Nuclear norm is the tightest convex envelop of rank function within the unit ball

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : Janurary 30, 2020
Last update : February 2, 2020

Rank and nuclear norm of matrix

Rank function

$$\mathrm{rank}(\mathbf{X}) : \mathbb{R}^{m \times n} \to \mathbb{N}.$$

- Input : a m-by-n real (or in general, complex) matrix $\mathbf{X}$
- Output : an (positive) integer, which is the number of linearly independent columns of $\mathbf{X}$ (or the number of linearly independent rows of $\mathbf{X}$)

Nuclear norm (or a less standard name "trace norm")

$$\|\mathbf{X}\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{X}) : \mathbb{R}^{m \times n} \to \mathbb{R}$$

- Input : a m-by-n real (or in general, complex) matrix $\mathbf{X}$
- Output : an (non-negative) real number, which is the sum of the singular value of $\mathbf{X}$

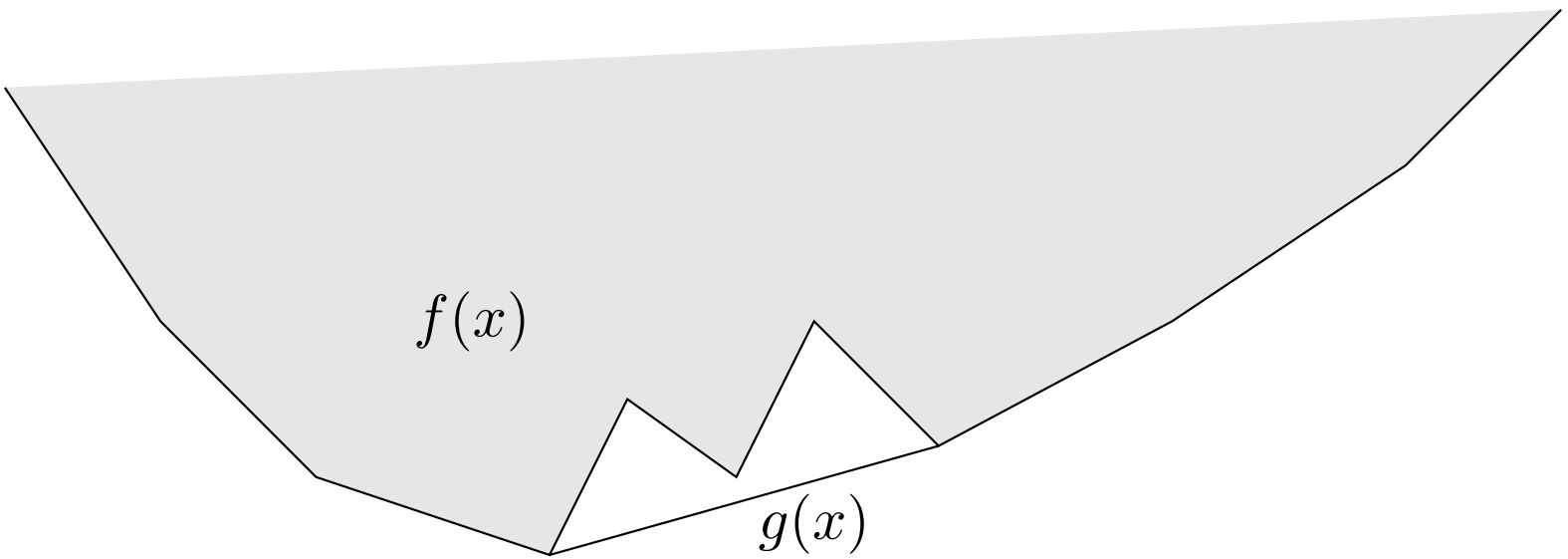
The concept of convex envelope

Given a function $f : C \to \mathbb{R}$, a function g is called the (largest) *convex envelope* of f if and only if g is convex and

$$g(\mathbf{x}) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in C.$$

i.e., g is a convex function *below* f that is *closest point-wise* to f.

Illustration :



(The LaTeX's Tikz code to generate this figure is from here.)

