Reinforcement Learning

Temporal Difference Algorithms

Lecturer: *Nicolò Cesa-Bianchi* version February 9, 2023

We consider the discounted infinite horizon criterion and focus on MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$ such that $\mathcal{A}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$, transition kernel $\{p(\cdot \mid s, a) \,:\, s \in \mathcal{S},\, a \in \mathcal{A}\}$, and time-independent reward function $r : \mathcal{S} \times \mathcal{A} \to [-1, 1]$.

Fix a stationary deterministic policy π and consider the problem of estimating the state-value function V^π. Recall the system of linear equations that V^π satisfies,

$$V^\pi(s) = r\big(s, \pi(s)\big) + \gamma\, \mathbb{E}\big[V^\pi(s') \mid s\big] \qquad s \in \mathcal{S}$$

where $s' \sim p\big(\cdot \mid s, \pi(s)\big)$. Now, similarly to what we did in Q-learning, we can obtain a sequence $V_0, V_1, \ldots$ of approximations to V^π by running gradient descent on the square loss

$$\ell_t(V_t) = \frac{1}{2}\Big(V_t(s_t) - r\big(s_t, \pi(s_t)\big) - \gamma\, \mathbb{E}\big[V_t(s') \mid s\big]\Big)^2$$

which amounts to the update

$$V_{t+1}(s_t) = (1 - \eta_t) V_t(s_t) + \eta_t \Big(r\big(s_t, \pi(s_t)\big) + \gamma\, \mathbb{E}\big[V_t(s') \mid s_t\big]\Big)$$

Since, however, $\mathbb{E}\big[V_t(s') \mid s\big]$ is not directly accessible, we run gradient descent on a perturbed gradient,

$$V_{t+1}(s_t) = (1 - \eta_t) V_t(s_t) + \eta_t \Big(r\big(s_t, \pi(s_t)\big) + \gamma\, V_t(s_{t+1})\Big)$$

where $s_{t+1} \sim p\big(\cdot \mid s_t, \pi(s_t)\big)$. We call **temporal difference** the quantity

$$\Delta_t = r\big(s_t, \pi(s_t)\big) + \gamma\, V_t(s_{t+1}) - V_t(s_t)$$

and write the above update equivalently as

$$V_{t+1}(s_t) = V_t(s_t) + \eta_t \Delta_t$$

The algorithm based on this update is known as TD(0). Similarly to what we did for Q-learning, we can prove the convergence of TD(0) when η_t is a function $\eta_t : \mathcal{S} \to [0, 1]$ of the states defined by

$$\eta_t(s) = \frac{\mathbb{I}\{s = s_t\}}{N_t(s)} \qquad \text{where} \qquad N_t(s) = \sum_{\tau=0}^{t} \mathbb{I}\{s_\tau = s\}$$

Because we focus on deterministic policies, the learning rate η_t can depend only on states rather than on state-action pairs.

Theorem 1 *Assume that* TD(0) *is run with stationary deterministic policy* π *such that, for all* $s \in \mathcal{S}$,

$$\sum_{t\geq 0} \eta_t(s) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty \qquad \textit{and} \qquad \sum_{t\geq 0} \eta_t(s)^2 = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

Then

$$\lim_{t\to\infty} V_t(s) = V^{\pi}(s) \qquad s \in \mathcal{S}$$

with probability 1.

The conditions on η are satisfied when the Markov chain induced by π on the MDP is **irreducible**. That is, for any two distinct states $s, s' \in \mathcal{S}$ there is an integer n such that the probability of going from s to s' in n steps is positive.

The update of TD(0) is based on a 1-step lookahead $R_t^{(1)}(s_t) = r(s_t, a_t) + \gamma\, V_t(s_{t+1})$ so that $\Delta_t = R_t^{(1)}(s_t) - V_t(s_t)$. Using the identity

$$V^{\pi}(s) = r\big(s, \pi(s)\big) + \mathbb{E}\left[\sum_{\tau=1}^{\infty} \gamma^{\tau} r\big(s_\tau, \pi(s_\tau)\big) \,\middle|\, s\right] \qquad s \in \mathcal{S}$$

where $s_\tau \sim p\big(\cdot \mid s_{\tau-1}, \pi(s_{\tau-1})\big)$ and $s_0 = s$, TD(0) can be easily generalized to a n-step lookahead

$$R_t^{(n)}(s_t) = \sum_{\tau=0}^{n-1} \gamma^{\tau} r\big(s_{t+\tau}, \pi(s_{t+\tau})\big) + \gamma^n\, V_t(s_{t+n})$$

