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

Lecture 13

Newton's Method

Lecturer: Bin Hu, Date:10/16/2018

So far we have talked about first-order optimization methods that only require evaluating the gradient. In this lecture, we talk about Newton's method which uses Hessian to accelerate the convergence. The pure form of Newton's method iterates as

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

Depending on how efficient one can compute Hessian for a given practical problem., using Hessian may or may not be a good idea in general. The main advantage of Newton's method is that it achieves superlinear convergence when it is initialized closed enough to a local min. The convergence bound has a form:

$$\|x_{k+1} - x^*\| \le c\|x_k - x^*\|^2 < 1$$

This is much faster than the linear convergence rate bound $\|x_{k+1} - x^*\| \le \rho\|x_k - x^*\|$. Therefore, if we already have a rough solution for a local min, we can quickly refine this rough solution and get an accurate solution by applying Newton's method.

13.1 Interpretations of Newton's Method

The gradient method $x_{k+1} = x_k - \alpha \nabla f(x_k)$ can be interpreted as follows. At each step k, we are actually solving a quadratic minimization problem

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^p} \left\{ f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2\alpha}\|x - x_k\|^2 \right\}$$

The quadratic cost $\left\{ f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2\alpha}\|x - x_k\|^2 \right\}$ is just the sum of the Taylor expansion of f at x_k and an ℓ_2 regularizar. If we know f is L-smooth, then we know

$$f(x) \le f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{L}{2}\|x - x_k\|^2$$

So the gradient method with $\alpha = \frac{1}{L}$ is actually minimizing the above quadratic upper bound for f at each k.

Can we improve the optimization process by minimizing a better quadratic estimation of f at each k? This natural question leads to Newton's method. Specifically, the pure form of Newton's method iterates as

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^p} \left\{ f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2}(x - x_k)^\mathsf{T} \nabla^2 f(x_k)(x - x_k) \right\} \tag{13.1}$$

At each step k, we estimate f by its second-order Taylor expansion $f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2}(x - x_k)^\mathsf{T}\nabla^2 f(x_k)(x - x_k)$ and then minimize this quadratic estimation. Intuitively, the Taylor expansion gives a good local estimate for f but may not give a good global estimation. Consequently, Newton's method sometimes does not even converge if the iterates are too far away from the optimal points.

13.2 A Rough Analysis of Newton's Method

Here we sketch some non-rigorous analysis of Newton's Method to offer some high level ideas about how superlinear convergence is achieved. We have

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}\nabla f(x_k)\| \\
&= \|(\nabla^2 f(x_k))^{-1}\left(\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k)\right)\| \\
&\le \|(\nabla^2 f(x_k))^{-1}\|\|\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k) + \nabla f(x^*)\|
\end{aligned}$$

Consider a local min x^* satisfying $\nabla f(x^*) = 0$ and $\nabla f(x^*) > 0$. When x_k is really closed to x^*, we expect that $\nabla f(x_k)$ is also positive definite and $\|(\nabla^2 f(x_k))^{-1}\|$ is bounded above by some positive constant β_1. Let's assume $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le M\|x - y\|$. Then we can bound the second term using the following relation:

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x_k + \gamma(x^* - x_k))\dot{(}x_k - x^*)d\gamma$$

Specifically, we have

$$\begin{aligned}
&\|\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k) + \nabla f(x^*)\| \\
=&\|\int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x_k + \gamma(x^* - x_k)))\dot{(}x_k - x^*)d\gamma\| \\
\le&\int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x_k + \gamma(x^* - x_k))\|\|x_k - x^*\|d\gamma \\
\le& M(\int_0^1 \gamma d\gamma)\|x_k - x^*\|^2 \\
=&\frac{M}{2}\|x_k - x^*\|^2
\end{aligned}$$

Eventually we have

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}\nabla f(x_k)\| \\
&= \|(\nabla^2 f(x_k))^{-1}\left(\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k)\right)\| \\
&\le \|(\nabla^2 f(x_k))^{-1}\|\|\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k) + \nabla f(x^*)\| \\
&\le \frac{M\beta_1}{2}\|x_k - x^*\|^2
\end{aligned}$$

The above analysis is not rigorous and roughly explains the superlinear convergence of Newton's method.

13.3 Issues of Newton's Method and Some Fixes

The pure form of Newton's method has some disadvantages.

1. The computation of the Hessian matrix can be expensive.

2. Denote $d_k = (\nabla^2 f(x_k))^{-1}\nabla f(x_k)$. To obtain d_k, one typically needs to solve a linear equation $\nabla^2 f(x_k)d_k = \nabla f(x_k)$. This can be expensive.

3. The Hessian may not be positive definite. Even worse, the Hessian may be singular and $(\nabla^2 f(x_k))^{-1}$ does not exist.

4. The pure form of Newton's method does not have global convergence guarantees. If it is initialized far away from the optimal solution, it may even diverge for a smooth strongly-convex function f. See Page 107 of Moritz Hardt's note on Convex Optimization and Approximation for such an example.

In the next lecture, we will talk about the Quasi-Newton methods which partially fix the first two issues. For the third issue, one typically perturb the Hessian matrix as $\nabla^2 f(x_k)+\delta_k I$ where δ_k is some positive number. When δ_k is sufficiently large, $\nabla^2 f(x_k) + \delta_k I$ is going to become positive definite. There is a trade-off here. If we choose δ_k to be too large, then we are almost just doing the gradient descent method and the Hessian information is not efficiently used. In practice, one needs to carefully select δ_k for singular Hessian. To fix the fourth issue, one can use the damped Newton $x_{k+1} = x_k - \alpha_k(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$ where α_k is determined by the Armijo rule. So we just choose $\alpha_k = \beta^m$ where m is the smallest integer such that

$$f\left(x_k - \beta^m(\nabla^2 f(x_k))^{-1}\nabla f(x_k)\right) \leq f(x_k) - \sigma\beta^m \nabla f(x_k)(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$$

Here $\beta < 1$ is some prescribed positive number. Then one can show the damped Newton has global convergence guarantees for smooth strongly-convex functions and self-concordant functions.

13.4 Some Variants of Newton's Method

We briefly mention a few variants for Newton's method. One issue for Newton's method is that the quadratic function $f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2}(x - x_k)^\mathsf{T}\nabla^2 f(x_k)(x - x_k)$ may only be a good estimate for f when x is not far from x_k. What if we enforce x_{k+1} to be not far from x_k in the update? This is the idea of the trust region method. At each step k, the trust region method updates x_{k+1} as

$$x_{k+1} = \underset{\|x-x_k\|\leq\Delta_k}{\arg\min}\left\{f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2}(x - x_k)^\mathsf{T}\nabla^2 f(x_k)(x - x_k)\right\} \tag{13.2}$$

So we restrict x_{k+1} to be in a trust region $\|x - x_k\| \leq \Delta_k$. The parameter Δ_k can be tuned. When Δ_k is large, the trust region update behaves more similarly to Newton's method. The trust region method fixes the global convergence issue of Newton's method to some extent. It also gets a lot of recent attention due to its ability to escape saddle points. One can actually show that the trust region method can escape strict saddle points under some assumptions.

One can also add higher order term $\|x - x_k\|^3$ to the quadratic estimation $f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2}(x - x_k)^\mathsf{T}\nabla^2 f(x_k)(x - x_k)$. This is the idea of cubic regularization.