01DROS
Seminar Course on Dynamic Decision Making

2017/2018

Lecture 10

25.4

Temporal Difference

Recall VI:

$$V_{k+1}(s) = \max_a \left[R(a,s) + \gamma \sum_{s'} T(s'|a,s) \cdot V_k(s') \right]$$

Temporal Difference Learning (TD):

Let we have experience (s', a, r, s). Find $\updownarrow$ update $V_{new}(s) \leftarrow V_{old}(s)$

$$V_{new}(s) = (1-\alpha) V_{old}(s) + \alpha [r + \gamma V_{OLD}(s')]$$

$$= V_{OLD}(s) + \alpha [r + \gamma V_{OLD}(s') - V_{OLD}(s)] \quad (*)$$

$(*)$ is a stochastic variant of DP (converges with prob. 1).

Reinforcement TD:

$r > \gamma V_{OLD}(s') - V_{OLD}(s) \implies$ increase $V(s)$

$r < \gamma V_{OLD}(s') - V_{OLD}(s) \implies$ decrease $V(s)$

s1	s2	s3	s4

$a = \{left, right\}$

'Left' in S1 → reward = 100 } terminal states.
'right' in S4 → 0

$V_T = [0, 0, 0, 0]$. — terminal value.

$$V_{T-1}(s) = \max_a \left[r(s,a) + \gamma \cdot \sum_{s'} T(s'|s,a) \cdot V_T(s') \right]$$
$$= \max_a \left[r(s,a) + \gamma \cdot V_T(s') \right.$$

$$V_{T-1}(s1) = \max_a \left[r(s1,a) + \gamma \cdot V_T(s0) \right] = 100 + 0.7 \cdot 0 = 100.$$
$$V_{T-1}(s2) = V_{T-1}(s3) = V_{T-1}(s4) = 0.$$
$$V_{T-2}(s1) = \max_a \left[r(s1,a) + \gamma \cdot V_{T-1}(s0) \right] = 100 + 0.7 \cdot 0 = 100.$$
$$V_{T-2}(s2) = 0 + 0.7 \cdot V_{T-1}(s1) = 0.7 \cdot 100 = 70.$$
$$V_{T-2}(s3) = V_{T-2}(s4) = 0$$
$$V_{T-3}(s1) = \max_a \left[r(s1,a) + \gamma \cdot V_{T-2}(s0) \right] = 100$$
$$V_{T-3}(s2) = \max_a \left[r(s2,a) + \gamma \cdot V_{T-2}(s1) \right] = 0 + 0.7 \cdot 100 = 70.$$
$$V_{T-3}(s3) = \max_a \left[r(s3,a) + \gamma \cdot V_{T-2}(s2) \right] = 0 + 0.7 \cdot 70 = 49$$
$$V_{T-3}(s4) = 0$$
$$V_{T-4}(s1) = 100, \quad V_{T-4}(s2) = 70, \quad V_{T-4}(s3) = 49; \quad V_{T-4}(s4) = 0 + 0.7 \cdot V_{T-3}(s3)$$
$$= 0 + 0.7 \cdot 49 = 34.3.$$

$$V = [100, 70, 49, 34.3]$$

Link between linear quadratic control and MDP

Recap: Kalman filter

- POMDP
- belief state (degree of belief in s_{t+1}) :

$$b_{t+1}(s_{t+1}) = \Pr(s_{t+1} | o_{t+1}, a_t, b_t) \propto \Pr(o_{t+1} | s_{t+1}, a_t) \sum_{s \in S} T(s_{t+1} | a_t, s_t) b_t(s_t)$$

- belief dynamics is Gaussian => Kalman filter

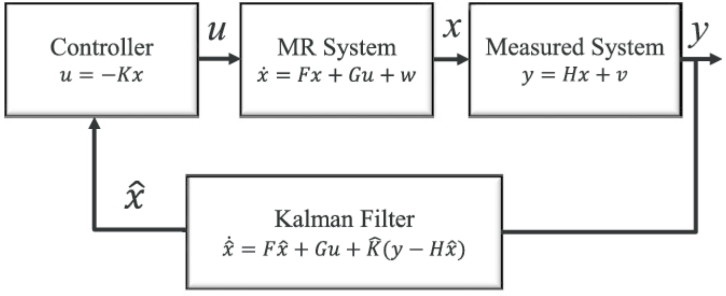
Control theory

Optimal control theory:

Given: cost function.

