Lecture 5

In Lecture 4 we found out that gradient descent works, and has linear convergence. In this lecture we introduce Newton's method, an algorithm that takes advantage of the second derivative and has quadratic convergence under certain circumstances. Throughout this lecture, $\|\cdot\|$ will refer to the 2-norm $\|\cdot\|_2$.

5.1 Newton's Method

Let $f \in C^2(\mathbb{R}^n)$ and let's look again at the unconstrained problem

$$\text{minimize} \quad f(\boldsymbol{x}).$$

Newton's method starts with a guess $\boldsymbol{x}_0$ and then proceeds to compute a sequence of points $\{\boldsymbol{x}_k\}_{k\geq 0}$ in $\mathbb{R}^n$ by the rule

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \nabla^2 f(\boldsymbol{x}_k)^{-1}\nabla f(\boldsymbol{x}_k), \quad k \geq 0. \tag{5.1}$$

The algorithm stops when $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < \varepsilon$ for some predefined tolerance $\varepsilon > 0$. In the context of the general scheme $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$, the step length is $\alpha_k = 1$, and the search direction is the inverse of the Hessian multiplied with the negative gradient.

Recall that the inner product $\langle \boldsymbol{p}, \nabla f(\boldsymbol{x})\rangle$ is the directional derivative of f, and that a *descent direction* is a direction in which the rate of change (slope) is negative. The following gives a criterion for the search direction in Newton's method to be a descent direction.

Lemma 5.1. *Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be a positive definite symmetric matrix and $f \in C^1(\boldsymbol{R}^n)$. Then $\boldsymbol{p} = -\boldsymbol{B}^{-1}\nabla f(\boldsymbol{x})$ is a descent direction of f at $\boldsymbol{x}$.*

Proof. If $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, then $\boldsymbol{B}^{-1}$ is also positive definite, since for all $\boldsymbol{v} \in \mathbb{R}^n$,

$$\boldsymbol{v}^\top \boldsymbol{B}^{-1}\boldsymbol{v} = (\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{v})^\top \boldsymbol{B}^{-1}\boldsymbol{v} = (\boldsymbol{B}^{-1}\boldsymbol{v})^\top \boldsymbol{B}^\top(\boldsymbol{B}^{-1}\boldsymbol{v}) = (\boldsymbol{B}^{-1}\boldsymbol{v})^\top \boldsymbol{B}(\boldsymbol{B}^{-1}\boldsymbol{v}) > 0.$$

(This can also be seen by noting that the eigenvalues of $\boldsymbol{B}^{-1}$ are the inverses of the eigenvalues of $\boldsymbol{B}$.) For $\boldsymbol{p} = -\boldsymbol{B}^{-1}\nabla f(\boldsymbol{x})$ we then get

$$\langle \boldsymbol{p}, \nabla f(\boldsymbol{x})\rangle = -\langle \boldsymbol{B}^{-1}\nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x})\rangle = -\nabla f(\boldsymbol{x})^\top \boldsymbol{B}^{-1}\nabla f(\boldsymbol{x}) < 0,$$

which shows that $\boldsymbol{p}$ is a descent direction. □

To better understand Newton's method, we first look at the one dimensional case.

Example 5.2. Let $f \in C^2(\mathbb{R})$. In this case Newton's method is described as

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}, \quad k \geq 0. \tag{5.2}$$

Newton's method looks for a local minimizer in the form of a point x^* such that $f'(x^*) = 0$ and $f''(x^*) > 0$. Setting $g(x) := f'(x)$, we are looking for a *root* x^*,

$$g(x^*) = 0.$$

One approach to find such a root is to approximate the function $g(x)$ at a point x_k by its tangent line,

$$g(x) \approx g(x_k) + g'(x_k)(x - x_k),$$

and then identify the next iterate x_{k+1} as the root of this linear approximation:

$$g(x_k) + g'(x_k)(x_{k+1} - x_k) = 0 \iff x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

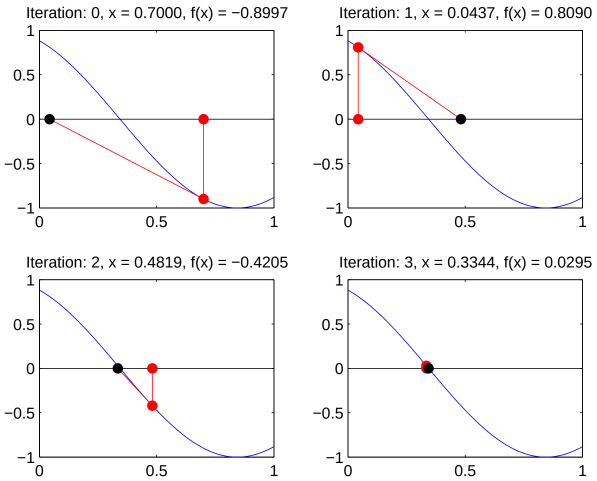

Figure 5.1: Newton's method

Geometrically this corresponds to taking the tangent to g at x_k and setting x_{k+1} to be the intersection of this tangent line with the x-axis, as shown in Figure 5.1. Replacing $g(x) = f'(x)$ gives precisely Newton's method (5.2).

