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

Lecture 12

Review of the Covered Materials

Lecturer: Bin Hu, Date:10/09/2018

Today we reviewed some course materials that are relevant to the midterm.

12.1 Stepsize Rule

First, let's go through the stepsize rule again. When implementing the gradient method with the Armijo rule, you just need to find the smallest integer m such that

$$f(x_k - \alpha_0\beta^m \nabla f(x_k)) \leq f(x_k) - \sigma\alpha_0\beta^m \|\nabla f(x_k)\|^2 \tag{12.1}$$

where $\beta < 1$ and $\sigma < 1$ are fixed in advance. The parameter α_0 is also set up in advance and is typically the largest stepsize value you want to try. You can run a few iterations of the gradient method with constant stepsize to determine α_0. Once α_0, β, and σ are set up, then at every iteration k just start with $m = 0$. If (12.1) does not hold for $m = 0$, then increase m and test (12.1) again. Keep on increasing m until (12.1) is satisfied and use that m. Eventually the stepsize at step k is set up as $\alpha_k = \alpha_0\beta^m$ where m is the smallest integer such that (12.1) holds. 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 (12.1) is satisfied. Since $\beta < 1$, there always exists m such that the Armijo rule can be used. From the condition $\beta^m \leq \frac{2(1-\sigma)}{\alpha_0 L}$, you can see ensuring $\sigma < 1$ is also important. If $\sigma = 1$, then $\beta^m > \frac{2(1-\sigma)}{\alpha_0 L} = 0$ for all m and there is no guarantee that we can find m such that (12.1) holds. You should get ready to calculate α_k based on Armijo rule for some simple f. The good thing for Armijo rule is that one only needs to test from $m = 0, 1, \ldots$ and usually will quickly find a value of m such that (12.1) holds. On the other hand, the direct line search method determines α_k based on greedily solving a one-dimensional optimization method

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

In the above optimization problem, α is the decision variable. You should also be able to analytically solve $\arg\min_{\alpha\in\mathbb{R}} f(x_k - \alpha\nabla f(x_k))$ for some simple f by taking the derivative with respect to α.

12.2 Optimality Condition

Recall that x^* is a stationary point for differentiable f when $\nabla f(x^*) = 0$.

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 $x \in U$. Similarly, x^* is a local max if there is a neighborhood U around x^* such that $f(x^*) \geq f(x)$ for all $x \in U$. Saddle points are stationary points that are not local min or max. Given a point x^*, how do we know whether it is a local min or a saddle point? We use optimality conditions.

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. For example, given a convex function $f = x^4$, we know $\nabla f(x^*) = 0$ guarantees x^* to be a global min (and then of course also a local min) due to the convexity. Therefore we know $x^* = 0$ is a local min of $f = x^4$ even we only have $\nabla^2 f(0) = 0$. However, given $f = x^3$, we also have $\nabla f(0) = 0$ and $\nabla^2 f(0) = 0$. But this time $x^* = 0$ is not a local min. Therefore, one cannot directly determine whether x^* is a local min or a saddle point if the only available information is $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \geq 0$. There is one simple case that one can immediately determine that x^* is a saddle point. If $\nabla^2 f(x^*)$ has both positive and negative eigenvalues, then x^* is a strict saddle point.

For twice-differentiable convex f, we know $\nabla^2 f(x) \geq 0$ for all x. Then $\nabla f(x^*) = 0$ guarantees x^* to be a global min. For m-strongly convex f, we have $\nabla^2 f(x) \geq mI$ for all x. Then we know there exists a unique point satisfying $\nabla f(x^*) = 0$ and this is the global min. Therefore, if we know more properties of f, we may be able to say more even only given $\nabla f(x^*)$. But for general f, we need to look at Hessian.

Given a simple f and a point x^*, you need to be able to calculate $\nabla^2 f(x^*)$ and use this matrix to check whether x^* is a local min or a saddle point.

12.3 Convergence rates proofs

We use the dissipation inequality technique to prove convergence rates. The dissipation inequality appeared in the previous lectures has the following form:

$$V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq S(\xi_k, u_k)$$

We can use the dissipation inequality to prove various results:

1. If $S(\xi_k, u_k) \leq 0$, then the dissipation inequality becomes $V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq 0$. This is a linear convergence in V. We have used this type of arguments to show the linear convergence of the gradient method.

2. If $S(\xi_k, u_k) \leq -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*))$, we have $V(\xi_{k+1}) + f(x_{k+1}) - f(x^*) \leq \rho^2(V(\xi_k) + f(x_k) - f(x^*))$. This is a linear convergence in $V(\xi_k) + f(x_k) - f(x^*)$. We have shown the linear convergence of Nesterov's method via this type of arguments.

3. If $S(\xi_k, u_k) \leq f(x^*) - f(x_k)$ and $\rho^2 = 1$, then the dissipation inequality leads to the inequality $V(\xi_{k+1}) - V(\xi_k) + f(x_k) - f(x^*) \leq 0$. Summing this inequality leads to $\sum_{t=0}^{k}(f(x_t) - f(x^*)) \leq V(\xi_0) - V(\xi_{k+1})$. We have used this argument to show that the gradient method is guaranteed to converge at the sublinear rate $O(1/k)$ when the objective function is smooth and convex.

Understanding the convergence rate proofs for the gradient method with $\alpha = \frac{1}{L}$ for both the convex and strongly-convex cases is the most important. Also take a look at the solution of Problem 2 in HW2 to see how practice aligns with the convergence rate theory.