3. Convergence of Newton's Method

We won't prove the following classic theorem:

Theorem (Local Convergence)

Let $F: \mathbb{R}^d \to \mathbb{R}^d$ be cont diff and assume $F(x^*) = 0$ for some x^*. If $\nabla F(x^*)$ is nonsingular, then some neighborhood S of x^* has any $x_0 \in S$ produce Newton steps

$$x_k \in S, \quad x_k \to x^*, \quad \underline{\nabla F(x_k) \text{ nonsingular}}.$$

Define $\|\overset{\in \mathbb{R}^{d \times d}}{A}\| = \max_{\|x\|_2 = 1} \|Ax\|_2$ (if A symmetric, like $\nabla^2 f$: $\max\{|\lambda_i|\}$)

Lemma If A is nonsingular and $\|A^{-1}(B-A)\| < 1$, then B is nonsingular with

$$\|B^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}(B-A)\|}.$$

(works for any norm $\|AB\| \le \|A\| \cdot \|B\|$)

Theorem (Quadratic Convergence) $F: \mathbb{R}^d \to \mathbb{R}^d$, $\nabla F(x) \in \mathbb{R}^{d \times d}$

Consider some $x^* \in \mathbb{R}^d$ with $F(x^*) = 0$ and $\nabla F(x^*)$ nonsingular.

Suppose some neighborhood $B(x^*, r)$ has $\nabla F(x)$ exist
and is Lipschitz continuous with constant L.

Then some $\varepsilon > 0$ has all $x_0 \in B(x^*, \varepsilon)$ produce Newton
steps with $\nabla F(x_k)$ nonsingular, $x_k \in B(x^*, \varepsilon)$,
and $\|x_{k+1} - x^*\| \leq c \cdot \|x_k - x^*\|^2$.

Proof. First, lets bound $\nabla F(x_0)$, showing they are nicely invertible.

We know $\nabla F(x^*)$ is nonsingular

$\Rightarrow$ Define $M = \|\nabla F(x^*)^{-1}\| < \infty$.

Lets look at the neighborhood $\varepsilon = \min(r, \frac{1}{2ML})$.

This ensures

$$
\begin{aligned}
\|\nabla F(x^*)^{-1}(\nabla F(x_0) - \nabla F(x^*))\| &\leq \|\nabla F(x^*)^{-1}\| \, \|\nabla F(x_0) - \nabla F(x^*)\| \\
&\leq M L \|x_0 - x^*\| \\
&\leq M L \varepsilon < 1/2 .
\end{aligned}
$$

$\Rightarrow$ Lemma applies.

$\Rightarrow$ First step is well-defined ($\nabla F(x_0)$ nonsingular)
and $\|\nabla F(x_0)^{-1}\| \leq 2M$.

Lets show first step converges quadratically.

$$
\begin{aligned}
x_1 - x^* &= x_0 - x^* - \nabla F(x_0)^{-1} F(x_0) \\
&= x_0 - x^* - \nabla F(x_0)^{-1} \left(F(x_0) - F(x^*) \right) \\
&= \nabla F(x_0)^{-1} \left[F(x^*) - \left(F(x_0) + \nabla F(x_0)(x^* - x_0) \right) \right]
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \|x_1 - x^*\| &\leq \|\nabla F(x_0)^{-1}\| \, \| F(x^*) - \left(F(x_0) + \nabla F(x_0)(x^* - x_0) \right) \| \\
&\leq 2M \cdot \frac{L}{2} \|x_0 - x^*\|^2 \\
&= ML \|x_0 - x^*\|^2
\end{aligned}
$$

, quadratic convergence $c = ML$.

To iductively apply this, we need $\|x_1 - x^*\| \leq \varepsilon$.

Reuse quadratic bound....

$$
\begin{aligned}
\|x_1 - x^*\| &\leq ML \|x_0 - x^*\| \cdot \|x_0 - x^*\| \\
&\leq ML \, \varepsilon \cdot \varepsilon \\
&\leq \tfrac{1}{2} \cdot \varepsilon = \varepsilon/2 < \varepsilon.
\end{aligned}
$$

$\square$

4. Problems with Newton

Convergence is only local in neighborhood of size $\sim \frac{1}{ML}$.