The corresponding updates are $V_{t+1}(s_t) = V_t(s_t) + \eta_t \Delta_t^{(n)}$ where $\Delta_t^{(n)} = R_t^{(n)}(s_t) - V_t(s_t)$. Note that

$$\Delta_t^{(n)} = \sum_{\tau=0}^{n-1} \gamma^{\tau} \Delta_{t+\tau}$$

Indeed,

$$\begin{aligned}
\sum_{\tau=0}^{n-1} \gamma^{\tau} \Delta_{t+\tau} &= \sum_{\tau=0}^{n-1} \gamma^{\tau} \Big(r\big(s_{t+\tau}, \pi(s_{t+\tau})\big) + \gamma\, V_t(s_{t+\tau+1}) - V_t(s_{t+\tau}) \Big) \\
&= \sum_{\tau=0}^{n-1} \gamma^{\tau}\, r\big(s_{t+\tau}, \pi(s_{t+\tau})\big) + \sum_{\tau=0}^{n-1} \Big(\gamma^{\tau+1}\, V_t(s_{t+\tau+1}) - \gamma^{\tau}\, V_t(s_{t+\tau}) \Big) \\
&= \sum_{\tau=0}^{n-1} \gamma^{\tau}\, r\big(s_{t+\tau}, \pi(s_{t+\tau})\big) + \gamma^n\, V_t(s_{t+n}) - V_t(s_t) \\
&= R_t^{(n)}(s_t) - V_t(s_t) = \Delta_t^{(n)}
\end{aligned}$$

It can be shown that if we run TD(0) with a n-step lookahead (for any given $n \geq 1$), then $V_t(s)$ converges to $V^{\pi}(s)$ for all $s \in \mathcal{S}$.

In case of deterministic T (finite horizon), we can choose $n = T$ and set $\gamma = 1$. The resulting algorithm is known as **Monte-Carlo sampling**.

In the discounted setting, the choice of n may impact the quality of the policy evaluation process. Instead of choosing a single value for n, we may average over all positive integers. A simple way of implementing this idea is through exponential averaging with a parameter $\lambda \in (0, 1)$. This implies that the weight assigned to each parameter n is $(1 - \lambda)\lambda^{n-1}$. This leads to the TD($\lambda$) algorithm.

Recall $\Delta_t = R_t^{(1)}(s_t) - V_t(s_t)$. The TD($\lambda$) update is defined by

$$V_{t+1}(s_t) = V_t(s_t) + (1 - \lambda)\eta_t \sum_{n=1}^{\infty} \lambda^{n-1} \Delta_t^{(n)}$$

The problem with this approach is that we have to compute an infite sum to make a single update. Luckily, there is an equivalent formulation that avoids this problem. The trick is to use the notion of **eligibility trace**

$$e_t(s) = \sum_{k=0}^{t} \eta_k (\lambda\gamma)^{t-k} \mathbb{I}\{s = s_k\}$$

Note that e_t can be recursively computed from $e_t(s) = 0$ and $e_t(s) = (\lambda\gamma)e_{t-1}(s) + \eta_t \mathbb{I}\{s = s_t\}$ for all $s \in \mathcal{S}$.

Now recall the definition of temporal difference,

$$\Delta_t = r\big(s_t, \pi(s_t)\big) + \gamma\, V_t(s_{t+1}) - V_t(s_t)$$

The backward temporal difference is just the standard temporal difference multiplied by the eligibiity trace,

$$\Delta_t^{\mathrm{B}}(s) = \Delta_t\, e_t(s)$$

The resulting algorithm is described below here. Note that in the backward view all states s get updated at each time step t.

Algorithm 1 TD(λ)

Input: Stationary deterministic policy π, initial state $s_0 \in \mathcal{S}$, parameter $\lambda \in (0, 1)$

Set $V_0(s) = 0$ and $e_0(s) = 0$ for all $s \in \mathcal{S}$
for $t = 0, 1, \ldots$ **do**
    Get $a_t = \pi(s_t)$ and observe $r(s_t, a_t)$, $s_{t+1} \sim p(\cdot \mid s_t, a_t)$
    Compute $\Delta_t = r(s_t, a_t) + \gamma\, V_t(s_{t+1}) - V_t(s_t)$
    for $s \in \mathcal{S}$ **do**
        Compute $e_t(s) = (\lambda\gamma)e_{t-1}(s) + \eta_t \mathbb{I}\{s = s_t\}$
        Update $V_{t+1}(s) = V_t(s) + e_t(s)\Delta_t$
    end for
end for

The following result shows that the forward and backward updates

$$V_{t+1}^{\mathrm{F}}(s_t) = V_t^{\mathrm{F}}(s_t) + (1 - \lambda)\eta_t \sum_{n=1}^{\infty} \lambda^{n-1} \Delta_t^{(n)} \qquad \text{and} \qquad V_{t+1}^{\mathrm{B}}(s) = V_t^{\mathrm{B}}(s) + \Delta_t^{\mathrm{B}}(s) \quad s \in \mathcal{S}$$

converge to the same limit.

