2b)

$$AA^T_{i,j} = \sum_{j=0}^{M}\sum_{k=0}^{N} A_{i,k}\,A_{j,k}$$

3a) $l(w_0, w) = \Sigma [t^{(n)} - y(x^{(n)})]^2$
$= \Sigma [y(x^{(n)}) - t^{(n)}]^2$
$= \Sigma [y^{(n)} - t^{(n)}]^2$

$y^{(n)}$ & $t^{(n)}$ can be written as vectors $\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \vec{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}$

Also, the summation can be written as the magnitude.

Thus,

$\Sigma [y^{(n)} - t^{(n)}]^2 = || y - t ||^2$, as required ☒

3b) $y(x) = w_0 + \Sigma w_m z_m$

$\Sigma w_m z_m$ can be written as Zw, where Z is a feature matrix with m features and n columns.

w can be written as a column vector, $w = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

Zw would produce an $m \times 1$ matrix.

so, $\Sigma w_m z_m = Zw$

To add w_0 to Zw we must turn w_0 into a column matrix with m rows. We can do this by multiplying w_0 by a column vector of 1's.

Thus,

$y = w_0 \vec{1} + Zw$, as required ☒

3c) $l(w_0, w) = \Sigma [t^{(n)} - y(x^{(n)})]^2 = \Sigma [y^{(n)} - t^{(n)}]^2$

$\frac{\partial l(w_0, w)}{\partial w} = 2 \Sigma [y^{(n)} - t^{(n)}] \cdot \frac{\partial}{\partial w} [y^{(n)} - t^{(n)}]$
$= 2(y - t) \cdot \frac{\partial}{\partial w} [(w_0 + \Sigma w_m z_m] - t^{(n)}]$
$= 2(y-t) \cdot \frac{\partial}{\partial w} (Zw)$
$= 2(y-t) \cdot Z^T$
$= 2$

3d) $l(w_0, w) = \sum [y^{(n)} - t^{(n)}]^2$

$$\frac{dl(w_0, w)}{dw_0} = 2 \sum [y^{(n)} - t^{(n)}] \cdot \frac{d}{dw_0} (y^{(n)} - \cancelto{0}{t^{(n)}})$$

$$= 2 \sum [y^{(n)} - t^{(n)}] \frac{d}{dw_0} [w_0 + \sum \cancelto{0}{z_m w_m}]$$

$$= 2 \sum [y^{(n)} - t^{(n)}] \cdot 1$$

$$= 2 \cdot \vec{1}^T (y - t)$$

Note: $\frac{d}{dw_0} [w_0 + \sum z_m w_m]$
$= 1$

4e)
The testing error is larger than the training error because the model is poor in general, due to the lack of basis functions. Although it is poor, it fits the training data much better than the test data.

4f)
The model begins to slowly overfit the training data, thus resulting in a large difference between the training error and testing error.

4g)
Just by looking at the graph, we can see that the model is highly overfit. The model goes through every point of the training data, instead of generalizing the data. The large delta between the training and testing error is a quantification of the figure.

5d)
It can be seen that the smallest gamma fails to regularize the model enough, and hence we see many inflection points. On the other hand, the largest gamma value regularizes the model too much and doesn’t have enough variability and thus turns the model into a parabola, which is inaccurate to our data. The optimal gamma regularizes the model just enough to take the general structure of how the points are laid out.

6a) $\tilde{L}(w_0, w) = L(w_0, w) + \gamma \sum w_j^2$

$$\frac{\partial \tilde{L}(w_0, w)}{\partial w} = \frac{\partial l}{\partial w} + \frac{\partial}{\partial w}\left(\lambda \sum_{j=1}^{n} w_j^2\right)$$

$$= 2Z^T(y-t) + 2\lambda \sum_{i=1}^{n} w_j$$

$$= 2\left[Z^T(y-t) + \lambda \sum_{i=1}^{n} w_j\right]$$

6b) $$\frac{\partial \tilde{L}(w_0, w)}{\partial w_0} = \frac{\partial L}{\partial w_0} + \cancel{\frac{\partial}{\partial w_0}\left(\lambda \sum_{i=1}^{n} w_j^2\right)}$$

$$= \frac{\partial L}{\partial w_0}$$

$$= 2\vec{1}^T(y-t)$$