Convex Optimization, Spring 2021
Homework 5

Due on June 9, 2021

Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- You are required to write down all the major steps towards making your conclusions; otherwise you may obtain limited points (20%) of the problem.
- Write your homework in English; otherwise you will get no points of this homework.
- Do your homework by yourself. Any form of plagiarism will lead to 0 point of this homework. If more than one plagiarisms during the semester are identified, we will prosecute all violations to the fullest extent of the university regulations, including but not limited to failing this course, academic probation, or expulsion from the university.
- If you have any doubts regarding the grading, you need to contact the instructor or the TAs within two days since the grade is announced.

I. Proximal Gradient Methods

A useful extension of Moreau decomposition can be stated as

Fact 1 (refer to Lecture 11)**.** *Suppose f is closed convex and $\lambda > 0$. Then*

$$\boldsymbol{x} = \operatorname{prox}_{\lambda f}(\boldsymbol{x}) + \lambda \operatorname{prox}_{\frac{1}{\lambda} f^*}(\boldsymbol{x}/\lambda). \tag{1}$$

Please verify this fact (the derivation process is required). (30 points)

Solution:

Proof. Using Moreau decomposition, for any $\boldsymbol{x}$,

$$\operatorname{prox}_{\lambda f}(\boldsymbol{x}) = \boldsymbol{x} - \operatorname{prox}_{(\lambda f)^*}(\boldsymbol{x}) = \boldsymbol{x} - \operatorname{prox}_{\lambda f^*(\cdot/\lambda)}(\boldsymbol{x}), \tag{2}$$

where the second equality follows by Theorem 4.14(a) in [1]. By Theorem 6.12 in [1],

$$\operatorname{prox}_{\lambda f^*(\cdot/\lambda)}(\boldsymbol{x}) = \lambda \operatorname{prox}_{\lambda^{-1} f^*}(\boldsymbol{x}/\lambda), \tag{3}$$

which, combined with (2), yields (1). □

II. Accelerated Gradient Methods

The momentum coefficient of the fast iterative shrinkage-thresholding algorithm (FISTA) (refer to lecture 11) at t-th iterate is given as

$$\theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2} \quad \text{with} \quad \theta_0 = 1.$$

(1) Show that

$$\theta_t \geq \frac{t+2}{2} \tag{4}$$

holds true for all $t \geq 1$. (20 points)

(2) Show that

$$\frac{\theta_t - 1}{\theta_{t+1}} = 1 - \frac{3}{t} + o(\frac{1}{t}). \tag{5}$$

(20 points)

Solution:

(1) *Proof.* The proof is by induction on t.

- Base case: Obviously, for $t = 0, \theta_0 = 1 \geq \frac{0+2}{2}$.
- Inductive step: Suppose that the claim holds for t, meaning that $\theta_t \geq \frac{t+2}{t}$. We now show that $\theta_{t+1} \geq \frac{t+3}{2}$. By the recursive relation defining the sequence and the induction assumption,

$$\theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2} \geq \frac{1 + \sqrt{1 + (k+2)^2}}{2} \geq \frac{1 + \sqrt{(k+2)^2}}{2} = \frac{t+3}{2} \tag{6}$$

This completes the proof. □

(2) *Proof.* Since $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2} = \frac{2\theta_t^2}{\sqrt{1+4\theta_t^2}-1}$, we have

$$\frac{\theta_t - 1}{\theta_{t+1}} = \frac{(\theta_t - 1)\left(\sqrt{1 + 4\theta_t^2} - 1\right)}{2\theta_t^2} \tag{7}$$

Due to the fact that $\theta_t \geq \frac{t+2}{2}, \forall t \geq 1$ (we require you to prove in the next part), we have $\frac{1}{\theta_t} \leq \frac{2}{t+2}$. Thus

$$\frac{(\theta_t - 1)\left(\sqrt{1 + 4\theta_t^2} - 1\right)}{2\theta_t^2} = \frac{(\theta_t - 1)(2\theta_t - 1)}{2\theta_t^2} + o\left(\frac{1}{t}\right) \tag{8}$$

$$= 1 - \frac{3}{2\theta_t} + o\left(\frac{1}{t}\right) \tag{9}$$

$$= 1 - \frac{3}{t} + o\left(\frac{1}{t}\right). \tag{10}$$

This completes the proof. □

III. Alternating Direction Method of Multipliers

Suppose we observe $\boldsymbol{M}$ (refer to lecture 14), which is superposition of low-rank component $\boldsymbol{L}$ and sparse outliers $\boldsymbol{S}$, provide the ADMM update (the derivation process is required) for each variable at the t-th iteration

