Chapter 13

Newton's Method

So far, we have dealt with methods that use first-order (gradient or subgradient) information about the objective function, or unbiased estimates of the gradient. We have shown that such algorithms can yield sequences of iterates that converge at linear or sublinear rates. We turn our attention in this chapter to methods that exploit second-derivative (Hessian) information. The canonical method here is Newton's method, named after Isaac Newton, who proposed a version of the method for polynomial equations in around 1670.

For many functions, including many that arise in data analysis, second-order information is not difficult to compute and manipulate. In comparing with first-order methods, there is a tradeoff. Second-order methods typically have locally superlinear convergence rates. (Once the iterates reach a neighborhood of a solution at which second-order sufficient conditions are satisfied, convergence is rapid.) Moreover, their global convergence properties are attractive. (With appropriate enhancements, they can provably avoid convergence to saddle points.) But the cost of calculating and handling the second-order information and computing the step is higher. Whether this tradeoff makes them appealing depends on the specifics of the application at hand.

We start by sketching the basic Newton method for the unconstrained smooth optimization problem min $f(x)$, and prove local convergence to a minimizer x^*. We discuss too Newton's method for the system of square nonlinear equations $F(x) = 0$, where $F : \mathbb{R}^n \to \mathbb{R}^n$, showing quadratic convergence to a solution x^* at which the Jacobian (the $n \times n$ matrix of first partial derivatives of the components of F) is nonsingular.

13.1 Basic Newton Method

Consider the problem

$$\min f(x), \tag{13.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz twice continuously differentiable function, where the Hessian has Lipschitz constant M, that is,

$$\|\nabla^2 f(x') - \nabla^2 f(x'')\| \leq M\|x' - x''\|. \tag{13.2}$$

Newton's method will generate a sequence of iterates $\{x^k\}_{k=0,1,2,\ldots}$. A second-order Taylor series approximation to f around the current iterate x^k is

$$f(x^k + p) \approx f(x^k) + \nabla f(x^k)^T p + \frac{1}{2} p^T \nabla^2 f(x^k) p. \tag{13.3}$$

When $\nabla^2 f(x^k)$ is positive definite, we can choose d to minimize the right-hand side explicitly. The minimizing value is

$$p^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k), \tag{13.4}$$

is the Newton step. Thus, in its most basic form, Newton's method is defined by the following iteration:

$$x^{k+1} = x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k). \tag{13.5}$$

We have the following local convergence result in the neighborhood of a point x^* satisfying second-order sufficient conditions.

Theorem 13.1. *Consider the problem* (13.1) *with f twice Lipschitz continuously differentiable with Lipschitz constant M defined in* (13.2). *Suppose that the second-order sufficient conditions are satisfied for the problem* (13.1) *at the point x^*, that is, $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq mI$ for some $m > 0$. Then if $\|x^0 - x^*\| \leq \frac{m}{2M}$, the sequence defined by* (13.5) *converges to x^* at a quadratic rate, with*

$$\|x^{k+1} - x^*\| \leq \frac{M}{m} \|x^k - x^*\|^2, \quad k = 0, 1, 2, \ldots . \tag{13.6}$$

Proof. From (13.4) and (13.5), and using $\nabla f(x^*) = 0$, we have

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \nabla^2 f(x^k)^{-1} \nabla f(x^k) \\ &= \nabla^2 f(x^k)^{-1} [\nabla^2 f(x^k)(x^k - x^*) - (\nabla f(x^k) - \nabla f(x^*))]. \end{aligned}$$

so that

$$\|x^{k+1} - x^*\| \leq \|\nabla f(x^k)^{-1}\| \|\nabla^2 f(x^k)(x^k - x^*) - (\nabla f(x^k) - \nabla f(x^*))\|. \tag{13.7}$$

By using Taylor's theorem (see (2.12) with $x = x^k$ and $p = x^* - x^k$), we have

$$\nabla f(x^k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^k + t(x^* - x^k))(x^k - x^*)\, dt.$$

By using this result along with the Lipschitz condition (13.2), we have

$$\begin{aligned} &\|\nabla^2 f(x^k)(x^k - x^*) - (\nabla f(x^k) - \nabla f(x^*))\| \\ &\quad = \left\| \int_0^1 [\nabla^2 f(x^k) - \nabla^2 f(x^k + t(x^* - x^k))](x^k - x^*)\, dt \right\| \\ &\quad \leq \int_0^1 \|\nabla^2 f(x^k) - \nabla^2 f(x^k + t(x^* - x^k))\| \|x^k - x^*\|\, dt \\ &\quad \leq \left(\int_0^1 Mt\, dt \right) \|x^k - x^*\|^2 = \tfrac{1}{2} M \|x^k - x^*\|^2. \end{aligned} \tag{13.8}$$

From the Weilandt-Hoffman inequality and (13.2), we have that

$$|\lambda_{\min}(\nabla^2 f(x^k)) - \lambda_{\min}(\nabla^2 f(x^*))| \leq \|\nabla^2 f(x^k) - \nabla^2 f(x^*)\| \leq M \|x^k - x^*\|,$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a symmetric matrix. Thus for

$$\|x^k - x^*\| \leq \frac{m}{2M}, \tag{13.9}$$

we have

$$\lambda_{\min}(\nabla^2 f(x^k)) \geq \lambda_{\min}(\nabla^2 f(x^*)) - M\|x^k - x^*\| \geq m - M\frac{m}{2M} \geq \frac{m}{2},$$

so that $\|\nabla^2 f(x^k)^{-1}\| \leq 2/m$. By substituting this result together with (13.8) into (13.7), we obtain

$$\|x^{k+1} - x^*\| \leq \frac{2}{m}\frac{M}{2}\|x^k - x^*\|^2 = \frac{M}{m}\|x^k - x^*\|^2,$$

verifying the locally quadratic convergence rate. By applying (13.9) again, we have

$$\|x^{x+1} - x^*\| \leq \left(\frac{M}{m}\|x^k - x^*\|\right)\|x^k - x^*\| \leq \frac{1}{2}\|x^k - x^*\|,$$

so, by arguing inductively, we see that the sequence converges to x^* provided that x^0 satisfies (13.9), as claimed. □

