## Question 1 EM for Probabilistic PCA

**(a)** *E-step. Calculate the statistics of the posterior distribution $q(z) = p(z|\boldsymbol{x})$ which you'll need for the M-step.*

From the Appendix, we know how to get the distribution of $z$ given $\mathbf{x}$, where $z$ is drawn from Gaussian distribution and $\mathbf{x}$ is drawn from a spherical Gaussian distribution.
In our setting,

$$p(z) = \mathcal{N}(z|0, 1)$$
$$p(\mathbf{x}|z) = \mathcal{N}(\mathbf{x}|z\mathbf{u}, \sigma^2\mathbf{I})$$

To apply the parameters in the formulae of the Appendix, we have

$$\begin{aligned}
&\boldsymbol{\mu} = 0, \boldsymbol{\Sigma} = 1, \\
&\boldsymbol{A} = \boldsymbol{u}, \boldsymbol{B} = 0, \boldsymbol{S} = \sigma^2\boldsymbol{I} \\
&\boldsymbol{C} = (1 + \boldsymbol{u}^T(\sigma^2)^{-1}\boldsymbol{u})^{-1} = \frac{\sigma^2}{\sigma^2 + \boldsymbol{u}^T\boldsymbol{u}}
\end{aligned}$$

Thus, we can obtain the folowing formulae:

$$\begin{aligned}
p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x}|0, \boldsymbol{u}^T\boldsymbol{u} + \sigma^2) \\
p(z|\boldsymbol{x}) &= \mathcal{N}(z|\boldsymbol{C}(\boldsymbol{u}^T(\sigma^2)^{-1}\boldsymbol{x}), \boldsymbol{C}) \\
&= \mathcal{N}(z|\frac{\boldsymbol{u}^T\boldsymbol{x}}{\sigma^2 + \boldsymbol{u}^T\boldsymbol{u}}, \frac{\sigma^2}{\sigma^2 + \boldsymbol{u}^T\boldsymbol{u}})
\end{aligned}$$

As a result,

$$\begin{aligned}
m = E[z|\boldsymbol{x}] &= \frac{\boldsymbol{u}^T\boldsymbol{x}}{\sigma^2 + \boldsymbol{u}^T\boldsymbol{u}} \\
Var[z|\boldsymbol{x}] &= \frac{\sigma^2}{\sigma^2 + \boldsymbol{u}^T\boldsymbol{u}} \\
s = E[z^2|\boldsymbol{x}] &= Var[z|\boldsymbol{x}] + E[z|\boldsymbol{x}]^2 \\
&= \frac{\sigma^4 + \sigma^2\boldsymbol{u}^T\boldsymbol{u} + (\boldsymbol{u}^T\boldsymbol{x})^2}{(\sigma^2 + \boldsymbol{u}^T\boldsymbol{u})^2}
\end{aligned}$$

**(b)** *M-step. Re-estimate the parameters, which consist of the vector $\boldsymbol{u}$. derive a formula for $\boldsymbol{u}_{new}$ that maximizes the expected log-likelihood, i.e.,*

$$\boldsymbol{u}_{new} = \arg\max_{\boldsymbol{u}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{q(z^{(i)})}[\log p(z^{(i)}, \boldsymbol{x}^{(i)})]$$

Denote the function to be maximized as

$$\begin{aligned}
\mathbb{F} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{q(z^{(i)})}[\log p(z^{(i)}, \boldsymbol{x}^{(i)})] \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{q(z^{(i)})}[\log q(z^{(i)})p(\boldsymbol{x}^{(i)})]
\end{aligned}$$

Question 1 [(b)] continued on next page. . .

