5 Monte Carlo RL

Melih Kandemir

Özyeğin University
Computer Science Department
melih.kandemir@ozyegin.edu.tr

24 Oct 2017

Monte Carlo RL

- (+) learns directly from episodes of experience.
- (+) is *model-free* (i.e. requires no knowledge of MDP transitions and rewards).
- (+) is based only on generated sample transitions, not complete distributions of all possible transitions.
- (-) works only for *episodic* tasks.
- (o) applies Monte Carlo integration to value approximation.

Monte Carlo Integration

$$\mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz$$
$$\approx \sum_{l=1}^{L} f(z^{(l)}),$$

where

$$z^{(1)}, z^{(2)}, \cdots z^{(L)} \sim p(z).$$

First-visit Monte Carlo method

- **Visit:** Each occurrence of state s in an episode.
- **First-visit MC:** estimate $v_\pi(s)$ as the average return following the first visits to s.

For a certain state s, assume we observe three episodes

$$S_1^{(1)}, A_1^{(1)}, R_2^{(1)}, S_2^{(1)}, A_2^{(1)}, R_3^{(1)}, S_3^{(1)} = s, A_3^{(1)}, R_4^{(1)}, S_5^{(1)} = s_{end}$$
$$S_1^{(2)}, A_1^{(2)}, R_2^{(2)}, S_2^{(2)}, A_2^{(2)}, R_3^{(2)}, S_3^{(2)}, A_3^{(2)}, R_4^{(2)},$$
$$S_4^{(2)} = s, A_4^{(2)}, R_5^{(2)}, S_5^{(2)} = s, A_5^{(2)}, R_6^{(2)}, S_6^{(2)} = s_{end}$$
$$S_1^{(3)}, A_1^{(3)}, R_2^{(3)}, S_2^{(3)} = s, A_2^{(3)}, R_3^{(3)}, S_3^{(3)} = s_{end},$$

the first-visit estimate of the value function is

$$v_\pi(s) \approx \frac{1}{3}\Big[R_4^{(1)} + R_5^{(2)} + R_6^{(2)} + R_3^{(3)}\Big].$$

Every-visit Monte Carlo method

Estimate $v_\pi(s)$ as the average return following *all* visits to s within an episode

$$v_\pi(s) \approx \frac{1}{3}\left[R_4^{(1)} + R_5^{(2)} + R_6^{(2)} + R_6^{(2)} + R_3^{(3)}\right].$$

Convergence

- Both first-visit and every-visit Monte Carlo converge to $v_\pi(s)$ as the number of visits go to infinity.
- Both averages are *unbiased* estimators and their standard error converges quadratically $(1/\sqrt{n})$.

Advantages of MC over DP

- MC can learn from the agent's own experience.
- MC can learn from simulation
  (as big data as the computer can generate and as cheap data as the electricity).
- MC estimates for each state are independent
  (computational expense of estimating the value of one state does not depend on the state space size).

First-visit MC algorithm

Initialize:

$\pi \leftarrow$ policy to be evaluated

$V \leftarrow$ an arbitrary state-value function

$Returns(s) \leftarrow$ an empty list $\forall s \in \mathcal{S}$

repeat forever

Generate an episode using π

foreach state s appearing in the episode

$G \leftarrow$ return following the first occurrence of s

Append G to $Returns(s)$

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

MC Action Value Estimation with Exploring Starts

- Greedy policy improvement over $V(s)$ requires the model of the MDP

$$\pi'(s) \leftarrow \underset{a}{\operatorname{argmax}} \sum_{s',r} p(s', r|s, a)\Big[r + \gamma V(s')\Big].$$

- Greedy policy improvement over $Q(s, a)$ is model-free

$$\pi'(s) \leftarrow \underset{a}{\operatorname{argmax}}\, Q(s, a).$$

- Suppose the policy is greedy and the MDP is deterministic, then the entire episode following (s, a) is determined.
  Nothing to average!
- **Remedy:** Choose random (s, a).
  This is called *Exploring Starts (ES)*.

