15 Newton's method

So far we studied optimization algorithms that rely on first-order information of the objective function only, i.e., gradients. Today we look at Newton's method which relies on second derivatives. Consider the unconstrained minimization problem:

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

where f is a convex function that is twice differentiable. The Hessian of f at x is the $n \times n$ symmetric matrix defined by:

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right]_{1\leq i,j\leq n}.$$

Since f is convex we know that $\nabla^2 f(x)$ is positive semidefinite for all x. We further assume that it is nonsingular for all x. Newton's method to solve (1) is described by the following iteration rule:

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

There are different ways to interpret the rule (2).

- *Sequential quadratic approximation:* Consider the second-order approximation of f around the current point x_k:

  $$f(x_k + h) \approx f(x_k) + \nabla f(x_k)^T h + \frac{1}{2} h^T \nabla^2 f(x_k) h.$$

  Minimizing the right-hand side gives $h = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$. This means that x_{k+1} minimizes the second-order approximation of f at x_k. In particular it implies that if f is a quadratic function, then Newton's method converges in one step.

- *Gradient method in the local metric:* For any point $x \in \mathbb{R}^n$, since $\nabla^2 f(x) \succ 0$ (positive definite) we can consider the metric (i.e., inner product)

  $$\langle u, v\rangle_x = u^T \nabla^2 f(x) v$$

  for any $u, v \in \mathbb{R}^n$. The *gradient* of f at x, *in the local metric* $\langle\cdot,\cdot\rangle_x$ is the vector g such that the following identity is true:

  $$f(x + h) = f(x) + \langle g, h\rangle_x + o(h).$$

  It is not hard to verify that $g = \nabla^2 f(x)^{-1}\nabla f(x)$. This means that the Newton iteration (2) is nothing but a gradient descent method (with unit step size), where the gradient is taken with respect to the local metric $\langle\cdot,\cdot\rangle_x$.

- *Newton-Raphson method for solving $\nabla f(x) = 0$:* The Newton-Raphson for solving a system of nonlinear equations $F(x) = 0$ were $F : \mathbb{R}^n \to \mathbb{R}^n$ is defined by the iteration rule: $x_{k+1} = x_k - JF(x_k)^{-1}F(x_k)$, where $JF(x_k)$ is the Jacobian of F at x_k. This iteration rule is based on a successive linearization of F. It is not hard to see that Newton's method (2) is a Newton-Raphson method for the system of nonlinear equations $\nabla f(x) = 0$.

Remarks:

- Invariance under linear transformation: Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix and consider the change of coordinates $y = Ax$. Let $\tilde{f}(y) = f(A^{-1}y)$ and consider Newton's method applied to $\tilde{f}$. Let $(x_k)_{k\geq 0}$ be the iterates of Newton's method applied to f, and let $(y_k)_{k\geq 0}$ be the iterates of Newton's method applied to $\tilde{f}$ starting from $y_0 = Ax_0$. Using the fact that $\nabla\tilde{f}(y) = A^{-T}\nabla f(A^{-1}y)$ and $\nabla^2\tilde{f}(y) = A^{-T}\nabla f(A^{-1}y)A^{-1}$, it is not hard to check that we have, for all $k \geq 1$, $y_k = Ax_k$. In other words, the behavior of Newton's method does not depend on the choice of coordinates made.

  This independence property is not true for other algorithms we saw before, e.g., the gradient method. Recall that if we apply the gradient method to $f(x) = (x_1^2 + x_2^2)/2$ with exact line search we converge to $x^* = (0, 0)$ in one iteration from any starting point; whereas if we apply it to $\tilde{f}(y) = (y_1^2 + \gamma y_2^2)/2$ with $\gamma \neq 1$ this is not the case.

- Iteration complexity of Newton's method: Each iteration of Newton's method requires inverting the Hessian $\nabla^2 f(x)$ (or solving $\nabla^2 f(x)h = \nabla f(x)$). Solving a generic linear system of size $n \times n$ requires $\approx n^3$ floating point operations. This can be prohibitive for very large n. Note that the iteration complexity of the gradient method is simply $O(n)$.

One characteristic of Newton's method is that it is very fast if you start close enough to the optimal solution. This is the object of the next theorem. Note that $\|\nabla f(x)\|_2$ is used as a measure proximity to the optimal solution.

Theorem 15.1 (Quadratic convergence of Newton's method)**.** *Let f be a m-strongly convex function that is C^2 such that $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq M\|x - y\|_2$.*[1] *Given $x \in \mathbb{R}^n$ let $x^+ = x - (\nabla^2 f(x))^{-1}\nabla f(x)$. Then*

$$\frac{M}{2m^2}\|\nabla f(x^+)\|_2 \leq \left(\frac{M}{2m^2}\|\nabla f(x)\|_2\right)^2.$$

Proof. Let $h = -\nabla^2 f(x)^{-1}\nabla f(x)$ be the Newton step. Then we can write

$$\begin{aligned}\nabla f(x^+) &= \nabla f(x) + \int_0^1 \nabla^2 f(x + th)hdt \\ &= \int_0^1 (\nabla^2 f(x + th) - \nabla^2 f(x))hdt\end{aligned}$$

where we used the fact that $\nabla f(x) = -\nabla^2 f(x)h$. Thus

$$\begin{aligned}\|\nabla f(x^+)\|_2 &\leq \int_0^1 \|(\nabla^2 f(x + th) - \nabla^2 f(x))h\|_2 dt \\ &\leq \int_0^1 Mt\|h\|_2^2 dt = \frac{M}{2}\|h\|_2^2 = \frac{M}{2}\|\nabla^2 f(x)^{-1}\nabla f(x)\|_2^2 \leq \frac{M}{2m^2}\|\nabla f(x)\|_2^2\end{aligned}$$

where in the last inequality we used $\nabla^2 f(x) \succeq mI$. □

Let $r_k = \frac{M}{2m^2}\|\nabla f(x_k)\|_2$. The theorem above tells us that $r_{k+1} \leq r_k^2$. This implies that $r_k \leq r_0^{2^k}$. If $r_0 < 1$ then $r_k \to 0$ at a very fast rate. For example if $r_0 = 1/10$, we get:

$$r_0 = 0.1,\ r_1 = 0.01,\ r_2 = 0.0001,\ r_3 = 0.00000001,\ ...$$

This is called quadratic convergence.

[1] For a matrix H we let $\|H\|_2 = \sup_{\|x\|_2=1}\|Hx\|_2$ be the operator norm of H

Remark 1. *Using strong convexity one can get a bound on $f(x_k) - f^*$ in terms of $\|\nabla f(x_k)\|_2$. Indeed, one can prove that*

$$f(x_k) - f^* \leq \frac{1}{2m}\|\nabla f(x_k)\|_2^2. \tag{3}$$

(Write $f(x_k) - f^* = f(x_k) - \min f \leq f(x_k) - \min_h \left\{ f(x_k) + \nabla f(x_k)^T h + \frac{m}{2}\|h\|_2^2 \right\} = \frac{1}{2m}\|\nabla f(x_k)\|_2^2$.) This shows that if $\|\nabla f(x_k)\|_2$ converges to 0 at a quadratic convergence rate, then so does $f(x_k) - f^*$.

It can be shown that, outside the region of quadratic convergence, Newton's method with a non-unit step size achieves steady progress, i.e.,

$$f(x_{k+1}) - f(x_k) \leq -\gamma$$

for some positive $\gamma > 0$, assuming that $\|\nabla f(x_k)\|_2 \geq \beta$. See Exercise sheet 3.