Lecture 3: Model-Free Policy Evaluation: Policy Evaluation Without Knowing How the World Works[1]

Emma Brunskill

CS234 Reinforcement Learning

Winter 2022

[1] Material builds on structure from David SIlver's Lecture 4: Model-Free Prediction. Other resources: Sutton and Barto Jan 1 2018 draft Chapter/Sections: 5.1; 5.5; 6.1-6.3

Refresh Your Knowledge L3 [Polleverywhere Poll]

- What is the max number of iterations of policy iteration in a tabular MDP?
  1. $|A||S|$
  2. $|S|^{|A|}$
  3. $|A|^{|S|}$
  4. Unbounded
  5. Not sure
- In a tabular MDP asymptotically value iteration will always yield a policy with the same value as the policy returned by policy iteration
  1. True.
  2. False
  3. Not sure
- Can value iteration require more iterations than $|A|^{|S|}$ to compute the optimal value function? (Assume $|A|$ and $|S|$ are small enough that each round of value iteration can be done exactly).
  1. True.
  2. False
  3. Not sure

Refresh Your Knowledge L3

- What is the max number of iterations of policy iteration in a tabular MDP?
  Answer: $|A|^{|S|}$: There are only $|A|^{|S|}$ policies in a tabular MDP and each policy can only be considered at most once, since policy improvement either results in a policy with a higher value or returns the same policy if the optimal policy has been found.

- In a tabular MDP asymptotically value iteration will always yield a policy with the same value as the policy returned by policy iteration
  Answer. True. Both are guaranteed to converge to the optimal value function and a policy with an optimal value

- Can value iteration require more iterations than $|A|^{|S|}$ to compute the optimal value function? (Assume $|A|$ and $|S|$ are small enough that each round of value iteration can be done exactly).
  Answer: True. As an example, consider a single state, single action MDP where $r(s,a) = 1$, $\gamma = .9$ and initialize $V_0(s) = 0$. $V^*(s) = \frac{1}{1-\gamma}$ but after the first iteration of value iteration, $V_1(s) = 1$.

Today's Plan

check your understanding polls
due Sunday at 6pm
Stanford
time

- Last Time:
  - Markov reward / decision processes
  - Policy evaluation & control when have true model (of how the world works)
- **Today**
  - **Policy evaluation without known dynamics & reward models**

    dynamics
    reward

    tabular
- Next Time:
  - Control when don't have a model of how the world works

This Lecture: Policy Evaluation

- Estimating the expected return of a particular policy if don't have access to true MDP models
- Monte Carlo policy evaluation
  - Policy evaluation when don't have a model of how the world work
    - Given on-policy samples
- Temporal Difference (TD)
- Certainty Equivalence with dynamic programming
- Metrics to evaluate and compare algorithms

Recall

- Definition of Return, G_t (for a MRP)
  - Discounted sum of rewards from time step t to horizon

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots \gamma^{H-1} r_{H}$$

- Definition of State Value Function, $V^\pi(s)$
  - Expected return from starting in state s under policy π

$$V^\pi(s) = \mathbb{E}_\pi[G_t | s_t = s] = \mathbb{E}_\pi[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots | s_t = s]$$

- Definition of State-Action Value Function, $Q^\pi(s, a)$
  - Expected return from starting in state s, taking action a and then following policy π

$$\begin{aligned} Q^\pi(s, a) &= \mathbb{E}_\pi[G_t | s_t = s, a_t = a] \\ &= \mathbb{E}_\pi[\underbrace{r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots | s_t = s, a_t = a}] \end{aligned}$$

This Lecture: Policy Evaluation

- Estimating the expected return of a particular policy if don't have access to true MDP models
- Monte Carlo policy evaluation
  - Policy evaluation when don't have a model of how the world work
    - Given on-policy samples
- Temporal Difference (TD)
- Certainty Equivalence with dynamic programming
- Metrics to evaluate and compare algorithms

Monte Carlo (MC) Policy Evaluation

- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ in MDP M under policy π
- $V^\pi(s) = \mathbb{E}_{T \sim \pi}[G_t | s_t = s]$
  - Expectation over trajectories T generated by following π
- Simple idea: Value = mean return
- If trajectories are all finite, sample set of trajectories & average returns

Monte Carlo (MC) Policy Evaluation

- If trajectories are all finite, sample set of trajectories & average returns
- Does not require MDP dynamics/rewards
- Does not assume state is Markov
- Can be applied to episodic MDPs
  - Averaging over returns from a complete episode
  - Requires each episode to terminate

Monte Carlo (MC) On Policy Evaluation

- Aim: estimate $V^\pi(s)$ given episodes generated under policy π
  - $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ where the actions are sampled from π
- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ in MDP M under policy π
- $V^\pi(s) = \mathbb{E}_\pi[G_t | s_t = s]$
- MC computes empirical mean return
- Often do this in an incremental fashion
  - After each episode, update estimate of V^π

First-Visit Monte Carlo (MC) On Policy Evaluation

counts

Initialize $N(s) = 0$, $G(s) = 0 \;\forall s \in S$

Loop

$\pi(s_{i,1})$

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$ as return from time step t onwards in ith episode
- For each time step t till the end of the episode i
  - If this is the **first** time t that state s is visited in episode i
    - Increment counter of total first visits: $N(s) = N(s) + 1$
    - Increment total return $G(s) = G(s) + G_{i,t}$
    - Update estimate $V^\pi(s) = G(s)/N(s)$

