Newton's Method and Gradient Descent Method

As already stated, one of the basic tasks of optimization is to solve

$$\begin{cases} \min f(x) \\ x \in S \end{cases}$$

where S is a closed set in $\mathbb{R}^d$ and $f\colon \mathbb{R}^d \to \mathbb{R}$ a nice function. There are two general algorithms we can used:

(1) Since the minimum will most likely be also a local minimum and thus a critical point, we can look for the points where $\nabla f = 0$. This can be done by way of **Newton's method**.
(2) We can address the minimization directly by "crawling" S by a trajectory of points at which f is gradually smaller. This is the basis of **Gradient descent method**.

We will now describe these methods is some detail.

Newton's method

Algorithm: This is a method for finding roots of a function f, i.e., points $\hat{x}$ where $f(\hat{x}) = 0$. The best way to describe its algorithm is as follows: We pick a point x_1, find the tangent line to $y = f(x)$ at $x = x_1$, look for the intersection of the tangent with the x axis and call this point x_2. Then we repeat this starting from x_2 instead of x_1 and so on.

This result in a sequence of points $\{x_n\}$ such that x_{n+1} is the intersection of the tangent line

$$y = f(x_n) + f'(x_n)(x - x_n)$$

with x-axis $y = 0$. A bit of algebra gives

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

which is well defined whenever $f'(x_n) \neq 0$. As we assume to start close to a root $\hat{x}$, we may guarantee this by requiring that $f'(\hat{x}) \neq 0$ and assuming that f' is continuous (and thus non-zero even in a neighborhood of $\hat{x}$).

Convergence rate: It is worthwhile to study the convergence rate of the method. For this we subtract $\hat{x}$ on both sides of the above equation to get

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

and so, assuming that $|f'(x_n)| \geq c_1 > 0$, we get

$$|x_{n+1} - \hat{x}| \leq \frac{1}{c_1}\big|f(x_n) + (\hat{x} - x_n)f'(x_n)\big|$$

Taylor's theorem with remainder gives us

$$f(\hat{x}) - \left[f(x_n) + (\hat{x} - x_n)f'(x_n)\right] = \int_{x_n}^{\hat{x}} (x - x_n) f''(s) \mathrm{d}s$$

Assuming that $|f''(s)| \le c_2$ for s between $\hat{x}$ and x_n, and using that $f(\hat{x}) = 0$, we get

$$\left|f(x_n) + (\hat{x} - x_n)f'(x_n)\right| \le c_2 \left|\int_{x_n}^{\hat{x}} (s - x_n) \mathrm{d}s\right| = \frac{c_2}{2} |x_n - \hat{x}|^2$$

Using this in the above equation we obtain

$$\boxed{|x_{n+1} - \hat{x}| \le C|x_n - \hat{x}|^2 \quad \text{where} \quad C := \frac{c_2}{2c_1}}$$

This is referred as **quadratic** convergence, although as we will show next, the decay of $|x_n - \hat{x}|$ is in fact **doubly exponential**.

Decay estimate: In analyzing the consequence of this **recursive bound**, note that

$$C|x_1 - \hat{x}| < 1 \quad \Rightarrow \quad |x_{n+1} - \hat{x}| < |x_n - \hat{x}| \text{ and so } C|x_n - \hat{x}| < 1$$

So starting the iterations anywhere in the set $\{x\colon |x - \hat{x}| < 1/C\}$, the sequence will never leave this set. Now if $|x_1 - \hat{x}| < 1/C$, then there is $\kappa > 0$ such that $|x_1 - \hat{x}| = \frac{1}{C}\mathrm{e}^{-2\kappa}$. By induction we then show

$$|x_1 - \hat{x}| = \frac{1}{C}\mathrm{e}^{-2\kappa} \quad \Rightarrow \quad |x_n - \hat{x}| \le \frac{1}{C}\mathrm{e}^{-2^n\kappa}$$

This is an extremely fast approach to the limit. Indeed, if, for instance $C = 1$ and $\kappa = 1/2$, then $|x_2 - \hat{x}| \le 1/\mathrm{e}^{-2} \approx 0.1$, $|x_4 - \hat{x}| \le 1/\mathrm{e}^{-8} \approx 0.0001$ and $|x_6 - \hat{x}| \le 1/\mathrm{e}^{-32} \approx 10^{-15}$, etc.'

Failure if started too far: The method understandably fails if x_1 is too far from the root $\hat{x}$. An example for this is $f(x) = \tanh(x)$ which has a unique root at $x = 0$. Since $f'(x) = \cosh(x)^{-2}$, the iterations then correspond to

$$x_{n+1} = x_n - \frac{1}{2}\sinh(2x_n).$$

As $\frac{1}{2}\sinh(2x_n) \ge x_n$ for $x_n > 0$ (and opposite inequality holds for $x_n < 0$) the signs of x_n alternate. For the sequence of absolute values we then get

$$|x_{n+1}| = \frac{1}{2}\sinh(2|x_n|) - |x_n|$$

Letting $h(a) = \frac{1}{2}\sinh(2a) - a$, we have $|x_{n+1}| = h(|x_n|)$. The equation $h(a) = a$ has two non-negative solutions: $a = 0$ and $a = a^\star > 0$ such that $\sinh(2a^\star) = 4a^\star$. The graphical analysis of the trajectory shows that

$$|x_1| < a^\star \quad \Rightarrow \quad x_n \to 0 \text{ (the method does find the root)}$$
$$|x_1| > a^\star \quad \Rightarrow \quad |x_n| \to \infty \text{ (the method fails)}$$