The MC-ES Algorithm

Initialize for all $s \in \mathcal{S}, a \in \mathcal{A}$:

$Q(s,a) \;\leftarrow\; \text{arbitrary}$

$\pi(s) \;\leftarrow\; \text{arbitrary}$

$Returns(s,a) \leftarrow\; \text{empty list}$

repeat forever

Choose $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}$ s.t. all pairs have probability > 0

Generate an episode starting from S_0, A_0, following π

For each pair (s,a) appearing in the episode:

$G \leftarrow$ return following the first occurrence of s, a

Append G to $Returns(s,a)$

$Q(s,a) \leftarrow \text{average}(Returns(s,a))$

For each s in the episode:

$\pi(s) \leftarrow \underset{a}{\text{argmax}}\, Q(s,a)$

The policy improvement theorem for MC-ES

For a given q, the corresponding greedy policy is

$$\pi(s) = \operatorname*{argmax}_a q(s,a).$$

Then,

$$\begin{aligned}
q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \operatorname*{argmax}_a q_{\pi_k}(s,a)) \\
&= \max_a \; q_{\pi_k}(s,a) \\
&\geq q_{\pi_k}(s, \pi_k(s)) \\
&\geq v_{\pi_k}(s).
\end{aligned}$$

Convergence of Monte Carlo ES

- Evaluation and improvement steps alternate on an episode-by-episode basis.
- Intuitively, Monte Carlo ES can converge only to the optimal policy.
- Convergence to a suboptimal policy would follow convergence to the related value function.
- The next step would again be an inevitable policy improvement.
- To date, the theoretical reasons for this nice property are yet unknown!

How to avoid ES?

- ES is not a plausible assumption. It is hard to target important states in large state spaces.
- Classify RL methods into two:
  - **On-policy** methods generate data from the policy being learned.
  - **Off-policy** methods use different policies for learning and data generation.
- On-policy methods should assume **soft** policies to assure exploration (i.e. $\pi(a|s) > 0,\ \forall s, a$).
- Off-policy methods survive from ES by definition (choose the data generation policy accordingly).

$\epsilon-$*greedy* policies

$$\pi(a|s) = \begin{cases} \epsilon/|\mathcal{A}(s)| + 1 - \epsilon, & \text{if } a_* = \underset{a}{\operatorname{argmax}}\, q(s,a) \\ \epsilon/|\mathcal{A}(s)|, & \text{otherwise} \end{cases}$$

- Assure that the action probabilities sum up to 1.
- $\epsilon-$greedy policies are examples of $\epsilon-$soft policies, as $\pi(a|s) \geq \dfrac{\epsilon}{|\mathcal{A}(s)|}$ for all (s,a).

On-policy first-visit MC control

Initialize for all $s \in \mathcal{S}, a \in \mathcal{A}$:

$Q(s,a) \leftarrow$ arbitrary

$Returns(s,a) \leftarrow$ empty list

$\pi(s|a) \leftarrow$ an arbitrary $\epsilon-$soft policy

repeat forever

(a) Generate an episode using π

(b) For each pair (s,a) appearing in the episode:

$G \leftarrow$ return following the first occurrence of s, a

Append G to $Returns(s,a)$

$Q(s,a) \leftarrow \mathrm{average}(Returns(s,a))$

For each s in the episode:

$A^* \leftarrow \underset{a}{\mathrm{argmax}}\, Q(s,a)$

For all $a \in \mathcal{A}$:

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

$\epsilon-$*greedy* policy improvement theorem

Theorem. For any $\epsilon-$greedy policy π, the $\epsilon-$greedy policy π' wrt q_π is an improvement, $v_{\pi'} \geq v_\pi$.

Proof.

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_a \pi'(a|s) q_\pi(s,a) \\
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \max_a \; q_\pi(s,a) \\
&\geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} \; q_\pi(s,a) \\
&= \sum_a \pi(a|s) q_\pi(s,a) = v_\pi(s) \qquad \blacksquare
\end{aligned}
$$

Off-policy prediction via importance sampling

Dilemma of all learning control methods:

- learn $q_*(s, a)$, hence behavior of π_* should be observed
- need to behave non-optimally to explore *all* actions

Solution is to use two policies instead of one:

- **target policy:** policy being learned (π)
- **behavior policy:** policy that generates behavior (b)

Because $\pi \neq b$, we call this approach *off-policy* RL.

Pros and cons of off-policy RL

- Off-policy methods incur higher variance, hence converge slower than on-policy methods.
- Off-policy methods have on-policy methods as their special case, hence they are more general and powerful.
- Off-policy methods can learn from a non-learning controller (e.g. a human expert), on-policy methods cannot.