Rank and nuclear norm are related by convex envelope

Theorem (Fazel02) For $f(\mathbf{X}) = \mathrm{rank}(\mathbf{X})$ over the set

$$S := \{\mathbf{X} \in \mathbb{R}^{m \times n} \mid \|\mathbf{X}\|_2 \leq 1\},$$

the convex envelope of f is the nuclear norm $\|\mathbf{X}\|_*$.

- The theorem only applies to those $\mathbf{X}$ inside the unit ball. It tells nothing if $\mathbf{X}$ is outside the unit ball. In fact, we will see, if $\mathbf{X}$ is outside the unit ball, the output of the convex envelope will be at ∞.
- If we have $S' := \{\mathbf{X} \in \mathbb{R}^{m \times n} \mid \|\mathbf{X}\|_2 \leq M\}$, we can do a scaling to reuse the theorem : in this case the convex envelope of f on S' will be $\frac{1}{M}\|\mathbf{X}\|_*$, for $M > 0$.
- The tool to prove the theorem is basically the convex conjugate.

Idea of the proof : Legendre-Fenchel's convex conjugate

Given a function $f : C \to \mathbb{R}$, the convex conjugate of f, denoted as f^*, is defined as

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in C} \ \langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}).$$

To prove the envelope theorem, the key is to use the fact that

$$\text{convex envelope of } f = (f^*)^* = f^{**}.$$

The proof to be presented next follows p.56-59 of **Fazel02**[1]. The proof basically shows $\big(\mathrm{rank}(\mathbf{X})\big)^{**}$ is the nuclear norm.

[1] M. Fazel, "Matrix rank minimization with application", Stanford University, 2002.

Tools for the proof

- The set of unit ball

$$S := \{\mathbf{X} \in \mathbb{R}^{m\times n} \mid \|\mathbf{X}\|_2 \leq 1\}. \tag{1}$$

- Convex conjugate : for $f : C \to \mathbb{R}$,

$$f^*(\mathbf{y}) = \sup_{\mathbf{x}\in C} \langle \mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x}). \tag{2}$$

- Inner product in $\mathbb{R}^{m\times n}$

$$\langle \mathbf{Y}, \mathbf{X}\rangle = \sum_{ij} \mathbf{Y}_{ij}\mathbf{X}_{ij} = \mathrm{Tr}(\mathbf{Y}^\top \mathbf{X}). \tag{3}$$

- von Neumman's trace inequality : let $\mathbf{A}, \mathbf{B}$ in $\mathbb{R}^{m\times n}$. Let $\sigma_i(\mathbf{A})$ be the ith largest singular value of $\mathbf{A}$ and let $q = \min\{m, n\}$, then

$$\mathrm{Tr}(\mathbf{A}^\top \mathbf{B}) \leq \underbrace{\left|\mathrm{Tr}(\mathbf{A}^\top \mathbf{B})\right| \leq \sum_{i=1}^{q} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})}_{\text{the original von Neumman's trace inequality}} . \tag{4}$$

Equality holds when singular vectors of $\mathbf{A}$ and that of $\mathbf{B}$ are equal.

The proof ... 1/12

$$
\begin{aligned}
f^*(\mathbf{Y}) &\overset{(2)}{=} \sup_{\mathbf{X}\in C} \langle \mathbf{Y}, \mathbf{X}\rangle - f(\mathbf{X}) \\
&\overset{\text{def of } f}{=} \sup_{\mathbf{X}\in S} \langle \mathbf{Y}, \mathbf{X}\rangle - \operatorname{rank}(\mathbf{X}), \text{ where } S \text{ is defined in (1)} \\
&\overset{(1)}{=} \sup_{\|\mathbf{X}\|_2\le 1} \langle \mathbf{Y}, \mathbf{X}\rangle - \operatorname{rank}(\mathbf{X}) \\
&\overset{(3)}{=} \sup_{\|\mathbf{X}\|_2\le 1} \operatorname{Tr}(\mathbf{Y}^\top \mathbf{X}) - \operatorname{rank}(\mathbf{X}). \\
&\overset{(4)}{\le} \sup_{\|\mathbf{X}\|_2\le 1} \sum_{i=1}^{q} \sigma_i(\mathbf{Y})\sigma_i(\mathbf{X}) - \operatorname{rank}(\mathbf{X}). \\
&\overset{\text{see } *}{=} \sup_{\|\mathbf{X}\|_2\le 1} \sum_{i=1}^{q} \sigma_i(\mathbf{Y})\sigma_i(\mathbf{X}) - \operatorname{rank}(\mathbf{X}). \qquad (5)
\end{aligned}
$$