Multivariate version: We have so far only addressed one function of one variable. The application (finding critical points in unconstraint optimization) with require finding a **common root** of d-functions of d variables,

$$f_i(\hat{x}) = 0, \quad i = 1, \dots, d$$

(These functions will themselves be components of the gradient of the function we are minimizing.) Starting from a point $x_n \in \mathbb{R}^d$, we thus get d tangent lines,

$$y_i = f_i(x_n) - (x - x_n) \cdot \nabla f_i(x_n)$$

that intersect the set where $y_i = 0$ for all $i = 1, \ldots, d$ at the point $x_{n+1} \in \mathbb{R}^d$ with equations

$$0 = f_i(x_n) - (x_{n+1} - x_n) \cdot \nabla f_i(x_n), \quad i = 1, \ldots, d$$

or

$$(x_{n+1} - x_n) \cdot \nabla f_i(x_n) = -f_i(x_n), \quad i = 1, \ldots, d.$$

We can write these as a matrix equation: Let $\nabla \vec{f}(x)$ be the matrix with i-th column being the vector $\nabla f_i(x)$ and $\vec{f}(x)$ being the row vector with i-th entry being $f_i(x)$ then

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

Multiplying on the left by the inverse of $\nabla \vec{f}(x_n)$, we get

$$(x_{n+1} - x_n)^{\mathrm{T}} = -\vec{f}(x_n)^{\mathrm{T}} \left[\nabla \vec{f}(x_n)\right]^{-1}$$

If we prefer to write this using transposed vectors, this reads

$$\boxed{x_{n+1} = x_n - \left[\nabla \vec{f}(x_n)\right]^{-\mathrm{T}} \vec{f}(x_n)}$$

This is the recursion corresponding to multivariate Newthon's method. **Word of advice:** Derive this in each problem again to make sure all transposes are done right.

Gradient Descent Method

Algorithm: The goal here is to address directly the process of minimizing function f. We will only discuss the **unconstrained** version, where all directions are feasible. As $-\nabla f(x)$ is the direction of **steepest descent** of f at x, we set

$$\boxed{x_{n+1} = x_n - h \nabla f(x_n)}$$

for some $h > 0$.

Picking the step length: The step length h was chosen to be independent of n, although one can play with other choices as well. The question is how to select h in order to make the best gain of the method. Let us discuss this first in the case of a function of a single variable. Then

$$x_{n+1} = x_n - h f'(x_n)$$

and so $f(x_{n+1}) = f(x_n - h f'(x_n))$. To turn the right-hand side into a more manageable form, we gain invoke Taylor's theorem:

$$f(x+t) = f(x) + t f'(x) + \int_x^{x+t} (s - x) f''(s) \mathrm{d}s$$

Assuming that $f''(s) \le L$, this gives us

$$f(x+t) \le f(x) + t f'(x) + \frac{t^2}{2} L$$

Using this for $x = x_n$ and $t = -hf'(x_n)$, we thus get

$$\begin{aligned} f(x_{n+1}) &= f\big(x_n - hf'(x_n)\big) \\ &\le f(x_n) - hf'(x_n) f'(x_n) + \frac{1}{2} L \big[hf'(x_n)\big]^2 \\ &= f(x_n) - [f'(x_n)]^2 \Big(h - \frac{L}{2} h^2\Big). \end{aligned}$$

The gain from the method will be best if $h - \frac{L}{2}h^2$ is maximal. This happens at the point

$$\boxed{h = \frac{1}{L}}$$

In the situation when f is a function of many variables, the same derivation applies except that $f'(x_n)$ has to be replaced by $\nabla f(x_n)$ and L by

$$\boxed{L := \sup_{x:\ f(x) \le f(x_1)} \ \sup_{v \in \mathbb{R}^d \setminus \{0\}} \frac{v^{\mathrm{T}} \,\mathrm{Hess}_f(x)\, v}{|v|^2}}$$

where $|v|$ is the Euclidean length of vector v. The supremum over v achieved by the largest eigenvalue of the Hessian. The supremum over x can be reduced to the set where $f(x) \le f(x_1)$ because the function f decreases along each linear segment connecting x_n to x_{n+1}.

Remarks on convergence: It is obvious that the method defines a sequence of points $\{x_n\}$ along which $f(x_n)$ decreases. If f is bounded from below and the level sets of f are bounded, $f(x_n)$ converges. But the above derivation (we use for convenience only the one-dimensional version) shows

$$f(x_{n+1}) \le f(x_n) - \frac{1}{2L} \big[f'(x_n)\big]^2$$

or, after some algebra,

$$\big[f'(x_n)\big]^2 \le 2L\big[f(x_n) - f(x_{n+1})\big]$$

Since $f(x_n) - f(x_{n+1}) \to 0$, also $f'(x_n) \to 0$. Now if the level sets of f are also bounded, $\{x_n\}$ contains a subsequence that converges to a point $\hat{x}$. By continuity of f', we then have $f'(\hat{x}) = 0$, i.e., $\hat{x}$ is a critical point. The method thus generally finds a critical point but that could still be a local minimum or a saddle point. Which it is cannot be decided at this level of generality.