Newton's Method

Recall Newton (or "Newton-Raphson") method for finding solution to non-linear equation:

$$g(x) = 0 \quad \text{---} \ (*)$$

with $g : \mathbb{R} \to \mathbb{R}$. Given x_k, find x_{k+1} to solve $(*)$

$$0 = g(x_{k+1}) \approx g(x_k) + g'(x_k)(x_{k+1} - x_k)$$

Assuming $g'(x_k) \neq 0$, set

$$x_{k+1} = x_k - (g'(x_k))^{-1} g(x_k)$$

Generalization to optimization

In optimization goal is to get to x s.t. $\nabla f(x) = 0$

Given x_k, we want find x_{k+1} s.t. $\nabla f(x_{k+1}) = 0$

Taylor's Approx: $\nabla f(x_{k+1}) \approx \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)$

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

assuming this is non-singular

- This is called Newton's Method for optimization
- Can be viewed as general GD: $x_{k+1} = x_k + \alpha_k d_k$ with $\alpha_k = 1$, and $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$

If $\nabla^2 f(x_k) > 0$, then

$$\nabla f(x_k)^T d_k < 0$$

Convergence of Newton's Method

Let x^* be s.t. $\nabla f(x^*) = 0$. Then

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

$$= \|x_k - x^* - (\nabla^2 f(x_k))^{-1} (\nabla f(x_k) - \nabla f(x^*))\|$$

By Taylor's Theorem,

$$\nabla f(x_k) = \nabla f(x^*) + \nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*)$$

for some $\beta \in [0,1]$

Thus,

$$\|x_{k+1} - x^*\| = \|x_k - x^* - (\nabla^2 f(x_k))^{-1} \nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*)\|$$

$$= \|(\nabla^2 f(x_k))^{-1} \nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*) - (x_k - x^*)\|$$

$$= \|\underbrace{(\nabla^2 f(x_k))^{-1}}_{n \times n \text{ matrix}} \underbrace{(\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k))}_{n \times n \text{ matrix}} \underbrace{(x_k - x^*)}_{n \times 1 \text{ vector}}\|$$

Recall Matrix Norm,

$$\|A\|_{n \times n} = \max_{x : \|x\| = 1} \|Ax\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Also, if A is symmetric, $\|A\| = \lambda_{\max}(A)$.

Note that $\frac{\|Ax\|}{\|x\|} \leq \|A\|$ for all $x \neq 0$

$\Rightarrow \|Ax\| \leq \|A\| \|x\|$ for all x

For $A_{n \times n}$, $B_{n \times n}$, $\|ABx\| \leq \|A\| \|B\| \|x\|$

Thus,

$$\|x_{k+1}-x^*\| \le \|\nabla^2 f(x_k)^{-1}\| \, \|\nabla^2 f(x^*+\beta(x_k-x^*)) - \nabla^2 f(x_k)\| \cdot \|x_k - x^*\|$$

Now suppose f is locally strongly convex near x^*.

Then $\nabla^2 f(x^*) \succcurlyeq mI$, with $m>0$

$$\Rightarrow \lambda_{\min}(\nabla^2 f(x^*)) \ge m > 0$$

For symmetric $A \overset{n\times n}{\succ} 0$, with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n > 0$, A^{-1} has eigenvalues, $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_n}$.

Proof If u_i is an eigenvector for λ_i,

$Au_i = \lambda_i u_i \Rightarrow A^{-1}Au_i = \lambda_i A^{-1}u_i \Rightarrow \lambda_i^{-1}u_i = A^{-1}u_i$ ●

This implies $\|A^{-1}\| = \lambda_{\max}(A^{-1}) = \lambda_{\min}^{-1}(A)$.

