First order Taylor series approximation

$$h(w) = g(w^0) + \frac{d}{dw}g(w^0)\cdot(w - w^0)$$

$$h(\bar{w}) = g(\bar{w}^0) + \nabla g(\bar{w}^0)^T(\bar{w} - w^0)$$

Gradient descent

$$\bar{w}^k = \bar{w}^{k-1} - \alpha\nabla g(\bar{w}^{k-1})$$

tions

Second order Taylor series approximation

$$h(w) = g(v) + \left(\frac{d}{dw}g(v)\right)(w-v) + \frac{1}{2}\left(\frac{d^2}{dw^2}g(v)\right)(w-v)^2$$

$$h(\bar{w}) = g(\bar{v}) + \nabla g(\bar{v})^T(\bar{w}-\bar{v}) + \frac{1}{2}(\bar{w}-\bar{v})^T\nabla^2 g(\bar{v})(\bar{w}-\bar{v})$$

$$\frac{dh(w)}{dw} = 0 \Rightarrow w^* = v - \frac{\frac{d}{dw}g(v)}{\frac{d^2}{dw^2}g(v)}$$

$$\nabla h(\bar{w}) = \bar{0} \Rightarrow \bar{w}^* = \bar{v} - (\nabla^2 g(\bar{v}))^{-1} g(\bar{v})$$

Newton's method.

$$h(w) = g(w^{k-1}) + \left(\frac{d}{dw}g(w^{k-1})\right)(w - w^{k-1}) + \frac{1}{2}\left(\frac{d^2}{dw^2}g(w^{k-1})\right)(w-w^{k-1})^2$$

$$\Rightarrow w^k = w^{k-1} - \frac{\frac{d}{dw}g(w^{k-1})}{\frac{d^2}{dw^2}g(w^{k-1})}$$

$$h(\bar{w}) = g(\bar{w}^{k-1}) + \nabla g(\bar{w}^{k-1})^T(\bar{w} - \bar{w}^{k-1}) + \frac{1}{2}(\bar{w} - \bar{w}^{k-1})^T \nabla^2 g(\bar{w}^{k-1})(\bar{w} - \bar{w}^{k-1})$$

$$\Rightarrow \bar{w}^k = \bar{w}^{k-1} \underbrace{- (\nabla^2 g(w^{k-1}))^{-1}\nabla g(w^{k-1})}_{\bar{d}_1^k}$$

$$\alpha = 1$$

2.1 Gradient descent

Assume $\bar{w}^k = \bar{w}^{k-1} + \eta \vec{m} \Rightarrow \eta \vec{m} = \bar{w}^k - \bar{w}^{k-1}$

where η is the step length (we assume η is a constant whatever the value of k)

Goal: $\min g(\bar{w})$. According to First order taylor series approximation

$\because h(w^k) = g(\bar{w}^{k-1}) + \eta \vec{m}^T \nabla g(\bar{w}^{k-1})$

$\therefore \min g(\bar{w}) = \min \{ g(\bar{w}^{k-1}) + \eta \underbrace{\vec{m}^T \nabla g(\bar{w}^{k-1})} \}$

Goal: inner product is smallest.

$\because \vec{m}^T \nabla g(\bar{w}^{k-1}) = \|\vec{m}^T\| \, \|\nabla g(\bar{w}^{k-1})\| \cos\theta$.

$\therefore \vec{m} = -\nabla g(\bar{w}^{k-1})$

which means $\vec{m}$ is converse to the direction of $\nabla g(\bar{w}^{k-1})$.

$\therefore \bar{w}^k = \bar{w}^{k-1} + \eta \vec{m}$

$= \bar{w}^{k-1} - \eta \nabla g(\bar{w}^{k-1})$

Here we regard the length of $\vec{m}$ as the same of $\nabla g(\bar{w}^{k-1})$

So the algorithm of Newton's method is:
Input: Tutee differentiable function g and initial point $\bar{w}^0$; $k=1$
Repeat until stopping condition is met:

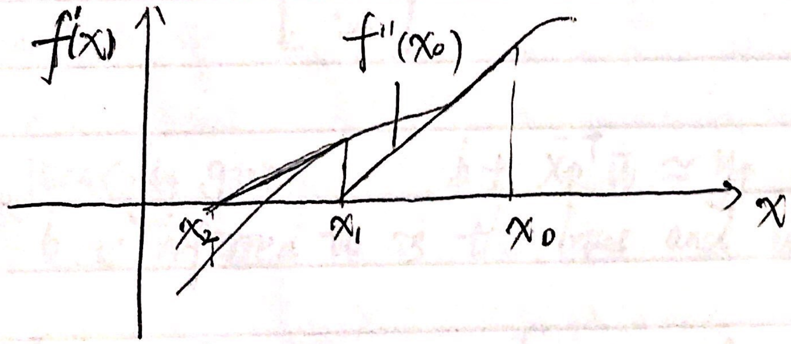
If $\nabla^2 g(\bar{w}^{k-1})$ is invertible:

$$\bar{w}^k = \bar{w}^{k-1} - [\nabla^2 g(\bar{w}^{k-1})]^{-1} \nabla g(\bar{w}^{k-1})$$

If one dimentional:

$$w^k = w^{k-1} - \frac{g'(w^{k-1})}{g''(w^{k-1})}$$

To make it more clear, with lower-level: (one dimention)



goal: $f'(x)=0$

$$a x_n + b = f'(x_n)$$

$$f''(x_n) x_n + b = f'(x_n) \Rightarrow b = f'(x_n) - f''(x_n) x_n$$

$$\therefore \quad f''(x_n) x_{n+1} + f'(x_n) - f''(x_n) x_n = 0$$

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

Newton's method produces a sequence of points $\vec{w}^1$, $\vec{w}^2$. ,… that minimize cost function g by repeatedly creating the second order Taylor series quadratic approximation to the function, and traveling to a stationary point of this quadratic. Because Newton's method uses quadratic as opposed to linear approximation at each step, with a quadratic more closely mimicking the associated function, it is often much more effective than gradient descent in the sense that it requires for fewer steps for convergence.

$Q_{N\times N}$ symmetric

If it satisfies $\bar{z}^T Q \bar{z} \geq 0$, for all $\bar{z}$, then it must have all non-negative eigenvalues.

Proof: $Q = \sum_{i=1}^{P} \underbrace{\bar{x}_i \bar{x}_i^T}_{N\times N}$

$$\bar{z}^T \left(\sum_i \bar{x}_i \bar{x}_i^T\right) z^T = \sum_i \bar{z}^T \bar{x}_i \bar{x}_i^T \bar{z}$$

if. $\bar{b}_i = \bar{x}_i^T \bar{z}$

$$= \sum_i \bar{b}_i^T \bar{b}_i = \sum_i \| b_i \|^2 \geq 0$$

$\therefore$ Q positive definite.

Take the general multi-input quadratic function.

$$g(\bar{w}) = a + \bar{b}^T \bar{w} + \bar{w}^T \bar{C} \bar{w}$$

Where C is an $N\times N$ symmetric matrix, $\bar{b}$ is an $N\times 1$ vector, and a is a scalar. Computing the first derivative (gradient) we have

$$\nabla g(\bar{w}) = 2\bar{C}\bar{w} + b \quad \left(\bar{C}\bar{w} = -\tfrac{1}{2}\bar{b} \right)$$

$$\nabla^2 g(\bar{w}) = 2\bar{C}.$$