TEMPORAL DIFFERENCE LEARNING

MONTE CARLO SUMMARY

REVIEW

Needed a way to estimate q_π and v_π and improve the policy without knowing p

Monte Carlo Methods:

- Wait till the end of an episode and use G_t as prediction targets

Benefits:

- Unbiased: given an infinite amount of data, we can recover q_π, or v_π

Drawbacks:

- Only works for episodic MDPs
- High variance target: Var(G_t) can have high variance because it sums up many random variables
- Offline only (cannot update every step)

AN ONLINE APPROACH

MOTIVATION FOR TEMPORAL DIFFERENCE LEARNING

$$V_{n+1}(S_t) = V_n(S_t) + \alpha_n \left(G_t - V_n(S_t) \right)$$

$$G_t = R_{t+1} + \underbrace{\gamma R_{t+2} + \gamma^2 R_{t+3} + \dots}_{=\gamma G_{t+1}}$$

Know after one step

Have to wait for the future to know these

We want to update every step

AN ONLINE APPROACH

TEMPORAL DIFFERENCE LEARNING FOR V

$$V_{n+1}\left(S_t\right) = V_n\left(S_t\right) + \alpha_n\left(G_t - V_n\left(S_t\right)\right)$$

AN ONLINE APPROACH

TEMPORAL DIFFERENCE LEARNING FOR V

$$
\begin{aligned}
V_{n+1}\left(S_t\right) &= V_n\left(S_t\right) + \alpha_n\left(G_t - V_n\left(S_t\right)\right) \\
&= V_n\left(S_t\right) + \alpha_n\left(R_{t+1} + \gamma G_{t+1} - V_n\left(S_t\right)\right)
\end{aligned}
$$

