Solution to Written Assignment 2
SENG 474/CSC 578D

Question 1

(a) $CountMap$ vector on the first pass looks like this: $[0, 5, 5, 3, 6, 1, 3, 2, 6, 3, 2]$. Frequent buckets are those with $CountMap[h] \geq 4$, that is buckets in $\{1, 2, 4, 8\}$.

(b) The pairs that are counted in the second pass are those that are hashed to frequent buckets. $\{1, 2\}, \{1, 4\}, \{2, 4\}, \{2, 6\}, \{3, 4\}, \{3, 5\}, \{4, 6\}, \{5, 6\}$

Question 2

For simplicity, we ignore the memory needed to hold the CountMap in the first pass. Observe that:

- The memory needed to hold all frequent items is $25 * 10^4 * 4 = 10^6$ bytes.
- The memory needed to count the support of truly frequent item pair in 2nd pass is 12×10^6 bytes, since for each item pair, we need 8 bytes to hold the pair and 4 bytes for the counter.

Assume that we construct a BitMap of B bits in the 1st pass. There are two extremes:

- If $B = 10^6$, every location of the BitMap would be 1, thus, in the 2nd pass, we need to count the support for all P pairs, that takes $12P$ bytes of memory. The total memory used by the algorithm is $13 \times 10^6 + 10^6/4 + 12P$ bytes
- If $B = 10^6 + P$, every location of the BitMap, except for those of truly frequent itempairs, would be 0, thus, in the 2nd pass, we don't need to count the support for all P pairs. The total memory used by the algorithm is $13 \times 10^6 + (10^6 + P)/4$ bytes.

So it does seem like we need $13 \times 10^6 + (10^6 + P)/4$ bytes of memory as in the second case. However, we can do much better than that.

In the following, we show how to determine B so that the total memory of the algorithm is minimum. Let $B = 10^6 \times b$ and $P = 10^6 \times p$ for some positive numbers b, p. Note that $b \geq 1$ since $B \geq 10^6$. Since there are $10^6 + P = 10^6(p + 1)$ candidate pairs, the number of candidate pairs hold by each location in the CountMap is:

$$\frac{10^6(p+1)}{B} = \frac{p+1}{b} \tag{1}$$

Note that since $P = 10^6 p < (25 * 10^4)^2/2$, we have $p < 3.125 \times 10^4$. Thus, $\frac{p+1}{b} \leq p + 1 < 25 \times 10^4$. Hence, except for 10^6 locations of the BitMap corresponding to truly frequent itempairs, other

locations are set to 0. Hence, the number of candidate pairs we need to count the support in 2nd phase is:

$$\begin{aligned} 10^6 + P - (\frac{p+1}{b})(B - 10^6) &= 10^6 + 10^6 p - (\frac{p+1}{b})(b-1)10^6 \\ &= 10^6 \frac{p+1}{b} \end{aligned} \tag{2}$$

which costs $12 \times 10^6 \frac{p+1}{b} = 12 \times 10^6 \frac{p+1}{b}$ bytes. Hence, the total memory (in bytes) of the algorithm is:

$$13 \times 10^6 + 10^6 b/4 + 12 \times 10^6 \frac{p+1}{b} = 10^6 (13 + \frac{b}{4} + 12\frac{p+1}{b}) \tag{3}$$

Obviously, we should choose b to minimize $\frac{b}{4} + \frac{12(p+1)}{b}$. Observe that $\frac{b}{4} + \frac{12(p+1)}{b}$ is minimum when $b = 2\sqrt{12(p+1)}$ and the minimum memory (in bytes) is:

$$10^6 (13 + \sqrt{12(p+1)}) \tag{4}$$

However, note that $B < 10^6 + P$, since otherwise, we can go with the solution in the second extreme case above. This implies:

$$10^6 b < 10^6 + 10^6 p \tag{5}$$

that implies $2\sqrt{12(p+1)} < (p+1)$. Solving this inequality, we get $p > 47$. In conclusion, the minimum memory is:

$$S = \begin{cases} 13 \times 10^6 + \frac{10^6 + P}{4} & \text{if } P \leq 47 \times 10^6 \\ 10^6 (13 + \sqrt{12(\frac{P}{10^6} + 1)}) & \text{otherwise} \end{cases}$$

Question 3

By cyclic property, we have $Tr(CAB) = Tr(ABC)$.

$$BC = \begin{bmatrix} -1 & 2 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 8 & -2 \end{bmatrix}$$

It immediately follows:

$$ABC = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 8 & -2 \end{bmatrix} = \begin{bmatrix} 2a_{11} + 8a_{12} & 3a_{11} - 2a_{12} \\ 2a_{12} + 8a_{22} & 3a_{21} - 2a_{22} \end{bmatrix}$$

So $Tr(ABC) = 2a_{11} + 8a_{12} + 3a_{21} - 2a_{22}$. That implies:

$$\nabla_A Tr(ABC) = \begin{bmatrix} 2 & 8 \\ 3 & -2 \end{bmatrix}$$

Another way to compute $\nabla_A Tr(ABC)$ is using the fact that $\nabla_A Tr(AB) = B^T$, we have: $\nabla_A Tr(CAB) = (BC)^T = C^T B^T = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 8 & -2 \end{bmatrix} = \begin{bmatrix} 2 & 8 \\ 3 & -2 \end{bmatrix}$.