Evaluation of the Quality of a Policy Estimation Approach: Bias, Variance and MSE

- Consider a statistical model that is parameterized by θ and that determines a probability distribution over observed data $P(x|\theta)$
- Consider a statistic $\hat{\theta}$ that provides an estimate of θ and is a function of observed data x
  - E.g. for a Gaussian distribution with known variance, the average of a set of i.i.d data points is an estimate of the mean of the Gaussian
- Definition: the bias of an estimator $\hat{\theta}$ is:

$$Bias_\theta(\hat{\theta}) = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta$$

- Definition: the variance of an estimator $\hat{\theta}$ is:

$$Var(\hat{\theta}) = \mathbb{E}_{x|\theta}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

- Definition: mean squared error (MSE) of an estimator $\hat{\theta}$ is:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias_\theta(\hat{\theta})^2$$

Evaluation of the Quality of a Policy Estimation Approach: Consistent Estimator

- Consider a statistical model that is parameterized by θ and that determines a probability distribution over observed data $P(x|\theta)$
- Consider a statistic $\hat{\theta}$ that provides an estimate of θ and is a function of observed data x
- Definition: the bias of an estimator $\hat{\theta}$ is:

$$Bias_\theta(\hat{\theta}) = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta$$

- Let n be the number of data points x used to estimate the parameter θ and call the resulting estimate of θ using that data $\hat{\theta}_n$
- Then the estimator $\hat{\theta}_n$ is consistent if, for all $\epsilon > 0$

$$\lim_{n \to \infty} Pr(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

- Quick check: if an estimator is unbiased (bias $= 0$) is it consistent?

First-Visit Monte Carlo (MC) On Policy Evaluation

Initialize $N(s) = 0$, $G(s) = 0 \ \forall s \in S$
Loop

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$ as return from time step t onwards in ith episode
- For each time step t till the end of the episode i
  - If this is the **first** time t that state s is visited in episode i
    - Increment counter of total first visits: $N(s) = N(s) + 1$
    - Increment total return $G(s) = G(s) + G_{i,t}$
    - Update estimate $V^\pi(s) = G(s)/N(s)$

Properties:

- V^π estimator is an unbiased estimator of true $\mathbb{E}_\pi[G_t | s_t = s]$
- By law of large numbers, as $N(s) \to \infty$, $V^\pi(s) \to \mathbb{E}_\pi[G_t | s_t = s]$

consistent

Every-Visit Monte Carlo (MC) On Policy Evaluation

Initialize $N(s) = 0$, $G(s) = 0 \ \forall s \in S$
Loop

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$ as return from time step t onwards in ith episode
- For each time step t till the end of the episode i
  - state s is the state visited at time step t in episodes i
  - Increment counter of total visits: $N(s) = N(s) + 1$
  - Increment total return $G(s) = G(s) + G_{i,t}$
  - Update estimate $V^\pi(s) = G(s)/N(s)$

Every-Visit Monte Carlo (MC) On Policy Evaluation

Initialize $N(s) = 0$, $G(s) = 0 \ \forall s \in S$
Loop

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$ as return from time step t onwards in ith episode
- For each time step t till the end of the episode i
  - state s is the state visited at time step t in episodes i
  - Increment counter of total visits: $N(s) = N(s) + 1$
  - Increment total return $G(s) = G(s) + G_{i,t}$
  - Update estimate $V^\pi(s) = G(s)/N(s)$

Properties:

- V^π every-visit MC estimator is a **biased** estimator of V^π
- But consistent estimator and often has better MSE

Worked Example First Visit MC On Policy Evaluation

Initialize $N(s) = 0,\ G(s) = 0\ \forall s \in S$
Loop

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$
- For each time step t till the end of the episode i
  - If this is the **first** time t that state s is visited in episode i
    - Increment counter of total first visits: $N(s) = N(s) + 1$
    - Increment total return $G(s) = G(s) + G_{i,t}$
    - Update estimate $V^\pi(s) = G(s)/N(s)$
- Mars rover: $\mathrm{R(s)} = [\ \overset{s_1}{1}\ 0\ 0\ 0\ 0\ 0\ \overset{s_7}{+10}]$
- $\pi(s) = a_1\ \forall s,\ \gamma = 1.$ any action from s_1 and s_7 terminates episode
- Trajectory $= (s_3, a_1, 0, s_2, a_1, 0, s_2, a_1, 0, s_1, a_1, 1,$ terminal$)$

Worked Example MC On Policy Evaluation

Initialize $N(s) = 0,\ G(s) = 0\ \forall s \in S$

Loop

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- $G_{i,t} = \underline{r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i-1} r_{i,T_i}}$
- For each time step t till the end of the episode i
  - If this is the **first** time t that state s is visited in episode i
    - Increment counter of total first visits: $N(s) = N(s) + 1$
    - Increment total return $G(s) = G(s) + G_{i,t}$
    - Update estimate $V^\pi(s) = G(s)/N(s)$

$t = 1$ s_3 $G_{1,1} = 1$

$t = 2$ s_2 $G_{1,2} = 1$

$t = 3$ s_2

$t = 4$ s_1

- Mars rover: R(s) = [1 0 0 0 0 0 +10]
- Trajectory = (s_3, a_1, 0, s_2, a_1, 0, s_2, a_1, 0, s_1, a_1, 1, terminal)
- Let $\gamma = 1$. First visit MC estimate of V of each state?
- $V(s_3) = 1 = V(s_2) = V(s_1)$ $\quad V = [1\ 1\ 1\ 0\ 0\ 0\ 0]$
- Now let $\gamma = 0.9$. Compare the first visit & every visit MC estimates of s_2.
- $V(s_3)$ $\quad G_{1,1} = 0 + \gamma \cdot 0 + \gamma^2 \cdot 0 + \gamma^3$

