STATS 701 – Theory of Reinforcement Learning
Monte Carlo Methods

Ambuj Tewari

Associate Professor, Department of Statistics, University of Michigan
tewaria@umich.edu
https://ambujtewari.github.io/stats701-winter2021/

Slide Credits: Prof. M. Vidyasagar @ IIT Hyderabad, India

Winter 2021

Outline

Outline

Outline

Setting

- A MDP with known state space $\mathcal{X}$ and action space $\mathcal{U}$, but unknown state transition matrices A^{u_k}, $u_k \in \mathcal{U}$, and unknown reward function $R : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$.
- A known and fixed policy π. So the MDP becomes a Markov Reward Process.
- Under π, the Markov process has a known set of absorbing states, i.e, states x_i such that

$$\Pr\{X_{t+1} = x_i | X_t = x_i\} = 1,$$

  Note: A^π is still unknown
- Recall: An episode is any sample path that terminates in an absorbing state.
- Assumption: Once the sample path terminates in an absorbing state, the Markov process can be restarted from an arbitrary state.

Approach

Data: A sample path $\{(X_t, U_t, W_t\}_{t \geq 0}$ that terminates in an absorbing state, where X_t = state, U_t = action, and W_t = reward.

Note: $U_t = \pi(X_t)$, where π is known. So U_t does not add any information.

Recall:

$$V(x_i) = E\left[\sum_{t=0}^{\infty} \gamma^t R(X_t, U_t) | X_t = x_i, \pi\right].$$

So if $X_t = x_i$ for some t, then we can estimate $V(x_i)$ via

$$G_t = \sum_{i=0}^{\infty} \gamma^i W_{t+i}.$$

Approach (Cont'd)

By tradition, take $R(x_j) = 0$ if x_j is an absorbing state. So if an episode starts at $t = 0$, passes through the state of interest x_i at time τ, and terminates at time T, then

$$\sum_{i=0}^{T-\tau} \gamma^i W_{\tau+i}$$

gives an estimate for $V(x_i)$. Then we average the estimates over the number of episodes.

First-Time vs. Everytime Estimates

What if a sample path passes through the state of interest x_i more than once?

We can compute

$$\sum_{i=0}^{T-\tau} \gamma^i W_{\tau+i}$$

only the first time during the episode that $X_\tau = x_i$, or every time. These are called the first-time and everytime estimates.

First-Time vs. Everytime Estimates: Example

Suppose $n = 3$, and for convenience label the states as A, B, C. Suppose further that $R(A) = 3$, $R(B) = 2$ and C is an absorbing state for the policy under study, so that $R(C) = 0$. Suppose there are three episodes (all terminating at C):

$$\mathcal{E}_1 = ABABBC, \mathcal{E}_2 = BBC, \mathcal{E}_3 = BAABC.$$

Suppose the discount factor $\gamma = 0.9$. We wish to estimate the value $V(A)$.

First-Time vs. Everytime Estimates: Example (Cont'd)

Recall:

$$\mathcal{E}_1 = ABABBC, \mathcal{E}_2 = BBC, \mathcal{E}_3 = BAABC.$$

The episode $\mathcal{E}_2$ does not interest us because A does not occur in it. So we can form the following quantities:

$$H_{11} = 3 + 2 \cdot (0.9) + 3 \cdot (0.9)^2 + 2 \cdot (0.9)^3 + 2 \cdot (0.9)^4,$$

$$H_{12} = 3 + 2 \cdot (0.9) + 2 \cdot (0.9)^2,$$

$$H_{31} = 3 + 3 \cdot (0.9) + 2 \cdot (0.9)^2, H_{32} = 3 + 2 \cdot (0.9).$$

Then $(H_{11} + H_{31})/2$ is the first-time estimate for $V(A)$, while $(H_{11} + H_{12} + H_{31} + H_{32})/4$ is the everytime estimate for $V(A)$.

