Homework 3

Part 1 :

Gradient calculation :

Sigmoid function

Given $f(w) = \dfrac{1}{1+e^{-w^T x}}$ since x and w are vectors.

$$\frac{\partial f}{\partial w_i} = \frac{0 \cdot (1+e^{-w^T x}) - 1 \cdot \frac{\partial}{\partial w_i}(1+e^{-w^T x})}{(1+e^{-w^T x})^2}$$

$$= \frac{-\frac{\partial}{\partial w_i}\left(1+e^{-w^T x}\right)}{(1+e^{-w^T x})^2}$$

$$= \frac{e^{-w^T x} \cdot \frac{\partial}{\partial w_i}\left(-w^T x\right)}{(1+e^{-w^T x})^2}$$

$$= \frac{(e^{-w^T x})\, x_i}{(1+e^{-w^T x})^2} \quad //$$

$$= f(w)\left[1 - f(w)\right] x_i$$

Logistic Loss Function:

$$F(w) = \log(1+e^{-yxw})$$

$$\frac{\partial F(w)}{\partial w} = \frac{1}{1+e^{-yxw}} \cdot (-yx)\, e^{-yxw}$$

$$= -yx \times \frac{e^{-yxw}}{1+e^{-yxw}}$$

$$= yx\left(1 - \frac{1}{1+e^{-yxw}}\right)$$

Part 2:

1. $F(w) = \frac{1}{2} \| y - Xw \|_2^2$

Gradient of the given function:

$$\nabla_w f(w) = \frac{2}{2} (-X^T)(y - Xw)$$

Calculating the gradient

$$\nabla_w f(w) = -X^T y + X^T X w$$

is the required gradient.

$$Hf(x) = \nabla_w^2 f(w) = X^T X$$

From the above question we can see that if we have a vector V such that V is the eigen vector vector of X

Then we can write,

$$V^T X^T X V = (XV)^T XV \geq 0$$

which means that $X^T X \geq 0$

And if the Hessian is positive semi definite it is convex.

2. We stick with least square formulation because according to Gauss-Markov's theorem the least squared estimator has the smallest variance among all linear unbiased estimator.
Futhermore we know that the least squared estimator is similar to the maximum likelihood estimator (MLE).
If we assume the linear model to be normally distributed error, we notice that the MLE never explicitly introduces squared error.

3. For $F(w)$ to be strongly convex $X^T X$ should be a positive definite matrix
ie. X should be a full rank matrix.