CS 4/5789 9 Feb 2022 Prof Sarah Dean

Lecture 6: The Linear Quadratic Regulator

Last time, we discussed finite horizon MDPs with continuous state & action spaces. We also introduced linear dynamics (transitions). Today we consider a continuous MDP problem with linear dynamics and quadratic costs: the Linear Quadratic Regulator.

1) LQR LQR(A, B, Q R)

For this continuous MDP,

$\mathcal{S} = \mathbb{R}^{n_s}$, $\mathcal{A} = \mathbb{R}^{n_a}$, $f(s_t, a_t, w_t) = As_t + Ba_t + w_t$

$w_t \sim \mathcal{N}(0, \sigma^2 I)$

$c(s, u) = s^T Q s + u^T R u$ quadratic costs, $Q, R \succ 0$ symetric.

Horizon H & initial distribution μ_0.

Putting this all together, Optimal Control Problem:

$$\min_{\pi} \mathbb{E}\left[s_H^T Q s_H + \sum_{t=0}^{H-1} s_t^T Q s_t + a_t^T R a_t \;\middle|\; \begin{array}{l} s_{t+1} = As_t + Ba_t + w_t,\ x_0 \sim \mu_0 \\ a_t = \pi_t(s_t),\ w_t \sim \mathcal{N}(0, \sigma^2 I) \end{array} \right]$$

Example: 1d robot from last lecture

$s_t = \begin{bmatrix} p_t \\ v_t \end{bmatrix}$ position, velocity $s_{t+1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} s_t + \begin{bmatrix} 0 \\ 1/m \end{bmatrix} a_t$

goal: move to goal position $p=0$ and be still $v=0$ without using too much force can use quatratic cost:

$$c(s,a) = \gamma_p p^2 + \gamma_v v^2 + \gamma_a a^2$$

$$= s^T \underbrace{\begin{bmatrix} \gamma_p & 0 \\ 0 & \gamma_v \end{bmatrix}}_{Q} s + \underbrace{\gamma_a}_{R} a^2$$

Depending on the relative weighting of γ_p, γ_v, and γ_a, optimal policy will be more or less aggressive.

Value and Q functions "cost to go"

$$V_t^{\pi}(s) = \mathbb{E}\left[s_H^T Q s_H + \sum_{k=t}^{H-1} s_k^T Q s_k + a_k^T R a_k \;\middle|\; \begin{array}{l} s_{k+1} = As_k + Ba_k + w_k \\ a_k = \pi_k(s_k) \\ w_k \sim \mathcal{N}(0, \sigma^2 I) \\ s_t = s \end{array} \right]$$

$$Q_t^{\pi}(s,a) = \mathbb{E}\left[s_H^T Q s_H + \sum_{k=t}^{H-1} s_k^T Q s_k + a_k^T R a_k \;\middle|\; \begin{array}{l} s_{k+1} = As_k + Ba_k + w_k \\ a_k = \pi_k(s_k) \\ w_t \sim \mathcal{N}(0, \sigma^2 I) \\ s_t = s,\ a_t = a \end{array} \right]$$

$$V_t^{\pi}(s) = Q_t^{\pi}(s, \pi(s))$$

Notice that due to termal cost, $V_H(s)$ is nonzero

2) Optimal LQR Policy

We can derive the optimal LQR Policy via Dynamic Programming.

DP For optimal control:

$$\pi^* = (\pi^*_0, \pi^*_1, \dots, \pi^*_{H-1})$$

Start: $V^*_H(s) = c_H(s)$

(due to terminal cost, initialize $V^*_H(s) \neq 0$ unlike before, but the algorithm is the same).

for $t = H-1, H-2, \dots, 0$:

$$Q^*_t(s,a) = c_t(s,a) + \mathbb{E}_{s' \sim P(s,a)}\left[V^*_{t+1}(s')\right]$$

$= Q^*_{t+1}(s, \pi^*_{t+1}(s))$

$$\pi^*_t = \arg\min_a Q^*_t(s,a)$$

using "mimize cost" convention

Theorem (LQR optimal value fn. & Policy):

For LQR(A, B, Q, R), The optimal value function is quadratic:

$$V^*_t(s) = s^T P_t s + p_t$$

and the optimal policy is linear

$$\pi^*_t(s) = -K^*_t s$$

where (P_t, p_t, K^*_t) can be computed exactly from (A, B, Q, R).

<u>Proof:</u> We prove by induction, using DP.

<u>claim 1</u>: (Base case) $V_H^*(s) = s^T P_H s + p_H$ is quadratic. $\forall s$

<u>claim 2</u>: (induction) Assume $V_{t+1}^*(s) = s^T P_{t+1} s + p_{t+1}$ $\forall s$. Then

1) $Q_t^*(s,a)$ is quadratic in s, a

2) $\pi_t^*(a) = \arg\min_a Q_t^*(s,a)$ is linear in s

Therefore, $V_t^*(s) = s^T P_t s + p_{t+1}$ is quadratic.

Then by induction, V is quatratic & π linear.