Then,

$$\begin{aligned}
\log p(\boldsymbol{x}^{(i)})q(z^{(i)}) &= \log \frac{1}{\sqrt{2\pi(\boldsymbol{u}^T\boldsymbol{u}+\sigma^2)}} e^{-\frac{\boldsymbol{x}^{(i)2}}{2(\boldsymbol{u}^T\boldsymbol{u}+\sigma^2)}} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{\boldsymbol{u}^T\boldsymbol{u}+\sigma^2}}} e^{-\frac{(z^{(i)}-\frac{\boldsymbol{u}^T\boldsymbol{x}^{(i)}}{\sigma^2+\boldsymbol{u}^T\boldsymbol{u}})^2}{2\frac{\sigma^2}{\boldsymbol{u}^T\boldsymbol{u}+\sigma^2}}} \\
&\propto -\frac{\boldsymbol{x}^{(i)2}}{2(\boldsymbol{u}^T\boldsymbol{u}+\sigma^2)} - \frac{(z^{(i)}-\frac{\boldsymbol{u}^T\boldsymbol{x}^{(i)}}{\sigma^2+\boldsymbol{u}^T\boldsymbol{u}})^2}{2\frac{\sigma^2}{\boldsymbol{u}^T\boldsymbol{u}+\sigma^2}} \\
&\propto -\frac{\boldsymbol{x}^{(i)2}\sigma^2 + [\boldsymbol{u}^T\boldsymbol{x}^{(i)} - (\sigma^2+\boldsymbol{u}^T\boldsymbol{u})z^{(i)}]^2}{2\sigma^2(\boldsymbol{u}^T\boldsymbol{u}+\sigma^2)} \\
&\propto -\frac{z^{(i)2}(\sigma^2+\boldsymbol{u}^T\boldsymbol{u})}{2\sigma^2} + \frac{z^{(i)}\boldsymbol{u}^T\boldsymbol{x}^{(i)}}{\sigma^2} \\
&\propto -\frac{z^{(i)2}\boldsymbol{u}^T\boldsymbol{u}}{2\sigma^2} + \frac{z^{(i)}\boldsymbol{u}^T\boldsymbol{x}^{(i)}}{\sigma^2}
\end{aligned}$$

Apply the liearity of expectation,

$$\begin{aligned}
\mathbb{F} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\log p(\boldsymbol{x}^{(i)})q(z^{(i)})] \\
&= \frac{1}{N}\sum_{i=1}^{N}[-\frac{\mathbb{E}[z^{(i)2}|\boldsymbol{x}^{(i)}]\boldsymbol{u}^T\boldsymbol{u}}{2\sigma^2} + \frac{\mathbb{E}[z^{(i)}|\boldsymbol{x}^{(i)}]\boldsymbol{u}^T\boldsymbol{x}^{(i)}}{\sigma^2}] \\
&= \frac{1}{N}\sum_{i=1}^{N}[-\frac{s^{(i)}\boldsymbol{u}^T\boldsymbol{u}}{2\sigma^2} + \frac{m^{(i)}\boldsymbol{u}^T\boldsymbol{x}^{(i)}}{\sigma^2}]
\end{aligned}$$

To get the gradient with repect to $\boldsymbol{u}$,

$$\begin{aligned}
\frac{\partial \mathbb{F}}{\partial \boldsymbol{u}} = -\frac{1}{N}\sum_{i=1}^{N}[\frac{s^{(i)}\boldsymbol{u}}{\sigma^2} + \frac{m^{(i)}\boldsymbol{x}^{(i)}}{\sigma^2}] = 0 \\
\boldsymbol{u} \leftarrow \frac{\frac{1}{N}\sum_{i=1}^{N} m^{(i)}\boldsymbol{x}^{(i)}}{\frac{1}{N}\sum_{i=1}^{N} s^{(i)}} \\
\boldsymbol{u} \leftarrow \frac{\sum_{i=1}^{N} m^{(i)}\boldsymbol{x}^{(i)}}{\sum_{i=1}^{N} s^{(i)}}
\end{aligned}$$

