SI251 - Convex Optimization, Spring 2022
Homework 3

Due on May 22, 2022, before class

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.
- 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.
- No late submission will be accepted.
- 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.
- Handwritten assignment is acceptable, but we prefer a LaTex version.

I. Proximal Gradient

1.Given $y \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, recall lasso criterion:

$$f(\beta) = \frac{1}{2}\|y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1,$$

where we can define $g(\beta) = \frac{1}{2}\|y - \boldsymbol{X}\beta\|_2^2$ and $h(\beta) = \lambda\|\beta\|_1$

(a) (10 points) Find the proximal operator and write down the proximal gradient update. (Hint: Iterative shrinkage-thresholding algorithms (ISTA))

Solution: The proximal operator is the soft-thresholding as

$$\left(\text{prox}_{\lambda h}(\beta)\right)_i = S_\lambda(\beta)_i = \left(|\beta_i| - \lambda\right)_+ \text{sign}\left(\beta_i\right)$$

and the proximal gradient update is given by

$$\beta^{t+1} = S_{\lambda\eta}\left(\beta^t - \eta\boldsymbol{X}^T\left(\boldsymbol{X}^T\beta - y\right)\right),$$

where η is the appropriate stepsize.

(b) (10 points) Please give the convergence analysis of ISTA.(Hint: Detail can be found in paper [1])

Solution: For the constant stepsize, the Lipschiz constant of the gradient ∇f is $L = \lambda_{\max}\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ for each iteration. By [1, Lemma 2.3], we obtain

$$\begin{aligned}\frac{2}{L}\left(f\left(\beta^*\right) - f\left(\beta^{t+1}\right)\right) &\geq \left\|\beta^{t+1} - \beta^t\right\|^2 + 2\left\langle\beta^t - \beta^*, \beta^{t+1} - \beta^t\right\rangle \\ &= \left\|\beta^{t+1} - \beta^*\right\|^2 - \left\|\beta^t - \beta^*\right\|^2.\end{aligned}$$

Summing this inequality over $t = 0, \ldots, T-1$ gives

$$\frac{2}{L}\left(Tf\left(\beta^*\right) - \sum_{t=0}^{T-1} f\left(\beta^{t+1}\right)\right) \geq \left\|\beta^t - \beta^*\right\|^2 - \left\|\beta^0 - \beta^*\right\|^2.$$

By [1, Lemma 2.3], we also have

$$\frac{2}{L}\left(f\left(\beta^t\right) - f\left(\beta^{t+1}\right)\right) \geq \left\|\beta^t - \beta^{t+1}\right\|^2.$$

Multiplying this inequality by t and summing over $t = 0, \ldots, T-1$ gives

$$\frac{2}{L}\sum_{t=0}^{T-1}\left(tF\left(\beta^t\right) - (t+1)F\left(\beta^{t+1}\right) + F\left(\beta^{t+1}\right)\right) \geq \sum_{t=0}^{T-1} t\left\|\beta^t - \beta^{t+1}\right\|^2,$$

which can be simplified as

$$\frac{2}{L}\left(-Tf\left(\beta^t\right) + \sum_{t=0}^{T-1} f\left(\beta^{t+1}\right)\right) \geq \sum_{t=0}^{T-1} t\left\|\beta^t - \beta^{t+1}\right\|^2.$$

Together we can obtain

$$\frac{2T}{L}\left(f\left(\beta^*\right) - f\left(\beta^t\right)\right) \geq \left\|\beta^t - \beta^*\right\|^2 - \left\|\beta^0 - \beta^*\right\|^2 + \sum_{t=0}^{T-1} t\left\|\beta^t - \beta^{t+1}\right\|^2,$$

an hence it follows that

$$f\left(\beta^t\right) - f\left(\beta^*\right) \leq \frac{L\left\|\beta^0 - \beta^*\right\|^2}{2T}.$$

II. Dual Methods

2.(20 points) This problem consider the following composite convex problem:

$$\boldsymbol{x}_* := \underset{\boldsymbol{x}}{argmin}\{H(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{A}\boldsymbol{x})\},$$

where both $f : \mathbb{R}^n - > (-\infty, +\infty]$ and $g : \mathbb{R}^m - > (-\infty, +\infty]$ are proper, closed and convex extended real-valued functions, while the function f is further assumed to be σ-strongly convex for $\sigma > 0$, and $\boldsymbol{A}$ is a $m \times n$ matrix. Due to the strong convexity, the problem above has a unique optimal solution $\boldsymbol{x}_*$.

The above problem has the following equivalent constrained form:

$$\boldsymbol{x}_* = \underset{\boldsymbol{x}}{argmin}\, \underset{\boldsymbol{z}}{min}\{\hat{H}(\boldsymbol{x}, \boldsymbol{z}) := f(\boldsymbol{x}) + g(\boldsymbol{z}) : \boldsymbol{A}\boldsymbol{x} - \boldsymbol{z} = 0\}, \qquad (P)$$

where $H(\boldsymbol{x}) = \hat{H}(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x})$.