* : the sup is on $\mathbf{X}$, we are "free to pick" whatever $\mathbf{Y}$ we want to make the sup as large as possible. By equality condition of (4), we can pick $\mathbf{Y}$ such that the singular vectors of $\mathbf{Y}$ equals to that of $\mathbf{X}$.
(Why we want equal sign : because we want to show $\big(\operatorname{rank}(\mathbf{X})\big)^{**}$ **equals** to nuclear norm — we don't want to have $\le$ or $\ge$)

The proof ... 2/12

Now, we only need to consider $\mathbf{X}$ in the unit ball with rank at most q. So we can let $\mathrm{rank}(\mathbf{X}) = r$ for $r \in \{0, q\}$. So

$$f^*(\mathbf{Y}) \overset{(5)}{=} \sup_{\|\mathbf{X}\|_2 \leq 1} \sum_{i=1}^{q} \sigma_i(\mathbf{Y})\sigma_i(\mathbf{X}) - \mathrm{rank}(\mathbf{X}).$$

becomes

$$f^*(\mathbf{Y}) = \sup_{\|\mathbf{X}\|_2 \leq 1} \sum_{i=1}^{r} \sigma_i(\mathbf{Y})\sigma_i(\mathbf{X}) - r, \quad r \in \{0, q\}.$$

Now consider the subscript of $\sup$. $\|\mathbf{X}\|_2 \leq 1$ means $\sigma_r(\mathbf{X}) \leq \cdots \leq \sigma_1(\mathbf{X}) \leq 1$, so

$$f^*(\mathbf{Y}) = \sup \sum_{i=1}^{r} \sigma_i(\mathbf{Y}) - r, \quad r \in \{0, q\}. \tag{6}$$

where the equality holds due to $\sup$. Note that we are focusing on $\mathbf{X}$ that fulfils $\|\mathbf{X}\| \leq 1$, so we drop the subscript of $\sup$.

The proof ... 3/12

As $r \in \{0, \ldots, q\}$, we have to consider all the cases for $r = 0$ to $r = q$ in (5) :

$$f^*(\mathbf{Y}) \overset{(5),(6)}{=} \max\Big(\underbrace{0}_{\text{case } r=0}, \underbrace{\sigma_1(\mathbf{Y}) - 1}_{\text{case } r=1}, \ldots, \underbrace{\sum_{i=1}^{q} \sigma_i(\mathbf{Y}) - q}_{\text{case } r=q} \Big)$$

$$= \max\Big(0, \sigma_1(\mathbf{Y}) - 1, \ldots, \sum_{i=1}^{q} \sigma_i(\mathbf{Y}) - q\Big). \tag{7}$$

To solve (7), we cannot change $\mathbf{X}$ any more : we already do so in order to arrive at (7). What is left are those $\sigma_i(\mathbf{Y})$. So next we consider the case for $\|\mathbf{Y}\| > 1$ or $\|\mathbf{Y}\| \leq 1$.

If $\|\mathbf{Y}\| \leq 1$, then $\sigma_i(\mathbf{Y}) \leq 1$ and all the subtractions give negative value and so $f^*(\mathbf{Y}) = 0$. So

$$f^*(\mathbf{Y}) = \begin{cases} 0 & \|\mathbf{Y}\| \leq 1 \\ \max\Big(0, \sigma_1(\mathbf{Y}) - 1, \ldots, \sum_{i=1}^{q} \sigma_i(\mathbf{Y}) - q\Big) & \|\mathbf{Y}\| > 1 \end{cases}$$

In which the terms inside max can be compactly represented by using r again, see next page.

