Newton's Method

* We have seen that GD was derived from a 1^{st} order Taylor approximation.

$$f(x) \approx f(x^{(t)}) + \underbrace{\nabla f(x^{(t)})^T (x - x^{(t)})}_{\text{want this to be as negative as possible}}$$

want this to be small

$$\Rightarrow \boxed{x^{(t+1)} = x^{(t)} - \mu \nabla f(x^{(t)})}$$

This gives

$$f(x^{(t+1)}) \approx f(x^{(t)}) - \mu \|\nabla f(x^{(t)})\|^2$$

* If instead of a 1^{st} order Taylor, we use a 2^{nd} order Taylor approximation, we get Newton's method:

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

We expect that $\nabla f = 0$ at a minimum

So taking derivatives on both sides

$$\underbrace{\nabla f(x)}_{0} \approx 0 + \nabla f(x^{(t)}) + \nabla^2 f(x^{(t)})(x - x^{(t)})$$

$\Rightarrow$ when the LHS is 0 (i.e. $\nabla f = 0$) we have

$$\Rightarrow x - x^{(t)} \approx -\left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})$$

at a minimizer

So we can set

Newton's Method

$$\boxed{x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})}$$

Example: Let $f : \mathbb{R}^+ \to \mathbb{R}$ be given by

$$f(x) = x - \ln(x)$$

Then $\nabla f(x) = f'(x) = 1 - \frac{1}{x}$

& $\nabla^2 f(x) = f''(x) = \frac{1}{x^2}$

Newton's method initialized to $x^{(0)} = 0.5$ gives

$$x^{(t+1)} = x^{(t)} - \left[f''(x^{(t)})\right]^{-1} f'(x^{(t)})$$

$$= x^{(t)} - (x^{(t)})^2 \left(1 - \frac{1}{x^{(t)}}\right)$$

$$= 2x^{(t)} - (x^{(t)})^2$$

$$\Rightarrow x^{(1)} = 2x^{(0)} - (x^{(0)})^2$$

$$= 2 \times 0.5 - 0.5^2$$

$$= 0.75$$

$$\Rightarrow x^{(2)} = \cdots = 0.9375$$

$$x^{(3)} = \cdots = 0.9961$$

$$x^{(4)} = \cdots = 0.9998$$

(In fact the optimum is at $x^* = 1$)

If this appears fast, it is <u>not</u> a coincidence!

We'll need a def'n to discuss the convergence of Newton's method.

Def'n: For a matrix M

$$\|M\| = \max_{x \neq 0} \frac{\|Mx\|}{\|x\|}$$

length of Mx

length of x

$\|M\|$ measures how much M stretches vectors

Consequence of the def'n:

$$\forall z: \ \|Mz\| \leq \|M\| \, \|z\|$$

Theorem: (Conv. of Newton's Method)

Let f be twice continuously differentiable & suppose that x^* has $\nabla f(x^*) = 0$. Suppose further that:

$$\begin{cases} \|\nabla^2 f(x^*)^{-1}\| \leq \frac{1}{h} & \text{for some } h > 0 \\ \|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq L \|x - x^*\| & \text{for all } x \end{cases}$$

Then if $\|x^{(0)} - x^*\| \leq \frac{2h}{3L}$

& $x^{(t+1)} = x^{(t)} - [\nabla^2 f(x^{(t)})]^{-1} \nabla f(x^{(t)})$

we have

$$\begin{cases} \|x^{(t)} - x^*\| \leq 2h/3L & \forall t \\ \|x^{(t+1)} - x^*\| \leq \frac{3L}{2h} \|x^{(t)} - x^*\|^2 & \forall t \end{cases}$$

Loose interpretation: If we start close to a local minimizer and the fct is nice, we converge quickly to the minimum.

Example: $f(x) = x_1^4 + 2x_1^2 x_2^2 + x_2^4$

To use Newton's method, need $\nabla f(x)$, $\nabla^2 f(x)$:

$$\nabla f(x) = \begin{pmatrix} 4x_1^3 + 4x_1 x_2^2 \\ 4x_1^2 x_2 + 4x_2^3 \end{pmatrix}$$

$$\nabla^2 f(x) = \begin{pmatrix} 12x_1^2 + 4x_2^2 & 8x_1 x_2 \\ 8x_1 x_2 & 4x_1^2 + 12x_2^2 \end{pmatrix}$$

Suppose $x^{(0)} = (1, 1)$

Then: $x^{(1)} = x^{(0)} - \left[\nabla^2 f(x^{(0)})\right]^{-1} \nabla f(x^{(0)})$

$$= \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 16 & 8 \\ 8 & 16 \end{pmatrix}^{-1} \begin{pmatrix} 8 \\ 8 \end{pmatrix}$$

$$= \begin{pmatrix} 2/3 \\ 2/3 \end{pmatrix}$$

Continuing in this way

$$x^{(2)} = \cdots = \begin{pmatrix} (2/3)^2 \\ (2/3)^2 \end{pmatrix}$$

$\vdots$

$$x^{(t)} = (2/3, 2/3)^t \xrightarrow[t \to \infty]{} 0$$

(Converges to zero exponentially fast in t)!

