1 The hypothesis formula

According to the hypothesis interpretation, we have:

$$h_\theta(x) = S(x \cdot \theta) = P(y = 1|x, \theta)$$

where $P(y = 1|x, \theta)$ is the probability that $y = 1$ given the vector of features x, parameterized by θ.
Since the classification is binary, we can consider each result (label) as a ***Bernoulli Random Variable***: $Y \sim Ber(p)$ with $p = \theta \cdot x$. Using the **Bernoulli** distribution mass function, we rewrite:

$$h_\theta(x) = P(Y = y|X = x) = p^y \cdot (1-p)^{1-y} = (\sigma(\theta^T \cdot x))^y \cdot [1 - \sigma(\theta^T \cdot x)]^{1-y}$$

$$h_\theta(x) = (\sigma(\theta^T \cdot x))^y \cdot [1 - \sigma(\theta^T \cdot x)]^{1-y}$$

2 Maximum Likelihood Method

2.1 The Log-Likelihood

We can apply the Maximum Likelihood Method. As the training labels are independent, the likelihood mass function can be expressed as:

$$L(\theta) = \prod_{i=1}^{m} P(Y = y^i|X = x^i) = \prod_{i=1}^{m} (\sigma(\theta^T \cdot x))^y \cdot [1 - \sigma(\theta^T \cdot x)]^{1-y} \tag{1}$$

Applying the natural logarithm on both sides of 1:

$$LL(\theta) = \sum_{i=1}^{m} (y^i \cdot \log(\sigma(\theta^T \cdot x^i)) + (1 - y^i) \cdot \log(1 - \sigma(\theta^T \cdot x^i)) \tag{2}$$

2.2 Gradient Descent

2.2.1 Loss Function Derivative

We can now consider the cost function of the **Machine Learning** algorithm as

$$J(\theta) = -\tfrac{1}{m} \textstyle\sum_{i=1}^{m} (y^i \cdot \log(\sigma(\theta^T \cdot x^i)) + (1 - y^i) \cdot \log(1 - \sigma(\theta^T \cdot x^i))$$

the gradient descent can be used to minimize J. Therefore, we need to compute the derivative of J. As the derivative of sum is the sum of derivatives, let's consider the derivative of a single term in the sum.

$$\frac{\delta LL(\theta)}{\delta \theta_j} = \frac{\delta LL(\theta)}{\delta p} \cdot \frac{\delta p}{\delta \theta_j} \qquad \text{p denotes } \theta^T \cdot X$$

$$\frac{\delta LL(\theta)}{\delta \theta_j} = \frac{\delta LL(\theta)}{\delta p} \cdot \frac{\delta p}{\delta z} \cdot \frac{\delta z}{\theta_j} \qquad \text{where z denotes } \theta^T \cdot X$$

Let's consider each of these terms independently, starting with the first term.

$$\begin{aligned}
LL(\theta) &= y \log p + (1-y) \cdot \log(1-p) && \text{p} = \sigma(\theta^T \cdot X) \\
\frac{\delta LL(\theta)}{\delta p} &= \frac{y}{p} - \frac{1-y}{1-p} && \text{taking the derivative with respect to } p
\end{aligned}$$

As for the second term

$$\begin{aligned}
p &= \sigma(z) && \text{where z} = \theta^T \cdot X \\
\frac{\delta p}{\delta z} &= \sigma(z) \cdot [1 - \sigma(z)] && \text{taking the derivative of p with respect to z}
\end{aligned}$$

and the third term

$$\begin{aligned}
z &= \theta^T \cdot X && \text{as defined above} \\
\frac{\delta z}{\delta \theta_j} &= x_j && \text{as the only interaction is with } x_j
\end{aligned}$$

The derivative of a single term can be calculated as follows:

$$\begin{aligned}
\frac{\delta LL(\theta)}{\delta \theta_j} &= \frac{\delta LL(\theta)}{\delta p} \cdot \frac{\delta p}{\delta z} \cdot \frac{\delta z}{\theta_j} \\
&= [\frac{y}{p} - \frac{1-y}{1-p}] \cdot \sigma(z) \cdot [1-\sigma(z)] \cdot x_j && \text{substituting each term} \\
&= [\frac{y}{p} - \frac{1-y}{1-p}] \cdot p \cdot [1-p] \cdot x_j && \text{since p } = \sigma(z) \\
&= [y - \sigma(\theta^T x)] \cdot x_j && \text{expand and substitute p by } \sigma(\theta^T x)
\end{aligned}$$

the final form of the derivative can be expressed as:

$$\frac{\delta J(\theta)}{\delta \theta_j} = \frac{1}{m} \sum_{i=1}^{m} [\sigma(\theta^T x^i) - y^i] \cdot x_j \tag{3}$$

2.2.2 vectorized form

The algorithm can be expressed as follows
repeat until convergence {
    $\theta_i := \theta_i - \alpha \cdot \frac{dJ}{d\theta_i}$
    for $i = 0, 1, \ldots\, n$
}

having

$$\frac{\delta J}{\delta \theta_k} = X^T \cdot (\sigma(X \cdot \theta) - y) \tag{4}$$

where $\sigma(X \cdot \theta)$ is the resulting matrix of applying the sigmoid function element-wise to the matrix $(X \cdot \theta)$