



AIL 7022: Reinforcement Learning

Lecture 8: Model-Free PE & Finding Policies

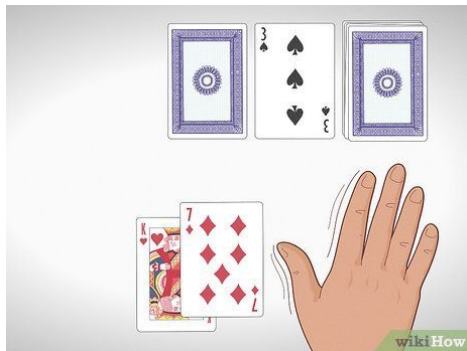
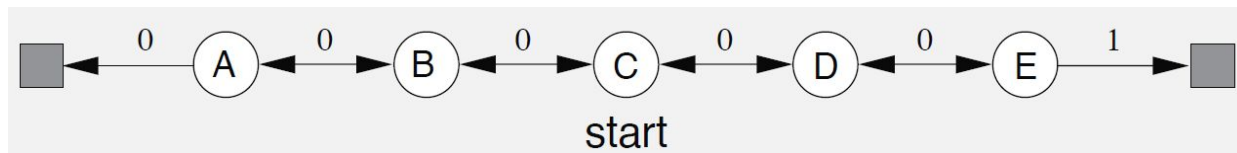
Instructor: Raunak Bhattacharyya



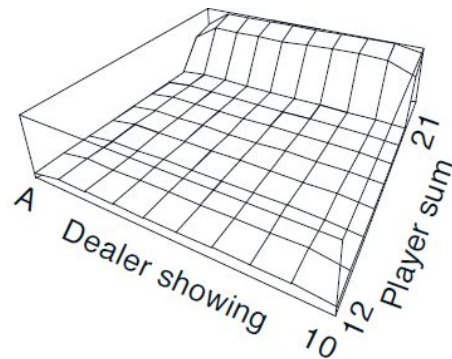
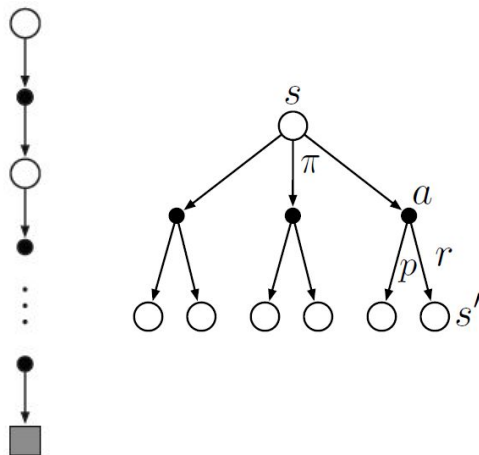
ScAI

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Towards Model-Free Policy Evaluation



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



Monte Carlo Policy Evaluation

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

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 $Returns(S_t)$

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

Off-Policy Prediction via Importance Sampling

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

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

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

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

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

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

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

Plan

- Off-policy prediction using IS
- Incremental MC algorithm
- Finding policies
- Assignment Overview

Off-Policy Prediction via Importance Sampling

$$\Pr\{A_t, S_{t+1}, A_{t+1}, \dots, S_T \mid S_t, A_{t:T-1} \sim \pi\}$$

Coverage assumption

$$= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1})$$

$$= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)$$

$$\rho_{t:T-1} \doteq \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)}$$

$$= \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

IS Monte Carlo

Incremental Approach

expectation of a single variable X from m samples:

$$\begin{aligned}\hat{x}_m &= \frac{1}{m} \sum_{i=1}^m x^{(i)} \\ \hat{x}_m &= \frac{1}{m} \left(x^{(m)} + \sum_{i=1}^{m-1} x^{(i)} \right) \\ &= \frac{1}{m} \left(x^{(m)} + (m-1)\hat{x}_{m-1} \right) \\ &= \hat{x}_{m-1} + \frac{1}{m} \left(x^{(m)} - \hat{x}_{m-1} \right)\end{aligned}$$

Incremental Approach

$$\hat{x}_m = \hat{x}_{m-1} + \alpha(m) \left(x^{(m)} - \hat{x}_{m-1} \right)$$

$$\sum_{m=1}^{\infty} \alpha(m) = \infty \quad \sum_{m=1}^{\infty} \alpha^2(m) < \infty$$

$$\hat{x} \leftarrow \hat{x} + \alpha(x - \hat{x})$$

$$NewEstimate \leftarrow OldEstimate + StepSize \left[Target - OldEstimate \right]$$

Incremental Model-Free Policy Evaluation

$$\hat{V}_m^\pi(s^j) \leftarrow \hat{V}_{m-1}^\pi(s^j) + \alpha [G^{(m)} - \hat{V}_{m-1}^\pi(s^j)]$$

$$\hat{V}_m^\pi(s^j) \leftarrow \hat{V}_{m-1}^\pi(s^j) + \alpha [G^{(m)} - \hat{V}_{m-1}^\pi(s^j)]$$

New estimate

Old estimate

Target

Monte Carlo vs. Dynamic Programming

$$V^{\pi}(s^j) = \mathbb{E}_{p_{\theta}(\tau)} \left[G_t \middle| s_t = s^j \right]$$

Target for Monte Carlo

$$V^{\pi}(s^j) = \mathbb{E}_{p_{\theta}(\tau)} \left[r_{t+1} + \gamma \cdot V^{\pi}(s_{t+1}) \middle| s_t = s^j \right]$$

Target for DP

Why do both approaches give estimates and not the actual value function until they converge?

