When does TD have lower variance than MC?

Sam Greydanus
sgrey@google.com

September 11, 2018

Abstract

The goal of this writeup is to clarify the connection between Monte Carlo (MC) and Temporal Difference (TD) value estimation. I will also show that TD value estimates always have equal or lower variance than MC estimates.

1 Setup

Monte Carlo (MC). Monte Carlo methods estimate the value of a state s_t by computing the average return $G(s_t)$ at that state. A simple every-visit MC update would look like Equation 1 where $\alpha = \frac{1}{k}$ and k is the visit count of s_t.

$$V(s_t) \leftarrow V(s_t) + \alpha[G(s_t) - V(s_t)] \tag{1}$$

Temporal Differences (TD). Another way to estimate the value of a state is with *temporal difference* (TD) learning. The idea of TD learning is to express the value of s_t in terms of the values of its successor state s_{t+1} along a trajectory, plus the change in reward due to the $s_t \rightarrow s_{t+1}$ transition.

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \tag{2}$$

This approach is often called *bootstrapping* because each state uses the next state's value to update its own.

Implementation notes. Looking at Equations 1 and 2, we can see that the core difference between MC and TD updates is the substitution $G(s_t) = r_{t+1} + \gamma V(s_{t+1})$. Unfortunately, the classic algorithms for MC and TD value estimation are very different. In the textbook, *Reinforcement Learning: An Introduction* [1], the MC and TD algorithms are given on pages 92 and 120 respectively. For MC learning, the authors compute $G_0 ... G_N$ across each episode and average across episodes. For TD learning, the authors update $V(s_t)$ after each state transition in an episode.

In spite of these textbook differences, most modern implementations of MC and TD are strikingly similar. The value updates happen at the end of each episode and proceed in reverse order from the last episode in the trajectory to the first. We choose to follow this format (and thus depart slightly from textbook definitions) for two reasons. First, doing so enables us to make tighter analogies between MC and TD. Second, most real-world applications of RL algorithms use this format.

Pseudocode for our versions of MC and TD value estimation can be found in Figure 1.

Initialize:

$\pi \leftarrow$ policy to be evaluated

$V_{MC}, V_{TD} \leftarrow$ initialize all state-value pairs to zero

Repeat forever:

Generate an episode of N steps using π

For s_t in *Reversed*$(s_1 ... s_{N-1})$:

$G(s_t) \leftarrow$ return following the first occurrence of s_t

$V_{MC}(s_t) \leftarrow V_{MC}(s_t) + \alpha[G(s_t) - V_{MC}(s_t)]$

$V_{TD}(s_t) \leftarrow V_{TD}(s_t) + \alpha[r_{t+1} + \gamma V_{TD}(s_{t+1}) - V_{TD}(s_t)]$

Figure 1: Estimating value using Monte Carlo and Temporal Difference methods.

2 Proof

In this section, we will show why TD value estimates always have equal or lower variance than MC estimates. Note that this proof only works for tabular agents/environments; when state approximators are introduced, all bets are off.

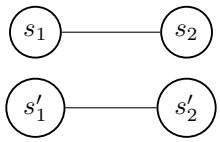

Figure 2: Adjacent states along two different trajectories.

Consider two adjacent states s_1 and s_2 along trajectory T, and two adjacent states s'_1 and s'_2 along trajectory T' (see Figure 2). We intend to perform a value update using first T and then T'.

Now, consider all cases for which Equation 3 holds:

$$V_{TD}(s_2) = G(s_2) \quad \text{and} \quad V_{TD}(s'_2) = G(s'_2) \qquad \text{assumption of Lemma 1} \tag{3}$$

Lemma 1. *Wherever Equation 3 is true, TD and MC updates are identical.* The way to see this is by expressing the TD updates along trajectory T' using Equation 2:

$$V_{TD}(s'_1) = V_{TD}(s_1) + \alpha(r_{12'} + \gamma V_{TD}(s'_2) - V_{TD}(s_1)) \qquad \text{TD update rule} \tag{4}$$

The next step is to recognize that the MC and TD value estimates after the first value update - the one along T - are identical:

$$
\begin{aligned}
V_{TD}(s_1) =& 0 + (r_{12} + \gamma V_{TD}(s_2) - 0) && \text{initial value } V(s_1) = 0 && (5)\\
=& r_{12} + \gamma G(s_2) = G(s_1) && \text{definition of return} && (6)\\
=& V_{MC}(s_1) && \text{definition of first MC update} && (7)
\end{aligned}
$$

Substituting Equation 7 into Equation 4, we have

$$
\begin{aligned}
V_{TD}(s_1') =& V_{MC}(s_1) + \alpha(r_{12'} + \gamma V_{TD}(s_2') - V_{MC}(s_1)) && (8)\\
=& V_{MC}(s_1) + \alpha(r_{12'} + \gamma G(s_2') - V_{MC}(s_1)) && (9)\\
=& V_{MC}(s_1) + \alpha(G(s_1') - V_{MC}(s_1)) && (10)\\
=& V_{MC}(s_1') && (11)
\end{aligned}
$$

Equations 8-11 tell us that MC and TD updates along T and T' are identical whenever Equation 3 is true.

Lemma 2. *When $s_1 \neq s_1'$ and $s_2 = s_2'$, the TD update has lower variance than the MC update.* To see this, we must first recall that value estimates are simply averages. The variance of these averages is proportional to $\frac{\sigma^2}{n}$ where n is the number of items being averaged over. Next, we need to introduce a count function, $N(s)$ which measures the number of times the agent has visited state s. In the case we are considering,

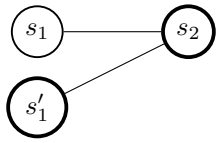

Figure 3: In Lemma 2, we consider the case where s_2 and s_2' are the same state but s_1 and s_1' are not.

$$
\begin{aligned}
& N(s_1) = N(s_1' = 1 \quad \text{and} \quad N(s_2) = N(s_2' = 2 && (12)\\
& Var[G(s_1)] = Var[G(s_1')] = \sigma^2 && (13)\\
& Var[G(s_2)] = Var[G(s_2')] = \frac{\sigma^2}{2} && (14)
\end{aligned}
$$

Now recall that the difference between a MC update and a TD update is given by the substitution $G(s_t) = r_{t+1} + \gamma V(s_{t+1})$. Let's compare the variances of these quantities at state s_1

$$
\begin{aligned}
Var[G(s_1)] \overset{?}{>} & Var[r_{12} + \gamma V_{TD}(s_2)] && \text{variance of both terms} && (15)\\
\overset{?}{>} & Var[r_{12}] + Var[\gamma V_{TD}(s_2)] && \text{assume independence} && (16)\\
> & 0 + \gamma Var[G(s_2)] && \text{assume deterministic rewards and } \gamma < 1 && (17)
\end{aligned}
$$

Proof. Whenever s_2 or s_2' is a terminal node, the assumption in Equation 3 is true. By induction, this assumption holds for each preceding state, except for the case where $s_1 \neq s_1'$ and $s_2 = s_2'$.

In this case, shown in Figure 2, $V_{TD}(s_2')$ gets updated during the first value update - the one along T - but this value update never gets reflected in the return $G(s_1')$. Thus $V_{TD}(s_1')$ and $V_{MC}(s_1')$ go out of sync. But we've shown in Lemma 2 that TD updates always have lower variance than MC updates in this case. This means that, regardless of how trajectories T and T' overlap, TD updates will always have equal or lower variance than MC updates. This argument extends to an arbitrary number of trajectories and updates.

References

[1] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 2 edition, 2018.