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

Lecture 3

Unconstrained Optimization for Smooth Strongly-Convex Functions, Part III

Lecturer: Bin Hu, Date:09/06/2018

Previously, we have shown that the gradient method $x_{k+1} = x_k - \alpha \nabla f(x_k)$ converges to the global min at a linear rate when the objective function f is L-smooth and m-strongly convex. In this lecture, we will do three things.

1. Show that a quadratic function with a positive definite Hessian is L-smooth and m-strongly convex.
2. Convert the convergence rate of the gradient method to some "iteration complexity" bounds.
3. Introduce two application examples with L-smooth and m-strongly convex objective functions: ridge regression and logistic regression.

In this lecture, we will need the following fact.

Fact 1: For a symmetric matrix A, one always has $\lambda_{\min}\|x\|^2 \leq x^{\mathsf{T}}Ax \leq \lambda_{\max}\|x\|^2$ where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of A, respectively.

More discussions about the above fact can be found in Prof. Srikant's note. The discussion is presented in the last three pages of "some linear algebra facts" listed on the following site: https://sites.google.com/site/ece490spring2017/lecture-notes

3.1 Positive Definite Quadratic Problem

Consider the following objective function

$$f(x) = \frac{1}{2}x^{\mathsf{T}}Qx + q^{\mathsf{T}}x + r \tag{3.1}$$

where $Q \in \mathbb{R}^{p\times p}$, $q \in \mathbb{R}^p$, and $r \in \mathbb{R}$. In addition, Q is assumed to be positive definite. Denote $\lambda_{\min}$ and $\lambda_{\max}$ as the smallest and largest eigenvalues of Q, respectively. Since Q is positive definite, we have $\lambda_{\max} \geq \lambda_{min} > 0$.

We will show that f is $\lambda_{\max}$-smooth and $\lambda_{\min}$-strongly convex.

- Smoothness: Notice $\nabla f(x) = Qx + q$ (verify this yourself!) and the largest eigenvalue of Q^2 is $\lambda_{\max}^2$ (why?). We have

  $$\|\nabla f(x) - \nabla f(y)\| = \|Q(x-y)\| = \sqrt{(x-y)^{\mathsf{T}}Q^2(x-y)} \leq \sqrt{\lambda_{\max}^2\|x-y\|^2} = \lambda_{\max}\|x-y\|$$

  Therefore, by definition f is $\lambda_{\max}$-smooth.

- Strong convexity: It is straightforward to verify that

$$f(x) - f(y) - \nabla f(y)(x-y) = \frac{1}{2}(x-y)^\mathsf{T} Q(x-y) \geq \frac{\lambda_{\min}}{2}\|x-y\|^2$$

  Therefore, we have $f(x) \geq f(y) + (\nabla f(y))^\mathsf{T}(x-y) + \frac{\lambda_{\min}}{2}\|x-y\|^2$ and f is $\lambda_{\min}$-strongly convex.

Therefore, if we apply the gradient method $x_{k+1} = x_k - \alpha \nabla f(x_k)$ to minimize f, the iterates x_k will converge to the unique global min x^* linearly. If we choose $\alpha = \frac{1}{\lambda_{\max}}$, we will have $\rho = 1 - \frac{\lambda_{\min}}{\lambda_{\max}}$.

In general, the ratio $\kappa := \frac{L}{m}$ is called the condition number. The condition number describes how "difficult" the optimization problem is. If the condition number is small, it means the problem is "well conditioned" and should be relatively easy. If the condition number is large, it means the problem is "ill conditioned" and should be more difficult.

3.2 From convergence rate to iteration complexity

The convergence rate ρ naturally leads to an iteration number T guaranteeing the algorithm to achieve the so-called ε-optimality, i.e. $\|x_T - x^*\| \leq \varepsilon$ [1].