(a) (10 points) Find the dual problem of problem P.
Solution: The primal problem is

$$\min_{x,z} f(x) + g(z), \quad \text{s.t. } Ax - z = 0.$$

The Lagrange function is given by

$$\mathcal{L}(x, z, u) = f(x) + g(z) + u^T(Ax - z).$$

The dual problem is given by

$$\max_u -f^*\left(-A^T u\right) - g^*(u),$$

(b) (10 points) Using proximal gradient method to optimize problem (a) and write down the proximal gradient update. (Hint: Refer to paper [2])
Solution: Let $F(\boldsymbol{y}) = f^*\left(\boldsymbol{A}^T\boldsymbol{y}\right)$, suppose F has a lipschitz continuous gradient due to the strong convexity of f with $L_F = \frac{\|A\|_2^2}{\sigma}$, we have the following proximal gradient update for iteration t :

$$\begin{aligned}\boldsymbol{y}^{t+1} &= \arg\min_y \left\{ F\left(\boldsymbol{y}^t\right) + \left\langle \boldsymbol{y} - \boldsymbol{y}^t, \nabla F\left(\boldsymbol{y}^t\right)\right\rangle + \frac{L_F}{2}\left\|\boldsymbol{y} - \boldsymbol{y}^t\right\|_2^2 + g^*(-\boldsymbol{y})\right\} \\ &= \text{prox}_{1/L_F g^*}\left(\boldsymbol{y}^t - \frac{1}{L_F}\nabla F\left(\boldsymbol{y}^t\right)\right)\end{aligned}$$

III. ADMM

3.(a) (15 points) Consider the basis pursuit problem

$$\begin{aligned}&\text{minimize} \|x\|_1 \\ &\text{subject to } Ax = b\end{aligned}$$

This problem can be rewritten as the following form:

$$\begin{aligned}&\text{minimize } f(x) + g(y) \\ &\text{subject to } x - y = 0\end{aligned} \qquad (1)$$

where $f(x)$ is an indicator function of the set $\{x : Ax = b\}$, and $g(y) = \|y\|_1$. Recall the indicator function satisfies $f(x) = 0$ if $Ax = b$ and $f(x) = \infty$ if $Ax \neq b$. The augmented Lagrangian function is given as

$$L_\rho(x, y, \lambda) = f(x) + g(y) + \lambda^\top(x - y) + \frac{\rho}{2}\|x - y\|^2$$

Your task is to write out the ADMM update formula for (1) using the projection operator onto the set $\{x : Ax = b\}$ and the shrinkage operator. Specifically, write out x_{k+1}, y_{k+1}, and λ_{k+1} as functions of x_k, y_k, and λ_k. Simplify the arg min operation using the projection operator and the shrinkage operator.

(b) (15 points) Consider the following least square problem $\min_x \sum_{i=1}^n \frac{1}{2}\left(a_i^\top x - b_i\right)^2$ where $a_i \in \mathbb{R}^p$, $b_i \in \mathbb{R}$, and $x \in \mathbb{R}^p$. This problem can be rewritten as

$$\begin{aligned}&\text{minimize} \sum_{i=1}^n f_i\left(x^i\right) \\ &\text{subject to } x^i - y = 0, \forall i \in \{1, 2, \ldots, n\}\end{aligned}$$

where $f_i\left(x^i\right) = \frac{1}{2}\left(a_i^\top x^i - b_i\right)^2$, and $x^i \in \mathbb{R}^p$ is a vector having the same dimension as a_i. The augmented Lagrangian is given by

$$L_\rho = \sum_{i=1}^n \left\{ f_i\left(x^i\right) + \left(\lambda^i\right)^\top\left(x^i - y\right) + \frac{\rho}{2}\left\|x^i - y\right\|^2\right\}$$

Your task is to write out the ADMM update formula for the above problem. Specifically, express x_{k+1}^i, y_{k+1}, and λ_{k+1}^i as functions of $x_k^i, y_k, \lambda_k^i, a_i, b_i$ and ρ.
Solution: (a) ADMM updates x_{k+1} as follows:

$$\begin{aligned} x_{k+1} &= \arg\min_x L_\rho\left(x, y_k, \lambda_k\right) \\ &= \arg\min_{x:Ax=b}\left\{\lambda_k^\top\left(x-y_k\right)+\frac{\rho}{2}\left\|x-y_k\right\|^2\right\} \\ &= \arg\min_{x:Ax=b}\left\{\frac{\rho}{2}\left\|x-y_k+\lambda_k/\rho\right\|^2\right\} \end{aligned}$$