Theorem 2 *Assume*

$$\lim_{t\to\infty} \mathbb{I}\{s_t = s\} = \infty \qquad \textit{and} \qquad \lim_{t\to\infty} V^{\mathrm{F}}(s) = V^{\pi}(s)$$

for all $s \in \mathcal{S}$ *with probability* 1. *Let* $V_0^{\mathrm{F}}(s) = 0$ *and* $V_0^{\mathrm{B}}(s) = 0$ *for all* $s \in \mathcal{S}$. *Then*

$$\lim_{t\to\infty} V^{\mathrm{B}}(s) = V^{\pi}(s) \qquad s \in \mathcal{S}$$

PROOF. Fix any $s \in \mathcal{S}$. Since $V_0^{\mathrm{F}}(s) = 0$, $s_t = s$ occurs for infinitely many t with probability 1, and $V_{t+1}^{\mathrm{F}}(s) = V_t^{\mathrm{F}}(s)$ when $s_t \neq s$, we have that

$$\lim_{t\to\infty} V^{\mathrm{F}}(s) = \sum_{t=0}^{\infty} \Big(V_{t+1}^{\mathrm{F}}(s) - V_t^{\mathrm{F}}(s)\Big)\mathbb{I}\{s_t = s\}$$

Likewise, using $V_0^{\mathrm{B}}(s) = 0$,

$$\lim_{t\to\infty} V^{\mathrm{B}}(s) = \sum_{t=0}^{\infty} \Big(V_{t+1}^{\mathrm{B}}(s) - V_t^{\mathrm{B}}(s)\Big)$$

Therefore, we are left to prove that

$$\sum_{t=0}^{\infty} \Big(V_{t+1}^{\mathrm{F}}(s) - V_t^{\mathrm{F}}(s)\Big)\mathbb{I}\{s_t = s\} = \sum_{t=0}^{\infty} \Big(V_{t+1}^{\mathrm{B}}(s) - V_t^{\mathrm{B}}(s)\Big)$$

We have the following chain of equalities

$$\begin{aligned}
\sum_{t=0}^{\infty} \Big(V_{t+1}^{\mathrm{F}}(s) - V_t^{\mathrm{F}}(s)\Big)\mathbb{I}\{s_t = s\} &= (1-\lambda)\sum_{t=0}^{\infty}\eta_t\mathbb{I}\{s_t = s\}\sum_{n=1}^{\infty}\lambda^{n-1}\Delta_t^{(n)} \\
&= (1-\lambda)\sum_{t=0}^{\infty}\eta_t\mathbb{I}\{s_t = s\}\sum_{n=1}^{\infty}\lambda^{n-1}\sum_{\tau=0}^{n-1}\gamma^{\tau}\Delta_{t+\tau} \\
&= (1-\lambda)\sum_{t=0}^{\infty}\eta_t\mathbb{I}\{s_t = s\}\sum_{n=0}^{\infty}\lambda^{n}\sum_{\tau=0}^{n}\gamma^{\tau}\Delta_{t+\tau} \\
&= (1-\lambda)\sum_{t=0}^{\infty}\eta_t\mathbb{I}\{s_t = s\}\sum_{\tau=0}^{\infty}\sum_{n=\tau}^{\infty}\lambda^{n}\gamma^{\tau}\Delta_{t+\tau} \\
&= (1-\lambda)\sum_{t=0}^{\infty}\eta_t\mathbb{I}\{s_t = s\}\sum_{k=t}^{\infty}\sum_{n=k-t}^{\infty}\lambda^{n}\gamma^{k-t}\Delta_{k} \\
&= (1-\lambda)\sum_{t=0}^{\infty}\eta_t\mathbb{I}\{s_t = s\}\sum_{k=t}^{\infty}(\lambda\gamma)^{k-t}\Delta_{k}\sum_{n=k-t}^{\infty}\lambda^{n-k+t} \\
&= \sum_{t=0}^{\infty}\eta_t\mathbb{I}\{s_t = s\}\sum_{k=t}^{\infty}(\lambda\gamma)^{k-t}\Delta_{k} \\
&= \sum_{k=0}^{\infty}\Delta_{k}\sum_{t=0}^{k}\eta_t\mathbb{I}\{s_t = s\}(\lambda\gamma)^{k-t} \\
&= \sum_{k=0}^{\infty}\Delta_k^{\mathrm{B}}(s) = \sum_{k=0}^{\infty}\Big(V_{k+1}^{\mathrm{B}}(s) - V_k^{\mathrm{B}}(s)\Big)
\end{aligned}$$

Since we chose s arbitrarily, the proof is concluded. $\square$