spectral norm of the matrix (the largest singular value).

The existence of the dual certificate with the required properties can be proven via the ingenious golfing scheme invented by David Gross [4]. We refer the reader to [6]. We complete this lecture with the proof of Lemma 2.

Proof. The inner product between two matrices is given by $\langle \mathbf{X}, \mathbf{Z}\rangle = \operatorname{trace}(\mathbf{X}^{\mathsf{T}}\mathbf{Z})$.

Consider any feasible solution $\mathbf{X}$ to (1) with $\mathcal{P}_{\Omega}(\mathbf{X}) = \mathcal{P}_{\Omega}(\mathbf{A})$. Let $\mathbf{G}$ be an $n \times n$ matrix with satisfies

$$\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{G})\| = 1 \tag{3}$$

and

$$\langle \mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{G}), \mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{X} - \mathbf{A})\rangle = \|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{X} - \mathbf{A})\|_*. \tag{4}$$

Such a matrix $\mathbf{G}$ always exist by the duality between the nuclear norm and the spectral norm (exercise)[1]. Because $\mathbf{U}\mathbf{V}^{\mathsf{T}} + \mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{G})$ is a sub-gradient[2] of $\|\mathbf{Z}\|_*$ at $\mathbf{Z} = \mathbf{A}$ (exercise), we get

$$\|\mathbf{X}\|_* - \|\mathbf{A}\|_* \geq \left\langle \mathbf{U}\mathbf{V}^{\mathsf{T}} + \mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{G}), \mathbf{X} - \mathbf{A} \right\rangle.$$

We also have $\langle \mathbf{Y}, \mathbf{X} - \mathbf{A}\rangle = \langle \mathcal{P}_{\Omega}(\mathbf{Y}), \mathcal{P}_{\Omega}(\mathbf{X} - \mathbf{A})\rangle = 0$ since $\mathcal{P}_{\Omega}(\mathbf{Y}) = \mathbf{Y}$. It follows that

$$\begin{aligned}
\|\mathbf{X}\|_* - \|\mathbf{A}\|_* &\geq \left\langle \mathbf{U}\mathbf{V}^{\mathsf{T}} + \mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{G}) - \mathbf{Y}, \mathbf{X} - \mathbf{A} \right\rangle \\
&\overset{(a)}{=} \|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{X} - \mathbf{A})\|_* + \left\langle \mathbf{U}\mathbf{V}^{\mathsf{T}} - \mathcal{P}_{\mathcal{T}}(\mathbf{Y}), \mathbf{X} - \mathbf{A} \right\rangle - \langle \mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{Y}), \mathbf{X} - \mathbf{A}\rangle \\
&\overset{(b)}{\geq} \|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{X} - \mathbf{A})\|_* - \|\mathbf{U}\mathbf{V}^{\mathsf{T}} - \mathcal{P}_{\mathcal{T}}(\mathbf{Y})\|_F \|\mathcal{P}_{\mathcal{T}}(\mathbf{X} - \mathbf{A})\|_F - \|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{Y})\| \|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{X} - \mathbf{A})\|_* \\
&\overset{(c)}{\geq} \frac{1}{2}\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{X} - \mathbf{A})\|_* - \frac{1}{4n}\|\mathcal{P}_{\mathcal{T}}(\mathbf{X} - \mathbf{A})\|_F
\end{aligned}$$

[1] Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The associated dual norm, is defined as $\|\mathbf{z}\|_{dual} = \sup\{\langle \mathbf{z}, \mathbf{x}\rangle \ : \ \|\mathbf{x}\| \leq 1\}$.
[2] Vector $\mathbf{g} \in \mathbb{R}^n$ is a subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x}$ if for all $\mathbf{z}$, $f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{g}^{\mathsf{T}}(\mathbf{z} - \mathbf{x})$.

where (a) we used that $\mathbf{Y} = \mathcal{P}_{\mathcal{T}}(\mathbf{Y}) + \mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Y})$ and (4), (b) follows because

$$\langle \mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Y}), \mathbf{X} - \mathbf{A} \rangle = \langle \mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Y}), \mathcal{P}_{\mathcal{T}^\perp}(\mathbf{X} - \mathbf{A}) \rangle \,,$$

$$\mathbf{U}\mathbf{V}^\mathsf{T} = \mathcal{P}_{\mathcal{T}}(\mathbf{U}\mathbf{V}^\mathsf{T})$$

so that $\left\langle \mathbf{U}\mathbf{V}^\mathsf{T} - \mathcal{P}_{\mathcal{T}}(\mathbf{Y}), \mathbf{X} - \mathbf{A} \right\rangle = \left\langle \mathbf{U}\mathbf{V}^\mathsf{T} - \mathcal{P}_{\mathcal{T}}(\mathbf{Y}), \mathcal{P}_{\mathcal{T}}(\mathbf{X} - \mathbf{A}) \right\rangle$, the Frobenius norm is self-dual and the duality between the nuclear norm and the spectral norm, and in the (c) we used conditions 2) in the Lemma.

