ECE 414/517 Reinforcement Learning

LECTURE 11: MONTE CARLO METHODS

SEP. 29 2022

Monte Carlo (Exploring Starts)

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:
- $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
- $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
- $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
- Choose $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}(S_0)$ such that all pairs have probability > 0
- Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$
- Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
  - $G \leftarrow \gamma G + R_{t+1}$
  - Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
    - Append G to $Returns(S_t, A_t)$
    - $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
    - $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

No Exploring Starts

1. Exploring starts not always possible.
   - Why?
2. On the other hand can't simply follow greedy approach – no exploration.
3. We can employ an ϵ-greedy policy instead, where $\pi(s,a) > 0 \quad \forall \quad s,a$:
   - With probability $1-\epsilon$ choose the greedy action
   - With probability ϵ – explore uniformly
4. Probabilities of actions:
   - For non max actions: $\frac{\epsilon}{|\mathcal{A}(s)|}$
   - For max action: $1-\epsilon+\frac{\epsilon}{|\mathcal{A}(s)|}$

Policy Improvement under ϵ-greedy

Show that the policy improvement theory holds.

$$
\begin{aligned}
Q^{\pi_k}(s, \pi_{k+1}(s)) &= \sum_a \pi_{k+1}(a|s)\, Q^{\pi_k}(s,a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a Q^{\pi_k}(s,a) + (1-\epsilon) \max_a Q^{\pi_k}(s,a) \\
&\geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a Q^{\pi_k}(s,a) + (1-\epsilon) \sum_a \frac{\pi_k(a|s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} Q^{\pi_k}(s,a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a Q^{\pi_k}(s,a) - \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a Q^{\pi_k}(s,a) + \pi_k(a|s) Q^{\pi_k}(s,a) \\
&= V^{\pi_k}(s)
\end{aligned}
$$

On policy Monte Carlo

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

  $\pi \leftarrow$ an arbitrary ε-soft policy

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

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

Repeat forever (for each episode):

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

  $G \leftarrow 0$

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

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

    Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

      Append G to $Returns(S_t, A_t)$

      $Q(S_t, A_t) \leftarrow$ average$(Returns(S_t, A_t))$

      $A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

      For all $a \in \mathcal{A}(S_t)$:

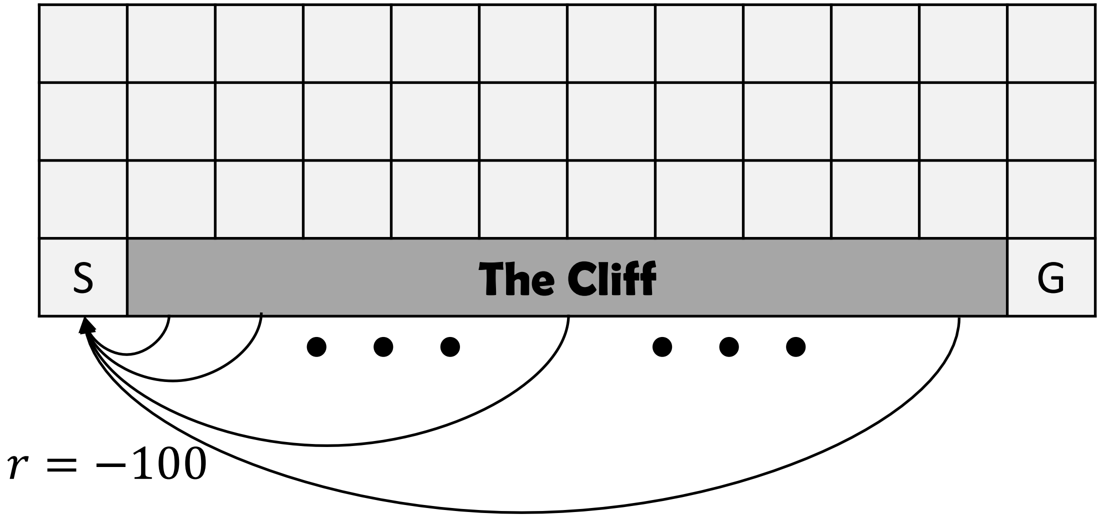
$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

On-policy Vs. Off-policy

1. **On-policy methods,** are methods where you evaluate and improve the policy that is used to create the episodes.
   - To ensure all actions are selected we need to employ an ϵ-soft policy
2. Instead we can use **off-Policy methods.** In this case the policy we use to generate the episodes is different than the behavior we are trying to evaluate and improve.
   - Behavior policy $b(a|s)$ – The policy used to generate episodes
   - Estimation policy $\pi(a|s)$ – The policy being evaluated and improved.
3. This has some advantages:
   - Include On-policy as special case ($b(a|s) = \pi(a|s)$)
   - Estimation policy can be deterministic (e.g. greedy) and therefore optimal vs near optimal (ϵ−soft).
   - Can learn from observing the behavior of other non-learning agents (such as humans)
4. Disadvantages:
   - More complex and slower to converge.

