1 Newton's Method and Steepest Descent (July 7)

1.1 Motivation for Newton's Method

Consider a twice-differentiable function $f : I \to \mathbb{R}$ defined on an interval $I \subseteq \mathbb{R}$. We would like to find the minima of f. We shall do so by considering quadratic approximations of f.

Let us start at a point $x_0 \in I$. Consider

$$q(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2,$$

the (best) quadratic approximation to f at x_0. Note that $q(x_0) = f(x_0)$, $q'(x_0) = f'(x_0)$ and $q''(x_0) = f''(x_0)$. We will now find the local minimizer x_1 for the quadratic q. That is, we would like to find x_1 such that

$$0 = q'(x_1) = f'(x_0) + f''(x_0)(x_1 - x_0),$$

implying that, so long as $f''(x_0) \neq 0$,

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}.$$

The idea of Newton's method is to iterate this procedure. (Consider the Newton's method for finding roots of functions; this is the same as finding the root of the derivative of the function.)

1.2 Newton's Method in One Dimension

Precisely, we pick a starting point $x_0 \in I$. Then we recursively define

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}.$$

We hope that the sequence x_n converges to a minimizer of f. For the sake of the rest of the lecture, let $g = f'$. With this notation we may write Newton's method as

$$x_0 \in I$$
$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}.$$

Theorem 1.1. *(Convergence of Newton's Method) Let $g \in C^2(I)$ (i.e. $f \in C^3(I)$). Suppose there is an $x_* \in I$ satisfies $g(x_*) = 0$ and $g'(x_*) \neq 0$. If x_0 is sufficiently close to x_*, then the sequence x_n generated by Newton's method converges to x_*.*

Proof. Since $g'(x_0) \neq 0$, there is, by continuity of g', an $\alpha > 0$ such that

1. $|g'(x_1)| > \alpha$ for all x_1 in a neighbourhood of x_0, and

2. $|g''(x_2)| < \frac{1}{\alpha}$ for all x_2 in the neighbourhood of x_0.

The proof of the first claim is a simple continuity argument. The proof of the second claim follows from continuity of g'' and the extreme value theorem applied to this neighbourhood('s closure). (That is, we can choose an α to bound $|g'|$ from below, and then shrink it possibly to ensure $1/\alpha$ bounds $|g''|$ from above.)

Since $g(x_*) = 0$, the formula of Newton's method now implies

$$x_{n+1} - x_* = x_n - x_* - \frac{g(x_n) - g(x_*)}{g'(x_n)} = -\frac{g(x_n) - g(x_*) - g'(x_n)(x_n - x_*)}{g'(x_n)}. \qquad (*)$$

By the second order mean value theorem, there exists a ξ sufficiently close to x_* such that

$$g(x_*) = g(x_n) + g'(x_n)(x_* - x_n) + \frac{1}{2}g''(\xi)(x_* - x_n)^2.$$

Then (*) becomes

$$x_{n+1} - x_* = \frac{1}{2}\frac{g''(\xi)}{g'(x_n)}(x_n - x_*)^2.$$

The bounds on g' and g'' we found at the start of the proof imply that

$$|x_{n+1} - x_*| < \frac{1}{2\alpha^2}|x_n - x_*|^2. \qquad (**)$$

Let ρ be the constant $\rho = \frac{1}{\alpha^2}|x_0 - x_*|$. Choose x_0 close enough to x_* so that $\rho < 1$. Then (**) implies

$$|x_1 - x_*| < \frac{1}{2\alpha^2}|x_0 - x_*||x_0 - x_*| = \rho|x_0 - x_*| < |x_0 - x_*|.$$

Similarly, (**) gives

$$|x_2 - x_*| < \frac{1}{2\alpha^2}|x_1 - x_*|^2 < \frac{1}{2\alpha^2}\rho^2|x_0 - x_*|^2 < \rho^2|x_0 - x_*|.$$

Continuing in the same way we obtain

$$|x_n - x_*| < \rho^n|x_0 - x_*|,$$

implying that Newton's method converges in our neighbourhood. □

1.3 Newton's Method in Higher Dimensions

Consider a function $f : \Omega \to \mathbb{R}$ defined on an open set $\Omega \subseteq \mathbb{R}^n$. We choose a starting point $x_0 \in \Omega$, and recursively define

$$x_{n+1} = x_n - \nabla^2 f(x_n)^{-1}\nabla f(x_n).$$

For a general f, the algorithm requires that $\nabla^2 f(x_n)$ is invertible. The algorithm stops if $\nabla f(x_n) = 0$ at some point (that is, the sequence given by Newton's method becomes constant if $\nabla f(x_n) = 0$ for some x_n.) Our main result is

Theorem 1.2. *(Convergence of Newton's Method) Suppose $f \in C^3(\Omega)$. Suppose also that there is an $x_* \in \Omega$ such that $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*)$ is invertible. Then the sequence x_n defined by*

$$x_{n+1} = x_n - \nabla^2 f(x_n)^{-1} \nabla f(x_n)$$

converges for all x_0 sufficiently close to x_*.

The goal of Newton's method was to find a minimizer of f, but it is possible for it to fail, for it only searches for *critical points*, not necessarily extrema.

