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

Lecture 9

Unconstrained Optimization of Smooth Convex Functions, Part I

Lecturer: Bin Hu, Date:09/27/2018

In the previous lectures, we have talked about optimization of smooth strongly-convex functions. What if the objective function f is just convex (not strongly-convex)? Recall a differentiable function f is convex if the following inequality holds for all x, y

$$f(x) \geq f(y) + \nabla f(y)^\mathsf{T}(x - y)$$

You can think convex functions as "0-strongly convex functions," although the m-strong convexity typically implicitly assume $m > 0$.

We need to answer three questions here.

1. Does the global min x^* exists and is it unique for convex f? No! The global min may not even exist. A trivial example is the linear function $f(x) = x$. Clearly $f(x) = f(y) + \nabla f(y)^\mathsf{T}(x - y)$ and f is convex. But there does not exist a global min for this function. When x^* exists, there may be multiple global mins. Just think about the function $f(x) = 0$. This function is convex and achieves its global min at any point x.

2. What algorithm shall we use? Suppose $\nabla f(x^*) = 0$. Then by the definition of convexity we have $f(x) \geq f(x^*) + \nabla f(x^*)^\mathsf{T}(x - x^*) = f(x^*)$ for any x. So x^* is a global min. Therefore any algorithm designed to solve $\nabla f(x^*) = 0$ can be applied. Again, we will discuss first-order methods including the gradient method and the momentum methods.

3. What performance guarantees can we say about these algorithms? For the gradient method, we can show $f(x_k) - f(x^*) = O(1/k)$. For Nesterov's accelerated method, we can show $f(x_k) - f(x^*) = O(1/k^2)$. We don't have linear convergence anymore. The convergence rate $O(1/k)$ and $O(1/k^2)$ are significantly slower than the linear convergence rate $O(\rho^{-k})$, and categorized as "sublinear convergence rates." We will discuss how to modify the dissipation inequality approach to show such sublinear convergence rates.

9.1 A Revisit of Dissipation Inequality

The dissipation inequality has the following form

$$V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq S(\xi_k, u_k)$$

where V is non-negative. Depending on the properties of S, the dissipation inequality reaches different conclusions.

1. If $S(\xi_k, u_k) \leq 0$, then we have $V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq 0$. This is a linear convergence in V. We used this argument to show the linear convergence of the gradient method.

2. If $S(\xi_k, u_k) \leq -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*))$, we have $V(\xi_{k+1}) + f(x_{k+1}) - f(x^*) \leq \rho^2(V(\xi_k) + f(x_k) - f(x^*))$. This is also linear convergence. We used this argument to show the linear convergence of Nesterov's method.

3. How about having the condition $S(\xi_k, u_k) \leq f(x^*) - f(x_k)$ and $\rho^2 = 1$? Then the dissipation inequality leads to the inequality $V(\xi_{k+1}) - V(\xi_k) + f(x_k) - f(x^*) \leq 0$. Summing this inequality leads to

$$\sum_{t=0}^{k}(f(x_t) - f(x^*)) \leq V(\xi_0) - V(\xi_{k+1}) \leq V(\xi_0)$$

   The last step relies on the fact $V \geq 0$. If we know $f(x_t) \leq f(x_{t-1})$, then the left side of the above inequality can be lower bounded by $(k+1)(f(x_k) - f(x^*))$. Therefore, we eventually have

$$f(x_k) - f(x^*) \leq \frac{V(\xi_0)}{k+1} \tag{9.1}$$

   This is a sublinear rate result. We will use this argument to show that the gradient method is guaranteed to converge at the sublinear rate $O(1/k)$ when the objective function is smooth and convex.

9.2 Sublinear Convergence Rate of Gradient Method

The dissipation inequality is constructed by solving the following condition

$$\begin{bmatrix} A^\mathsf{T} PA - \rho^2 P & A^\mathsf{T} PB \\ B^\mathsf{T} PA & B^\mathsf{T} PB \end{bmatrix} - X \leq 0. \tag{9.2}$$

When f is smooth and convex, we want to show the gradient method converges at the rate $O(1/k)$. We have $A = I$, $B = -\alpha I$, and $\rho^2 = 1$. The key issue is how to choose X such that

$$\begin{bmatrix} x_k - x^* \\ \nabla f(x_k) \end{bmatrix}^\mathsf{T} X \begin{bmatrix} x_k - x^* \\ \nabla f(x_k) \end{bmatrix} \leq f(x^*) - f(x_k) \tag{9.3}$$

If f is L-smooth and convex, the following inequality (actually we have used this in HW1) holds for all x and y

$$f(x) \geq f(y) + \nabla f(y)^\mathsf{T}(x - y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$$

We set $x = x^*$ and $y = x_k$ in the above inequality. This leads to

$$f(x^*) \geq f(x_k) + \nabla f(x_k)^\mathsf{T}(x^* - x_k) + \frac{1}{2L}\|\nabla f(x^*) - \nabla f(x_k)\|^2$$

which can be rewritten in the form of (9.3) with $X = \begin{bmatrix} 0 & -\frac{1}{2}I \\ -\frac{1}{2}I & \frac{1}{2L}I \end{bmatrix}$. We can set $P = pI$ and the condition (9.2) just becomes

$$\begin{bmatrix} 0 & -\alpha p \\ -\alpha p & \alpha^2 p \end{bmatrix} - \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2L} \end{bmatrix} \leq 0 \tag{9.4}$$

When $\alpha = \frac{1}{L}$, we can set $p = \frac{L}{2}$. The left side of the above inequality just becomes a zero matrix and the testing condition is met. Now the dissipation inequality holds. In addition, we have $f(x_{k+1}) - f(x_k) \leq -\left(\alpha - \frac{L\alpha^2}{2}\right)\nabla f(x_k) \leq 0$. [1] So we do have $f(x_{k+1}) \leq f(x_k)$, and (9.1) follows as a consequence. Finally, we have shown the following bound holds for the gradient method with a smooth and convex objective function

$$f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2(k+1)}$$

9.3 Extension for Nesterov's Method

In the next lecture, we will modify Nesterov's method for smooth and convex objective functions. In this case, we will use time-varying parameters, i.e. α_k and β_k. Consequently we will have a time-varying optimization model:

$$\begin{aligned} \xi_{k+1} &= A_k\xi_k + B_k u_k \\ v_k &= C_k \xi_k \\ u_k &= \nabla f(v_k) \end{aligned} \tag{9.5}$$

Now (A, B, C) just depend on k. Nesterov's method can achieve a rate $O(1/k^2)$. This is faster than $O(1/k)$. The proof relies on solving a modified testing condition

$$\begin{bmatrix} A_k^\mathsf{T} P_{k+1} A_k - \rho^2 P_k & A_k^\mathsf{T} P_{k+1} B_k \\ B_k^\mathsf{T} P_{k+1} A_k & B_k^\mathsf{T} P_{k+1} B_k \end{bmatrix} - X_k \leq 0.$$

In the class I tried to briefly talk about this approach but clearly I confused a lot of people. So let's go through this in more details in the next lecture.

[1] In the proof of 1a in HW2, just set $\beta = 0$ and we directly get this result as a special case.