SI251 - Convex Optimization, Fall 2021
Homework 3

Due on Dec 19, 2021, 23:59 UTC+8

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 ($\leq 20\%$) of the problem.
- Write your homework in English; otherwise you will get no points of this homework.
- If you want to submit a handwritten version, scan it clearly. Camscanner is recommended.
- 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 . Mirror Descent Methods

(1) Let φ be proper convex and differentiable. Suppose $\mathbf{y} = \nabla\varphi(\mathbf{x})$, from conjugate subgradient theorem, show that

$$\varphi(\mathbf{x}) + \varphi^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle,$$

where $\varphi^*(\mathbf{y})$ is Fenchel conjugate of φ. (15 points)

Hint: The conjugate subgradient theorem states that if f is closed proper convex, then the following statements are equivalent for a pair of vectors $(\mathbf{x}, \mathbf{y})$: (i) $\langle \mathbf{x}, \mathbf{y} \rangle = f(\mathbf{x}) + f^*(\mathbf{y})$; (ii) $\mathbf{y} \in \partial f(\mathbf{x})$; (iii) $\mathbf{x} \in \partial f^*(\mathbf{y})$, where $f^*(\mathbf{y})$ is Fenchel conjugate of f.

(2) Using Bregman divergence, show that mirror descent has an alternative form, which reads

$$\mathbf{x}^{t+1} = \nabla\varphi^*(\nabla\varphi(\mathbf{x}^t) - \eta_t \mathbf{g}^t),$$

where $\varphi^*(\mathbf{x})$ is Fenchel-conjugat of φ, $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$ and $\eta_t > 0$ is the stepsize. (For simplicity, assume the constraints set $\mathcal{C} = \mathbb{R}^n$.) (15 points)

Solutions:

(1) Since φ is convex and $\mathbf{y} = \nabla\varphi(\mathbf{x})$, we have

$$\begin{gathered}
\varphi(\mathbf{m}) \geq \varphi(\mathbf{x}) + \langle \mathbf{y}, \mathbf{m} - \mathbf{x} \rangle \\
\Updownarrow \\
\langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \geq \langle \mathbf{y}, \mathbf{m} \rangle - \varphi(\mathbf{m}) \\
\Updownarrow \\
\langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \geq \sup_{\mathbf{m}} \langle \mathbf{y}, \mathbf{m} \rangle - \varphi(\mathbf{m}) \\
\Updownarrow \\
\langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \geq \varphi^*(\mathbf{y})
\end{gathered}$$

On the other hand

$$\begin{aligned}
\varphi^*(\mathbf{y}) &= \sup_{\mathbf{x}} \langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \\
&\geq \langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x})
\end{aligned}$$

Therefore, arranging the term, we have $\varphi(\mathbf{x}) + \varphi^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$. This completes the proof.

(2)

$$\begin{aligned}
\mathbf{x}^{t+1} &= \underset{\mathbf{x}\in\mathcal{C}}{\arg\min} \left\{ \langle \mathbf{g}^t, \mathbf{x} - \mathbf{x}^t \rangle + \frac{1}{\eta_t} \left(\varphi(\mathbf{x}) - \varphi\left(\mathbf{x}^t\right) - \langle \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \rangle \right) \right\} \\
&= \underset{\mathbf{x}\in\mathcal{C}}{\arg\min} \left\{ \left\langle \mathbf{g}^t - \frac{1}{\eta_t} \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \right\rangle + \frac{1}{\eta_t} \left(\varphi(\mathbf{x}) - \varphi\left(\mathbf{x}^t\right) \right) \right\} \\
&= \underset{\mathbf{x}\in\mathcal{C}}{\arg\min} \left\{ \left\langle \mathbf{g}^t - \frac{1}{\eta_t} \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \right\rangle + \frac{1}{\eta_t} \varphi(\mathbf{x}) \right\} \\
&= \underset{\mathbf{x}\in\mathcal{C}}{\arg\min} \left\{ \langle \eta_t \mathbf{g}^t - \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \rangle + \varphi(\mathbf{x}) \right\} \\
&= \underset{\mathbf{x}\in\mathcal{C}}{\operatorname{argmax}} \left\{ \nabla\varphi\left(\mathbf{x}^t\right) - \eta_t \mathbf{g}^t, \mathbf{x} - \mathbf{x}^t \rangle - \varphi(\mathbf{x}) \right\}
\end{aligned}$$

Therefore, we have $x^{t+1} = \nabla\varphi^*\left(\nabla\varphi\left(x^t\right) - \eta_t g^t\right)$

II . **Proximal Algorithm**

For each of the following convex functions, compute the proximal operator prox_f.