AN ONLINE APPROACH

TEMPORAL DIFFERENCE LEARNING FOR V

$$V_{n+1}(S_t) = V_n(S_t) + \alpha_n \left(G_t - V_n(S_t) \right)$$

$$= V_n(S_t) + \alpha_n \left(R_{t+1} + \gamma G_{t+1} - V_n(S_t) \right)$$

Replace with a prediction of G_{t+1}

AN ONLINE APPROACH

TEMPORAL DIFFERENCE LEARNING FOR V

$$V_{n+1}(S_t) = V_n(S_t) + \alpha_n \left(G_t - V_n(S_t) \right)$$

$$= V_n(S_t) + \alpha_n \left(R_{t+1} + \gamma G_{t+1} - V_n(S_t) \right)$$

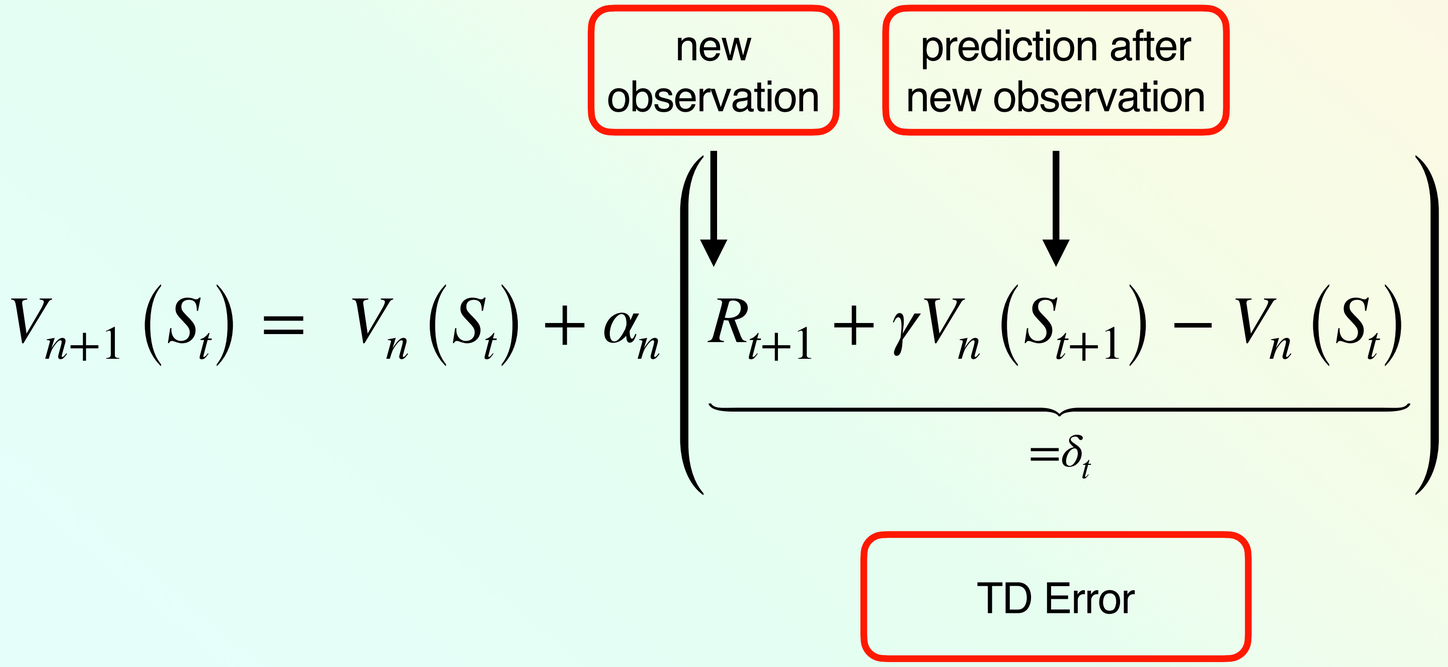
Replace with a prediction of G_{t+1}

$$= V_n(S_t) + \alpha_n \left(R_{t+1} + \gamma V_n(S_{t+1}) - V_n(S_t) \right)$$

Prediction of $v_\pi(S_{t+1})$

TEMPORAL DIFFERENCE LEARNING

THE ANATOMY



TD(0) ALGORITHM

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until S is terminal

EXAMPLE

Example 6.1: Driving Home Each day as you drive home from work, you try to predict how long it will take to get home. When you leave your office, you note the time, the day of week, the weather, and anything else that might be relevant. Say on this Friday you are leaving at exactly 6 o'clock, and you estimate that it will take 30 minutes to get home. As you reach your car it is 6:05, and you notice it is starting to rain. Traffic is often slower in the rain, so you reestimate that it will take 35 minutes from then, or a total of 40 minutes. Fifteen minutes later you have completed the highway portion of your journey in good time. As you exit onto a secondary road you cut your estimate of total travel time to 35 minutes. Unfortunately, at this point you get stuck behind a slow truck, and the road is too narrow to pass. You end up having to follow the truck until you turn onto the side street where you live at 6:40. Three minutes later you are home.

EXAMPLE

The sequence of states, times, and predictions is thus as follows:

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

EXAMPLE

The sequence of states, times, and predictions is thus as follows:

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

$$\sum_{k=0}^{t} R_{k+1}$$

$$V_n(S_t)$$

$$\sum_{k=0}^{t-1} R_{k+1} + V_n(S_t)$$

EXAMPLE

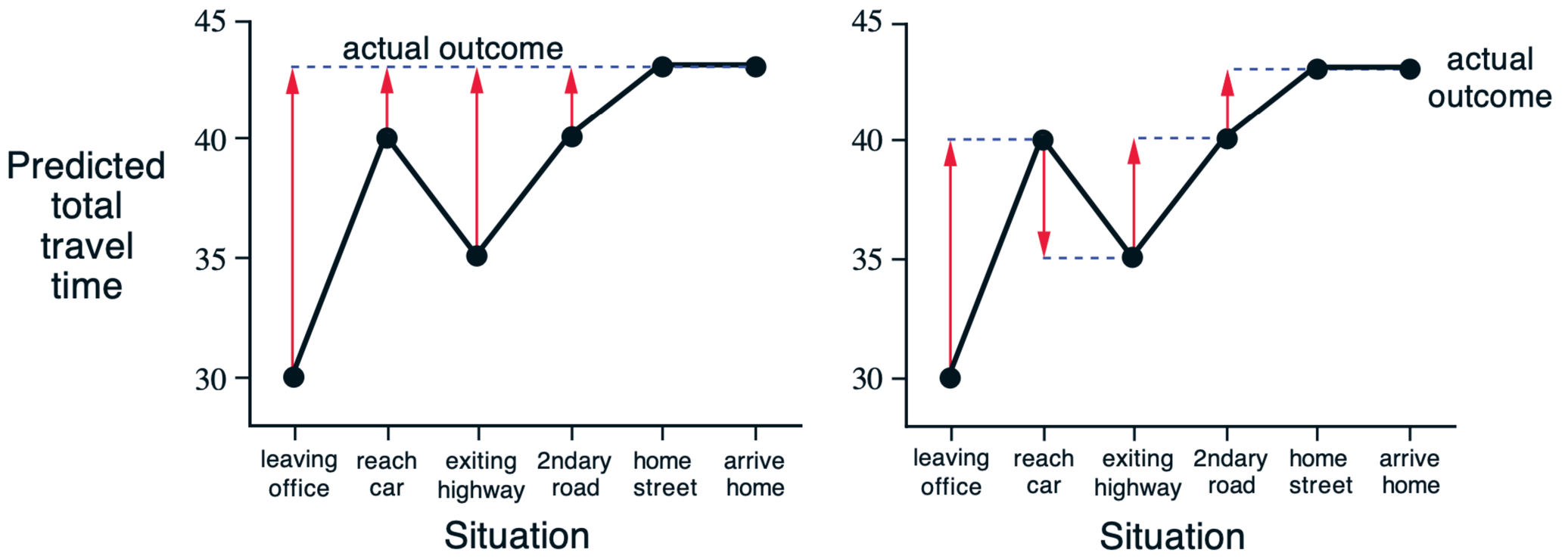


Figure 6.1: Changes recommended in the driving home example by Monte Carlo methods (left) and TD methods (right).

TEMPORAL DIFFERENCE LEARNING

VARIANTS

$$V_{n+1}\left(S_t\right) = V_n\left(S_t\right) + \alpha_n\left(R_{t+1} + \gamma V_n\left(S_{t+1}\right) - V_n\left(S_t\right)\right)$$

$$Q_{n+1}\left(S_t, A_t\right) = Q_n\left(S_t, A_t\right) + \alpha_n\left(R_{t+1} + \gamma Q_n\left(S_{t+1}, A_{t+1}\right) - Q_n\left(S_t, A_t\right)\right)$$

TEMPORAL DIFFERENCE LEARNING

VARIANTS

$$V_{n+1}\left(S_t\right) = V_n\left(S_t\right) + \alpha_n\left(R_{t+1} + \gamma V_n\left(S_{t+1}\right) - V_n\left(S_t\right)\right)$$

$$Q_{n+1}\left(S_t, A_t\right) = Q_n\left(S_t, A_t\right) + \alpha_n\left(R_{t+1} + \gamma Q_n\left(S_{t+1}, A_{t+1}\right) - Q_n\left(S_t, A_t\right)\right)$$

$$Q_{n+1}\left(S_t, A_t\right) = Q_n\left(S_t, A_t\right) + \alpha_n\left(R_{t+1} + \gamma \mathbb{E}\left[Q_n\left(S_{t+1}, A_{t+1}\right) \mid S_{t+1}\right] - Q_n\left(S_t, A_t\right)\right)$$

TEMPORAL DIFFERENCE LEARNING

VARIANTS

$$V_{n+1}\left(S_t\right) = V_n\left(S_t\right) + \alpha_n\left(R_{t+1} + \gamma V_n\left(S_{t+1}\right) - V_n\left(S_t\right)\right)$$

$$Q_{n+1}\left(S_t, A_t\right) = Q_n\left(S_t, A_t\right) + \alpha_n\left(R_{t+1} + \gamma Q_n\left(S_{t+1}, A_{t+1}\right) - Q_n\left(S_t, A_t\right)\right)$$

$$Q_{n+1}\left(S_t, A_t\right) = Q_n\left(S_t, A_t\right) + \alpha_n\left(R_{t+1} + \gamma \mathbb{E}\left[Q_n\left(S_{t+1}, A_{t+1}\right) \mid S_{t+1}\right] - Q_n\left(S_t, A_t\right)\right)$$

$$Q_{n+1}\left(S_t, A_t\right) = Q_n\left(S_t, A_t\right) + \alpha_n\left(R_{t+1} + \gamma V_n\left(S_{t+1}\right) - Q_n\left(S_t, A_t\right)\right)$$