Worked Example MC On Policy Evaluation

Initialize $N(s) = 0$, $G(s) = 0 \ \forall s \in S$

Loop

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$
- For each time step t till the end of the episode i
  - If this is the **first** time t that state s is visited in episode i
    - Increment counter of total first visits: $N(s) = N(s) + 1$
    - Increment total return $G(s) = G(s) + G_{i,t}$
    - Update estimate $V^\pi(s) = G(s)/N(s)$

- Mars rover: R = [1 0 0 0 0 0 +10] for any action
- γ^2 γ
  Trajectory = (s_3, a_1, 0, s_2, a_1, 0, s_2, a_1, 0, s_1, a_1, 1, terminal)
- Let $\gamma = 1$. First visit MC estimate of V of each state?
  V = [1 1 1 0 0 0 0]

- Now let $\gamma = 0.9$. Compare the first visit & every visit MC estimates of s_2.
  First visit: $V^{MC}(s_2) = \gamma^2$, Every visit: $V^{MC}(s_2) = \frac{\gamma^2 + \gamma}{2}$

Incremental Monte Carlo (MC) On Policy Evaluation

looping

After each episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots$

- Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots$ as return from time step t onwards in ith episode
- For state s visited at time step t in episode i
  - Increment counter of total visits: $N(s) = N(s) + 1$
  - Update estimate

$$V^\pi(s) = V^\pi(s)\frac{N(s)-1}{N(s)} + \frac{G_{i,t}}{N(s)} = V^\pi(s) + \frac{1}{N(s)}(G_{i,t} - V^\pi(s))$$

Incremental Monte Carlo (MC) On Policy Evaluation

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$
- for $i = 1 : T_i$ where T_i is the length of the i-th episode
    - $V^\pi(s_{it}) = V^\pi(s_{it}) + \alpha(G_{i,t} - V^\pi(s_{it}))$

weight

Check Your Understanding L3N1: Polleverywhere Poll Incremental MC (State if each is True or False)

2

First or Every Visit MC

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$
  - For all s, for **first or every** time t that state s is visited in episode i
    - $N(s) = N(s) + 1$, $G(s) = G(s) + G_{i,t}$ $G_{old}(s) + G_{i,t}$
    - Update estimate $V^\pi(s) = G(s)/N(s)$ $=$ $\frac{G_{old}(s)}{N} + \frac{G_{it}}{N}$

$(N-1) V^\pi_{old}(s)$

Incremental MC

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$ $= \frac{V_{old}(N-1)}{N} + \frac{G_{it}}{N}$
- $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$
- for $t = 1 : T_i$ where T_i is the length of the i-th episode
  - $V^\pi(s_{it}) = V^\pi(s_{it}) + \alpha(G_{i,t} - V^\pi(s_{it}))$ $(1-\alpha) V_{old} + \alpha G_{it}$

$\left(\frac{N-1}{N}\right)$ $\frac{1}{N}$

1. Incremental MC with $\alpha = 1$ is the same as first visit MC
2. Incremental MC with $\alpha = \frac{1}{N(s_{it})}$ is the same as first visit MC
3. Incremental MC with $\alpha = \frac{1}{N(s_{it})}$ is the same as every visit MC
4. Incremental MC with $\alpha > \frac{1}{N(s_{it})}$ could be helpful in non-stationary domains

Break

- When we come back, continue with Monte Carlo policy evaluation

Check Your Understanding L3N1: Polleverywhere Poll Incremental MC Answers

First or Every Visit MC

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i - 1} r_{i,T_i}$
    - For all s, for **first or every** time t that state s is visited in episode i
        - $N(s) = N(s) + 1$, $G(s) = G(s) + G_{i,t}$. Update estimate $V^\pi(s) = G(s)/N(s)$

Incremental MC

- Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T_i}$
- for $t = 1 : T_i$ where T_i is the length of the i-th episode
    - $V^\pi(s_{it}) = V^\pi(s_{it}) + \alpha(G_{i,t} - V^\pi(s_{it}))$

