CMSC5743 2021F Homework 2

Due: Nov. 11, 2021

Solutions

All solutions should be submitted to the blackboard in the format of **PDF/MS Word**.

Q1 (13%)

(a) (4%) Consider a formulation as follows.

$$\min_{\beta_1,\beta_2,\beta_3} \quad \frac{1}{2}\left\|\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}3 & 4 & 5\\6 & 7 & 8\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\end{bmatrix}\right\|_2^2 + 3(|\beta_1| + |\beta_2| + |\beta_3|).$$

Please transfer above formulation as ADMM formulation form, that is

$$\begin{aligned}\min_{\boldsymbol{x},\boldsymbol{z}} \quad & f(\boldsymbol{x}) + g(\boldsymbol{z})\\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}.\end{aligned}$$

Show that in your ADMM formulation form, $f(\cdot)$ and $g(\cdot)$ are convex.

(b) (7%) Please use ADMM to handle the formulation in (a). The stopping criterion is set to 2 iterations. All variables are initialized to be 0.

(c) (2%) Compare the coordinate descent method and ADMM to handle the formulation in (a) by discussing the advantages and disadvantages.

A (a)

$$\min_{\beta_1,\beta_2,\beta_3} \frac{1}{2}\left\|\left[\begin{array}{c}1\\2\end{array}\right] - \left[\begin{array}{ccc}3 & 4 & 5\\6 & 7 & 8\end{array}\right]\left[\begin{array}{c}\beta_1\\\beta_2\\\beta_3\end{array}\right]\right\|_2^2 + 3\left(|\beta_1| + |\beta_2| + |\beta_3|\right)$$

$$= \min_{\beta} \frac{1}{2}\|\left[\begin{array}{c}1\\2\end{array}\right] - \left[\begin{array}{ccc}3 & 4 & 5\\6 & 7 & 8\end{array}\right]\beta\|_2^2 + 3\sum_{i=1}^{3}|\beta_i|$$

$$\min_{\beta,\alpha} \frac{1}{2}\|\left[\begin{array}{c}1\\2\end{array}\right] - \left[\begin{array}{ccc}3 & 4 & 5\\6 & 7 & 8\end{array}\right]\beta\|_2^2 + 3\sum_{i=1}^{3}|\alpha_i| \quad \text{st. } \beta - \alpha = 0$$

(b) The answer is not unique since it relies on your predefined ρ. Let $\rho = 1$, and initializes all variables to zero.

Iteration 1:

$$\beta^1 = \left(\left[\begin{array}{cc}-0.2598425197 & 0.188976378\\-0.007874015748 & 0.05118110236\\0.2440944882 & -0.08661417323\end{array}\right]y\right)$$

$$\alpha^1 = \left(\left[\begin{array}{cc}-0.7795275591 & 0.5669291339\\-0.02362204724 & 0.1535433071\\0.7322834646 & -0.2598425197\end{array}\right]y\right)$$

$$w^1 = \left(\left[\begin{array}{cc}-0.5196850394 & 0.3779527559\\-0.0157480315 & 0.1023622047\\0.4881889764 & -0.1732283465\end{array}\right]y\right)$$

Iteration 2:

$$\beta^2 = \left(\begin{bmatrix} -0.4772769546 & 0.3148986299 \\ -0.02985305972 & 0.06412362791 \\ 0.4175708352 & -0.1866513731 \end{bmatrix} y \right)$$
$$= \left(\begin{bmatrix} -0.4772769546 & 0.3148986299 \\ -0.02985305972 & 0.06412362791 \\ 0.4175708352 & -0.1866513731 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$
$$= \begin{bmatrix} 0.1525203052 \\ 0.0983941961 \\ 0.044268089 \end{bmatrix}$$

$$\alpha^2 = \left(\begin{bmatrix} -2.990885982 & 2.078554157 \\ -0.1368032736 & 0.4994574978 \\ 2.717279435 & -1.079639159 \end{bmatrix} y \right)$$
$$= \left(\begin{bmatrix} -2.990885982 & 2.078554157 \\ -0.1368032736 & 0.4994574978 \\ 2.717279435 & -1.079639159 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$
$$\begin{bmatrix} 1.166222332 \\ 0.862111722 \\ 0.558001117 \end{bmatrix}$$

