



AIL 7022: Reinforcement Learning

Lecture 7: Model-Free Policy Evaluation

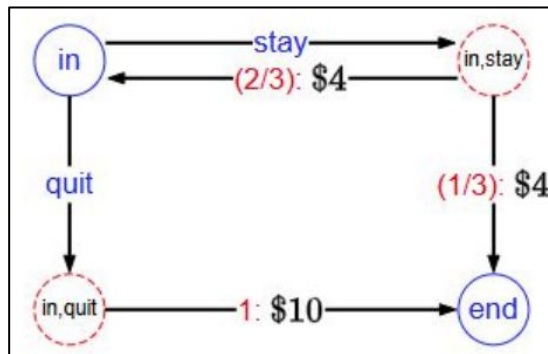
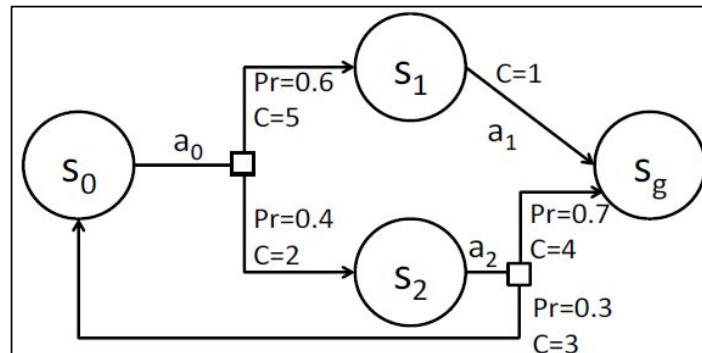
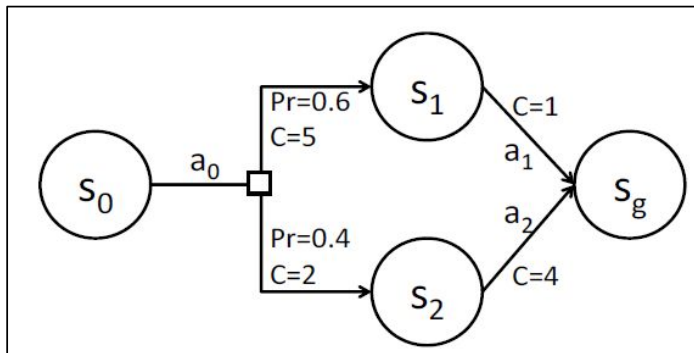
Instructor: Raunak Bhattacharyya



ScAI

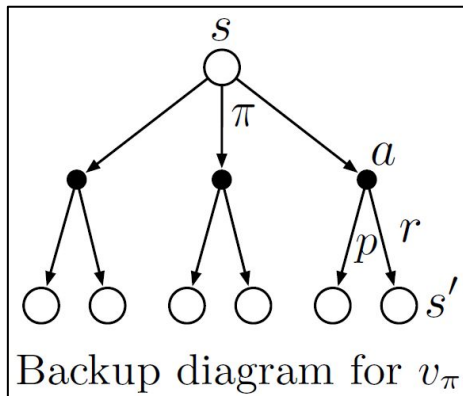
YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Policy Evaluation



Policy Evaluation: Iterative Solution

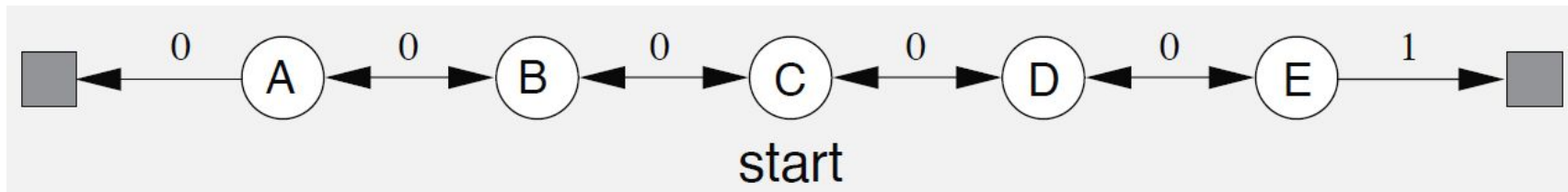
$$\begin{aligned} v_{k+1}(s) &\doteq \mathbb{E}_{\pi}[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) [r + \gamma v_k(s')] \end{aligned}$$