Importance Sampling (IS)

Intuition: Sample from a different distribution from the one being integrated.

$$\mathbb{E}_{p(z)}\Big[f(z)\Big] = \int f(z)p(z)dz \qquad = \int f(z)\frac{p(z)}{q(z)}q(z)dz$$

then do Monte Carlo integration

$$\mathbb{E}_{p(z)}\Big[f(z)\Big] \approx \frac{1}{K}\sum_{k=1}^{K} f(z^{(k)}) \times \underbrace{\frac{p(z^{(k)})}{q(z^{(k)})}}_{\text{Importance weight}}$$

for a set of $z^{(k)} \sim q(z)$.

IS applied to MC-RL

We require a behavior policy such that

$$\pi(a|s) > 0 \Rightarrow b(a|s) > 0, \quad \forall(s,a)$$

which is called the *coverage* assumption.
Given a starting state S_t, the probability of the subsequent state-action trajectory

$$A_t, S_{t+1}, A_{t+1}, \cdots, S_T$$

realized under policy π is

$$\begin{aligned}
&P(A_t, S_{t+1}, A_{t+1}, \cdots, S_T | S_t, A_{t:T} \sim \pi) \\
&\qquad = \pi(A_t|S_t)P(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1}) \\
&\qquad = \prod_{k=t}^{T-1} \pi(A_k|S_k)P(S_{k+1}|S_k, A_k).
\end{aligned}$$

IS applied to MC-RL

In our application

- $f(z) \leftarrow G$
- $p(z) \leftarrow \prod_{k=t}^{T-1} \pi(A_k|S_k)P(S_{k+1}|S_k, A_k)$
- $q(z) \leftarrow \prod_{k=t}^{T-1} b(A_k|S_k)P(S_{k+1}|S_k, A_k)$

Then the importance weight reads

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

which does not depend on the MDP!

Some definitions

- $t = [\underbrace{1, 2 \cdots, 103}_{episode\ 1}, \underbrace{104, 105, \cdots, 248}_{episode\ 2}, 249, \cdots]$
- $\mathcal{T}(s) \triangleq \{t | S_t = s\}$ (for the every-visit case)
- $T(t)$ is first-time termination after t
- G_t is return from t to $T(t)$
- $\{G_t\}_{t \in \mathcal{T}(s)}$ are returns for state s
- $\{\rho_{t:T(t)-1}\}_{t \in \mathcal{T}(s)}$ are corresponding importance weights

Ordinary vs Weighted IS

The value function $v_\pi(s)$ can be estimated in different ways.

Ordinary IS

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

Weighted IS

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

Ordinary vs Weighted IS

- Ordinary IS is unbiased ($\mathbb{E}[V(s)] = v_\pi(s)$), but its variance is unbounded (due to the importance weight).
- Weighted IS is biased ($\mathbb{E}[V(s)] = v_b(s) \neq v_\pi(s)$), but its variance is bounded.
- Weighted IS is preferred more often.
- Bias of Weighted IS converges to zero. Hence, it is asymptotically unbiased.
- Ordinary IS has poor convergence properties.
- Unbounded variance of IS is a headache especially if the trajectory contains loops.

Example: Infinite variance

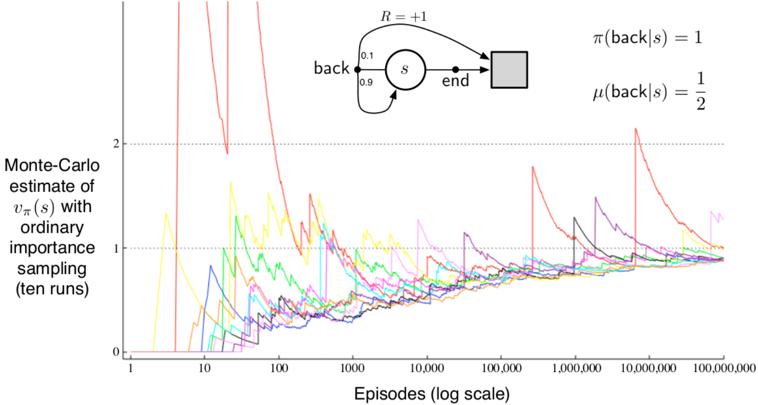

Figure. R. Sutton and A. Barto, MIT Press, 2017

Example: Infinite variance

Variance is defined as

$$
\begin{aligned}
Var[X] &\triangleq E[(X-\bar{X})^2] \\
&= E[X^2] - 2E[X\bar{X}] + E[\bar{X}^2] \\
&= E[X^2] - 2\underbrace{E[X]}_{\bar{X}}\bar{X} + \bar{X}^2 \\
&= E[X^2] - \bar{X}^2.
\end{aligned}
$$

Given that $\bar{X}$ is infinite as in our case, then $Var[X]$ is infinite if $E[X^2]$ is infinite!