$$\begin{aligned} \underset{\boldsymbol{L},\boldsymbol{S}}{\text{minimize}} \quad & \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \\ \text{subject to} \quad & \boldsymbol{L}+\boldsymbol{S}=\boldsymbol{M}, \end{aligned} \tag{11}$$

where $\|\boldsymbol{L}\|_* := \sum_{i=1}^{n}\sigma_i(\boldsymbol{L})$ is nuclear norm, and $\|\boldsymbol{S}\|_1 = \sum_{i,j}|S_{i,j}|$ is entrywise ℓ_1-norm. (30 points)

Solution:

By introducing dual variable $\boldsymbol{\Lambda}$, ADMM update rules for solving (11) are given by

$$\boldsymbol{L}^{t+1} = \arg\min_{\boldsymbol{L}} \left\{ \|\boldsymbol{L}\|_* + \frac{\rho}{2}\left\|\boldsymbol{L}+\boldsymbol{S}^t-\boldsymbol{M}+\frac{1}{\rho}\boldsymbol{\Lambda}^t\right\|_{\mathrm{F}}^2 \right\} \tag{12}$$

$$\boldsymbol{S}^{t+1} = \arg\min_{\boldsymbol{S}} \left\{ \lambda\|\boldsymbol{S}\|_1 + \frac{\rho}{2}\left\|\boldsymbol{L}^{t+1}+\boldsymbol{S}-\boldsymbol{M}+\frac{1}{\rho}\boldsymbol{\Lambda}^t\right\|_{\mathrm{F}}^2 \right\} \tag{13}$$

$$\boldsymbol{\Lambda}^{t+1} = \boldsymbol{\Lambda}^t + \rho\left(\boldsymbol{L}^{t+1}+\boldsymbol{S}^{t+1}-\boldsymbol{M}\right) \tag{14}$$

Define the function $h_t(\boldsymbol{L}) = \|\boldsymbol{L}\|_* + \frac{\rho}{2}\left\|\boldsymbol{L}+\boldsymbol{S}^t-\boldsymbol{M}+\frac{1}{\rho}\boldsymbol{\Lambda}^t\right\|_{\mathrm{F}}^2$. Since $h_t(\boldsymbol{L})$ is strictly convex, it is easy to see that there exists a unique minimizer. Now $\hat{\boldsymbol{L}}$ minimizes h_t if and only if $\boldsymbol{0}$ is a subgradient of the functional h_t at the point $\hat{\boldsymbol{L}}$, i.e.,

$$\boldsymbol{0} \in \hat{\boldsymbol{L}} - \left(\boldsymbol{M}-\boldsymbol{S}^t-\frac{1}{\rho}\boldsymbol{\Lambda}^t\right) + \frac{1}{\rho}\partial\|\hat{\boldsymbol{L}}\|_*, \tag{15}$$

where $\partial\|\hat{\boldsymbol{L}}\|_*$ is the set of subgradients of the nuclear norm. Let $\boldsymbol{L}$ be an arbitrary matrix and $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be its SVD. It is known [2] that

$$\partial\|\boldsymbol{L}\|_* = \left\{\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W} : \boldsymbol{U}^*\boldsymbol{W}=0, \quad \boldsymbol{W}\boldsymbol{V}=0, \quad \|\boldsymbol{W}\|_2 \le 1\right\}. \tag{16}$$

Set $\hat{\boldsymbol{L}} = \mathrm{SVT}_{\rho^{-1}}\left(\boldsymbol{M}-\boldsymbol{S}^t-\frac{1}{\rho}\boldsymbol{\Lambda}^t\right)$. Now we need to show that $\hat{\boldsymbol{L}}$ obeys (15). Decompose the SVD of $\boldsymbol{M}-\boldsymbol{S}^t-\frac{1}{\rho}\boldsymbol{\Lambda}^t$ as $\boldsymbol{M}-\boldsymbol{S}^t-\frac{1}{\rho}\boldsymbol{\Lambda}^t = \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^* + \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$, where $\boldsymbol{U}_0, \boldsymbol{V}_0$ (resp. $\boldsymbol{U}_1, \boldsymbol{V}_1$ are the singular vectors associated with singular values greater than $1/\rho$ (resp. smaller than or equal to $1/\rho$). With these notations, we have $\hat{\boldsymbol{L}} = \boldsymbol{U}_0\left(\boldsymbol{\Sigma}_0 - 1/\rho\boldsymbol{I}\right)\boldsymbol{V}_0^*$, and therefore,

