8.2.2 확률적 LQR

확률적 LQR 문제에서는 동역학 모델이 다음과 같이 선형 확률 동적 시스템으로 주어진다.

$$
\begin{aligned}
x_{t+1} &= A_t x_t + B_t u_t + f_{ct} + n, t = 0, \ldots T \\
&= F_t \begin{bmatrix} x_t \\ u_t \end{bmatrix} + f_{ct} + n_t
\end{aligned}
$$

여기서 $F_t = [A_t, B_t]$ 이고 A_t, B_t 는 확정된 행렬.

n_t 는 프로세스 노이즈로 평균이 0이고 분산이 Σ_i 인 가우시안 불포를 갖는 화이트 시퀀스 즉,

$$p(n_t) = N(0, \Sigma_t), \quad E[n_i n_j^T] = \Sigma_i \delta_{ij}$$

여기서 δ_{ij} 는 크로네커 델타 함수

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

프로세스 노이즈가 랜덤 시퀀스이기 때문에 상태변수 x_t 는 랜덤 시퀀스가 됨

초기값 x_0 도 n_t 와 독립이며 가우시안 분포를 갖는 것으로 가정

$$p(x_0) = N(\bar{x_0}, X_0), \quad E[(x_0 - \bar{x_0}) n_i^T] = 0$$

선형 확률 동적 시스템의 상태천이 확률밀도함수는 다음과 같이 전파된다.

$$p(x_{t+1} | x_t, u_t) = N(A_t x_t + B_t u_t + f_{ct}, \Sigma_t)$$

확률적 시스템에서는 무작위적 특성 때문에 비용함수의 값도 랜덤 변수 $\rightarrow$ 목적함수를 항상 최소화할 수 있는 제어 또는 행동 시퀀스를 구할 수 없음 $\rightarrow$ 비용함수의 기대값을 최소로 만드는 제어 시퀀스를 구하는 것을 목적으로 한다.

$$J_0 = E_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} c(x_t, u_t) \right]$$

확정적 시스템에서는 최적제어의 시퀀스를 구했다면 확률적 시스템에서는 현재 상태 변수가 주어진 조건에서 최적 정책 $\pi_t(u_t | x_t)$ 의 시퀀스 $(\pi_0, \pi_1, \ldots \pi_T)$ 를 구하는 것이 목적.

정책이 시변함수이기 때문에 목적함수는 다음과 같이 시변 상태가치 함수가 된다.

$$V(x_0) = E_{\tau_{u_0} \sim p(\tau_{u_0} | x_0)} \left[\sum_{t=0}^{T} c(x_t, u_t) \right]$$

$$J_0 = E_{x_0 \sim p(x_0)} [V(x_0)]$$

DP 를 적용하기 위해 상태 가치 함수를 시간 스텝 t를 기준으로 전개

$$
\begin{aligned}
V(\mathrm{x}_t) &= E_{\tau_{\mathrm{u}_t} \sim p(\tau_{\mathrm{u}_t}|\mathrm{x}_t)}\left[\sum_{k=t}^{T} c(\mathrm{x}_k,\ \mathrm{u}_k)\right] \\
&= \int_{\tau_{\mathrm{u}_t}} \left(\sum_{k=t}^{T} c(\mathrm{x}_k,\ \mathrm{u}_k)\right) p(\tau_{\mathrm{u}_t}|\mathrm{x}_t)\, d\tau_{\mathrm{u}_t} \\
&= \int_{\mathrm{u}_t}\left[\int_{\tau_{\mathrm{x}_{t+1}}} \left(\sum_{k=t}^{T} c(\mathrm{x}_k,\ \mathrm{u}_k)\right) p(\tau_{\mathrm{x}_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t)\, d\tau_{\mathrm{x}_{t+1}}\right] \pi(\mathrm{u}_t|\mathrm{x}_t)\, d\mathrm{u}_t \\
&= E_{\mathrm{u}_t \sim \pi(\mathrm{u}_t|\mathrm{x}_t)}[Q(\mathrm{x}_t,\ \mathrm{u}_t)]
\end{aligned}
\tag{8.41}
$$

대괄호 항은 어떤 상태변수 x_t 에서 제어 u_t 를 선택하고 그로부터 정책 π 로 기대할 수 있는 미래 비용의 기대값이므로 시변 행동 가치 함수 $Q(x_t, u_t)$ 가 된다.

행동 가치 함수 $Q(x_t, u_t)$ 를 한 스텝 더 전개 한다.

$$
\begin{aligned}
Q(\mathrm{x}_t,\ \mathrm{u}_t) &= \int_{\tau_{\mathrm{x}_{t+1}}} c(\mathrm{x}_t,\ \mathrm{u}_t) p(\tau_{\mathrm{x}_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t)\, d\tau_{\mathrm{x}_{t+1}} \\
&\qquad + \int_{\tau_{\mathrm{x}_{t+1}}} \left(\sum_{k=t+1}^{T} c(\mathrm{x}_k,\ \mathrm{u}_k)\right) p(\tau_{\mathrm{x}_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t)\, d\tau_{\mathrm{x}_{t+1}} \\
&= c(\mathrm{x}_t,\ \mathrm{u}_t) + Q_1
\end{aligned}
\tag{8.42}
$$

Q_1 을 정리해 보면

$$
\begin{aligned}
Q_1 &= \int_{\mathrm{x}_{t+1}} \int_{\tau_{\mathrm{u}_{t+1}}} \left(\sum_{k=t+1}^{T} c(\mathrm{x}_k,\ \mathrm{u}_k)\right) p(\tau_{\mathrm{u}_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t,\ \mathrm{x}_{t+1}) p(\mathrm{x}_{t+1}|\mathrm{x}_t,\ \mathrm{u}_t)\, d\tau_{\mathrm{u}_{t+1}} d\mathrm{x}_{t+1} \\
&= \int_{\mathrm{x}_{t+1}} \left[\int_{\tau_{\mathrm{u}_{t+1}}} \left(\sum_{k=t+1}^{T} c(\mathrm{x}_k,\ \mathrm{u}_k)\right) p(\tau_{\mathrm{u}_{t+1}}|\mathrm{x}_{t+1})\, d\tau_{\mathrm{u}_{t+1}}\right] p(\mathrm{x}_{t+1}|\mathrm{x}_t,\ \mathrm{u}_t)\, d\mathrm{x}_{t+1}
\end{aligned}
\qquad [8.4
$$

위의 식에서 대괄호는 $V(x_{t+1})$ 이므로

$$
Q_1 = \int_{x_{t+1}} V(x_{t+1}) p(x_{t+1}|x_t, u_t) dx_{t+1}
$$

이다. 따라서 행동가치 함수 $Q(x_t, u_t)$ 는 다음과 같이 된다.

$$
\begin{aligned}
Q(\mathrm{x}_t,\ \mathrm{u}_t) &= c(\mathrm{x}_t,\ \mathrm{u}_t) + \int_{\mathrm{x}_{t+1}} V(\mathrm{x}_{t+1}) p(\mathrm{x}_{t+1}|\mathrm{x}_t,\ \mathrm{u}_t)\, d\mathrm{x}_{t+1} \\
&= c(\mathrm{x}_t,\ \mathrm{u}_t) + E_{\mathrm{x}_{t+1} \sim p(\mathrm{x}_{t+1}|\mathrm{x}_t,\ \mathrm{u}_t)}[V(\mathrm{x}_{t+1})]
\end{aligned}
\qquad [8.4
$$