Example: Infinite variance

- Show that $E[X^2]$ is infinite
- Discard all episodes ending with the right action. The target policy will never take it.
- Only consider episodes full of left actions, all perform a self transition except the last one moving to the end state and terminating the episode.

Example: Infinite variance

$$
\begin{aligned}
\mathbb{E}_b & \left[\left(\prod_{t=0}^{T-1} \frac{\pi(A_t|S_t)}{b(A_t|S_t)} G_0 \right)^2 \right] \\
= & \frac{1}{2} \cdot 0.1 \cdot \left(\frac{1}{0.5} \right)^2 && \text{(episode of length 1)} \\
& + \frac{1}{2} \cdot 0.9 \cdot \frac{1}{2} \cdot 0.1 \cdot \left(\frac{1}{0.5} \frac{1}{0.5} \right)^2 && \text{(episode of length 2)} \\
& + \frac{1}{2} \cdot 0.9 \frac{1}{2} \cdot 0.9 \cdot \frac{1}{2} \cdot 0.1 \cdot \left(\frac{1}{0.5} \frac{1}{0.5} \frac{1}{0.5} \right)^2 \\
& + \cdots \\
= & 0.1 \sum_{k=0}^{\infty} 0.9^k \cdot 2^k \cdot 2 = 0.2 \sum_{k=0}^{\infty} 1.8^k = \infty \qquad \blacksquare
\end{aligned}
$$

Incremental averaging

The average return of n episodes can be calculated in batch as follows

$$V_n \triangleq \frac{G_1 + G_2 + \cdots + G_n}{n}.$$

The same average can also be calculated incrementally, without storing the list of all individual rewards

$$\begin{aligned}
V_n \triangleq & \frac{1}{n}\sum_{k=1}^{n} G_k \\
= & \frac{1}{n}\left[G_n + \sum_{k=1}^{n-1} G_k\right] = \frac{1}{n}\left[G_n + (n-1)\frac{1}{1-n}\sum_{k=1}^{n-1} G_k\right] \\
= & \frac{1}{n}\left[G_n + (n-1)V_{n-1}\right] = \frac{1}{n}\left[G_n + nV_{n-1} - V_{n-1}\right] \\
= & V_{n-1} + \frac{1}{n}\left[G_n - V_{n-1}\right]
\end{aligned}$$

Let us interpret the update

$$V_n \quad \leftarrow \quad V_{n-1} \; + \; \underbrace{\frac{1}{n}}_{learning\ rate} \underbrace{\left[\underbrace{G_n}_{target} - \underbrace{V_{n-1}}_{estimate} \right]}_{estimation\ error}$$

Let us generalize the update

$$V_n \leftarrow V_{n-1} + \underbrace{\color{red}{\alpha}}_{learning\ rate} \underbrace{\left[\underbrace{G_n}_{target} - \underbrace{V_{n-1}}_{estimate} \right]}_{estimation\ error}$$

- Generalize the learning rate from the number of samples to an arbitrary number such that $\alpha \in (0, 1]$.
- Now the update calculates a **running average** (i.e. forgets the past after $1/\alpha$ steps back).

Incremental Ordinary IS

Given a sequence $G_1, G_2, \cdots, G_{n-1}$, all starting with the same state and having the corresponding importance weights $W_1, W_2, \cdots, W_{n-1}$, where $W_i = \rho_{t:T(t)-1}$, we can calculate the average

$$V_n \triangleq \frac{1}{n} \sum_{k=1}^{n} W_k G_k, \qquad n \geq 2,$$

following the update rule

$$V_n \leftarrow V_{n-1} + \frac{1}{n} \Big[W_n G_n - V_{n-1} \Big], \qquad n \geq 1.$$

Incremental Weighted IS

Given a sequence $G_1, G_2, \cdots, G_{n-1}$, all starting with the same state and having the corresponding importance weights $W_1, W_2, \cdots, W_{n-1}$, where $W_i = \rho_{t:T(t)-1}$, we aim to estimate

$$V_n \triangleq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \quad n \geq 2$$

using incremental updates.

