Question 2

2(b)

The equation for computing AA^T without computing ^{AT} is:

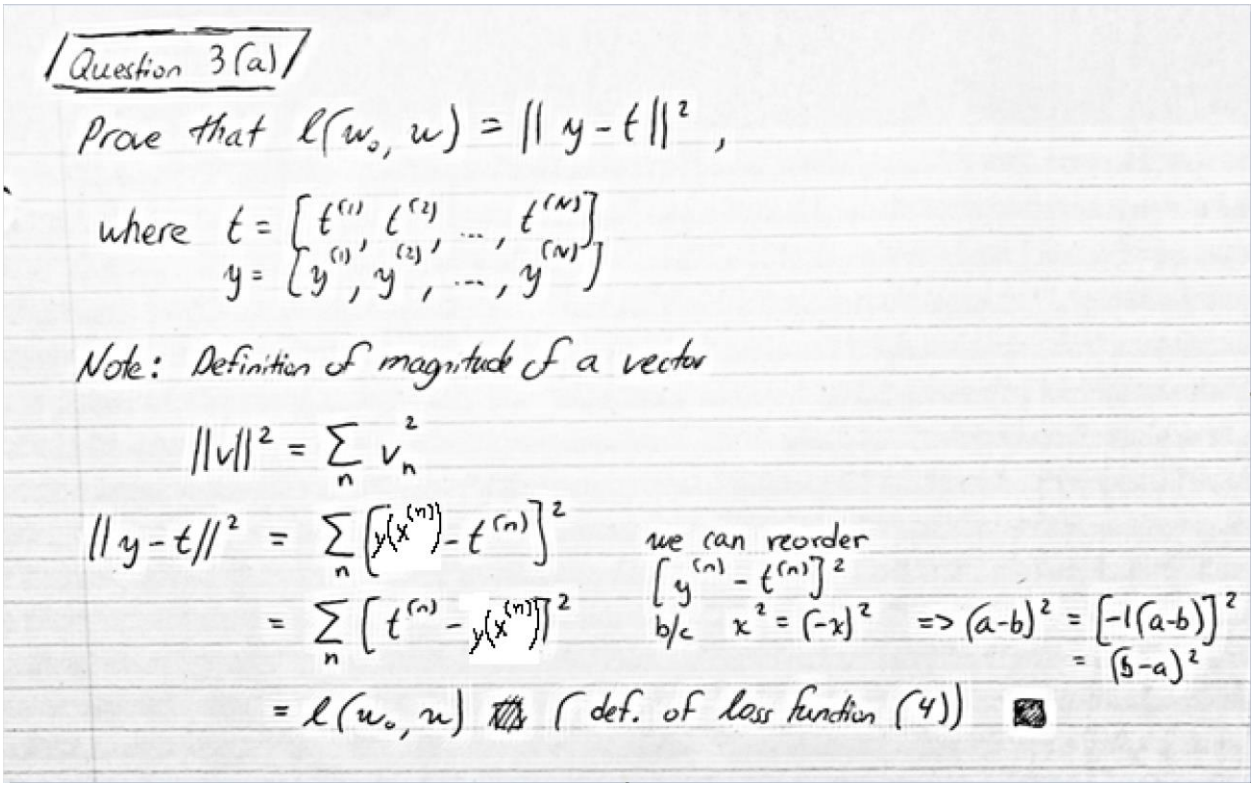
$$\mathrm{AA}^{\mathrm{T}}_{ij} = \sum_k \mathrm{A}_{ik}\, \mathrm{A}_{jk}$$

2(d)

myfun performs MN^2 floating-point multiplications and $MN(N-1)$ floating-point additions.

