

REINFORCEMENT

LEARNING

INTRODUCTION

REINFORCEMENT LEARNING

- Born in Computer Science community
- Many similarities with Optimal Control, but RL
 - tries to find globally optimal policy
 - assumes dynamics is unknown
 - initially focused on finite size state/control spaces
 - uses different terminology
 - typically assumes dynamics is stochastic
 - we will not in this course, to simplify equations
 - typically uses infinite horizon

TERMINOLOGY

RL

State s

Action a

Environment

Reward

Return

Maximize

Value function

Optimal Value function

OC

State x

Control u

Plant

Cost

Cost to go

Minimize

Cost to go (of a policy)

Value function

MARKOV DECISION PROCESSES (MDP)

- Describe environment for RL problem
- Fully observable
- Markov property: Future is independent of the past given the present

$$P(x_{t+1} | x_t) = P(x_{t+1} | x_1, \dots, x_t)$$

- State transition probability: $P_{xx'} = P(x_{t+1} = x' | x_t = x)$

- State transition matrix:

- Each row sums to 1

- In deterministic systems

P contains only 0 and 1

$$P = \begin{bmatrix} P_{11} & \dots & P_{1n} \\ \vdots & & \vdots \\ P_{n1} & \dots & P_{nn} \end{bmatrix}$$

PROBABILY MATRIX vs DYNAMICS FUNCTION

In OCP for deterministic systems dynamics was encoded as:

$$x^+ = f(x)$$

In MP dynamics is encoded as probability density func.:

$$IP(x^+ | x)$$

If system is deterministic, both approaches can be used.

We will favor the first one.

INFINITE HORIZON AND DISCOUNTING

- Total discounted cost from time t to ∞ :

$$J_t = l_t + \gamma l_{t+1} + \dots = \sum_{k=0}^{\infty} \gamma^k l_{t+k}$$

- γ represents preference for later costs over immediate costs
 - γ close to 0 $\Rightarrow$ myopic evaluation
 - γ close to 1 $\Rightarrow$ far-sighted evaluation
- Why discount?
 - Avoid infinite cost
 - Uncertainty about the future
 - Immediate cost could require interest

γ = discount
factor

$$0 \leq \gamma \leq 1$$

VALUE FUNCTION (COST-TO-GO)

- Cost starting from state x :

$$V(x) = J_t (x_t = x) = \sum_{n=0}^{\infty} \gamma^n l_{t+n} = l_t + \gamma \underbrace{\sum_{n=0}^{\infty} \gamma^n l_{t+n+1}}_{J_{t+1}}$$

- Can be decomposed in two parts:

$$V(x) = l(x) + \gamma V(f(x))$$

- This is the Bellman Equation

$$\begin{array}{c} J_{t+1} \\ \parallel \\ V(x_{t+1}) \end{array}$$

OPTIMAL VALUE FUNCTION

- Minimum Value function over all policies:

$$V^*(x) = \min_{\pi} V^{\pi}(x)$$

- Optimal policy:

$$\pi^* \leq \pi \quad \forall \pi$$

$$\left[\text{where } \pi \leq \pi' \quad \text{if } V^{\pi}(x) \leq V^{\pi'}(x) \quad \forall x \right]$$

BELLMAN OPTIMALITY EQUATIONS

$$V^*(x) = \min_u l(x,u) + \gamma V^*(f(x,u))$$

- No closed-form solution
- Iterative algorithms