### Question 2 Contraction Maps

**(a)** *Show that the Bellman backup operator $T^\pi$ is a contraction map in the $||\cdot||_\infty$ norm.*

Our claim is that the Bellman backup operator $T^\pi$ is a contraction map, which means

$$||T^\pi Q_1 - T^\pi Q_2||_\infty \leq \gamma ||Q_1 - Q_2||_\infty$$

By applying Bellman equation:

$$Q_{k+1}(s,a) \leftarrow r(s,a) + \gamma \sum_{s'} P(s'|a,s) \sum_{a'} \pi(a'|s') Q_k(s',a')$$

Question 2 [(a)] continued on next page...

we have

$$
\begin{aligned}
&|T^{\pi}Q_1(s,a) - T^{\pi}Q_2(s,a)|_\infty \\
=&|[r(s,a) + \gamma \sum_{s'} P(s'|a,s) \sum_{a'} \pi(a'|s')Q_1(s',a')] - [r(s,a) + \gamma \sum_{s'} P(s'|a,s) \sum_{a'} \pi(a'|s')Q_2(s',a')]|_\infty \\
=&\gamma|\sum_{s'} P(s'|a,s) \sum_{a'} \pi(a'|s')[Q_1(s',a') - Q_2(s',a')]|_\infty \\
\leq&\gamma \sum_{s'} P(s'|a,s) \sum_{a'} \pi(a'|s')|Q_1(s',a') - Q_2(s',a')|_\infty \\
\leq&\gamma|Q_1(s',a') - Q_2(s',a')|_\infty \sum_{s'} P(s'|a,s) \sum_{a'} \pi(a'|s') \\
=&\gamma|Q_1(s',a') - Q_2(s',a')|_\infty
\end{aligned}
$$

This is true for any $(s,a)$, so

$$||T^{\pi}Q_1 - T^{\pi}Q_2||_\infty \leq \gamma||Q_1 - Q_2||_\infty$$

which is what we wanted to show.

## Question 3 Q-Learning

**(a)** *Determine the optimal policy and the Q-function for the optimal policy.*

The optimal policy will be

$$\pi(Stay|s_1) = 0; \pi(Switch|s_1) = 1; \pi(Stay|s_2) = 1; \pi(Switch|s_2) = 2;$$

Then, by applying Bellman equation

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|a,s) \max_{a'} Q^*(s',a')$$

, we get

$$
\begin{aligned}
Q^*(s_1, Stay) &= R(s_1) + 0.9 \max_{a'} Q^*(s_1, a') \\
Q^*(s_1, Switch) &= R(s_1) + 0.9 \max_{a'} Q^*(s_2, a') \\
Q^*(s_2, Stay) &= R(s_2) + 0.9 \max_{a'} Q^*(s_2, a') \\
Q^*(s_2, Switch) &= R(s_2) + 0.9 \max_{a'} Q^*(s_1, a')
\end{aligned}
$$

Applying to this question's setting,

$$
\begin{aligned}
R(s_1) + 0.9 \max_{a' \in \mathcal{A}} Q^*(s_1, a') - Q^*(s_1, Stay) = 0 \\
R(s_1) + 0.9 \max_{a' \in \mathcal{A}} Q^*(s_2, a') - Q^*(s_1, Switch) = 0 \\
R(s_2) + 0.9 \max_{a' \in \mathcal{A}} Q^*(s_2, a') - Q^*(s_2, Stay) = 0 \\
R(s_2) + 0.9 \max_{a' \in \mathcal{A}} Q^*(s_1, a') - Q^*(s_2, Switch) = 0
\end{aligned}
$$

that is,

$$
\begin{aligned}
1 + 0.9 \max\{Q^*(s_1, Stay), Q^*(s_1, Switch)\} - Q^*(s_1, Stay) = 0 \\
1 + 0.9 \max\{Q^*(s_2, Stay), Q^*(s_2, Switch)\} - Q^*(s_1, Switch) = 0 \\
2 + 0.9 \max\{Q^*(s_2, Stay), Q^*(s_2, Switch)\} - Q^*(s_2, Stay) = 0 \\
2 + 0.9 \max\{Q^*(s_1, Stay), Q^*(s_1, Switch)\} - Q^*(s_2, Switch) = 0
\end{aligned}
$$

