# Optimization algorithms (part 1)

$$w_{t+1} = w_t - \alpha \nabla l(w_t) \qquad \eta \Rightarrow lr$$

$l(w) = 4w^2 - 12w + 9$ $\qquad w = 1.5$

We have $l'(w) = 8w - 12$ $\qquad \nabla l(w_t) = 8w_t - 12$

$l''(w) = 8$

We start with $w_0 = 3$ & follow the update rule

$$w_{k+1} = w_k - \eta\, l'(w_k)$$

$k \Rightarrow$ No. of steps

$$\Rightarrow w_{k+1} = w_k - \eta(8w_k - 12)$$

$$\Rightarrow \boxed{w_{k+1} = w_k - 0.2(8w_k - 12)}$$ We assume $\eta = 0.2$

The Taylor series expansion of a function $f(x)$ that is differentiable at $a$ is given by

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

We consider a loss function $\ell(\omega)$ & replace $f(\cdot)$ by $\ell(\cdot)$, $x$ by $\omega_{k+1}$ & $a$ by $\omega_k$. We get -

$$\ell(\omega_{k+1}) = \ell(\omega_k) + \frac{\ell'(\omega_k)}{1!}[\omega_{k+1} - \omega_k] + \frac{\ell''(\omega_k)}{2!}(\omega_{k+1} - \omega_k)^2 + \cdots$$

**Assumption** :- Loss function in any bounded region can be approximated with a Quadratic function.

$$\Rightarrow \ell(\omega_{k+1}) = \ell(\omega_k) + (\omega_{k+1} - \omega_k)\,\ell'(\omega_k) + \frac{(\omega_{k+1} - \omega_k)^2}{2}\,\ell''(\omega_k)$$

We need to find $\omega_{k+1}$ & we partially differentiate wrt $\omega_{k+1}$

$$\frac{d}{d\omega_{k+1}}\ell(\omega_{k+1}) = \frac{d}{d\omega_{k+1}}\ell(\omega_k) + \frac{d}{d\omega_{k+1}}\left[(\omega_{k+1} - \omega_k)\,\ell'(\omega_k)\right] + \frac{d}{d\omega_{k+1}}\left[\frac{(\omega_{k+1} - \omega_k)^2}{2}\,\ell''(\omega_k)\right]$$

$$= 0 + \ell'(\omega_k)\frac{d}{d\omega_{k+1}}(\omega_{k+1} - \omega_k) + \frac{\ell''(\omega_k)}{2}\underbrace{\frac{d}{d\omega_{k+1}}(\omega_{k+1} - \omega_k)^2}_{\hookrightarrow\, 2(\omega_{k+1} - \omega_k)}$$

$$= \ell'(\omega_k) + 2\frac{\ell''(\omega_k)}{2}[\omega_{k+1} - \omega_k]$$

$$\boxed{\frac{d\,\ell(\omega_{k+1})}{d\omega_{k+1}} = \ell'(\omega_k) + \ell''(\omega_k)[\omega_{k+1} - \omega_k]}$$

$$= \underline{\underline{l'(\omega_K)}} + \frac{2\,l''(\omega_K)}{2}\left[\omega_{K+1} - \omega_K\right]$$

$$\boxed{\frac{d\,l(\omega_{K+1})}{d\omega_{K+1}} = \underline{l'(\omega_K) + l''(\omega_K)\left[\omega_{K+1} - \omega_K\right]}}$$

For extremum points; $\dfrac{d\,l(\omega_{K+1})}{d\omega_{K+1}} = 0$

$$\Rightarrow \underline{l'(\omega_K)} + l''(\omega_K)\left[\omega_{K+1} - \omega_K\right] = 0$$

$$\Rightarrow \underline{\underline{l''(\omega_K)}}\left[\omega_{K+1} - \omega_K\right] = -l'(\omega_K)$$

$$\Rightarrow \omega_{K+1} - \omega_K = -\frac{1}{l''(\omega_K)}\, l'(\omega_K)$$

$$\Rightarrow \omega_{K+1} = \omega_K - \frac{1}{l''(\omega_K)}\, l'(\omega_K)$$

$$\Rightarrow \boxed{\omega_{K+1} = \omega_K - \underline{\underline{\eta}}\, l'(\omega_K)} \quad \text{where } \underline{\eta = \frac{1}{l''(\omega_K)}}$$

Further, $\dfrac{d^2 l(\omega_{K+1})}{d\omega_{K+1}^2} = l''(\omega_K) > 0$

$\Rightarrow \omega_{K+1}$ is $\underline{\underline{\text{local}}}$ minima. $l(\omega_{K+1})$