Cliff walking

1. Need to go from S to G.
2. Actions are standard S,N,E,W
3. Reward is -1 for every move until goal is reached
4. In addition any move into “the cliff” results in a -100 reward and sends the agent back to the start.



Off Policy Prediction – Importance Sampling

1. In order to be able to estimate $\pi(a|s)$ from $\mathrm{b}(a|s)$ we require **coverage**. That is:
$$\pi(a|s) > 0 \quad \Rightarrow \quad b(a|s) > 0$$
   - Implies that $b(a|s)$ must be stochastic in actions not in $\pi(a|s)$
2. On the other hand $\pi(a|s)$ can be a deterministic greedy policy with respect to the current estimate of the action value function.
3. How to estimate action values for $\pi(a|s)$ from $\mathrm{b}(a|s)$?
   - **Importance sampling:** weigh returns according to the relative probability of a trajectory occurring under the target and behavior policies.
4. Probability of a trajectory:

$$P(A_t, S_{t+1}, A_{t+1}, \dots, S_T \,|S_t, A_{t:T-1} \sim \pi)$$
$$= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t) \times \pi(A_{t+1}|S_{t+1})p(S_{t+2}|S_{t+1}, A_{t+1}) \dots p(S_T|S_{T-1}, A_{T-1})$$
$$= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)$$

Off Policy Prediction – Importance Sampling

1. We then can define the relative probability of the trajectory under the target and behavior policies:

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

2. Using this we can calculate $V^\pi(s)$ given $V^b(s)$:

$$V^b(s) = \mathbb{E}_b(G_t|S_t = s)$$
$$V^\pi(s) = \mathbb{E}_b(\rho_{t:T-1} \times G_t|S_t = s)$$

3. Why does this work:

$$\mathbb{E}_b(\rho_{t:T-1} \times G_t|S_t = s) = \mathbb{E}_b\left(\frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)} \times G_t \middle| S_t = s\right) =$$

$$\sum_{\{trajectories\}} \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)} \times G_t \times \prod_{k=t}^{T-1} b(A_k|S_k) = \sum \prod_{k=t}^{T-1} \pi(A_k|S_k) \times G_t = \mathbb{E}_\pi(G_t|S_t = s)$$

Off Policy Prediction – Importance Sampling

Ordinary importance sampling: $V^{\pi}(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} \times G_t}{|\mathcal{T}(s)|}$

Weighted importance sampling: $V^{\pi}(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} \times G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$

Where $T(t)$ is the time of the end of an episode which step s_t is in, $\mathcal{T}(s)$ is all the times we visit state s (or only first visits), and G_t are the returns from t to $T(t)$.

Use weighted importance sampling to ensure variance is finite.

- Ordinary importance sampling: unbiased, but variance is unbounded because ratios are unbounded
- Weighted importance sampling: biased, but bias converges to zero asymptotically and variance is bounded

Off Policy Prediction – Importance Sampling

1. Estimate the value of hitting for anything under 20, given a 50%-50% policy.
2. We do this just for one state: (13,2,1)
3. Use both ordinary and weighted importance sampling