($M \sim$ how invertible the Jacobian is
$L \sim$ how nicely differentiable we are)

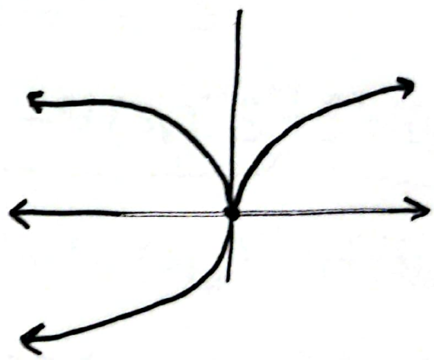
$F: \mathbb{C} \to \mathbb{C}$ goes wild when not local (fractals show up).

Newton may not converge when not initialized close enough. Examples $F: \mathbb{R} \to \mathbb{R}$

$$F(x) = \sqrt{|x|} \quad \text{or} \quad \begin{cases} \sqrt{x} & \text{if } x > 0 \\ -\sqrt{-x} & \text{if } x \le 0 \end{cases}$$

Initialize $x_0 = 1$

Newton alternates ± 1.



$F(x) = |x|^{1/4}$

Initialize $x_0 = 1$

Newton diverges.

Check $F(x) = x^2$, Newton converges but quadratically.

Sign Invariant. Behaves the same on $F(x)=0$ and $-F(x)=0$.

In optimization $\min f(x) \Rightarrow \nabla f(x)=0$

$\max -f(x) \Rightarrow -\nabla f(x)=0.$

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

Going uphill is not desirable.

Scale Invariant. For invertible $S \in \mathbb{R}^{d\times d}$

$$\min_{x\in\mathbb{R}^d} f(x) \iff \min_{y\in\mathbb{R}^d} \underbrace{f(Sy)}_{=h(y)}$$

$$\nabla h(y) = S^T \nabla f(Sy)$$

$$\nabla^2 h(y) = S^T \nabla^2 f(Sy) S$$

$$\nabla^2 h(y)^{-1} = S^{-1} \nabla^2 f(Sy)^{-1} S^{-T}$$

$$y_{k+1} = y_k - \underbrace{S^{-1}\nabla^2 f(Sy_k)^{-1} S^{-T}}\ \underbrace{S^T \nabla f(Sy_k)}$$

$$= y_k - S^{-1}\nabla^2 f(Sy_k)^{-1} \nabla f(Sy_k)$$

$$\underbrace{Sy_{k+1}}_{x_{k+1}} = \underbrace{Sy_k}_{x_k} - \nabla^2 f(\underbrace{Sy_k}_{x_k})^{-1} \nabla f(\underbrace{Sy_k}_{x_k})$$

Iteration Cost / Computational Complexity

(works millions) Compute gradient $\qquad O(d)$ memory, time

(works $10^{4.5}$) Compute Hessian $\qquad O(d^2)$ memory, time

(works 10^3) Solve $\quad \nabla F(x_k)^2 p = -F(x_k)$

worse than $O(d^2)$

d^3 if directly.

Interior Points Method (next semester).

Modified / Quasi-Newton Methods

1. Issues with Eigenvalues
2. Modified Newton
3. Convergence Guarantees
4. Computational Concerns
5. Approximating Hessians / Secant Equations
6. Quasi-Newton Method (BFGS)
7. Quasi-Newton Superlinear Convergence. $(10^{4.5} - 10^{6})$

1. Issues with Eigenvalues

Newton's Method repeatedly moves to the stationary point of our 2^{nd} order model.

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \underline{\nabla^2 f(x_k)} (x - x_k)$$

Lets work more generally with models

$$m_k(x) = \underbrace{f_k}_{\mathbb{R}} + \underbrace{g_k^T}_{\mathbb{R}^d} (x - x_k) + \frac{1}{2}(x - x_k)^T \underbrace{B_k}_{\mathbb{R}^{d \times d}} (x - x_k)$$

When $\begin{cases} f_k = f(x_k) \\ g_k = \nabla f(x_k) \\ B_k = \nabla^2 f(x_k) \end{cases}$, this is just 2^{nd} order model used by Newton

If instead $B_k = LI$, we can recover the quadratic from our characterization of Lipz ∇f.

(L-Lipz grad $\Leftrightarrow \nabla^2 f(x) \preceq LI$)

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{L}{2} \|x - x_k\|^2$$

The minimum/unique stationary point

is $x_k - \frac{1}{L} \nabla f(x_k)$.

(recover GD).