SI251 Convex Optimization
Homework 3

Instructor: Professor Ye Shi

Due on 11 Dec 23:59 UTC+8

Note:

- Please provide enough calculation process to get full marks.
- Please submit your homework to Gradescope with entry code: **J7DK2D**.
- Please check carefully whether the question number on the gradescope corresponds to each question.

Exercise 1. (Proximal Operator)

Question 1 (5 pts)

Let $f(X) = \lambda\|X\|_*$, where $X \in \mathbb{R}^{d\times m}$ is a matrix, $\|X\|_*$ denotes the nuclear norm, and $\lambda \in \mathbb{R}_+$ is the regularization parameter. Please derive proximal operator of f, i.e., $\text{prox}_f(\mathbf{X})$.

Solution:

$$\text{prox}_{\lambda||\cdot||_*}(X) = \arg\ \min_Z\{\lambda||Z||_* + \frac{1}{2}||Z - X||_F^2\}$$

We can perform SVD decomposition on matrices Z and X respectively

$$Z = U_Z\Sigma_Z V_Z^T, \quad X = U\Sigma V$$

The nuclear norm $||Z||_*$ is the sum of the singular values of Z:

$$||Z||_* = \sum_i \sigma_i(Z)$$

$$||\mathbf{A}||_F^2 = (\text{tr}(\mathbf{A}^T\mathbf{A})) = (\sum_i [\mathbf{A}^T\mathbf{A})]_{ii}) = (\sum_i(\sum_j A_{ij}^T A_{ji})) = (\sum_{i,j} A_{ij}^2)$$

We can prove that the square of the Frobenius norm is equal to the sum of the squares of its elements, so we can get

$$||Z - X||_F^2 = \sum_{i,j}(Z_{ij} - X_{ij})^2$$

Since U, V are orthogonal matrix and we have $||UAV||_F = ||A||_F$. Meanwhile, the Frobenius norm is monotonic to singular values, so we can transform the problem into an optimization problem for singular values. we can get:

$$||Z - X||_F^2 = ||U_Z\Sigma_Z V_Z^T - U\Sigma V||_F^2 \Rightarrow ||\Sigma_Z - \Sigma||_F^2 = \sum_i(\sigma_i(Z) - \sigma_i)^2$$

So the objective function becomes:

$$\min_{\sigma_i(Z)}\{\lambda\sum_i \sigma_i(Z) + \frac{1}{2}\sum_i(\sigma_i(Z) - \sigma_i)^2\}$$

This objective function is independent for each $\sigma_i(Z)$. The objective function becomes:

$$\min_{\sigma_i(Z)}\{\lambda\sigma_i(Z) + \frac{1}{2}(\sigma_i(Z) - \sigma_i)^2\}$$

Take the derivative of $\sigma_i(Z)$ and set the derivative to 0

$$\frac{d}{d\sigma_i(Z)}(\lambda\sigma_i(Z) + \frac{1}{2}(\sigma_i(Z) - \sigma_i)^2) = \lambda + (\sigma_i(Z) - \sigma_i) = 0 \Rightarrow \sigma_i(Z) = \sigma_i - \lambda$$

Since the eigenvalues obtained by SVD decomposition of the matrix are all positive numbers, there is an implicit condition $\sigma_i(Z) \geq 0, \sigma_i \geq 0$ here. So we get:

$$\sigma_i(Z) = \max(\sigma_i - \lambda, 0)$$

So we can get:

$$\text{prox}_{\lambda||\cdot||_*}(X) = U\text{diag}(\max(\sigma_i - \lambda, 0))V^T, \text{ where } X = U\Sigma V^T$$

Question 2 (5 pts)

If $f(x) = g(ax + b)$ with $a \neq 0$, please prove that

$$\text{prox}_f(x) = \frac{1}{a}(\text{prox}_{a^2 g}(ax + b) - b) \tag{1}$$

Solution:

$$\text{prox}_f(x) = \arg\min_z \{\frac{1}{2}||z - x||_2^2 + g(ax + b)\} \tag{2}$$

By change-of-variables, Assume $u = az + b$, $\arg\min_z = \frac{\arg\min_u - b}{a}$

$$\begin{aligned} \text{prox}_f(x) &= \frac{1}{a}(\arg\min_u \{\frac{1}{2}||\frac{u - b}{a} - x||_2^2 + g(u)\} - b) \\ &= \frac{1}{a}(\arg\min_u \{\frac{1}{2a^2}||u - (ax + b)||_2^2 + g(u)\} - b) \\ &= \frac{1}{a}(\arg\min_u \{\frac{1}{2}||u - (ax + b)||_2^2 + a^2 g(u)\} - b) \\ &= \frac{1}{a}(\text{prox}_{a^2 g}(ax + b) - b) \end{aligned} \tag{3}$$

Exercise 2. (Conjugate Function)(5 pts)

Given arbitrary function $f(x)$, there's an obvious convexity for its conjugate function,

$$f^*(y) = \sup_{x \in \mathrm{dom}\, f} \left(y^T x - f(x) \right),$$

where $f^*(y)$ represents the combination of series of point by point upper bounds. Please provide a detailed proof to the convexity of conjugate functions.

