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

Lecture 4

Unconstrained Optimization for Smooth Strongly-Convex Functions, Part IV

Lecturer: Bin Hu, Date:09/11/2018

In last lecture, we talked about two application examples: ridge regression and logistic regression. In these two example, the objective function is L-smooth and m-strongly convex. Therefore, you can use the gradient method to achieve the iteration complexity $O(\kappa \log(\frac{1}{\varepsilon}))$. This means that if we want to guarantee $\|x_T - x^*\| \leq \varepsilon$, then we need to scale T linearly with κ. In this lecture, we will introduce momentum methods that can accelerate the optimization of smooth strongly-convex functions. Specifically, Nesterov's accelerated method can improve the iteration complexity from $O(\kappa \log(\frac{1}{\varepsilon}))$ to $O(\sqrt{\kappa} \log(\frac{1}{\varepsilon}))$. This improvement is significant. Just consider $\kappa = 10000$. Then $\sqrt{\kappa} = 100$. This states that Nesterov's method is roughly 100 times faster than the gradient method in this case.

4.1 Further Comments on Gradient Descent Method

Suppose the objective function is L-smooth and m-strongly convex. In previous lectures, we have showed that the convergence rate of the gradient method with $\alpha = \frac{1}{L}$ is $\rho = 1 - \frac{1}{\kappa}$. A natural question is whether we can refine our analysis and prove an improved convergence rate for the gradient method. The answer is no. There exists a function f being L-smooth and m-strongly convex and an associated initial condition x_0 such that $\|x_k - x^*\| = \left(1 - \frac{1}{\kappa}\right)^k \|x_0 - x^*\|$. So there is no way to improve ρ for the gradient method when seeking for a convergence rate guarantee for all the smooth strongly-convex functions.

To find a such f, just consider a quadratic function $f = \frac{1}{2} x^\mathsf{T} Q x$ with a positive definite $Q > 0$. We have $\nabla f(x_k) = Q x_k$. Clearly the global min is $x^* = 0$. The gradient method $x_{k+1} = x_k - \alpha \nabla f(x_k)$ just becomes $x_{k+1} = (I - \alpha Q) x_k$. Now we use the following fact.

Fact. If λ is an eigenvalue of Q, then $1 - \alpha\lambda$ is the eigenvalue of $I - \alpha Q$.

Please verify the above fact by yourself.

Remember m is the smallest eigenvalue of Q. When $\alpha = \frac{1}{L}$, the matrix $I - \alpha Q$ has an eigenvalue at $1 - \frac{m}{L}$. Choose x_0 as the eigenvector associated with this eigenvalue, we have $\|x_k - x^*\| = \left(1 - \frac{1}{\kappa}\right)^k \|x_0 - x^*\|$.

Basically the iteration complexity $O(\kappa \log(\frac{1}{\varepsilon}))$ is tight for the gradient method.

4.2 Motivations for Accelerated Methods

Recall the convergence rate analysis we have done for the gradient method. The iteration complexity result $O(\kappa \log(\frac{1}{\varepsilon}))$ only requires the following inequality to hold for a particular

x^* and all x

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

We do not even require the following inequality to hold for all x and y

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

It is likely that the gradient method does not fully explore the properties of smooth strongly-convex functions and this leads to a slow convergence. There is a possibility that we can refine the optimization method to explore L-smooth and m-strongly convex properties better so that we can eventually achieve an improved accelerated rate. This is actually the case. Now we introduce such accelerated methods.

4.3 Momentum Methods

Momentum methods use the gradient information and the one-step memory x_{k-1}. One such example is the Heavy-ball method that iterates as

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \tag{4.1}$$

The extra term $\beta(x_k - x_{k-1})$ is the so-called "momentum term." One needs to choose the stepsize α and the momentum β, and also initialize the method at x_0 and x_{-1}. Then based on this iteration, one can compute x_1, x_2, ...

With well-chosen α and β, the Heavy-ball method achieves faster convergence rate than the gradient method for a positive definite quadratic minimization problem. However, the same choice of α and β may not work for other smooth strongly-convex functions. On the other hand, Nesterov's method is proved to have an improved iteration complexity $O(\sqrt{\kappa}\log(\frac{1}{\varepsilon}))$ for all the functions being L-smooth and m-strongly convex.