Incremental Weighted IS

Define $C_n \leftarrow C_{n-1} + W_n$ with $C_0 \triangleq 0$, then

$$\begin{aligned} C_n V_n &= G_n W_n + C_{n-1} V_{n-1} \\ &= G_n W_n + (C_n - W_n) V_{n-1} \\ &= G_n W_n + C_n V_{n-1} - W_n V_{n-1} \end{aligned}$$

Divide both sides by C_n

$$\begin{aligned} V_n &= \frac{G_n W_n + C_n V_{n-1} - W_n V_{n-1}}{C_n} \\ &= V_{n-1} + \frac{W_n}{C_n} \Big[G_n - V_{n-1} \Big], \end{aligned}$$

brings us to the conventional update format.

Off-policy MC prediction

Initialize for all $s \in \mathcal{S}, a \in \mathcal{A}$:

$Q(s,a) \leftarrow$ arbitrary

$C(s,a) \leftarrow 0$

repeat forever

$b \leftarrow$ any policy with coverage of π

Generate an episode using b:

$$S_0, A_0, R_1, \cdots, S_{T-1} A_{T-1}, R_T, S_T$$

$G \leftarrow 0$

$W \leftarrow 1$

For $t = T-1, T-2, \cdots$ down to 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 break

Off-policy MC control

Initialize for all $s \in \mathcal{S}, a \in \mathcal{A}$:

$$Q(s,a) \leftarrow \text{ arbitrary, } C(s,a) \leftarrow 0, \ \pi(s) \leftarrow \underset{a}{\text{argmax}}\, Q(S_t, a)$$

repeat forever

$b \leftarrow$ any soft policy

Generate an episode using b:

$$S_0, A_0, R_1, \cdots, S_{T-1} A_{T-1}, R_T, S_T$$

$G \leftarrow 0$

$W \leftarrow 1$

For $t = T-1, T-2, \cdots$ down to 0:

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

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$

$\pi(S_t) \leftarrow \underset{a}{\text{argmax}}\, Q(S_t, a)$

If $A_t \neq \pi(S_t)$ then break

$W \leftarrow W(1/b(A_t|S_t))$

Discount-aware IS

- Take into account the internal structure of the return (sum of discounted rewards)
- The full return G_0 of an episode of 100 steps and $\gamma = 0$ is $G_0 = R_1$, but the importance weight is

$$\frac{\pi(A_0|S_0)}{b(A_0|S_0)} \frac{\pi(A_1|S_1)}{b(A_1|S_1)} \cdots \frac{\pi(A_{99}|S_{99})}{b(A_{99}|S_{99})}$$

- Ordinary IS will scale the return by the entire product, which adds up a huge variance to the outcome.
- All factors other than $\frac{\pi(A_0|S_0)}{b(A_0|S_0)}$ are irrelevant, as the return is determined at the first timestep.

Discount-aware IS

Define a **flat partial return** as

$\bar{G}_{t:h} = R_{t+1} + R_{t+2} + \cdots + R_h,\ 0 \leq t < h \leq T$

Flat: No discounting **Partial:** Truncate the episode

Express the full return as function of the partial return

$$
\begin{aligned}
G_t \triangleq & R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T \\
= & (1-\gamma) R_{t+1} + (1-\gamma)\gamma(R_{t+1} + R_{t+2}) \\
& + (1-\gamma)\gamma^2(R_{t+1} + R_{t+2} + R_{t+3}) \\
& \vdots \\
& + (1-\gamma)\gamma^{T-t-2}(R_{t+1} + R_{t+2} + \cdots + R_{T-1}) \\
& + \gamma^{T-t-1}(R_{t+1} + R_{t+2} + \cdots + R_T) \\
& = (1-\gamma) \sum_{h=t+1}^{T-1} \gamma^{h-t-1} \bar{G}_{t:h} + \gamma^{T-t-1} \bar{G}_{t:T}
\end{aligned}
$$

Discount-aware IS

Intuition: Discount is the probability of termination (the degree of partial termination)

Ordinary

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} \bar{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1} \bar{G}_{t:T(t)} \right)}{|\mathcal{T}(s)|}$$

Weighted

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} \bar{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1} \bar{G}_{t:T(t)} \right)}{\left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} \bar{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1} \bar{G}_{t:T(t)} \right)}$$