Another way to understand Newton's method is to view it in contrast with gradient descent. While gradient descent corresponds to working with a *linear approximation*

$$f(\boldsymbol{x}_{k+1}) \approx f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle,$$

Newton's method is based on the *quadratic approximation*,

$$f(\boldsymbol{x}_{k+1}) \approx f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{1}{2} \langle \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle.$$

Example 5.3. Consider the function f on $\mathbb{R}^2$,

$$f(\boldsymbol{x}) = \frac{1}{2}(x_1^2 + 10x_2^2).$$

Starting with $\boldsymbol{x}_0 = (10, 1)^\top$, gradient descent takes 84 iterations to reach accuracy 10^{-6}, while Newton's method, unsurpringly, takes only one.

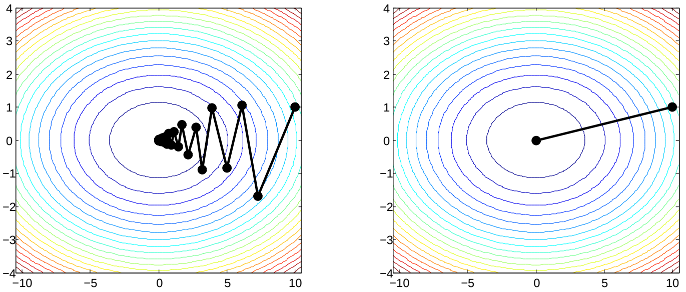

Figure 5.2: Gradient descent vs. Newton's method on a quadratic function.

In practice, when implementing Newton's method one does not explicitly compute the inverse of the Hessian. The reason is that one does not need the inverse itself, but only the product $\nabla^2 f(\boldsymbol{x}_k)^{-1}\nabla f(\boldsymbol{x}_k)$, which is the solution of a system of equations $\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{y} = \nabla f(\boldsymbol{x}_k)$ that can be solved much more efficiently. One therefore replaces the update step (5.1) with the following two steps:

$$\begin{aligned} \text{Find } \boldsymbol{y} \text{ such that} \quad \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{y} &= -\nabla f(\boldsymbol{x}_k), \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \boldsymbol{y}. \end{aligned}$$

There is a lot that can go wrong with Newton's method. In particular, the matrix $\nabla^2 f(\boldsymbol{x})$ has to be non-singular, or invertible, at every step. If, however, we start at a point $\boldsymbol{x}_0$ that is not too far from a local minimizer $\boldsymbol{x}^*$ with $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\nabla^2 f(\boldsymbol{x}^*)$ positive definite, then we are on the safe side.

Lemma 5.4. *Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be such that $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite. Then there exists an open neighbourhood U of $\boldsymbol{x}^*$ such that for all $\boldsymbol{x} \in U$, $\nabla^2 f(\boldsymbol{x})$ is positive definite.*

For the main result of this lecture, namely the quadratic convergence of Newton's method, we make the additional assumption that the Hessian $\nabla^2 f(\boldsymbol{x})$ is Lipschitz continuous as a function of $\boldsymbol{x}$.

Definition 5.5. A function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous on a domain $\Omega \subseteq \mathbb{R}^n$ with respect to a pair of norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ if there is a constant $L > 0$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$,

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

The constant L is called the *Lipschitz constant* of the map.

In particular, the Hessian of a function $f \in C^2(\mathbb{R}^n)$, considered as a map from $\mathbb{R}^n$ to $\mathbb{R}^{n\times n}$, is Lipschitz continuous with respect to a norm on $\mathbb{R}^n$ and the corresponding operator norm on $\mathbb{R}^{n\times n}$, if for any $\boldsymbol{x}, \boldsymbol{y}$ we have

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Theorem 5.6. *Let $f \in C^2(\mathbb{R}^n)$ and $\boldsymbol{x}^* \in \mathbb{R}^n$ a local minimizer with $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^*) > 0$. Then for $\boldsymbol{x}_0$ sufficiently close to $\boldsymbol{x}^*$, Newton's method has quadratic convergence.*

Proof. (Optional) Assume that $\nabla^2 f(\boldsymbol{x}_k)$ is positive definite. Consider the difference