A Toy Example

Key Reference: S. P. Singh and R. S. Sutton, "Reinforcement learning with replacing eligibility traces," *Machine Learning*, 22(1–3):123–158, 1996.

A Markov process with just two states, S and A (absorbing), with state transition matrix

	S	A
S	$1-p$	p
A	0	1

So all trajectories starting in S look like $SS\cdots SA$, where there are say l occurrences of S.

Let $\gamma = 1$ (undiscounted case) – permissible with absorbing state Markov processes.

A Toy Example (Cont'd)

- If we define $R(S) = 1, R(A) = 0$, then the first-time estimate for $V(S)$ would be I, the length of the sample path before hitting A.
- The analysis of hitting times shows that the average length of this sample path is $1/p$. So first-time estimate gives an unbiased estimate.
- There are I everytime estimates of V for such a trajectory Their sum is $I(I+1)/2$ and the average is $(I+1)/2$.
- Hence the expected value of the everytime estimate is $((1/p)+1)/2$, which is erroneous by a factor of 2 if p is very small.

Convergence Theorems

- Fact: In a Markov process with absorbing states, each sample path that terminates in an absorbing state is statistically independent of every other such path.
- Consequence: All first-time estimates for $V(x_i)$ can be viewed as statistically independent samples of $V(x_i)$.
- We can use Hoeffding's inequality to estimate the accuracy and confidence of the estimates.

Convergence Theorems (Cont'd)

- First-time estimates of $V(x_i)$ are unbiased estimates and converge to the true value as the number of episodes approaches ∞.
- Everytime estimates of $V(x_i)$ are biased but consistent estimates. Their expected value is not the true value of $V(x_i)$, but they converge to the true value as the number of episodes approaches ∞.
- Everytime estimates can have lower variance for a small number of episodes
- However, eventually first-time estimates will always have lower MSE

Convergence Theorems (Cont'd)

Theorem

Given an absorbing state Markov process and let $x_i \in \mathcal{X}$ be the state of interest. Given a series of episodes that contain x_i, discard the episode before the occurrence of x_i. Form the maximum-likelihood estimate A_{ML} of the state transition matrix based on these "reduced" episodes. Then undiscounted first-visit estimate of $V(x_i)$ is the same as $V(x_i)$ for the Markov process with the state transition matrix A_{ML}.

See Singh & Sutton, Theorem 5, Appendix A1 for the proof.

Summary of Statistical Properties

From Singh & Sutton:

Table 1. Summary of Statistical Results

Algorithm	Convergent	Unbiased	Short MSE	Long MSE	Reduced-ML
First-Visit MC	Yes	Yes	Higher	Lower	Yes
Every-Visit MC	Yes	No	Lower	Higher	No

"We suspect that the first-visit estimate is always the more useful one, even when it is worse in terms of MSE. Our other theoretical results are consistent with this view, but it remains a speculation and a topic for future research."

Shortcomings of Monte Carlo Method

- Number of samples of $V(x_i)$ equals the number of episodes containing x_i, not total length of sample path.
- Because estimates are based on a long sample path, variance in estimates is very high.
- Partial episodes are of no use.

Outline

Monte Carlo Method for Estimating Action-Value Function

- We can also use Monte Carlo methods to estimate the action-value function $Q(x_i, u_k)$ instead of the value function $V(x_i)$.
- This is based on episodes where every pair (x_i, u_k) is visited.
- Assumption that A is ergodic (or primitive) ensures that $A^n > 0$. So every state of interest x_i gets visited in an episode with high probability.
- If policy π is deterministic, then only state-action pairs $(x_i, \pi(x_i))$ occur in the episode. Therefore the use of probabilistic policies is a must.

Target and Behavior Policies

Monte Carlo methods cannot be used to estimate Q_π if $\pi \in \Pi_d$, or if $\pi(u_k|x_i) = 0$ for some pairs (x_i, u_k).

