CMSE 820 Homework 5. Due 13 Oct, 2020.

I. Write $\mathbf{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$ for the column vector whose components are all $1's$. Define the centering matrix H by

$$H := I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T,$$

where I_n is the $n \times n$ identity matrix, and $\mathbf{1}\mathbf{1}^T$ denotes the matrix multiplication of the two vectors $\mathbf{1}$ and $\mathbf{1}^T$.

(a) Prove that $H^T = H$ and $H^2 = H$. (Remark: this means H is an orthogonal projection matrix).

Solution. Let Y be a matrix defined by the multiplication of $\mathbf{1}\mathbf{1}^T$. We can write

$$Y = \mathbf{1}\mathbf{1}^T = (1)_{ij} \in \mathbb{R}^{n \times n},$$

where all of the ijh elements of Y are 1. Thus,

$$H = I_n - \frac{1}{n}(1)_{ij} = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \ldots & & & \\ -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \ldots & & \\ -\frac{1}{n} & -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \ldots & \\ \vdots & & & \ddots & & \\ & & & \ldots & -\frac{1}{n} & 1-\frac{1}{n} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

From inspection, we see H is symmetric, so $H^T = H$.

Additionally,

$$H^2 = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \ldots & & & \\ -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \ldots & & \\ -\frac{1}{n} & -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \ldots & \\ \vdots & & & \ddots & & \\ & & & \ldots & -\frac{1}{n} & 1-\frac{1}{n} \end{pmatrix} \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \ldots & & & \\ -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \ldots & & \\ -\frac{1}{n} & -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \ldots & \\ \vdots & & & \ddots & & \\ & & & \ldots & -\frac{1}{n} & 1-\frac{1}{n} \end{pmatrix}.$$

From this, we see that the diagonal elements of the matrix multiplication are equal to $(1 - \frac{1}{n})^2 + (n-1)/n^2$ and all of the off-diagonal elements are $-\frac{2}{n}(1 - \frac{1}{n}) + (n-2)/n^2$. Simplifying these expressions, we see that

$$(H^2)_{ij} = \begin{cases} 1 - \frac{1}{n}, & i = j \\ -\frac{1}{n}, & i \neq j \end{cases}.$$

Therefore, $H^2 = H$. ■

(b) Prove that $\text{Ker}(H) = \text{span}\{\mathbf{1}\}$ and $\text{Ran}(H) = \{u \in \mathbb{R}^n : u^T\mathbf{1} = 0\}$.

Solution. Let $x \in \mathbb{R}^n$ be a vector from the kernel of H such that $Hx = 0$. Let $x_1, \ldots, x_n$ be the elements of x. It follows that

$$\begin{aligned} Hx &= \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)x \\ &= Ix - \frac{1}{n}(1)_{ij}x = 0 \\ \Rightarrow x &= \frac{1}{n}(1)_{ij}x. \end{aligned}$$

Thus,

$$x_j = \sum_{i=1}^{n} \frac{x_i}{n}.$$

x_j is therefore the average value of the vector x. For this expression to hold true for any j, all of the elements of x need to be the same, $x_i = x_j, \forall i$. We can therefore write x as $x = c\mathbf{1}$ where $c \in \mathbb{R}$, which implies $x \in \text{span}\{\mathbf{1}\}$, and $\text{Ker}(H) \subseteq \text{span}\{\mathbf{1}\}$.

Suppose now instead that $x \in \text{span}\{\mathbf{1}\}$ which means $x = c(1)_j$ where $c \in \mathbb{R}$ and the jth element of x is 1. We can therefore show that the jth row of H times the column vector x yields,

$$\left(1 - \frac{1}{n}\right) c - \frac{(n-1)c}{n} = c - \frac{c}{n} - c + \frac{c}{n} = 0.$$

Therefore, $Hx = 0$ which implies $x \in \text{Ker}(H)$ and $\text{span}\{\mathbf{1}\} \subseteq \text{Ker}(H)$. ■

Suppose $x \in \text{Ran}(H)$. In homework 1, we proved that if $x \in \text{Ran}(A)$ where A is an orthogonal projection matrix, $Ax = x$. Using this property

$$Hx = x$$
$$x - \frac{1}{n}\mathbf{1}\mathbf{1}^T x = x.$$

This implies that $\frac{1}{n}\mathbf{1}\mathbf{1}^T x = 0 \Rightarrow \mathbf{1}\mathbf{1}^T x = 0$. This further implies that the sum of the element of x must be zero. Suppose we care about the non trivial solution where x is not the zero vector. We can conclude that $x \in \{u \in \mathbb{R}^n : u^T\mathbf{1} = 0\}$ since $u^T\mathbf{1} = \mathbf{1}^T u = 0$, and $\text{Ran}(H) \subseteq \{u \in \mathbb{R}^n : u^T\mathbf{1} = 0\}$.

