ECE 490: Introduction to Optimization **Fall 2018**

Homework 2

Instructor: Bin Hu Due date: September 27, 2018

1. In this problem, you will be asked to perform several calculations, and these calculations eventually lead to the convergence rate proof for Nesterov's accelerated method applied to smooth strongly-convex objective functions. Recall Nesterov's method is

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f((1+\beta)x_k - \beta x_{k-1})$$

which can also be written as

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ v_k &= C\xi_k \\ u_k &= \nabla f(v_k) \end{aligned}$$

where $A = \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix}$, $B = \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix}$, $C = \begin{bmatrix} (1+\beta)I & -\beta I \end{bmatrix}$, and $\xi_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$.

(a) Assume f is L-smooth and m-strongly convex. By L-smoothness and m-strong convexity, we have

$$\begin{aligned} f(x_k) - f(x_{k+1}) &= f(x_k) - f(v_k) + f(v_k) - f(x_{k+1}) \\ &\geq \nabla f(v_k)^\mathsf{T}(x_k - v_k) + \frac{m}{2}\|x_k - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2 \\ &= \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X_1 \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \end{aligned}$$

The last step in the above derivation requires substituting $x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha \nabla f(v_k)$ and $v_k = C\xi_k$ into the second-to-last line $\nabla f(v_k)^\mathsf{T}(x_k - v_k) + \frac{m}{2}\|x_k - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2$ and rewriting the resultant quadratic function. You task is figuring out this symmetric matrix X_1.

(b) Similarly, by L-smoothness and m-strong convexity, we have

$$\begin{aligned} f(x^*) - f(x_{k+1}) &= f(x^*) - f(v_k) + f(v_k) - f(x_{k+1}) \\ &\geq \nabla f(v_k)^\mathsf{T}(x^* - v_k) + \frac{m}{2}\|x^* - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2 \\ &= \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X_2 \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \end{aligned}$$

The last step in the above derivation requires substituting $x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha\nabla f(v_k)$ and $v_k = C\xi_k$ into the second-to-last line $\nabla f(v_k)^\mathsf{T}(x^* - v_k) + \frac{m}{2}\|x^* - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2$ and rewriting the resultant quadratic function. You task is figuring out this symmetric matrix X_2.

(c) Now based on the inequalities in (a) and (b), you can simply choose $X = \rho^2 X_1 + (1-\rho^2)X_2$ for any $0 < \rho < 1$, and we have

$$\begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \leq -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*))$$

Based on the testing condition presented in the class, if there exists $P \geq 0$ such that

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - X \leq 0 \tag{1}$$

then the following inequality holds

$$\begin{aligned}(\xi_{k+1} - \xi^*)^\mathsf{T}P(\xi_{k+1} - \xi^*) - \rho^2(\xi_k - \xi^*)^\mathsf{T}P(\xi_k - \xi^*) &\leq \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \\ &\leq -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*))\end{aligned}$$

which directly leads to the linear convergence rate for Nesterov's method:

$$(\xi_{k+1} - \xi^*)^\mathsf{T}P(\xi_{k+1} - \xi^*) + f(x_{k+1}) - f(x^*) \leq \rho^2\left((\xi_k - \xi^*)^\mathsf{T}P(\xi_k - \xi^*) + f(x_k) - f(x^*)\right). \tag{2}$$

Finding P to satisfy (1) is not trivial. Your task is to show that (1) holds for $P = \frac{1}{2}\begin{bmatrix} \sqrt{L}I \\ (\sqrt{m} - \sqrt{L})I \end{bmatrix}\begin{bmatrix} \sqrt{L}I & (\sqrt{m} - \sqrt{L})I \end{bmatrix} \geq 0$, $\rho^2 = 1 - \sqrt{\frac{m}{L}}$, and $X = \rho^2 X_1 + (1-\rho^2)X_2$ (X_1 and X_2 are the answers you get in (a) and (b)) when $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}}$. In your calculation, you are allowed to directly use (without proof) the following fact:

The matrix $\begin{bmatrix} c_1 & c_2 & c_3 \\ c_2 & c_4 & c_5 \\ c_3 & c_5 & c_6 \end{bmatrix}$ is negative semidefinite if and only if $\begin{bmatrix} c_1 I & c_2 I & c_3 I \\ c_2 I & c_4 I & c_5 I \\ c_3 I & c_5 I & c_6 I \end{bmatrix}$ is negative semidefinite.