Applying Lemma 3 below, we obtain

$$\|\mathbf{X}\|_* - \|\mathbf{A}\|_* \geq \frac{1}{2}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{X} - \mathbf{A})\|_* - \frac{1}{4n}\sqrt{2n}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{X} - \mathbf{A})\|_* > \frac{1}{8}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{X} - \mathbf{A})\|_*.$$

The right hand side is strictly positive for all $\mathbf{X}$ with $\mathcal{P}_\Omega(\mathbf{X} - \mathbf{A}) = 0$ and $\mathbf{X} \neq \mathbf{A}$. Otherwise we must have $\mathcal{P}_{\mathcal{T}}(\mathbf{X} - \mathbf{A}) = \mathbf{X} - \mathbf{A}$ and $\mathcal{P}_{\mathcal{T}}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}}(\mathbf{X} - \mathbf{A}) = 0$, contradicting the inequality $\|\mathcal{P}_{\mathcal{T}}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\|_{op} \leq \frac{1}{2}$. This proves that $\mathbf{A}$ is the unique optimal solution to the program (1). □

It remains to show the following technical lemma:

Lemma 3. If $p \geq \frac{1}{n}$ and $\|\mathcal{P}_{\mathcal{T}}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\|_{op} \leq \frac{1}{2}$, then

$$\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F \leq \sqrt{2n}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z})\|_* \;\text{ for all } \mathbf{Z} : \mathcal{R}_\Omega(\mathbf{Z}) = 0.$$

Proof. Observe that

$$\begin{aligned}
\|\sqrt{p}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F &= \sqrt{\langle(\mathcal{P}_{\mathcal{T}}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}})(\mathbf{Z}), \mathcal{P}_{\mathcal{T}}(\mathbf{Z})\rangle + \langle\mathcal{P}_{\mathcal{T}}(\mathbf{Z}), \mathcal{P}_{\mathcal{T}}(\mathbf{Z})\rangle} \\
&\geq \sqrt{\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F^2 - \|\mathcal{P}_{\mathcal{T}}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\|_{op}\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F^2} \\
&\geq \frac{1}{\sqrt{2}}\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F
\end{aligned}$$

where in the first equality we used the properties of orthogonal projection operators and the relationship between $\mathcal{R}_\Omega$ and $\mathcal{P}_\Omega$ and the last inequality follows from the assumption $\|\mathcal{P}_{\mathcal{T}}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\|_{op} \leq \frac{1}{2}$. On the other hand $\mathcal{P}_\Omega(\mathbf{Z}) = 0$ implies $\mathcal{R}_\Omega(\mathbf{Z}) = 0$ and thus

$$\begin{aligned}
\|\sqrt{p}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F &= \|\sqrt{p}\mathcal{R}_\Omega(\mathbf{Z} - \mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z}))\|_F \\
&= \|\sqrt{p}\mathcal{R}_\Omega\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z})\|_F \\
&\leq \frac{1}{\sqrt{p}}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z})\|_F \leq \sqrt{n}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z})\|_F
\end{aligned}$$

where in the second equation we used that $\mathcal{R}_\Omega(\mathbf{Z}) = 0$. Combining the last two display equations gives

$$\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F \leq \sqrt{2n}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z})\|_F \leq \sqrt{2n}\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{Z})\|_*.$$

□

References

[1] Y. Chi, "Low-rank matrix completion," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 178–181, 2018.

[2] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, 2009.

[3] E. Candès, "Mathematics of sparsity (and a few other things)," 2014.

[4] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, pp. 1548–1566, Feb. 2011.

[5] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2053 – 2080, Mar. 2010.

[6] Y. Chen, "Incoherence-optimal matrix completion," vol. 61, no. 5, pp. 2909–2923, 2015.