Problem: optimise sum of cost at any step and terminal cost.

Result: optimal control sequence and optimal state trajectory

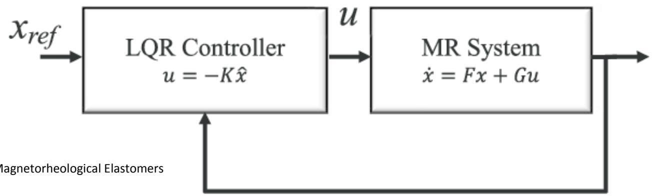


(a)

Problem in practice: what control action should we apply to move a system to a desired state?

Given: desired state (or trajectory).

Result: optimal control trajectory.



schema is from the article : P.X. Do, A State-of-Art on Control Schemes for Application Systems Featuring Magnetorheological Fluids and Magnetorheological Elastomers

Consider a discrete-time continuous state system:

Dynamics: (Linear)

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

Cost per step: (quadratic)

$$c(s_t, a_t) = s_t^T Q s_t + a_t^T R a_t \quad , R > 0 , \; Q \geq 0$$

$$c_T(s_T) = s_T^T Q s_T \quad \text{— terminal cost}$$

Control objective:

Choose $\pi(s_1 \dots s_T \; a_1 \dots a_{T-1})$ to minimise

$$\sum_{t=1}^{T-1} c(s_t, a_t) + \underbrace{c_T(s_T)}_{\text{can be assumed} = 0}$$

This is "classical" control problem, In LQR:

→ state s evolves linearly (due to A) and any deviations are corrected by a (Note: it is limited by B)

→ the cost c(s,a) penalises any deviation of either a or s.

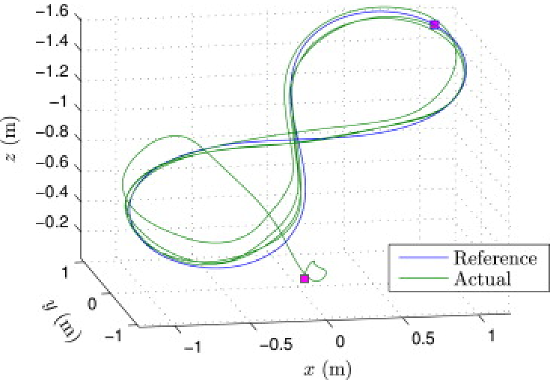
LQR (linear quadratic regulation problem) can

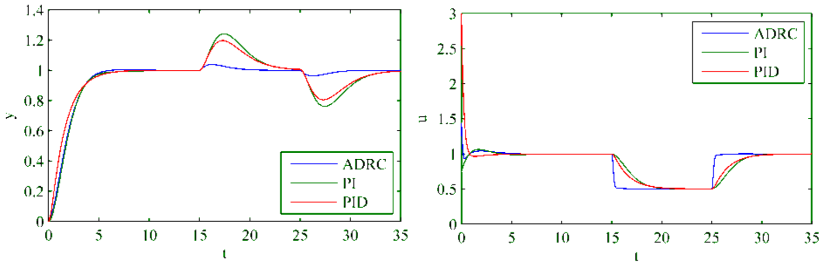
→ keep state s close to original value
regulation problem

→ keep state s close to a reference trajectory $\{\bar{s}_t\}$
tracking problem

in that case cost reads:

$$c(s_t, a_t) = \| s_t - \bar{s}_t \|_Q + \| a \|_R$$





(a) (b)

deterministic LQR and MDP

	MDP dynamic	Deterministic LQR
System Dynamics	$T(s_{t+1} \| a_t , s_t)$	$s_{t+1} = As_t + Ba_t$
DM rule (controller structure)	$a_t = \pi(s_t)$	$a_t = \pi(s_t)$
Objective function (utility)	$E[\Sigma_{t=0} c_t(s_t, a_t) + c_T(s_T) \| s_t, a_t]$	$\Sigma_{t=0} c_t(s_t, a_t) + c_T(s_T)$

Using Markov strategies does not lead to any loss of optimality.