Nesterov's accelerated method has the form

$$y_k = x_k + \beta(x_k - x_{k-1}) \tag{4.2}$$
$$x_{k+1} = y_k - \alpha \nabla f(y_k) \tag{4.3}$$

We can simply rewrite Nesterov's method as

$$x_{k+1} = x_k - \alpha \nabla f((1+\beta)x_k - \beta x_{k-1}) + \beta(x_k - x_{k-1}) \tag{4.4}$$

This looks very similar to Heavy-ball method. The difference is that Nesterov's accelerated method uses a gradient evaluated at $(1+\beta)x_k - \beta x_{k-1}$ while Heavy-ball method uses a gradient evaluated at x_k.

It is worth mentioning that both Heavy-ball method and Nesterov's method only use the first-order derivative (gradient) and do not require evaluating the second-order derivative (Hessian). Hence they belong to "first-order optimization methods."

We will not directly present the convergence rate proofs for Nesterov's method. We will first introduce a general model for first-order optimization methods. Then in later lectures we will present a unified analysis for the general model and then the iteration complexity results of Nesterov's method will be obtained as a special case of our general analysis.

4.4 A General Model for First-Order Methods

A general model for first-order optimization methods is the following

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ v_k &= C\xi_k \\ u_k &= \nabla f(v_k) \end{aligned} \tag{4.5}$$

where A, B, and C are matrices with compatible dimensions. In this general model, we can choose (A, B, C) accordingly to recover various first-order methods.

1. For gradient method, we choose $A = I$, $B = -\alpha I$, $C = I$, and $\xi_k = x_k$. Then $v_k = C\xi_k = x_k$, and $u_k = \nabla f(v_k) = \nabla f(x_k)$. The iteration $\xi_{k+1} = A\xi_k + Bu_k$ just becomes $x_{k+1} = Ax_k + Bu_k = x_k - \alpha\nabla f(x_k)$, which is exactly the gradient method.

2. For Heavy-ball method, we choose $A = \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix}$, $B = \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix}$, $C = \begin{bmatrix} I & 0 \end{bmatrix}$, and $\xi_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$. Then $v_k = C\xi_k = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} = x_k$, and $u_k = \nabla f(v_k) = \nabla f(x_k)$. The iteration $\xi_{k+1} = A\xi_k + Bu_k$ becomes
$$\begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} = \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} + \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix} \nabla f(x_k) = \begin{bmatrix} (1+\beta)x_k - \beta x_{k-1} - \alpha\nabla f(x_k) \\ x_k \end{bmatrix}$$
which is exactly the iteration for Heavy-ball method.

3. For Nesterov's accelerated method, we choose $A = \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix}$, $B = \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix}$, $C = \begin{bmatrix} (1+\beta)I & -\beta I \end{bmatrix}$, and $\xi_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$. Then $v_k = C\xi_k = \begin{bmatrix} (1+\beta)I & -\beta I \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} = (1+\beta)x_k - \beta x_{k-1}$, and $u_k = \nabla f(v_k) = \nabla f((1+\beta)x_k - \beta x_{k-1})$. The iteration $\xi_{k+1} = A\xi_k + Bu_k$ becomes
$$\begin{aligned} \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} &= \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} + \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix} \nabla f(v_k) \\ &= \begin{bmatrix} (1+\beta)x_k - \beta x_{k-1} - \alpha\nabla f((1+\beta)x_k - \beta x_{k-1}) \\ x_k \end{bmatrix} \end{aligned}$$

which is exactly the iteration (4.4) for Nesterov's accelerated method.

We can see that the only difference between Nesterov's accelerated method and Heavy-ball method is the choice of C. The different choices of C lead to completely different performance guarantees for these two methods when applied to smooth strongly-convex objective functions. In later lectures, we will provide some unified analysis routine for the general model (4.5). Then we will recover the iteration complexity $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ for Nesterov's method as a special case of our general analysis.