STAT 234 – Sequential Decision Making Feb 12, 2019

Lecture 5: Temporal Difference Methods

Instructor: Susan Murphy *Scribe: Kelly Zhang*

1 Review: Bellman Equation

Proposition 1. *(Bellman Equation) For a given policy π, its value function $V^\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_t = s]$ satisfies for all $s \in \mathcal{S}$ (state space)*

$$V^\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma V^\pi(S_{t+1}) | S_t = s]$$

1.1 Temporal Difference Error

From the Bellman equation we have that

$$\begin{aligned} 0 &= \mathbb{E}_\pi[R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) | S_t = s] \\ &= \mathbb{E}_\pi[\mathbb{1}_{S_t=s}(R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t))] \end{aligned}$$

Summing over all timesteps in the episode we get that

$$0 = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \mathbb{1}_{S_t=s}(R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t))\right]$$

Since the equality holds for all states $s \in \mathcal{S}$, we have that

$$0 = \sum_{s \in \mathcal{S}} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \mathbb{1}_{S_t=s}(R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t))\right]$$

If we define a vector of indicator random variables $\mathbf{1}_{S_t=s} \in \mathbb{R}^{|\mathcal{S}|}$ as

$$\mathbf{1}_{S_t=s} := [\mathbb{1}_{S_t=\mathcal{S}_{(1)}}, \mathbb{1}_{S_t=\mathcal{S}_{(2)}}, ..., \mathbb{1}_{S_t=\mathcal{S}_{(|\mathcal{S}|)}}]$$

we can also write the summation over $\mathcal{S}$ as a system of $|\mathcal{S}|$ equations

$$0 = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \mathbf{1}_{S_t=s}(R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t))\right]$$

Definition 1. *(Temporal Difference Error) The temporal difference error is the difference between the sum of discounted future rewards after visiting a state (our target) and our current estimate of that state's value:*

$$\text{TD Error} = [R_{t+1} + \gamma V^\pi(S_{t+1})] - V^\pi(S_t)$$

1.2 Episode Level Update Methods

Suppose we have the data for trajectories $i \in [1, n]$: $(S_0^i, A_0^i, R_1^i), (S_1^i, A_1^i, R_2^i), ..., (S_{T-1}^i, A_{T-1}^i, R_T^i)$. We assume we are in the episodic setting where $T \sim Geometric(\gamma)$.

Since $0 = \mathbb{E}_\pi\big[\sum_{t=0}^{\infty} \mathbb{1}_{S_t=s}(R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t))\big]$ and $\mathbb{E}_\pi[R_{t+1} + \gamma V^\pi(S_{t+1})] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+1+k}]$ thus for all $s \in \mathcal{S}$,

$$V^\pi(s) = \frac{\mathbb{E}_\pi\big[\sum_{t=0}^{\infty} \mathbb{1}_{S_t=s} \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}\big]}{\mathbb{E}_\pi\big[\sum_{t=0}^{\infty} \mathbb{1}_{S_t=s}\big]}$$

Using the method of moments we can use the sample trajectories to approximate the expectation to get the Monte Carlo estimator.

Estimator 1. *(Multiple episode update rule (Monte Carlo)) For all $s \in \mathcal{S}$,*

$$\hat{V}^\pi(s) = \frac{\mathbb{P}_n\big(\sum_{t=0}^{T-1} \mathbb{1}_{S_t=s} \sum_{k=0}^{T-1-t} \gamma^k R_{t+1+k}\big)}{\mathbb{P}_n\big(\sum_{t=0}^{T-1} \mathbb{1}_{S_t=s}\big)}$$

where $\mathbb{P}_n[f(X)] = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. This estimator is considered a *batch update rule* because the update is a function of multiple episodes.

Estimator 2. *(Between-episode update rule) For all $s \in \mathcal{S}$,*

$$\hat{V}_{n+1}^\pi(s) := \hat{V}_n^\pi(s) + \alpha_n \sum_{t=0}^{T-1} \mathbb{1}_{S_t=s}\left(\sum_{k=0}^{T-1-t} \gamma^k R_{t+1+k} - \hat{V}_n^\pi(s)\right)$$

Note that this update rule is equivalent to the batch update rule when $\alpha_n := \frac{1}{(n+1)\mathbb{P}_{n+1}(\sum_{t=0}^{T-1} \mathbb{1}_{S_t=s})}$.

2 Temporal Difference Methods

Note that the between-episode update rule requires us to finish the entire episode before updating our value function estimate. Temporal difference methods allow us to update the value function estimate *within* an episode, which may lead to faster learning.

