Convergence of Newton's Method

Suppose that $f(\boldsymbol{x})$ is strongly convex,

$$m\mathbf{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq m\mathbf{I}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N,$$

and that its Hessian is Lipschitz,

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \ \leq \ L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

(The norm on the left-hand side above is the standard operator norm.) We will show that the Newton algorithm coupled with an exact line search[1] converges to precision ϵ:

$$f(\boldsymbol{x}^{(k)}) - p^\star \leq \epsilon,$$

for a number of iterations

$$k \ \geq \ C_1 \left(f(\boldsymbol{x}^{(0)}) - p^\star \right) + \log_2 \log_2(\epsilon_0/\epsilon),$$

where we can take the constants above to be $C_1 = M^2 L^2 / m^5$ and $\epsilon_0 = 2m^3/L^2$. Qualitatively, this says that Newton's method takes a constant number of iterations to converge to any reasonable precision — we can bound $\log_2 \log_2(\epsilon_0/\epsilon) \leq 6$ (say) for ridiculously small values of ϵ.

To establish this result, we break the analysis into two stages. In the first, the *damped Newton stage*, we are far from the solution (as measured by $\|\nabla f(\boldsymbol{x}^{(k)})\|_2$), but we make constant progress towards the answer. Specifically, we will show that in this stage,

$$f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) \ \leq \ 1/C_1.$$

[1] These results are easily extended to backtracking line searches; we are just using an exact line search to make the exposition easier. See [**?**, Sec. 9.5.3] for the analysis with backtracking.

It is clear, then, that the number of damped Newton steps is no greater than $C_1\left(f(\boldsymbol{x}^{(0)}) - p^\star\right)$.

We will then show that when $\|\nabla f(\boldsymbol{x}^{(k)}\|_2$ is small enough, the gap closes dramatically at every iteration. We call this the *quadratic convergence stage*, as we will be able to show that once the algorithm enters this stage at iteration ℓ, for all $k > \ell$,

$$\|\nabla f(\boldsymbol{x}^{(k)})\|_2 \;\leq\; C_2 \cdot 2^{-2^{k-\ell}},$$

where $C_2 = L/(2m^2)$ is another constant.

Damped phase: We are in this stage when

$$\|\nabla f(\boldsymbol{x}^{(k)})\|_2 \geq m^2/L.$$

We take $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + t_{\text{exact}}\boldsymbol{d}^{(k+1)}$, where

$$\boldsymbol{d}^{(k+1)} = -\nabla^2 f(\boldsymbol{x}^{(k)})^{-1}\nabla f(\boldsymbol{x}^{(k)}),$$

and t_{exact} is the result of an exact line search[2]:

$$t_{\text{exact}} = \arg\min_{0\leq t\leq 1} f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)}).$$

With the current Newton decrement denoted as

$$\lambda_k^2 = -\nabla f(\boldsymbol{x}^{(k)})^{\mathrm{T}}\boldsymbol{d}^{(k+1)} = \|\boldsymbol{d}^{(k+1)}\|_2^2,$$

we know that

$$\begin{aligned} f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)}) &\leq f(\boldsymbol{x}^{(k)}) - t\lambda_k^2 + \frac{M}{2}\|t\boldsymbol{d}^{(k+1)}\|_2^2 \\ &\leq f(\boldsymbol{x}^{(k)}) - t\lambda_k^2 + \frac{M}{2m}t^2\lambda_k^2, \end{aligned}$$

[2]For convenience, we are not letting t be larger than 1, just as in a backtracking method.

where the second step follows from the fact that the largest eigenvalue of $[\nabla^2 f(\boldsymbol{x}^{(k)})]^{-1}$ is at most $1/m$. Plugging in $t = m/M$ above yields

$$
\begin{aligned}
f(\boldsymbol{x}^{(k)} + t_{\text{exact}}\boldsymbol{d}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) &\leq -\frac{m}{M}\lambda_k^2 \\
&\leq -\frac{m}{M^2}\|\nabla f(\boldsymbol{x}^{(k)})\|_2^2 \\
&\leq -\frac{m^5}{L^2 M^2}.
\end{aligned}
$$