Therefore, we have

$$x_{k+1} = \mathrm{proj}_X\left(y_k - \frac{\lambda_k}{\rho}\right)$$

where X is the set $\{x : Ax = b\}$. Similarly, we can show

$$\begin{aligned} y_{k+1} &= S_{1/\rho}\left(x_{k+1}+\frac{\lambda_k}{\rho}\right) \\ \lambda_{k+1} &= \lambda_k + \rho\left(x_{k+1}-y_{k+1}\right) \end{aligned}$$

where $S_{1/\rho}$ is the shrinkage operator that shrinks every value between $-1/\rho$ and $1/\rho$ to 0.
(b) By definition, ADMM iterates as

$$\begin{aligned} x_{k+1}^i &= \arg\min_{x^i}\left\{f_i\left(x^i\right)+\left(\lambda_k^i\right)^\top\left(x^i-y_k\right)+\frac{\rho}{2}\left\|x^i-y_k\right\|^2\right\} \\ y_{k+1} &= \arg\min_y\left\{\sum_{i=1}^n\left(-\left(\lambda_k^i\right)^\top y+\frac{\rho}{2}\left\|x^i-y\right\|^2\right)\right\} \\ \lambda_{k+1}^i &= \lambda_k^i+\rho\left(x_{k+1}^i-y_{k+1}\right) \end{aligned}$$

Since $f_i\left(x^i\right) = \frac{1}{2}\left(a_i^\top x^i - b_i\right)^2$, we eventually have

$$\begin{aligned} x_{k+1}^i &= \left(a_i a_i^\top+\rho I\right)^{-1}\left(a_i b_i+\rho y_k-\lambda_k^i\right) \\ y_{k+1} &= \frac{1}{n}\sum_{i=1}^n\left(x_{k+1}^i+\lambda_k^i/\rho\right) \\ \lambda_{k+1}^i &= \lambda_k^i+\rho\left(x_{k+1}^i-y_{k+1}\right) \end{aligned}$$

IV. Stochastic Gradient

4. (10 points) In this problem, we study a stochastic gradient method with a projection step. Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and m-strongly convex, and let C be a closed, convex set. Consider the projected stochastic gradient method

$$x^{l+1} = \Pi_C\left(x^l - \alpha^l G\left(x^l\right)\right),$$

where $G\left(x^l\right)$ is an unbiased estimate of $\nabla f\left(x^l\right)$. Assume that the randomness in $G\left(x^l\right)$ is independent of all past randomness in the algorithm. Letting $x^* = \arg\min_{x\in C} f(x)$, prove that the iterates satisfy the bound

$$\mathbb{E}\left\|x^{l+1}-x^*\right\|_2^2 \le \left(1-2\alpha^l m\right)\mathbb{E}\left\|x^l-x^*\right\|_2^2+\left(\alpha^l\right)^2 B^2$$

where $B^2 = \sup_{x\in C}\mathbb{E}\|G(x)\|_2^2$.
Solution: We use non-expansiveness of the projection operator and the fact that $x^\ell \in$ C to obtain

$$\begin{aligned} \left\|x^{\ell+1}-x^*\right\|_2^2 &= \left\|\Pi_C\left(x^\ell-\alpha^\ell G\left(x^\ell\right)\right)-\Pi_C\left(x^*\right)\right\|_2^2 \le \left\|x^\ell-x^*-\alpha^\ell G\left(x^\ell\right)\right\|^2 \\ &= \left\|x^\ell-x^*\right\|_2^2+\left(\alpha^\ell\right)^2\left\|G\left(x^\ell\right)\right\|^2-2\alpha^\ell\left\langle G\left(x^\ell\right), x^\ell-x^*\right\rangle \\ &\le \left\|x^\ell-x^*\right\|_2^2+\left(\alpha^\ell\right)^2\left\|G\left(x^\ell\right)\right\|^2-2\alpha^\ell\left\langle G\left(x^\ell\right)-G\left(x^*\right)\cdot x^\ell-x^*\right\rangle \end{aligned}$$

The third lines follows from optimality of x^*, whereby we know that $\langle G\left(x^*\right), x-\ x^*\rangle \ge 0$ for all $x \in C$. Now taking the expectations on both sides conditioned on $x^{\bar{e}}$, we obtain by unbiasedness of G

$$\begin{aligned} \mathbb{E}\left\|x^{\ell+1}-x^*\right\|_2^2 &\le \mathbb{E}\left\|x^\ell-x^*\right\|_2^2+\left(\alpha^\ell\right)^2 B^2-2\alpha^\ell\left\langle\nabla f\left(x^\ell\right)-\nabla f\left(x^*\right), x^\ell-x^*\right\rangle \\ &\le \left(1-2\alpha^\ell m\right)\mathbb{E}\left\|x^\ell-x^*\right\|_2^2+\left(\alpha^\ell\right)^2 B^2 \end{aligned}$$

where the second line follows by m-strong convexity of f. The result now follows via the tower property.