위이 식을 상태 가치 식 8.41에 대입하면

$$
V(x_t) = E_{u_t \sim \pi_t(u_t|x_t)}[c(x_t, u_t) + E_{x_{t+1}, \sim p(x_{t+1}|x_t, u_t)}[V(x_{t+1})]]
$$

확정적 정책 $u_t = \pi(x_t)$ 를 가정하면

$$
V(x_t) = c(x_t, u_t) + E_{x_{t+1}, \sim p(x_{t+1}|x_t, u_t)}[V(x_{t+1})]
$$

벨만의 최적성 원리에 의하면 현재 시간스텝 t 의 최적 제어는 다음을 만족해야 한다.

$$
V^*(x_t) = \underset{u_t}{min}\ ((x_t, u_t) + E_{x_{t+1}, \sim p(x_{t+1}|x_t, u_t)}[V(x_{t+1})])
$$

확률적 LQR문제에서는 비용함수가 다음과 같이 2차 함수로 주어진다.

$$c(x_t, u_t) = \frac{1}{2}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^T C_T \begin{bmatrix} x_t \\ u_t \end{bmatrix} + \begin{bmatrix} x_t \\ u_t \end{bmatrix}^T c_t$$

최종 시간 $t = T$ 에서의 목적함수 는

$$\begin{aligned} V(\mathrm{x}_T) &= \frac{1}{2}\begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{C}_T \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix} + \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{c}_T \\ &= \frac{1}{2}\begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{Q}_T \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix} + \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{q}_T \\ &= Q(\mathrm{x}_T,\ \mathrm{u}_T) \end{aligned} \tag{8.50}$$

여기서 $Q_T = C_T,\ q_T = c_T$ 로 놓았다. 위의 목적함수를 최소로 만드는 최적제어를 구하기 위해 다음과 같이 미분한다.

$$\begin{aligned} 0 &= \frac{\partial\ V(\mathrm{x}_T)}{\partial\ \mathrm{u}_T} = \frac{\partial\ Q(\mathrm{x}_T,\ \mathrm{u}_T)}{\partial\ \mathrm{u}_T} \\ &= Q_{\mathrm{ux}T}\mathrm{x}_T + Q_{\mathrm{uu}T}\mathrm{u}_T + Q_{\mathrm{u}T} \end{aligned} \tag{8.51}$$

여기서

$$Q_T = \begin{bmatrix} Q_{xxT} & Q_{xuT} \\ Q_{uxT} & Q_{uuT} \end{bmatrix}$$
$$q_T = \begin{bmatrix} Q_{xT} \\ Q_{uT} \end{bmatrix}$$

이다. 그러면 u'_T 는

$$u'_T = -Q_{uuT}^{-1} Q_{uxT} x_T - Q_{uuT}^{-1} Q_{uT}$$

칼만 게인을 다음과 같이 정의를 하면

$$\begin{aligned} \mathrm{K}_T &= -Q_{\mathrm{uu}T}^{-1} Q_{\mathrm{ux}T} \\ \mathrm{k}_T &= -Q_{\mathrm{uu}T}^{-1} Q_{\mathrm{u}T} \end{aligned} \tag{8.54}$$

$t = T$ 에서의 최적제어는 다음과 같다.

$$u'_T = K_T x_T + k_T$$

위의 식을 이용해 $t = T$ 에서의 최소 목적함수의 값을 구하면 다음과 같다.

$$V^*(\mathrm{x}_T) = \frac{1}{2}\begin{bmatrix} \mathrm{x}_T \\ \mathrm{K}_T\mathrm{x}_T + \mathrm{k}_T \end{bmatrix}^T \mathrm{Q}_T \begin{bmatrix} \mathrm{x}_T \\ \mathrm{K}_T\mathrm{x}_T + \mathrm{k}_T \end{bmatrix} + \begin{bmatrix} \mathrm{x}_T \\ \mathrm{K}_T\mathrm{x}_T + \mathrm{k}_T \end{bmatrix}^T \mathrm{q}_T$$
$$= \frac{1}{2}\mathrm{x}_T^T\mathrm{V}_T\mathrm{x}_T + \mathrm{x}_T^T\mathrm{v}_T + const \tag{8.56}$$

여기서 const는

$$\mathrm{V}_T = Q_{\mathrm{xx}T} + Q_{\mathrm{xu}T}\mathrm{K}_T + \mathrm{K}_T^T Q_{\mathrm{ux}T} + \mathrm{K}_T^T Q_{\mathrm{uu}T}\mathrm{K}_T$$
$$\mathrm{v}_T = Q_{\mathrm{x}T} + \mathrm{K}_T^T Q_{\mathrm{u}T} + Q_{\mathrm{xu}T}\mathrm{k}_T + \mathrm{K}_T^T Q_{\mathrm{uu}T}\mathrm{k}_T \tag{8.57}$$

다음으로 역방향 시간시스템 $t = T - 1$ 에서의 목적함수는 다음과 같다.

$$V(\mathrm{x}_{T-1}) = \frac{1}{2}\begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T \mathrm{C}_{T-1} \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix} + \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T \mathrm{c}_{T-1}$$
$$+ E_{\mathrm{x}_T \sim p(\mathrm{x}_T|\mathrm{x}_{T-1}, \mathrm{u}_{T-1})}\left[\frac{1}{2}\mathrm{x}_T^T\mathrm{V}_T\mathrm{x}_T + \mathrm{x}_T^T\mathrm{v}_T + const\right]$$
$$= \frac{1}{2}\begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T (\mathrm{C}_{T-1} + \mathrm{F}_{T-1}^T\mathrm{V}_T\mathrm{F}_{T-1}) \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}$$
$$+ \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T (\mathrm{c}_{T-1} + \mathrm{F}_{T-1}^T\mathrm{V}_T\mathrm{f}_{cT-1} + \mathrm{F}_{T-1}^T\mathrm{v}_T) + const$$
$$+ E_{\mathrm{x}_T \sim p(\mathrm{x}_T|\mathrm{x}_{T-1}, \mathrm{u}_{T-1})}\left[\frac{1}{2}\mathrm{n}_{T-1}^T\mathrm{V}_T\mathrm{n}_{T-1}\right] \tag{8.58}$$

$$E_{\mathrm{x}_T \sim p(\mathrm{x}_T|\mathrm{x}_{T-1}, \mathrm{u}_{T-1})}\left[\frac{1}{2}\mathrm{n}_{T-1}^T\mathrm{V}_T\mathrm{n}_{T-1}\right]$$
$$= E_{\mathrm{x}_T \sim p(\mathrm{x}_T|\mathrm{x}_{T-1}, \mathrm{u}_{T-1})}\left[\frac{1}{2}tr(\mathrm{V}_T\mathrm{n}_{T-1}\mathrm{n}_{T-1}^T)\right]$$
$$= \frac{1}{2}tr(\mathrm{V}_T E_{\mathrm{x}_T \sim p(\mathrm{x}_T|\mathrm{x}_{T-1}, \mathrm{u}_{T-1})}[\mathrm{n}_{T-1}\mathrm{n}_{T-1}^T])$$
$$= \frac{1}{2}tr(\mathrm{V}_T\Sigma_{T-1}) \tag{8.59}$$

여기서 식 8.59에 의하면 또 다른 상수항을 추가한 것에 불가하므로 목적함수는 다음과 같이 된다.