Suppose now instead that $x \in \{u \in \mathbb{R}^n : u^T\mathbf{1} = 0\}$ is a vector whose elements sum to 0. It follows that

$$Hx = (I - 1/n(1)_{ij})x = x - \frac{1}{n}0 = x.$$

Since $Hx = x$, from the proof in homework 1, we can conclude $x \in \text{Ran}(H)$, and $\{u \in \mathbb{R}^n : u^T\mathbf{1}\} \subseteq \text{Ran}(H)$. ■

II. Let $X \in \mathbb{R}^{p\times n}$ be a sample matrix whose columns consist of sample points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$. We assume the sample points are not centered, i.e., their mean $\mu := \frac{1}{n}(x^{(1)} + \cdots + x^{(n)})$ is not the zero vector in $\mathbb{R}^p$. Let H be the centering matrix defined as above.

(a) Define $Y := XH$ and denote the columns of Y by $y^{(1)}, \ldots, y^{(n)}$. Prove that the sample points $y^{(1)}, \ldots, y^{(n)}$ are centered. (Remark: this means right multiplication by H centers the sample points in the sample matrix X.)

Solution. Let $\overline{x} = x^{(1)} + \cdots + x^{n} \in \mathbb{R}^p$ be the sum of all of the feature vectors (column sum of the sample matrix X) such that $\mu = \frac{1}{n}\overline{x}$. Thus,

$$\begin{aligned}
Y &= XH \\
&= X(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) \\
&= X - \frac{1}{n}\begin{pmatrix} x^{(1)} & x^{(2)} & \dots & x^{(n)} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \\
&= X - \frac{1}{n}\left(x^{(1)} + x^{(2)} + \dots + x^{(n)}\right)\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \\
&= X - \frac{1}{n}\overline{x}\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \\
&= X - \mu\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}.
\end{aligned}$$

We can write the second term in the expression as

$$\mu\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} = \begin{pmatrix} \mu & \mu & \dots & \mu \end{pmatrix} \in \mathbb{R}^{p\times n},$$

so

$$Y = XH = X - \begin{pmatrix} \mu & \mu & \dots & \mu \end{pmatrix}.$$

Since $y^{(i)} = x^{(i)} - \mu$,

$$\frac{1}{n}\sum_{i=1}^{n} y^{(i)} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} - \mu = \frac{1}{n}(\overline{x} - n\mu) = \mu - \mu = 0.$$

The matrix Y is thus centered. ■

(b) Recall that $A \subset \mathbb{R}^p$ is said to be a d-dimensional affine subspace if and only if there is a d-dimensional vector subspace $S \subset \mathbb{R}^p$ and a fixed vector $a \in \mathbb{R}^p$ such that

$$A = \{a + v : v \in S\}.$$

For example, a 1D affine subspace of $\mathbb{R}^2$ is a line. In contrast, a $1D$ vector subspace of $\mathbb{R}^2$ is a line that passes through the origin.

Denote by A_d the d-dimensional affine space that "best fit" the non-centered sample points $x^{(1)}, \dots, x^{(n)}$ ("best fit" means the sum of the squared distances from these sample points to the affine space is minimized.) Prove that

$$A_d = \{\mu + v : v \in \text{span}\{u^{(1)}, \dots, u^{(d)}\}\}$$

where $u^{(1)}, \dots, u^{(d)}$ are the eigenvectors associated to the d largest eigenvalues of the matrix XHX^T.

Solution. Let A_d be the d-dimensional affine space that minimizes the sum of squared distances from the sample points. Suppose $S = \{a - \mu : a \in A\}$ and the mean of the sample points is μ. Let $Y = XH$, or $y^{(i)} = x^{(i)} - \mu$ for $i = 1, \dots, n$ be the centered sample points. Shifting A to S by subtracting μ does not effect the sum of squared distances to the sample points. In this case, the sum of squared distances from $y^{(i)}$ to S is equal to the sum of squared distances from $x^{(i)}$ to A. Additionally, since S is now centered, it must contain the zero vector and is therefore a vector subspace. Thus, S must be the d-dimensional vector subspace that minimizes the sum of squared distances from the centered sample points $y^{(i)}$ to the vector subspace.