2.1 TD(0)

The *temporal difference error* in the between-episode update rule is the difference between the entire future reward after visiting a state and the estimated value of that state:

$$\hat{V}_{n+1}^\pi(s) := \hat{V}_n^\pi(s) + \alpha_n \sum_{t=0}^{T-1} \mathbb{1}_{S_t=s}\Bigg(\underbrace{\sum_{k=0}^{T-1-t} \gamma^k R_{t+1+k} - \hat{V}_n^\pi(s)}_{\text{TD error}}\Bigg)$$

Suppose though that instead of updating only at the end of episodes, we want to update the value function at every timestep *within* each episode. For the between-episode update rule, we sum over all time-steps because we have to see all future rewards in order to estimate the value of any given state anyway. In the within-episode update rule, we remove the summation over all time-steps in the episode by finding an alternative estimate of the sum of future rewards for a state. Recall the Bellman equation,

$$V^\pi(S_t) = \mathbb{E}_\pi\left[\sum_{k=0}^{T-1-t} \gamma^k R_{t+1+k} \middle| S_t\right] = \mathbb{E}_\pi[R_{t+1} + \gamma V^\pi(S_{t+1})|S_t]$$

We can "bootstrap" by using our current estimate of the value function $\hat{V}^\pi$ to make the following approximation:

$$\mathbb{E}_\pi\left[\sum_{k=0}^{T-1-t} \gamma^k R_{t+1+k} \middle| S_t\right] \approx \mathbb{E}_\pi[R_{t+1} + \gamma \hat{V}^\pi(S_{t+1})|S_t]$$

Using $R_{t+1} + \gamma \hat{V}^\pi(S_{t+1})$ as the target in our temporal difference error is called TD(0) and it allows us to update the value function at every timestep within the episode (after seeing the first non-zero reward).

Estimator 3. *(TD(0) update rule) We let m index over updates of $\hat{V}^\pi$ and t index over timesteps within the current episode. Note we can update our value function estimate at each timestep for all n episodes in succession. For all $s \in \mathcal{S}$,*

$$\hat{V}^\pi_{m+1}(s) := \hat{V}^\pi_m(s) + \alpha_m \mathbb{1}_{S_t=s}\left(R_{t+1} + \gamma \hat{V}^\pi_m(S_{t+1}) - \hat{V}^\pi_m(s)\right)$$

Note that there is no α_m that in general makes the TD(0) update rule equivalent to the batch update rule.[1]

2.2 k-step and λ Returns

In TD(0) our new formulation of the target in our TD error term is based on the Bellman equation after applying recursion once, which allows us to update the value function after just one timestep within an episode. More generally though we can apply the Bellman recursion k times and get an update of the value function after seeing k timesteps.

$$\begin{aligned} V^\pi(S_t) &= \mathbb{E}_\pi[R_{t+1} + \gamma V^\pi(S_{t+1})|S_t] \\ &= \mathbb{E}_\pi\big[R_{t+1} + \gamma \mathbb{E}_\pi[R_{t+2} + \gamma V^\pi(S_{t+2})|S_{t+1}]|S_t\big] \end{aligned}$$

[1] TD(0) has been proven to converge to the true value function for cases when $\hat{V}^\pi$ is tabular or a linear function. See sections 6.2 and 9.2 in Sutton and Barto for more details.

$$= \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 V^\pi(S_{t+1})|S_t]$$
$$= \mathbb{E}_\pi\left[\sum_{j=0}^{k-1} \gamma^j R_{t+1+j} + \gamma^k V^\pi(S_{t+k})\middle| S_t\right]$$

Definition 2. *(k-step return)* For $k \in \mathbb{N}$,

$$R_t^{(k)} := \sum_{j=0}^{k-1} \gamma^j R_{t+1+j} + \gamma^k V^\pi(S_{t+k})$$

Estimator 4. *(k-step return update rule) For all $s \in \mathcal{S}$,*

$$\hat{V}_{m+1}^\pi(s) := \hat{V}_m^\pi(s) + \alpha_m \mathbb{1}_{S_t=s}\left(R_{t+1}^{(k)} - \hat{V}_m^\pi(s)\right)$$

An alternative target in the TD error is the exponentially weighted average of the k-step returns for all $k \in [1, \infty)$. This target is called the λ-return.

Definition 3. *(λ-return)* For $0 \leq \lambda < 1$,

$$R_t^{(\lambda)} := (1-\lambda)\sum_{k=1}^{\infty} \lambda^{k-1} R_t^{(k)}$$

Note that assuming bounded rewards, $\mathbb{E}_\pi[R_t^\lambda|S_t] = V^\pi(S_t)$ because

$$\mathbb{E}_\pi\left[(1-\lambda)\sum_{k=1}^{\infty} \lambda^{k-1} R_t^{(k)}\middle| S_t\right] = (1-\lambda)\sum_{k=1}^{\infty} \lambda^{k-1}\mathbb{E}_\pi\left[R_t^{(k)}|S_t\right] = V^\pi(S_t)$$

Note that we must wait until the end of the episode to compute the λ-return. In the next section we describe the advantages of TD(λ), which uses a modified version of the λ-return as the target.[2]

2.3 TD(λ)

As with the Monte Carlo estimator, the λ-return update rule requires the episode to end before we can do an update. We show that we can modify the λ-return from an between episode into a within episode upate called TD(λ). Moreover, TD(λ) will update all recently visited states with every non-zero reward received.

