1) $(h_t^*(x), B_t^*) = \text{argmin } (e^{B_t} - e^{-B_t})\epsilon_t + e^{-B_t}$

a) $\frac{d}{dB_t}(h_t^*(x), B_t^*) = \text{argmin } (e^{B_t} + e^{-B_t})\epsilon_t - e^{-B_t}$

$$e^{B_t}\epsilon_t + e^{-B_t}(\epsilon_t - 1) = 0$$

$$e^{B_t}\epsilon_t = e^{-B_t}(1-\epsilon_t)$$

$$\log(\epsilon_t) + \log e^{B_t} = \log e^{-B_t} + \log(1-\epsilon_t)$$

$$\log \epsilon_t + B_t = -B_t + \log(1-\epsilon_t)$$

$$2B_t = \log(1-\epsilon_t) - \log(\epsilon_t)$$

$$B_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

b) A hard margin SVM has a slack of 0 thus $\epsilon_t = 0$. Therefore

$$B_t = \lim_{\epsilon_t \to 0} \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right) = \infty$$

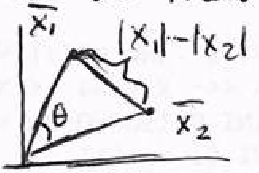
2) a) $\mu_k = \dfrac{\sum z_{nk} x_n}{\sum z_{nk}}$

If $\mu_k = \sum a_{nk} x_n$, then

$a_{nk} = \dfrac{z_{nk}}{\sum_n z_{nk}}$

Therefore a_{nk} is computed with all z's.

b) We can find the distance between 2 points if we treat the vectors as edges of a triangle.



Law of Cosines:

$$(|x_1 - x_2|)^2 = (|x_1|)^2 + (|x_2|)^2 - 2|x_1||x_2|\cos\theta$$

$$(|x_1|)^2 = \langle x_1, x_1 \rangle$$

$$(|x_2|)^2 = \langle x_2, x_2 \rangle$$

$$2|x_1||x_2|\cos\theta = 2\langle x_1, x_2 \rangle$$

Therefore the distance between x_1 and x_2 can be computed using only inner products.

3)

a) If the cluster centers are placed at $X = 1.5, 5,$ and 7 then we will have optimal clustering.

$$(1-1.5)^2 + (2-1.5)^2 + (5-5)^2 + (7-7)^2 = 0.5$$

b) If we chose our cluster centers to be $\{1, 2, 6\}$ then our objective function is $(1-1)^2 + (2-1)^2 + (5-6)^2 + (7-6)^2 = 2$. Because this cluster arrangement can't be improved, Lloyd's algorithm has converged to a suboptimal cluster assignment. In the case of points 5 and 7, they will not change their cluster from center at 6 because 6 is the closest centroid. Thus the algorithm will converge at this assignment as there are no better centroid options that are closer to ~~ass~~ reassign the points to. This arrangement is suboptimal because we showed in part 3a that there was an arrangement with a lower objective function.

c) $\|x_n - \mu_k\|^2 = \|x_n - \sum_n a_{nk} x_n\|^2$

$$= \langle x_n, x_n \rangle + \langle \sum_n a_{nk} x_n, \sum_n a_{nk} x_n \rangle + 2\langle x_n, \sum_n a_{nk} x_n \rangle$$

$\langle x_n, x_n \rangle$ is obviously an inner product.

$\langle \sum_n a_{nk} x_n, \sum a_{nk} x_n \rangle = \langle a_{1k} x_1 + a_{2k} x_2 + \dots a_{nk} x_n, a_{1k} x_1 + a_{2k} x_2 + \dots a_{nk} x_n \rangle$

$$= (a_{1k} x_1 + a_{2k} x_2 + \dots + a_{nk} x_n) \cdot (a_{1k} x_1 + a_{2k} x_2 + \dots + a_{nk} x_n)$$

This is an inner product.

$\langle x_n, \sum_n a_{nk} x_n \rangle = \sum_n a_{nk} \langle x_n, x_n \rangle$

This is composed of an inner product too.

Therefore $\|x_n - \mu_k\|^2$ in this alternative representation is made up of a linear combination of inner products.

4)

a) Transition probabilities:

$q_{21} = P(q_{t+1}=2 | q_t=1) = 1 - q_{11} = 0$

$q_{22} = P(q_{t+1}=2 | q_t=2) = 1 - q_{12} = 0$

Output probabilities:

$e_1(B) = P(O_t = B | q_t = 1) = 1 - e_1(A) = 0.01$

$e_2(A) = P(O_t = A | q_t = 2) = 1 - e_2(B) = 0.49$

b) $P(A) = \pi_1 e_1(A) + \pi_2 e_2(A)$

$= 0.49 \cdot 0.99 + 0.51 \cdot 0.49$

$= 0.735$

$P(B) = \pi_1 e_1(B) + \pi_2 e_2(B)$

$= 0.49 \cdot 0.01 + 0.51 \cdot 0.51$

$= 0.265$

Therefore A should show up more frequently since it has a higher probability from initial state probabilities.

c) We know that our HMM most frequently starts at outputting A. Because of $e_1(A) = 0.99$, we know an output of A means the hidden state was most likely 1. We also know that from q_{11}, once we get a 1 in our state, we will always continue getting 1's. Thus once we get into this infinite loop of 1's, our output is most likely to be an A associated with the 1. This our 3 output symbols are most likely "AAA".