

Model: predict what env will do next

Modeled: dynamic programming
find smallest $\pi \rightarrow \pi^*$

$$\pi \xleftarrow[\pi \rightarrow \text{greedy}(V)]{V \rightarrow V_\pi} V$$

Modeless: Monte Carlo
Temporal-difference
Q-learning

} on-policy

off -

Value func: $V_\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} + \dots | s_t = s]$
prediction of future reward

Reward(R): $R_s = E[R_{t+1} | s_t = s]$

$$\text{Return}(G) = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

State trans prob(P) = $P(s'|s, a)$
state trans matrix: $[\# \text{ of states} \times \# \text{ of states}]$

Markov reward process $\langle S, P, R, \gamma \rangle$

Bellman eq = immediate reward + successor reward

$$V(s) = E[R_{t+1} + \gamma V(s_{t+1}) | s_t] \\ = R + \gamma P V_{t+1} = (I - \gamma P)^{-1} R$$

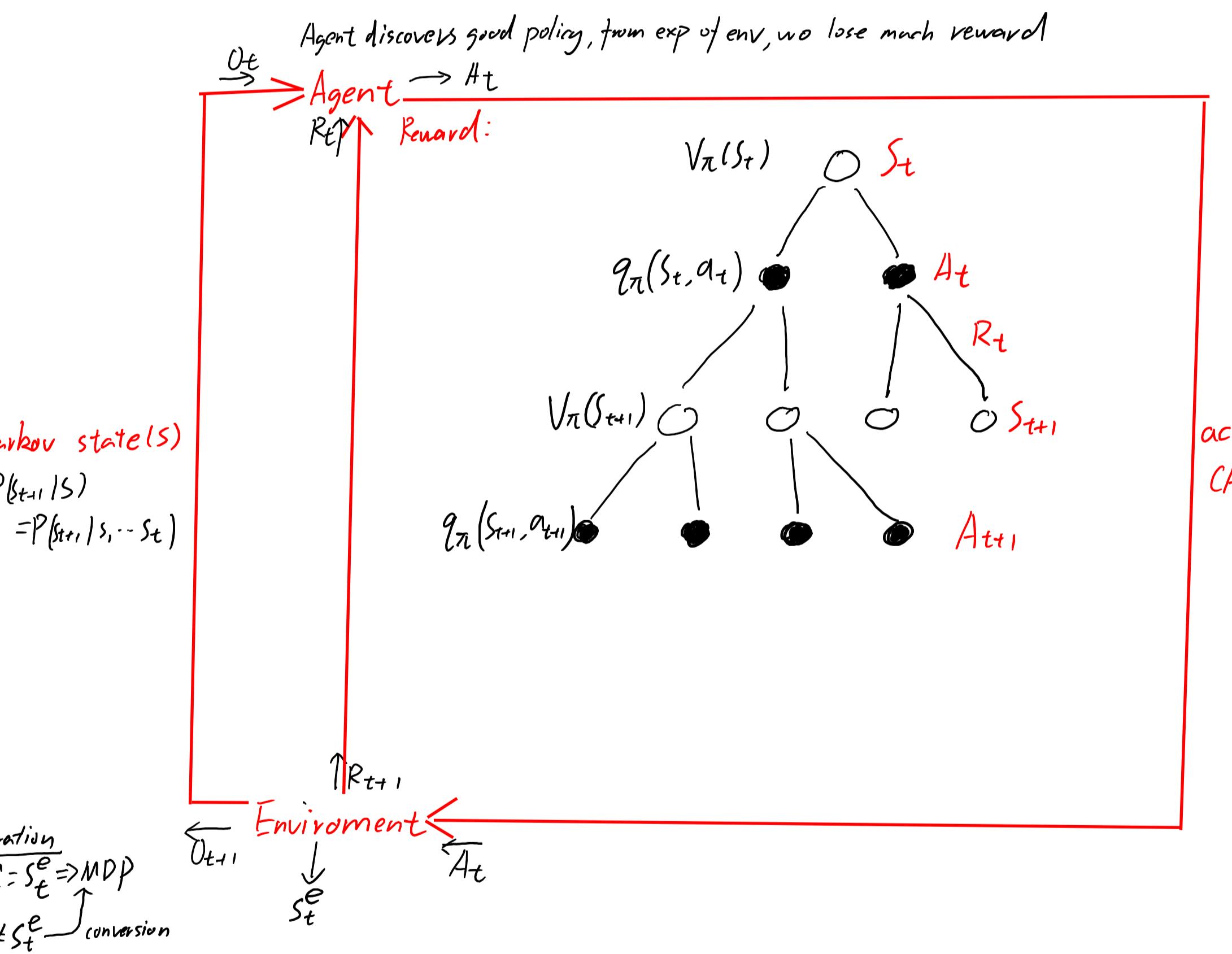
$O(\# \text{ of states}^3)$

DP

\Rightarrow finite states \Rightarrow MC
TD

infinite states \Rightarrow policy gradient

Policy (π): map from state to action; deterministic: $a = \pi(s)$
stochastic: $\pi(a|s) = P(a_t | s_t)$
strict sense stationary (time independent)



Markov decision process $\langle S, P, R, \gamma, A \rangle$

Only depends on current state

$$P_{ss'}^\pi = \sum_{a \in A} \pi(a|s) P_{ss'}^a, \quad R_s^\pi = \sum_{a \in A} \pi(a|s) R_s^a$$

State value func:

$$\begin{aligned} \text{expected return following policy } \pi &= E_\pi[G_t | s_t] = E_\pi[R_{t+1} + \gamma V_\pi(s_{t+1}) | s_t] \\ &= \sum_{a \in A} \pi(a|s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s') \right) \end{aligned}$$

action value func:

$$\begin{aligned} \text{expected return following action then policy} &= q_\pi(s, a) = E_\pi[G_t | s, a_t] = E_\pi[R_{t+1} + \gamma q_\pi(s_{t+1}, a_{t+1}) | s_t, a_t] \\ &= R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a' \in A} \pi(a'|s') q_\pi(s', a') \end{aligned}$$