04. Convergence

Introduction:

- TD with control is about actually learning the values of different actions we might take to figure out the right way to behave, instead of just learning the values with update rules like TD.
- So, the difference between using RL with control and RL without control is whether or not there's actions that are being chosen by the learner.

Bellman Equations:

- This is the Bellman equation without actions:

$$V(s) = R(s) + \gamma \sum_{s'} T(s, s')V(s')$$

- Talking about TD, we used the notion of sequence of states/rewards (TD(0)):

$$V_t(s_{t-1}) = V_{t-1}(s_{t-1}) + \alpha_t(r_t + \gamma V_{t-1}(s_t) - V_{t-1}(s_{t-1}))$$

- Now, we'll modify these equations to account for actions:

$$Q(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s')\, max_{a'} Q(s', a')$$

$$Q_t(s_{t-1}, a_{t-1}) = Q_{t-1}(s_{t-1}, a_{t-1}) + \alpha_t(r_t + \gamma max_{a'} Q_{t-1}(s_t, a') - V_{t-1}(s_{t-1}, a_{t-1}))$$

- This Q – learning update rule handles two kinds of approximations at the same time:
  - If we knew the model, we can update the Q values by applying the Bellman Equations:

$$Q(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s')\, max_{a'} Q_{t-1}(s', a')$$

  - If we knew Q^* (the Q – value for each state-action pair that would be obtained by following the optimal policy after taking action a in state s), we can use that to learn Q^*:

$$Q_T(s_{t-1}, a_{t-1}) = Q_{T-1}(s_{t-1}, a_{t-1}) + \alpha_T(r_t + \gamma max_{a'} Q^*(s_t, a') - V_{t-1}(s_{t-1}, a_{t-1}))$$

    Since we're using a true value (Q^*), we'll eventually converge.

Bellman Operator and Contraction Mapping:

- Let B be an operator, or mapping from value functions to value functions.

$$[BQ](s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s')\, max_{a'} Q(s', a')$$

  - So, we give this Bellman Operator a Q function, and it gives back a different Q function that is equal to the immediate reward plus the discounted expected value of the next state where we look up the value of the next state using whichever Q function given as an input to the operator.

- Using the Bellman operator:
  1. $Q^* = BQ^*$ given us the Bellman Equation.
  2. $Q_t = BQ_{t-1}$ gives us Value Iteration.

- Contraction mapping:
  - If, for all F, G and $0 \leq \gamma < 1$:
    $$\|BF - BG\|_\infty \leq \gamma \|F - G\|_\infty \qquad where \quad \|Q\|_\infty = max_{s,a} Q(s,a)$$
    Then B is a contraction mapping.
  - In other words, if for all value functions F, G, the distance between B applied to F, and B applied to G is smaller than or equal the original distance between F and G multiplied by γ, then we say B is a contraction mapping. Which means that applying B makes the distance between the resulting functions closer than the original ones.

- Contraction properties: If B is a contraction mapping:
  1. $F^* = BF^*$ has a unique solution.
  2. $F_t = BF_{t-1}$: $F_t \rightarrow F^*$ We can converge to F^* by value iteration using B.
  - If we have two value function F_{t-1} and F^*:
    $$\|BF_{t-1} - BF^*\|_\infty \leq \gamma \|F_{t-1} - F^*\|_\infty$$
    If B is a contraction mapping, then the two properties above apply and we get the following:
    $$\|F_t - F^*\|_\infty \leq \gamma \|F_{t-1} - F^*\|_\infty$$
    So, we're getting closer to F^*.