V. **Second-Order Methods**

5. Consider the function $f(x) = \|x\|_2^\beta$ for some $\beta > 1$.

(a) (15 points) Show that the Newton direction d^l for this function at an iterate $x^l \neq 0$ is given by $d^l = -\frac{1}{\beta-1}x^l$. Hint: The matrix inversion lemma may be useful:

$$\left(A + CBC^T\right)^{-1} = A^{-1} - A^{-1}C\left(B^{-1} + C^T A^{-1} C\right)^{-1} C^T A^{-1}$$

for square matrices A and B, and a third matrix C of appropriate dimension (assuming that all relevant inverses exist).

(b) (5 points) Suppose that we apply Newtons method with constant stepsize equal to 1 . For what starting points x^0 and values of $\beta > 1$ does the method converge to $x^* = 0$?

Solution: (a) Note that $\nabla\|x\| = \frac{x}{\|x\|}$ for $x \neq 0$. Using chain rule, we compute the gradient and Hessian:

$$\begin{aligned}
\nabla f(x) &= \beta\|x\|^{\beta-1}\frac{x}{\|x\|} = \beta\|x\|^{\beta-2}x \\
\nabla^2 f(x) &= \beta(\beta-2)\|x\|^{\beta-4}xx^T + \beta\|x\|^{\beta-2}I \\
&= \beta\|x\|^{\beta-4}\left[\|x\|^2 I + (\beta-2)xx^T\right].
\end{aligned}$$

(As a check, it can be seen that these equations are correct in the quadratic case $\beta = 2$). The given form of the Newton direction can be computed by using the given matrix inversion formula, and different choices of A, B and C are possible. (If $\beta = 2$, the problem is quadratic and so can be solved directly.) One choice that will work for $\beta \neq 2$

$$A = \|x\|^2 I, \quad B = \beta - 2, \quad C = x.$$

We then have

$$\begin{aligned}
\left[\nabla^2 f(x)\right]^{-1} &= \frac{1}{\beta\|x\|^{\beta-4}}\left[\frac{1}{\|x\|^2}I - \frac{1}{\|x\|^2}x\left(\frac{1}{\beta-2} + x^T\frac{I}{\|x\|^2}x\right)^{-1}x^T\frac{1}{\|x}\right] \\
&= \frac{1}{\beta\|x\|^{\beta-2}}\left[I - \frac{\beta-2}{\beta-1}\frac{xx^T}{\|x\|^2}\right]
\end{aligned}$$

Thus,

$$\begin{aligned}
d = -\left[\nabla^2 f(x)\right]^{-1}\nabla f(x) &= -\frac{1}{\beta\|x\|^{\beta-2}}\left[I - \frac{\beta-2}{\beta-1}\frac{xx^T}{\|x\|^2}\right]\beta\|x\|^{\beta-2}x \\
&= -\left(1 - \frac{\beta-2}{\beta-1}\right)x = -\frac{1}{\beta-1}x
\end{aligned}$$

An alternative and somewhat more direct way to verify the given descent direction is to check that $\nabla^2 f(x)d = -\nabla f(x)$ as follows:

$$\begin{aligned}
\nabla^2 f(x)d &= \left[\beta(\beta-2)\|x\|^{\beta-4}xx^T + \beta\|x\|^{\beta-2}I\right]\left[-\frac{1}{\beta-1}x\right] \\
&= -\beta\|x\|^{\beta-2}x\left[\frac{(\beta-2)}{\beta-1} + \frac{1}{\beta-1}\right] = -\nabla f(x).
\end{aligned}$$

With this descent direction $d^\ell = -\frac{1}{\beta-1}x^\ell$, the Newton update with stepsize 1 has the form $x^{\ell+1} = x^\ell + d^\ell = \frac{\beta-2}{\beta-1}x^\ell$.

(b) The pure Newton update converges to $x^* = 0$ for any initial point $x^0 \in \mathbb{R}^d$ as long as $\left|\frac{\beta-2}{\beta-1}\right| < 1$, or equivalently $\beta > \frac{3}{2}$. For $\beta \in \left(1, \frac{3}{2}\right]$, it converges only for $x^0 = 0$; for all non-zero initializations $x^0 \neq 0$, it either oscillates (for $\beta = 3/2$) or it diverges (for $\beta \in (1, 3/2)$).

References

[1] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[2] D. Kim and J. A. Fessler, "Fast dual proximal gradient algorithms with rate $\mathbf{O}(1/k^{1.5})$ for convex minimization," *arXiv preprint arXiv:1609.09441*, 2016.