(1) $f(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, where $\mathbf{x} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}_+$ is the regularization parameter. (15 points)

(2) $f(\mathbf{X}) = \lambda\|\mathbf{X}\|_*$, where $\mathbf{X} \in \mathbb{R}^{d\times m}$ is a matrix and $\lambda \in \mathbb{R}_+$ is the regularization parameter. (25 points)

Solutions:

(1) By definition of the prox,

$$\mathrm{prox}_f(\mathbf{x}) = \underset{\mathbf{u}\in\mathbb{R}^d}{\arg\min} \left\{ \lambda\|\mathbf{u}\|_1 + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 \right\}$$

This formulation takes the form as Lasso. Then we see the minimizer $\mathbf{u}^*$ is a soft-thresholding operator of x at λ

$$\left[\mathrm{prox}_f(\mathbf{x})\right]_i = [\mathbf{u}^*]_i = \mathcal{S}_\lambda(\mathbf{u}) = \begin{cases} \mathbf{u}_i - \lambda & \text{if} \quad \mathbf{u}_i > \lambda \\ 0 & \text{if} \quad |\mathbf{u}_i| \le \lambda \\ \mathbf{u}_i + \lambda & \text{if} \quad \mathbf{u}_i < -\lambda \end{cases}$$

(2) Let $\mathbf{X} = \mathbf{U\Sigma V}^T$ be the SVD. Since the Frobenius norm is rotation-invariant, we write

$$\begin{aligned} \mathrm{prox}_f(\mathbf{X}) &= \underset{\mathbf{Y}\in\mathbb{R}^{d\times m}}{\arg\min} \left\{ f(\mathbf{Y}) + \frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 \right\} \\ &= \underset{\mathbf{Y}\in\mathbb{R}^{d\times m}}{\arg\min} \left\{ f(\mathbf{Y}) + \frac{1}{2}\left\|\mathbf{\Sigma}-\mathbf{U}^T\mathbf{Y}\mathbf{V}\right\|_F^2 \right\} \\ &= \mathbf{U}\left(\underset{\tilde{\mathbf{Y}}\in\mathbb{R}^{d\times m}}{\arg\min} \left\{ f(\tilde{\mathbf{Y}}) + \frac{1}{2}\|\mathbf{\Sigma}-\tilde{\mathbf{Y}}\|_F^2 \right\} \right)\mathbf{V}^T \\ &= \mathbf{U}\left(\mathrm{prox}_f(\mathbf{\Sigma})\right)\mathbf{V}^T \end{aligned} \tag{1}$$

where we used rotational invariance of the Frobenius norm in the second equality of (1) and reparameterized the problem using $\tilde{\mathbf{Y}} = \mathbf{U}^T\mathbf{Y}\mathbf{V}$ in third equality of (1) since $f(\tilde{\mathbf{Y}}) = f(\mathbf{Y})$ We still have to show how to compute prox $_f(\mathbf{\Sigma})$ for a non-negative diagonal matrix $\mathbf{\Sigma}$ of singular values. Clearly the minimizer $\mathbf{Y}$ is diagonal, since every non-zero off-diagonal term in $\mathbf{Y}$ gets a positive penalty from the Frobenius norm. This yields

$$\begin{aligned} \mathrm{prox}_f(\mathbf{D}) &= \underset{\mathbf{Y}\in\mathbb{R}^{d\times m}}{\arg\min} \left\{ f(\mathbf{Y}) + \frac{1}{2}\|\mathbf{D}-\mathbf{Y}\|_F^2 \right\} \\ &= \underset{\mathbf{Y}\in\mathbb{R}^{d\times m}}{\arg\min} \left\{ \lambda \sum_{j=1}^{\min(d,m)} \sigma_j(\mathbf{Y}) + \frac{1}{2}\left(\sigma_j(\mathbf{D}) - \sigma_j(\mathbf{Y})\right)^2 \right\} \\ &= \underset{\mathbf{y}\in\mathbb{R}^{\min(d,m)}}{\arg\min} \ \lambda\|\mathbf{y}\|_1 + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \end{aligned} \tag{2}$$

where $\mathbf{y}, \mathbf{x}$ are the vectors of singular values of $\mathbf{Y}$ and $\mathbf{X}$ respectively and the last equality of (2) follows by $\sigma_j(\mathbf{X}) = \sigma_j(\mathbf{D})$ for all j. After the problem reduced to a ℓ_1 regularized problem in the vector space, we see that the minimizer $\mathbf{y}^*$ is a soft-thresholding version of $\mathbf{x}$ at λ and thus we have

$$\left(\mathrm{prox}_f(\mathbf{D})\right)_{ij} = \begin{cases} 0 & i \neq j \\ \max\{\sigma_i(\mathbf{X}) - \lambda, 0\} & i = j \end{cases}$$

Thus the whole procedure for computing the proximal operator involves taking the SVD and soft-thresholding the singular values.