The proof ... 4/12

For the case $\|\mathbf{Y}\| > 1$, the largest case is the one that sums all positive $\sigma_i(\mathbf{Y}) - 1$ terms

$$f^*(\mathbf{Y}) = \begin{cases} 0 & \|\mathbf{Y}\| \leq 1 \text{ or } r = 0 \\ \max\limits_{\sigma_i(\mathbf{Y})} \sum\limits_{i=1}^{r} \sigma_i(\mathbf{Y}) - r & \sigma_r(\mathbf{Y}) > 1 \text{ and } \sigma_{r+1}(\mathbf{Y}) \leq 1, r \in \{1, q\} \end{cases} \tag{8}$$

Note the condition

$$\sigma_r(\mathbf{Y}) > 1 \text{ and } \sigma_{r+1}(\mathbf{Y}) \leq 1, \ r \in \{1, q\} \tag{9}$$

is important.

Using $(x)_+ = \max(x, 0)$, (8) can be compactly represented as

$$f^*(\mathbf{Y}) = \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1)_+, \quad r \in \{1, q\}. \tag{10}$$

So we now have $f^*(\mathbf{Y})$ expressed by (8),(9),(10).

The first half of the proof is finished. The second half of the proof is to use $f^*(\mathbf{Y})$ to compute $f^{**}(\mathbf{Y})$ and show this is the nuclear norm.

The proof ... 5/12

To avoid confusion of notation, we consider $f^{**}(\mathbf{Z})$ on the variable $\mathbf{Z} \in \mathbb{R}^{m \times n}$.

$$f^{**}(\mathbf{Z}) \overset{(2)}{=} \sup_{\mathbf{Y} \in C_{f^*}} \langle \mathbf{Z}, \mathbf{Y} \rangle - f^*(\mathbf{Y}),$$

where C_{f^*} is the domain of $f^*(\mathbf{Y})$, in which it is just $\mathbb{R}^{m \times n}$ (no constraint on $\mathbf{Y}$ in $f^*(\mathbf{Y})$), so we can remove this and get

$$\begin{aligned} f^{**}(\mathbf{Z}) &\overset{(3),(4)}{\leq} \sup_{\mathbf{Y}} \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - f^*(\mathbf{Y}) \\ &\overset{\text{see } *}{=} \sup_{\mathbf{Y}} \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - f^*(\mathbf{Y}) \\ &\overset{(10)}{=} \sup_{\sigma_i(\mathbf{Y})} \left\{ \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1)_+ \right\}, r \in \{1, q\} \end{aligned}$$

* : again we pick $\mathbf{Z}$ such that the singular vectors of it are equal to that of $\mathbf{Y}$ to maximize the inequality.

The proof ... 6/12

Recall that the theorem is to show **within the unit ball**, convex envelope of rank function is nuclear norm.

We already have the expression of the convex envelope, which is

$$\begin{aligned} f^{**}(\mathbf{Z}) &= \sup_{\mathbf{Y}} \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - f^{*}(\mathbf{Y}) \\ &= \sup_{\sigma_i(\mathbf{Y})} \left\{ \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1)_+ \right\}, r \in \{1, q\} \quad (11) \end{aligned}$$

So we consider the two cases on $\mathbf{Z}$: inside or outside the unit ball.
That is : $\|\mathbf{Z}\| > 1$ or $\|\mathbf{Z}\| \leq 1$.

We first consider the case $\|\mathbf{Z}\| > 1$, showing that $f^{**}(\mathbf{Z}) = \infty$.
And then we consider the case $\|\mathbf{Z}\| \leq 1$, showing that $f^{**}(\mathbf{Z}) = \|\mathbf{Y}\|_*$ for whatever possible $\mathbf{Y}$.

The proof ... 7/12

$$f^{**}(\mathbf{Z}) \overset{(11)}{=} \sup_{\sigma_i(\mathbf{Y})} \left\{ \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1)_+ \right\}, r \in \{1, q\}$$

If $\|\mathbf{Z}\|_2 = \sigma_1(\mathbf{Z}) > 1$, in this case we can choose $\sigma_1(\mathbf{Y})$ sufficiently large such that $f^{**}(\mathbf{Z}) \to \infty$. (we are free to do so as there is no constraint under $\sup$).

