Question 1

For more details, please see Murphy 8.3.6

In logistic regression, we prefer MAP estimation to MLE estimation. We may overfit if we use MLE, placing too much probability mass on our training data and creating solutions that do not generalise well to test data.

Even in data rich settings, we prefer regularised models (in similar vein to regularising linear regression models with methods such as ridge regression, lasso, elastic net etc). Optimising with respect to MLE may lead to brittle solutions that do not generalise well. Murphy explains in his book:

Suppose we have linearly separable data. MLE is obtained when $\|\mathbf{w}\| \to \infty$, which is an infinitely steep sigmoid function $\mathbf{w}^2\mathbf{x} > \mathbf{w}_0 \cdots$ linear threshold unit).

This assigns maximal amount of probability mass to the training data, and accordingly will overfit and not generalise well. To prevent this, we can use l2 regularisation.

Question 2

Since the question asks to add Gaussian noise to each data point, let $x_{ij} = x_{ij} + \epsilon_{ij}$ where i is the data point and j is the feature of that data point. Therefore

$$y_i(x, w) = w_o + \sum_{j=1}^{D} w_j(x_{ij} + \epsilon_{ij}) \tag{16}$$

Substituting that into the cost function,

$$\begin{aligned} E_D(w) &= \frac{1}{2}\sum_{i=1}^{N}\Big[w_o + \sum_{j=1}^{D} w_j(x_{ij}+\epsilon_{ij}) - t_i\Big]^2 \\ &= \frac{1}{2}\sum_{i=1}^{N}\Big[w_o + \sum_{j=1}^{D} w_j x_{ij} - t_i + \sum_{j=1}^{D} w_j\epsilon_{ij}\Big]^2 \end{aligned}$$

Let $A_i = w_o + \sum_{j=1}^{D} w_j x_{ij} - t_i$ and $B_i = \sum_{j=1}^{D} w_j\epsilon_{ij}$

$$E_D(w) = \frac{1}{2}\sum_{i=1}^{N}\Big[A_i + B_i\Big]^2 \tag{17}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\Big[A_i^2 + 2A_iB_i + B_i^2\Big] \tag{18}$$

$$= \frac{1}{2}\sum_{i=1}^{N} A_i^2 + \sum_{i=1}^{N} A_iB_i + \frac{1}{2}\sum_{i=1}^{N} B_i^2 \tag{19}$$

Taking the expectation of the cost function

$$\begin{aligned} \mathbb{E}[E_D(w)] &= \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{N} A_i^2 + \sum_{i=1}^{N} A_iB_i + \frac{1}{2}\sum_{i=1}^{N} B_i^2\right] \\ &= \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{N} A_i^2\right] + \mathbb{E}\left[\sum_{i=1}^{N} A_iB_i\right] + \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{N} B_i^2\right] \end{aligned}$$

Now we need to simplify each of these expectation terms. Looking at each of the individual components

$$\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{N} A_i^2\right] = \frac{1}{2}\sum_{i=1}^{N}\Big[(w_o + \sum_{j=1}^{D} w_j x_{ij} - t_i)^2\Big]$$

Since each of the terms w x and t are fixed w.r.t to the cost function. Note that this is our original cost function.

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{2}\sum_{i=1}^{N} 2A_iB_i\Big] &= \mathbb{E}\Big[\sum_{i=1}^{N} A_iB_i\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{N} A_i \sum_{j=1}^{D} w_j\epsilon_{ij}\Big] \\
&= \sum_{i=1}^{N} A_i \sum_{j=1}^{D} w_j \mathbb{E}\Big[\epsilon_{ij}\Big] \\
&= 0
\end{aligned}$$

since $\mathbb{E}\Big[\epsilon_{ij}\Big] = 0$

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{N} B_i^2 &= \frac{1}{2}\sum_{i=1}^{N}\Big[\sum_{j=1}^{D} w_j\epsilon_{ij}\Big]^2 \\
&= \frac{1}{2}\sum_{i=1}^{N}\Big[\sum_{j=1}^{D}\sum_{k=1}^{D} w_jw_k\epsilon_{ij}\epsilon_{ik}\Big] \\
&= \frac{1}{2}\sum_{i=1}^{N}\Big[\sum_{j=1}^{D} w_j^2\sigma^2\Big] \\
&= \frac{1}{2}\sigma^2\sum_{j=1}^{D} w_j^2
\end{aligned}$$

We take the expectation to get to the third line since.

Bringing it all back together, the expected loss function is given as

$$\mathbb{E}\Big[E_D\Big] = \frac{1}{2}\sum_{i=1}^{N}\Big[(w_o + \sum_{j=1}^{D} w_jx_{ij} - t_i)^2\Big] + \frac{N}{2}\sigma^2\sum_{j=1}^{D} w_j^2 \tag{20}$$

You may have noticed that the regulariser term increases with the number of data points. This is not true. Generally the amount of regularisation should decrease as N increases.

To remove the dependence of N on the regularisation term, in this situation, we can instead minimise the mean of squares error instead of the sum of squares error, so the loss becomes

$$E_D(w) = \frac{1}{2N}\sum_{i=1}^{N}(y(x_i, w) - t_i)^2 \tag{21}$$

i.e. we divide by a factor of N. By doing this, we make the regularisation term independent of the number of data points.

$$\mathbb{E}\Big[E_D\Big] = \frac{1}{2N}\sum_{i=1}^{N}\Big[(w_o + \sum_{j=1}^{D} w_jx_{ij} - t_i)^2\Big] + \frac{1}{2}\sigma^2\sum_{j=1}^{D} w_j^2 \tag{22}$$