Error

error: $L(w,b) = \sum_i (y_i - wx_i - b)^2$

$\frac{\delta L}{\delta w} = 2\sum_i (y_i - wx_i - b)(-x_i) = 0$
$\frac{\delta L}{\delta b} = 2\sum_i (y_i - wx_i - b)(-1) = 0$

$\begin{bmatrix} \frac{\delta L}{\delta W} \\ \frac{\delta L}{\delta b} \end{bmatrix} = \nabla L(w,b) = 0$
(gradient = minimizing loss)

Optimization

$L(w) = \frac{1}{2}\left(w^T x - y\right)^2$
$\frac{\partial L}{\partial w} = (w^T x - y)x$

$w_{\text{new}} = w_{\text{old}} - \eta \frac{\partial L}{\partial w}$ (gradient descent)
$\quad = w_{\text{old}} - \eta(w^T x - y)x$
$\Delta w = -\eta(\text{error})x$
$\min_w L(w)$

Calculus

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

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}\right) \in \mathbb{R}$$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \in \mathbb{R}^d$$

Examples

$F(\mathbf{w}) = 3w_1 w_2 + w_3$

$$\nabla F(\mathbf{w}) = \begin{bmatrix} 3w_2 \\ 3w_1 \\ 1 \end{bmatrix}$$

$F(\mathbf{w}) = \mathbf{w}^T \mathbf{x} = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$

$$\nabla F(\mathbf{w}) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \mathbf{x}$$

$F(\mathbf{w}) = \|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = w_1^2 + w_2^2 + \cdots + w_d^2$

$$\nabla F = \begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_d \end{bmatrix}$$

Taylor Series

$f(x + \Delta x) \approx f(x) + \Delta x f'(x) + \frac{1}{2}(\Delta x)^2 f''(x)$
$\frac{f(x+\Delta x) - f(x)}{\Delta x} \approx f'(x)$
$f(x + \Delta x) \approx f(x) + (\Delta x)^T \nabla f + \frac{1}{2}(\Delta x)^T H (\Delta x)$
$f'' = H \in \mathbb{R}^{d \times d}$

$\nabla f = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \frac{\delta f}{\delta x_2} \end{bmatrix}$
$\nabla^2 f = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_2 \delta x_2} \\ \frac{\delta^2 f}{\delta x_1 \delta x_2} & \frac{\delta^2 f}{\delta x_2^2} \end{bmatrix}$
$\frac{\delta^2 f}{\delta x_2 \delta x_1} = \frac{\delta^2 f}{\delta x_1 \delta x_2} \to$ symmetric

Directional Derivative
$\lim_{x\to 0} \frac{1}{\alpha} \left(f(x+\alpha d) - f(x)\right) = f'_d(x) = \langle d, \nabla f(x) \rangle$

$f(x) = w^T x$
$\lim_{\alpha\to 0} \frac{1}{\alpha} \left(w^T(x+\alpha d) - w^T x\right) = \frac{1}{\alpha}\left(\alpha w^T d\right) = w^T d$
$\nabla f(x) = w = \langle d, w \rangle$

$f(x) = \frac{1}{2} x^T A x,\ A = A^T,\ x \in \mathbb{R}^d,\ A \in \mathbb{R}^{d\times d},\ \frac{1}{2} A x^2 \to Ax$
directional derivative: $\langle d, Ax \rangle$, $\nabla f(x) = Ax$

<u>Function minimization</u>
$\max(f) = -\min(-f)$
local minima vs global minima
criteria for local minima : $f(x) \leq f(x+\Delta x)$
for convex functions : local = global

1st Order
$$\begin{aligned} f(x+\Delta x) &\approx f(x) + (\Delta x)^T \nabla f(x) \geq f(x) \\ &\approx (\Delta x)^T \nabla f(x) \geq 0 \\ &\approx -(\nabla f)^T \nabla f(x) \geq 0 \\ &\approx ||\nabla f||^2 \leq 0 \to ||\nabla f|| = 0 \to \nabla f = 0 \end{aligned}$$
f has a minima at $x \to \nabla f(x) = 0$

2nd order
$f(x+\Delta x) \approx f(x) + \frac{1}{2}(\Delta x)^T H \Delta x \geq f(x)$
$(\Delta x)^T H \Delta x \geq 0 \to H = \frac{d^2 f}{dx^2} \geq 0 \to H$ is PSD

Gradient Descent
$x_{\text{new}} = x_{\text{old}} - \eta \nabla f(x_{\text{old}})$
$L(w):\ w_{\text{new}} = w_{\text{old}} - \eta \nabla f(w_{\text{old}})$
$\eta =$ learning rate (too high $\to$ divergence, too low $\to$ slow convergence)

$x_{t+1} = x_t - \eta \nabla f(x_t)$
$\frac{\Delta x}{\Delta t} = -\eta \nabla f(x_t) \approx \frac{dx}{dy}$
$x_t \to$ exponential decay

$d = -f \nabla f(x) \to \langle -\nabla f(x), \nabla f(x) \rangle = -\|\nabla f(x)\|^2$
$x_{t+1} = x_t + \alpha d$
$x_{t+1} = x_t - \alpha \nabla f(x_t)$

Newton's Method (2nd Order)
$$\begin{aligned} F(x+\Delta x) &= F(x) + (\Delta x)^T \nabla f + \tfrac{1}{2} \Delta x^T H \Delta x \text{ (gradient goes towards 0)} \\ &= F(x) + \tfrac{1}{2} \Delta x^T H \Delta x \text{ (quadratic form)} \end{aligned}$$
$\nabla F(x+\Delta x) = \nabla F(x) + H\Delta x$
$0 = \nabla F(x_k) + H(x_{k+1} - x_k)$
$H(x_{k+1} - x_k) = -\nabla F(x_k)$
$x_{k+1} - x_k = -H^{-1}(\nabla F(x_k))$
$x_{k+1} = x_k - H^{-1}(\nabla F(x_k))$

pro: quadratic convergence (faster), compared to linear
con: computationally expensive