$$w^2 = \left(\begin{bmatrix} 1.993923988 & -1.385702772 \\ 0.09120218243 & -0.3329716652 \\ -1.811519623 & 0.7197594391 \end{bmatrix} y \right)$$
$$= \left(\begin{bmatrix} 1.993923988 & -1.385702772 \\ 0.09120218243 & -0.3329716652 \\ -1.811519623 & 0.7197594391 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$
$$= \begin{bmatrix} -0.777481556 \\ -0.574741148 \\ -0.3720007448 \end{bmatrix}$$

(c) Coordinate descent: Easy to implement The algorithm is scalable since no need to

- lasso problem:

$$\text{minimize} \quad (1/2)\|Ax - b\|_2^2 + \lambda\|x\|_1$$

- ADMM form:

$$\begin{array}{ll} \text{minimize} & (1/2)\|Ax - b\|_2^2 + \lambda\|z\|_1 \\ \text{subject to} & x - z = 0 \end{array}$$

- ADMM:

$$\begin{aligned} x^{k+1} &:= (A^TA + \rho I)^{-1}(A^Tb + \rho z^k - y^k) \\ z^{k+1} &:= S_{\lambda/\rho}(x^{k+1} + y^k/\rho) \\ y^{k+1} &:= y^k + \rho(x^{k+1} - z^{k+1}) \end{aligned}$$

Figure 1: Q1 answer

read the whole dataset into memory Cannot solve two convex functions problem ADMM: ADMM is often slow to converge to high accuracy ADMM has a penalty value ρ which requires to be carefully tuned

Q2 (12%)

(a) (4%) Considering the matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 5 & 8 \\ -4 & 3 & -1 & 2 \end{bmatrix}.$$

Write down the singular values, and corresponding left and right singular vectors for $\boldsymbol{A}$.

(b) (4%) Show $\boldsymbol{A}$ in orthogonal rank 1 form, that is, show $\boldsymbol{A}$ as a sum of outer products that are mutually orthogonal.

(c) (4%) Show the 2-norm and the Frobenius norm of the error in replacing A by the rank 1 approximation in (b).

A (a) $\sigma_1 = \sqrt{62 + \sqrt{1193}}$. Left singular vector: $\begin{bmatrix} 0.981444490820115 \\ 0.191746476992005 \end{bmatrix}$ Right singular vector:

$$\begin{bmatrix} 0.021826804688847 \\ 0.258321361638367 \\ 0.479923777349034 \\ 0.838132944498311 \end{bmatrix}$$

$$\sigma_2 = \sqrt{62 - \sqrt{1193}}$$

Left singular vector:

$$\begin{bmatrix} -0.191746476992005 \\ 0.981444490820115 \end{bmatrix}$$

Right singular vector:

$$\begin{bmatrix} -0.785750364850703 \\ 0.488687171358743 \\ -0.370245449695075 \\ 0.08185059356211 \end{bmatrix}$$

(b) Rank-one form to $A \approx \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$ where σ_1 is the first singular value, u_1 is the first left singular vector, and v_1 is the first right singular vector of A

$$A \approx \begin{bmatrix} 1.003 & 1.995 & 4.999 & 7.994 \\ -3.999 & 3 & -0.996 & 2.002 \end{bmatrix}$$

(c) 2-norm of the error:

$$\|A - \widehat{A}\|_2 \\ = 5.24$$

Frobenius norm of the error:

$$\|A - \widehat{A}\|_F \\ = 5.24$$

Q3 (13%)

(a) (4%) Assume $\boldsymbol{A} = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$ and rank is 2. Computing Singular Value Decomposition for $\boldsymbol{A}$, that is $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.

(b) (4%) Given a matrix $\boldsymbol{A}$, prove that $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ have the same singular values.

(c) (5%) Given a matrix $\boldsymbol{A}$, prove that $\sigma_1 \geq |\lambda|_{\max}$, namely, its largest singular value dominates all eigenvalues.