Assuming $\nabla^2 f(x)$ is continuous, if $\|x_k - x^*\|$ is small, then $\lambda_{\min}(\nabla^2 f(x_k))$ is close to $\lambda_{\min}(\nabla^2 f(x^*))$

i.e. $\lambda_{\min}(\nabla^2 f(x_k)) \ge \tilde{\gamma} > 0$. Then

$$\|\nabla^2 f(x_k)^{-1}\| = \lambda_{\min}^{-1}(\nabla^2 f(x_k)) \le \frac{1}{\tilde{\gamma}} \triangleq \gamma$$

Furthermore, assume that $\nabla^2 f$ is L-Lipschitz in a neighborhood $\mathcal{S}$ of x^*, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L\|x-y\| \quad \text{for } x, y \in \mathcal{S}.$$

Thus, $\|x_{k+1}-x^*\| \le \gamma L \|x^*+\beta(x_k-x^*)-x_k\| \|x_k-x^*\|$

$= \gamma L \|(\beta-1)(x_k-x^*)\| \|x_k-x^*\|$

$\because \beta \in [0,1] \longrightarrow \le \gamma L \|x_k-x^*\|^2$

Thus: $\|x_{k+1}-x^*\| \le \gamma L \|x_k-x^*\|^2$

Now suppose x_0 is close enough to x^* s.t.

$$\|x_0-x^*\| \gamma L = \delta < 1$$

Then,

$\|x_1-x^*\| \le \delta \|x_0-x^*\|$,

$\|x_2-x^*\| \le \gamma \cdot L \|x_1-x^*\|^2$

$\le \gamma L \delta^2 \|x_0-x^*\|^2 = \delta^3 \|x_0-x^*\|$

$\|x_3-x^*\| \le \gamma L \|x_2-x^*\|^2$

$\le \gamma L \delta^6 \|x_0-x^*\|^2 = \delta^7 \|x_0-x^*\|$

By induction,

$$\|x_N-x^*\| \le \delta^{2^N-1} \|x_0-x^*\|$$

Assuming ∇f is M-Lipshitz in neighborhood of x^*,

$f(x_N)-f(x^*) \le \nabla f(x^*)^{=0}(x_N-x^*) + \frac{M}{2}\|x_N-x^*\|^2$

$\le \frac{M}{2} \delta^{(2^{N+1}-2)} \|x_0-x^*\|^2$

Thus to make $f(x_N)-f(x^*) < \varepsilon$, need $N \sim O\left(\log\left(\log\left(\frac{1}{\varepsilon}\right)\right)\right)$

Order-2 or super-linear convergence!

- Newton's method is super-fast close to local min if function is strongly convex around min.

- If the function is quadratic, Newton's method converges in one step!

$$f(x) = \frac{1}{2} x^T Q x + b x + c, \quad Q \succ 0.$$

$$\nabla f(x) = Qx + b, \quad \nabla^2 f(x) = Q.$$

Global min x^* satisfies $Qx^* + b = 0$

$$\Rightarrow x^* = -Q^{-1} b$$

Newton's method: For any $x_0 \in \mathbb{R}^n$,

$$x_1 = x_0 - (\nabla^2 f(x_0))^{-1} \nabla f(x_0)$$

$$= x_0 - Q^{-1}(Qx_0 + b) = -Q^{-1}b = x^*$$

- But Newton's method has several drawbacks:

(1) $\nabla^2 f(x)^{-1}$ may fail to exist, i.e. $\nabla^2 f(x)$ is singular, e.g. in region where f is linear

(2) It is not necessarily a general GD method since $\nabla^2 f(x_k)$ may not be $\succ 0$.

(3) It is not a descent method, $f(x_{k+1})$ may be $> f(x_k)$

(4) It may stop at local max. or saddlepoints.

Modifications to Newton's Method to Ensure Global Convergence

(a) Try Newton's method. If either $\nabla^2 f(x_k)$ is singular or $f(x_{k+1}) > f(x_k)$ then use (b)

(b) Find δ_k s.t.

$$\left(\delta_k I + \nabla^2 f(x_k)\right) > 0$$

and $\lambda_{\min}\left(\delta_k I + \nabla^2 f(x_k)\right) \geq \Delta > 0$

so that $\delta_k I + \nabla^2 f(x_k)$ is easily invertible

Then set $d_k = -\left(\delta_k I + \nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$

This ensures that $d_k^T \nabla f(x_k) < 0$

Then use $x_{k+1} = x_k + \alpha_k d_k$ with α_k chosen using Armijo's Rule.

- If at any point $\nabla^2 f(x_k) > 0$ go back to Newton's method and check if $f(x_{k+1}) < f(x_k)$. Continue Newton's method as long as $\nabla^2 f(x_k) > 0$ and $f(x_{k+1}) < f(x_k)$.