[Note :- learning rate $(\eta) = \dfrac{1}{l''(\omega_K)}$ ]

In multivariate case;

$$l'(w_K) = \nabla l(w_K) = g$$

$w_K \Rightarrow$ vector

$$l''(w_K) = H \quad [\text{Hessian matrix of } l \text{ at } w_K]$$

$H \Rightarrow$ matrix

Also, $\underbrace{(w_{K+1} - w_K)^2 l''(w_K)} = \underbrace{(w_{K+1} - w_K)^T} H \underbrace{(w_{K+1} - w_K)}$

Under the same set of assumptions,

$$l(w_{K+1}) = l(w_K) + (w_{K+1} - w_K)^T g + \frac{1}{2} \underbrace{(w_{K+1} - w_K)^T H (w_{K+1} - w_K)}$$

$\rightarrow (w_{K+1} - w_K)^2 l''(w_K)$

$\rightarrow (w_{K+1} - w_K) l'(w_K)$

Partially differentiating wrt $w_{K+1}$;

$$\nabla_{w_{K+1}} l(w_{K+1}) = \nabla_{w_{K+1}} l(w_K) + g\left[\nabla (w_{K+1} - w_K)^T\right] + \nabla_{w_{K+1}} \left[\frac{1}{2} (w_{K+1} - w_K)^T H (w_{K+1} - w_K)\right]$$

$$= g + H (w_{K+1} - w_K)$$

$$\nabla_{w_{K+1}} l(w_K) = 0$$

Reason :-

1) $\nabla_{w_{K+1}} (w_{K+1} - w_K)^T = 1$

2) $\nabla_{w_{K+1}} [w_{K+1} - w_K]^T H [w_{K+1} - w_K] = H (w_{K+1} - w_K)$

$$\Rightarrow \boxed{\nabla_{w_{K+1}} l(w_{K+1}) = g + H (w_{K+1} - w_K)}$$

For extremum points; $\nabla_{w_{k+1}} l(w_{k+1}) = 0$

$$\Rightarrow g + H(w_{k+1} - w_k) = 0$$

$$\Rightarrow H(w_{k+1} - w_k) = -g$$

$$\Rightarrow H^{-1} H(w_{k+1} - w_k) = -H^{-1} g$$

$\eta$

$$\Rightarrow w_{k+1} - w_k = -H^{-1} g$$

$$\Rightarrow w_{k+1} = w_k - H^{-1} g$$

$\frac{1}{l''(w_k)} = \eta$

$$\Rightarrow \boxed{w_{k+1} = w_k - H^{-1} \nabla l(w_k)}$$

$\eta = H^{-1}$

We have,

$$l(w_{k+1}) = l(w_k) + (w_{k+1} - w_k)^T \nabla l(w_k) + \frac{1}{2}(w_{k+1} - w_k)^T H (w_{k+1} - w_k)$$

We assume $\cancel{w}_{k+1} = w_k - \eta \nabla l(w_k)$ where $\eta$ is unknown learning rate.

$w_{k+1} - w_k = -\eta \nabla l(w_k)$
$= -\eta g$

$$\Rightarrow l(w_k - \eta \nabla l(w_k)) = l(w_k) + (w_k - \eta \nabla l(w_k) - w_k)^T \nabla l(w_k) + \frac{1}{2}(w_k - \eta \nabla l(w_k) - w_k)^T H (w_k - \eta \nabla l(w_k) - w_k)$$

$$= l(w_k) - \eta g^T g + \frac{1}{2}(-\eta g^T) H (-\eta g)$$

$$= l(w_k) - \eta g^T g + \frac{\eta^2}{2} g^T H g$$

We partially differentiate wrt. $\eta$.

$$\frac{\partial}{\partial \eta} \, l(w_k - \eta \nabla l(w_k)) = -g^T g + \eta \, g^T H g$$

for extremum points,

$$\frac{\partial}{\partial \eta} \, l(w_k - \eta \nabla l(w_k)) = 0$$

$$\Rightarrow -g^T g + \eta \, g^T H g = 0$$

$$\Rightarrow \boxed{\eta = \frac{g^T g}{g^T H g}}$$

As $H = \underbrace{\Phi D \Phi^T}_{\text{eigen dec. Th.}}$ where $D = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix}$

$\Phi$ is eigenvector matrix where each column of $\Phi$ corresponds to one eigenvector of $H$.

