16 Newton's method (continued)

Recall Newton's method:

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

where $t_k > 0$ is the step size.

Assume that f is m-strongly convex, and that $\nabla^2 f(x)$ is M-Lipschitz with respect to the operator norm.

Convergence of Newton's method We saw last lecture that with $t_k = 1$, the iterates satisfy

$$\frac{M}{2m^2}\|\nabla f(x_{k+1})\|_2 \leq \left(\frac{M}{2m^2}\|\nabla f(x_k)\|_2\right)^2.$$

In particular, if at some iteration i we have $\frac{M}{2m^2}\|\nabla f(x_i)\|_2 = 1-\delta < 1$ then we get $\|\nabla f(x_k)\|_2 \to 0$ at a quadratic rate, i.e., $\|\nabla f(x_k)\|_2 \leq \frac{2m^2}{M}(1-\delta)^{2^{k-i}}$.

Unfortunately Newton's method with unit step size $t_k = 1$ does not always converge. Here is an example (from [Pol]): consider a convex function $f(x)$ so that

$$f(x) = \begin{cases} (x-1)^2 & \text{if } x \leq -1 \\ (x+1)^2 & \text{if } x \geq 1 \end{cases}$$

and on $[-1,1]$ it is chosen (arbitrarily) so that overall the function is smooth and convex (see figure below).

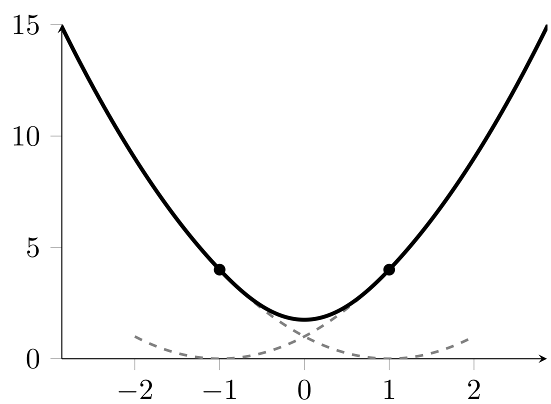

Figure 1: Example where Newton's iterations with $t_k = 1$ do not converge. If $x_0 = 1$, the sequence of iterates produced is $x_k = (-1)^k$.

We see on this function that if $x_0 = +1$, the quadratic approximation of f at x is $(x+1)^2$ whose minimum is at $x = -1$, and thus $x_1 = -1$. With the same reasoning we get $x_2 = +1$, and we see that Newton's method with unit step size oscillates between the points $+1$ and -1.

For this reason, we need to introduce a non-unit step size, at least for the first iterations of the algorithm.

We can prove the following:

Proposition 16.1. *Assume f is m-strongly and L-smooth. The Newton's method with step size $t_k = m/L$ satisfies $f(x^+) - f(x) \leq -c\|\nabla f(x)\|_2^2$ with $c = m/(2L^2)$.*

Proof. We have, since f is L-smooth (and using notation $\lambda_f(x)^2 = \left\langle \nabla f(x), \nabla^2 f(x)^{-1} \nabla f(x) \right\rangle$):

$$\begin{aligned}
f(x^+) &\leq f(x) + \left\langle \nabla f(x), x^+ - x \right\rangle + \frac{L}{2}\|x^+ - x\|_2^2 \\
&= f(x) - t\left\langle \nabla f(x), \nabla^2 f(x)^{-1} \nabla f(x) \right\rangle + \frac{L}{2} t^2 \|\nabla^2 f(x)^{-1} \nabla f(x)\|_2^2 \\
&\leq f(x) - t\lambda_f(x)^2 + \frac{L}{2m} t^2 \|\nabla^2 f(x)^{-1/2} \nabla f(x)\|_2^2 \\
&= f(x) - \left(t - \frac{L}{2m} t^2 \right) \lambda_f(x)^2
\end{aligned}$$

where in the second inequality we used that $\nabla^2 f(x)^{-1} \preceq (1/m)I$. With $t = m/L$ we thus get $f(x^+) - f(x) \leq -\frac{m}{2L}\lambda_f(x)^2 \leq -\frac{m}{2L^2}\|\nabla f(x)\|_2^2$ where the last inequality follows from $\nabla^2 f(x)^{-1} \succeq \frac{1}{L} I$. $\square$

We can now summarize the behaviour of Newton's method. Fix $\gamma = m^2/M$.

- Phase 1: $\|\nabla f(x_k)\|_2 \geq \gamma$, then by using a step size $t_k = m/L$ we get $f(x_{k+1}) - f(x_k) \leq -c\gamma^2$.
- Phase 2: $\|\nabla f(x_k)\|_2 \leq \gamma$: we have $M/(2m^2)\|\nabla f(x_k)\|_2 \leq 1/2$ and so we get quadratic convergence from this iteration onwards, i.e., $\|\nabla f(x_k)\|_2 \leq 2\gamma(1/2)^{2^{k-k_2}}$ where k_2 is the first iteration of phase 2.

Note that the number of iterations of phase 1 is at most $\frac{f(x_0)-f^*}{c\gamma^2}$. So the total number of iterations to reach $\|\nabla f(x)\|_2 \leq \epsilon$ is at most $\frac{f(x_0)-f^*}{c\gamma^2} + \log\log(2\gamma/\epsilon)$.

References

[Pol] Boris T Polyak. Introduction to optimization. 1987. *Optimization Software, Inc, New York.* 1