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - \nabla^2 f(\boldsymbol{x}_k)^{-1}\nabla f(\boldsymbol{x}_k)\| \\ &\overset{(1)}{=} \|\boldsymbol{x}_k - \boldsymbol{x}^* - \nabla^2 f(\boldsymbol{x}_k)^{-1}(\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*))\| \\ &= \|\nabla^2 f(\boldsymbol{x}_k)^{-1}(\nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}_k - \boldsymbol{x}^*) - (\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*))\|\end{aligned} \tag{5.3}$$

where (1) follows from $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. The Fundamental Theorem of Calculus tells us

$$\begin{aligned}\nabla f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}_k) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\nabla f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k))\,\mathrm{d}t \\ &= \int_0^1 \nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k))(\boldsymbol{x}^* - \boldsymbol{x}_k)\,\mathrm{d}t.\end{aligned}$$

Continuing from (5.3), by inserting this identity,

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| &= \left\|\nabla^2 f(\boldsymbol{x}_k)^{-1}\int_0^1 \left[\nabla^2 f(\boldsymbol{x}_k) - \nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k))\right](\boldsymbol{x}_k - \boldsymbol{x}^*)\,\mathrm{d}t\right\| \\ &\leq \|\nabla^2 f(\boldsymbol{x}_k)^{-1}\|\cdot \\ &\quad \int_0^1 \|\nabla^2 f(\boldsymbol{x}_k) - \nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k))\|\,\mathrm{d}t \cdot \|(\boldsymbol{x}_k - \boldsymbol{x}^*)\|.\end{aligned}$$

Applying the Lipschitz bound to the term inside the integral gives

$$\|\nabla^2 f(\boldsymbol{x}_k) - \nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k))\| \leq Lt\|\boldsymbol{x}_k - \boldsymbol{x}^*\|.$$

Integrating this out, we end up with the bound

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq \frac{L}{2}\|\nabla^2 f(\boldsymbol{x}_k)^{-1}\| \cdot \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2.$$

The only remaining issue is that the "constant" on the right-hand side is not a constant. However, the Lipschitz continuity implies that $\nabla^2 f(\boldsymbol{x}_k)$ converges to $\nabla^2 f(\boldsymbol{x}^*)$ if $\boldsymbol{x}_k$ converges to $\boldsymbol{x}^*$. Since the inversion of a matrix is a continuous operation, also $\nabla^2 f(\boldsymbol{x}_k)^{-1}$ converges to $\nabla^2 f(\boldsymbol{x}^*)^{-1}$. In particular, if $\boldsymbol{x}_k$ is sufficiently close to $\boldsymbol{x}^*$, we have $\|\nabla^2 f(\boldsymbol{x}_k)^{-1}\| \leq 2\|\nabla^2 f(\boldsymbol{x}^*)\|$. Setting $M := L\|\nabla^2 f(\boldsymbol{x}^*)^{-1}\|/2$, we end up with the bound

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq M \cdot \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2.$$

By Lemma 5.4 there exists an open neighbourhood around $\boldsymbol{x}^*$ in which $\nabla^2 f(\boldsymbol{x})$ is positive definite, and within this neighbourhood there is an $\boldsymbol{x}_0$ such that $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 < 1/M$, which ensures that all following iterates remain in U. This shows quadratic convergence. □

Note that the conditions for quadratic convergence in an open neighbourhood U of $\boldsymbol{x}^*$ are precisely that f is convex on U.

5.2 Quasinewton methods

One drawback of Newton's method is that it requires the computation of the Hessian matrix, which can be expensive. Quasinewton methods use an approximation of the Hessian, $\boldsymbol{B}_k \approx \nabla f(\boldsymbol{x}_k)$, at each step of the algorithm. These are construction in a way that $\boldsymbol{B}_{k+1}$ can easily be computed from $\boldsymbol{B}_k$. A popular method, that is used often in practical applications because of its efficiency, is the **Broyden-Fletcher-Shanno-Goldfarb** (**BFGS**) method. The BFGS method may be described as follows.

- Start with $\boldsymbol{x}_0$, $\boldsymbol{B}_0$.
- For $k \geq 0$, compute

$$\begin{aligned}
\boldsymbol{p}_k &= \boldsymbol{B}_k^{-1}\nabla f(\boldsymbol{x}_k)\\
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k \text{ for a suitable step length } \alpha_k\\
\boldsymbol{s}_k &= \alpha_k \boldsymbol{p}_k\\
\boldsymbol{y}_k &= \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)\\
\boldsymbol{B}_{k+1} &= \boldsymbol{B}_k + \frac{\boldsymbol{y}_k \boldsymbol{y}_k^\top}{\boldsymbol{y}_k^\top \boldsymbol{s}_k} - \frac{\boldsymbol{B}_k \boldsymbol{s}_k \boldsymbol{s}_k^\top \boldsymbol{B}_k}{\boldsymbol{s}_k^\top \boldsymbol{B}_k \boldsymbol{s}_k}.
\end{aligned}$$

- Stop if $\|\nabla f(\boldsymbol{x}_k)\| < \varepsilon$ for some tolerance ε, or if $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < \varepsilon$.