Question 3

3(a)



3(b)

Question 3(b)

Prove that $y = w_0\vec{1} + Zw$,

where y and w are treated as column vectors

$$w_0\vec{1} + Zw = \begin{bmatrix} \phi_1(x^{(1)}) & \phi_2(x^{(1)}) & \cdots & \phi_M(x^{(1)}) \\ \phi_1(x^{(2)}) & & & \\ \vdots & & \ddots & \\ \phi_1(x^{(N)}) & & & \phi_M(x^{(N)}) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix} + w_0\vec{1}$$

$$= \begin{bmatrix} w_1\phi_1(x^{(1)}) + w_2\phi_2(x^{(1)}) + \ldots + w_M\phi_M(x^{(1)}) \\ w_1\phi_1(x^{(2)}) + w_2\phi_2(x^{(2)}) + \ldots + w_M\phi_M(x^{(2)}) \\ \vdots \\ w_1\phi_1(x^{(N)}) + w_2\phi_2(x^{(N)}) + \ldots + w_M\phi_M(x^{(N)}) \end{bmatrix} + w_0\vec{1}$$

let a_i represent row i of Zw where $i \in [1, N]$

$$= \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} + \begin{bmatrix} w_0 \\ w_0 \\ \vdots \\ w_0 \end{bmatrix} = \begin{bmatrix} w_0 + \sum_{m=1}^{M} w_m z_m \\ w_0 + \sum_{m=1}^{M} w_m z_m \\ \vdots \\ w_0 + \sum_{m=1}^{M} w_m z_m \end{bmatrix}$$ (by eqn. #5)

where $x = x^{(1)}$

where $x = x^{(2)}$

where $x = x^{(N)}$

$$= \begin{bmatrix} y(x^{(1)}) \\ y(x^{(2)}) \\ \vdots \\ y(x^{(N)}) \end{bmatrix} = y$$ ∎

Note: $y(x) = w_0 + \sum_{m=1}^{M} w_m z_m$ (5)

3(c)

Question 3(c)

Prove that $\frac{\partial \ell(w_0, w)}{\partial w} = 2Z^T(y-t)$

$$\frac{\partial \ell(w_0, w)}{\partial w} = \frac{\partial \sum_{n=1}^{N} \left[y(x^{(n)}) - t^{(n)} \right]^2}{\partial w}$$

$$= \sum_{n=1}^{N} \frac{\partial \left[y(x^{(n)}) - t^{(n)} \right]^2}{\partial w}$$

$$= \sum_{n=1}^{N} 2\left(y(x^{(n)}) - t^{(n)} \right) \cdot \frac{\partial \left(y(x^{(n)}) - t^{(n)} \right)}{\partial w}$$

$$= 2 \sum_{n=1}^{N} \left(y(x^{(n)}) - t^{(n)} \right) \cdot \frac{\partial \left(\sum_{m=1}^{M} w_m z_m - t^{(n)} \right)}{\partial w}$$

$$= 2 \sum_{n=1}^{N} \left(y(x^{(n)}) - t^{(n)} \right) \cdot \phi_m(x^{(n)})$$

$$= 2 \sum_{n=1}^{N} \left(y(x^{(n)}) - t^{(n)} \right) \cdot z_{nm}$$

$$= 2 \begin{bmatrix} \sum_{n=1}^{N} \left(y(x^{(n)}) - t^{(n)} \right) z_{n0} \\ \vdots \\ \sum_{n=1}^{N} \left(y(x^{(n)}) - t^{(n)} \right) z_{nM} \end{bmatrix}$$

$$= 2 \begin{bmatrix} \sum_{n=1}^{N} \left(y(x^{(n)}) - t^{(n)} \right) \phi_0(x^{(n)}) \\ \vdots \\ \sum \left(y(x^{(n)}) - t^{(n)} \right) \phi_M(x^{(n)}) \end{bmatrix}$$

Note: $\phi_0(x) = 1$

$$= 2 \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ \phi_M(x^{(1)}) & \cdots & \phi_M(x^{(n)}) \end{bmatrix} \cdot \begin{bmatrix} y(x^{(1)}) - t^{(1)} \\ \vdots \\ y(x^{(n)}) - t^{(n)} \end{bmatrix}$$

$$= 2Z^T(y-t) \quad \blacksquare$$

3(d)

Question 3(d)