Now. $g^T H g = g^T \Phi D \Phi^T g$

$= X^T D X$ [we assume $X = \Phi^T g$]

$$\boxed{g^T H g = \sum_{i=1}^{n} \lambda_i x_i^2}$$

Let $\lambda_{max} = \max_i \lambda_i$

$\Rightarrow \lambda_i \leq \lambda_{max}$ $\qquad x_i^2 > 0$

$\Rightarrow \lambda_i x_i^2 \leq \lambda_{max} x_i^2$

$\Rightarrow \sum_{i=1}^{n} \lambda_i x_i^2 \leq \lambda_{max} \sum_{i=1}^{n} x_i^2$

$\Rightarrow \sum_{i=1}^{n} \lambda_i x_i^2 \leq \lambda_{max} \sum_{i=1}^{n} x_i^2 \rightarrow X^T X$ $\qquad X = \Phi^T g$

[Reason :- $H$ is a symmetric matrix

$\Rightarrow \Phi$ is orthogonal. [eigenvectors of symmetric matrix]

$\Rightarrow \Phi^T \Phi = I \Rightarrow g^T \Phi^T \Phi g^T = g^T g = g^T \Phi \Phi^T g$

.]

$\Phi^T \Phi = \Phi \Phi^T = I$

1. $\Rightarrow \Phi^T\Phi = I \Rightarrow \ldots$

$$\Rightarrow \quad g^T H g \leq \lambda_{max}\, g^T \underbrace{\Phi\Phi^T}\, g$$

$$\Rightarrow \quad g^T H g \leq \lambda_{max}\, g^T g$$

$$\Rightarrow \quad \frac{g^T H g}{g^T H g} \leq \lambda_{max} \frac{g^T g}{g^T H g}$$

$$\Rightarrow \quad \frac{g^T g}{g^T H g} \geq \frac{1}{\lambda_{max}}$$

$$\Rightarrow \quad \boxed{\eta \geq \frac{1}{\lambda_{max}}}$$

$$\varphi^T\varphi = \varphi\varphi^T = I$$

$$g^T\varphi^T\varphi = g^T\varphi\varphi^T = g^T$$

$$g^T\varphi^T\varphi g = g^T\varphi\varphi^T g = g^T g$$

$$\frac{1}{\text{highest e.v. of } H}$$

$$l(\omega) = 7 + (\omega_1 - 1)^2 + (\omega_2 - 1)^6$$

Now $\nabla_\omega l(\omega) = \begin{pmatrix} 2(\omega_1 - 1) \\ 6(\omega_2 - 1) \end{pmatrix} = \begin{pmatrix} = \frac{\partial l(\omega)}{\partial \omega_1} \\ \frac{\partial l(\omega)}{\partial \omega_2} \end{pmatrix}$

$$H(l(\omega)) = \begin{pmatrix} \frac{\partial^2 l(\omega)}{\partial \omega_1^2} & \frac{\partial^2 l(\omega)}{\partial \omega_2 \partial \omega_1} \\ \frac{\partial^2 l(\omega)}{\partial \omega_1 \partial \omega_2} & \frac{\partial^2 l(\omega)}{\partial \omega_2^2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$$

We have $H = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$ The eigenvalues of $H$ are $2, 6$.

As $\eta \geq \frac{1}{\lambda_{max}} \Rightarrow \boxed{\eta \geq \frac{1}{6}}$

We consider optimal learning rate $(\eta) = 1/6$.

Gradient update rule

$$\omega_{k+1} = \omega_k - \frac{1}{6} \nabla_{\omega_k} l(\omega_k)$$

$$= \begin{pmatrix} \omega_1^k \\ \omega_2^k \end{pmatrix} - \frac{1}{6} \begin{pmatrix} 2\omega_1^k - 2 \\ 6\omega_2^k - 6 \end{pmatrix}$$

$$= \begin{pmatrix} \omega_1^k - \frac{1}{3}\omega_1^k + \frac{1}{3} \\ \omega_2^k - \omega_2^k + \frac{6}{6} \end{pmatrix} = \begin{pmatrix} \frac{2}{3}\omega_1^k + \frac{1}{3} \\ 1 \end{pmatrix}$$

---

$$\Rightarrow \begin{pmatrix} \omega_1^{k+1} \\ \omega_2^{k+1} \end{pmatrix} = \begin{pmatrix} \frac{2}{3}\omega_1^k + \frac{1}{3} \\ 1 \end{pmatrix}$$