Reinforcement Learning
Solving MDPs

Policy: \( a_t \sim \pi(o_t) \)

Most General Case

More Specific Case

Fully Observed System

Known Transition Function

Known Reward Function

\[
o_t = s_t \\
s_{t+1} \sim T(s_t, a_t) \\
R(s_{t+1}, s_t, a_t)
\]
Recap

Computing $V_*(s)$ and $Q_*(s, a)$ for known MDPs.

Backup diagrams, Bellman equations

$$V_\pi(s) = \sum_a \pi(a | s) \sum_{s', r} p(s', r | s, a) \left( r + \gamma V_\pi(s') \right)$$

Policy Evaluation, Improvement

Value Iteration

$$V_{k+1}(s) = \max_a \sum_{s', r} p(s', r | s, a) \left( r + \gamma V_k(s') \right)$$
Solving MDPs

Policy: $a_t \sim \pi(o_t)$

Most General Case

Fully Observed System:

$o_t = s_t$

More Specific Case

Fully Observed System:

$o_t = s_t$

Known Transition Function:

$s_{t+1} \sim T(s_t, a_t)$

Known Reward Function:

$R(s_{t+1}, s_t, a_t)$
6.2. Advantages of TD Prediction Methods

Example 6.2 Random Walk

In this example we empirically compare the prediction abilities of TD(0) and constant-MC when applied to the following Markov reward process:

We will often use MRPs when focusing on the prediction problem, in which there is no need to distinguish the dynamics due to the environment from those due to the agent. In this MRP, all episodes start in the center state, C, and proceed left or right by one state on each step, with equal probability. Episodes terminate either on the extreme left or the extreme right. When an episode terminates on the right, a reward of +1 occurs; all other rewards are zero. For example, a typical episode might consist of the following state-and-reward sequence:

Because this task is undiscounted, the true value of each state is the probability of terminating on the right if starting from that state. Thus, the true value of the center state is $v(C) = 0.5$. The true values of all the states, A through E, are $1/6$, $2/6$, $3/6$, $4/6$, and $5/6$.
MC Backup

Lecture 4: Model-Free Prediction

Temporal-Difference Learning

Unified View

Monte-Carlo Backup

$$V(S_t) = V(S_t) + \gamma G_t$$
TD(0) Backup

\[ V(S_t) = V(S_t) + \gamma V(S_{t+1}) + \tau_t r_{t+1} \]

\( S_t \)
\( S_{t+1} \)
\( r_{t+1} \)
Dynamic Programming Backup

\[ V(S_t) = \mathbb{E} \left[ R_{t+1} + V(S_{t+1}) \right] \]
Lecture 4: Model-Free Prediction

Temporal-Difference Learning

Unified View

Unified View of Reinforcement Learning

Dynamic programming

Exhaustive search

Monte Carlo

Bootstrap, $\lambda$

Shallow backups

Deep backups

Full backups

Sample backups
Recap

• Model Free Policy Evaluation
  • Monte Carlo, TD(0), TD(\(\lambda\))

• Model Free Control
  • On-policy: \(\epsilon\)-greedy, SARSA, SARSA(\(\lambda\))
  • Off-policy: Q-Learning
Model Free RL

Model Free Policy Evaluation

Model Free Control

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih  Koray Kavukcuoglu  David Silver  Alex Graves  Ioannis Antonoglou
Daan Wierstra  Martin Riedmiller