Prove that $\frac{\partial \ell(w_0, w)}{\partial w_0} = 2\vec{1}^T(y-t)$

$$\frac{\partial \ell(w_0, w)}{\partial w_0} = \frac{\partial \sum_{n=1}^{N} [y(x^{(n)}) - t^{(n)}]^2}{\partial w_0}$$

$$= \sum_{n=1}^{N} \frac{\partial [y(x^{(n)}) - t^{(n)}]^2}{\partial w_0}$$

$$= \sum_{n=1}^{N} 2(y(x^{(n)}) - t^{(n)}) \cdot \frac{\partial (y(x^{(n)}) - t^{(n)})}{\partial w_0}$$

$$= 2\sum_{n=1}^{N} (y(x^{(n)}) - t^{(n)}) \cdot \frac{\partial \left(\sum_{m=0}^{M} w_m z_m - t^{(n)}\right)}{\partial w_0}$$

$$= 2\sum_{n=1}^{N} (y(x^{(n)}) - t^{(n)}) \cdot \phi_0(x^{(n)})$$

$$= 2\sum_{n=1}^{N} (y(x^{(n)}) - t^{(n)}) \cdot z_{n0}$$

Note: $\phi_0(x) = 1$

$$= 2\sum_{n=1}^{N} (y(x^{(n)}) - t^{(n)}) \phi_0(x^{(n)})$$

$$= 2 \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \cdot \begin{bmatrix} y(x^{(1)}) - t^{(1)} \\ \vdots \\ y(x^{(n)}) - t^{(n)} \end{bmatrix}$$

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

Question 4

4(e)

The test error is larger than the training error because the model was trained to fit the points in the training data to minimize specifically this training error. In other words, a fitted model adapts to the training data while the test data is new and it will have more difficulty predicting it.

4(f)

The model is starting to overfit the training data, which is evident by the training error being much lower than the test error. In other words, the model has adapted to the training data specifically, so much that it does not generalize well for other inputs.

4(g)

The model is now a perfect fit for the training data, which is a bad thing if you want to use it to predict anything other than the training data. The training error is very low because the plot of the model passes through every point in the training data. The test error is very high because the model is perfectly adapted to the training data and cannot predict the new inputs of the test data.

5(d)

The curve with the weights for the smallest gamma looks like a wavelet. While it has many weights that are very negative, there are also many weights that are very large so it does not limit the effective model complexity much. This means that many weights decay to zero but many are also scaled up by a large amount. This leads to the corresponding plot overfitting because its model complexity is not limited.

The curve with the optimal weights looks similar to the curve in 5(b), which has 19 basis functions and a gamma of 10^{-9}. The optimal choice of gamma decays the weights that are not supported by the data and effectively limits the model complexity. This leads to the corresponding plot fitting nicely.

The curve with the weights for the largest gamma looks like a parabola that has been flipped over the x-axis. Its weights are very small and this limits the effective model complexity to being linear like. This leads to the corresponding plot underfitting.

Question 6

6(a)

Question 6(a)

Derive $\frac{\partial \hat{L}(w_0, w)}{\partial w}$

$$= \frac{\partial L(w_0, w)}{\partial w} + \frac{\partial \left(\gamma \sum_{j=1}^{M} w_j^2\right)}{\partial w} \quad \text{(by (6))}$$

$$\propto 2Z^T(y-t) + \gamma 2w \quad \text{(by 3(c) and by Aside)}$$

Aside:

$$\frac{\partial \left(\gamma \sum_{j=1}^{M} w_j^2\right)}{\partial w_j} = \gamma 2 w_i$$

$$\Rightarrow \frac{\partial \left(\gamma \sum_{j=1}^{M} w_j^2\right)}{\partial w} = \gamma 2 \vec{w}$$

6(b)

Question 6(b)

Derive $\frac{\partial \tilde{\ell}(w_0, w)}{\partial w_0}$

$$= \frac{\partial \ell(w_0, w_*)}{\partial w_0} + \frac{\partial \left(\gamma \sum_{j=1}^{M} w_j^2\right)}{\partial w_0} \quad \text{(by (6))}$$

$$= 2\vec{1}^T(y-t) + \gamma \frac{\partial \sum_{j=1}^{M} w_j^2}{\partial w_0} \quad \text{(by 3(b'))}$$

$$= 2\vec{1}^T(y-t) + 0 \quad \text{(by Aside)}$$

Aside:

$$\gamma \frac{\partial \sum_{j=1}^{M} w_j^2}{\partial w_0}$$

$$= \gamma \frac{\partial (w_1 + w_2 + \ldots + w_M)}{\partial w_0}$$

$$= \gamma(0)$$

$$= 0$$