CMSE 820 HW 4

Hao Lin

October 4, 2018

1 Question 1

The dual norm $\|\cdot\|^*$ associated with a norm $\|\cdot\|$ on $\mathbb{R}^{p\times n}$ is defined as

$$\|G\|^* = \max_{\|B\|\leq 1} \mathrm{Tr}(B^T G).$$

1.1 a.

Prove that the dual norm associated with the spectral norm is the nuclear norm.
Proof: Let $A = U\Sigma V^T$ be the compact SVD for $A \in \mathbb{R}^{p\times n}$. Consider $B = UV^T$. By construction, $\|B\|_2 = \sigma_1(B) = 1$. (Incidentally, note that **the spectral norm coincides with the matrix 2-norm $\|\cdot\|_2$ induced from ℓ_2 norm of vectors**.) It follows that

$$\|A\|_2^* \geq \mathrm{Tr}(B^T A) = \mathrm{Tr}(VU^TU\Sigma V^T) = \mathrm{Tr}(V^TV\Sigma) = \mathrm{Tr}(\Sigma) = \|A\|_*. \, (*)$$

In order to complete the rest of the proof, we first need to prove the following **lemma**:

$$\boxed{\forall A, B \in \mathbb{R}^{p\times n}, \mathrm{Tr}(B^T A) \leq \|B\|_2 \|A\|_*.}$$

Proof of the lemma: Let $A = U\Sigma V^T$ be the compact SVD for A, where $\Sigma \in \mathbb{R}^{r\times r}$ for some $r \leq \min\{p, n\}$.

$$\mathrm{Tr}(B^T A) = \mathrm{Tr}(B^T U\Sigma V^T) = \mathrm{Tr}(V^T B^T U\Sigma) = \mathrm{Tr}((U^T BV)^T \Sigma) = \mathrm{Tr}(S^T\Sigma),$$

where $S = U^T BV \in \mathbb{R}^{r\times r}$. Note that $\|S\|_2 = \|U^TBV\|_2 \leq \|U^T\|_2\|B\|_2\|V\|_2 = \|B\|_2$. (The inequality follows immediately from the definition of induced matrix norms.) Applying Von Neumann's trace inequality for square matrices, we have

$$\mathrm{Tr}(B^T A) = \mathrm{Tr}(S^T\Sigma) \leq |\mathrm{Tr}(S^T\Sigma)| \leq \sum_{i=1}^r \sigma_i(S)\sigma_i(\Sigma) \leq \|S\|_2 \sum_{i=1}^r \sigma_i(A) \leq \|B\|_2\|A\|_*.$$

It follows immediately that

$$\forall B \in \mathbb{R}^{p\times n} \text{ such that } \|B\|_2 \leq 1, \mathrm{Tr}(B^T A) \leq \|A\|_*. \implies \|A\|_2^* \leq \|A\|_* (**)$$

(*) and (**) together imply that $\|A\|_2^* = \|A\|_*$.

1.2 b.

Prove that the dual norm associated with nuclear norm is the spectral norm.
Proof: Following from the lemma proven in (a), $\forall A, B \in \mathbb{R}^{p\times n}$ with $\|B\|_* \le 1$, we have

$$\mathrm{Tr}(B^T A) = \mathrm{Tr}((A^T B)^T) = \mathrm{Tr}(A^T B) \le \|A\|_2 \|B\|_* \le \|A\|_2. \, (***)$$

Let $A = U\Sigma V^T$ be the compact SVD of A. In particular, $A\vec{v}_1 = \sigma_1(A)\vec{u}_1$, where $\vec{u}_1$ and $\vec{v}_1$ are the first column vectors of U and V, respectively. Let $B = \vec{v}_1\vec{u}_1^T$ be a rank-1 matrix. The definition of B is an SVD of B itself, so $\|B\|_* = 1$.

$$\mathrm{Tr}(B^T A) = \mathrm{Tr}(\vec{v}_1\vec{u}_1^T U\Sigma V^T) = \mathrm{Tr}((V^T\vec{v}_1)(\vec{u}_1^T U)\Sigma) = \Sigma_{11} = \|A\|_2.$$

That is, the equality in (***) can be achieved. It follows that the dual norm associated with $\|\cdot\|_*$ is $\|\cdot\|_2$.

1.3 c.

Firstly, we check that the nuclear norm is indeed a norm.