- Bellman Operator contracts:
  $$[BQ](s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s')\, max_{a'} Q(s',a')$$
  Given Q_1, Q_2:
  $$\|BQ_1 - BQ_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$
  $$\|BQ_1 - BQ_2\| = max_{s,a} |[BQ_1](s,a) - [BQ_2](s,a)|$$
  $$= max_{s,a} \left| (R(s,a) + \gamma \sum_{s'} T(s,a,s')\, max_{a'} Q_1(s',a')) - (R(s,a) + \gamma \sum_{s'} T(s,a,s')\, max_{a'} Q_2(s',a')) \right|$$
  Since we're computing the different between Q values for the same state-action pair, they will have the same reward:
  $$\|BQ_1 - BQ_2\| = max_{s,a} \left| \gamma \sum_{s'} T(s,a,s')\, (max_{a'} Q_1(s',a') - max_{a'} Q_2(s',a')) \right|$$
  Since this is a weighted average, it can't be larger than the biggest difference, instead of considering the weight of the probability of reaching s', we'll just consider the s' where the difference is larger:
  $$\|BQ_1 - BQ_2\| \leq \gamma max_{s'} |max_{a'} Q_1(s',a') - max_{a'} Q_2(s',a')|$$
  Although we have a different a' for each of Q_1 and Q_2, because the max operator is a non-expansion (see next section), we can substitute using a single common a', removing the $max_{a'}$ inside the absolute:
  $$\|BQ_1 - BQ_2\| \leq \gamma max_{s',a'} |Q_1(s',a') - Q_2(s',a')| = \gamma \|Q_1 - Q_2\|_\infty$$

max is a Non-Expansion:

- For all f, g:

$$|max_a f(a) - max_a g(a)| \leq max_a |f(a) - g(a)|$$

- Proof: Without loss of generality, we'll assume that $max_a f(a) \geq max_a g(a)$, then:

$$|max_a f(a) - max_a g(a)| = max_a f(a) - max_a g(a)$$

Since we've a different a for each function, we'll assume:

$$a_1 = argmax_a f(a)$$
$$a_2 = argmax_a g(a)$$

Then:

$$max_a f(a) - max_a g(a) = f(a_1) - g(a_2)$$

Since a_2 was chosen to maximize the g function, any other a substituted instead of a_2, for example a_1, can only make the term $g(a_2)$ smaller or at least no bigger. So, we can conclude:

$$f(a_1) - g(a_2) \leq f(a_1) - g(a_1)$$

Now, we can reintroduce the max operator again:

$$f(a) - g(a) \leq max_a |f(a) - g(a)|$$

Then:

$$|max_a f(a) - max_a g(a)| \leq max_a |f(a) - g(a)|$$

Convergence Theorem:

- Theorem: Let B be a contraction mapping and $Q^* = BQ^*$ be its fixed point. Let Q_0 be a Q-function and define $Q_{t+1} = [B_t Q_t]Q_t$. Then, $Q_t \rightarrow Q^*$ if:
  1. The learning algorithm will update all state-action values on all time steps. However, if it's a state-action value that doesn't correspond to the current state-action pair that we just experienced, we'll set the learning rate to zero.

$$\alpha_t(s,a) = 0 \quad if \quad s_t \neq s \; and \; a_t \neq a$$

  2. Non-expansion property → For all Q-functions (U_1, U_2) and state-action pairs (s,a):

$$|([B_t U_1]Q^*)(s,a) - ([B_t U_2]Q^*)(s,a)| \leq (1 - \alpha_t(s,a))|U_1(s,a) - U_2(s,a)|$$

  3. Contraction property → For all Q-functions (Q, U) and state-action pairs (s,a):

$$|([B_t U]Q^*)(s,a) - ([B_t U]Q)(s,a)| \leq (\gamma\alpha_t(s,a))|Q^*(s,a) - Q(s,a)|$$

  4. Learning rate condition:

$$\sum_t \alpha_t = \infty \; and \; \sum_t {\alpha_t}^2 < \infty$$

Given the 1st assumption ($\alpha_t(s,a) = 0 \quad if \quad s_t \neq s \; and \; a_t \neq a$), this definition will not converge unless we visit every state-action pair infinitely often. If we don't, $\sum_t \alpha_t$ will be finite.