$V^\pi(s_{it}) = G_{i,t}$

1. Incremental MC with $\alpha = 1$ is the same as first visit MC

   false

2. Incremental MC with $\alpha = \frac{1}{N(s_{it})}$ is the same as first visit MC

   false

3. Incremental MC with $\alpha = \frac{1}{N(s_{it})}$ is the same as every visit MC

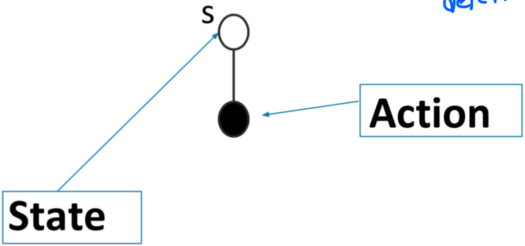
   true

4. Incremental MC with $\alpha > \underbrace{\frac{1}{N(s_{it})}}$ could help in non-stationary domains

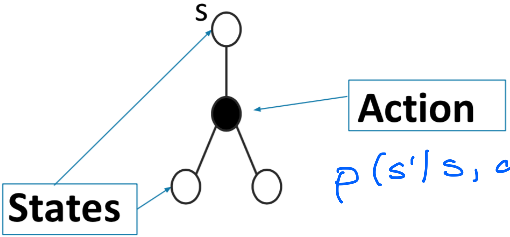
   true

Policy Evaluation Diagram

deterministic π



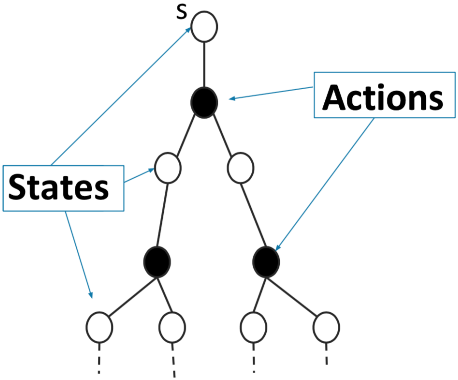
Policy Evaluation Diagram



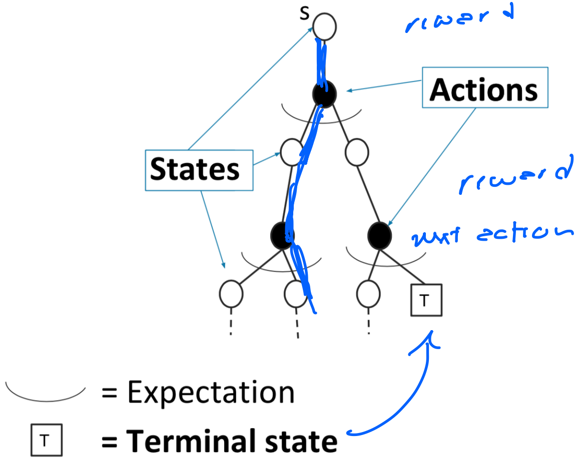
$\pi(s)$

$p(s'|s,a)$

Policy Evaluation Diagram



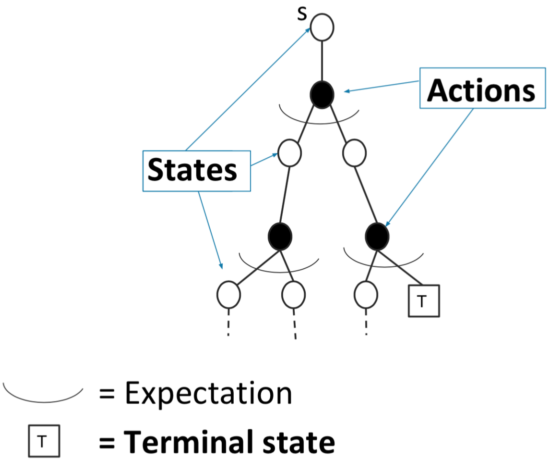
Policy Evaluation Diagram



MC Policy Evaluation

$(1-\alpha)V^\pi(s) + \alpha G_{i,t}$

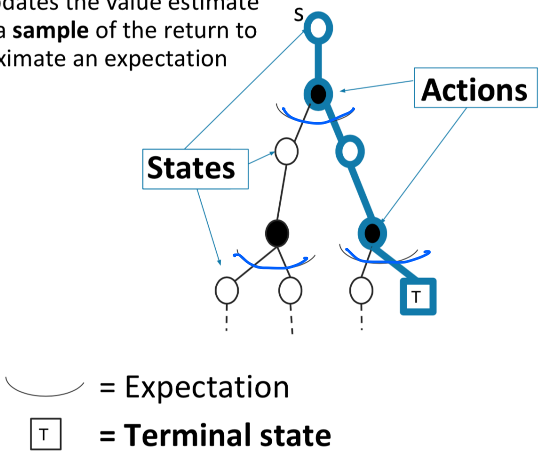
$$V^\pi(s) = V^\pi(s) + \alpha(\underbrace{G_{i,t}} - V^\pi(s))$$



MC Policy Evaluation

$$V^{\pi}(s) = V^{\pi}(s) + \alpha(G_{i,t} - V^{\pi}(s))$$

MC updates the value estimate using a **sample** of the return to approximate an expectation



Monte Carlo (MC) Policy Evaluation Key Limitations

- Generally high variance estimator
  - Reducing variance can require a lot of data
  - In cases where data is very hard or expensive to acquire, or the stakes are high, MC may be impractical
- Requires episodic settings
  - Episode must end before data from episode can be used to update V

Monte Carlo (MC) Policy Evaluation Summary

- Aim: estimate $V^\pi(s)$ given episodes generated under policy π
  - $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ where the actions are sampled from π
- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ under policy π
- $V^\pi(s) = \mathbb{E}_\pi[G_t | s_t = s]$
- Simple: Estimates expectation by empirical average (given episodes sampled from policy of interest)
- Updates V estimate using **sample** of return to approximate the expectation
- Does not assume Markov process
- Converges to true value under some (generally mild) assumptions

Break

- (End of Monte Carlo policy evaluation)

This Lecture: Policy Evaluation

- Estimating the expected return of a particular policy if don't have access to true MDP models
- Monte Carlo policy evaluation
  - Policy evaluation when don't have a model of how the world work
    - Given on-policy samples
- **Temporal Difference (TD)**
- Certainty Equivalence with dynamic programming
- Metrics to evaluate and compare algorithms