$$V(\mathrm{x}_{T-1}) = \frac{1}{2}\begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T (\mathrm{C}_{T-1} + \mathrm{F}_{T-1}^T\mathrm{V}_T\mathrm{F}_{T-1}) \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}$$
$$+ \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T (\mathrm{c}_{T-1} + \mathrm{F}_{T-1}^T\mathrm{V}_T\mathrm{f}_{cT-1} + \mathrm{F}_{T-1}^T\mathrm{v}_T) + const \tag{8.60}$$
$$= \frac{1}{2}\begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T \mathrm{Q}_{T-1} \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix} + \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{u}_{T-1} \end{bmatrix}^T \mathrm{q}_{T-1} + const$$
$$= Q(\mathrm{x}_{T-1},\ \mathrm{u}_{T-1})$$

$$
\begin{aligned}
\mathrm{Q}_{T-1} &= \mathrm{C}_{T-1} + \mathrm{F}_{T-1}^{T}\mathrm{V}_{T}\mathrm{F}_{T-1} \\
\mathrm{q}_{T-1} &= \mathrm{c}_{T-1} + \mathrm{F}_{T-1}^{T}\mathrm{V}_{T}\mathrm{f}_{\mathrm{c}T-1} + \mathrm{F}_{T-1}^{T}\mathrm{v}_{T}
\end{aligned} \tag{8.61}
$$

목적함수를 최소로 만드는 최적제어를 구하기 위해 다음과 같이 미분을 수행한다.

$$
\begin{aligned}
0 &= \frac{\partial\ V(\mathrm{x}_{T-1})}{\partial\ \mathrm{u}_{T-1}} \\
&= Q_{\mathrm{ux}\,T-1}\mathrm{x}_{T-1} + Q_{\mathrm{uu}\,T-1}\mathrm{u}_{T-1} + Q_{\mathrm{u}T-1}
\end{aligned} \tag{8.62}
$$

$$
Q_{T-1} = \begin{bmatrix} Q_{xxT-1} & Q_{xuT-1} \\ Q_{uxT-1} & Q_{uuT-1} \end{bmatrix}
$$

$$
q_{T-1} = \begin{bmatrix} Q_{xT-1} \\ Q_{uT-1} \end{bmatrix}
$$

$$
u'_{T-1} = -Q_{uuT-1}^{-1}Q_{uuT-1}x_{T-1} = Q_{uuT-1}^{-1}Q_{uT-1}
$$

칼만 게임을 다음과 같이 정의 하면!

$$
K_{T-1} = -Q_{uuT-1}^{-1}Q_{uxT-1}
$$

$$
k_{T-1} = -Q_{uuT-1}^{-1}Q_{uT-1}
$$

$t = T - 1$ 에서의 최적제어는

$$
u'_{T=1} = K_{T-1}x_{T-1} + k_{T-1}
$$

위 식을 이용해 $t = T - 1$ 에서의 최소 목적함수 값을 구하면 다음과 같다.

$$
\begin{aligned}
V^{*}(\mathrm{x}_{T-1}) &= \frac{1}{2}\begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{K}_{T-1}\mathrm{x}_{T-1} + \mathrm{k}_{T-1} \end{bmatrix}^{T} \mathrm{Q}_{T-1} \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{K}_{T-1}\mathrm{x}_{T-1} + \mathrm{k}_{T-1} \end{bmatrix} \\
&\quad + \begin{bmatrix} \mathrm{x}_{T-1} \\ \mathrm{K}_{T-1}\mathrm{x}_{T-1} + \mathrm{k}_{T-1} \end{bmatrix}^{T} \mathrm{q}_{T-1} + const \\
&= \frac{1}{2}\mathrm{x}_{T-1}^{T}\mathrm{V}_{T-1}\mathrm{x}_{T-1} + \mathrm{x}_{T-1}^{T}\mathrm{v}_{T-1} + const
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{V}_{T-1} &= Q_{\mathrm{xx}\,T-1} + Q_{\mathrm{xu}\,T-1}\mathrm{K}_{T-1} + \mathrm{K}_{T-1}^{T}Q_{\mathrm{ux}\,T-1} + \mathrm{K}_{T-1}^{T}Q_{\mathrm{uu}\,T-1}\mathrm{K}_{T-1} \\
\mathrm{v}_{T-1} &= Q_{\mathrm{x}T-1} + \mathrm{K}_{T-1}^{T}Q_{\mathrm{u}T-1} + Q_{\mathrm{xu}\,T-1}\mathrm{k}_{T-1} + \mathrm{K}_{T-1}^{T}Q_{\mathrm{uu}\,T-1}\mathrm{k}_{T-1}
\end{aligned}
$$

확정적 LQR과 비교해 보면 상후살ㅇ에 또다른 상수항 $\frac{1}{2}tr(V, \Sigma_{t-1})$ 만 추가 됐고 나머지는 동일하다는 것을 알 수 있다.

결론

확률적 LQR 업데이트 식은 프로세스 노이즈의 공분산 Σ_t 와 초기 상태 변수의 공분산 $\backslash X_0$ 와는 무관하며 확정적 LQR 식과 동일하다는 것을 알 수 있다.

8.2.3 가우시안 LQR

- 확률적 LQR의 확정적 정책을 $\pi(u_t|x_t) = N(x_t|\mu_t, S_t)$ 와 같은 확률밀도 함수를 갖는 확률적 정책으로 확장 시킨 것
- 동일한 x_t 에 대해 다른 제어 μ_t 가 선택될 수 있다.
- 목적함수

$$J_0 = E_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} (c(x_t, u_t) + log\, \pi(u_t|x_t)) \right]$$

위의 식을 전개 하면

$$\begin{aligned} J_0 &= E_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} c(\mathrm{x}_t,\ \mathrm{u}_t) \right] + \sum_{t=0}^{T} E_{\mathrm{x}_t \sim p(\mathrm{x}_t), \mathrm{u}_t \sim \pi(\mathrm{u}_t|\mathrm{x}_t)} [\log \pi(\mathrm{u}_t|\mathrm{x}_t)] \\ &= E_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} c(\mathrm{x}_t,\ \mathrm{u}_t) \right] - \sum_{t=0}^{T} E_{\mathrm{x}_t \sim p(\mathrm{x}_t)} [H(\pi(\mathrm{u}_t|\mathrm{x}_t))] \end{aligned} \tag{8.71}$$

$$H(\pi(u_t|x_t)) = -\int \pi(u_t|x_t) log\, \pi(u_t|x_t) du_t$$

확률적 LQR의 목적함수에 **한계확률밀도함수 $p(x_t)$ 에 대한 정책 엔트로피 기대값 도입됨**

기존 목적함수에 화률적 정책의 엔트로피가 추가함으로써 가우시안 LQR 제어기는 **비용을 최소화함과 동시에 정책의 무작위성을 최대화** 하는 특성이 있다.

확률적 시스템에서는 상태 변수가 랜덤 변수. $\rightarrow$ 시간스텝 t 에서의 제어는 상태변수 x_t 를 기반으로 구축해야 한다는 추가 조건 필요 $\rightarrow$ 목적함수를 다음과 같이 초기 상태변수 x_0 에 대한 조건부 함수로 바꾼다.

$$V(x_0) = E_{\tau_{u_0} \sim p(\tau_{u_0}|x_0)} \left[\sum_{t=0}^{T} (c(x_t, u_t) + log\, \pi(u_t|x_t)) \right]$$

새로운 목적함수는 상태가치 함수에 엔트로피를 추가한 것이므로 **소프트 상태가치 함수(soft state-value function)**이라 부른다.