CONNECTION TO DYNAMIC PROGRAMMING

EXPECT TD TARGET

$$\mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s\right] = \sum_a \Pr\left(A_t = a \,|\, S_t = s\right) \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s, A_t = a\right]$$

CONNECTION TO DYNAMIC PROGRAMMING

EXPECT TD TARGET

$$\mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s\right] = \sum_a \Pr\left(A_t = a \,|\, S_t = s\right) \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s, A_t = a\right]$$

$$= \sum_a \pi(a \,|\, s) \sum_{s'} \Pr(S_{t+1} = s' \,|\, S_t = s, A_t = a) \mathbb{E}\left[R_{t+1} + \gamma V_n(s') \middle| S_t = s, A_t = a, S_{t+1} = s'\right]$$

CONNECTION TO DYNAMIC PROGRAMMING

EXPECT TD TARGET

$$
\begin{aligned}
\mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s\right] &= \sum_a \Pr\left(A_t = a \,|\, S_t = s\right)\right) \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s, A_t = a\right] \\
&= \sum_a \pi(a \,|\, s) \sum_{s'} \Pr(S_{t+1} = s' \,|\, S_t = s, A_t = a) \mathbb{E}\left[R_{t+1} + \gamma V_n(s') \middle| S_t = s, A_t = a, S_{t+1} = s'\right] \\
&= \sum_a \pi(a \,|\, s) \sum_{s'} p(s' \,|\, s, a)\Big(r(s, a, s') + \gamma V_n(s')\Big)
\end{aligned}
$$

CONNECTION TO DYNAMIC PROGRAMMING

EXPECT TD TARGET

$$
\begin{aligned}
\mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s\right] &= \sum_a \Pr\left(A_t = a \,|\, S_t = s\right) \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \middle| S_t = s, A_t = a\right] \\
&= \sum_a \pi(a \,|\, s) \sum_{s'} \Pr(S_{t+1} = s' \,|\, S_t = s, A_t = a) \mathbb{E}\left[R_{t+1} + \gamma V_n(s') \middle| S_t = s, A_t = a, S_{t+1} = s'\right] \\
&= \sum_a \pi(a \,|\, s) \sum_{s'} p(s' \,|\, s, a)\Big(r(s, a, s') + \gamma V_n(s')\Big)
\end{aligned}
$$

Same target as the Bellman operator!

TD VS MC

BIAS — MC

$$V_{n+1}(s) = V_n(s) + \alpha_n \mathbb{E}\left[G_t - V_n(s) \middle| S_t = s\right]$$

TD VS MC

BIAS — MC

$$
\begin{aligned}
V_{n+1}(s) = & \; V_n(s) + \alpha_n \mathbb{E}\left[G_t - V_n(s) \middle| S_t = s \right] \\
= & \; V_n(s) + \alpha_n \left(\mathbb{E}\left[G_t \middle| S_t = s \right] - V_n(s) \right)
\end{aligned}
$$

TD VS MC

BIAS — MC

$$
\begin{aligned}
V_{n+1}(s) = \; & V_n(s) + \alpha_n \mathbb{E}\left[G_t - V_n(s) \middle| S_t = s \right] \\
= \; & V_n(s) + \alpha_n \left(\mathbb{E}\left[G_t \middle| S_t = s \right] - V_n(s) \right) \\
= \; & V_n(s) + \alpha_n \left(v_\pi(s) - V_n(s) \right)
\end{aligned}
$$