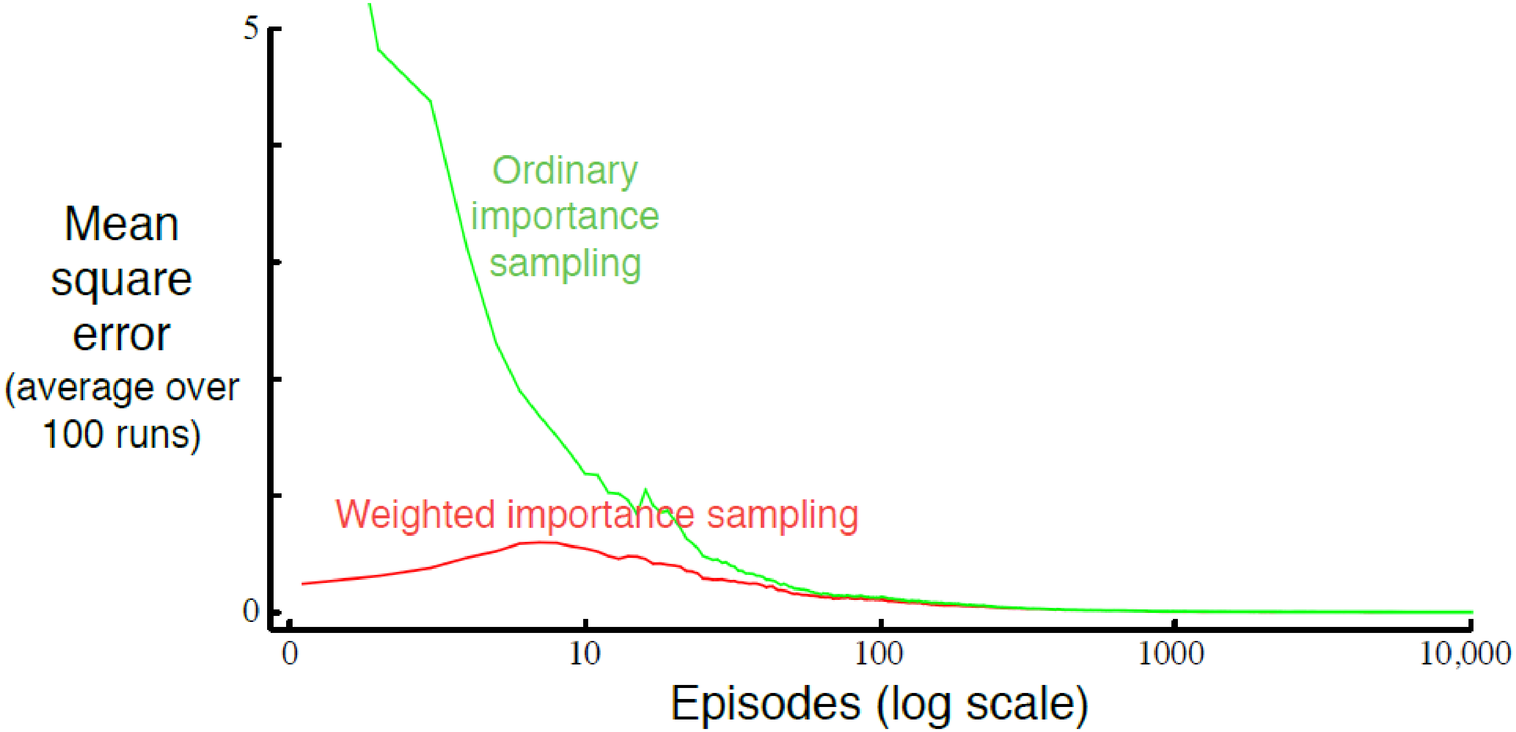
$$\rho_{t:T-1} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)}$$

$$V^{\pi}(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} \times G_t}{|\mathcal{T}(s)|} \quad \text{or}$$

$$V^{\pi}(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} \times G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

Off Policy Prediction – Importance Sampling

1. Estimate the value of hitting for anything under 20, given a 50%-50% policy.
2. We do this just for one state: (13,2,1)
3. Use both ordinary and weighted importance sampling



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

$$V^\pi(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1} \times G_t}{|\mathcal{T}(s)|} \text{ or}$$

$$V^\pi(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1} \times G_t}{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1}}$$

Importance Sampling – Incremental Implementation

1. Instead of storing all possible returns we can use similar methods to those described in Chapter 2.

$$NewEstimate \leftarrow OldEstimate + StepSize[NewData - OldEstimate]$$

2. Ordinary Sampling:
   - $V_{n+1} = V_n + \frac{1}{n}[\rho G_n - V_n]$

3. Weighted Sampling:
   - $V_n = \frac{\sum_{k=1}^{n-1} \rho_k G_k}{\sum_{k=1}^{n-1} \rho_k}, \quad \mathrm{n} \geq 2$
   - Define: $C_{n+1} = C_n + \rho_{n+1}$
   - And therefore: $V_{n+1} = V_n + \frac{\rho_n}{C_n}[G_n - V_n]$

Off Policy MC Prediction

Off-policy MC prediction (policy evaluation) for estimating $Q \approx q_\pi$

Input: an arbitrary target policy π
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$

Loop forever (for each episode):
    $b \leftarrow$ any policy with coverage of π
    Generate an episode following b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$
        If $W = 0$ then exit For loop

Off Policy MC Prediction

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
 $Q(s,a) \in \mathbb{R}$ (arbitrarily)
 $C(s,a) \leftarrow 0$
 $\pi(s) \leftarrow \arg\max_a Q(s,a)$ (with ties broken consistently)

Loop forever (for each episode):
 $b \leftarrow$ any soft policy
 Generate an episode using b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
 $G \leftarrow 0$
 $W \leftarrow 1$
 Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
  $G \leftarrow \gamma G + R_{t+1}$
  $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
  $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
  $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
  If $A_t \neq \pi(S_t)$ then exit For loop
  $W \leftarrow W \frac{1}{b(A_t|S_t)}$