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

Lecture 2

Unconstrained Optimization for Smooth Strongly-Convex Functions, Part II

Lecturer: Bin Hu, Date:09/04/2018

This lecture focuses on the performance of the gradient method for the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^p} \; f(x) \tag{2.1}$$

where $f:\mathbb{R}^p\to\mathbb{R}$ is a differentiable function being L-smooth and m-strongly convex. We know there exists a unique global min x^* such that $f(x^*)\le f(x)$ for all $x\in\mathbb{R}^p$. The gradient method iterates as follows

$$x_{k+1} = x_k - \alpha\nabla f(x_k) \tag{2.2}$$

In the last lecture, it was mentioned that the gradient method satisfies $\|x_k - x^*\| \le \rho^k\|x_0 - x^*\|$ for some $0<\rho<1$ if a reasonable stepsize α is used. The smaller ρ is, the faster the gradient method converges to the optimal point x^*. However, ρ cannot be arbitrarily small (which means the gradient method cannot converge as fast as we want). In this lecture, we will study how ρ depends on m, L, and α.

The main theorem describing how ρ depends on m, L, and α is stated as follows.

Theorem 2.1. *Suppose f is L-smooth and m-strongly convex. Let x^* be the unique global min. Given a stepsize α, if there exists $0<\rho<1$ and $\lambda\ge 0$ such that*

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} \tag{2.3}$$

is a negative semidefinite matrix, then the gradient method satisfies $\|x_k - x^*\| \le \rho^k\|x_0 - x^*\|$.

The above theorem presents a sufficient testing condition for the linear convergence of the gradient method. We will use the theorem to analyze the convergence rate of the gradient method. First, we prove the theorem using a general trick called dissipation inequality (we will explain this terminology in Lecture 4).

2.1 A Useful Lemma

Denote the $p\times p$ identity matrix as I. The following lemma is very helpful and will be used to prove Theorem 2.1.

Lemma 2.2. *Suppose the sequences* $\{\xi_k \in \mathbb{R}^p : k = 0, 1, \ldots\}$ *and* $\{u_k \in \mathbb{R}^p : k = 0, 1, 2, \ldots\}$ *satisfy* $\xi_{k+1} = \xi_k - \alpha u_k$. *In addition, assume the following inequality holds for all* k

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0. \tag{2.4}$$

If there exist $0 < \rho < 1$ *and* $\lambda \geq 0$ *such that*

$$\begin{bmatrix} (1-\rho^2)I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M \tag{2.5}$$

is a negative semidefinite matrix, then the sequence $\{\xi_k : k = 0, 1, \ldots\}$ *satisfies* $\|\xi_k\| \leq \rho^k \|\xi_0\|$.

Proof: The key relation is

$$\|\xi_{k+1}\|^2 = \|\xi_k - \alpha u_k\|^2 = \|\xi_k\|^2 - 2\alpha(\xi_k)^\mathsf{T} u_k + \alpha^2 \|u_k\|^2 = \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \tag{2.6}$$

Since (2.5) is negative semidefinite, we have

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} (1-\rho^2)I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} + \lambda M \right) \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0 \tag{2.7}$$

We just expand the above inequality as

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} -\rho^2 I & 0_p \\ 0_p & 0_p \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0 \tag{2.8}$$

Applying the key relation (2.6), the above inequality can be rewritten as

$$\|\xi_{k+1}\|^2 - \rho^2 \|\xi_k\|^2 + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0 \tag{2.9}$$

Due to the condition (2.4) and the non-negativity of λ, we have

$$\|\xi_{k+1}\|^2 - \rho^2 \|\xi_k\|^2 \leq -\lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Hence $\|\xi_{k+1}\| \leq \rho \|\xi_k\|$ for all k. Therefore, we have $\|\xi_k\| \leq \rho \|\xi_{k-1}\| \leq \rho^2 \|\rho_{k-2}\| \leq \ldots \leq \rho^k \|\xi_0\|$. ■

It is emphasized that the condition (2.4) does not state that M is a positive semidefinite matrix. The inequality (2.4) is only assumed to hold for the two given sequences $\{\xi_k \in \mathbb{R}^p : k = 0, 1, \ldots\}$ and $\{u_k \in \mathbb{R}^p : k = 0, 1, 2, \ldots\}$. In addition, the relation $\xi_{k+1} = \xi_k - \alpha u_k$ is equivalent to

$$\xi_{k+1} = \begin{bmatrix} I & -\alpha I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}$$

which states that ξ_{k+1} is a linear function of (ξ_k, u_k). This is the reason why $\|\xi_{k+1}\|^2$ is just a quadratic form of (ξ_k, u_k) as shown in (2.6).