TD VS MC

BIAS — MC

$$
\begin{aligned}
V_{n+1}(s) = \; & V_n(s) + \alpha_n \mathbb{E}\left[G_t - V_n(s) \,\middle|\, S_t = s \right] \\
= \; & V_n(s) + \alpha_n \left(\mathbb{E}\left[G_t \,|\, S_t = s \right] - V_n(s) \right) \\
= \; & V_n(s) + \alpha_n \left(v_\pi(s) - V_n(s) \right)
\end{aligned}
$$

Unbiased — update moves towards the correct value in expectation

TD VS MC

BIAS — TD

$$V_{n+1}(s) = V_n(s) + \alpha_n \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) - V_n(s) \middle| S_t = s\right]$$

TD VS MC

BIAS — TD

$$
\begin{aligned}
V_{n+1}(s) = \; & V_n(s) + \alpha_n \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) - V_n(s) \,\middle|\, S_t = s\right] \\
= \; & V_n(s) + \alpha_n \left(\mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \,|\, S_t = s\right] - V_n(s)\right)
\end{aligned}
$$

TD VS MC

BIAS — TD

$$
\begin{aligned}
V_{n+1}(s) = \; & V_n(s) + \alpha_n \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) - V_n(s) \,\middle|\, S_t = s \right] \\
= \; & V_n(s) + \alpha_n \left(\mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \,|\, S_t = s \right] - V_n(s) \right) \\
= \; & V_n(s) + \alpha_n \left(\sum_a \pi(a \,|\, s) \sum_{s'} p(s' \,|\, s, a) \Big(r(s, a, s') + \gamma V_n(s') \Big) - V_n(s) \right)
\end{aligned}
$$

TD VS MC

BIAS — TD

$$
\begin{aligned}
V_{n+1}(s) &= V_n(s) + \alpha_n \mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) - V_n(s) \,\middle|\, S_t = s\right] \\
&= V_n(s) + \alpha_n \left(\mathbb{E}\left[R_{t+1} + \gamma V_n\left(S_{t+1}\right) \,|\, S_t = s\right] - V_n(s)\right) \\
&= V_n(s) + \alpha_n \left(\sum_a \pi(a \,|\, s) \sum_{s'} p(s' \,|\, s, a)\Big(r(s, a, s') + \gamma V_n(s')\Big) - V_n(s) \right)
\end{aligned}
$$

Biased —update does not move towards v_π in expectation.

Unbiased if $V_n = v_\pi$

The expected update is the Bellman operator, so updates still move toward v_π as $n \to \infty$ if $\alpha_n = 1/n$

TD VS MC

VARIANCE OF MC

$$Z = X + Y$$

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

TD VS MC

VARIANCE OF MC

$$\begin{aligned}\text{Var}(G_t) &= \text{Var}(R_{t+1} + \gamma G_{t+1}) \\ &= \text{Var}(R_{t+1}) + \text{Var}(\gamma G_{t+1}) + 2Cov(R_{t+1}, \gamma G_{t+1})\end{aligned}$$

TD VS MC

VARIANCE OF MC

$$\begin{aligned}\text{Var}(G_t) &= \text{Var}(R_{t+1} + \gamma G_{t+1}) \\ &= \text{Var}(R_{t+1}) + \text{Var}(\gamma G_{t+1}) + 2Cov(R_{t+1}, \gamma G_{t+1})\end{aligned}$$

$$\text{Var}\left(R_{t+1} + \gamma V_n(S_{t+1})\right) = \text{Var}(R_{t+1}) + \text{Var}\left(\gamma V_n(S_{t+1})\right) + 2Cov\left(R_{t+1}, \gamma V_n(S_{t+1})\right)$$

TD VS MC

VARIANCE OF MC

$$\begin{aligned}\text{Var}(G_t) &= \text{Var}(R_{t+1} + \gamma G_{t+1}) \\ &= \text{Var}(R_{t+1}) + \text{Var}(\gamma G_{t+1}) + 2Cov(R_{t+1}, \gamma G_{t+1})\end{aligned}$$

$$\text{Var}\left(R_{t+1} + \gamma V_n(S_{t+1})\right) = \text{Var}(R_{t+1}) + \text{Var}\left(\gamma V_n(S_{t+1})\right) + 2Cov\left(R_{t+1}, \gamma V_n(S_{t+1})\right)$$

$$\text{Var}(G_t) - \text{Var}\left(R_{t+1} + \gamma V_n(S_{t+1})\right) = \text{Var}(\gamma G_{t+1}) + 2\text{Cov}(R_{t+1}, \gamma G_{t+1}) - \text{Var}\left(\gamma V_n(S_{t+1})\right) - 2\text{Cov}\left(R_{t+1}, \gamma V_n(S_{t+1})\right)$$