원래 목적함수와 소프트 상태가치 함수와의 관계

$$J_0 = E_{x_0 \sim p(x_0)} [V(x_0)]$$

DP를 적용하기 위해 소프트 상태가치 함수를 시간스텝 t를 기준으로 전개 해 보자. (식 8.75)

$$
\begin{aligned}
V(\mathrm{x}_t) &= E_{\tau_{u_t} \sim p(\tau_{u_t}|\mathrm{x}_t)}\left[\sum_{k=t}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_k|\mathrm{x}_k))\right] \\
&= E_{\tau_{u_t} \sim p(\tau_{u_t}|\mathrm{x}_t)}\left[\log\pi(\mathrm{u}_t|\mathrm{x}_t)+\sum_{k=t}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_{k+1}|\mathrm{x}_{k+1}))\right] \\
&= \int_{\tau_{u_t}} \log\pi(\mathrm{u}_t|\mathrm{x}_t)p(\tau_{u_t}|\mathrm{x}_t)d\tau_{u_t} \\
&\quad + \int_{\tau_{u_t}}\left(\sum_{k=t}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_{k+1}|\mathrm{x}_{k+1}))\right)p(\tau_{u_t}|\mathrm{x}_t)d\tau_{u_t} \\
&= \int_{u_t}\int_{\tau_{x_{t+1}}} \log\pi(\mathrm{u}_t|\mathrm{x}_t)p(\tau_{x_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t)\pi(\mathrm{u}_t|\mathrm{x}_t)d\tau_{x_{t+1}}d\mathrm{u}_t \\
&\quad + \int_{u_t}\left[\int_{\tau_{x_{t+1}}}\left(\sum_{k=t}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_{k+1}|\mathrm{x}_{k+1}))\right)p(\tau_{x_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t)d\tau_{x_{t+1}}\right]\pi(\mathrm{u}_t|\mathrm{x}_t)d\mathrm{u}_t \\
&= E_{u_t \sim \pi(u_t|x_t)}[\log\pi(\mathrm{u}_t|\mathrm{x}_t)+Q(\mathrm{x}_t,\ \mathrm{u}_t)]
\end{aligned}
\tag{8.75}
$$

여기서 $\tau_{u_t} = (u_t, x_{t+1}, u_{t+1}, \ldots, x_T, u_T)$, $\tau_{x_{t+1}} = (x_{t+1}, u_{t+1}, \ldots, x_T, u_T)$ 다. 정리를 하면,

$$V(x_t) = E_{u_t \sim \pi(u_t|x_t)}[log\,\pi(u_t|x_t) + Q(x_t, u_t)]$$

$Q(x_t, u_t)$: **소프트 행동가치 함수**

$$
\begin{aligned}
Q(\mathrm{x}_t,\ \mathrm{u}_t) &= \int_{\tau_{x_{t+1}}}\left(c(\mathrm{x}_t,\ \mathrm{u}_t)+\sum_{k=t+1}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_k|\mathrm{x}_k))\right)p(\tau_{x_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t)d\tau_{x_{t+1}} \\
&= c(\mathrm{x}_t,\ \mathrm{u}_t)+\int_{\tau_{x_{t+1}}}\left(\sum_{k=t+1}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_k|\mathrm{x}_k))\right)p(\tau_{x_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t)d\tau_{x_{t+1}} \\
&= c(\mathrm{x}_t,\ \mathrm{u}_t)+Q_1
\end{aligned}
\tag{8.77}
$$

Q 를 정리해 보면,

$$
\begin{aligned}
Q_1 &= \int_{x_{t+1}}\int_{\tau_{u_{t+1}}}\left(\sum_{k=t+1}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_k|\mathrm{x}_k))\right)p(\tau_{u_{t+1}}|\mathrm{x}_t,\ \mathrm{u}_t,\ \mathrm{x}_{t+1})p(\mathrm{x}_{t+1}|\mathrm{x}_t,\ \mathrm{u}_t)d\tau_{u_{t+1}}d\mathrm{x}_{t+1} \\
&= \int_{x_{t+1}}\left[\int_{\tau_{u_{t+1}}}\left(\sum_{k=t+1}^{T}(c(\mathrm{x}_k,\ \mathrm{u}_k)+\log\pi(\mathrm{u}_k|\mathrm{x}_k))\right)p(\tau_{u_{t+1}}|\mathrm{x}_{t+1})d\tau_{u_{t+1}}\right]p(\mathrm{x}_{t+1}|\mathrm{x}_t,\ \mathrm{u}_t)d\mathrm{x}_{t+1}
\end{aligned}
\tag{8.78}
$$

대괄호는 $V(x_{t+1})$ 이므로

$$Q_1 = \int_{x_{t+1}} V(x_{t+1})p(x_{t+1}|x_t, u_t)dx_{t+1}$$

따라서 소프트 행동가치 함수는 다음과 같이 된다.

$$
\begin{aligned}
Q(\mathrm{x}_t,\ \mathrm{u}_t) &= c(\mathrm{x}_t,\ \mathrm{u}_t)+\int_{x_{t+1}} V(\mathrm{x}_{t+1})p(\mathrm{x}_{t+1}|\mathrm{x}_t,\ \mathrm{u}_t)d\mathrm{x}_{t+1} \\
&= c(\mathrm{x}_t,\ \mathrm{u}_t)+E_{x_{t+1}\sim p(x_{t+1}|x_t,u_t)}[V(\mathrm{x}_{t+1})]
\end{aligned}
\tag{8.80}
$$

위의 식을 식(8.75)에 대입하면 다음과 같다.

$$V(x_t) = E_{u_t \sim \pi(u_t|x_t)}\left[log\,\pi(u_t|x_t) + c(x_t, u_t) + E_{u_{t+1} \sim p(x_{t+1}|x_t, u_t)}\left[V(x_{t+1})\right]\right]$$

벨만의 최적성 원리에 의하면 현재 시간 스텝 t 에서 최종 시간스텝 T 까지 최소의 비용함수를 실현 하는 현재 시간스텝 t 의 최적제어는 다음 식을 만족해야 하다.

$$V(x_t) = \min_{\pi} E_{u_t \sim \pi(u_t|x_t)} \left[\log \pi(u_t|x_t) + c(x_t, u_t) + E_{u_{t+1} \sim p(x_{t+1}|x_t, u_t)} \left[V(x_{t+1}) \right] \right]$$

최종 시간인 $t = T$ 에서는 소프트 상태 가치 함수는 다음과 같다.

$$\begin{aligned} V(\mathrm{x}_T) &= E_{\mathrm{u}_T \sim p(\mathrm{u}_T|\mathrm{x}_T)} \left[\log \pi(\mathrm{u}_T|\mathrm{x}_T) + \frac{1}{2} \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{C}_T \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix} + \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{c}_T \right] \\ &= E_{\mathrm{u}_T \sim p(\mathrm{u}_T|\mathrm{x}_T)} \left[\log \pi(\mathrm{u}_T|\mathrm{x}_T) + \frac{1}{2} \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{Q}_T \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix} + \begin{bmatrix} \mathrm{x}_T \\ \mathrm{u}_T \end{bmatrix}^T \mathrm{q}_T \right] \\ &= E_{\mathrm{u}_T \sim p(\mathrm{u}_T|\mathrm{x}_T)} \left[\log \pi(\mathrm{u}_T|\mathrm{x}_T) + Q(\mathrm{x}_T, \mathrm{u}_T) \right] \end{aligned} \qquad (8.83)$$