Monte Carlo vs. Dynamic Programming

$$V^{\pi}(s^j) = \mathbb{E}_{p_{\theta}(\tau)} \left[G_t \middle| s_t = s^j \right]$$

Target for Monte Carlo

Uses a sample of the real thing

But not the expectation, just a single sample

$$V^{\pi}(s^j) = \mathbb{E}_{p_{\theta}(\tau)} \left[r_{t+1} + \gamma \cdot V^{\pi}(s_{t+1}) \middle| s_t = s^j \right]$$

Target for DP

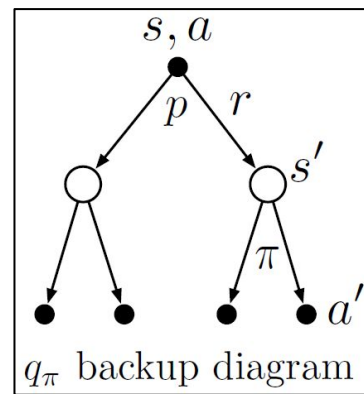
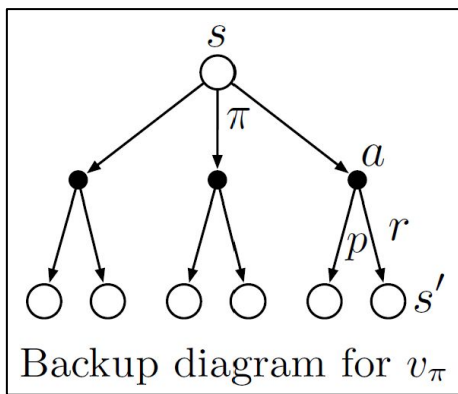
Does compute expectations. Uses known model

Target not created using the real value function but rather the current estimate: Bootstrapping

Finding Policies

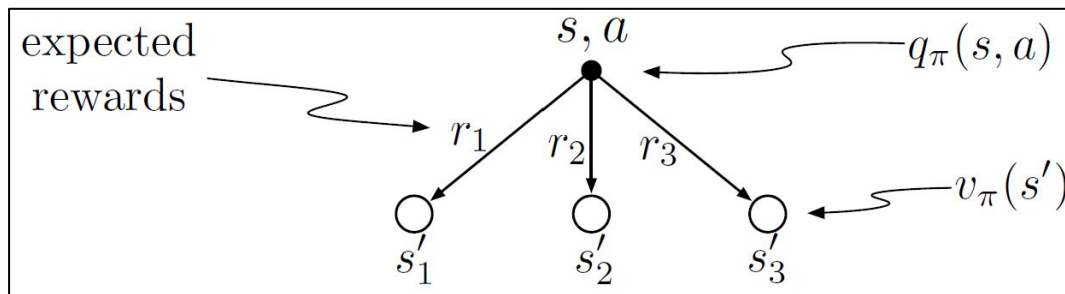
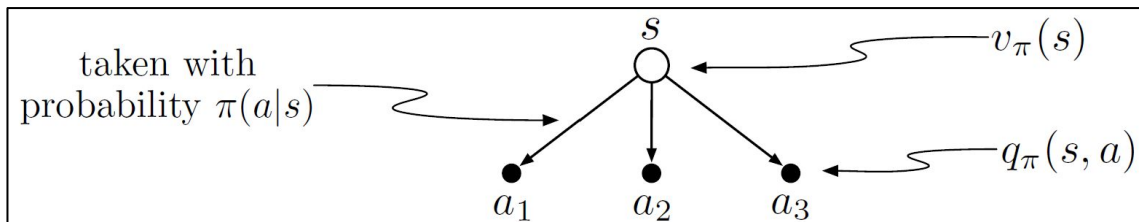
q: Action-Value function

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



v & q

$$\begin{aligned}
 q_{\pi}(s, a) &\doteq \mathbb{E}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s, A_t = a] \\
 &= \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_{\pi}(s') \right].
 \end{aligned}$$



Policy Improvement

Let π and π' be any pair of deterministic policies such that, for all $s \in \mathcal{S}$,

$$q_{\pi}(s, \pi'(s)) \geq v_{\pi}(s)$$

Then the policy π' must be as good as, or better than, π .

$$v_{\pi'}(s) \geq v_{\pi}(s)$$

Policy Improvement

$$\begin{aligned} v_{\pi}(s) &\leq q_{\pi}(s, \pi'(s)) \\ &= \mathbb{E}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s, A_t = \pi'(s)] \\ &= \mathbb{E}_{\pi'}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s] \\ &\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_{\pi}(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s] \\ &= \mathbb{E}_{\pi'}[R_{t+1} + \gamma \mathbb{E}[R_{t+2} + \gamma v_{\pi}(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})] \mid S_t = s] \\ &= \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_{\pi}(S_{t+2}) \mid S_t = s] \\ &\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_{\pi}(S_{t+3}) \mid S_t = s] \\ &\vdots \\ &\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \mid S_t = s] \\ &= v_{\pi'}(s). \end{aligned}$$

Optimal Policies and Value Functions

$$v_*(s) \doteq \max_{\pi} v_{\pi}(s), \text{ for all } s \in \mathcal{S}.$$

$$q_*(s, a) \doteq \max_{\pi} q_{\pi}(s, a), \text{ for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A}(s)$$

$$\begin{aligned} v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a) \\ &= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\ &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \max_a \sum_{s', r} p(s', r \mid s, a) [r + \gamma v_*(s')]. \end{aligned}$$

Bellman Optimality Equations

$$\begin{aligned} q_*(s, a) &= \mathbb{E} \left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a \right] \\ &= \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right]. \end{aligned}$$