1. Since singular values are nonnegative, $\forall A \in \mathbb{R}^{p\times n}$, $\|A\|_* \ge 0$.
2. $\|A\|_* = 0 \Leftrightarrow \sigma_i = 0$, for $i = 1, ..., \min(\mathrm{p,n}) \Leftrightarrow A = 0$.
3. Consider a matrix $A \in \mathbb{R}^{p\times n}$ and let $A = U\Sigma V^T$ be an SVD of A. $\forall \alpha \in \mathbb{R}$, $\alpha A = (\mathrm{sign}(\alpha)U)(|\alpha|\Sigma)V^T$ is a valid SVD for the matrix αA. Hence, $\|\alpha A\|_* = |\alpha|\|A\|_*$.
4. The triangle inequality: Consider some $A, B \in \mathbb{R}^{p\times n}$. Since $\|\cdot\|_*$ is the dual norm associated with $\|\cdot\|_2$, there exists a matrix $X \in \mathbb{R}^{p\times n}$ with $\|X\|_2 \le 1$ such that

$$\|A + B\|_* = \mathrm{Tr}(X^T(A + B)) = \mathrm{Tr}(X^T A) + \mathrm{Tr}(X^T B) \le \|A\|_* + \|B\|_*.$$

The convexity of $\|\cdot\|_*$ follows naturally from properties of norms:

$$\forall A, B \in \mathbb{R}^{p\times n}, \forall t \in [0, 1], \|tA + (1 - t)B\|_* \le t\|A\|_* + (1 - t)\|B\|_*.$$

2 Question 2

The first M entries of $x \in \mathbb{R}^p$ are missing. $\mu \in \mathbb{R}^p$ and $U \in \mathbb{R}^{p\times d}$ are known. The optimization problem is

$$\underset{y,\, x_U}{\arg\min} \|x - \mu - Uy\|^2.$$

The corresponding Lagrangian function is $\mathcal{L}(y, x_U) = \|x - \mu - Uy\|^2$, whose partial derivatives are

$$\frac{\partial \mathcal{L}}{\partial y} = -2U^T(x - \mu - Uy) = 0, \; \frac{\partial \mathcal{L}}{\partial x} = 2(x - \mu - Uy) = 0.$$

Rewrite

$$x = \begin{bmatrix} X_U \\ X_O \end{bmatrix}, \mu = \begin{bmatrix} \mu_U \\ \mu_O \end{bmatrix}, U = \begin{bmatrix} U_U \\ U_O \end{bmatrix}$$

. Then the necessary conditions for the optimization problem become

$$\begin{aligned} &U_U^T(x_U - \mu_U) + U_O^T(x_O - \mu_O) = y, \\ &x_U - \mu_U - U_U y = 0, \\ &x_O - \mu_O - U_O y = 0. \end{aligned}$$

If we further assume that U_O is of full rank (so that $U_O^T U_O$ is invertible), we can solve the equations above simultaneously,

$$y = (I_d - U_U^T U_U)^{-1} U_O^T(x_O - \mu_O) = (U_O^T U_O)^{-1} U_O^T(x_O - \mu_O),$$

$$x_U = \mu_U + U_U(U_O^T U_O)^{-1} U_O^T(x_O - \mu_O).$$

Note that a necessary condition for U_O to be full rank is that the number of observed entries is at least d.

3 Question 3

Properties of the $\ell_{2,1}$ Norm.

3.1

a. If $\mathbf{x} \neq \mathbf{0}$, $\|\mathbf{x}\|_2$ is a convex differentiable function of $\mathbf{x}$, and hence

$$\partial\|\mathbf{x}\|_2 = \nabla\|\mathbf{x}\|_2 = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}.$$

b. If $\mathbf{x} = \mathbf{0}$, we want to show that $\partial\|\mathbf{x}\|_2(\mathbf{x} = \mathbf{0}) = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq 1\}$. $\forall\mathbf{y}$ and $\forall\mathbf{w}$ such that $\|\mathbf{w}\|_2 \leq 1$, by the the Cauchy-Schwarz inequality,

$$\mathbf{y}^T\mathbf{w} \leq |\mathbf{y}^T\mathbf{w}| \leq \|\mathbf{y}\|_2\|\mathbf{w}\|_2 \leq \|\mathbf{y}\|_2.$$

On the other hand, $\forall\mathbf{w}'$ such that $\|\mathbf{w}'\|_2 > 1$, $\exists\mathbf{y}' = \mathbf{w}'/\|\mathbf{w}'\|_2$ such that $\mathbf{y}'^T\mathbf{w}' > \|\mathbf{y}'\|_2$. Therefore, $\partial\|\mathbf{x}\|_2(\mathbf{x} = \mathbf{0}) = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq 1\}$.