여기서 $Q_T = C_T$, $q_T = c_T$ 로 놓았다. 최적정책을 가우시안 확률민도 함수로 가정했으므로 소프트 상태 가치 함수는 다음과 같이 전개 된다.

$$\begin{aligned} V(\mathrm{x}_T) = E_{\mathrm{u}_T \sim \pi(\mathrm{u}_T|\mathrm{x}_T)} & \left[\begin{aligned} & -\frac{m}{2} \log(2\pi) - \frac{1}{2} \log(\det \mathrm{S}_T) \\ & -\frac{1}{2} (\mathrm{u}_T - \mu_T)^T \mathrm{S}_T^{-1} (\mathrm{u}_T - \mu_T) \\ & + \frac{1}{2} \mathrm{x}_T^T Q_{\mathrm{xx}T} \mathrm{x}_T + \mathrm{x}_T^T Q_{\mathrm{xu}T} \mathrm{u}_T + \frac{1}{2} \mathrm{u}_T^T Q_{\mathrm{uu}T} \mathrm{u}_T \\ & + \mathrm{x}_T^T Q_{\mathrm{x}T} + \mathrm{u}_T^T Q_{\mathrm{u}T} \end{aligned} \right] \\ = & -\frac{m}{2} \log(2\pi) - \frac{1}{2} \log(\det \mathrm{S}_T) - \frac{1}{2} tr(\mathrm{I}) \\ & + \frac{1}{2} \mathrm{x}_T^T Q_{\mathrm{xx}T} \mathrm{x}_T + \mathrm{x}_T^T Q_{\mathrm{xu}T} \mu_T \\ & + \frac{1}{2} \mu_T^T Q_{\mathrm{uu}T} \mu_T + \frac{1}{2} tr(Q_{\mathrm{uu}T} \mathrm{S}_T) \\ & + \mathrm{x}_T^T Q_{\mathrm{x}T} + \mu_T^T Q_{\mathrm{u}T} \end{aligned} \qquad (8.84)$$

여기서 m 은 u_T 의 차원이고 I 는 단위행렬이며

$$Q_T = \begin{bmatrix} Q_{xxT} & Q_{xuT} \\ Q_{uxT} & Q_{uuT} \end{bmatrix}$$
$$q_T = \begin{bmatrix} Q_{xT} \\ Q_{uT} \end{bmatrix}$$

소프트 상태가치 함수를 최소로 만드는 최적 정책은 다음과 같이 가우시안 분포의 평균과 공분산에 대해서 각각 미분하여 구할 수 있다.

$$\frac{\partial V(\mathrm{x}_T)}{\partial \mu_T} = Q_{\mathrm{uu}T} \mu_T + Q_{\mathrm{ux}T} \mathrm{x}_T + Q_{\mathrm{u}T} = 0 \qquad (8.86)$$

$$\frac{\partial V(\mathrm{x}_T)}{\partial \mathrm{S}_T} = -\frac{1}{2} \mathrm{S}_T^{-1} + \frac{1}{2} Q_{\mathrm{uu}T} = 0$$

그러면 μ_T^* 와 S_T^* 는 다음과 같이 구해진다.

$$\mu^{*}_{T} = -Q^{-1}_{uuT} Q_{uxT} x_T - Q^{-1}_{uuT} Q_{uT}$$
$$S^{*}_{T} = Q^{-1}_{uuT} \tag{8.87}$$

칼만 게인을 다음과 같이 정의를 하면

$$K_T = -Q^{-1}_{uuT} Q_{uxT}$$
$$k_T = -Q^{-1}_{uuT} Q_{uT} \tag{8.88}$$

t=T에서의 최적 가우시안 LQR의 평균제어는 다음과 같이 쓸 수 있다.

$$\mu^{*}_{T} = K_T x_T + k_T$$

위 식을 이용해 $t = T$에서의 최소 소프트 상태가치 값을 구하면 다음과 같다.

$$\begin{aligned} V^{*}(x_T) &= \frac{1}{2} x_T^T Q_{xxT} x_T + x_T^T Q_{xuT} K_T x_T + x_T^T Q_{xuT} k_T \\ &\quad + \frac{1}{2} x_T^T K_T^T Q_{uuT} K_T x_T + x_T^T K_T^T Q_{uuT} k_T \\ &\quad + x_T^T Q_{xT} + x_T^T K_T^T Q_{uT} + const \\ &= \frac{1}{2} x_T^T V_T x_T + x_T^T v_T + const \end{aligned} \tag{8.90}$$

여기서

$$\begin{aligned} V_T &= Q_{xxT} + Q_{xuT} K_T + K_T^T Q_{uxT} + K_T^T Q_{uuT} K_T \\ v_T &= Q_{xT} + K_T^T Q_{uT} + Q_{xuT} k_T + K_T^T Q_{uuT} k_T \end{aligned} \tag{8.91}$$

다음으로 시간의 다음 역방향 단계인 $t = T - 1$ 에서는 소프트 상태가치 함수가 다음과 같다.

$$\begin{aligned} V(x_{T-1}) &= E_{u_{T-1} \sim \pi(x_{T-1}|x_{T-1})} \left[\begin{aligned} &\log \pi(u_{T-1}|x_{T-1}) + \\ &+ \frac{1}{2} \begin{bmatrix} x_{T-1} \\ u_{T-1} \end{bmatrix}^T C_{T-1} \begin{bmatrix} x_{T-1} \\ u_{T-1} \end{bmatrix} + \begin{bmatrix} x_{T-1} \\ u_{T-1} \end{bmatrix}^T c_T \\ &+ \frac{1}{2} x_T^T V_T x_T + x_T^T v_T + const \end{aligned} \right] \\ &= E_{u_{T-1} \sim \pi(x_{T-1}|x_{T-1})} \left[\begin{aligned} &\log \pi(u_{T-1}|x_{T-1}) + \\ &+ \frac{1}{2} \begin{bmatrix} x_{T-1} \\ u_{T-1} \end{bmatrix}^T Q_{T-1} \begin{bmatrix} x_{T-1} \\ u_{T-1} \end{bmatrix} + \begin{bmatrix} x_{T-1} \\ u_{T-1} \end{bmatrix}^T q_{T-1} + const \end{aligned} \right] \\ &= E_{u_{T-1} \sim \pi(x_{T-1}|x_{T-1})} \left[\log \pi(u_{T-1}|x_{T-1}) + Q(x_{T-1}, u_{T-1}) \right] \end{aligned} \tag{8.92}$$

여기서

$$
\begin{aligned}
Q_{T-1} &= C_{T-1} + F_{T-1}^T V_T F_{T-1} \\
q_{T-1} &= c_{T-1} + F_{T-1}^T V_T F_{T-1} + F_{T-1}^T v_T
\end{aligned}
$$

최적정책이 가우시안 확률밀도 함수이므로 소프트 상태가치 함수는 다음과 같이 전개된다.