Example: $f(x) = \frac{x^4}{4} - x^2 + 2x + 1$

start at $x^{(0)} = 0$

$$\nabla f(x) = f'(x) = x^3 - 2x + 2$$

$$\nabla^2 f(x) = f''(x) = 3x^2 - 2$$

$$\Rightarrow x^{(1)} = x^{(0)} - \left[\nabla^2 f(x^{(0)})\right]^{-1} \nabla f(x^{(0)})$$

$$= 0 - (-2)^{-1} \cdot 2$$

$$= 1$$

$$\Rightarrow x^{(2)} = x^{(1)} - \left[\nabla^2 f(x^{(1)})\right]^{-1} \nabla f(x^{(1)})$$

$$= 1 - 1^{-1} \cdot 1 = 0$$

$$= x^{(0)}$$

it sent me back to $x^{(0)}$, so we entered a cycle!

$\Rightarrow$ Newton's method may not always converge

Why does this not contradict our theorem?

Continuing with Newton's Method

$$\boxed{x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})}$$

Example: $f(x) = \frac{x^4}{4} - x^2 + 2x + 1$

$f: \mathbb{R} \to \mathbb{R}$

Start Newton's Method at $x^{(0)} = 0$.

$$\nabla f(x) = x^3 - 2x + 2 \qquad (= f'(x))$$

$$\nabla^2 f(x) = 3x^2 - 2 \qquad (= f''(x))$$

$$\Rightarrow x^{(1)} = x^{(0)} - \left[f''(x^{(0)})\right]^{-1} f'(x^{(0)})$$

$$= 0 - (-2)^{-1}(2)$$

$$= 1$$

$$
\begin{aligned}
x^{(2)} &= x^{(1)} - \left[f''(x^{(1)})\right]^{-1} f'(x^{(1)}) \\
&= 1 - (1)^{-1}(1) = 0 \\
&= x^{(0)}
\end{aligned}
$$

$$
\begin{tikzcd}
x^{(0)} \arrow[r] \arrow[d, squiggly] & x^{(1)} \arrow[r] \arrow[d] & x^{(2)} \arrow[d] \\
0 & 1 & 0
\end{tikzcd}
$$

Back to $x^{(0)}$, so we entered a cycle!

- So, Newton's method need not always converge
- Why does this not contradict the theorem? (exercise).

Some Remarks on Newton's Method:

(1) The theorem tells us that Newton's method can converge very fast (in terms of the number of iterations)

(2) On the other hand, finding the inverse of the Hessian can be expense if n is large.

Instead, in practice, the following observation is useful

$$x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})$$

$$\Rightarrow \nabla^2 f(x^{(t)})\left(x^{(t+1)} - x^{(t)}\right) = -\nabla f(x^{(t)})$$

$$\Rightarrow \underbrace{\nabla^2 f(x^{(t)})}_{\text{known}} \underbrace{x^{(t+1)}}_{\text{unknown}} = \underbrace{\nabla^2 f(x^{(t)})}_{\text{known}} \underbrace{x^{(t)}}_{\text{known}} - \underbrace{\nabla f(x^{(t)})}_{\text{know}}$$

$\Rightarrow$ we have a system that looks like $Ax^{(t+1)} = b$ and we want to solve for $x^{(t+1)}$.

So we can use linear algebra techniques to solve for $x^{(t+1)}$.

(3) We can modify Newton's method, for example, to include a step-size

$$x^{(t+1)} = x^{(t)} - \mu^{(t)} \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})$$

- can choose a fixed μ
- Can choose via backtracking line search.

Quasi-Newton Methods:

(very briefly)

Recall that GD had an interpretation whereby

$$f(x) \approx f(x^{(t)}) + \nabla f(x^{(t)})^T (x - x^{(t)}) + \frac{1}{2\mu^{(t)}} \|x - x^{(t)}\|^2$$

then minimizing the RHS gave us

$$\boxed{x^{(t+1)} = x^{(t)} - \mu^{(t)} \nabla f(x^{(t)}).} \leftarrow \text{GD}$$

Meanwhile Newton's method approximates

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

As before, minimizing the RHS gives us

$$\boxed{x^{(t+1)} = x^{(t)} - \left[\nabla^2 f(x^{(t)})\right]^{-1} \nabla f(x^{(t)})} \leftarrow \text{Newton's method}$$

So GD approximates the Hessian with $(1/\mu^{(t)})I$.

Quasi-Newton methods approximate the Hessian with some matrix $B^{(t)}$ which may change from iteration to iteration, so that

$$x^{(t+1)} = x^{(t)} - \mu^{(t)}[B^{(t)}]^{-1}\nabla f(x^{(t)})$$

There are several such methods with different choices of $B^{(t)}$

We won't cover them here, but examples are

BFGS methods
Broyden method
⋮