$\min L = -\frac{1}{N}\sum_{i=1}^{n}\left(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right)$

$\hat{y}_i = \sigma(w^T x_i + b)$ — convex, differentiable

$a_i^{(i)} = \sigma(w^T x_i + b)$

Algorithm : Gradient Descent

Batch Gradient Descent $\quad A \Rightarrow \hat{Y}$

1) Compute gradients:

$$\frac{\partial L}{\partial w} = \frac{1}{N} X (A - Y)^T, \quad \frac{\partial L}{\partial b} = \frac{1}{N}\sum_{i=1}^{n}\left(a_i^{(i)} - y_i^{(i)}\right)$$

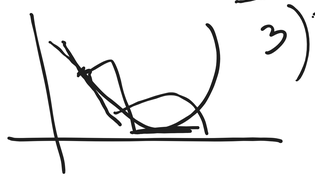
2) Update Parameters (learning rate)

$$w_{t+1} = w_t - \eta \frac{\partial L}{\partial w}$$

$$b_{t+1} = b_t - \eta \frac{\partial L}{\partial b}$$

$t=0$

$t=1$



→ No. of epochs (known domain)

→ Threshold

Vanilla Gradient Descent

$$w_{t+1} = w_t - \eta \frac{dL}{\partial w}$$

1) Batch (Entire Data)

2) Stochastic (One observation per update)

3) Mini-Batch $\quad 1 < B < N$

1) Momentum (SGD)

$$w_{t+1} = w_t - v_t$$

$$v_t = \gamma v_{t-1} + \eta \nabla L(w)$$

$0.9 < 1$