$$
\begin{aligned}
V(\mathrm{x}_{T-1}) &= E_{u_{T-1} \sim \pi(u_T|x_T)} \left[\begin{aligned} & -\frac{1}{2}\log(\det \mathrm{S}_{T-1}) - \frac{1}{2}(\mathrm{u}_{T-1} - \mu_{T-1})^T \mathrm{S}_T^{-1} (\mathrm{u}_{T-1} - \mu_{T-1}) \\ & + \frac{1}{2}\mathrm{x}_{T-1}^T Q_{\mathrm{xx}T-1}\mathrm{x}_{T-1} + \mathrm{x}_{T-1}^T Q_{\mathrm{xu}T-1}\mathrm{u}_{T-1} + \frac{1}{2}\mathrm{u}_{T-1}^T Q_{\mathrm{uu}T-1}\mathrm{u}_{T-1} \\ & + \mathrm{x}_{T-1}^T Q_{\mathrm{x}T-1} + \mathrm{u}_{T-1}^T Q_{\mathrm{u}T-1} + const \end{aligned} \right] \\
&= -\frac{1}{2}\log(\det \mathrm{S}_{T-1}) - \frac{1}{2}tr(\mathrm{I}) + \frac{1}{2}\mathrm{x}_{T-1}^T Q_{\mathrm{xx}T-1}\mathrm{x}_{T-1} + \mathrm{x}_{T-1}^T Q_{\mathrm{xu}T-1}\mu_{T-1} \\
&\quad + \frac{1}{2}\mu_{T-1}^T Q_{\mathrm{uu}T-1}\mu_{T-1} + \frac{1}{2}tr(Q_{\mathrm{uu}T-1}\mathrm{S}_{T-1}) \\
&\quad + \mathrm{x}_{T-1}^T Q_{\mathrm{x}T-1} + \mu_{T-1}^T Q_{\mathrm{u}T-1} + const
\end{aligned}
\tag{8.94}
$$

여기서

$$
\begin{aligned}
Q_{T-1} &= \begin{bmatrix} Q_{xxT-1} & Q_{xuT-1} \\ Q_{uxT-1} & Q_{uuT-1} \end{bmatrix} \\
q_{T-1} &= \begin{bmatrix} Q_{xT-1} \\ Q_{uT-1} \end{bmatrix}
\end{aligned}
$$

평균과 공분산에 대해서 각각 미분하여 계산하면

$$
\begin{aligned}
\frac{\partial V(\mathrm{x}_{T-1})}{\partial \mu_{T-1}} &= Q_{\mathrm{uu}T-1}\mu_{T-1} + Q_{\mathrm{ux}T-1}\mathrm{x}_{T-1} + Q_{\mathrm{u}T-1} = 0 \\
\frac{\partial V(\mathrm{x}_{T-1})}{\partial \mathrm{S}_{T-1}} &= -\frac{1}{2}\mathrm{S}_{T-1}^{-1} + \frac{1}{2}Q_{\mathrm{uu}T-1} = 0
\end{aligned}
\tag{8.96}
$$

그러면 μ_{T-1}^* 와 S_{T-1}^* 는 다음과 같이 구해진다.

$$
\begin{aligned}
\mu_{T-1}^* &= -Q_{\mathrm{uu}T-1}^{-1} Q_{\mathrm{ux}T-1}\mathrm{x}_{T-1} - Q_{\mathrm{uu}T-1}^{-1} Q_{\mathrm{u}T-1} \\
\mathrm{S}_{T-1}^* &= Q_{\mathrm{uu}T-1}^{-1}
\end{aligned}
\tag{8.97}
$$

칼만 게인을 다음과 같이 정의를 하면

$$
\begin{aligned}
\mathrm{K}_{T-1} &= -Q_{\mathrm{uu}T-1}^{-1} Q_{\mathrm{ux}T-1} \\
\mathrm{k}_{T-1} &= -Q_{\mathrm{uu}T-1}^{-1} Q_{\mathrm{u}T-1}
\end{aligned}
\tag{8.98}
$$

t=T-1에서의 최적 가우시안 LQR의 평균제어는 다음과 같이 쓸 수 있다.

$$
\mu_{T-1}^* = K_{T-1} x_{T-1} + k_{T-1}
$$

위의 식을 이용해 $t = T - 1$ 에서의 최소 소프트 상태가치 값을 구하면 다음과 같다.

$$
\begin{aligned}
V^*(\mathrm{x}_{T-1}) &= \frac{1}{2}\mathrm{x}_{T-1}^T Q_{\mathrm{xx}T-1}\mathrm{x}_{T-1} + \mathrm{x}_{T-1}^T Q_{\mathrm{xu}T-1}\mathrm{K}_{T-1}\mathrm{x}_{T-1} + \mathrm{x}_{T-1}^T Q_{\mathrm{xu}T-1}\mathrm{k}_{T-1} \\
&\quad + \frac{1}{2}\mathrm{x}_{T-1}^T \mathrm{K}_{T-1}^T Q_{\mathrm{uu}T-1}\mathrm{K}_{T-1}\mathrm{x}_{T-1} + \mathrm{x}_{T-1}^T \mathrm{K}_{T-1}^T Q_{\mathrm{uu}T-1}\mathrm{k}_{T-1} \\
&\quad + \mathrm{x}_{T-1}^T Q_{\mathrm{x}T-1} + \mathrm{x}_{T-1}^T \mathrm{K}_{T-1}^T Q_{\mathrm{u}T-1} + const \\
&= \frac{1}{2}\mathrm{x}_{T-1}^T \mathrm{V}_{T-1}\mathrm{x}_{T-1} + \mathrm{x}_{T-1}^T \mathrm{v}_{T-1} + const
\end{aligned}
\tag{8.100}
$$

여기서

$$
\begin{aligned}
\mathrm{V}_{T-1} &= Q_{\mathrm{xx}T-1} + Q_{\mathrm{xu}T-1}\mathrm{K}_{T-1} + \mathrm{K}_{T-1}^T Q_{\mathrm{ux}T-1} + \mathrm{K}_{T-1}^T Q_{\mathrm{uu}T-1}\mathrm{K}_{T-1} \\
\mathrm{v}_{T-1} &= Q_{\mathrm{x}T-1} + \mathrm{K}_{T-1}^T Q_{\mathrm{u}T-1} + Q_{\mathrm{xu}T-1}\mathrm{k}_{T-1} + \mathrm{K}_{T-1}^T Q_{\mathrm{uu}T-1}\mathrm{k}_{T-1}
\end{aligned}
\tag{8.101}
$$

위의 전개식을 확정적 LQR과 비교를 해보면

- 가우시안 정책이 산출하는 평균은 확정적 LQR의 제어값과 동일하고
- 공분산은 $S_t = Q^- 1_{uut}$ 임을 알수 있다.

$$\pi(u_t|x_t) = N(K_t x_t + k_t, Q^- 1_u ut)$$

8.2.4 반복적 LQR (iLQR)

- 비선형 시스템에 LQR를 적용한 것
- LQR은 선형 시스템과 2차 함수로 된 비용함수에만 적용할 수 있다.
- iLQR은 현재의 궤적을 기준으로 비선형 시스템을 1차 시스템으로 근사하고 비용함수를 2차 함수로 근사한 후에 LQR을 적용하는 방법
- 궤적이 수렴할 때 까지 반복한다.

