Proof 1: Derivation of the cost function

Cost function: \[ J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \]

Optimization algorithm: \[ \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \]

Derivative of the cost function:
\[
\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j} (h_\theta(x^{(i)}) - y^{(i)})^2 \\
= \frac{1}{2m} \sum_{i=1}^{m} 2 (h_\theta(x^{(i)}) - y^{(i)}) \frac{\partial}{\partial \theta_j} (h_\theta(x^{(i)}) - y^{(i)}) \\
= \frac{1}{2m} \sum_{i=1}^{m} 2 (h_\theta(x^{(i)}) - y^{(i)}) \frac{\partial}{\partial \theta_j} \sum_{i=0}^{n} (\theta x_j - y) \\
= \frac{1}{2m} \sum_{i=1}^{m} 2 (h_\theta(x^{(i)}) - y) x_j
\]

Final update rule:
\[ \theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \]

Proof 2: Normal equations

Cost function: \[ J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \]

\[
J(\theta) = \frac{1}{2m} (h_\theta(x) - y)^T (h_\theta(x) - y) \\
J(\theta) = \frac{1}{2m} (X\theta - y)^T (X\theta - y) \\
J(\theta) = \frac{1}{2m} ((X\theta)^T X - y^T) (X\theta - y) \\
J(\theta) = \frac{1}{2m} (X\theta)^T X\theta - (X\theta)^T y + y^T X\theta + y^T y
\]

Note that \( X\theta \) is a vector, and so is \( y \). So when we multiply one by another, it doesn't matter what the order is (as long as the dimensions work out). So we can further simplify:

\[
J(\theta) = \frac{1}{2m} ((X\theta)^T X\theta - 2(X\theta)^T y + y^T y) \\
J(\theta) = \frac{1}{2m} ((\theta^T X^T X\theta - 2(X\theta)^T y + y^T y)
\]

We are trying to solve for \( \theta \) because it is unknown. As with basic calculus we will take derivative of the above equation with respect to theta and then we will set it equal to 0. That will give us the minimum. We derive by each component of the vector, and then combine the resulting derivatives into a vector again. The result is:

\[
\frac{\partial}{\partial \theta_0} = \frac{1}{2m} (2 X^T X\theta - 2 X^T y) = 0 \\
\frac{\partial}{\partial \theta_j} = X^T X\theta - X^T y = 0
\]
\[ X^T X \theta = X^T y \]

Hence, normal equation: \( \theta = (X^T X)^{-1} X^T y \)

**Proof 3: Loss of the logistic regression**

Loss: \( J(\theta) = -\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \)

Gradient if the loss:

\[
\frac{\partial J}{\partial \theta_j} = -\sum_{i=1}^{m} y^{(i)} \frac{\partial \log(h_\theta(x^{(i)}))}{\partial \theta_j} + (1 - y^{(i)}) \frac{\partial \log(1 - h_\theta(x^{(i)}))}{\partial \theta_j} \\
= \sum_{i=1}^{m} y^{(i)} \frac{\partial (h_\theta(x^{(i)}))}{\partial \theta_j} \frac{1}{h_\theta(x^{(i)})} + (1 - y^{(i)}) \frac{\partial (1 - h_\theta(x^{(i)}))}{\partial \theta_j} \frac{1}{1 - h_\theta(x^{(i)})} \\
= \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}
\]

**Proof 4: Scheme of the chain rule**

\[
\begin{align*}
Y &= f(X) \\
Z &= g(Y)
\end{align*}
\]

Local gradients:

\[
\begin{align*}
\frac{\partial Z}{\partial Y} &= \frac{\partial g(Y)}{\partial Y} = g'(Y) \\
\frac{\partial Y}{\partial X} &= \frac{\partial f(X)}{\partial X} = f'(X)
\end{align*}
\]

Chain rule gradient:

\[
\frac{\partial Z}{\partial X} = \frac{\partial Y}{\partial X} \frac{\partial Z}{\partial Y} = f'(X)g'(Y) = f'(X)g'(f(X))
\]

Application: backpropagation

**Proof 5: Bias - Variance**

We know that: \( Y = f + \epsilon \). We fit \( Y = \hat{f} \). 
Here $\hat{f}$ is our estimator of $f$ based on our data. Note that $\hat{f}$ is a random variable because it is trained on the data which itself is a random variable.

We assume that $\epsilon$ is such that $\epsilon$ is independent from $\hat{f}$ with mean 0 and variance $\sigma^2$.

Furthermore, $f$ is deterministic. Therefore $E[f] = f$, $Var[f] = 0$.

Note: we have the following fact: $\forall X$ random variable, $Var[X] = E[X^2] - E[X]^2$.

Therefore, the error is:

$$E[\|Y - \hat{f}\|^2] = E[\|f - \hat{f} + \epsilon\|^2] = E[\|f - \hat{f}\|^2] + 2E[(f - \hat{f})\epsilon] + E[\epsilon^2]$$

$$= Var[f - \hat{f}] + E[(f - \hat{f})^2] + 2E[(f - \hat{f})E[\epsilon]] + E[\epsilon^2] \text{ because } \epsilon \text{ is independent from } \hat{f}$$

$$= Var[\hat{f}] + Bias^2 + \sigma^2 \text{ because } f \text{ is deterministic and } \epsilon \text{ has zero mean}$$