A (a) $\boldsymbol{U} = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 & -3 \\ 3 & 1 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{45} & \\ & \sqrt{5} \end{bmatrix}$, and $\boldsymbol{V} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$.

(b) Assume $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top$, and $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top$. By the above two equations, we can have a conclusion that $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ have the same singular values.

(c) Recall that multiplying by an orthogonal matrix ($\boldsymbol{Q}$) does not change length. In other words, $\|\boldsymbol{Q}\boldsymbol{x}\| = \|\boldsymbol{x}\|$ since $\|\boldsymbol{Q}\boldsymbol{x}\|^2 = \boldsymbol{x}^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{x} = \|\boldsymbol{x}\|^2$. With $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, we can write

$$\|\boldsymbol{A}\boldsymbol{x}\| = \left\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{x}\right\| = \left\|\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{x}\right\| \leq \sigma_1 \left\|\boldsymbol{I}\boldsymbol{V}^\top\boldsymbol{x}\right\| = \sigma_1\|\boldsymbol{x}\|, \tag{1}$$

where $\boldsymbol{I}$ is an identity matrix. An eigenvector has $\|\boldsymbol{A}\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$. So Eq. (1) says that $|\lambda|\|\boldsymbol{x}\| \leq \sigma_1\|\boldsymbol{x}\|$. Then $\sigma_1 \geq |\lambda|_{\max}$.

Q4 (12%)

(a) (4%) Given $\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 5 & 6 & 7 \\ 8 & 9 & 10 \end{bmatrix}$. Calculate $\boldsymbol{A} \otimes \boldsymbol{B}$ and $\boldsymbol{B} \otimes \boldsymbol{A}$ (Kronecker product).

(b) (4%) Let us have a rank-1 tensor $\boldsymbol{\mathcal{X}} = \left[\begin{array}{cc|cc} 1 & \sqrt{2} & \sqrt{2} & 2 \\ \sqrt{2} & 2 & 2 & 2\sqrt{2} \end{array} \right]$. Calculating $\|\boldsymbol{\mathcal{X}}\|_F$ (Frobenius norm)

(c) (4%) Write down the 1-flattening of $\boldsymbol{\mathcal{X}}$ (1-flattening means only the first dim to flatten).

A (a) $\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} 5 & 6 & 7 & 10 & 12 & 14 \\ 8 & 9 & 10 & 16 & 18 & 20 \\ 15 & 18 & 21 & 20 & 24 & 28 \\ 24 & 27 & 30 & 32 & 36 & 40 \end{bmatrix}$ and $\boldsymbol{B} \otimes \boldsymbol{A} = \begin{bmatrix} 5 & 10 & 6 & 12 & 7 & 14 \\ 15 & 20 & 18 & 24 & 21 & 28 \\ 8 & 16 & 9 & 18 & 10 & 20 \\ 24 & 32 & 27 & 36 & 30 & 40 \end{bmatrix}$.

(b) $\|\mathcal{X}\|_F = \sqrt{1^2 + (\sqrt{2})^2 + (\sqrt{2})^2 + 2^2 + (\sqrt{2})^2 + 2^2 + 2^2 + (2\sqrt{2})^2} = 3\sqrt{3}$.

(c) $\mathcal{X}_{(1)} = \begin{bmatrix} 1 & \sqrt{2} & \sqrt{2} & 2 \\ \sqrt{2} & 2 & 2 & 2\sqrt{2} \end{bmatrix}$.

Q5 (12%) This exercise provide an example to show the benefit of SVD over Eigen-decomposition.

Suppose $\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$. Assume in real-world scenarios, due to some trouble, the $A[4,1]$ entry changes from zero to $\frac{1}{60000}$ and mark the new matrix as $\boldsymbol{A}'$. Now $\boldsymbol{A}'$ is a full-rank matrix.

(a) (4%) Calculate the eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{A}'$.

(b) (4%) Calculate the singular value of $\boldsymbol{A}$ and $\boldsymbol{A}'$.

(c) (4%) What do you observe from the calculation results.