Question 3 [(a)] continued on next page. . . 

so

$$Q^*(s_2, Stay) = 1 + Q^*(s_1, Switch)$$
$$Q^*(s_2, Switch) = 1 + Q^*(s_1, Stay)$$

Apply them,

$$1 + 0.9\max\{Q^*(s_1, Stay), Q^*(s_1, Switch)\} - Q^*(s_1, Stay) = 0$$
$$1.9 + 0.9\max\{Q^*(s_1, Stay), Q^*(s_1, Switch)\} - Q^*(s_1, Switch) = 0$$

so

$$Q^*(s_1, Switch) = Q^*(s_1, Stay) + 0.9$$
$$Q^*(s_2, Stay) = Q^*(s_1, Stay) + 1.9$$
$$Q^*(s_2, Switch) = Q^*(s_1, Stay) + 1$$

Finally, we have

$$Q^*(s_1, Stay) = 18.1$$
$$Q^*(s_1, Switch) = Q^*(s_1, Stay) + 0.9 = 19$$
$$Q^*(s_2, Stay) = Q^*(s_1, Stay) + 1.9 = 20$$
$$Q^*(s_2, Switch) = Q^*(s_1, Stay) + 1 = 19.1$$

or

| | Stay | Switch |
|---|---|---|
| $s_1$ | 18.1 | 19 |
| $s_2$ | 20 | 19.1 |

**(b)** *Now suppose we apply Q-learning, except that instead of the $\epsilon$-greedy policy, the agent follows the greedy policy which always chooses $\pi(s) = \arg\max_a Q(s, a)$. Assume the agent starts in state $S_0 = s_1$. Give an example of a Q-function that is in equilibrium (i.e. it will never change after the Q-learning update rule is applied), but which results in a suboptimal policy.*

To be in equilibrium, the expected change in $Q(S, A)$ should be zero, i.e.

$$\mathbb{E}[R + \gamma \max_{a' \in \mathcal{A}} Q(S', a') - Q(S, A) | S, A] = 0$$

According to Q-Learning, we should initialize $Q(s, a)$ for all the $(s, a) \in \mathcal{S} \times \mathcal{A}$.
We could assign them as 10 and 0's, where the Q-function is shown as

| | Stay | Switch |
|---|---|---|
| $s_1$ | 10 | 0 |
| $s_2$ | 0 | 0 |

For the time step $t = 0$,
Choose $A_t$ according to the greedy policy, i.e., $A_0 \leftarrow \arg\max_{a \in \mathcal{A}} Q(S_0, a) = \arg\max_{a \in \mathcal{A}} Q(s_1, a)$
Because $Q(s_1, Stay) = 10$ and $Q(s_1, Switch) = 0$, we will choose $A_0 = Stay$.
Then, the new state becomes $S_1 = s_1$, as we choose to *Stay*.
Also, $R_0 = r(s_1, Stay) = 1$.
Finally, we update the action-value function at state-action $(s_1, Stay)$ as

$$\begin{aligned} Q(s_1, Stay) &\leftarrow Q(s_1, Stay) + \alpha[R_0 + \gamma \max_{a' \in \mathcal{A}} Q(s_1, a') - Q(s_1, Stay)] \\ &= 10 + \alpha[1 + 0.9 \times 10 - 10] \\ &= 10 \end{aligned}$$

The table of Q-function still is

| | Stay | Switch |
|---|---|---|
| $s_1$ | 10 | 0 |
| $s_2$ | 0 | 0 |

Continue the algorithm, we will find it choose *Stay* all the time and will never visit $s_2$.
And the expected change in $Q(S, A)$ should be zero, i.e.

$$\mathbb{E}[R(s_1) + \gamma \max_{a' \in \mathcal{A}} Q(s_1, a') - Q(s_1, Stay) | s_1, Stay] = 0$$

It is an equilibrium situation, but never converging to the optimal policy.