It follows that S must be spanned by the eigenvectors associated to the d largest eigenvalues of YY^T. However,

$$YY^T = (XH)(XH)^T = XHH^TX^T = XHHX^T = XHX^T,$$

so S is spanned by the d eigenvectors associated to the d largest eigenvalues of XHX^T. Let $u^{(i)}$ be the eigenvector associated to the ith largest eigenvalue of XHX^T. We can write S as $S = \text{span}\{u^{(1)}, \dots, u^{(d)}\}$. Since $S = \{a - \mu : a \in A\}$, it follows that $A_d = \{v + \mu : v \in S\}$, and finally $A_d = \{v + \mu : v \in \text{span}\{u^{(1)}, \dots, u^{(d)}\}\}$. ■

III. Let $X \in \mathbb{R}^{p\times n}$ $(p < n)$ be a sample matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ where r is the rank of X. The nuclear norm of X is defined by

$$\|X\|_* := \sigma_1 + \cdots + \sigma_r.$$

We write $I_{p,n}$ for the $p \times n$ matrix with 1 on the "diagonal," that is,

$$I_{p,n} = \begin{pmatrix} 1 & & & & 0 & \dots & 0 \\ & 1 & & & 0 & \dots & 0 \\ & & \ddots & & 0 & \dots & 0 \\ & & & 1 & 0 & \dots & 0 \end{pmatrix}_{p\times n}.$$

(a) Prove that

$$\|X\|_* = \max_{Q=UI_{p,n}V^T} \text{tr}(Q^TX),$$

where U and V are $p\times p$ and $n\times n$ orthogonal matrices respectively. Here, the maximum is taken over all the matrices Q which can be written as $Q = UI_{p,n}V^T$ for some orthogonal matrices U, V. (Hint: first show $\text{tr}(Q^TX) \leq \|X\|_*$ for an arbitrary Q of this form, then examine when the maximum is achieved.)

Solution. Let the SVD of X be written as $X = U\Sigma V^T$ where U and V are orthogonal matrices. Suppose we fix Q to be $Q = UI_{p,n}V^T$. Thus, $\text{tr}(Q^TX) = \text{tr}(VI_{n,p}U^TU\Sigma V^T) = \text{tr}(VI_{n,p}\Sigma V^T)$. Let

$$\Sigma' = I_{n,p}\Sigma = \begin{pmatrix} \sigma_1(X) & & & & & & \\ & \sigma_2(X) & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r(X) & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix}_{n\times n}.$$

By the properties of the trace $\mathrm{tr}(V\Sigma' V^T) = \mathrm{tr}(V^T V \Sigma') = \mathrm{tr}(\Sigma') = \sum_{i=1}^r \sigma_i(X) = \|X\|_*$. Since $\mathrm{tr}(Q^T X) = \|X\|_*$ for this choice of Q, it follows that $\max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T X) \geq \|X\|_*$.

Consider now an arbitrary Q.

$$\begin{aligned}
\mathrm{tr}(Q^T X) &= \mathrm{tr}(Q^T U \Sigma V^T) \\
&= \mathrm{tr}(V^T Q^T U \Sigma) \\
&= \mathrm{tr}((U^T Q V)^T \Sigma) \\
&\leq \sum_{i=1}^r \sigma_i(X)\sigma_i(Q),
\end{aligned}$$

where we utilized the properties of the trace (invariance under cyclic permutations) to equate $\mathrm{tr}(Q^T U\Sigma V^T) = \mathrm{tr}(V^T Q^T U \Sigma)$ and Von-Neumann's inequality in the final line. Since $Q = UI_{p,n}V^T$ for some arbitrary orthogonal matrices U, V (not necessarily equal to the U, V in the SVD of X), we know $\sigma_i(Q) = 1, \forall i = 1, \ldots, r$. Therefore, $\mathrm{tr}(Q^T X) \leq \sum_{i=1}^r \sigma_i(X) = \|X\|_*$. It thus follows that $\max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T X) \leq \sum_{i=1}^r \sigma_i(X) = \|X\|_*$.

Therefore,

$$\max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T X) = \|X\|_*. \blacksquare$$

(b) Prove that $\|X\|_*$ is a convex function of X.

Solution. Consider two arbitrary matrices $A, B \in \mathbb{R}^{p\times n}$.

$$\begin{aligned}
\|A+B\|_* &= \max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T(A+B)) \\
&= \max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T A + Q^T B).
\end{aligned}$$

Since the trace is linear,

$$\|A+B\|_* = \max_{Q=UI_{p,n}V^T} (\mathrm{tr}(Q^T A) + \mathrm{tr}(Q^T B)),$$

which implies that

$$\begin{aligned}
\|A+B\|_* &= \max_{Q=UI_{p,n}V^T} (\mathrm{tr}(Q^T A) + \mathrm{tr}(Q^T B)) \\
&\leq \max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T A) + \max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T B) \\
\|A+B\|_* &\leq \|A\|_* + \|B\|_*.
\end{aligned}$$