A (a) Eigenvalue of $\boldsymbol{A}$ is 0, 0, 0, 0.
Eigenvalue of $\boldsymbol{A}'$ is $\frac{1}{10}, \frac{i}{10}, -\frac{1}{10}, -\frac{i}{10}$.
(b) Singular Value of $\boldsymbol{A}$ is 3, 2, 1, 0.
Singlar value of $\boldsymbol{A}'$ is 3, 2, 1, $\frac{1}{60000}$.
(c) The change of the singular value is more stable even if the entry in $\boldsymbol{A}$ change.

Q6 (13%) Construct a rank-1 matrix $\boldsymbol{A}$ satisfying all the following conditions.

- $\boldsymbol{A}\boldsymbol{v} = 12\boldsymbol{u}$;
- $\boldsymbol{v} = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$;
- $\boldsymbol{u} = \frac{1}{3}\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}$.

A A matrix with $\boldsymbol{A}\boldsymbol{v} = 12\boldsymbol{u}$ would have $\boldsymbol{u}$ in its column space. Since rank of matrix $\boldsymbol{A}$ is 1, $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{w}^\top$ for some vector $\boldsymbol{w}$. Since $\boldsymbol{v}$ is a unit vector and $\boldsymbol{A}\boldsymbol{v} = 12\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{v}$, then

$$\boldsymbol{A} = 12\boldsymbol{u}\boldsymbol{v}^\top = \begin{bmatrix} 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \\ 2 & 2 & 2 & 2 \end{bmatrix}$$

Q7 (12%) Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a matrix, and let $\boldsymbol{A}^\top$ be the transposed matrix of $\boldsymbol{A}$.

(a) (4%) Show that both $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ are positive semidefinite.

(b) (4%) Show that $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ have exactly the same nonzero eigenvalues.

(c) (4%) If we know that $m = n$ and $\boldsymbol{A}$ is positive semidefinite, show that the eigenvalues and singular values of $\boldsymbol{A}$ are exactly the same.

A Answer

(a) Take any $\boldsymbol{x} \in \mathbb{R}^n$, we have $\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \geq 0$. Similarly, take any $\boldsymbol{x} \in \mathbb{R}^m$, we have $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{x} = \|\boldsymbol{A}^\top\boldsymbol{x}\|_2^2 \geq 0$.

(b) Let $\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, where $\lambda \neq 0, \boldsymbol{x} \neq 0$. Absolutely, $\boldsymbol{A}\boldsymbol{x} \neq 0$, otherwise $\lambda\boldsymbol{x} = 0$, which is a contradition. Obviously, We have $\boldsymbol{A}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}) = \lambda(\boldsymbol{A}\boldsymbol{x})$. Therefore, λ is also an eigenvalue of $\boldsymbol{A}\boldsymbol{A}^\top$, and $\boldsymbol{A}\boldsymbol{x}$ is the corresponding eigenvector. It is very similar to show that eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$ are also eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$.

(c) $\boldsymbol{A}$ is PSD, so $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}^2$. Let $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$, then $\boldsymbol{A}^2 \boldsymbol{x} = \lambda \boldsymbol{A}\boldsymbol{x} = \lambda^2 \boldsymbol{x}$. Therefore λ^2 is an eigenvalue of $\boldsymbol{A}^2$. Obviously $\lambda \geq 0$ as $\boldsymbol{A}$ is PSD, therefore the corresponding singular value $\sigma = \sqrt{\lambda^2} = \lambda$.

Q8 (13%) Assume that $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^n$ is a matrix, and $\mathrm{tr}(\cdot)$ is the trace function, *i.e.* $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^n a_{ii}$ is the sum of diagonal entries.

(a) (3%) Show that $\|\boldsymbol{A}\|_F^2 = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})$.

(b) (3%) Assume that $\boldsymbol{M} \in \mathbb{R}^n$ is positive semidefinite, show that $\mathrm{tr}(\boldsymbol{M}) = \sum_{i=1}^n \lambda_i$, where $\lambda_1, \cdots, \lambda_n$ are eigenvalues of $\boldsymbol{M}$.