Bootstrapping

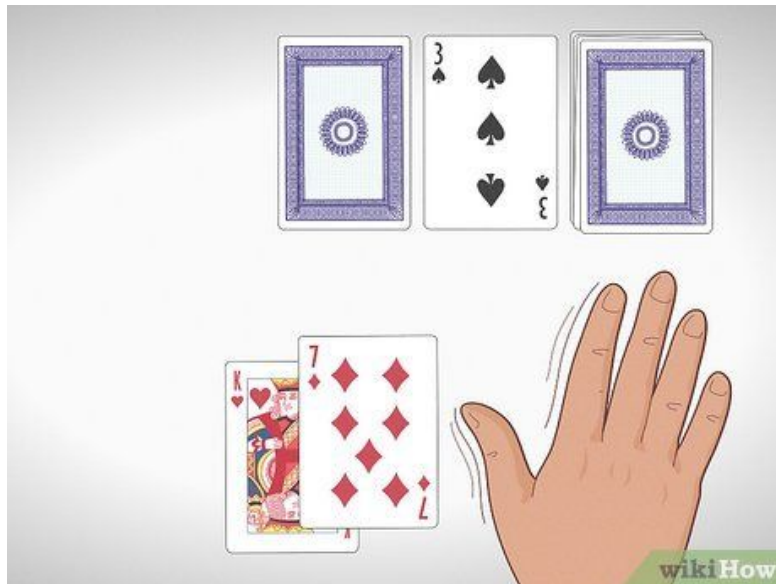
Prediction (vs. control)

Example: Random Walk



What is the value function?

Blackjack



[Source](#), [Simulator](#)

Formulation

Dynamic programming hard

Model-Free Approaches

- MDP model is unknown, but we can sample from it

**Autonomous vehicle in
traffic**

**Robotic navigation in
unknown envs**

**Advertising with
unknown user behavior**

- MDP model is known, but it's easier to sample

Climate models

Robotic navigation

Game playing

Plan

- Monte Carlo
- Importance Sampling
- Off-policy
- Temporal Difference

Monte-Carlo Policy Evaluation

Monte Carlo Estimation

Setting

Episodes and visits

$$\mathbb{E}[f(x)] \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^N f(x_i)$$

where $x_1, x_2, \dots, x_N$ are i.i.d. random samples drawn from the distribution over x .

- Sample trajectories
- Store the obtained returns
- Average

Monte Carlo Policy Evaluation

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

$Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):

Generate an episode following π : $S_0, A_0, R_1, S_1, A_1, R_2, \dots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \dots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

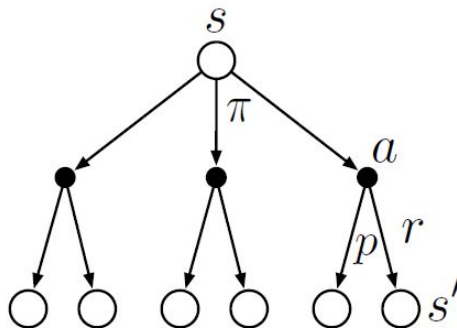
Unless S_t appears in $S_0, S_1, \dots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow \text{average}(Returns(S_t))$

Convergence

Backup diagram



- Estimate for one state does not build upon the estimate of any other state (this was the case in DP)
- Monte Carlo methods do not bootstrap

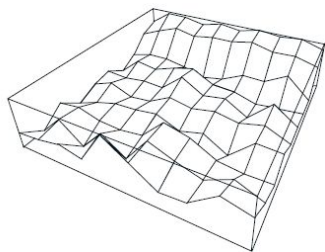
Computational expense of estimating the value of a state is independent of the number of states

Blackjack

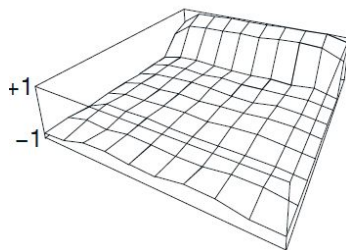
Policy: Stand only if 20 or 21

After 10,000 episodes

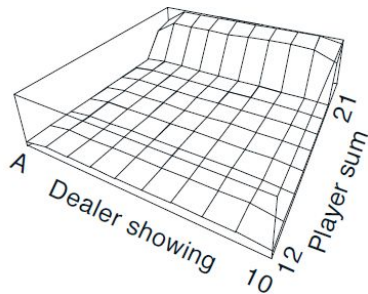
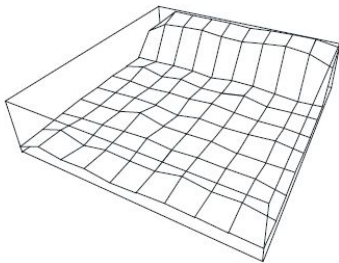
Usable
ace



After 500,000 episodes



No
usable
ace



- At convergence value jump for last 2
- Value drop off
- Frontmost values higher for usable ace
- Usable ace less certain and regular

Off-Policy Prediction via Importance Sampling

$$\mathbb{E}_p[z(x)] = \int z(x)p(x)dx$$

$$r = \mathbb{E}_p[z(x)]$$

$$\hat{r} = \frac{1}{n} \sum_{i=1}^n z(x_i)$$

$$\mathbb{E}_p[z(x)] = \int z(x)p(x)dx,$$

$$= \int z(x) \frac{p(x)}{q(x)} q(x) dx$$

$$= \mathbb{E}_q \left[z(x) \frac{p(x)}{q(x)} \right]$$

$$\hat{r} = \frac{1}{n} \sum_{i=1}^n z(x_i) \frac{p(x_i)}{q(x_i)}$$