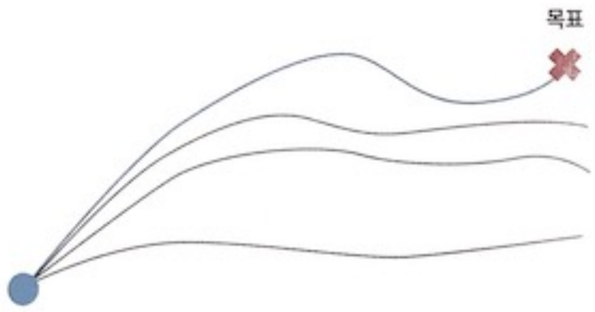

그림 8.3 반복적 LQR의 아이디어

다음과 같은 비선형 이산시간 시스템을 생각해 보자

$x_{t+1} = f(x_t, u_t)$

다음과 같은 목적 함수를 가정한다.

$J_0 = \sum_{t=0}^{T} c(x_t, u_t)$

테일러 시리즈를 이용해 명목(norminal) 궤적 $(\hat{x_t}, \hat{u_t})$ 를 기준으로 비선형 시스템을 선형화해 보자

$$
\begin{aligned}
x_{t+1} &= f(x_t, u_t) \\
&\approx f(\hat{x}_t, \hat{u}_t) + f_{x_t}(x_t - \hat{x}_t) + f_{u_t}(u_t - \hat{u}_t) \\
&= f_{x_t} x_t + f_{u_t} u_t + f(\hat{x}_t, \hat{u}_t) - f_{x_t}\hat{x}_t - f_{u_t}\hat{u}_t \\
&= f_{x_t} x_t + f_{u_t} + f_{c_t}
\end{aligned}
$$

여기서 $f_{x_t} = \triangledown_{x_t} f(\hat{x}_t, \hat{u}_t), \qquad f_{u_t} = \triangledown_{u_t} f(\hat{x}_t, \hat{u}_t), \qquad f_{c_t} = f(\hat{x}_t, \hat{u}_t) - f_{x_t}\hat{x}_t - f_{u_t}\hat{u}_t$ 이다.

용어 설명 비선형 시스템을 테일러 시리즈를 이용해 선형화할 때 기준이 되는 궤적을 명목궤적이라고 한다. 실제 궤적은 명목궤적과 큰 차이가 나지 않는 주변에 있다고 가정해 실제 궤적 x_t를 명목궤적 $\bar{\mathrm{x}}_t$와 섭동(perturbation) $\triangle \mathrm{x}_t$의 합으로 표현한다. 즉, 비선형 동적 시스템을 $\mathrm{y}_t = \mathrm{h}(\mathrm{x}_t)$라고 할 때 명목궤적 $\bar{\mathrm{x}}_t$를 기준으로 전개하면 다음과 같다.

$$
\begin{aligned}
\mathrm{y}_t &= \mathrm{h}(\mathrm{x}_t) = \mathrm{h}(\bar{\mathrm{x}}_t + \triangle \mathrm{x}_t) \\
&= \mathrm{h}(\bar{\mathrm{x}}_t) + \left.\frac{d\mathrm{h}}{d\mathrm{x}}\right|_{\mathrm{x}_t=\bar{\mathrm{x}}_t} \triangle \mathrm{x}_t + H.O.T. \\
&\approx \mathrm{h}(\bar{\mathrm{x}}_t) + \left.\frac{d\mathrm{h}}{d\mathrm{x}}\right|_{\mathrm{x}_t=\bar{\mathrm{x}}_t} \triangle \mathrm{x}_t
\end{aligned}
$$

여기서 $H.O.T.$는 $\triangle \mathrm{x}_t$의 고차항을 나타내며, $\triangle \mathrm{x}_t$가 작다고 가정하고 무시한다. 또한 $\left.\frac{d\mathrm{h}}{d\mathrm{x}}\right|_{\mathrm{x}_t=\bar{\mathrm{x}}_t} = \nabla_{\mathrm{x}} \mathrm{h}(\bar{\mathrm{x}}_t)$를 자코비안(Jacobian) 행렬이라고 한다.

비용함수도 2차함수로 근사해보자.

$$
\begin{aligned}
J_0 &\approx \sum_{t=0}^{T} \left(c_t(\hat{\mathrm{x}}_t,\ \hat{\mathrm{u}}_t) + \mathrm{b}_t^T \begin{bmatrix} \mathrm{x}_t - \hat{\mathrm{x}}_t \\ \mathrm{u}_t - \hat{\mathrm{u}}_t \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \mathrm{x}_t - \hat{\mathrm{x}}_t \\ \mathrm{u}_t - \hat{\mathrm{u}}_t \end{bmatrix}^T \mathrm{C}_t \begin{bmatrix} \mathrm{x}_t - \hat{\mathrm{x}}_t \\ \mathrm{u}_t - \hat{\mathrm{u}}_t \end{bmatrix} \right) \\
&= \sum_{t=0}^{T} \left(\begin{bmatrix} \mathrm{x}_t \\ \mathrm{u}_t \end{bmatrix}^T \mathrm{b}_t + \frac{1}{2} \begin{bmatrix} \mathrm{x}_t \\ \mathrm{u}_t \end{bmatrix}^T \mathrm{C}_t \begin{bmatrix} \mathrm{x}_t \\ \mathrm{u}_t \end{bmatrix} - \begin{bmatrix} \mathrm{x}_t \\ \mathrm{u}_t \end{bmatrix}^T \mathrm{C}_t \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix} \right) \\
&\qquad + \sum_{t=0}^{T} \left(c_t(\hat{\mathrm{x}}_t,\ \hat{\mathrm{u}}_t) + \frac{1}{2} \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix}^T \mathrm{C}_t \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix} - \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix}^T \mathrm{b}_t \right) \\
&= \sum_{t=0}^{T} \left(\frac{1}{2} \begin{bmatrix} \mathrm{x}_t \\ \mathrm{u}_t \end{bmatrix}^T \mathrm{C}_t \begin{bmatrix} \mathrm{x}_t \\ \mathrm{u}_t \end{bmatrix} + \begin{bmatrix} \mathrm{x}_t \\ \mathrm{u}_t \end{bmatrix}^T \mathrm{c}_t + \mathrm{d}_t \right)
\end{aligned}
$$

여기서

$$
\mathrm{C}_t = \nabla^2_{\mathrm{x}_t, \mathrm{u}_t} c_t(\hat{\mathrm{x}}_t,\ \hat{\mathrm{u}}_t),\ \mathrm{b}_t = \nabla_{\mathrm{x}_t, \mathrm{u}_t} c_t(\hat{\mathrm{x}}_t,\ \hat{\mathrm{u}}_t)
$$

$$
\mathrm{c}_t = \mathrm{b}_t - \mathrm{C}_t \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix},
$$

$$
\begin{aligned}
\mathrm{d}_t &= c_t(\hat{\mathrm{x}}_t,\ \hat{\mathrm{u}}_t) + \frac{1}{2} \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix}^T \mathrm{C}_t \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix} - \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix}^T \mathrm{b}_t \\
&= c_t(\hat{\mathrm{x}}_t,\ \hat{\mathrm{u}}_t) - \frac{1}{2} \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix}^T \mathrm{C}_t \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix} - \begin{bmatrix} \hat{\mathrm{x}}_t \\ \hat{\mathrm{u}}_t \end{bmatrix}^T \mathrm{c}_t
\end{aligned}
$$

d_t 는 상수 항으로서 2차 함수로 근사된 목적함수의 최적화에 영향을 미치지 않는다.