1.4 Things That May Go Wrong

It is possible for Newton's method to fail to converge even when f has a unique global minimizer x_* and the initial point x_0 can be taken arbitrarily close to x_*. Consider

$$f(x) = \frac{2}{3}|x|^{3/2} = \begin{cases} \frac{2}{3}x^{3/2} & x \geq 0 \\ \frac{2}{3}(-x)^{3/2} & x \leq 0 \end{cases}.$$

This function is differentiable, and its derivative is

$$f'(x) = \begin{cases} x^{1/2} & x \geq 0 \\ -(-x)^{1/2} & x \leq 0 \end{cases}$$

and its second derivative is

$$f''(x) = \begin{cases} \frac{1}{2}x^{-1/2} & x > 0 \\ \frac{1}{2}(-x)^{-1/2} & x < 0 \\ \text{N/A} & x = 0 \end{cases},$$

so $f \notin C^3$ (it is not even C^2). Let $x_0 = \varepsilon$. Then

$$x_1 = \varepsilon - \frac{f'(\varepsilon)}{f''(\varepsilon)} = \varepsilon - \frac{\varepsilon^{1/2}}{\frac{1}{2}\varepsilon^{-1/2}} = \varepsilon - 2\varepsilon = -\varepsilon,$$

and

$$x_2 = -\varepsilon - \frac{f'(-\varepsilon)}{f''(-\varepsilon)} = -\varepsilon - \frac{-\varepsilon^{1/2}}{\frac{1}{2}\varepsilon^{-1/2}} = -\varepsilon + 2\varepsilon = \varepsilon.$$

So Newton's method gives an alternating sequence $\varepsilon, -\varepsilon, \varepsilon, -\varepsilon, \ldots$. This definitely does not converge. This does not contradict the theorem of convergence because the function in question does not satisfy the conditions of the theorem.

Now we consider an example in which the function in question converges, just not to a minimizer. Consider $f(x) = x^3$, which has derivatives $f'(x) = 3x^2$ and $f''(x) = 6x$. Starting at x_0, we have

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} = x_n - \frac{3x_n^2}{6x_n} = x_n - \frac{1}{2}x_n = \frac{1}{2}x_n.$$

So Newton's method definitely converges to the critical point 0, no matter the choice of $x_0 \in \mathbb{R}$. However, the function f in question does not have a global minimizer, so, while Newton's method converges, it does not converge to an extrema of any sorts.

1.5 Motivation for Steepest Descent

Consider a C^1 function $f : \Omega \to \mathbb{R}$ defined on an open set $\Omega \subseteq \mathbb{R}^n$. The idea is: at every point in the "landscape" of f (the graph of f in $\mathbb{R}^{n+1}$), make a step "downwards" in the steepest direction. (If you're on a mountain and want to descend to the bottom as fast as possible, how do you do so? You, at your current position, take a step down in the steepest direction, and repeat until you're done.)

Since the gradient $\nabla f(x_0)$ represents the direction of greatest increase of f at x_0, the vector $-\nabla f(x_0)$ represents the direction of steepest decrease at x_0. We would therefore like to move in the direction of the negative gradient. We will do so, with the condition that we move until we have a minimizer in the direction of the negative gradient (at which point we will stop moving and repeat).

1.6 Steepest Descent

Here is the steepest descent algorithm:

$$\begin{aligned} x_0 &\in \Omega \\ x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \end{aligned}$$

where $\alpha_k \geq 0$ satisfies

$$f(x_k - \alpha_k \nabla f(x_k)) = \min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k)).$$

We call α_k the *optimal step*, since it is chosen so that x_{k+1} is the minimum of f sufficiently close to x_k. We also call x_{k+1} the *minimum point on the half-line* $x_k - \alpha \nabla f(x_k), \alpha \geq 0$. We now describe some properties of the method of steepest descent.

Theorem 1.3. *The steepest descent algorithm is actually descending; $f(x_{k+1}) < f(x_k)$ so long as $\nabla f(x_k) \neq 0$.*

Proof. We have

$$f(x_{k+1}) = f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k - s \nabla f(x_k))$$

for all $s \in [0, \alpha_k]$. Also,

$$\left.\frac{d}{ds}\right|_{s=0} f(x_k - s \nabla f(x_k)) = \nabla f(x_k) \cdot (-\nabla f(x_k)) = -\|\nabla f(x_k)\|^2 < 0.$$

Then for sufficiently small $s \geq 0$,

$$f(x_k - s \nabla f(x_k)) < f(x_k),$$

proving the claim. □

Theorem 1.4. *The steepest descent algorithm moves in perpendicular steps; for all k, we have $(x_{k+2} - x_{k+1}) \cdot (x_{k+1} - x_k) = 0$.*

Proof. We have

$$(x_{k+2} - x_{k+1}) \cdot (x_{k+1} - x_k) = \alpha_{k+1}\alpha_k \nabla f(x_{k+1}) \cdot \nabla f(x_k).$$

Recall that $\alpha_k \geq 0$. If $\alpha_k = 0$, then the whole expression is zero and we're done. Consider the possibility that $\alpha_k > 0$. Then

$$f(x_k - \alpha_k \nabla f(x_k)) = \min_{s>0} f(x_k - s\nabla f(x_k)),$$

implying that α_k is a minimizer of the function on the right in the above. Then

$$0 = \frac{d}{ds}\bigg|_{s=\alpha_k} f(x_k - s\nabla f(x_k)) = \nabla f(x_k - \alpha_k \nabla f(x_k)) \cdot (-\nabla f(x_k)) = -\nabla f(x_{k+1}) \cdot \nabla f(x_k),$$

proving the claim. □

The fact that the steepest descent algorithm moves in perpendicular steps implies that the method may converge very slowly. Consider the example of a quadratic function $f(x) = x^T Q x$ in $\mathbb{R}^2$ for Q positive definite, and its elliptical level sets.