Temporal Difference Learning

- "If one had to identify one idea as central and novel to reinforcement learning, it would undoubtedly be temporal-difference (TD) learning." – Sutton and Barto 2017
- Combination of Monte Carlo & dynamic programming methods
- Model-free
- Can be used in episodic or infinite-horizon non-episodic settings
- Immediately updates estimate of V after each (s, a, r, s') tuple

Temporal Difference Learning for Estimating V

- Aim: estimate $V^\pi(s)$ given episodes generated under policy π
- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ in MDP M under policy π
- $V^\pi(s) = \mathbb{E}_\pi[G_t | s_t = s]$
- Recall Bellman operator (if know MDP models)

deterministic π

$$B^\pi V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V(s')$$

- In incremental every-visit MC, update estimate using 1 sample of return (for the current ith episode)

$$V^\pi(s) = V^\pi(s) + \alpha(G_{i,t} - V^\pi(s))$$

- Insight: have an estimate of V^π, use to estimate expected return

$$V^\pi(s) = V^\pi(s) + \alpha([r_t + \gamma V^\pi(s_{t+1})] - V^\pi(s))$$

Temporal Difference [$TD(0)$] Learning

- Aim: estimate $V^\pi(s)$ given episodes generated under policy π
  - $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ where the actions are sampled from π
- Simplest TD learning: update value towards estimated value

$$V^\pi(s_t) = V^\pi(s_t) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s_t))$$

- TD error:

$$\delta_t = r_t + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)$$

- Can immediately update value estimate after (s, a, r, s') tuple
- Don't need episodic setting

Temporal Difference [$TD(0)$] Learning Algorithm

Input: α
Initialize $V^\pi(s) = 0$, $\forall s \in S$
Loop

- Sample **tuple** (s_t, a_t, r_t, s_{t+1})
- $V^\pi(s_t) = V^\pi(s_t) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s_t))$

Compute new V^π at the end of 1 trajectory

Input: α
Initialize $V^\pi(s) = 0, \forall s \in S$
Loop

- Sample **tuple** (s_t, a_t, r_t, s_{t+1})
- $V^\pi(s_t) = V^\pi(s_t) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s_t))$

Example Mars rover: $R^{(s)} = [\ 1\ 0\ 0\ 0\ 0\ 0\ +10]$ ~~for any action~~

- $\pi(s) = a_1\ \forall s$, $\gamma = 1$. any action from s_1 and s_7 terminates episode
- Trajectory $= (s_3, a_1, 0, s_2, a_1, 0, s_2, a_1, 0, s_1, a_1, 1, \text{terminal})$

1st $v(3)$? MC $V = [1\ 1\ 1\ 0\ 0\ 0\ 0]$

TD $(s_3\ a_1\ 0\ s_2)$ TD target: $0 + \gamma V^\pi(s_2) = 0$

$V(s_3) = 0 + \alpha(0 - 0) = 0$

$(s_2\ a_1\ 0\ s_2)$ TD target $0 + \gamma V^\pi(s_2) = 0$

... $(s_1\ a_1\ 1\ \text{terminal})$ TD target $1 + 0 = 1$

Worked Example TD Learning

Input: α
Initialize $V^\pi(s) = 0, \forall s \in S$
Loop

- Sample **tuple** (s_t, a_t, r_t, s_{t+1})
- $V^\pi(s_t) = V^\pi(s_t) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s_t))$

Example:

- Mars rover: R = [1 0 0 0 0 0 +10] for any action
- $\pi(s) = a_1 \ \forall s$, $\gamma = 1$. any action from s_1 and s_7 terminates episode
- Trajectory = $(s_3, a_1, 0, s_2, a_1, 0, s_2, a_1, 0, s_1, a_1, 1, \text{terminal})$
- TD estimate of all states (init at 0) with $\alpha = 1$?
  V = [1 0 0 0 0 0 0 0]
- First visit MC estimate of V of each state? [1 1 1 0 0 0 0]

Temporal Difference (TD) Policy Evaluation

Bellman operator

$$V^{\pi}(s_t) = r(s_t, \pi(s_t)) + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, \pi(s_t)) V^{\pi}(s_{t+1})$$

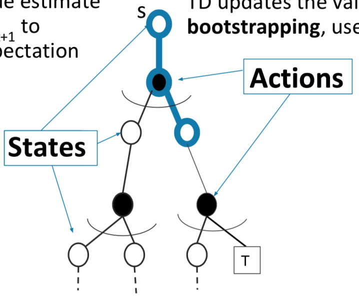
MC rest of future

DP

$$V^{\pi}(s_t) = V^{\pi}(s_t) + \alpha([r_t + \gamma V^{\pi}(s_{t+1})] - V^{\pi}(s_t))$$

TD updates the value estimate using a **sample** of s_{t+1} to approximate an expectation

TD updates the value estimate by **bootstrapping**, uses estimate of $V(s_{t+1})$



Check Your Understanding L3N2: Polleverywhere Poll Temporal Difference [$TD(0)$] Learning Algorithm

Input: α
Initialize $V^\pi(s) = 0, \forall s \in S$
Loop

- Sample **tuple** (s_t, a_t, r_t, s_{t+1})
- $V^\pi(s_t) = V^\pi(s_t) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s_t))$

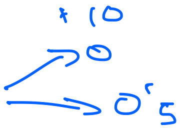
Select all that are true

1. If $\alpha = 0$ TD will weigh the TD target more than the past V estimate
2. If $\alpha = 1$ TD will update the V estimate to the TD target
3. If $\alpha = 1$ TD in MDPs where the policy goes through states with multiple possible next states, V may oscillate forever
4. There exist deterministic MDPs where $\alpha = 1$ TD will converge

