Question 1: a) calculate posterior distribution $q(z) = p(z|x)$

σ^2 is fixed noise variance

u is parameter vector.

$z \sim N(0, I)$

$x|z \sim N(zu, \sigma^2 I)$

From Appendix:

$p(z) = N(z|\mu, \Sigma)$

$p(x|z) = N(x|Az+b, S)$

$p(x) = N(x|A\mu+b, A\Sigma A^T + S)$

$p(z|x) = N[z|C(A^T S^{-1}(x-b) + \Sigma^{-1}\mu), C]$

$C = (\Sigma^{-1} + A^T S^{-1} A)^{-1}$

From Appendix we know $p(\vec{x}|z) = N(\vec{x}|A\vec{z}+\vec{b}, S)$

$x|z \sim N(z\vec{u}, \sigma^2 I)$

$\therefore \boxed{A = \vec{u},\ \vec{b} = 0}$ $\quad S = \sigma^2 I$

$z \sim N(0, I)$

$p(z) = N(z|\vec{\mu}, \Sigma)$

$\therefore \boxed{\Sigma = I \ \ \therefore \Sigma^{-1} = I \ \ \vec{\mu} = 0}$

$\therefore C = [\Sigma^{-1} + A^T S^{-1} A]^{-1} = [I + \vec{u}^T S^{-1} \vec{u}]^{-1} = [I + \vec{u}^T \sigma^{-2} I \vec{u}]^{-1}$

$\therefore p(z|\vec{x}) = N[z|C(A^T S^{-1}(x-b) + \Sigma^{-1}\mu), C]$

($A^T S^{-1}$: mean; C: variance)

$\therefore m = E[z|\vec{x}] = C \cdot [A^T S^{-1}(x-b) + \Sigma^{-1}\mu]$

$\boxed{m = [I + \vec{u}^T \sigma^{-2} I \vec{u}]^{-1} \cdot [\vec{u} \cdot \sigma^{-2} I \vec{x}]}$

$\because Var(Y) = E(Y^2) - E(Y)^2$

$\therefore S = E(z^2|x] = Var(z|x) + (E[z|x])^2$

$= C + \{C[A^T S^{-1}(x-b) + \Sigma^{-1}\mu]\}^2$

$= C + [C(\vec{u} \cdot \sigma^{-2} I \vec{x})]^2$

where $C = [I + \vec{u}^T \sigma^{-2} I \vec{u}]^{-1}$

$\therefore \boxed{S = (I + \vec{u}^T \sigma^{-2} I \vec{u})^{-1} + [(I + \vec{u}^T \sigma^{-2} I \vec{u})^{-1}(\vec{u} \cdot \sigma^{-2} I \vec{x})]^2}$

b> $\log p(z^i, x^i) = \log\left[p(x^i|z^i)\cdot p(z)\right] = \log p(x^i|z^i) + \log p(z^i)$

Equation $\frac{1}{N}\sum_{i=1}^{N} E_{q(z^{(i)})}\left[\log p(z^{(i)}, x^{(i)})\right]$. ①

$= \frac{1}{N}\sum_{i=1}^{N} E_{q(z^{(i)})}\left[\log p(x_i|z_i) + \log p(z_i)\right]$

$\log p(z_i) = \log\left(N(0,1)\right) = \log \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} = \log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}z^{2(i)}$

$\log p(x_i|z_i) = \log\left(N(z^{(i)}u, \sigma^2 I)\right) = \log\left[\frac{1}{2\pi^{\frac{1}{2}}(\sigma^2 I)^{\frac{1}{2}}} e^{-\frac{(x_i - z_i u)^2}{2\sigma^2}}\right] = \log\left[\frac{1}{2\pi^{\frac{1}{2}}(\sigma^2)^{\frac{1}{2}}}\right] - \frac{(x_i - z_i u)^2}{2\sigma^2}$

by using linear expertion

$E\left[\log(z^{(i)}, x^{(i)}\right] = \text{const} + \sum_{i=1}^{N}\left[-\frac{1}{2}E[z^{(i)2}] - \frac{1}{2\sigma^2}E(x^{(i)} - z^{(i)}u)^2\right]$

$= \text{const} - \frac{1}{2}\sum_{i=1}^{N} E[z^{(i)2}] + \frac{1}{\sigma^2}E\left[x^{(i)2} - 2x^{(i)}z^{(i)}u + z^{(i)2}u^2\right]$

take derivative of u_j:

$$\frac{\partial E\left[\log p(z^{(i)}, x^{(i)})\right]}{\partial u_j} = \frac{\partial\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N} E\left[-2x^{(i)}z^{(i)}u + z^{(i)2}u^2\right]\right]}{\partial u_j}$$

$$= \frac{\partial\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[-2x^{(i)}m^{(i)}u + u^2 s^{(i)}\right]\right.}{\partial u_j} = 0$$

$\therefore \sum_{i=1}^{N}\left[-2x_j^{(i)}m^{(i)} + 2u_j s^{(i)}\right] = 0$

$\therefore \sum_{i=1}^{N}\left[x_j^{(i)}m^{(i)}\right] = \sum_{i=1}^{N}\left[u_j s^{(i)}\right]$

$$u_j = \sum_{i=1}^{N}\left[x_j^{(i)}m^{(i)}\right]\cdot\left[\sum_{i=1}^{N} s^{(i)}\right]^{-1}$$

question 2: Show that the Bellman backup operator T^π is a contraction map

For an operator f is a contraction map if $\|f(x_1) - f(x_2)\| \le \alpha \|x_1 - x_2\|$