VI: $V_{T-1}(s) = \max_{a} \left[r + \gamma \sum_{s'} T(s'|s,a) \cdot V_T(s \right.$

Consider finite-horizon problem:

$$\{a_t\}^{opt}_{t \in N} = \underset{\{a_t\}_{t \in N} \in K^N}{\operatorname{argmin}} \sum_{t=0}^{T} c(s_t, a_t), \quad \text{where } s_0 \text{ is given.}$$

VI can be used for finite-horison by optimally solving the problem at time T, then use the results for $T-1$ and so on.

At time T (terminal state): $V_T(s_T) = \min_{a_T} c(s_T, a_T) = \min_{a_T} \left(s_T^T Q s_T + a_T^T R a_T \right)$

$$= s_T^T Q s_T := s^T P_T s$$

a_T is not important (mostly it is 0)

$$VI: \quad V_{T-1}(s) = \max_a \left[r + \gamma \sum_{s'} T(s'|s,a) \cdot V_T(s') \right].$$

In our case model is deterministic $\Rightarrow$ no need to sum over s', so:

$$V_{T-1}(s) = \min_a \left[c(s,a) + V_T(s_T) \right] = \min_a \left[c(s,a) + V_T(A \cdot s + B \cdot u) \right]$$

$$= \min_a \left[s^T Q s + a^T R a + (A \cdot s + B \cdot a)^T P_T (A \cdot s + B \cdot a) \right]$$

$$= s^T Q \cdot s + (A \cdot s)^T \cdot P_T (A \cdot s) +$$

$$+ \min_a \Big[\underbrace{a^T R a + (Ba)^T P_T (Ba) + (As)^T \cdot P_T (As)}_{\Phi(a,s)} \Big]$$

$$a^{opt} = \underset{a}{\operatorname{argmin}} \; \Phi(a,s)$$

$$\Phi(a,s) = a' (R + B^T \cdot P_T B) a + (As)^T P_T \cdot Bu$$

optimal action $\quad a^{opt} = \underset{a}{\operatorname{argMin}} \; \Phi(a,s) = - (R + B^T P_T B)^{-1} \cdot B^T P_T A \cdot s$

$$V_{T-1}(s) = s^T \underbrace{\left(Q + A^T P_T A - A^T P_T B \left(R + B^T P_T B\right)^{-1} B^T \cdot P_T \cdot A\right)}_{P_{T-1}} s$$

$V_{T-1}(s) = S^T \cdot P_{T-1} \cdot S$. → recursive definition of VF which is parameterized by P_{T-1}

$$P_{T-1} = Q + A^T P_T A - A^T P_T B \left(R + B^T P_T B\right)^{-1} B^T \cdot P_T \cdot A$$

it converges to optimal VF:

$$V(s) = s^T \cdot P \cdot s.$$

Optimal policy: $\pi^{opt} = -H \cdot s$, $H = \left(Q + B^T \cdot P \cdot B\right)^{-1} B^T \cdot P \cdot A$

This result can be obtained using Riccati equation (control theory).

Riccati equations are named after count Jacopo Francesco Riccati (1676-1754) who studied the differential equations of the form $y'(t)=c_0(t)+c_1(t)\, y(t)+c_2(t)\, y^2(t)$ and its variations. In modern control, such equations arise in the calculus of variations and optimal filtering. The discrete-time version of these equations are also named after Riccati. Covariance matrices in Kalman filter are computed and are given by the forward Riccati difference equation.

Take home message

- MDP and LQR is linked: iterative vs non-iterative proof.
- Similar approach can be used in LQG (Linear-Quadratic-Gaussian) control when the controller acts on states s predicted by Kalman filter (see Kalman filter lecture)
- The optimal controller in LQG is the same as in the deterministic case. The only effect of the noise is to increase the value function (This phenomenon is unique to LQG systems).