$p(s'|s,a) = ?$

Break

Check Your Understanding L3N2: Polleverywhere Poll Temporal Difference [$TD(0)$] Learning Algorithm



Input: α
Initialize $V^\pi(s) = 0, \forall s \in S$
Loop

- Sample **tuple** (s_t, a_t, r_t, s_{t+1})
- $V^\pi(s_t) = \cancel{V^\pi(s_t)} + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - \cancel{V^\pi(s_t)})$

Answers. If $\alpha = 1$ TD will update to the TD target. If $\alpha = 1$ TD in MDPs where the policy goes through states with multiple possible next states, V may oscillate forever. There exist deterministic MDPs where $\alpha = 1$ TD will converge.

Summary: Temporal Difference Learning

- Combination of Monte Carlo & dynamic programming methods
- Model-free
- **Bootstraps and samples**
- Can be used in episodic or infinite-horizon non-episodic settings
- Immediately updates estimate of V after each (s, a, r, s') tuple

This Lecture: Policy Evaluation

- Estimating the expected return of a particular policy if don't have access to true MDP models
- Monte Carlo policy evaluation
    - Policy evaluation when don't have a model of how the world work
        - Given on-policy samples
- Temporal Difference (TD)
- Certainty Equivalence with dynamic programming
- Metrics to evaluate and compare algorithms

Recall: Dynamic Programming for Policy Evaluation

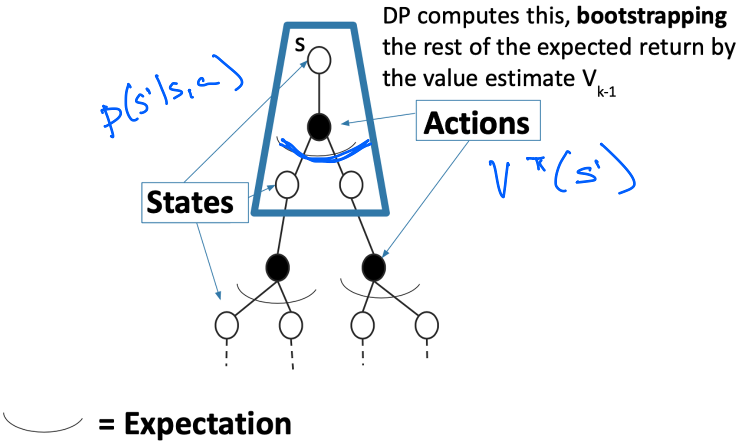
- If we knew dynamics and reward model, we can do policy evaluation
- Initialize $V_0^\pi(s) = 0$ for all s
- For $k = 1$ until convergence
  - For all s in S

$$V_k^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{k-1}^\pi(s')$$

- $V_k^\pi(s)$ is exactly the k-horizon value of state s under policy π
- $V_k^\pi(s)$ is an **estimate of the infinite horizon** value of state s under policy π

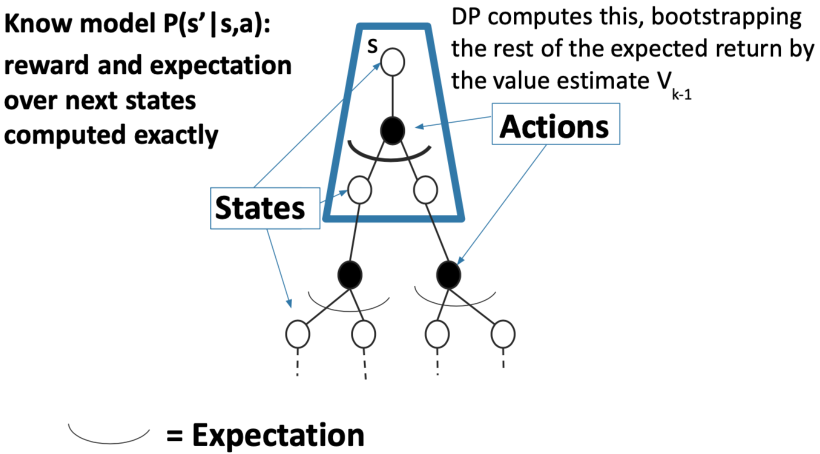
$$V^\pi(s) = \mathbb{E}_\pi[G_t|s_t = s] \approx \mathbb{E}_\pi[r_t + \gamma V_{k-1}|s_t = s]$$

Dynamic Programming Policy Evaluation $V^\pi(s) \leftarrow \mathbb{E}_\pi[r_t + \gamma V_{k-1} | s_t = s]$



- Bootstrapping: Update for V uses an estimate

Dynamic Programming Policy Evaluation $V^\pi(s) \leftarrow \mathbb{E}_\pi[r_t + \gamma V_{k-1} | s_t = s]$

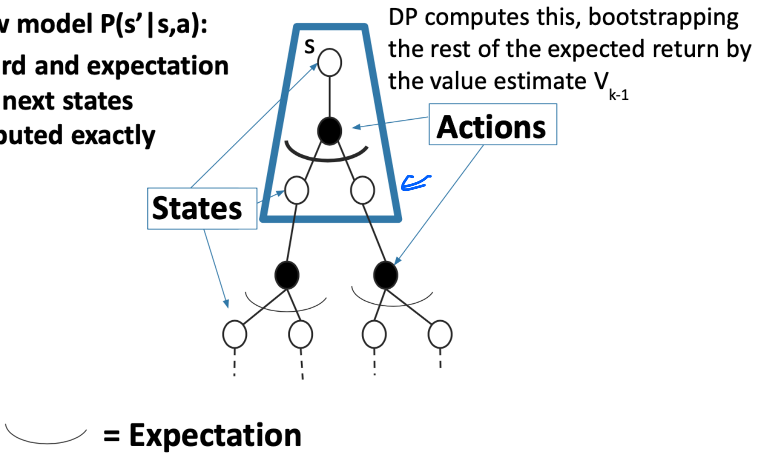


- Bootstrapping: Update for V uses an estimate

What about when we don't know the models?