이제 선형화된 시스템과 2차 함수로 근사된 목적함수에 LQR 알고리즘을 적용할 수 있다.

- LQR의 역방향 패스를 적용하면 칼반 게인 시퀀스 $(K_0, K_1, \ldots, K_T)$ 와 $(k_0, k_1, \ldots, k_T)$ 를 계산할 수 있다.
- 그런 다음, 순방향 패스를 계산하면 새로운 최적 제어 $(u_0, u_1, \ldots, u_t)$ 와 같은 상태 시퀀스 $(x_0, x_1, \ldots, x_T)$ 를 얻을 수 있다.
- 그리고 명목 제어 시퀀스와 상태 시퀀스를 다음과 같이 업데이트 한 후, 위 과정을 수렴할 때 까지 반복하면 된다.

$$
\begin{aligned}
(\hat{u_0}, \hat{u_1}, \ldots, \hat{u_T}) &\leftarrow (u_0, u_1, \ldots, u_T) \\
(\hat{x_0}, \hat{x_1}, \ldots, \hat{x_T}) &\leftarrow (x_0, x_1, \ldots, x_T)
\end{aligned}
$$

1. 초기 상태 x_0에 대해서 명목 제어 $(\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_T)$와 명목 상태 $(\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_T)$ 초기화

2. Repeat {

[1] $\hat{J}_0 = \sum_{t=0}^{T} c(\hat{x}_t, \hat{u}_t)$ 계산

[2] $f_{xt} = \nabla_{x_t} f(\hat{x}_t, \hat{u}_t)$. $f_{ut} = \nabla_{u_t} f(\hat{x}_t, \hat{u}_t)$. $f_{ct} = f(\hat{x}_t, \hat{u}_t) - f_{xt}\hat{x}_t - f_{ut}\hat{u}_t$ 계산

[3] $C_t = \nabla^2_{x_t, u_t} c_t(\hat{x}_t, \hat{u}_t)$, $b_t = \nabla_{x_t, u_t} c_t(\hat{x}_t, \hat{u}_t)$ 계산

[4] $c_t = b_t - C_t \begin{bmatrix} \hat{x}_t \\ \hat{u}_t \end{bmatrix}$ 계산

[5] LQR 역방향 패스

[6] LQR 순방향 패스

for t = 0 : T {

(1) $u_t = K_t x_t + k_t$ 계산

(2) $x_{t+1} = f(x_t, u_t)$ 계산

} end

[7] 명목 궤적과 제어 업데이트

$$(\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_T) \leftarrow (u_0, u_1, \ldots, u_T)$$
$$(\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_T) \leftarrow (x_0, x_1, \ldots, x_T)$$

[8] 목적함수를 계산한다.

$$J_0 = \sum_{t=0}^{T} c_t(x_t, u_t)$$

} 수렴할 때($|\hat{J}_0 - J_0| \le \epsilon$)까지 반복

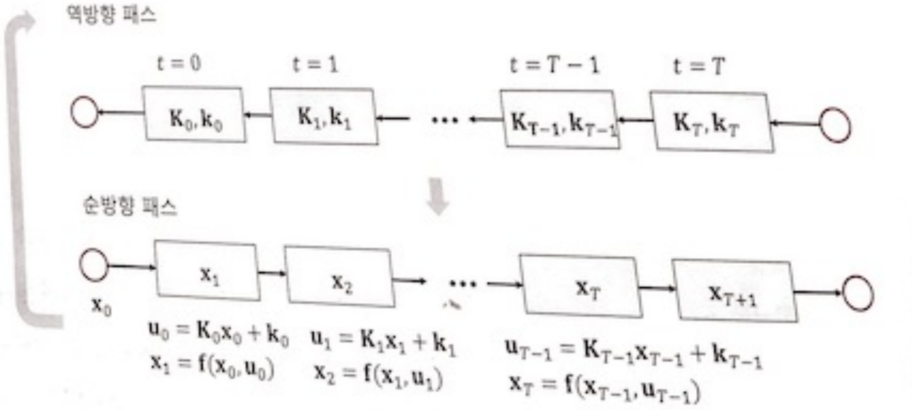
역방향 패스
t = 0
t = 1
t = T − 1
t = T
K_0, k_0
K_1, k_1
K_{T-1}, k_{T-1}
K_T, k_T
순방향 패스
x_1
x_2
x_T
x_{T+1}
x_0
$u_0 = K_0x_0 + k_0$
$x_1 = f(x_0, u_0)$
$u_1 = K_1x_1 + k_1$
$x_2 = f(x_1, u_1)$
$u_{T-1} = K_{T-1}x_{T-1} + k_{T-1}$
$x_T = f(x_{T-1}, u_{T-1})$

그림 8.4 반복적 LQR

8.3 모델 학습 방법

모델 강화 학습에서는 에이전트 환경과 상호 작용 하면서 얻은 샘플로 시스템의 동역학 모델을 지도학습 한다.

가장 간단한 알고리즘

1. 랜덤 정책이나 기본 정책을 실행해 상태천이 데이터세트 $D = \{(X_t, u_t, x_{t+1})_j\}$ 를 수집한다.
2. $x_{t+1} \approx f(x_t, u_t)$ 가 되도록 함수 $f(x_t, u_t)$ 를 학습한다.
3. 동역학 모델을 이용해 정책을 계산한다.

- 시스템의 동역하 구조는 알지만 일부 파라미터 값이 불확실한 경우에 유용
- <u>데이터를 수집하는 데 사용한 정책과 추정된 모델을 이용해 계산하는 정책이 다르다.</u>

이 책에서 사용할 모델 학습 방법

1. 랜덤 정책이나 기본 정책을 실행해 상태천이 데이터세트 $D = \{(X_t, u_t, x_{t+1})_j\}$ 를 수집한다.
2. $x_{t+1} \approx f(x_t, u_t)$ 가 되도록 함수 $f(x_t, u_t)$ 를 학습한다.
3. 동역학 모델을 이용해 정책을 계산한다.
4. 계산된 정책을 실행해 새로운 궤적을 발생시키고 상태 천이 데이터세트 D에 추가한다.
5. 2번으로 돌아가 절차를 반복한다.

시스템의 동역학 모델 $x_{t+1} \approx f(x_t, u_t)$ 를 사용할 수 있지만, 가우시안 프로세스, 가우시안 혼합모델 (GMM), **신경망**과 같은 일반적인 모델로도 표현할 수 있다.

글로벌 모델

- 모든 상태 공간에서 작동하는 단일 모델
- 상태 공간 전체에서 연속성을 갖고 있다는 장점이 있다.
- **시스템의 운동이 매우 복잡하다면** 전체 상태 공간에서 매우 복잡한 운동 모델을 고려해야 하고, 이 모델을 학습하기 위해서는 **많은 데이터**가 필요할 것이다.
- 가우시안 프로세스나 신경망 등으로 모델링

로컬 모델

- 상태 공간 일부에서만 작동하는 모델
- 단순하기 때문에 적은 수의 데이터세트 만으로도 추정하기 쉽다는 장점이 있다.
- 로컬 모델 기반으로 계산한 정책이 업데이트가 되면 해당 모델이 부정확해진다는 단점이 있다. $\rightarrow$ 명목 궤적 근처에서만 국지적으로 유효하기 때문

- 이 단점을 극복한다면 로컬 모델은 모델 기반 강화학습 기법을 실제 문제에 적용하는데 있어서 매우 유용한 수단이 될 수 있다.