Quadratic convergence: When

$$
\|\nabla f(\boldsymbol{x}^{(k)})\|_2 < m^2/L,
$$

we start to settle things very quickly. We will assume that in this stage, we choose the step size to be $t = 1$. In fact, you can show that under very mild assumptions on the backtracking parameter ($\alpha < 1/3$, to be specific), backtracking will indeed not backtrack at all and return $t = 1$ (see [**?**, p. 490]).

We start by pointing out that by construction,

$$
\nabla^2 f(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k+1)} = -\nabla f(\boldsymbol{x}^{(k)}),
$$

and so by the Taylor theorem

$$
\begin{aligned}
\nabla f(\boldsymbol{x}^{(k+1)}) &= \nabla f(\boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)}) - \nabla^2 f(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k+1)} \\
&= \int_0^1 \nabla^2 f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)})\boldsymbol{d}^{(k+1)}\,\mathrm{d}t - \nabla^2 f(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k+1)} \\
&= \int_0^1 \left[\nabla^2 f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)}) - \nabla^2 f(\boldsymbol{x}^{(k)})\right] \boldsymbol{d}^{(k+1)}\,\mathrm{d}t.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\|\nabla f(\boldsymbol{x}^{(k+1)})\|_2 &\leq \int_0^1 \|\nabla^2 f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)}) - \nabla^2 f(\boldsymbol{x}^{(k)})\| \cdot \|\boldsymbol{d}^{(k+1)}\|_2 \,\mathrm{d}t \\
&\leq \int_0^1 t^2 L \|\boldsymbol{d}^{(k+1)}\|_2^2 \,\mathrm{d}t \\
&= \frac{L}{2} \|\nabla f(\boldsymbol{x}^{(k)})^{-1} \nabla f(\boldsymbol{x}^{(k)})\|_2^2 \\
&\leq \frac{L}{2m^2} \|\nabla f(\boldsymbol{x}^{(k)})\|_2^2.
\end{aligned}
$$

Since $\|\nabla f(\boldsymbol{x}^{(k)})\|_2 \leq m^2/L$, we have

$$
\frac{L}{2m^2}\|\nabla f(\boldsymbol{x}^{(k+1)})\|_2 \;\leq\; \left(\frac{L}{2m^2}\|\nabla f(\boldsymbol{x}^{(k)})\|_2\right)^2 \;\leq\; \left(\frac{1}{2}\right)^2.
$$

That is, at every iteration, we are **squaring** the error (which is less than 1/2). If we entered this stage at iteration ℓ, this means

$$
\frac{L}{2m^2}\|\nabla f(\boldsymbol{x}^{(k)})\|_2 \;\leq\; \left(\frac{L}{2m^2}\|\nabla f(\boldsymbol{x}^{(\ell)})\|_2\right)^{2^{k-\ell}} \;\leq\; \left(\frac{1}{2}\right)^{2^{k-\ell}}.
$$

Then using the strong convexity of f,

$$
f(\boldsymbol{x}^{(k)}) - p^\star \;\leq\; \frac{1}{2m}\|\nabla f(\boldsymbol{x}^{(k)})\|_2^2 \;\leq\; \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{k-\ell+1}}.
$$

The right hand side above is less than ϵ when

$$
k - \ell + 1 \geq \log_2 \log_2(\epsilon_0/\epsilon), \quad \epsilon_0 = 2m^3/L^2,
$$

so we spend no more than $\log_2 \log_2(\epsilon_0/\epsilon)$ iterations in this phase.

Note that

$$
\epsilon = 10^{-20}\epsilon_0 \quad \Rightarrow \quad \log_2 \log_2(\epsilon_0/\epsilon) = 6.0539.
$$