Remedy: Choose a probabilistic policy $\phi \in \Pi_p$ such that $\phi(u_k|x_i) > 0$ for all pairs (x_i, u_k), and use it to generate a sample path $\{(X_t, U_t, W_t)\}_{t=0}^{T}$.

Challenge: How can this sample path be used to estimate $Q_\pi(x_i, u_k)$?

Reminder: π is called the "target" policy and ϕ is called the "behavior" policy. Note that this is "off-policy" estimation.

Importance Sampling: Preliminaries

Given a sample path, for each policy π, there is an associated likelihood

$$\Pr\{U_t, X_{t+1}, U_{t+1}, \cdots, X_T | X_t, U_t^{T-1} \sim \pi\},$$

where $U_t^{T-1} \sim \pi$ means that $\Pr\{U_\tau | X_\tau\}$ has the distribution $\pi(X_\tau)$, for $t \leq \tau \leq T-1$.

This quantity can be expressed as

$$\mathcal{P}_\pi = \prod_{\tau=t}^{T-1} \Pr\{X_{\tau+1} | X_\tau, U_\tau\} \pi(U_\tau | X_\tau).$$

Unfortunately, we don't know the transition probabilities $\Pr\{X_{\tau+1} | X_\tau, U_\tau\}$.

Importance Sampling: Preliminaries (Cont'd)

However, because the sample path is generated using the policy ϕ, what we can actually measure is

$$\mathcal{P}_\phi = \prod_{\tau=t}^{T-1} \Pr\{X_{\tau+1}|X_\tau, U_\tau\}\phi(U_\tau|X_\tau).$$

Now note that

$$\rho_{[t,T-1]} := \frac{\mathcal{P}_\pi}{\mathcal{P}_\phi} = \prod_{\tau=t}^{T-1} \frac{\pi(U_\tau|X_\tau)}{\phi(U_\tau|X_\tau)}.$$

The unknown transition probabilities $\Pr\{ X_{\tau+1}|X_\tau, U_\tau\}$ simply cancel out. So $\rho_{[t,T-1]}$ can be computed because π, ϕ are known policies, and U_τ, X_τ can be observed.

Consequence of This Formula

Recall:

$$\rho_{[t,T-1]} := \frac{\mathcal{P}_\pi}{\mathcal{P}_\phi} = \prod_{\tau=t}^{T-1} \frac{\pi(U_\tau|X_\tau)}{\phi(U_\tau|X_\tau)}.$$

Consequence: Given a sample path $\{(X_t, U_t, W_t\}$, if we know its likelihood according to ϕ, we can compute its likelihood according to π, without knowing the dynamics of the MDP.

Given the return G_t per episode (which as per ϕ), it can be multiplied by $\rho_{[t,T-1]}$ to get an estimate of the return as per π.

Ordinary Importance Sampling Estimate

If we sample according to π, we estimate $\hat{V}_\pi(x_i)$ via

$$\hat{V}_\pi(x_i) = \frac{1}{|J(x_i)|} \sum_{t \in J(x_i)} G_t.$$

where G_t is the discounted reward.

The estimate

$$\hat{V}_\pi(x_i) = \frac{1}{|J(x_i)|} \sum_{t \in J(x_i)} \rho_{[t,T-1]} G_t$$

is called the "ordinary importance sampling" estimate. The factor $\rho_{[t,T-1]}$ compensates for off-policy sampling.

Weighted Importance Sampling Estimate

The estimate

$$\hat{V}_\pi(x_i) = \frac{\sum_{t \in J(x_i)} \rho_{[t,T-1]} G_t}{\sum_{t \in J(x_i)} \rho_{[t,T-1]}}$$

is called the "weighted importance sampling" estimate.

Replacing $\rho_{[t,T-1]}$ by one leads to the ordinary estimate.

The ordinary importance sampling estimate is unbiased but can have very large variance.

The weighted importance sampling estimate is biased but consistent. It also has lower variance than the ordinary estimate.