Question 4

(a) We want to prove that $Tr(AB) = Tr(BA)$. Using the definition of trace, for two $n \times n$ matrices A and B we have:

$$\begin{aligned} Tr(AB) &= \sum_{i=1}^{n} (AB)[i,i] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} (A[i,j]B[j,i]) \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} (B[j,i]A[i,j]) \\ &= \sum_{j=1}^{m} (BA)[j,j] \\ &= Tr(BA) \end{aligned}$$

(b) By part (a), we have:

$$Tr(AB) = \sum_{j=1}^{m} \sum_{i=1}^{n} (B[j,i]A[i,j])$$

Thus, $\frac{\partial Tr(AB)}{\partial A[i,j]} = B[j,i]$. Hence, $\nabla_A Tr(AB) = B^T$.

Question 5

(a) $\gamma(L) = min(L.P)$. Also note that $d(x_1, L_1) = \frac{|w^T x_1 + b|}{\|w\|_2}$. Using this, for L_1, we have $w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, so $\|w\|_2 = \sqrt{2}$. It follows that:

$$\begin{aligned} d(x_1, L_1) &= \frac{1}{\sqrt{2}} \\ d(x_2, L_1) &= \frac{1}{\sqrt{2}} \\ d(x_3, L_1) &= \frac{3}{\sqrt{2}} \\ d(x_4, L_1) &= \frac{3}{\sqrt{2}} \end{aligned}$$

The smallest value is $\frac{1}{\sqrt{2}}$. Thus, $\gamma(L_1) = \frac{1}{\sqrt{2}}$. Performing the same for other lines, we get $\gamma(L_2) = \sqrt{2}$, $\gamma(L_3) = \frac{1}{\sqrt{2}}$, $\gamma(L_4) = \frac{1}{\sqrt{5}}$.

(b) Observe that for any two points p, q on the plane, the line that separating two points and has maximum margin w.r.t two points will going through the midpoint of the segment pq and perpendicular to pq.

Apply the observation to two points $(\begin{bmatrix}1\\2\end{bmatrix}, -1), (\begin{bmatrix}3\\4\end{bmatrix}, 1)$. The line that has maximum w.r.t these two points will go through the point $\begin{bmatrix}2\\3\end{bmatrix}$ and is perpendicular to the vector $\begin{bmatrix}1\\1\end{bmatrix}$, which happens to be L_2. The margin of L_2 w.r.t the two points is $\sqrt{2}$. Other line separating this two points would have margin less than $\sqrt{2}$. Note that any line separating four points would separate the two points as well. Thus, for any separating line $(L) : w_1x_1 + w_2x_2 + b = 0$ of the four points, $\gamma(L_2) \geq \gamma(L)$.

Question 6

A good starting value of b and vector w would be $b = 0$ and $w = [-1, 1]$ where it classifies 4 of the given points correctly. To perform gradient descent, each step follows update rule as below:

$$w[j] = w[j] - \eta(w[j] + C\sum_{i=1}^{n} -y_i x_i[j].\,\mathbf{1}\left[y_i(w^T x_i + b) \leq 1\right])$$

$$b = b - \eta(C\sum_{i=1}^{n} -y_i.\,\mathbf{1}\left[y_i(w^T x_i + b) \leq 1\right])$$

The result of first phase:

$$w[1] = -1 - 0.2(-1 + 0.1\sum_{i=1}^{6} -y_i x_i[j].\,\mathbf{1}\left[y_i(w^T x_i + b) \leq 1\right]) = -0.7$$

$$w[2] = -1 - 0.2(1 + 0.1\sum_{i=1}^{6} -y_i x_i[j].\,\mathbf{1}\left[y_i(w^T x_i + b) \leq 1\right]) = 0.7$$

$$b = 0 - 0.2(0.1\sum_{i=1}^{6} -y_i.\,\mathbf{1}\left[y_i(w^T x_i + b) \leq 1\right]) = 0$$

Doing the same for step two yields $w = [-0.4, 0.6]$ and $b = 0.04$.