To guarantee $\|x_T - x^*\| \leq \varepsilon$, we can use the bound $\|x_T - x^*\| \leq \rho^T \|x_0 - x^*\|$. If we choose T such that $\rho^T \|x_0 - x^*\| \leq \varepsilon$, then we guarantee $\|x_T - x^*\| \leq \varepsilon$. Denote $c = \|x_0 - x^*\|$. Then $c\rho^k \leq \varepsilon$ is equivalent to

$$\log c + k \log \rho \leq \log(\varepsilon) \tag{3.2}$$

Notice $\rho < 1$ and $\log \rho < 0$. The above inequality is equivalent to

$$k \geq \log\left(\frac{\varepsilon}{c}\right) / \log \rho = \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho) \tag{3.3}$$

So if we choose $T = \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$, we guarantee $\|x_T - x^*\| \leq \varepsilon$.

Notice $\log \rho \leq \rho - 1 < 0$ (this can be proved using the concavity of log function and we will talk about concavity in later lectures), so $\frac{1}{1-\rho} \geq -\frac{1}{\log \rho}$ and we can also choose $T = \log\left(\frac{c}{\varepsilon}\right) / (1-\rho) \geq \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$ to guarantee $\|x_T - x^*\| \leq \varepsilon$.

Another interpretation for $T = \log\left(\frac{c}{\varepsilon}\right) / (1-\rho)$ is that a first-order Taylor expansion of $-\log \rho$ at $\rho = 1$ leads to $-\log \rho \approx 1 - \rho$. So $\log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$ is roughly equal to $\log\left(\frac{c}{\varepsilon}\right) / (1-\rho)$ when ρ is close to 1.

Clearly the smaller T is, the more efficient the optimization method is. The iteration number T describes the "ε-optimal iteration complexity" of the gradient method for smooth strongly-convex objective functions.

[1] In many situations people require ε-optimal solution x_T to satisfy $f(x_T) - f(x^*) \leq \varepsilon$. We will talk about this case in late lectures. Typically this ends up with the same iteration complexity since we have $f(x) - f(x^*) = O(\|x - x^*\|^2)$ in many cases.

- For the gradient method with $\alpha = \frac{1}{L}$, we have $\rho = 1 - \frac{m}{L} = 1 - \frac{1}{\kappa}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) = \kappa \log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log(\frac{1}{\varepsilon})\right)$. [2] Here we use the big O notation to highlight the dependence on κ and ε and hide the dependence on the constant c.

- For the gradient method with $\alpha = \frac{2}{L+m}$, we have $\rho = \frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa+1}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) = \frac{\kappa+1}{2}\log\left(\frac{c}{\varepsilon}\right)$. Although $\frac{\kappa+1}{2} \leq \kappa$, we still have $\frac{\kappa+1}{2}\log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log(\frac{1}{\varepsilon})\right)$. Therefore, the stepsize $\alpha = \frac{2}{m+L}$ can only improve the constant C hidden in the big O notation of the iteration complexity. People call this "improvement of a constant factor".

- In general, when ρ has the form $\rho = 1 - 1/(a\kappa + b)$, the resultant iteration complexity is always $O\left(\kappa \log(\frac{1}{\varepsilon})\right)$.

How shall we interpret the iteration complexity $O\left(\kappa \log(\frac{1}{\varepsilon})\right)$? It states that the required iteration T scales with the condition number κ. For larger κ, more iterations are required. This is consistent with our intuition since larger κ means the problem is ill-conditioned and more difficult to solve. In later lectures, we will introduce more sophisticated algorithms to decrease the iteration complexity for unconstrained optimization problems with smooth strongly-convex objective functions. Specifically, we will discuss Nesterov's method which decreases the iteration complexity from $O\left(\kappa \log(\frac{1}{\varepsilon})\right)$ to $O\left(\sqrt{\kappa} \log(\frac{1}{\varepsilon})\right)$.

3.3 Two application examples

Finally we will discuss two application examples for unconstrained optimization with smooth strongly-convex objective functions.

3.3.1 Ridge regression

The ridge regression is formulated as an unconstrained minimization problem with the following objective function

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}(a_i^\mathsf{T} x - b_i)^2 + \frac{\lambda}{2}\|x\|^2 \tag{3.4}$$

where $a_i \in \mathbb{R}^p$ and $b_i \in R$ are data points used to fit the linear model x.