Outline

Greedy Policy Optimization: Motivation

Suppose π is a given policy and that we have an estimate for $Q_\pi(x_i, u_k)$ for each state-action pair (x_i, u_k), where

$$Q_\pi(x_i, u_k) = E[R(X_t, U_t) + \gamma V_\pi(X_{t+1}) | X_t = x_i, U_t = u_k].$$

What if we choose the "next" policy to improve Q_π for each state-action pair?

This is called "greedy policy optimization." But does it work?

Policy Improvement Theorem

Theorem

(Policy Improvement Theorem) Suppose $\pi, \phi \in \Pi_p$, *and moreover*

$$Q_\pi(x_i, \phi(x_i)) \geq Q_\pi(x_i, \pi(x_i)) = V_\pi(x_i), \ \forall x_i \in \mathcal{X}. \tag{1}$$

Then

$$V_\phi(x_i) \geq V_\pi(x_i), \ \forall x_i \in \mathcal{X}. \tag{2}$$

Policy Improvement Theorem: Corollary

Theorem

Moreover, suppose there is a state $x_i \in \mathcal{X}$ such that (1) holds with strict inequality, that is

$$Q_\pi(x_i, \phi(x_i)) > Q_\pi(x_i, \pi(x_i)).$$

Then there is a state $x_j \in \mathcal{X}$ such that (2) holds with strict inequality, that is,

$$V_\phi(x_j) > V_\pi(x_j).$$

Greedy Policy Optimization: Formal Statement

Start with an initial policy and corresponding action-value function Q_π. Define

$$k^* := \arg\max_{u_k \in \mathcal{U}} Q_\pi(x_i, u_k), \phi(x_i) = u_{k^*}, \ \forall x_i \in \mathcal{X}.$$

Then it follows from the Policy Improvement theorem that $V_\phi(x_i) \geq V_\pi(x_i)$ for all $x_i \in \mathbf{x}$. Repeat.

If $Q_\pi(x_i, \pi(x_i))$ cannot be improved for any $x_i \in \mathcal{X}$, then π is optimal.

This may be called a "pure" greedy policy.

ϵ-Greedy Policies: Combining Exploration with Exploitation

Notation: If $\pi \in \Pi_p$, then

$$\pi(u_k|x_i) := \Pr\{U_t = u_k | X_t = x_i\}.$$

A policy $\pi \in \Pi_p$ is said to be ϵ-soft if

$$\pi(u_k|x_i) \geq \epsilon, \ \forall u_k \in \mathcal{U}, x_i \in \mathcal{X}.$$

Suppose $\pi \in \Pi_p$ is the current ϵ-soft policy. Can we can generate an updated ϵ-soft policy ϕ that is better?

ϵ-Greedy Updating

Suppose $\pi \in \Pi_p$ is the current ϵ-soft policy. We can generate an updated ϵ-soft policy ϕ as follows: Define a deterministic policy $\psi \in \Pi_d$ by

$$k^* := \arg\max_{u_k \in \mathcal{U}} Q_\pi(x_i, u_k), \psi(x_i) = u_{k^*}, \ \forall x_i \in \mathcal{X}.$$

Now define the ϵ-soft policy $\phi \in \Pi_p$ by

$$\phi(u_k|x_i) = \begin{cases} \frac{\epsilon}{|\mathcal{U}|} & k \neq k^* \\ \frac{\epsilon}{|\mathcal{U}|} + (1-\epsilon) & k = k^*. \end{cases}$$

Properties of ϵ-Greedy Updating

Theorem

With ϕ defined as above, we have that

$$Q_\pi(x_i, \phi(x_i)) \geq Q_\pi(x_i, \pi(x_i)) = V_\pi(x_i), \ \forall x_i \in \mathcal{X}.$$

The import of the above Theorem is that the above ϵ-greedy updating rule will eventually converge to the optimal ϵ-soft policy.

The ϵ-greedy policy and updating rule can be combined with a "schedule" for reducing ϵ to zero, which would presumably converge to the optimal policy.