(Hint: The calculation here can be lengthy. So you are allowed to use some symbolic toolbox to help as long as you turn in the code.)

(d) Now based on your calculations, you know (2) holds. Since $P \geq 0$, you can then get

$$f(x_k) - f(x^*) \leq \left(1 - \sqrt{\frac{m}{L}}\right)^k \left((\xi_0 - \xi^*)^\mathsf{T}P(\xi_0 - \xi^*) + f(x_0) - f(x^*)\right)$$

Your task is using the above bound to show that one can choose $T = O(\sqrt{\frac{L}{m}}\log(\frac{1}{\varepsilon}))$ to guarantee $f(x_T) - f(x^*) \leq \varepsilon$.

2. Programming Assignment

(a) First, you are asked to implement the gradient method, Nesterov's method, and Heavy-ball method to solve the positive definite quadratic minimization problem:

$$\min_{x\in\mathbb{R}^p} \frac{1}{2}x^\mathsf{T}Qx + q^\mathsf{T}x + r \tag{3}$$

where Q is positive definite. A Matlab script that you can use to generate Q, q, and r is provided on the course website. In the code, you can specify the problem dimension p, the strong convexity parameter m, and the smoothness parameter L (Notice always choosing $m < L$). Then the code will generate (Q, q, r) for your given values of (m, L). You are also allowed to use any other scientific computing language. But then you are asked to generate (Q, q, r) using your own code if you choose to use other languages. For gradient method, you should experiment two cases: $\alpha = \frac{1}{L}$ and $\alpha = \frac{2}{m+L}$. For Nesterov's method, choose $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$. For Heavy-ball method, choose $\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$ and $\beta = \left(\frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}\right)^2$. Always start from the initial condition $x_0 = x_{-1} = (1; 1; \ldots; 1)^\mathsf{T}$ You are asked to turn in plots of the progression of objective values (relative to the minimum) for various problem sizes ($p = 100$; 500) and (m, L) values ($m = 1$, $L = 10$; $m = 0.1$, $L = 1000$). Notice for this quadratic problem, the optimal point $x^* = -Q^{-1}q$ can be directly computed when the dimension p is not that high. This can be used when you plot the progression of objective values relative to the minimum. The y axis for your plots should be in log scale. You can try different iteration number (e.g. $k = 1000$) until the algorithm converges. Then briefly discuss your findings relative to the convergence rate theory.

(b) In (a), you should observe that Heavy-ball method works well for the quadratic problem. Here you are asked to run the three algorithms for another problem. Consider a one-dimensional function whose gradient is defined as: $\nabla f(x) = 25x$ for $x < 1$, $\nabla f(x) = x + 24$ for $1 \leq x < 2$, and $\nabla f(x) = 25x - 24$ for $x \geq 2$. This function is 25-smooth and 1-strongly convex ($L = 25$ and $m = 1$). Again, for gradient method, you should experiment two cases: $\alpha = \frac{1}{L}$ and $\alpha = \frac{2}{m+L}$. For Nesterov's method, choose $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$. For Heavy-ball method, choose $\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$ and $\beta = \left(\frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}\right)^2$. Always start from the initial condition $3.07 \leq x_0 = x_{-1} \leq 3.46$. You are asked to turn in plots of the progression of objective values. No need to plot things in log scale this time. Then briefly discuss your findings. Does Heavy-ball method still converge?