<u>Proof of Claim 1:</u>

$V_H^*(s) = c_H(s) = s^T Q s$. So $P_H = Q$, $p_H = 0$. ✓
(symmetric)

<u>Proof of claim 2:</u>

Part 1) $Q_t^*(s,a) = s^T Q s + a^T R a + \mathbb{E}_{s' \sim P(s,a)}\left[V_{t+1}^*(s')\right]$

$$\mathbb{E}_{s'}\left[V_{t+1}^*(s')\right] = \mathbb{E}_{w \sim N(0,1)}\left[V_{t+1}^*(As+Ba+w)\right]$$

$$\begin{aligned} V_{t+1}^*(As+Ba+w) = {} & (As)^T P_{t+1}(As) + (As)^T P_{t+1} Ba + (As)^T P_{t+1} w \\ & + (Ba)^T P_{t+1} As + (Ba)^T P_{t+1} Ba + (Ba)^T P_{t+1} w \\ & + w^T P_{t+1} As + w^T P_{t+1} Ba + w^T P_{t+1} w + p_{t+1} \end{aligned}$$

also P_{t+1} symmetric

once we take expectation, many terms $=0$ because $\mathbb{E}w = 0$.

$$\mathbb{E}_{s'}\left[V_{t+1}^*(s')\right] = s^T A^T P_{t+1} A s + 2 s^T A^T P_{t+1} B a + a^T B^T P_{t+1} B a + \mathbb{E}_w\left[w^T P_{t+1} w\right] + p_{t+1}$$

to simplify the remaining expectation, recall

$$W^T P W = \mathrm{Tr}(W^T P W) = \mathrm{Tr}(P W W^T).$$ ← cyclic property of trace

$$\mathbb{E}[W^T P W] = \mathbb{E}[\mathrm{Tr}(P W W^T)] = \mathrm{Tr}(P\, \mathbb{E}[W W^T])$$ ← linearity of expectation

$$= \sigma^2 \mathrm{Tr}(P)$$

Finally,

$$Q_t^*(s,a) = s^T (Q + A^T P_{t+1} A) s + a^T (R + B^T P_{t+1} B) a$$

$$+ 2 s^T A^T P_{t+1} B a + \sigma^2 \mathrm{tr}(P) + p_{t+1} \quad \checkmark$$

Done with part 1 because this is a quadratic function.

Part 2

$$\pi_t^*(s) = \underset{a}{\operatorname{argmin}}\ Q_t^*(s,a)$$

First let's derive the minimization for a generic quadratic function.

symmetric, not symmetric

$$Q(s,a) = s^T M_1 s + a^T M_2 a + 2 s^T M_3 a + c$$

minimum must occur at a critical point.

$$\nabla_a Q(s,a) = \nabla_a (s^T M_1 s) + \nabla_a (a^T M_2 a) + \nabla_a (s^T M_3 a)$$

$$= 0 + 2 M_2 a + 2 M_3^T s$$

Then $\nabla_a Q(s,a) = 0$ when $M_2 a = -M_3 s$

for now assuming invertibility, $\underbrace{a = -M_2^{-1} M_3^T s}_{\text{linear function of } s}$

Going back to $Q_t^*(s,a)$

$$M_2 = \underbrace{\underset{\succ 0}{R} + \underset{\succeq 0}{B^T P_{t+1} B}}_{\text{invertible}} \quad \& \quad M_3 = A^T P_{t+1} B$$

Therefore, $$\pi_t^*(s) = -\underbrace{(R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A}_{\text{Define } K_t^*}$$

Last peice: check that $V_t^*(s) = Q_t^*(s, \pi_t^*(s))$ is quadratic & derive equations for P_t and p_t

Rather than plug in directly, recall our general quadratic function

$$Q(s, a^*) = s^T M_1 s + s^T M_3 M_2^{-1} \cancel{M_2 M_2^{-1}} M_3^T s$$
$$-2\, s^T M_3 M_2^{-1} M_3^T s + c$$
$$= s^T \left(M_1 - M_3 M_2^{-1} M_3^T \right) s + c$$

Therefore, this is the form of $V_t^*(s)$. Plugging M_1, M_2, M_3, c in, we have

$$V_t^*(s) = s^T \Big(\overbrace{Q + A^T P_{t+1} A - A^T P_{t+1} B (R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A}^{P_t} \Big) s$$
$$+ \overbrace{\sigma^2 \operatorname{tr}(P_{t+1}) + p_{t+1}}^{p_t}$$

✓

This concludes the proof of Claim 2 and therefore the proof of Theorem. □

Collecting the iterative Definitions together:

$P_H = Q, \quad p_H = 0$

for $t = H-1, \ldots, 0$:

$$P_t = Q + A^T P_{t+1} A - A^T P_{t+1} B (R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A$$
$$p_t = p_{t+1} + \sigma^2 \operatorname{tr}(P_{t+1})$$
$$K_t^* = (R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A$$

Some straightforward Extensions:

1) time-varying costs / dynamics

e.g. $s_{t+1} = A_t s_t + B_t a_t + w_t$

$c_t(s,a) = s_t^T Q_t s_t + a_t^T R_t a_t$

2) non-stochastic disturbance

$s_{t+1} = A s_t + B a_t + w_t + v_t$

where v_t is known a priori

3) trajectory tracking

$c_t(s,a) = (s - s^*_t)^T Q (s - s^*_t) + (a - a^*_t)^T R (a - a^*_t)$

for desired trajectory $(s^*_0, a^*_0, \ldots)$ known a priori.

(this case can be reduced to case 2 if substitute $s \leftarrow s - s^*_t$ and $a \leftarrow a - a^*_t$)