(Hint: I suggest that you follow the instructions below to complete this proof in order to gain a full understanding of the geometric meaning of conjugate functions and dual problems.

1. Transform a closed convex function to its epigraph.
2. Transform the epigraph to a hyperplane of which normal vector is $(y, -1)$.
3. Try to clip subsets of the hyperplanes, then associate each hyperplane with some crossing points)

Solution:
Given a set $A \subseteq \mathbb{R}^n$, its support function is as follows,

$$h_A(x) \equiv \sup\{x \cdot a : a \in A\},$$

where describes a set of its support hyperplanes. For every vector $x \in \mathbb{R}^n$, we project the set A onto x in order to fetch the largest value of the projection norm. Thus we get hyperplanes with each x representing the plane's normal vector,

$$\{y \in \mathbb{R}^n : x \cdot y = h_A(x)\}.$$

If A is convex, there is a bijection between A and h_A.

With the above prerequisite, now consider a function $f : \mathbb{R}^n \to \tilde{\mathbb{R}}$ and its epigraph:

$$\mathrm{epi}(f) \equiv \left\{(x, \alpha) \in \mathbb{R}^{n+1} : f(x) \leq \alpha\right\} \subseteq \mathbb{R}^{n+1}.$$

A function is convex if its epigraph is convex. The supporting function of the epigraph is as follows:

$$h_{\mathrm{epi}(f)}\left(y, y^*\right) = \sup_{\mathbf{x} \in \mathbb{R}^n, \alpha \geq f(\mathbf{x})} \left(y^T \cdot \mathbf{x} + y^* \alpha\right),$$

where the normal vector for each hyperplane is (y, y^*).

We don't need too many normal directions in our proof, it's sufficient to only look at the normal vectors in the form of $(y, -1)$. Then we can get the subset of the above hyperplane set, which is:

$$h_{\mathrm{epi}(f)}\left(y, -1\right) \equiv \sup_{\mathbf{x} \in \mathbb{R}^n, \alpha \geq f(\mathbf{x})} \left(y^T \cdot \mathbf{x} - \alpha\right).$$

From the above defination, we reformulate the inequality constraint $f(x) \leq \alpha$ to,

$$-\alpha \leq -f(x),$$

where we can make substitution between maximising $-\alpha$ and $-f(x)$, thus we derive that:

$$h_{\mathrm{epi}(f)}\left(y, -1\right) \equiv \sup_{\mathbf{x} \in \mathbb{R}^n} \left(y^T \cdot \mathbf{x} - f(\mathbf{x})\right),$$

which is clearly the definition of the conjugate function.

Finally, we can conclude that the conjugate function of f is the set of hpyerplanes to the epigraph of f, and it is clearly convex.

Exercise 3. (ADMM)(5 pts)

Consider the following robust PCA problem, which try to decompose matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ into low rank matrix $\mathbf{L} \in \mathbb{R}^{n\times m}$ and sparse matrix $\mathbf{S} \in \mathbb{R}^{n\times m}$:

$$\min_{\mathbf{L},\mathbf{S}} \quad ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1 \tag{4}$$

$$\text{s.t.} \quad \mathbf{L} + \mathbf{S} = \mathbf{M} \tag{5}$$

where $||\mathbf{L}||_* := \sum_{i=1}^{n} \sigma_i(\mathbf{L})$ is the nuclear norm, and $||\mathbf{S}||_1 := \sum_{i,j} |S_{ij}|$ is the entrywise ℓ_1 norm. Please prove that when the dual variable is $\mathbf{\Lambda}$, the ADMM update of robust PCA problem is

$$\mathbf{L}^{(t+1)} = \mathrm{SVT}_{\rho^{-1}}\left(\mathbf{M} - \mathbf{S}^{(t)} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\right) \tag{6}$$

$$\mathbf{S}^{(t+1)} = \mathrm{ST}_{\lambda\rho^{-1}}\left(\mathbf{M} - \mathbf{L}^{(t+1)} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\right) \tag{7}$$

$$\mathbf{\Lambda}^{(t+1)} = \mathbf{\Lambda}^{(t)} + \rho\left(\mathbf{L}^{(t+1)} + \mathbf{S}^{(t+1)} - \mathbf{M}\right) \tag{8}$$

where for any $\mathbf{X}$ with SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{\Sigma} = \mathrm{diag}(\{\sigma_i\}))$, one has

$$\mathrm{SVT}_\tau(\mathbf{X}) = \mathbf{U}\mathrm{diag}(\{(\sigma_i - \tau)_+\})\mathbf{V}^\top \tag{9}$$

and

$$(\mathrm{ST}_\tau(\mathbf{X}))_{i,j} = \begin{cases} X_{i,j} - \tau, & \text{if } X_{i,j} > \tau, \\ 0, & \text{if } |X_{i,j}| \le \tau, \\ X_{i,j} + \tau, & \text{if } X_{i,j} < -\tau. \end{cases} \tag{10}$$

Hint: Please provide enough proof details, or you will lose points.