Know model P(s'|s,a):

reward and expectation over next states computed exactly



Alternative: Certainty Equivalence V^π MLE MDP Model Estimates

- Model-based option for policy evaluation without true models
- After each (s_i, a_i, r_i, s_{i+1}) tuple
  - Recompute maximum likelihood MDP model for (s, a)

$$\hat{P}(s'|s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{i} \mathbb{1}(s_k = s, a_k = a, s_{k+1} = s')$$

$$\hat{r}(s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{i} \mathbb{1}(s_k = s, a_k = a) r_k$$

  - Compute V^π using MLE MDP [1] (using any method from lecture 2))

[1] Requires initializing for all (s, a) pairs

s_1	s_2	s_3	s_4	s_5	s_6	s_7
$R(s_1) = +1$ *Okay Field Site*	$R(s_2) = 0$	$R(s_3) = 0$	$R(s_4) = 0$ 	$R(s_5) = 0$	$R(s_6) = 0$	$R(s_7) = +10$ *Fantastic Field Site*

- Mars rover: R = [1 0 0 0 0 0 +10] for any action
- $\pi(s) = a_1 \ \forall s$, $\gamma = 1$. any action from s_1 and s_7 terminates episode
- Trajectory = (s_3, a_1, 0, s_2, a_1, 0, s_2, a_1, 0, s_1, a_1, 1, terminal)
- First visit MC estimate of V of each state? [1 1 1 0 0 0 0]
- TD estimate of all states (init at 0) with $\alpha = 1$ is [1 0 0 0 0 0 0]
- What is the certainty equivalent estimate?
- $\hat{r} = [1\ 0\ 0\ 0\ 0\ 0\ 0]$, $\hat{p}(terminate|s_1, a_1) = \hat{p}(s_2|s_3, a_1) = 1$
- $\hat{p}(s_1|s_2, a_1) = .5$, $\hat{p}(s_2|s_2, a_1) = .5$, V = [1 1 1 0 0 0 0]

Alternative: Certainty Equivalence V^π MLE MDP Model Estimates

- Model-based option for policy evaluation without true models
- After each (s, a, r, s') tuple MLE
  - Recompute maximum likelihood MDP model for (s, a)

$$\hat{P}(s'|s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{K} \sum_{t=1}^{L_k-1} 1(s_{k,t} = s, a_{k,t} = a, s_{k,t+1} = s')$$

$$\hat{r}(s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{K} \sum_{t=1}^{L_k-1} 1(s_{k,t} = s, a_{k,t} = a) r_{t,k}$$

  - Compute V^π using MLE MDP
- Cost: Updating MLE model and MDP planning at each update ($O(|S|^3)$ for analytic matrix solution, $O(|S|^2|A|)$ for iterative methods)
- Very data efficient and very computationally expensive
- Consistent (will converge to right estimate for Markov models)
- Can also easily be used for off-policy evaluation

This Lecture: Policy Evaluation

- Estimating the expected return of a particular policy if don't have access to true MDP models
- Monte Carlo policy evaluation
  - Policy evaluation when don't have a model of how the world work
    - Given on-policy samples
- Temporal Difference (TD)
- Certainty Equivalence with dynamic programming
- **Metrics to evaluate and compare algorithms**

Check Your Understanding L3N3: Properties of Algorithms for Evaluation.

	DPCE	MC	TD
Can use w/out access to true MDP models			
Usable in continuing (non-episodic) setting			
Assumes Markov process			
Converges to true value in limit[2]			
Unbiased estimate of value			

- DPCE = Dynamic Programming w/certainty equivalence estimates, MC = Monte Carlo, TD = Temporal Difference

[2] For tabular representations of value function. More on this in later lectures

Check Your Understanding L3N3: Properties of Algorithms for Evaluation.

	DPCE	MC	TD
Can use w/out access to true MDP models	X	X	X
Usable in continuing (non-episodic) setting	X		X
Assumes Markov process	X		X
Converges to true value in limit[3]	X	X	X
Unbiased estimate of value		X	

α correctly

- DPCE = Dynamic Programming w/certainty equivalence estimates, MC = Monte Carlo, TD = Temporal Difference

[3] For tabular representations of value function. More on this in later lectures

Some Important Properties to Evaluate Model-free Policy Evaluation Algorithms

- Bias/variance characteristics
- Data efficiency
- Computational efficiency
- Mostly focus on comparing MC and TD methods but we will connect back to dynamic programming with certainty equivalence methods later