Additionally, if we consider $\|\lambda A\|_*$ for some positive real value λ, we see that

$$\|\lambda A\|_* = \max_{Q=UI_{p,n}V^T} \mathrm{tr}(Q^T \lambda A) = \max_{Q=UI_{p,n}V^T} \lambda \mathrm{tr}(Q^T A) = \lambda \|A\|_*,$$

again by the linearity of the trace. From these properties, we know

$$\|\lambda A + (1-\lambda)B\| \leq \lambda\|A\|_* + (1-\lambda)\|B\|_*,$$

and therefore the nuclear norm is a convex function. $\blacksquare$

IV. Let $X \in \mathbb{R}^{p\times n}$ be a sample matrix with SVD $X = U\Sigma V^T$ and singular values $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_r(X)$. Find the minimizer of the following optimization problem

$$\min_A \|X - A\|_F^2 + \lambda\|A\|_*$$

in terms of $U, V, \sigma_i(X), i = 1, \ldots, r$. Here $\lambda > 0$ is a constant and $\|A\|_*$ is the nuclear norm of A.

Solution. Let $A \in \mathbb{R}^{p\times n}$ be a matrix whose singular values decomposition can be written as $A = U_2\Sigma_2V_2^T$ where U_2 and V_2 are orthogonal matrices. We can rewrite $\|X - A\|_F^2$ as,

$$\begin{aligned}\|X - A\|_F^2 &= \|U\Sigma V^T - U_2\Sigma_2V_2^T\|_F^2 \\ &= \|\Sigma - U^TU_2\Sigma_2V_2^TV\|_F^2 \\ &= \|\Sigma - \tilde{U}\Sigma_2\tilde{V}^T\|_F^2 \\ &= \|\Sigma\|_F^2 + \|\Sigma_2\|_F^2 - 2(\Sigma, \tilde{U}\Sigma_2\tilde{V}^T)_F,\end{aligned}$$

where in the above expression $\tilde{U} = U^TU_2$ and $\tilde{V}^T = V_2^TV$. Therefore,

$$\|X - A\|_F^2 + \lambda\|A\|_* = \|\Sigma\|_F^2 + \|\Sigma_2\|_F^2 - 2(\Sigma, \tilde{U}\Sigma_2\tilde{V}^T)_F + \lambda\sum_{i=1}^{\min\{p,n\}}\sigma_i(A),$$

where $\sigma_i(A)$ is the ith singular value of A. Using the Von-Neumann inequality,

$$\begin{aligned}\|X - A\|_F^2 + \lambda\|A\|_* &\geq \|\Sigma\|_F^2 + \|\Sigma_2\|_F^2 - 2\sum_{i=1}^{\min\{p,n\}}\sigma_i(X)\sigma_i(A) + \lambda\sum_{i=1}^{\min\{p,n\}}\sigma_i(A) \\ &= \|\Sigma\|_F^2 + \sum_{i=1}^{\min\{p,n\}}\sigma_i(A)^2 - 2\sigma_i(X)\sigma_i(A) + \lambda\sigma_i(A).\end{aligned}$$

In order to minimize $\|X - A\|_F^2 + \lambda\|A\|_*$, it suffices to minimize the sum term on the right side of the above inequality. Since the expression being summed over is a quadratic function in terms of the singular values of A and we can assume X is some given, fixed matrix, we can minimize the expression by minimizing each term of the sum. Since each term is a polynomial in terms of the singular values of A, the derivative is smooth and continuous. Thus, we can minimize the terms by solving

$$\begin{aligned}\frac{\partial}{\partial\sigma_i(A)}\left(\sigma_i(A)^2 - 2\sigma_i(X)\sigma_i(A) + \lambda\sigma_i(A)\right) &= 0 \\ 2\sigma_i(A) - 2\sigma_i(X) + \lambda &= 0 \\ \sigma_i(A) &= \sigma_i(X) - \frac{\lambda}{2}.\end{aligned}$$

We must constrain $\sigma_i(A) > 0$, and thus $\sigma_i(A) = (\sigma_i(X) - \frac{\lambda}{2})_+$. Therefore, the right hand side of the inequality obtained using Von-Neumann's inequality is minimized when the ith singular value of A is equal to the ith singular value of X adjusted by $\frac{\lambda}{2}$. Additionally, we know that equality of the two expressions obtained using Von-Neumann's inequality hold when $\tilde{U} = I$ and $\tilde{V}^T = I$, where I represents the identity matrix in the corresponding dimension of each matrix respectively. This implies, $U_2 = U$ and $V_2^T = V^T$. Thus, the matrix A that minimizes $\|X - A\|_F^2 + \lambda \|A\|_*$ is

$$A = U \left(\sigma_i(X) - \frac{\lambda}{2} \right)_+ V^T = A = U S_{\lambda/2}(\sigma_i(X)) V^T. \blacksquare$$