- What is this problem about? The purpose of this problem is to fit a linear relationship between a and b. One wants to predict b from a as $b = a^\mathsf{T} x$. The ridge regression gives a way to find such x based on the observed pairs of (a_i, b_i).

[2] For any functions $h(\varepsilon, \kappa)$ and $g(\varepsilon, \kappa)$, we say $h(\varepsilon, \kappa) = O(g(\varepsilon, \kappa))$ if there exists a constant C such that $|h(\varepsilon, \kappa)| \leq C|g(\varepsilon, \kappa)|$.

- Why is there a term $\frac{\lambda}{2}\|x\|^2$? The term $\frac{\lambda}{2}\|x\|^2$ is called ℓ_2-regularizer. It confines the complexity of the linear predictors you want to use. The high-level idea is that you want x to work for all (a, b), not just the observed pairs (a_i, b_i). This is called "generalization" in machine learning. So adding such a term can induce the so-called stability and helps the predictor x to "generalize" for the data you have not seen. You need to take a machine learning course if you want to learn about generalization.

- What is λ? λ is a hyperparameter which is tuned to trade off training performance and generalization. For the purpose of this course, let's say λ is a fixed positive number. In practice, λ is typically set as a small number between 10^{-8} and 0.1.

Now we will show that f is L-smooth and m-strongly convex.
It is straightforward to verify that

$$f(x) = \frac{1}{2}x^\mathsf{T}\left(\frac{2}{n}\sum_{i=1}^n a_i a_i^\mathsf{T} + \lambda I\right)x - \left(\frac{2}{n}\sum_{i=1}^n b_i a_i\right)^\mathsf{T} x + \frac{1}{n}\sum_{i=1}^n b_i^2$$

which is a special case of (3.1) with (Q, q, r) defined as

$$\begin{aligned} Q &= \frac{2}{n}\sum_{i=1}^n a_i a_i^\mathsf{T} + \lambda I \\ q &= \frac{2}{n}\sum_{i=1}^n b_i a_i \\ r &= \frac{1}{n}\sum_{i=1}^n b_i^2 \end{aligned}$$

Notice Q is positive definite (why?). Therefore, we can apply gradient method to ridge regression, and obtain an iteration complexity $O(\kappa \log(\frac{1}{\varepsilon}))$ where κ is the condition number of the positive definite matrix $\frac{2}{n}\sum_{i=1}^n a_i a_i^\mathsf{T} + \lambda I$.

3.3.2 ℓ_2-Regularized Logistic regression

The ℓ_2-regularized logistic regression is formulated as an unconstrained minimization problem with the following objective function

$$f(x) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-b_i a_i^\mathsf{T} x}) + \frac{\lambda}{2}\|x\|^2 \tag{3.5}$$

where $a_i \in \mathbb{R}^p$ and $b_i \in \{-1, 1\}$ are data points used to fit the linear model x.

- What is this problem about? The purpose of this problem is to fit a linear "**classifier**" between a and b. Let's say you have collected a lot of images for cats and dogs. You

augment the pixels of any such image into a vector a and wants to predict whether the image is a cat or a dog. Let's say $b = 1$ if the image is a cat, and $b = -1$ if the image is a dog. So you want to predict b based on a. You want to find x such that $b = 1$ when $a^\mathsf{T} x \geq 0$, and $b = -1$ when $a^\mathsf{T} x < 0$. The logistic regression gives a way to find such x based on the observed feature/label pairs of (a_i, b_i). You may want to take a statistics course or a machine learning course if you want to learn more about logistic regression.

- Why is there a term $\frac{\lambda}{2}\|x\|^2$? Again, the term $\frac{\lambda}{2}\|x\|^2$ is the ℓ_2-regularizer. It is used to induce generalization and help x work on all the (a, b) not just the observed data points (a_i, b_i).

The function (3.5) is also L-smooth and m-strongly convex. Hence the gradient method can be applied here to achieve an iteration complexity of $O(\kappa \log(\frac{1}{\varepsilon}))$.