$$\boldsymbol{M}-\boldsymbol{S}^t-\frac{1}{\rho}\boldsymbol{\Lambda}^t-\hat{\boldsymbol{L}} = 1/\rho\left(\boldsymbol{U}_0\boldsymbol{V}_0^* + \boldsymbol{W}\right), \quad \boldsymbol{W} = \rho\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^* \tag{17}$$

By definition, $\boldsymbol{U}_0^*\boldsymbol{W}=0, \boldsymbol{W}\boldsymbol{V}_0=0$ and since the diagonal elements of $\boldsymbol{\Sigma}_1$ have magnitudes bounded by $1/\rho$, we also have $\|\boldsymbol{W}\|_2 \le 1$. Hence, $\boldsymbol{M}-\boldsymbol{S}^t-\frac{1}{\rho}\boldsymbol{\Lambda}^t-\hat{\boldsymbol{L}} \in 1/\rho\partial\|\hat{\boldsymbol{L}}\|_*$, which implies

$$\boldsymbol{L}^{t+1} = \mathrm{SVT}_{\rho^{-1}}\left(\boldsymbol{M}-\boldsymbol{S}^t-\frac{1}{\rho}\boldsymbol{\Lambda}^t\right). \tag{18}$$

Define the function $g_t(\boldsymbol{S}) = \lambda\|\boldsymbol{S}\|_1 + \frac{\rho}{2}\left\|\boldsymbol{L}^{t+1}+\boldsymbol{S}-\boldsymbol{M}+\frac{1}{\rho}\boldsymbol{\Lambda}^t\right\|_{\mathrm{F}}^2$. Similarly, $\hat{\boldsymbol{S}}$ should satisfy

$$\boldsymbol{0} \in \hat{\boldsymbol{S}} - \left(\boldsymbol{M}-\boldsymbol{L}^{t+1}-\frac{1}{\rho}\boldsymbol{\Lambda}^t\right) + \frac{\lambda}{\rho}\partial\|\hat{\boldsymbol{S}}\|_1, \tag{19}$$

where $\partial\|\hat{\boldsymbol{S}}\|_1$ is the set of subgradients of the ℓ_1-norm. Since $\|\boldsymbol{S}\|_1 = \sum_{i,j}|S_{i,j}|$, we have

$$\sum_{i,j}\mathrm{sign}(S_{i,j})\boldsymbol{E}_{i,j} \in \partial\|\hat{\boldsymbol{S}}\|_1, \tag{20}$$

where $\boldsymbol{E}_{i,j}$ is a matrix with only $E_{i,j}=1$. That combines with (19) implies

$$\boldsymbol{S}^{t+1} = \hat{\boldsymbol{S}} = \mathrm{ST}_{\lambda\rho^{-1}}\left(\boldsymbol{M}-\boldsymbol{L}^{t+1}-\frac{1}{\rho}\boldsymbol{\Lambda}^t\right). \tag{21}$$

Hence, the exact the ADMM rules are summarized as

$$\boldsymbol{L}^{t+1} = \mathrm{SVT}_{\rho^{-1}}\left(\boldsymbol{M} - \boldsymbol{S}^t - \frac{1}{\rho}\boldsymbol{\Lambda}^t\right) \tag{22}$$

$$\boldsymbol{S}^{t+1} = \mathrm{ST}_{\lambda\rho^{-1}}\left(\boldsymbol{M} - \boldsymbol{L}^{t+1} - \frac{1}{\rho}\boldsymbol{\Lambda}^t\right) \tag{23}$$

$$\boldsymbol{\Lambda}^{t+1} = \boldsymbol{\Lambda}^t + \rho\left(\boldsymbol{L}^{t+1} + \boldsymbol{S}^{t+1} - \boldsymbol{M}\right). \tag{24}$$

References

[1] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017.

[2] A. S. Lewis, "The mathematics of eigenvalue optimization," *Math. Program.*, vol. 97, no. 1-2, pp. 155–176, 2003.