1. Derive the updates for gradient descent applied to $L_2$-regularized logistic loss. Describe how this update compares to $L_2$-regularized hinge-loss and exponential loss.

Take derivatives with respect to $w_i$ and $b$. Consider the logistic loss function for a fixed example $x_n$. It is easiest to take derivatives by using the chain rule. First, note that, for a fixed example $x_n$:

$$\hat{y}_n = w \cdot x_n + b = \sum_{i=1}^{D} w_i x_{ni} + b$$

$$\frac{\partial \hat{y}_n}{\partial w_i} = x_{ni} \quad \frac{\partial \hat{y}_n}{\partial b} = 1$$

Taking derivatives of logistic loss function, for a fixed example $x_n$, and applying the chain rule:

$$l(y_n, \hat{y}_n) = \frac{1}{\log 2} \log [1 + \exp (-y_n \hat{y}_n)]$$

$$\frac{\partial l(y_n)}{\partial \hat{y}_n} = \frac{1}{\log 2} \frac{-y_n \exp (-y_n \hat{y}_n)}{1 + \exp (-y_n \hat{y}_n)} = \frac{-y_n \sigma (-y_n \hat{y}_n)}{\log 2}$$

$$\frac{\partial l(y_n, \hat{y}_n)}{\partial w_i} = \frac{\partial l(y_n, \hat{y}_n)}{\partial \hat{y}_n} \frac{\partial \hat{y}_n}{\partial w_i} = \frac{-y_n x_{ni} \sigma (-y_n \hat{y}_n)}{\log 2}$$

$$\frac{\partial l(y_n, \hat{y}_n)}{\partial b} = \frac{\partial l(y_n, \hat{y}_n)}{\partial \hat{y}_n} \frac{\partial \hat{y}_n}{\partial b} = \frac{-y_n \sigma (-y_n \hat{y}_n)}{\log 2}$$

Finally, forming the $L_2$ regularized logistic loss function, and forming the gradient using the derivatives above:

$$\mathcal{L}(w, b) = \sum_{n=1}^{N} l(y_n, \hat{y}_n) + \frac{\lambda}{2} ||w||^2$$

$$\nabla_w \mathcal{L} = \sum_{n=1}^{N} \frac{-y_n x_n \sigma (-y_n \hat{y}_n)}{\log 2} + \lambda w$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{n=1}^{N} \frac{-y_n \sigma (-y_n \hat{y}_n)}{\log 2}$$
We use the gradients to form the updates for gradient descent:

\[ w \leftarrow w + \eta \sum_{n=1}^{N} y_n x_n \sigma(-y_n \hat{y}_n) \frac{1}{\log 2} - \eta \lambda w \]

\[ b \leftarrow b + \eta \sum_{n=1}^{N} y_n \sigma(-y_n \hat{y}_n) \frac{1}{\log 2} \]

The updates for gradient descent with \( L_2 \)-regularized hinge loss for logistic loss, hinge-loss, and exponential loss all have the form:

\[ w \leftarrow w + \eta \sum_{n=1}^{N} c_n y_n x_n - \eta \lambda w \]

\[ b \leftarrow b + \eta \sum_{n=1}^{N} c_n y_n \]

where the weight \( c_n \) given to data point \( n \) depends on the loss function:

\[ c_n = \frac{\sigma(-y_n \hat{y}_n)}{\log 2} \text{ for logistic loss} \]

\[ c_n = 1(y_n \hat{y}_n < 1) \text{ for hinge loss} \]

\[ c_n = \exp(-y_n \hat{y}_n) \text{ for exponential loss} \]

2. Derive the updates for (sub)-gradient descent applied to \( L_1 \)-regularized hinge loss. First, you need to compute the subgradient for the \( L_1 \)-norm. Second you derive the updates. Third, discuss the effect of \( \lambda \) on the update. How does it compare to the update with \( L_2 \)-regularization?

Compute the subgradient of the \( L_1 \) norm:

\[ ||w|| = \sum_{n=1}^{N} |w_i| \]

\[ \frac{\partial |w_i|}{\partial w_i} = \mathbb{1}(w_i > 0) - \mathbb{1}(w_i < 0) \quad (i \text{th component of the subgradient}) \]

The subgradient for hinge loss:

\[ l(y, \hat{y}) = \max\{0, 1 - y \hat{y}\} \]

\[ \frac{\partial l}{\partial \hat{y}} = -y \mathbb{1}(y \hat{y} < 1) \]
Using the gradients to construct the updates:

\[
    w_i \leftarrow w_i + \eta \sum_{n=1}^{N} y_n \mathbb{1}(y_n \hat{y}_n < 1) - \eta \lambda [\mathbb{1}(w_i > 0) - \mathbb{1}(w_i < 0)]
\]

\[
    b \leftarrow b + \eta \sum_{n=1}^{N} y_n \mathbb{1}(y_n \hat{y}_n < 1)
\]

All components of the weight vector \( w \) are adjusted towards zero with equal step \( \lambda \eta \) regardless of whether the sign of \( \hat{y}_n \) is correct. \( \lambda \) has the effect of reducing the norm of the weight vector. For hinge loss with \( L2 \)-regularization, each component of \( w \) is instead adjusted by a step proportional to the component's magnitude.

3. What is the relationship between negative log-likelihood of a conditional Bernoulli model and the logistic loss function?

They are the same, up to a scaling factor. See section 7.6 of CIML, which points out that the logistic loss function was originally developed from a conditional Bernoulli generative model.