III . Alternating Direction Method of Multipliers

Consider the least squares regression problem with ℓ_2-norm regularization (ridge penalty),

$$\underset{\mathbf{z}\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{A}_i\mathbf{z}-\mathbf{b}_i\|_2^2+(\lambda/2)\|\mathbf{z}\|_2^2 \tag{3}$$

where $\mathbf{A}_i \in \mathbb{R}^{n_i\times d}$ is the data matrix, $\mathbf{b}_i \in \mathbb{R}^{n_i}$ is the measurement vector, λ is the regularization parameter.

Please write the **exact** ADMM update steps for this problem. (30 points)

Solution:

We first rewrite the problem (3) with local variables $\{\mathbf{x}_i\}$ and a common global variable $\mathbf{z}$ as

$$\begin{aligned}\underset{\mathbf{x}_i,\mathbf{z}\in\mathbb{R}^d}{\text{minimize}} \quad & \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{A}_i\mathbf{x}_i-\mathbf{b}_i\|_2^2+(\lambda/2)\|\mathbf{z}\|_2^2\\ \text{subject to} \quad & \mathbf{x}_i-\mathbf{z}=0,\end{aligned}$$

which can be further rewritten as

$$\begin{aligned}\underset{\mathbf{x}_i,\mathbf{z}\in\mathbb{R}^d}{\text{minimize}} \quad & \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{A}_i\mathbf{x}_i-\mathbf{b}_i\|_2^2+(\lambda/2)\|\mathbf{z}\|_2^2\\ \text{subject to} \quad & \begin{bmatrix}\mathbf{x}_1\\ \vdots\\ \mathbf{x}_N\end{bmatrix}-\begin{bmatrix}\mathbf{z}\\ \vdots\\ \mathbf{z}\end{bmatrix}=\mathbf{0}.\end{aligned}$$

By introducing Lagrange multipliers $\{\mathbf{y}_i\}$, the augmented Lagrangian $L_\rho(\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z},\mathbf{y}_1,\ldots,\mathbf{y}_N)$ is given by

$$L_\rho(\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z},\mathbf{y}_1,\ldots,\mathbf{y}_N)=\frac{1}{2}\sum_{i=1}^{N}\|\mathbf{A}_i\mathbf{x}_i-\mathbf{b}_i\|_2^2+g(\mathbf{z})+\sum_{i=1}^{N}\left(\mathbf{y}_i^\top(\mathbf{x}_i-\mathbf{z})+(\rho/2)\|\mathbf{x}_i-\mathbf{z}\|_2^2\right).$$

The ADMM update rules can be given by

$$\begin{aligned}\mathbf{x}_i^{k+1} &:= \underset{\mathbf{x}_i}{\text{argmin}}\left(\|\mathbf{A}_i\mathbf{x}_i-\mathbf{b}_i\|_2^2+(\mathbf{y}_i^k)^\top(\mathbf{x}_i-\mathbf{z}^k)+(\rho/2)\left\|\mathbf{x}_i-\mathbf{z}^k\right\|_2^2\right)\\ \mathbf{z}^{k+1} &:= \underset{\mathbf{z}}{\text{argmin}}\left((\lambda/2)\|\mathbf{z}\|_2^2+\sum_{i=1}^{N}\left(-(\mathbf{y}_i^k)^\top\mathbf{z}+(\rho/2)\|\mathbf{x}_i^{k+1}-\mathbf{z}\|_2^2\right)\right)\\ \mathbf{y}_i^{k+1} &:= \mathbf{y}_i^k+\rho\left(\mathbf{x}_i^{k+1}-\mathbf{z}^{k+1}\right).\end{aligned}$$

Since the objective function is convex, we can exploit the optimality condition to obtain the exact update rules as follows:

$$\begin{aligned}\mathbf{x}_i^{k+1} &= (\mathbf{A}_i^\top\mathbf{A}_i+\rho I)^{-1}(\mathbf{A}_i^\top\mathbf{b}_i-\mathbf{y}_i^k+\rho\mathbf{z}^k)\\ \mathbf{z}^{k+1} &= \frac{1}{\lambda+\rho N}\sum_{i=1}^{N}(\rho\mathbf{x}_i^{k+1}+\mathbf{y}_i^k)\\ \mathbf{y}_i^{k+1} &:= \mathbf{y}_i^k+\rho\left(\mathbf{x}_i^{k+1}-\mathbf{z}^{k+1}\right).\end{aligned}$$