2.2 Proof of Theorem 2.1

When f is L-smooth and m-strongly convex (definitions are provided in the note for Lecture 1), one can prove the following inequality holds for $x, y \in \mathbb{R}^p$

$$(\nabla f(x) - \nabla f(y))^\mathsf{T}(x-y) \geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2 \tag{2.10}$$

This is the so-called co-coercivity property. You will be asked to prove this inequality in homework. This inequality can be rewritten as

$$\begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0. \tag{2.11}$$

Setting $y = x^*$ and noticing $\nabla f(x^*) = 0$, the above inequality leads to

$$\begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix} \geq 0. \tag{2.12}$$

The gradient method $x_{k+1} = x_k - \alpha \nabla f(x_k)$ can be rewritten as $x_{k+1} - x^* = x_k - x^* - \alpha \nabla f(x_k)$. We set $\xi_k = x_k - x^*$, and $u_k = \nabla f(x_k)$. Then the gradient method is exactly $\xi_{k+1} = \xi_k - \alpha u_k$ where (ξ_k, u_k) satisfies

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0. \tag{2.13}$$

The above inequality is just a restatement of (2.12). Therefore, we can choose $M = \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix}$ and apply Lemma 2.2 to directly prove Theorem 2.1. The final fact required for the proof is that $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is negative semidefinite if and only if $\begin{bmatrix} aI & bI \\ bI & cI \end{bmatrix}$ is negative semidefinite (verify this!).

2.3 Convergence Rates of Gradient Method

Now we apply Theorem 2.1 to obtain the convergence rate ρ for the gradient method with various stepsize choices.

- Case 1: If we choose $\alpha = \frac{1}{L}$, $\rho = 1 - \frac{m}{L}$, and $\lambda = \frac{1}{L^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} -\frac{m^2}{L^2} & \frac{m}{L^2} \\ \frac{m}{L^2} & -\frac{1}{L^2} \end{bmatrix} = \frac{1}{L^2}\begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix} \tag{2.14}$$

The right side is clearly negative semidefinite due to the fact that $\begin{bmatrix} a \\ b \end{bmatrix}^\mathsf{T} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = -(ma-b)^2 \leq 0$. Therefore, the gradient method with $\alpha = \frac{1}{L}$ converges as

$$\|x_k - x^*\| \leq \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\| \tag{2.15}$$

- Case 2: If we choose $\alpha = \frac{2}{m+L}$, $\rho = \frac{L-m}{L+m}$, and $\lambda = \frac{2}{(m+L)^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{2.16}$$

  The zero matrix is clearly negative semidefinite. Therefore, the gradient method with $\alpha = \frac{2}{m+L}$ converges as

$$\|x_k - x^*\| \leq \left(\frac{L-m}{L+m}\right)^k \|x_0 - x^*\| \tag{2.17}$$

Notice $L \geq m > 0$ and hence $1 - \frac{m}{L} \geq \frac{L-m}{L+m}$. This means the gradient method with $\alpha = \frac{2}{m+L}$ converges slightly faster than the case with $\alpha = \frac{1}{L}$. However, m is typically unknown in practice. The step choice of $\alpha = \frac{1}{L}$ is also more robust (we will discuss this in later sections). The most popular choice for α is still $\frac{1}{L}$.

We will give interpretations for the above convergence rates in the next lecture.

Finally, in Homework 1, you will be asked to express ρ as a function of α. Hence you have to choose λ carefully for a given α. Notice $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is negative semidefinite if and only if $c \leq 0$ and $ac - b^2 \geq 0$. So $\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix}$ is negative semidefinite if and only if

$$(1-\rho^2-2mL\lambda)(\alpha^2-2\lambda) - (\lambda(m+L)-\alpha)^2 \geq 0 \tag{2.18}$$

$$\alpha^2 - 2\lambda \leq 0 \tag{2.19}$$

which is equivalent to

$$\rho^2 \geq 1 - 2mL\lambda - \frac{(\lambda(m+L)-\alpha)^2}{\alpha^2-2\lambda} \tag{2.20}$$

$$\lambda \geq \frac{\alpha^2}{2} \tag{2.21}$$

In the homework, you will be guided to use the above formula to express ρ as a function of α after setting λ to some function of α.