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

Lecture 11

Basics of Non-convex Optimization and Some Stepsize Rules

Lecturer: Bin Hu, Date:10/04/2018

So far we have talked about optimization of smooth convex functions. What if the functions are not convex? Let's talk about this topic.

11.1 One-Point Convexity

In general, the guarantees for optimization of all non-convex functions are weak. However, some of the non-convex functions still have nice properties that can be exploited for obtaining a global guarantee. One such property is the so-called "one-point convexity." Recall that we have used the following inequality to prove the linear convergence of the gradient method:

$$\begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix} \geq 0. \tag{11.1}$$

When f is L-smooth and m-strongly convex, we have

$$\begin{bmatrix} x-y \\ \nabla f(x)-\nabla f(y) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x)-\nabla f(y) \end{bmatrix} \geq 0. \tag{11.2}$$

which is actually more general than (11.1). So we actually have proved the gradient method linearly converges not only for smooth strongly-convex f but also for all f satisfying (11.1). Comparing (11.1) with (11.2), we can see that we just replace the arbitrary vector y in (11.2) with a specific point x^* in (11.1). Hence (11.1) can be viewed as a "one-point convexity" condition. For functions satisfying one-point convexity, we can still use the gradient method which is guaranteed to achieve linear convergence. Notice the above one-point convexity condition does not even require smoothness.

In phase retrieval problems, a commonly-used condition is the regularity condition. The global regularity condition states that the following inequality holds for some positive μ and λ

$$\begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -\lambda I & I \\ I & -\mu I \end{bmatrix} \begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix} \geq 0. \tag{11.3}$$

This is an equivalent form of the one-point convexity and has been used to show linear convergence of the gradient method for phase retrieval problems. One technical issue is that usually this condition only holds locally for phase retrieval problems. So a lot of phase retrieval research focuses on how to develop good initialization techniques that guarantee the initial condition of the gradient method is in the region where the regularity condition holds.

Several other one-point convexity conditions include the Polyak-Lojasiewicz (PL) condition, Quadratic Growth (QG) condition, and the restricted secant inequality. We will not cover these conditions in details. But the take-home message is that you can expect the problem to be relatively "simple" if the function satisfies some sort of one-point convexity condition.

11.2 Optimization of General Non-Convex Functions

For general non-convex functions, even finding a local min is NP-hard in the worst case. If f is smooth and also bounded below by some constant C, the gradient method is guaranteed to converge to a point whose gradient is 0 (or equivalently just the so-called stationary point). To see this, notice the L-smoothness directly leads to the following

$$\begin{aligned} f(x_{k+1}) \leq & f(x_k) + \nabla f(x_k)^{\mathsf{T}}(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ = & f(x_k) - \left(\alpha - \frac{L\alpha^2}{2}\right)\|\nabla f(x_k)\|^2 \end{aligned}$$

Summing the above inequality from $k = 0$ to $k = T$ and canceling terms, we have

$$f(x_{T+1}) \leq f(x_0) - \left(\alpha - \frac{L\alpha^2}{2}\right)\sum_{k=0}^{T}\|\nabla f(x_k)\|^2$$

This states the following inequality holds for all T

$$\left(\alpha - \frac{L\alpha^2}{2}\right)\sum_{k=0}^{T}\|\nabla f(x_k)\|^2 \leq f(x_0) - f(x_{T+1}) \leq f(x_0) - C$$

As long as $\alpha - \frac{L\alpha^2}{2} > 0$, we know $\sum_{k=0}^{T}\|\nabla f(x_k)\|^2$ is bounded and increases as T increases. We know a bounded monotone sequence converges to one point. Hence $\sum_{k=0}^{\infty}\|\nabla f(x_k)\|^2$ exists and $\sum_{k=T}^{\infty}\|\nabla f(x_k)\|^2$ converges to 0 as T goes to ∞. Notice $x_{k+T} - x_k = \alpha\sum_{t=k}^{k+T-1}\nabla f(x_t)$. Hence we can show $\{x_k\}$ is a Cauchy sequence and converges to one point. This point has to have a zero gradient since $\|\nabla f(x_k)\| \to 0$. Therefore, we have shown the gradient method converges to a stationary point.