[2]For $\lambda = 0$, TD(λ) = TD(0).

Estimator 5. *(TD(λ) between episode update) For all $s \in \mathcal{S}$,*

$$\hat{V}^{\pi}_{n+1}(s) := \hat{V}^{\pi}_{n}(s) + \alpha_n \sum_{t=0}^{\infty} \mathbb{1}_{S_t=s}(R^{\lambda}_t - \hat{V}^{\pi}_{n}(s))$$

Proposition 2. *(λ-return temporal difference formulation)*

$$R^{\lambda}_t - V^{\pi}(S_t) = \sum_{k=0}^{\infty} (\lambda\gamma)^k (R_{t+k+1} + \gamma V^{\pi}(S_{t+1+k}) - V^{\pi}(S_{t+k}))$$

Proof:

$$\begin{aligned}
R^{\lambda}_t - V^{\pi}(S_t) &= -V^{\pi}(S_t) + (1-\lambda)\sum_{k=1}^{\infty} \lambda^{k-1} R^{(k)}_t \\
&= -V^{\pi}(S_t) + (1-\lambda)\lambda^0(R_{t+1} + \gamma V^{\pi}(S_{t+1})) \\
&\qquad + (1-\lambda)\lambda^1(R_{t+1} + \gamma R_{t+2} + \gamma^2 V^{\pi}(S_{t+2})) \\
&\qquad + (1-\lambda)\lambda^2(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V^{\pi}(S_{t+3})) \\
&\qquad + ... \\
&= -V^{\pi}(S_t) + R_{t+1} + (1-\lambda)\gamma V^{\pi}(S_{t+1}) \\
&\qquad\quad + (\lambda\gamma)(R_{t+2} + (1-\lambda)\gamma V^{\pi}(S_{t+2})) \\
&\qquad\quad + (\lambda\gamma)^2(R_{t+3} + (1-\lambda)\gamma V^{\pi}(S_{t+3})) + ... \\
&= \sum_{k=0}^{\infty} (\lambda\gamma)^k (R_{t+k+1} + \gamma V^{\pi}(S_{t+1+k}) - V^{\pi}(S_{t+k})) \qquad \blacksquare
\end{aligned}$$

Now let us derive the within episode TD(λ) update from the between episode update rule. For convenience we let $\hat{\delta}_{t+k,n} = R_{t+1+k} + \gamma \hat{V}^{\pi}_n(S_{t+1+k}) - \hat{V}^{\pi}_n(S_{t+k})$

$$\begin{aligned}
\hat{V}^{\pi}_{n+1}(s) &= \hat{V}^{\pi}_{n}(s) + \alpha_n \sum_{t=0}^{\infty} \mathbb{1}_{S_t=s}(R^{\lambda}_t - \hat{V}^{\pi}_{n}(s)) \\
&= \hat{V}^{\pi}_{n}(s) + \alpha_n \sum_{t=0}^{\infty} \mathbb{1}_{S_t=s} \sum_{k=0}^{\infty} (\lambda\gamma)^k \hat{\delta}_{t+k,n} \\
&= \hat{V}^{\pi}_{n}(s) + \alpha_n \sum_{t=0}^{\infty} \mathbb{1}_{S_t=s} \sum_{j=t}^{\infty} (\lambda\gamma)^{j-t} \hat{\delta}_{j,n} \\
&= \hat{V}^{\pi}_{n}(s) + \alpha_n \sum_{j=0}^{\infty} \hat{\delta}_{j,n} \underbrace{\sum_{t=0}^{j} \mathbb{1}_{S_t=s} (\lambda\gamma)^{j-t}}_{\varphi_j(s)}
\end{aligned}$$

Thus, for all $s \in \mathcal{S}$

$$\hat{V}^{\pi}_{n+1}(s) = \hat{V}^{\pi}_{n}(s) + \alpha_n \sum_{j=0}^{\infty} \hat{\delta}_{j,n} \varphi_j(s)$$

Note that $\varphi_j(s)$ will be non-zero for all previously seen states in the episode (older episodes are discounted more), this means that whenever a non-zero reward is received, the value of all states previously visited in the episode will be updated.

Just as we modified the Monte Carlo between episode update into the TD(0) update, we can now modify the between episode TD(λ) update to a between episode update by removing the summation over future residual errors.

Estimator 6. *(TD(λ) within episode update) We let m index over updates of $\hat{V}^{\pi}$ and t index over timesteps within the current episode. Note we can update our value function estimate at each timestep for all n episodes in succession. For all $s \in \mathcal{S}$,*

$$\hat{V}_{m+1}(s) := \hat{V}_{m+1}(s) + \alpha_m \hat{\delta}_{t,m} \varphi_t(s)$$