Consider the expression of f^{**} and focus on the terms with $\sigma_1(\mathbf{Y})$

$$\begin{aligned}
f^{**}(\mathbf{Z}) &= \sup_{\sigma_1(\mathbf{Y})} \sigma_1(\mathbf{Z})\sigma_1(\mathbf{Y}) - (\sigma_1(\mathbf{Y}) - 1)_+ + \text{other terms} \\
&\overset{\sigma_1(\mathbf{Y}) \gg 1}{=} \sup_{\sigma_1(\mathbf{Y})} \sigma_1(\mathbf{Z})\sigma_1(\mathbf{Y}) - \sigma_1(\mathbf{Y}) + \text{other terms} \\
&= \sup_{\sigma_1(\mathbf{Y})} \sigma_1(\mathbf{Y}) \underbrace{(\sigma_1(\mathbf{Z}) - 1)}_{>0} + \text{other terms} \\
&= \infty.
\end{aligned}$$

Such case $f^{**}(\mathbf{Z}) = \infty$ if $\|\mathbf{Z}\| > 1$ is OK for our purpose as the convex envelop theorem consider the matrix *with in the unit ball*.

The proof ... 8/12

Now we consider the case $\|\mathbf{Z}\| \leq 1$ and show $f^{**}(\mathbf{Z}) = \|\mathbf{Z}\|_*$.

We begin with (11) :

$$\begin{aligned} f^{**}(\mathbf{Z}) &= \sup_{\mathbf{Y}} \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - f^*(\mathbf{Y}) \\ &= \sup_{\sigma_i(\mathbf{Y})} \left\{ \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1)_+ \right\}, r \in \{1, q\} \end{aligned}$$

Again, $\sup_{\mathbf{Y}}$ has no restriction on $\mathbf{Y}$ so we can select whatever $\mathbf{Y}$ we want, so we consider again two cases : $\|\mathbf{Y}\| \leq 1$ and $\|\mathbf{Y}\| > 1$.

If $\|\mathbf{Y}\| \leq 1$, then $f^*(\mathbf{Y}) \overset{(8)}{=} 0$ so the supremum in (11) is achieved for $\sigma_i(\mathbf{Y}) = 1$, which gives

$$f^{**}(\mathbf{Z}) = \sum_{i=1}^{r} \sigma_i(\mathbf{Z}) \cdot 1 - \underbrace{\sum_{i=1}^{r} (1-1)_+}_{=0} = \sum_{i=1}^{r} \sigma_i(\mathbf{Z}) = \|\mathbf{Z}\|_*, \quad r \in \{1, q\}.$$

So lastly we need to show, for $\|\mathbf{Z}\| \leq 1$ and $\|\mathbf{Y}\| > 1$, $f^{**}(\mathbf{Z}) = \|\mathbf{Z}\|_*$, and the proof is completed.

The proof ... 9/12

To show $f^{**}(\mathbf{Z}) = \|\mathbf{Z}\|_*$ for $\|\mathbf{Z}\| \leq 1$ and $\|\mathbf{Y}\| > 1$, again we begin with (11) :

$$f^{**}(\mathbf{Z}) \quad = \quad \sup_{\sigma_i(\mathbf{Y})} \left\{ \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1)_+ \right\}, r \in \{1, q\}$$

The key is to show the red part is smaller than $\|\mathbf{Z}\|_*$, so the $\sup_{\mathbf{Y}}$ gives $\|\mathbf{Z}\|_*$.

Recall from analysis in case forgot : we have $\sup_{x<3} x$ equals to 3, even though x never touch 3 (however $\max_{x<3} x$ has no solution).

Similarly, if we can show the red part $< \|\mathbf{Z}\|_*$, even though the the red part never touch $\|\mathbf{Z}\|_*$, the sup of the red part is $\|\mathbf{Z}\|_*$.

