

Maximising Long-Term Reward Using Temporal Meta-Differences

David Silver

11th November, 2004

- Consider a reinforcement learning algorithm such as $TD(\lambda)$
- Each update can be considered a state transition in a giant MDP: from state s and value function V , to new state s' and new value function V'
- We want to find the sequence of updates that maximises our long-term reward

- Define this to be a *meta-MDP* with states $\bar{s} = \langle V, s \rangle$
- and meta-policy $\bar{\pi}$ corresponding to the algorithm used
- and meta-value-function $\bar{V}^{\bar{\pi}}$ corresponding to the expected return whilst following this algorithm

$$\bar{V}_t^{\bar{\pi}}(\langle V, s \rangle) = E_{\bar{\pi}}\{E_{\pi_t, \pi_{t+1}, \pi_{t+2}, \dots}\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots | s_t = s\} | V_t = V\}$$

(1)

- The meta-MDP has state space exponential in the size of original MDP!
- Instead, estimate the meta-value-function using the following approximation:

$$\bar{V}(\langle V, s \rangle) \approx V(s) + \bar{W}(V(s), s) \quad (2)$$

where $\bar{W}(v, s)$ is the *improvement function* at state s given a value function at that state, $v = V(s)$

The expected value of $\bar{W}_t^{\bar{\pi}}$ can be derived as follows:

$$\begin{aligned}
\bar{V}_t^{\bar{\pi}}(\langle V, s \rangle) &= E_{\bar{\pi}}\{E_{\pi_t, \pi_{t+1}, \pi_{t+2}, \dots}\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots | s_t = s\} | V_t = V\} \\
&= E_{\bar{\pi}} \left\{ \begin{array}{l} + E_{\pi_t} \{r_{t+1} | s_t = s\} \\ + \gamma E_{\pi_t, \pi_{t+1}} \{r_{t+2} | s_t = s\} \\ + \gamma^2 E_{\pi_t, \pi_{t+1}, \pi_{t+2}} \{r_{t+3} | s_t = s\} \\ \vdots \end{array} \middle| V_t = V \right\}
\end{aligned} \tag{3}$$

Approximate \bar{V}_t by truncating the sequence of policy improvements at the k^{th} step:

$$\widehat{V}_{t,k}^{\bar{\pi}}(\langle V, s \rangle) = E_{\bar{\pi}} \left\{ \begin{array}{l} E_{\pi_t} \{r_{t+1} | s_t = s\} \\ + \gamma E_{\pi_t, \pi_{t+1}} \{r_{t+2} | s_t = s\} \\ + \gamma^2 E_{\pi_t, \pi_{t+1}, \pi_{t+2}} \{r_{t+3} | s_t = s\} \\ \vdots \\ + \gamma^k E_{\pi_t, \pi_{t+1}, \dots, \pi_{t+k}} \{r_{t+k+1} + \gamma r_{t+k+2} + \gamma^2 r_{t+k+3} + \dots | s_t = s\} \end{array} \middle| V_t = V \right\} \quad (4)$$

This approximation converges in the limit:

$$\lim_{k \rightarrow \infty} \widehat{V}_{t,k}^{\bar{\pi}}(\langle V, s \rangle) = \bar{V}_t^{\bar{\pi}}(\langle V, s \rangle) \quad (5)$$

Expressing as an iterative sequence:

$$\begin{aligned}
\widehat{V}_{t,k+1}^{\bar{\pi}}(\langle V, s \rangle) &= \widehat{V}_{t,k}^{\bar{\pi}}(\langle V, s \rangle) \\
&+ E_{\bar{\pi}} \left\{ \begin{array}{l} \gamma^k E_{\pi_t, \pi_{t+1}, \dots, \pi_{t+k}} \{r_{t+k+1} | s_t = s\} \\ + \gamma^{k+1} E_{\pi_t, \pi_{t+1}, \dots, \pi_{t+k+1}} \{r_{t+k+2} + \gamma r_{t+k+3} + \gamma^2 r_{t+k+4} + \dots | s_t = s\} \\ - \gamma^k E_{\pi_t, \pi_{t+1}, \dots, \pi_{t+k}} \{r_{t+k+1} + \gamma r_{t+k+2} + \gamma^2 r_{t+k+3} + \dots | s_t = s\} \end{array} \middle| V_t = V \right\} \\
&= \widehat{V}_{t,k}^{\bar{\pi}}(\langle V, s \rangle) + \gamma^k E_{\bar{\pi}} \left\{ \begin{array}{l} + \frac{r_{t+k+1}}{\gamma V_{t+k+1}(s)} \\ - V_{t+k}(s) \end{array} \middle| V_t = V, s_t = s \right\} \\
&= \widehat{V}_{t,k}^{\bar{\pi}}(\langle V, s \rangle) + E_{\bar{\pi}} \{ \gamma^k \Delta V_{t+k}(s) | V_t = V, s_t = s \}
\end{aligned} \tag{6}$$

where ΔV_t represents the *temporal meta-difference* at time t , defined by:

$$\Delta V_t(s) = r_{t+k+1} + \gamma V_{t+k+1}(s) - V_{t+k}(s) \tag{7}$$

It follows that:

$$\begin{aligned}\bar{V}_t^{\bar{\pi}}(\langle V, s \rangle) &= \hat{V}_{t,0}^{\bar{\pi}}(\langle V, s \rangle) + E_{\bar{\pi}}\{\Delta V_0(s) + \gamma \Delta V_1(s) + \gamma^2 \Delta V_2(s) + \dots | V_t = V, s_t = s\} \\ &= V(s) + \bar{W}_t^{\bar{\pi}}(\langle V, s \rangle)\end{aligned}\tag{8}$$

and so the improvement function $\bar{W}_t^{\bar{\pi}}$ is:

$$\bar{W}_t^{\bar{\pi}}(\langle V, s \rangle) = E_{\bar{\pi}}\{\Delta V_0(s) + \gamma \Delta V_1(s) + \gamma^2 \Delta V_2(s) + \dots | V_t = V, s_t = s\}\tag{9}$$

This is like an expected return, where the reward \bar{r}_t is now $\Delta V_t(s)$

Also this only depends on s and $v = V(s)$, not the value function at other states.

The Bellman equation is:

$$\bar{W}^{\bar{\pi}}(\langle v, s \rangle) = E_{\bar{\pi}}\{\Delta V(s) + \gamma \bar{W}^{\bar{\pi}}(\langle v', s \rangle)\} \quad (10)$$

Which we can estimate using an update of the form:

$$\bar{W}(\langle v, s \rangle) \leftarrow \bar{W}(\langle v, s \rangle) + \beta[\Delta V(s) + \gamma \bar{W}(\langle v', s \rangle) - \bar{W}(\langle v, s \rangle)] \quad (11)$$

- If the meta-value function was known exactly, we could apply a greedy algorithm to give optimal exploration.
- But we only have an estimate.
- So we must also explore at the meta-level to make sure that our estimate continues to improve.

The algorithm illustrated here uses TD(0) on the original MDP, $TD(0)$ at the meta-level, and ϵ -greedy exploration:

Repeat forever:

Take action a from state s , observe r, s'

Choose a' from s' using π_t

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a') - Q(s, a)]$$

$$\pi(s, a) \leftarrow \epsilon\text{-greedy over } Q(s, a) + W(V(s), s, a)$$

$$V'(s) \leftarrow \sum_a \pi(s, a) Q(s, a)$$

$$\Delta V(s) \leftarrow r + \gamma V'(s) - V(s)$$

$$\bar{W}(V(s), s, a) \leftarrow \bar{W}(V(s), s, a) + \beta[\Delta V(s) + \gamma \bar{W}(V'(s), s, a) - \bar{W}(V(s), s, a)]$$

$$V(s) \leftarrow V'(s); s \leftarrow s'; a \leftarrow a'$$

Some possible extensions to the basic algorithm:

- Use a function approximator for \bar{W}
- Use options and SMDPs
- Model the uncertainty in the value function
- Use average-reward formulation instead of discounting

Assuming that the error in the meta-value function is normally distributed:

- Keep a running estimate of the variance
- Sample this distribution and add to the estimate of $\bar{V}^{\bar{\pi}}$
- Use a greedy policy over the noisy estimates of $\bar{V}^{\bar{\pi}}$

Using average-reward formulation has several key advantages:

- We can actually measure long-term reward (no discounting)
- Can take account of improvements across all states
- Can treat episodic tasks as continuing tasks without bias

Measure the improvement in the average reward ρ :

$$\Delta\rho_t = \rho_{t+1} - \rho_t \quad (12)$$

We know that the average improvement must converge on zero in the limit:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n E_{\bar{\pi}}(\Delta\rho_{t+k} | V_t = V, s_t = s) = 0 \quad (13)$$

Define the meta-value function:

$$\bar{V}_t^{\bar{\pi}}(\langle V, s \rangle) = V(s) + \bar{W}_t^{\bar{\pi}}(\langle V, s \rangle) \quad (14)$$

Where the improvement function is the transient improvement:

$$\bar{W}_t^{\bar{\pi}}(\langle V, s \rangle) = \sum_{k=0}^n E_{\bar{\pi}}(\Delta\rho_{t+k} | V_t = V, s_t = s) \quad (15)$$

This is just like calculating the value function, where reward is now $\Delta\rho$ and the average reward is now zero.

Repeat forever:

Take action a from state s , observe r, s'

Choose a' from s' using π_t

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r - \rho + Q(s', a') - Q(s, a)]$$

$$\Delta\rho \leftarrow r - \rho + Q(s', a') - Q(s, a)$$

$$\rho \leftarrow \rho + \mu\Delta\rho$$

$$W(s, a) \leftarrow W(s, a) + \nu[\Delta\rho + W(s', a') - W(s, a)]$$

$$\pi(s, a) \leftarrow \epsilon\text{-greedy over } Q(s, a) + W(s, a)$$

$$\rho \leftarrow \rho'; s \leftarrow s'; a \leftarrow a'$$