In general, a stationary point may not even be a local min. Recall that x^* is a local min if there is a neighborhood U around x^* such that $f(x^*) \leq f(x)$ for all $x \in U$. A point x^* is a saddle point if for all neighborhoods U around x^*, there are $x, y \in U$ such that $f(x) \leq f(x^*) \leq f(y)$. So a natural question is whether we can at least avoid converging to some kinds of saddle points. A lot of recent research papers focus on how to escape strict saddle points. Before talking about what strict saddle points are, we first review some optimality conditions for local min.

Now we only consider twice-differentiable f. A sufficient condition guaranteeing x^* being a local min is $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0$. A necessary condition required by every local min x^* is $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \geq 0$. Generally speaking, if a stationary point x^* has a positive semidefinite Hessian, it is non-trivial to decide whether this is a local min or a saddle point. If a saddle point has a positive semidefinite Hessian, then it is hard to handle. On the other hand, if the Hessian for a saddle point x^* has a negative minimum eigenvalue, then this saddle point is a strict saddle point and it is relatively easy to handle. By Stable Manifold theorem, we can guarantee the gradient method with a random initialization does not converge to such strict saddle points with probability one.

To summarize, one focus of the cutting edge theoretical research for non-convex optimization is on how to escape certain kinds of saddle points. Escaping saddle points is still an important research topic and many people are working on this. For exposure purpose, we briefly talked escaping saddle points here. This topic is not going to be tested in homework or exam. However, the optimality conditions for local min is something that you may be tested in HW or exam.

11.3 Stepsize Rules

So far we have talked about the theoretical side of optimization. The theory looks at stepsize depending on the smoothness parameter L. How about practice? How to implement things? Now we talk about some stepsize rules for implementation of the gradient method.

1. Trial and error: grid α and start with trying some larger α first. Intuitively larger stepsize leads to faster convergence (although this is not always true). If the larger stepsize fails, then divide it by a factor of constant and try it again. Keep on shrinking the stepsize until the function value starts to decrease and converge. This is the trial-and-error approach. So in practice, you may have to try various stepsizes before you find something that works.

2. Direct line search: The line search method involves solving a one-dimensional optimization at every step. Specifically, choose the stepsize as follows

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}} f(x_k - \alpha \nabla f(x_k))$$

   So at every step just try to decrease the function value as much as you can. Although we already know that being greedy at every step may not help in the long run (e.g. using momentum is helpful in the long run but may not be greedy at every step), the line search is still a popular heuristic.

3. Armijo rule: This is also known as the backtracking search. Fix positive $\beta < 1$ and $\sigma < 1$ in advance. Then find the smallest integer m such that

$$f(x_k - \alpha_0 \beta^m \nabla f(x_k)) \leq f(x_k) - \sigma \alpha_0 \beta^m \|\nabla f(x_k)\|^2 \tag{11.4}$$

Here, start with $\beta = 0$. Then increase m until the above inequality is satisfied and use that m. When f is L-smooth, there always exists an integer m such that the above inequality holds. To see this, notice L-smoothness means

$$\begin{aligned} f(x_k - \alpha_0\beta^m \nabla f(x_k)) &\leq f(x_k) + \nabla f(x_k)^\mathsf{T}(-\alpha_0\beta^m\nabla f(x_k)) + \frac{L}{2}\|-\alpha_0\beta^m\nabla f(x_k)\|^2 \\ &= f(x_k) - \left(\alpha_0\beta^m - \frac{L\beta^{2m}\alpha_0^2}{2}\right)\|\nabla f(x_k)\|^2 \end{aligned}$$

If we choose m such that $\alpha_0\beta^m - \frac{L\beta^{2m}\alpha_0^2}{2} \geq \sigma\alpha_0\beta^m$ (which is equivalent to $\beta^m \leq \frac{2(1-\sigma)}{\alpha_0 L}$), we guarantee the condition (11.4) is satisfied. Since $\beta < 1$, there always exists m such that the Armijo rule can be used.

For machine learning problems, the learning rate (stepsize) of SGD is typically tuned using the trial-and-error approach. Another popular choice is the adaptive stepsize rule such as ADAM, AMSGRAD, etc. The point is that the stepsize rules for deterministic optimization and stochastic optimization are quite different.