Bias/Variance of Model-free Policy Evaluation Algorithms

- Return G_t is an unbiased estimate of $V^\pi(s_t)$
- TD target $[r_t + \gamma V^\pi(s_{t+1})]$ is a biased estimate of $V^\pi(s_t)$
- But often much lower variance than a single return G_t
- Return function of multi-step sequence of random actions, states & rewards
- TD target only has one random action, reward and next state
- MC
  - Unbiased (for first visit)
  - High variance
  - Consistent (converges to true) even with function approximation
- TD
  - Some bias
  - Lower variance
  - TD(0) converges to true value with tabular representation
  - TD(0) does not always converge with function approximation

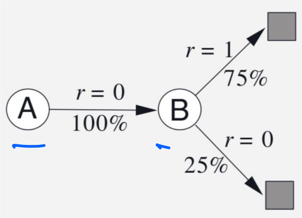
s_1	s_2	s_3	s_4	s_5	s_6	s_7
$R(s_1) = +1$ *Okay Field Site*	$R(s_2) = 0$	$R(s_3) = 0$	$R(s_4) = 0$ 	$R(s_5) = 0$	$R(s_6) = 0$	$R(s_7) = +10$ *Fantastic Field Site*

- Mars rover: R = [1 0 0 0 0 0 +10] for any action
- $\pi(s) = a_1\ \forall s$, $\gamma = 1$. any action from s_1 and s_7 terminates episode
- Trajectory = (s_3, a_1, 0, s_2, a_1, 0, s_2, a_1, 0, s_1, a_1, 1, terminal)
- First visit MC estimate of V of each state? [1 1 1 0 0 0 0]
- TD estimate of all states (init at 0) with $\alpha = 1$ is [1 0 0 0 0 0 0]
- TD(0) only uses a data point (s, a, r, s') once
- Monte Carlo takes entire return from s to end of episode

Batch MC and TD

- Batch (Offline) solution for finite dataset
  - Given set of K episodes
  - Repeatedly sample an episode from K se f
  - Apply MC or TD(0) to the sampled episode
- What do MC and TD(0) converge to?

AB Example: (Ex. 6.4, Sutton & Barto, 2018)



- Two states A, B with $\gamma = 1$
- Given 8 episodes of experience:
  - $A, 0, B, 0$ A → r=0 → B → r = 0
  - $B, 1$ (observed 6 times)
  - $B, 0$
- Imagine run TD updates over data infinite number of times
- $V(B) =$

AB Example: (Ex. 6.4, Sutton & Barto, 2018)

- TD Update: $V^\pi(s_t) = V^\pi(s_t) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s_t))$

$\frac{1}{n}$

↑ # updates

- Two states A, B with $\gamma = 1$

MC $V(B) = \frac{6}{8} = 3/4$

- Given 8 episodes of experience:
  - $A, 0, B, 0$
  - $B, 1$ (observed 6 times)
  - $B, 0$
- Imagine run TD updates over data infinite number of times
- $V(B) = 0.75$ by TD or MC
- What about $V(A)$?

MC

0

TD

3/4

$0 + \gamma V^\pi(B)$

AB Example: (Ex. 6.4, Sutton & Barto, 2018)

- TD Update: $V^\pi(s_t) = V^\pi(s_t) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s_t))$
- Two states A, B with $\gamma = 1$
- Given 8 episodes of experience:
  - $A, 0, B, 0$
  - $B, 1$ (observed 6 times)
  - $B, 0$
- Imagine run TD updates over data infinite number of times
- $V(B) = 0.75$ by TD or MC
- What about $V(A)$?
  $V^{MC}(A) = 0 \; V^{TD}(A) = .75$
  Markov

Batch MC and TD: Converges

- Monte Carlo in batch setting converges to min MSE (mean squared error)
  - Minimize loss with respect to observed returns
  - In AB example, $V(A) = 0$
- TD(0) converges to DP policy V^π for the MDP with the maximum likelihood model estimates
- Aka same as dynamic programming with certainty equivalence!
  - Maximum likelihood Markov decision process model

$$\hat{P}(s'|s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{i} \mathbb{1}(s_k = s, a_k = a, s_{k+1} = s')$$

$$\hat{r}(s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{i} \mathbb{1}(s_k = s, a_k = a) r_k$$

  - Compute V^π using this model
  - In AB example, $V(A) = 0.75$

Some Important Properties to Evaluate Model-free Policy Evaluation Algorithms

- Data efficiency & Computational efficiency
- In simplest TD, use (s, a, r, s') once to update $V(s)$
  - $O(1)$ operation per update
  - In an episode of length L, $O(L)$
- In MC have to wait till episode finishes, then also $O(L)$
- MC can be more data efficient than simple TD
- But TD exploits Markov structure
  - If in Markov domain, leveraging this is helpful
- Dynamic programming with certainty equivalence also uses Markov structure

Summary: Policy Evaluation

Estimating the expected return of a particular policy if don't have access to true MDP models. Ex. evaluating average purchases per session of new product recommendation system

- Monte Carlo policy evaluation
    - Policy evaluation when we don't have a model of how the world works
        - Given on policy samples
        - Given off policy samples
- Temporal Difference (TD)
- Dynamic Programming with certainty equivalence
- Metrics to evaluate and compare algorithms
    - Robustness to Markov assumption
    - Bias/variance characteristics
    - Data efficiency
    - Computational efficiency

Today's Plan

- Last Time:
  - Markov reward / decision processes
  - Policy evaluation & control when have true model (of how the world works)
- Today
  - Policy evaluation without known dynamics & reward models
- Next Time:
  - Control when don't have a model of how the world works