Usually $\text{Var}(\gamma G_{t+1}) > \text{Var}\left(\gamma V_n(S_{t+1})\right)$
So MC usually higher variance than TD

BATCH TD

METHOD

Collect a data set of s, r, s' transitions.

Average the update for each unique state s in the dataset

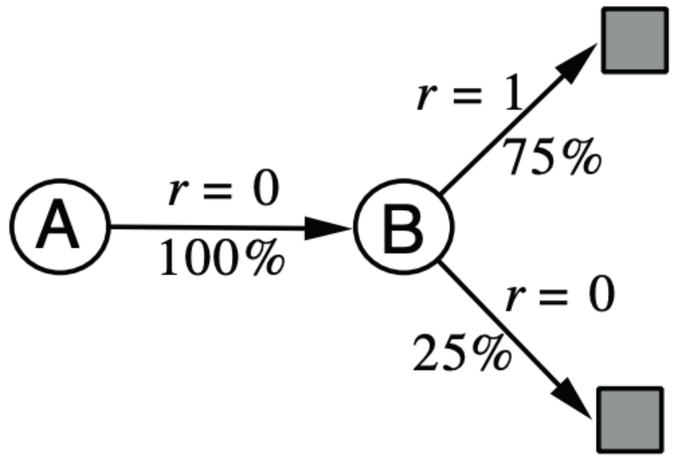
$$V_{n+1}(s) = V_n + \alpha \left(\frac{1}{k} \sum_{i=1}^{k} r_i + \gamma V_n(s_i') - v(s) \right)$$

Repeat for all states until convergence, i.e., $V_{n+1} = V_n$

EXAMPLE

Example 6.4: You are the Predictor Place yourself now in the role of the predictor of returns for an unknown Markov reward process. Suppose you observe the following eight episodes:

A, 0, B, 0	B, 1
B, 1	B, 1
B, 1	B, 1
B, 1	B, 0



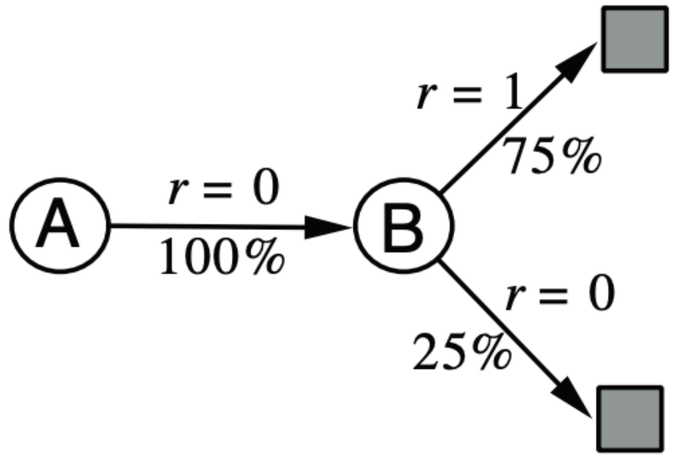
For MC and Batch TD what are

$V(A) = ?$
$V(B) = ?$

EXAMPLE

Example 6.4: You are the Predictor Place yourself now in the role of the predictor of returns for an unknown Markov reward process. Suppose you observe the following eight episodes:

A, 0, B, 0	B, 1
B, 1	B, 1
B, 1	B, 1
B, 1	B, 0



For MC and Batch TD what are

$V(A) = \underbrace{3/4}_{TD}$, or $\underbrace{0}_{MC}$

$V(B) = 3/4$

NEXT CLASS

WHAT YOU SHOULD DO

1. Quiz Due Tonight

Friday: More Temporal Difference Learning for Prediction