The proof ... 10/12

We now have : $\|\mathbf{Z}\| \leq 1$ and $\|\mathbf{Y}\| > 1$, and by (11) :

$$f^{**}(\mathbf{Z}) \quad = \quad \sup_{\sigma_i(\mathbf{Y})} \left\{ \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1)_+ \right\}, r \in \{1, q\}$$

First we simplify this expression by using (9) : as $\sigma_i(\mathbf{Y}) > 1$ for $i = 1, \ldots, r$, all the $\sigma_i(\mathbf{Y}) - 1$ are positive so we can drop $(\cdot)_+$ and get

$$f^{**}(\mathbf{Z}) \quad = \quad \sup_{\sigma_i(\mathbf{Y})} \left\{ \sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1) \right\}, r \in \{1, q\}$$

Doing so allow us to combine the summations.
Now consider the summation terms

$$\sum_{i=1}^{q} \sigma_i(\mathbf{Z})\sigma_i(\mathbf{Y}) - \sum_{i=1}^{r} (\sigma_i(\mathbf{Y}) - 1) = \sum_{i=1}^{q} \mu_i\sigma_i - \sum_{i=1}^{r} (\sigma_i - 1)$$

where $\mu_i = \sigma_i(\mathbf{Z})$, $\sigma_i = \sigma_i(\mathbf{Y})$.

The proof ... 11/12

A tricky step :

$$\sum_{i=1}^{q}\mu_i\sigma_i-\sum_{i=1}^{r}(\sigma_i-1) \quad = \quad \sum_{i=1}^{q}\mu_i\sigma_i-\sum_{i=1}^{r}(\sigma_i-1)-\sum_{i=1}^{q}\mu_i+\sum_{i=1}^{q}\mu_i$$

Expand the first three summations gives

$$\begin{aligned}&\mu_1\sigma_1+\mu_2\sigma_2+\cdots+\mu_r\sigma_r+\cdots+\mu_q\sigma_q \\ &-\Big((\sigma_1-1)+(\sigma_2-1)+\cdots+(\sigma_r-1)\Big) = \\ &-\Big(\mu_1+\mu_2+\cdots+\mu_r+\cdots+\mu_q\Big)\end{aligned} \quad \begin{aligned}&\sum_{i=1}^{r}(\sigma_i-1)(\mu_i-1) \\ &+\sum_{i=r+1}^{q}(\sigma_i-1)\mu_i\end{aligned}$$

So we have

$$\sum_{i=1}^{q}\mu_i\sigma_i-\sum_{i=1}^{r}(\sigma_i-1)=\sum_{i=1}^{r}(\sigma_i-1)(\mu_i-1)+\sum_{i=r+1}^{q}(\sigma_i-1)\mu_i+\sum_{i=1}^{q}\mu_i$$

The proof ... 12/12

$$\begin{aligned}
\sum_{i=1}^{q}\mu_i\sigma_i - \sum_{i=1}^{r}(\sigma_i - 1) &= \sum_{i=1}^{r}(\sigma_i - 1)(\mu_i - 1) + \sum_{i=r+1}^{q}(\sigma_i - 1)\mu_i + \sum_{i=1}^{q}\mu_i \\
&= \|\mathbf{Z}\|_* + \sum_{i=1}^{r}(\sigma_i - 1)(\mu_i - 1) + \sum_{i=r+1}^{q}(\sigma_i - 1)\mu_i \\
&\overset{\text{see } *}{\leq} \|\mathbf{Z}\|_*.
\end{aligned}$$

So we have $f^{**}(\mathbf{Z}) = \|\mathbf{Z}\|_*$. Together with other cases, we finished the second part of the proof of showing $f^{**}(\mathbf{Z}) = \|\mathbf{Z}\|_*$. □

$$* \qquad \sum_{i=1}^{r}(\sigma_i - 1)(\mu_i - 1) + \sum_{i=r+1}^{q}(\sigma_i - 1)\mu_i \leq 0$$

due to (9) : $\sigma_r(\mathbf{Y}) > 1$ and $\sigma_{r+1}(\mathbf{Y}) \leq 1,\ r \in \{1, q\}$.

End of document