(c) (7%) (**<u>Hard</u>**) You are given the following inequality,

$$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \sum_{i=1}^n \sigma_i(\boldsymbol{A})\sigma_i(\boldsymbol{B}),$$

where $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_n(\boldsymbol{A})$ and $\sigma_1(\boldsymbol{B}) \geq \cdots \geq \sigma_n(\boldsymbol{B})$ are singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Find an optimal solution to the following low rank approximation,

$$\min_{\substack{\boldsymbol{X} \in \mathbb{R}^{m \times n} \\ \mathrm{rank}(\boldsymbol{X}) \leq k}} \|\boldsymbol{X} - \boldsymbol{Y}\|_F^2,$$

where matrix $\boldsymbol{Y} \in \mathbb{R}^{m \times n}$ is fixed and has a full rank. The rank upper bound $k \leq \min\{m, n\}$. (Hint: use the results in the above questions, and consider singular values of the two matrices.)

A Answer

(a) The i-th diagonal entry of $\boldsymbol{A}^\top \boldsymbol{A}$ is the squared sum of the i-th column of $\boldsymbol{A}$, *i.e.* $\sum_{k=1}^n a_{ki}^2$. Then we have

$$\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A}) = \sum_{k=1}^n \sum_{i=1}^n a_{ki}^2 = \|\boldsymbol{A}\|_F^2.$$

(b) $\boldsymbol{M}$ is PSD, so we have the eigendecomposition $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ is a diagonal matrix with entries being eigenvalues of $\boldsymbol{M}$, and $\boldsymbol{Q}$ is orthogonal. Then we have

$$\mathrm{tr}(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top) = \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{Q}^\top \boldsymbol{Q}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^n \lambda_i.$$

(c) (**<u>Hard</u>**) Assume that the rank of matrix $\boldsymbol{X}$ is r, *i.e.* $\mathrm{rank}(\boldsymbol{X}) = r \leq k$. Let $l = \min\{m, n\}$ be the (full) rank of matrix $\boldsymbol{Y}$.

$$\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 = \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{Y}^\top \boldsymbol{Y}) - 2\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{Y})$$

Suppose that singular values of $\boldsymbol{X}$ are $\tilde{\sigma}_1 \geq \cdots \tilde{\sigma}_r > \tilde{\sigma}_{r+1} = \cdots = \tilde{\sigma}_l = 0$, and singular values of $\boldsymbol{Y}$ are $\sigma_1 \geq \cdots \sigma_l > 0$. According to the given inequality, we have

$$
\begin{aligned}
\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 &= \sum_{i=1}^{l}\tilde{\sigma}_i^2+\sum_{i=1}^{l}\sigma_i^2-2\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{Y}) \\
&\geq \sum_{i=1}^{l}\tilde{\sigma}_i^2+\sum_{i=1}^{l}\sigma_i^2-2\sum_{i=1}^{l}\sigma_i\tilde{\sigma}_i=\sum_{i=1}^{l}(\tilde{\sigma}_i-\sigma_i)^2 \\
&= \sum_{i=1}^{r}(\tilde{\sigma}_i-\sigma_i)^2+\sum_{i=r+1}^{l}\sigma_i^2 \geq \sum_{i=r+1}^{l}\sigma_i^2 \geq \sum_{i=k+1}^{l}\sigma_i^2.
\end{aligned}
$$

To make the equality hold, we must have $r=k$, and $\tilde{\sigma}_i=\sigma_i$ for $i=1,\cdots,k$. Assume that the SVD of $\boldsymbol{Y}$ is $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal. Take the first k columns of $\boldsymbol{U}$ as $\boldsymbol{U}_k$, the first k columns of $\boldsymbol{V}$ as $\boldsymbol{V}_k$, and let $\boldsymbol{\Sigma}_k=\mathrm{diag}(\sigma_1,\cdots,\sigma_k)$. You can check that $\boldsymbol{X}^*=\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^\top$ is of rank k, and the equality holds. Therefore, $\boldsymbol{X}^*=\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^\top$ must be an optimal solution.