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

Lecture 1

Unconstrained Optimization for Smooth Strongly-Convex Functions, Part I

Lecturer: Bin Hu, Date:08/30/2018

This lecture focuses on the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^p} \; f(x) \tag{1.1}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is a differentiable function being L-smooth and m-strongly convex. A point $x^* \in \mathbb{R}^n$ is a global min of f if for all $x \in \mathbb{R}^n$ the following holds

$$f(x^*) \le f(x). \tag{1.2}$$

We will answer the following three important questions:

1. Does a global min x^* exist for f being smooth and strongly-convex? If it exists, is it unique?
2. What algorithm shall we use to find the global min x^* if it exists?
3. How does the algorithm perform?

In this lecture, for any $x \in \mathbb{R}^n$, $\|x\|$ just denotes the 2-norm of x.

1.1 Some definitions

A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is L-smooth if for all $x, y \in \mathbb{R}^p$ the following inequality holds

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|. \tag{1.3}$$

One can prove that (this will be left as a homework problem) the definition of L-smoothness leads to the following useful inequality holding for all $x, y \in \mathbb{R}^p$

$$f(x) \le f(y) + \nabla f(y)^\mathsf{T}(x - y) + \frac{L}{2}\|x - y\|^2. \tag{1.4}$$

We say f is m-strongly convex (for some $m > 0$) if for all $x, y \in \mathbb{R}^p$ the following inequality holds

$$f(x) \ge f(y) + \nabla f(y)^\mathsf{T}(x - y) + \frac{m}{2}\|x - y\|^2. \tag{1.5}$$

We provide some interpretations for the above two properties. Given any fixed y, the right side of (1.4) provides a quadratic upper bound for f, and the right side of (1.5) provides a quadratic lower bound for f. This means that f more or less behaves similar to a quadratic function. Hence when f is L-smooth and m-strongly convex, the unconstrained optimization problem should not be that hard. This is basically true. We will give a comprehensive treatment for this case.

If f is L-smooth and m-strongly convex, one can show that (this is also going to be a homework problem) the following inequality (the so-called co-coercivity property) holds for all $x, y \in \mathbb{R}^n$

$$(\nabla f(x) - \nabla f(y))^\mathsf{T}(x-y) \geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2. \tag{1.6}$$

The above inequality is equivalent to

$$\begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mLI_p & (m+L)I_p \\ (m+L)I_p & -2I_p \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0$$

where the left side is in a quadratic form. The above inequality is going to be very useful!

1.2 Answer the first question: x^* exists and is unique!

The existence of x^* is typically shown by the following theorem, which is actually a corollary of the extreme value theorem.

Theorem 1.1. *If a continuous function f satisfies $\|f(x)\| \to \infty$ as $\|x\| \to \infty$, then a global min for f exists.*

When f is m-strongly convex, then it has a quadratic lower bound. Clearly $\|f(x)\| \to \infty$ as $\|x\| \to \infty$ (since the quadratic lower bound will blow to ∞ as $\|x\| \to \infty$). Then we know x^* exists.

Another useful theorem (called the optimality condition) is the following

Theorem 1.2. *Suppose f is differentiable and x^* is a global min of f. Then $\nabla f(x^*) = 0$.*

You will be asked to prove the above theorem in the homework.

Since we have already shown that x^* exists for strongly convex f, now we will further show x^* is unique. Assume x_1^* and x_2^* are both global min for f. We have $\nabla f(x_1^*) = \nabla f(x_2^*) = 0$. By (1.5), we have

$$f(x_1^*) \geq f(x_2^*) + \frac{m}{2}\|x_1^* - x_2^*\|^2$$
$$f(x_2^*) \geq f(x_1^*) + \frac{m}{2}\|x_1^* - x_2^*\|^2$$

Adding the above two inequalities leads to the conclusion $\|x_1^* - x_2^*\| \leq 0$. Hence $x_1^* = x_2^*$. Therefore the global min for f is unique.

Notice the existence and uniqueness of x^* do not require the smoothness property. However, non-smooth problems are more difficult to solve, as we will see in later lectures.

1.3 Answer the second question: Gradient method!

To find x^*, a classical method is the gradient decent method which iterates as

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \tag{1.7}$$

where α is a prescribed constant (called stepsize) one has to determine beforehand. One can choose any initial condition $x_0 \in \mathbb{R}^n$, and compute $x_1, x_2, \ldots, x_k, \ldots$. Here we assume given any x, one has the access to the first-order derivative information $\nabla f(x)$. Hence the gradient method is a first-order optimization method.

The gradient method has the advantage that it only requires the first-order derivative. In addition, when f is L-smooth and m-strongly convex, the gradient method is guaranteed to converge at a linear rate to the optimal point x^*.

1.4 Answer the third question: Linear convergence!

One can show that the gradient method satisfies

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\| \tag{1.8}$$

for some $0 \leq \rho < 1$. In controls literature, the above convergence behavior is called exponential convergence. However, in optimization literature, the above convergence behavior is called linear convergence. The reason is that if one takes the log of $\rho^k \|x_0 - x^*\|$, one gets $k \log \rho + \log \|x_0 - x^*\|$, which is a linear function of k. In this course, we will adopt the terminology "linear convergence."

Clearly the smaller ρ is, the faster x_k converges to x^*. However, ρ cannot be arbitrarily small. This means the convergence speed of the algorithm depends on the parameter choice α and also the function properties (m and L).

The following theorem describes the dependence between ρ and (α, m, L).

Theorem 1.3. *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 $\lambda \geq 0$ such that*

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} \tag{1.9}$$

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

The proof of the above theorem is very important and can be generalized in many different ways. So we will prove the above theorem in next lecture.

How to apply the above theorem? Given (m, L), one can use the theorem to determine what value of α leads to the smallest ρ (the fastest convergence rate) by showing (1.9) is a negative semidefinite matrix. Recall that $A \in \mathbb{R}^{p\times p}$ is positive semidefinite (p.s.d.) if $x^T A x \geq 0$ for any $x \in \mathbb{R}^p$. Similarly, $A \in \mathbb{R}^{p\times p}$ is negative semidefinite (n.s.d.) if $x^T A x \leq 0$

for any $x \in \mathbb{R}^p$. How to determine whether A is p.s.d? General methods involve calculating the eigenvalues of A or the leading principle minors. However, in many situations, it is just obvious. For example, the inequality $(a - cb)^2 \geq 0$ just states that $\begin{bmatrix} 1 & -c \\ -c & c^2 \end{bmatrix}$ is p.s.d. since $(a - cb)^2 = \begin{bmatrix} a \\ b \end{bmatrix}^\mathsf{T} \begin{bmatrix} 1 & -c \\ -c & c^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$. In later lectures, we will see a lot of examples of this flavor.