Lecture 18 Continuous State MDP & Model Simulation

1. Notation for continuous state MDP.

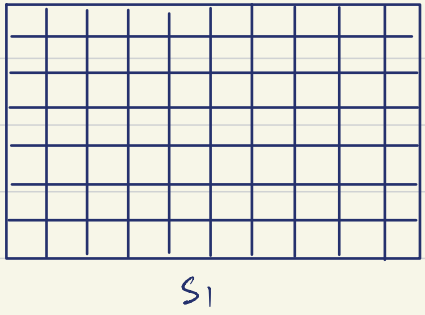
e.g. car : $S(x, y, \theta, \dot{x}, \dot{y}, \dot{\theta})$

θ: direction; $\dot{x}, \dot{y}, \dot{\theta}$: velocity on each direction

helicopter : $S(x, y, z, \phi, \theta, \psi, \dot{x}, \dot{y}, \dot{z}, \dot{\phi}, \dot{\theta}, \dot{\psi})$

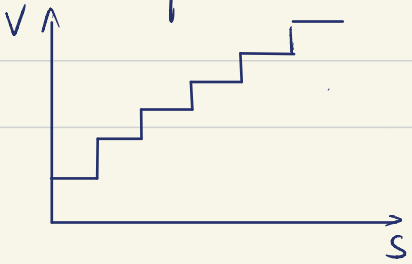
ϕ: roll; θ: pitch; ψ: yaw

2. Solution 1: Discretization



Problems:

① Naive representation for V^* (and π^*)

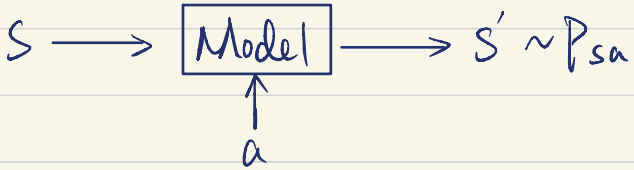


② Curse of Dimensionality

$S = \mathbb{R}^n$, and discretize each dimension into k values, get k^n discrete states.

If dimension is low (2-6), can choose to use discretization.

3. Model (simulator) of MDP:



① Physics simulator

② Learn model from data: model-based RL

$s_0^{(1)} \xrightarrow{a_0^{(1)}} s_1^{(1)} \xrightarrow{a_1^{(1)}} s_2^{(1)} \xrightarrow{a_2^{(1)}} \cdots s_T^{(1)}$

$s_0^{(2)} \xrightarrow{a_0^{(2)}} \cdots$

$s_0^{(m)} \xrightarrow{a_0^{(m)}} \cdots$

Apply supervised learning to estimate S_{t+1} as function of S_t, A_t

e.g. linear regression: $S_{t+1} = AS_t + BA_t$

$$\min_{A,B} \sum_{i=1}^{m} \sum_{t=0}^{T} \| S_{t+1}^{(i)} - (AS_t^{(i)} + BA_t^{(i)}) \|^2$$

Model:

$S_{t+1} = AS_t + BA_t$ ← deterministic (works well in simulator, but usually works bad in reality.)

or

✓ $S_{t+1} = AS_t + BA_t + \varepsilon_t$, $\varepsilon_t \sim N(0, \sigma^2 I)$ ← stochastic (random noise)

4. Fitted Value Iteration

Choose feature $\phi(s)$ of state s. $V(s) = \theta^T \phi(s)$

Previously Value Iteration:

$$V(s) := R(s) + \gamma \max_a \sum_{s'} P_{sa}(s') V(s')$$

$$= R(s) + \gamma \max_a E_{s' \sim P_{sa}}[V(s')]$$

$$y^{(i)} = \max_a \underbrace{E_{s' \sim P_{sa}}}_{q(a)} [R(s) + \gamma V(s')]$$

Sample $\{s^{(1)}, s^{(2)}, \cdots s^{(m)}\} \subseteq S$ randomly

Initialize $\theta := 0$

Repeat {

For $i = 1, \cdots, m$ {

For each action $a \in A$ {

Sample $s_1', s_2', \cdots s_k' \sim P_{s^{(i)}a}$ (using the simulated model.)

Set $q(a) = \frac{1}{k} \sum_{j=1}^{k} [R(s^{(i)}) + \gamma V(s_j')]$

$V(s_j') = \theta^T \phi(s_j')$ ← last iteration of θ

}

Set $y^{(i)} = \max_a q(a)$

}

$\theta := \arg\min_\theta \frac{1}{2} \sum_{i=1}^{m} (\theta^T \phi(s^{(i)}) - y^{(i)})^2$

Fitted VI gives approximation to V^*.

Implicitly define π^*: $\pi^* = \arg\max_a E_{s' \sim P_{sa}}[V^*(s')]$, use $s_1', s_2' \cdots s_k' \sim P_{sa}$ to approximate expectation.

* Say model is $S_{t+1} = f(S_t, a_t) + \varepsilon_t$

For deployment (run-time): set $\varepsilon_t = 0$, and $k = 1$

$\Rightarrow$ When in state s, pick action $\arg\max_a V(f(s,a))$

$s' \sim P_{sa}$

↖ simulation without noise