3.2

The $\ell_{2,1}$ norm

$$f(X) = \|X\|_{2,1} = \sum_j \|X_{.,j}\|_2 = \sum_j \sqrt{\sum_i X_{ij}^2}.$$

To show that $f(X)$ is a convex function of X, one needs only to show that $f(X)$ is a norm of X, as convexity follows directly from the definition of norms.
a. By construction, it's clear that for any matrix X, $f(X) \geq 0$.
b.

$$f(X) = 0 \Leftrightarrow \sum_j \sqrt{\sum_i X_{ij}^2} = 0 \Leftrightarrow \forall i,\, j,\, X_{ij} = 0 \Leftrightarrow X = 0.$$

c. $\forall \alpha \in \mathbb{R}$,

$$f(\alpha X) = \sum_j \sqrt{\sum_i (\alpha X_{ij})^2} = |\alpha| \sum_j \sqrt{\sum_i X_{ij}^2} = |\alpha| f(X).$$

d. $\forall X, Y$

$$f(X+Y) = \sum_j \|(X+Y)_{\cdot,j}\|_2 \leq \sum_j (\|X_{\cdot,j}\|_2 + \|Y_{\cdot,j}\|_2) = \sum_j \|X_{\cdot,j}\|_2 + \sum_j \|Y_{\cdot,j}\|_2 = f(X) + f(Y).$$

Note that the inequality follows from the triangle inequality for the ℓ_2 norm of vectors.

3.3

a. If $X_{\cdot,j} \neq \mathbf{0}$, $f(X)$ is differentiable and convex, so

$$(\partial \|X\|_{2,1})_{ij} = \frac{\partial f(X)}{\partial X_{ij}} = \frac{X_{ij}}{\|X_{\cdot,j}\|_2}.$$

b. If $X_{\cdot,j} = \mathbf{0}$, and $\forall Y$ such that only the j-th column $Y_{\cdot,j}$ differs from $X_{\cdot,j}$,

$$f(Y) - f(X) = \|Y_{\cdot,j}\|_2.$$

From (3.1), we already know that $\{\mathbf{w} : \forall Y_{\cdot,j}, \|Y_{\cdot,j}\|_2 \geq (\|Y_{\cdot,j}\|_2)^T \mathbf{w}\} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq 1\}$. Thus,

$$(\partial \|X\|_{2,1})_{ij}(X_{\cdot,j} = \mathbf{0}) = \{W_{ij} : \|W_{\cdot,j}\|_2 \leq 1\}.$$

3.4

The optimization problem

$$\min_A \frac{1}{2} \|X - A\|_F^2 + \tau \|A\|_{2,1}.$$

The corresponding Lagrangian function $\mathcal{L}(A) = \frac{1}{2}\|X - A\|_F^2 + \tau \|A\|_{2,1}$ with subgradient:

$$(\partial \mathcal{L})_{\cdot,j} = -(X_{\cdot,j} - A_{\cdot,j}) + \begin{cases} \tau \frac{A_{\cdot,j}}{\|A_{\cdot,j}\|_2},\ A.,j \neq \mathbf{0} \\ \tau W_{\cdot,j} : \|W_{\cdot,j}\|_2 \leq 1,\ A_{\cdot,j} = \mathbf{0}. \end{cases}$$

a. If $\|X_{.,j}\|_2 \geq \tau$, let $A_{.,j} = (1 - \frac{\tau}{\|X_{.,j}\|_2})X_{.,j}$. We can verify that

$$(\partial \mathcal{L})_{.,j} = -(X_{.,j} - (1 - \frac{\tau}{\|X_{.,j}\|_2})X_{.,j}) + \tau \frac{X_{.,j}}{\|X_{.,j}\|_2} = \mathbf{0}.$$

b. If $\|X_{.,j}\|_2 < \tau$, let $A_{.,j} = \mathbf{0}$ and choose $W_{.,j} = X_{.,j}/\tau$ ($\|W_{.,j}\|_2 < 1$), then

$$(\partial \mathcal{L})_{.,j} = -(X_{.,j}) + \tau \frac{X_{.,j}}{\tau} = \mathbf{0}.$$

Since $\| \cdot \|_F^2$ is strictly convex (note the power 2 in the superscript), the optimization problem is also strictly convex. The unique optimal solution is then given by

$$A = X \mathcal{S}_\tau(\mathrm{diag}(\mathbf{x}))\mathrm{diag}(\mathbf{x})^{-1},$$

where $x_j = \|X_{.,j}\|_2$ and the j-th entry of $\mathrm{diag}(\mathbf{x})^{-1}$ is zero if $x_j = 0$.