AML, 29 Oct 2019

Monte Carlo Methods

$\pi \rightarrow$ Calculate V_π (Policy Evaluation)

$\pi_0 \xrightarrow{e} V_{\pi_0} \xrightarrow{\text{greedy}} \pi_1 \xrightarrow{e} V_{\pi_1} \cdots$ Policy Iteration

May be done incrementally - Generalized PI

Given π - generate random episodes
- observe rewards
- work backwards & update V_π

Instead of v_π, estimate $q_\pi(s,a)$

- Ensure all (s,a) pairs are visited often enough
  - Exploring Starts : choose random initial (s,a)
  - Or, used ϵ-soft / ϵ-greedy strategies

ϵ-soft ⟶ $\pi(s,a) \geq \dfrac{\epsilon}{|N(s)|}$ always

$|N(s)|$ = # of actions at s

ϵ-greedy ⟹ Choose non greedy with $\dfrac{1}{|N(s)|} \cdot \epsilon$

Choose greedy with $\left(1 - \epsilon + \dfrac{1}{|N(s)|}\right)$

Want to converge to an optimum policy

- Typically policy is greedy / deterministic
- Explore vs exploit
  - Need to consciously deviate from current optimum

<u>Instead</u>

- Use a <u>different</u> policy to generate samples

<u>Off Policy</u> strategy vs <u>On policy</u> strategy

Use b to sample & update π

Same π to sample & update

Sample according to b, but update π

- Estimates via b samples generate expected values wrt b, not π!

Basic constraint

If $\pi(a|s) > 0$ then $b(a|s) > 0$

$\Downarrow$ Typically greedy, deterministic

$\Downarrow$ Must be stochastic, in general

Suppose we are at state S_t

Remaining trajectori $A_t, S_{t+1}, A_{t+1}, \ldots, A_{T-1}, S_T$

According to π

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

$$= \pi(A_t \mid S_t)\, p(S_{t+1} \mid S_t, A_t)\, \pi(A_{t+1} \mid S_{t+1}) \cdots$$

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

Instead, if we b rather than π

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

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

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

$$E\left[G_t \mid S_t = s\right] = v_b(s)$$

Instead

$$E\left[\rho_{t:T-1} \cdot G_t \mid S_t = s\right] = v_\pi(s)$$

Importance Sampling

All visits to s contribute to $V_\pi(s)$

$$V(s) = \frac{\sum_{t \in \tau(s)} \rho_{t:T(t)-1} G_t}{|\tau(s)|}$$

Adjusted observation

time points where s is visited

end point of the episode in which t occurs

Uniform count across episodes

$$e_1\ e_2 \ldots e_T \;\Big|\; e_{T+1}\ e_{T+2} \ldots e_{T+T'} \;\Big|$$

Episode 1 | episode 2

Previous ratio is called Ordinary importance sampling

Instead Weighted Importance Sampling

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

Consider a single episode

Single observatn

Ordinary

$$V(s) = \rho_{t:T-1} \cdot G_t$$

Lower bias

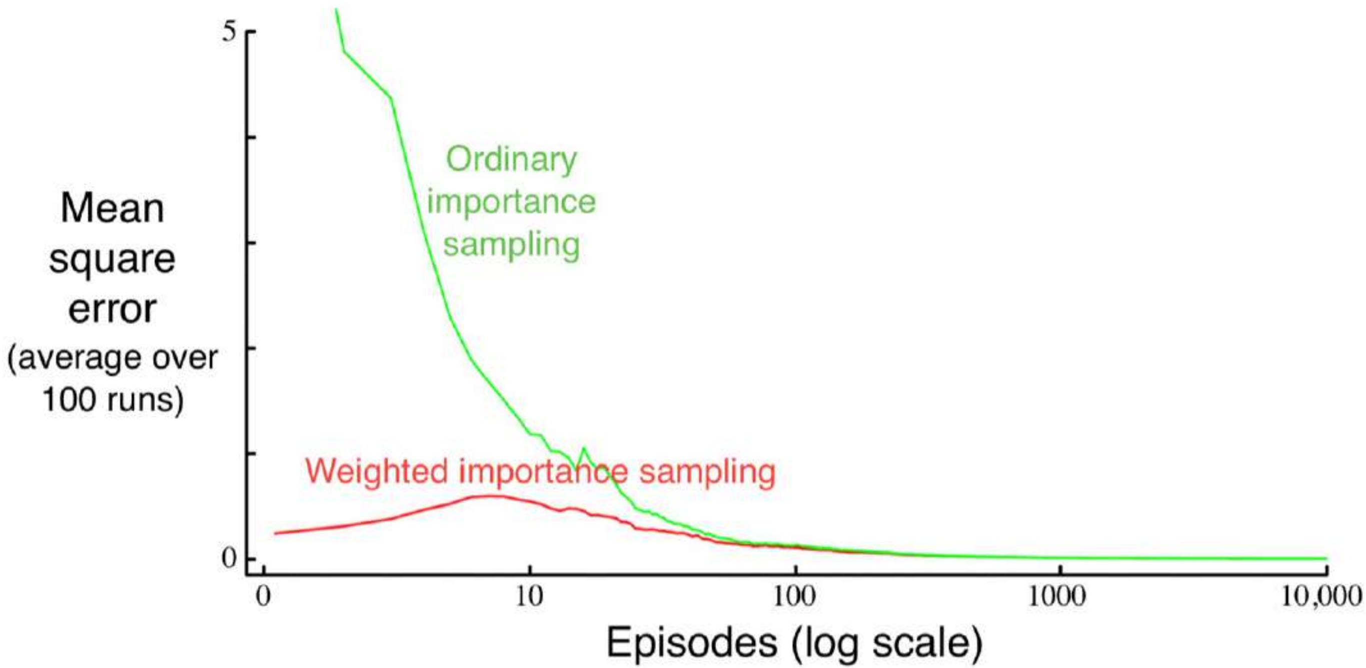
Higher variance

Weighted

$$V(s) = \frac{\cancel{\rho_{t:T-1}} \cdot G_t}{\cancel{\rho_{t:T-1}}}$$

Higher bias

Lower variance



<u>Exercise</u>: Weighted importance sampling can be computed incrementally

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, $t = T-1, T-2, \ldots, 0$, while $W \neq 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \underline{\frac{W}{C(S_t, A_t)}} [G - Q(S_t, A_t)]$
        $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

Generalize this to do policy iteration

Dynamic Programming

Update V_i to V_{i+1}

Bootstrapping

MDP theory

Monte Carlo

V_{i+1} replaces V_i

Not bootstrapped

Temporal Difference Learning

a "true" contribution of RL

Sampling + Bootstrapping

Monte Carlo

$$V(S_t) \leftarrow V(S_t) + \alpha\left[G_t - V(S_t)\right]$$

estimate

TD

$$V(S_t) \leftarrow V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right]$$

TD(0) - single step lookahead

$$V_\pi(s) \overset{\text{def}}{=} E\left[G_t \mid S_t = s\right]$$ — Monte Carlo

$$= E\left[R_{t+1} + \lambda G_{t+1} \mid S_t = s\right]$$ — Defn of rewards & values

$$= E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right]$$

— TD(0)

Example Driving time

Office $\longrightarrow$ Highway $\longrightarrow$ Side road $\longrightarrow$ Home

State	*Elapsed Time (minutes)*	*Predicted Time to Go*	*Predicted Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43