Lecture 12. RL and Control.
· Introduction:
(1) ND explicit supervision.
(2) only reward function.
(3) formalism: MDP
Markov Decision Process:
(1) Definition:
· S: state set
• A: action Set
o P(s,a): transition probability.
in state s. take action a.
* Y E CO. U: discount factor.
· R: SXA → R. reward function.
(2) dynamics:
$S_0 \xrightarrow{Q_0} S_1 \xrightarrow{Q_1} S_2 \cdots$
R= R(So, ao) + YR(S1, a1)+ 82 R(S2, a2) +
or only write the states:
$R = \sum_{k=0}^{k=0} J^k R(S^k)$
*: maximize target:
$E\left(\sum_{k=0}^{\infty} \gamma^{k} R(S_{k})\right].$

(3) Policy: $\pi: S \rightarrow A$. $Q = \pi(S)$ $V^{\pi}(s) = E[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \mid s_0 = s, \pi].$ X: Bellman Equation: $V^{\pi}(S) = R(S\delta) + \gamma \left[\sum P(S'|S,\pi(G)) V^{\pi}(S') \right].$ two terms: O immediate reward. e expected sum. (after first step) X: (S) is finite. (4) optimal value function: $V^*(S) = \max_{\pi} V^{\pi}(S)$ = $R(s) + \max_{\alpha \in A} \gamma \sum P(s'|s_{i}\alpha) V^*(s')$ X: maximum over a. $V_{\star}(z) = \Lambda_{\perp}(z) \leq \Lambda_{\perp}(z)$ $\pi^*(s) = \underset{\alpha \in A}{\text{arg max}} \sum_{s \in s} P(s'|s,a) V^*(s')$ · Value Iteration and Policy Iteration: (1) assumption: |S|< ∞. |A|< ∞. (2) value iteration: $V(s) := R(s) + \lambda \max \sum P(s'|s,a) V(s')$ Osynchronous update. @ asynchronous update.

USE V* to find optimal policy. (3) policy iteration: V:=V^ T(s)= arg max ≥ P(s'la.s) V(s') X: updating policy with current value function (4) for Small MDP: policy iteration is faster. for larger MDP: Is ↑ A ↑ Value iteration is preferred. • Learning a model for MDP: cu so far discussion: known: transition probability / reward function X: estimate from data. (2) trials → MLE. estimate P(s.a). average reward → R(s)	converge to V*	
T(s) = arg max \(\geq P(s' a.s) \) V(s') \(\for \) whole the current value function (4) for small MDP: Policy iteration is faster. for larger MPP: s \(\frac{1}{4} \) Value iteration is preferred. Learning a model for MPP: Co so far discussion: known: transition probability / reward function \(\frac{1}{4} \) estimate from data. (2) trials \(\to \) MLE. estimate \(\frac{1}{2} \) (s.a).	use V* to find optimal policy.	
T(s) = arg max \(\sigma \) P(s' a.s) \(V(s') \) \(\sigma \) \(\text{Lupdating policy with current value function} \) (4) \(\text{for Small MDP:} \) \(\text{policy iteration is faster.} \) \(\text{for larger MDP:} \ \sigma \) \(\left[A \) \) \(\text{value iteration is preferred.} \) \(\text{Learning a model for MDP:} \) \(\text{co so far discussion: known:} \) \(\text{transition probability / reward function} \) \(\text{\text{x: updating policy with current value function} \) \(\text{\text{transition probability / reward function} \) \(\text{\text{x: updating policy with current value function} \) \(\text{\text{transition probability / reward function} \) \(\text{\text{trials} -) \(MLE. \) \(\text{estimate P(s.a).} \)	(3) policy iteration:	
** updating Policy with current value function (9) for Small MDP: Policy iteration is faster. for larger MDP: ISI 1 A 1 value iteration is preferred. **Learning a model for MDP: cu so far discussion: known; transition probability / reward function **Estimate from data. (2) trials -> MLE. estimate P(s.a).	$V := V^{\pi}$	
** updating Policy with current value function (4) for Small MDP: Policy iteration is faster. for larger MDP: ISI 1 A 1 Value iteration is preferred. **Learning a model for MDP: CD SO far discussion: known; transition probability / reward function **X: estimate from data. (2) trials -> MLE. estimate P(S.a).	$\pi(s) = arg \max_{a \in A} \sum P(s' a,s) V(s')$	
policy iteration is faster. for larger MDP: ISI 1 A 1 value iteration is preferred. • Learning a model for MDP: co so far discussion: known: transition probability / reward function X: estimate from data. (2) trials -> MLE. estimate P(s.a).	• •	t value function
for larger MDP: ISIT AIT value iteration is preferred. • Learning a model for MDP: co so far discussion: known: transition probability / reward function X: estimate from data. (2) trials -> MLE. estimate P(s.a).	(4) for small MDP:	
value iteration is preferred. • Learning a model for MPP: co so far discussion: known: transition probability / reward function X: estimate from data. (2) trials -> MLE. estimate P(s.a).	policy iteration is faster.	
Learning a model for MPP: co so far discussion: known: transition probability / reward function X: estimate from data. (2) trials -> MLE. estimate P(s.a).	for larger MDP: ISI 1 AI1	
co so far discussion: known: transition probability / reward function X: estimate from data. (2) trials -> MLE. estimate P(s.a).	value iteration is preferred.	
(3) Sampling → optimizing exploration ← exploitation.	co so far discussion: known: transition probability / reward X: estimate from data. (2) trials -> MLE. estimate P(s.a).	function