Question 1

$\pi = (0.5\ 0.5)$, $A = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $B = \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$

Question 2

$\pi \cdot A = (0.5\ 0.5) \cdot \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} = (0.5\ 0.5)$

Question 3

$\pi \cdot A \cdot B = (0.5\ 0.5) \cdot \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} = (0.7\ 0.3)$

Question 4

This is a consequence of the chain rule, also known as the product rule, which states that $P(A, B) = P(B|A) \cdot P(A)$. In this example, $A = ((O_{1:t-1} = o_{1:t-1}) \cap (X_t = x_i))$ and $B = (O_t = o_t)$.

Question 5

The number of values of δ and δ^{idx} is $T \cdot N$.

Question 6

The di-gamma function is the probability of being in state $X_t = i$ and transitioning to state $X_{t+1} = j$ at time point t given the emission sequence and the HMM. As a result of this, you have to compute the way through $X_t = i$ to get to $X_{t+1} = j$ divided by all possible ways through the entire HMM, given the emission sequence, which is equal to $\sum_{k=1}^{N} \alpha_T(k)$. We basically want to calculate the ratio between one way and all ways.

Question 7

The alogrithm converges for both observation sequences. For the data sequence with the length of 1000 elements, it needed 956 iterations. For the T=10000 data sequence, 1524 iterations were needed. The criterium for convergence is whether the probability of the given emission sequence increases compared to the previous step. If this is not given, the algorithm should be aborted.

Question 8

We trained the model with the following values, which are the goal values:

$$A = \begin{pmatrix} 0.7 & 0.05 & 0.25 \\ 0.1 & 0.8 & 0.1 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.7 & 0.2 & 0.1 & 0.0 \\ 0.1 & 0.4 & 0.3 & 0.2 \\ 0.0 & 0.1 & 0.2 & 0.7 \end{pmatrix}$$

$\pi = (1.0\ 0.0\ 0.0)$

The output of the Baum-Welch algorithm is nearly the same as the intial values, but it differs largely in some array elements. The problem with estimating the difference between matrices is the fact that two distinct matrices can lead to the same probability for an emission sequence. Moreover, there are various ways to compare matrices. It is not clear which one is the best in a specific application case. These issues can be solved with using the probability of the emission sequence given the HMM as the only measurement of how good our estimation is.

Question 9

We trained the model with 2,4, and 5 states. In case of 2 states, we received for both emission sequence lengths a worse probability compare to the version with 3 states. In case of 4 states, we received for both emission sequence lengths a better probability compared to the version with 3 states. The 5-state solution provided an even better probability. In general, the more states we have the more transition the model can have. This allows to get the real world more precisely in the HMM (it is still just a model). Nevertheless, this requires a tremendous amount of computing power. The optimal setting can be determined by experimenting (change amount of states and the distribution). Basically, the more data we have the better will be our model.

Question 10

In case of the uniform distribution: It destroys the whole learning since Baum-Welch is based on calculating ratios. If the initial matrices are unformly distributed, the ratio will be the same for every element and we will get stuck If we have a diagonal matrix, this means that the HMM can never change a state. Therefore, no learning is possible. As previously mentioned, if we provide a HMM close to the solution HMM, we will not need many iterations and the probability of the emission sequence given the initial HMM is nearly optimal.