Of course, we do not need to explicitly identify a starting point x^0 in the stated region of convergence. Any sequence that approaches to x^* will eventually enter this region, and thereafter the quadratic convergence guarantees apply.

We consider now the smooth square system of nonlinear equations

$$F(x) = 0, \tag{13.10}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz continuously differentiable vector function. We use x^* to denote the solution of (13.12). The *Jacobian of* F is the matrix of first partial derivatives, the $n \times n$ matrix whose (i, j) entry is $\partial F_i/\partial x_j$. Similar to the case of scalar functions f, we model the vector function F near a point x^k by its first-order approximation:

$$F(x^k + p) \approx F(x^k) + J(x^k)p. \tag{13.11}$$

The accuracy of this approximation is given by another version of Taylor's theorem; see Theorem A.1.

Newton's method for nonlinear equations is simply the following recurrence:

$$x^{k+1} = x^k - J(x^k)^{-1}F(x^k). \tag{13.12}$$

We have the following quadratic convergence result for the case of nonsingular $J(x^*)$.

Theorem 13.2. *Suppose that x^* is a solution of (13.12) for which $J(x^*)$ is nonsingular. For x^0 sufficiently close to x^*, the sequence $\{x^k\}$ converges to x^* Q-superlinearly, that is,*

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \to 0. \tag{13.13}$$

If in addition J is Lipschitz continuous, that is,

$$\|J(x') - J(x'')\| \leq M\|x' - x''\|, \tag{13.14}$$

for some $M > 0$, we have that $\{x^k\}$ converges Q-quadratically, that is,

$$\|x^{k+1} - x^*\| \leq M\|J(x^*)^{-1}\|\|x^k - x^*\|^2. \tag{13.15}$$

Proof. We have from Theorem A.1 that

$$F(x^k) = F(x^k) - F(x^*) = J(x^k)(x^k - x^*) + \int_0^1 [J(x^* + t(x^k - x^*) - J(x^k)](x^k - x^*)\, dt. \quad (13.16)$$

For the second term on the right-hand side, we have

$$\begin{aligned}\left\| \int_0^1 [J(x^* + t(x^k - x^*) - J(x^k)](x^k - x^*)\, dt \right\| &\leq \int_0^1 \|J(x^* + t(x^k - x^*) - J(x^k)\| \|x^k - x^\| \, dt \\ &\leq \beta(x^k)\|x^k - x^*\|, \end{aligned} \quad (13.17)$$

where $\beta(x^k) \to 0$ as $x^k \to x^*$, and moreover we can choose $R_1 > 0$ to ensure that $\beta(x^k) \leq 1/(4\|J(x^*)^{-1}\|)$ for whenever $\|x^k - x^*\| \leq R_1$. Since J is nonsingular at x^* and continuous, we can choose a radius $R_2 > 0$ such that

$$\|J(x^k)^{-1}\| \leq 2\|J(x^*)^{-1}\|, \quad \text{whenever } \|x^k - x^*\| \leq R_2. \quad (13.18)$$

By multiplying both sides of (13.16) by $J(x^k)^{-1}$, and using (13.12), we have

$$x^k - x^{k+1} = (x^k - x^*) + J(x^k)^{-1} \int_0^1 [J(x^* + t(x^k - x^*) - J(x^k)](x^k - x^*)\, dt,$$

so by rearranging, taking norms, and using (13.17) and (13.18) we obtain

$$\|x^{k+1} - x^*\| \leq 2\|J(x^*)^{-1}\|\beta_k\|x^k - x^*\|. \quad (13.19)$$

Taking $k = 0$, and assuming that $\|x^0 - x^*\| \leq \min(R_1, R_2)$, we have from this expression that $\|x^1 - x^*\| \leq (1/2)\|x^0 - x^*\|$, so convergence of $\{x^k\}$ to x^* follows from an inductive argument. The superlinear rate (13.13) follows from (13.19) and the fact that $\beta(x^k) \to 0$ as $x^k \to x^*$.

For the quadratic rate, we have

$$\begin{aligned}\left\| \int_0^1 [J(x^* + t(x^k - x^*) - J(x^k)](x^k - x^*)\, dt \right\| &\leq \int_0^1 \|J(x^* + t(x^k - x^*) - J(x^k)\| \|x^k - x^\| \, dt \\ &\leq \int_0^1 M\|x^k - x^*\|^2 \int_0^1 t\, dt \\ &= \tfrac{1}{2} M\|x^k - x^*\|^2. \end{aligned}$$

Here we can set $\beta(x_k) = \frac{1}{2}M\|x^k - x^*\|$ in (13.17), so (13.15) follows immediately from (13.19). □

Global complexity for strongly convex?
Equivalent to steepest descent in a different metric (rescaled by Hessian).

13.2 Practical Newton Frameworks

Practical Newton frameworks: adding trust regions or line searches.