11.3 Newton's method revisited

Recall Newton's method. Start w/ first-order nec. cond. and TE:

$$0 = \nabla f(x^* + h) = \nabla f(x^*) + \nabla^2 f(x^*) h + O(\|h\|^2).$$

As $h \to 0$, approximate:

$$h \approx -\nabla^2 f(x^*)^{-1} \nabla f(x^*).$$

Hence, reasonable to set $p_n = -\nabla f(x_n)^{-1} \nabla f(x_n)$ in the iteration:

$$x_{n+1} = x_n + p_n \qquad (*)$$

to get:

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

Another way of thinking of Newton's method: Consider $(*)$.
Approximate f quadratically about x_n:

$$g_n(p) = f(x_n) + \nabla f(x_n)^T p + \frac{1}{2} p^T \nabla^2 f(x_n) p.$$

Then, by a TE:

$$f(x_n + p) = g_n(p) + O(\|p\|^3).$$

To find p, solve the quadratic program:

$$\underset{p}{\text{minimize}} \; g_n(p).$$

How to solve?

Check F.O.N.C.s:

$$\nabla_p \tfrac{1}{2} p^T A p = \tfrac{1}{2}(A + A^T) p$$

$$\nabla_p q_n(p) = \nabla f(x_n) + \nabla^2 f(x_n) p.$$

(gradient w.r.t. p — not x!)

Hence, same result:

$$p_n := p^* = -\nabla^2 f(x_n)^{-1} \nabla f(x_n).$$

In general. Newton's method may not converge. When it does, under certain circumstances, it converges quadratically. What does this mean?

Def: Let $\{x_k\}_{k=0}^{\infty} \subseteq \mathbb{R}^n$ be a sequence which converges to $x^* \in \mathbb{R}^n$. Then, $\{x_n\}$ converges w/ order of convergence $q \geq 1$ and rate of convergence μ if:

$$\lim_{k \to \infty} \frac{\| x_{k+1} - x^* \|}{\| x_k - x^* \|^q} = \mu < \infty.$$

Note: if $q = 1$, $\{x_n\}$ converges linearly. ← common
" $q = 2$, " " quadratically, ← uncommon
" $q = 3$, " " cubically. ← rare

There are some technical variations on this definition. We generally don't care too much about μ, usually only interested in q and its value.

Now a theorem:

Theorem: (Quadratic convergence of Newton's method.)

Let $S \subseteq \mathbb{R}^n$ be an open convex subset of S, containing x^*. Let $f : \mathbb{R}^n \to \mathbb{R}$ be $C^2(S)$ and assume that $\nabla^2 f$ is Lipschitz continuous with constant $L < \infty$ on S, i.e.:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$$

for all $x, y \in S$. Assume also that $\nabla^2 f$ is positive definite on S. Then, if $\|x_0 - x^*\|$ is small enough, Newton's method converges quadratically.

For a variety of reasons, we will often use an approximation of the Newton step at each iteration. More about this later.

Theorem 11.3 in Griva tells us that if $p_n \to -\nabla^2 f(x_n)^{-1} \nabla f(x_n)$ as $n \to \infty$, then $x_{n+1} = x_n + p_n$ converges to x^* superlinearly ($q = 1$ and $\mu = 0$). This is useful because frequently our approximations of Newton's method will be based on approximating $\nabla^2 f(x_n)$ with some consistent approximation. For example, quasi-Newton methods work this way.

The goal is to make an approximation which maintains superlinear convergence but is cheaper than Newton's method. This can often result in an algorithm which not only uses less memory than Newton's method, but runs faster (maybe more iterations, but each iteration faster to perform).

One reason to modify the Newton iteration is to ensure descent. We would like to modify the Newton step to ensure that it is a descent direction:

$$\nabla f(x_n)^T p_n < 0.$$

Observe that this is equivalent to:

$$\nabla f(x_n)^T \nabla^2 f(x_n)^{-1} \nabla f(x_n) > 0$$

since $p_n = -\nabla^2 f(x_n)^{-1} \nabla f(x_n)$ for the Newton iteration.

A <u>sufficient</u> condition for this to hold is $\nabla^2 f(x_n)$ to be positive definite. Why? Well, if $\nabla^2 f(x_n)$ is positive definite, since $\nabla^2 f(x_n)$ is symmetric, by the spectral theorem (see HW), we can write the orthogonal eigenvalue decomposition of $\nabla^2 f(x_n)$:

$$\nabla^2 f(x_n) = Q \Lambda Q^T, \quad Q = \underbrace{[q_1 \cdots q_n]}_{\text{eigenvectors}}, \quad \Lambda = \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}}_{\text{eigenvalues}}.$$

Then:

$$\nabla^2 f(x_n)^{-1} = (Q \Lambda Q^T)^{-1} = Q^{-T} \Lambda^{-1} Q^{-1} = Q \Lambda^{-1} Q^T.$$

But this is an orthogonal eigenvalue decomposition of $\nabla^2 f(x_n)^{-1}$! Hence, the eigenvalues of $\nabla^2 f(x_n)^{-1}$ are $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$, which are positive since $\lambda_i > 0$ for $i = 1, \ldots, n$.

What is happening if $\nabla^2 f(x_n)$ isn't positive definite? Well, recall that if $\nabla^2 f(x^*)$ is positive definite, then x^* is a local minimum. Indeed, the quadratic form $x \mapsto x^T A x$ has a single, unique global minimum if A is positive definite. Consider a positive semidefinite matrix $A \in \mathbb{R}^{2\times 2}$ with a zero eigenvalue. Let λ be the nonzero eigenvalue with u the corresponding normalized eigenvector. Then we can write:

$$A = \lambda u u^T = \lambda \begin{bmatrix} u_1^2 & u_1 u_2 \\ u_2 u_1 & u_2^2 \end{bmatrix}.$$

Let $v \neq 0$ such that $u^T v$. Then:

$$v^T A v = \lambda v^T u u^T v = \lambda (u^T v)^2 = 0.$$

And if $w = tu$, then:

$$w^T A w = \lambda t^2 u^T u u^T u = \lambda^2 t^2 (u^T u)^2 = \lambda^2 t^2.$$

So the quadratic form $x \mapsto x^T A x$ has quadratic variation in the u direction, and is zero in the v direction: