Newton's Method for Unconstrained Optimization

Robert M. Freund

February, 2004

1 Newton's Method

Suppose we want to solve:

$$\text{(P:)} \qquad \min f(x)$$

$$x \in \Re^n.$$

At $x = \bar{x}$, $f(x)$ can be approximated by:

$$f(x) \approx h(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^t H(\bar{x})(x - \bar{x}),$$

which is the quadratic Taylor expansion of $f(x)$ at $x = \bar{x}$. Here $\nabla f(x)$ is the gradient of $f(x)$ and $H(x)$ is the Hessian of $f(x)$.

Notice that $h(x)$ is a quadratic function, which is minimized by solving $\nabla h(x) = 0$. Since the gradient of $h(x)$ is:

$$\nabla h(x) = \nabla f(\bar{x}) + H(\bar{x})(x - \bar{x}) \ ,$$

we therefore are motivated to solve:

$$\nabla f(\bar{x}) + H(\bar{x})(x - \bar{x}) = 0 \ ,$$

which yields

$$x - \bar{x} = -H(\bar{x})^{-1} \nabla f(\bar{x}).$$

The direction $-H(\bar{x})^{-1} \nabla f(\bar{x})$ is called the *Newton direction*, or the *Newton step* at $x = \bar{x}$.

This leads to the following algorithm for solving (P):

Newton's Method:

Step 0 Given x^0, set $k \leftarrow 0$

Step 1 $d^k = -H(x^k)^{-1} \nabla f(x^k)$. If $d^k = 0$, then stop.

Step 2 Choose step-size $\alpha^k = 1$.

Step 3 Set $x^{k+1} \leftarrow x^k + \alpha^k d^k,\ k \leftarrow k+1$. Go to **Step 1**.

Note the following:

- The method assumes $H(x^k)$ is nonsingular at each iteration.
- There is no guarantee that $f(x^{k+1}) \leq f(x^k)$.
- Step 2 could be augmented by a line-search of $f(x^k + \alpha d^k)$ to find an optimal value of the step-size parameter α.

Recall that we call a matrix SPD if it is symmetric and positive definite.

Proposition 1.1 *If $H(x)$ is SPD and $d := -H(x)^{-1}\nabla f(x) \neq 0$, then d is a descent direction, i.e., $f(x + \alpha d) < f(x)$ for all sufficiently small values of α.*

Proof: It is sufficient to show that $\nabla f(x)^t d = -\nabla f(x)^t H(x)^{-1} \nabla f(x) < 0$. This will clearly be the case if $H(x)^{-1}$ is SPD. Since $H(x)$ is SPD, if $v \neq 0$,

$$0 < (H(x)^{-1}v)^t H(x)(H(x)^{-1}v) = v^t H(x)^{-1} v,$$

thereby showing that $H(x)^{-1}$ is SPD. ∎

1.1 Rates of convergence

A sequence of numbers $\{s_i\}$ exhibits *linear* convergence if $\lim_{i\to\infty} s_i = \bar{s}$ and

$$\lim_{i\to\infty} \frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|} = \delta < 1.$$

If $\delta = 0$ in the above expression, the sequence exhibits *superlinear* convergence.

A sequence of numbers $\{s_i\}$ exhibits *quadratic* convergence if $\lim_{i\to\infty} s_i = \bar{s}$ and

$$\lim_{i\to\infty} \frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|^2} = \delta < \infty.$$

1.1.1 Examples of Rates of Convergence

Linear convergence: $s_i = \left(\frac{1}{10}\right)^i$: 0.1, 0.01, 0.001, etc. $\bar{s} = 0$.

$$\frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|} = 0.1.$$

Superlinear convergence: $s_i = \frac{1}{i!}$: $1, \frac{1}{2}, \frac{1}{6}, \frac{1}{24}, \frac{1}{125}$, etc. $\bar{s} = 0$.

$$\frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|} = \frac{i!}{(i+1)!} = \frac{1}{i+1} \to 0 \text{ as } i \to \infty.$$

Quadratic convergence: $s_i = \left(\frac{1}{10}\right)^{(2^i)}$: 0.1, 0.01, 0.0001, 0.00000001, etc. $\bar{s} = 0$.

$$\frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|^2} = \frac{(10^{2^i})^2}{10^{2^{i+1}}} = 1.$$

1.2 Quadratic Convergence of Newton's Method

We have the following quadratic convergence theorem. In the theorem, we use the operator norm of a matrix M:

$$\|M\| := \max_x \{\|Mx\| \mid \|x\| = 1\} \ .$$

Theorem 1.1 (Quadratic Convergence Theorem) *Suppose $f(x)$ is twice continuously differentiable and x^* is a point for which $\nabla f(x^*) = 0$. Suppose that $H(x)$ satisfies the following conditions:*

- *there exists a scalar $h > 0$ for which $\|[H(x^*)]^{-1}\| \leq \frac{1}{h}$*
- *there exists scalars $\beta > 0$ and $L > 0$ for which $\|H(x) - H(x^*)\| \leq L\|x - x^*\|$ for all x satisfying $\|x - x^*\| \leq \beta$*

Let x satisfy $\|x - x^*\| < \gamma := \min\left\{\beta, \frac{2h}{3L}\right\}$, and let $x_N := x - H(x)^{-1}\nabla f(x)$. Then:

(i) $\|x_N - x^*\| \leq \|x - x^*\|^2 \left(\frac{L}{2(h - L\|x - x^*\|)}\right)$

(ii) $\|x_N - x^*\| < \|x - x^*\| < \gamma$

(iii) $\|x_N - x^*\| \le \|x - x^*\|^2 \left(\frac{3L}{2h}\right)$

■

Example 1: Let $f(x) = 7x - \ln(x)$. Then $\nabla f(x) = f'(x) = 7 - \frac{1}{x}$ and $H(x) = f''(x) = \frac{1}{x^2}$. It is not hard to check that $x^* = \frac{1}{7} = 0.142857143$ is the unique global minimum. The Newton direction at x is

$$d = -H(x)^{-1}\nabla f(x) = -\frac{f'(x)}{f''(x)} = -x^2\left(7 - \frac{1}{x}\right) = x - 7x^2.$$

Newton's method will generate the sequence of iterates $\{x^k\}$ satisfying:

$$x^{k+1} = x^k + (x^k - 7(x^k)^2) = 2x^k - 7(x^k)^2 \ .$$

Below are some examples of the sequences generated by this method for different starting points.

k	x^k	x^k	x^k	x^k
0	1.0	0	0.1	0.01
1	-5.0	0	0.13	0.0193
2	-185.0	0	0.1417	0.03599257
3	$-239,945.0$	0	0.14284777	0.062916884
4	-4.0302×10^{11}	0	0.142857142	0.098124028
5	-1.1370×10^{24}	0	0.142857143	0.128849782
6	-9.0486×10^{48}	0	0.142857143	0.1414837
7	-5.7314×10^{98}	0	0.142857143	0.142843938
8	$-\infty$	0	0.142857143	0.142857142
9	$-\infty$	0	0.142857143	0.142857143
10	$-\infty$	0	0.142857143	0.142857143

By the way, the "range of quadratic convergence" for Newton's method for this function happens to be

$$x \in (0.0 \ , \ 0.2857143) \ .$$

Example 2: $f(x) = -\ln(1 - x_1 - x_2) - \ln x_1 - \ln x_2$.

$$\nabla f(x) = \begin{bmatrix} \frac{1}{1-x_1-x_2} - \frac{1}{x_1} \\ \\ \frac{1}{1-x_1-x_2} - \frac{1}{x_2} \end{bmatrix},$$

$$H(x) = \begin{bmatrix} \left(\frac{1}{1-x_1-x_2}\right)^2 + \left(\frac{1}{x_1}\right)^2 & \left(\frac{1}{1-x_1-x_2}\right)^2 \\ \\ \left(\frac{1}{1-x_1-x_2}\right)^2 & \left(\frac{1}{1-x_1-x_2}\right)^2 + \left(\frac{1}{x_2}\right)^2 \end{bmatrix}.$$

$x^* = \left(\frac{1}{3}, \frac{1}{3}\right)$, $f(x^*) = 3.295836866$.

k	x_1^k	x_2^k	$\|x^k - x^*\|$
0	0.85	0.05	0.58925565098879
1	0.717006802721088	0.0965986394557823	0.450831061926011
2	0.512975199133209	0.176479706723556	0.238483249157462
3	0.352478577567272	0.273248784105084	0.0630610294297446
4	0.338449016006352	0.32623807005996	0.00874716926379655
5	0.333337722134802	0.333259330511655	$7.41328482837195e^{-5}$
6	0.333333343617612	0.33333332724128	$1.19532211855443e^{-8}$
7	0.333333333333333	0.333333333333333	$1.57009245868378e^{-16}$

Comments:

- The convergence rate is quadratic:

$$\frac{\|x_N - x^*\|}{\|x - x^*\|^2} \leq \frac{3L}{2h}$$

- We typically never know β, h, or L. However, there are some amazing exceptions, for example $f(x) = -\sum_{j=1}^{n} \ln(x_j)$, as we will soon see.
- The constants β, h, and L depend on the choice of norm used, but the method does not. This is a drawback in the concept. But we do not know β, h, or L anyway.
- In the limit we obtain $\frac{\|x_N - x^*\|}{\|x - x^*\|^2} \leq \frac{L}{2h}$

- We did not assume convexity, only that $H(x^*)$ is nonsingular and not badly behaved near x^*.
- One can view Newton's method as trying successively to solve

  $$\nabla f(x) = 0$$

  by successive linear approximations.
- Note from the statement of the convergence theorem that the iterates of Newton's method are equally attracted to local minima and local maxima. Indeed, the method is just trying to solve $\nabla f(x) = 0$.
- What if $H(x^k)$ becomes increasingly singular (or not positive definite)? In this case, one way to "fix" this is to use

  $$H(x^k) + \epsilon I \ . \tag{1}$$

- Comparison with steepest-descent. One can think of steepest-descent as $\epsilon \rightarrow +\infty$ in (1) above.
- The work per iteration of Newton's method is $O(n^3)$
- So-called "quasi-Newton methods" use approximations of $H(x^k)$ at each iteration in an attempt to do less work per iteration.

2 Proof of Theorem 1.1

The proof relies on the following two "elementary" facts. For the first fact, let $\|v\|$ denote the usual Euclidian norm of a vector, namely $\|v\| := \sqrt{v^T v}$. The operator norm of a matrix M is defined as follows:

$$\|M\| := \max_x \{\|Mx\| \mid \|x\| = 1\} \ .$$

Proposition 2.1 *Suppose that M is a symmetric matrix. Then the following are equivalent:*

1. *$h > 0$ satisfies $\|M^{-1}\| \leq \frac{1}{h}$*
2. *$h > 0$ satisfies $\|Mv\| \geq h \cdot \|v\|$ for any vector v*

■

You are asked to prove this as part of your homework for the class.

Proposition 2.2 *Suppose that $f(x)$ is twice differentiable. Then*

$$\nabla f(z) - \nabla f(x) = \int_0^1 \left[H(x + t(z-x)) \right] (z-x) dt \ .$$

Proof: Let $\phi(t) := \nabla f(x+t(z-x))$. Then $\phi(0) = \nabla f(x)$ and $\phi(1) = \nabla f(z)$, and $\phi^{'}(t) = \left[H(x + t(z-x)) \right] (z-x)$. From the fundamental theorem of calculus, we have:

$$\begin{aligned} \nabla f(z) - \nabla f(x) &= \phi(1) - \phi(0) \\ &= \int_0^1 \phi^{'}(t) dt \\ &= \int_0^1 \left[H(x + t(z-x)) \right] (z-x) dt \ . \ \blacksquare \end{aligned}$$

Proof of Theorem 1.1: We have:

$$\begin{aligned} x_N - x^* &= x - H(x)^{-1} \nabla f(x) - x^* \\ &= x - x^* + H(x)^{-1} \left(\nabla f(x^*) - \nabla f(x) \right) \\ &= x - x^* + H(x)^{-1} \int_0^1 \left[H(x + t(x^* - x)) \right] (x^* - x) dt \quad \text{(from Proposition 2.2)} \\ &= H(x)^{-1} \int_0^1 \left[H(x + t(x^* - x)) - H(x) \right] (x^* - x) dt \end{aligned}$$

Therefore

$$\begin{aligned}
\|x_N - x^*\| &\leq \|H(x)^{-1}\| \int_0^1 \| \left[H(x + t(x^* - x)) - H(x) \right] \| \|(x^* - x)\| dt \\
&\leq \|x^* - x\| \|H(x)^{-1}\| \int_0^1 L \cdot t \cdot \|(x^* - x)\| dt \\
&= \|x^* - x\|^2 \|H(x)^{-1}\| L \int_0^1 t dt \\
&= \frac{\|x^* - x\|^2 \|H(x)^{-1}\| L}{2}
\end{aligned}$$

We now bound $\|H(x)^{-1}\|$. Let v be any vector. Then

$$\begin{aligned}
\|H(x)v\| &= \|H(x^*)v + (H(x) - H(x^*))v\| \\
&\geq \|H(x^*)v\| - \|(H(x) - H(x^*))v\| \\
&\geq h \cdot \|v\| - \|H(x) - H(x^*)\| \|v\| \qquad \text{(from Proposition 2.1)} \\
&\geq h \cdot \|v\| - L\|x^* - x\| \cdot \|v\| \\
&= (h - L\|x^* - x\|) \cdot \|v\| \ .
\end{aligned}$$

Invoking Proposition 2.1 again, we see that this implies that

$$\|H(x)^{-1}\| \leq \frac{1}{h - L\|x^* - x\|} \ .$$

Combining this with the above yields

$$\|x_N - x^*\| \leq \|x^* - x\|^2 \frac{L}{2\left(h - L\|x^* - x\|\right)} \ ,$$

which is *(i)* of the theorem. Because $L\|x^* - x\| < \frac{2h}{3}$ we have:

$$\|x_N - x^*\| \leq \|x^* - x\| \frac{L\|x^* - x\|}{2\left(h - L\|x^* - x\|\right)} < \frac{\frac{2h}{3}}{2\left(h - \frac{2h}{3}\right)} \|x^* - x\| = \|x^* - x\| \ ,$$

which establishes *(ii)* of the theorem. Finally, we have

$$\|x_N - x^*\| \leq \|x^* - x\|^2 \frac{L}{2\left(h - L\|x^* - x\|\right)} \leq \|x^* - x\|^2 \frac{L}{2\left(h - \frac{2h}{3}\right)} = \|x^* - x\|^2 \frac{3L}{2h},$$

which establishes *(iii)* of the theorem. ∎

3 Newton's Method Exercises

1. (Newton's Method) Suppose we want to minimize the following function:

   $$f(x) = 9x - 4\ln(x - 7)$$

   over the domain $X = \{x \mid x > 7\}$ using Newton's method.

   (a) Give an exact formula for the Newton iterate for a given value of x.

   (b) Using a calculator (or a computer, if you wish), compute five iterations of Newton's method starting at each of the following points, and record your answers:

   - $x = 7.40$
   - $x = 7.20$
   - $x = 7.01$
   - $x = 7.80$
   - $x = 7.88$

   (c) Verify empirically that Newton's method will converge to the optimal solution for all starting values of x in the range $(7, 7.8888)$. What behavior does Newton's method exhibit outside of this range?

2. (Newton's Method) Suppose we want to minimize the following function:

   $$f(x) = 6x - 4\ln(x - 2) - 3\ln(25 - x)$$

   over the domain $X = \{x \mid 2 < x < 25\}$ using Newton's method.

   (a) Using a calculator (or a computer, if you wish), compute five iterations of Newton's method starting at each of the following points, and record your answers:

- $x = 2.60$
- $x = 2.70$
- $x = 2.40$
- $x = 2.80$
- $x = 3.00$

(b) Verify empirically that Newton's method will converge to the optimal solution for all starting values of x in the range $(2, 3.05)$. What behavior does Newton's method exhibit outside of this range?

3. (Newton's Method) Suppose that we seek to minimize the following function:

$$f(x_1, x_2) = -9x_1 - 10x_2 + \theta(-\ln(100 - x_1 - x_2) - \ln(x_1) - \ln(x_2) - \ln(50 - x_1 + x_2)),$$

where θ is a given parameter, on the domain $X = \{(x_1, x_2) \mid x_1 > 0, x_2 > 0, x_1 + x_2 < 100, x_1 - x_2 < 50\}$. This exercise asks you to implement Newton's method on this problem, first without a line-search, and then with a line-search. Run your algorithm for $\theta = 10$ and for $\theta = 100$, using the following starting points.

- $x_0 = (8, 90)^T$.
- $x_0 = (1, 40)^T$.
- $x_0 = (15, 68.69)^T$.
- $x_0 = (10, 20)^T$.

(a) When you run Newton's method without a line-search for this problem and with these starting points, what behavior do you observe?

(b) When you run Newton's method with a line-search for this problem, what behavior do you observe?

4. (Projected Newton's Method) Prove Proposition 6.1 of the notes on Projected Steepest Descent.

5. (Newton's Method) In class we described Newton's method as a method for finding a point x^* for which $\nabla f(x^*) = 0$. Now consider the following setting, where we have n nonlinear equations in n unknowns $x = (x_1, \ldots, x_n)$:

$$\begin{array}{rcl} g_1(x) & = & 0 \\ \vdots & \vdots & \vdots \\ g_n(x) & = & 0\ , \end{array}$$

which we conveniently write as

$$g(x) = 0\ .$$

Let $J(x)$ denote the Jacobian matrix (of partial derivatives) of $g(x)$. Then at $x = \bar{x}$ we have

$$g(\bar{x} + d) \approx g(\bar{x}) + J(\bar{x})d\ ,$$

this being the linear approximation of $g(x)$ at $x = \bar{x}$. Construct a version of Newton's method for solving the equation system $g(x) = 0$.

6. (Newton's Method) Suppose that $f(x)$ is a strictly convex twice-continuously differentiable function, and consider Newton's method with a line-search. Given $\bar{x}$, we compute the Newton direction $d = -[H(\bar{x})]^{-1}\nabla f(\bar{x})$ and the next iterate $\tilde{x}$ is chosen to satisfy:

   $$\tilde{x} := \arg\min_{\alpha} f(\bar{x} + \alpha d)\ .$$

   Prove that the iterates of this method converge to the unique global minimum of $f(x)$.

7. Prove Proposition 2.1 of the notes on Newton's method.

8. Bertsekas, Exercise 1.4.1, page 99.