∴ for T^π, $\|T^\pi Q_1 - T^\pi Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$

$\|\cdot\|_\infty$ means $\|x\|_\infty = \max_i |x_i|$

Answer:

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

$$\left\| (T^\pi Q_{k+1})(s,a) - (T^\pi Q_k)(s,a) \right\|_\infty$$

$$= \left\| r(s,a) + \gamma \sum_{s'} P(s'|a,s) \sum_{a'} \pi(a'|s') Q_{k+1}(s',a') \right] - \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a) \sum_{a'} \pi(a'|s') Q_k(s',a') \right] \|$$

$$= \gamma \left\| \left[\sum_{s'} P(s'|a,s) \left[\sum_{a'} \pi(a'|s') \max_{a'} Q_{k+1}(s',a') - \sum_{a'} \pi(a'|s') \max_{a'} Q_k(s',a') \right] \right] \right|$$

$$= \gamma \left| \sum_{s'} P(s'|a,s) \cdot \sum_{a'} \pi(a'|s') \left[\max_{a'} Q_{k+1}(s',a') - \max_{a'} Q_k(s',a') \right] \right|$$

$$\le \gamma \sum_{s'} P(s'|a,s) \cdot \sum_{a'} \pi(a'|s') \max_{a'} \left| Q_{k+1}(s',a') - Q_k(s',a') \right|$$

$$\le \gamma \max_{s',a'} \left| Q_{k+1}(s',a') - Q_k(s',a') \right| \underbrace{\sum_{s'} P(s'|a,s)}_{\text{sum}=1} \cdot \underbrace{\sum_{a'} \pi(a'|s')}_{\text{sum}=1}$$

$$= \gamma \max_{s',a'} \left| Q_{k+1}(s',a') - Q_k(s',a') \right|$$

$$= \gamma \|Q_{k+1}(s',a') - Q_k(s',a')\|_\infty$$

∴ $\|T^\pi Q_{k+1}(s,a) - T^\pi Q_k(s,a)\|_\infty \le \gamma \|Q_{k+1}(s',a') - Q_k(s',a')\|_\infty$

Question 3:

1) $S_1 \to a_1 \to S_1$ $\quad Q(S_1,a_1) = R(S_1,a_1) + 0.9 \times \max[Q(S_1,a_1), Q(S_1,a_2)]$

$S_1 \to a_2 \to S_1$ $\quad Q(S_1,a_2) = R(S_1,a_2) + 0.9 \times \max[Q(S_2,a_1), Q(S_2,a_2)]$

$S_2 \to a_1 \to S_2$ $\quad Q(S_2,a_1) = R(S_2,a_1) + 0.9 \times \max[Q(S_2,a_1), Q(S_2,a_2)]$

$S_2 \to a_2 \to S_1$ $\quad Q(S_2,a_2) = R(S_2,a_2) + 0.9 \times \max[Q(S_1,a_1), Q(S_1,a_2)]$

a1=stay a2 =switch

Q table:

	a_1	a_2
S_1	18.1	19
S_2	20	19.1

optimal policy: $S_1 \to a_2 \to S_2$

$S_2 \to a_1 \to S_2$

2) using Q-learning, and always chooses $\pi(s) = \arg\max_a Q(s,a)$

Assumption: the agent starts in state $S_0 = s_0$

learning rate = 1

$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + [R_t + \gamma \max_{a' \in A} Q(S_{t+1},a') - Q(S_t,A_t)]$

Q table initialization

	a_1	a_2
S_1	0	0
S_2	0	0

For example, after initialize the Q-table, Q-learning choose the first action to be a_1. Because, all the $\arg\max_{a \in A} Q(S_t,a)$ can choose a_1 or a_2. Here I will use $A_t = a_1$ to show why Q-function is in equilibrium.

step① $Q(S_t,A_t) \leftarrow 1 + 0.9 \times 0$

choose a_1, stay in S_1

$\Rightarrow$

	a_1	a_2
S_1	1	0
S_2	0	0

step② $Q(S_t,A_t) \leftarrow 1 + 0.9 \times 1$

choose a_1, stay in S_1

$\Rightarrow$

	a_1	a_2
S_1	1.9	0
S_2	0	0

step③ $Q(S_t,A_t) \leftarrow 1 + 0.9 \times 1.9$

choose a_1, stay in S_1

$\Rightarrow$

	a_1	a_2
S_1	2.71	0
S_2	0	0

⋮

when using $\arg\max_{a \in A} Q(s,a)$, we can find the trend that a function will always choose s_1, a_1

because $Q(s_1, a_1)$ is bigger than other Q value. Therefore, we can derive the final answer of Q table when a function always choose s_1, a_1.

$$1 + 0.9\, Q(s_1, a_1) \Rightarrow Q(s_1, a_1)$$

we also can find the regulaery $Q(s_1, a_1)$ at the n times $= 0.9^n + 0.9^{n-1} + 0.9^{n-2} \cdots 0.9^0$

$$\therefore Q(s_1, a_1) = \sum_{n=0}^{\infty} 0.9^n$$

$$= \frac{1}{1-0.9}$$

$$= 10$$

∴ the final Q table is

	a_1	a_2
s_1	10	0
s_2	0	0

policy $s_1 \rightarrow a_1 \rightarrow s_1$

From the Q table, we can see the result is a sub optimal policy. It gets stuck with state 1, and action 1 when the agent starts in state $S_0 = s_1$.