Solution: The primal variables $\mathbf{L}$ and $\mathbf{S}$ are augmented with a dual variable $\mathbf{\Lambda}$ to form the following augmented Lagrangian:

$$\mathcal{L}_\rho(\mathbf{L}, \mathbf{S}, \mathbf{\Lambda}) = \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \langle \mathbf{\Lambda}, \mathbf{L} + \mathbf{S} - \mathbf{M}\rangle + \frac{\rho}{2}\|\mathbf{L} + \mathbf{S} - \mathbf{M} - \frac{1}{\rho}\mathbf{\Lambda}\|_F^2,$$

The update for $\mathbf{L}$ involves minimizing the augmented Lagrangian with respect to $\mathbf{L}$ while keeping $\mathbf{S}$ and $\mathbf{\Lambda}$ fixed:

$$\mathbf{L}^{(t+1)} = \arg\min_{\mathbf{L}} \|\mathbf{L}\|_* + \langle \mathbf{\Lambda}^{(t)}, \mathbf{L} + \mathbf{S}^{(t)} - \mathbf{M}\rangle + \frac{\rho}{2}\|\mathbf{L} + \mathbf{S}^{(t)} - \mathbf{M} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\|_F^2.$$

By differentiating and setting the derivative to zero, we can identify the update rule as a **soft singular value thresholding** (SVT) operation. The result is:

$$\mathbf{L}^{(t+1)} = \mathrm{SVT}_{\frac{1}{\rho}}\left(\mathbf{M} - \mathbf{S}^{(t)} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\right),$$

where the soft singular value thresholding operator is defined as:

$$\mathrm{SVT}_\tau(\mathbf{X}) = \mathbf{U}\mathrm{diag}(\{(\sigma_i - \tau)_+\})\mathbf{V}^\top,$$

with $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ being the singular value decomposition (SVD) of $\mathbf{X}$, and $(\sigma_i - \tau)_+ = \max(\sigma_i - \tau, 0)$.

Step 2: Update for $\mathbf{S}$

Next, we update $\mathbf{S}$ by minimizing the augmented Lagrangian with respect to $\mathbf{S}$:

$$\mathbf{S}^{(t+1)} = \arg\min_{\mathbf{S}} \mathcal{L}_\rho(\mathbf{L}^{(t+1)}, \mathbf{S}, \mathbf{\Lambda}^{(t)}).$$

The term to minimize is:

$$\lambda\|\mathbf{S}\|_1 + \langle \mathbf{\Lambda}^{(t)}, \mathbf{L}^{(t+1)} + \mathbf{S} - \mathbf{M} \rangle + \frac{\rho}{2}\|\mathbf{L}^{(t+1)} + \mathbf{S} - \mathbf{M} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\|_F^2.$$

The minimization with respect to $\mathbf{S}$ leads to a **soft thresholding** operation:

$$\mathbf{S}^{(t+1)} = \mathrm{ST}_{\frac{\lambda}{\rho}}\left(\mathbf{M} - \mathbf{L}^{(t+1)} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\right),$$

where the soft thresholding operator is defined elementwise as:

$$(\mathrm{ST}_\tau(\mathbf{X}))_{i,j} = \begin{cases} X_{i,j} - \tau, & \text{if } X_{i,j} > \tau, \\ 0, & \text{if } |X_{i,j}| \le \tau, \\ X_{i,j} + \tau, & \text{if } X_{i,j} < -\tau. \end{cases}$$

Step 3: Update for $\mathbf{\Lambda}$
Finally, we update the dual variable $\mathbf{\Lambda}$ by gradient ascent on the augmented Lagrangian:

$$\mathbf{\Lambda}^{(t+1)} = \mathbf{\Lambda}^{(t)} + \rho\left(\mathbf{L}^{(t+1)} + \mathbf{S}^{(t+1)} - \mathbf{M}\right).$$

This update follows directly from the structure of the augmented Lagrangian.
Conclusion
Thus, the ADMM updates for the robust PCA problem are:

$$\mathbf{L}^{(t+1)} = \mathrm{SVT}_{\frac{1}{\rho}}\left(\mathbf{M} - \mathbf{S}^{(t)} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\right),$$

$$\mathbf{S}^{(t+1)} = \mathrm{ST}_{\frac{\lambda}{\rho}}\left(\mathbf{M} - \mathbf{L}^{(t+1)} - \frac{1}{\rho}\mathbf{\Lambda}^{(t)}\right),$$

$$\mathbf{\Lambda}^{(t+1)} = \mathbf{\Lambda}^{(t)} + \rho\left(\mathbf{L}^{(t+1)} + \mathbf{S}^{(t+1)} - \mathbf{M}\right).$$