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

Lecture 7

Unconstrained Optimization for Smooth Strongly-Convex Functions, Part VII

Lecturer: Bin Hu, Date:09/20/2018

In this lecture, we sketch out how to apply our routine to analyze Nesterov's method. Recall that Nesterov's method can be written as

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ v_k &= C\xi_k \\ u_k &= \nabla f(v_k) \end{aligned} \tag{7.1}$$

where $A = \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix}$, $B = \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix}$, and $C = \begin{bmatrix} (1+\beta)I & -\beta I \end{bmatrix}$. The convergence rate proof of Nesterov's method can be done by applying the dissipation inequality routine presented in Lecture 6.

1. Replace the nonlinear equation $u_k = \nabla f(v_k)$ in (7.1) by some quadratic inequality in the following form:

$$\begin{aligned} \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^\mathsf{T} X \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix} &\le -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*)) \\ &= \rho^2(f(x_k) - f(x_{k+1})) + (1-\rho^2)(f(x^*) - f(x_{k+1})) \end{aligned}$$

The key issue is how to figure out X. By L-smoothness and m-strong convexity of f, we have

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

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 will be asked to write out this symmetric matrix X_1 in Homework 2. Similarly, in

Homework 2 you will be asked to find X_2 such that

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

Then you can simply choose $X = \rho^2 X_1 + (1-\rho^2)X_2$ for any $0 < \rho < 1$, and we have

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

2. Test if there exists $P \geq 0$ such that

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - X \leq 0. \tag{7.2}$$

   If so, then the following inequality holds

$$(\xi_{k+1} - \xi^*)^\mathsf{T}P(\xi_{k+1} - \xi^*) - \rho^2(\xi_k - \xi^*)^\mathsf{T}P(\xi_k - \xi^*) \leq \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^\mathsf{T} X \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}$$

   which is exactly the so-called dissipation inequality $V_{k+1} - \rho^2 V_k \leq S(\xi_k, u_k)$ if we define $V_k = (\xi_k - \xi^*)^\mathsf{T}P(\xi_k - \xi^*)$ and $S(\xi_k, u_k) = \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^\mathsf{T} X \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}$. Clearly $V_k \geq 0$ due to the fact $P \geq 0$. In Homework 2, I will provide the value of P and you will be asked to verify that (7.2) holds with that P and $(\rho^2, \alpha, \beta) = (1 - \sqrt{\frac{m}{L}}, \frac{1}{L}, \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}})$

3. Now directly apply the supply rate condition to conclude $V_{k+1} + f(x_{k+1}) - f(x^*) \leq \rho^2\left(V_k + f(x_k) - f(x^*)\right)$. In Homework 2, you will be asked to convert this rate result into an ε-optimal iteration complexity result $O(\sqrt{\frac{L}{m}} \log \frac{1}{\varepsilon})$. Specifically, you will be asked to show that one can choose $T = O(\sqrt{\frac{L}{m}} \log \frac{1}{\varepsilon})$ to guarantee $f(x_T) - f(x^*) \leq \varepsilon$.

In Homework 2, you will be asked to flesh out all the detailed calculations for proving the accelerated rate of Nesterov's method.

Now we see that for L-smooth m-strongly convex objective function f, the iteration complexity can be improved from $O(\frac{L}{m} \log \frac{1}{\varepsilon})$ to $O(\sqrt{\frac{L}{m}} \log \frac{1}{\varepsilon})$. Is this the end of the story for optimization of smooth strongly-convex functions? The answer is no. Depending on the structure of f, sometimes new issues come up. For example, consider the ℓ_2-regularized logistic regression with the objective function $f = \frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-b_i a_i^\top x}) + \frac{\mu}{2}\|x\|^2$. In this case, there is a finite-sum structure $f = \frac{1}{n}\sum_{i=1}^{n} f_i$. If we directly apply Nesterov's method to this problem, at each iteration we need to calculate the full gradient $\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x)$. This full gradient evaluation requires calculating gradient on all f_i and then averaging. When n is large, the iteration cost is high. This motivates the application of stochastic gradient method. We will talk about this in the next lecture.