ECE 590D-001, Reinforcement Learning at Scale

Jay Hineman, Ph.D.

Geometric Data Analytics

2020

Mathematical details (restated) [?]

Following Sutton and Barto [?]

- t = time/iteration, $s, S_t \in \mathcal{S}$ = states, $a, A_t \in \mathcal{A}(s)$ = actions at state, $r, R_{t+1} \in \mathbb{R}$ = rewards, $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$ = trajectory
- transition function $= p(s', r'|s, a) := \text{Prob}\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\}$
- **probability fact:** $\sum_{s',r} p(s', r|s, a) = 1$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
- **return and discounted return:** $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$ and $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$
- **((board work))** from [?]

Mathematical specification of value and action-value (restated)

$$
\begin{aligned}
q_\pi(s) &= \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[\sum_{k=0} \gamma^k R_{(t+1)+k} | S_t = s, A_t = a], s \in \mathcal{S} \\
&= \text{Value of } s \text{ given the policy } \pi \\
v_\pi(s) &= \mathbb{E}_\pi[G_t | S_t = s] \\
&= \mathbb{E}_\pi[\sum_{k=0} \gamma^k R_{(t+1)+k} | S_t = s] \\
&= \text{action-value or } q\text{-value of } s, a \text{ given the policy } \pi
\end{aligned} \tag{1}
$$

- **((board work))** on *Bellman equation* for v_π from [**?**]

Bellman Equation

- **Bellman equation** for value function:
  $v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')$
- **Exercise:** Convince yourself, perhaps by making a more explicit notation, that (3.14) in [**?**] is simply an expectation.
- **Exercise:** Determine the Bellman equation for action-value. See Exercise 3.17 [**?**].

Optimal Policies and Value functions [?]

- *Solving a reinforcement learning task means, roughly, finding a policy that achieves a lot of reward over the long run*
- Since we can establish a partial ordering of policies we can talk about optimality: *A policy π is defined to be better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states. In other words, $\pi \geq \pi'$ if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in \mathcal{S}$.*
- Denote (maybe non-uniquely) optimal policies by π_*. *They share the same value function and action-value function!*

Optimal Policies and Value functions [?]

- **optimal state-value function:**

$$v_*(s) = \max_\pi v_\pi(s), \text{ for all } s \in \mathcal{S}$$

- **optimal action-value function:**

$$q_*(s, a) = \max_\pi q_\pi(s, a), \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

- Said another way:

$$\pi_* = \arg\max_\pi q_\pi(s, a) = \arg\max_\pi v_\pi(s), \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

Optimal Policies and Value functions [?]

- **Important identity:**

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

- v_* and q_* must satisfy Bellman equations. *Because they are optimal "value" functions, their consistency condition can be written in a special form without reference to any specific policy*.
- **Belleman optimality for** v_*

$$v_*(s) = \max_a \sum_{s',r} p(s', r|s,a)[r + \gamma v_*(s'), s \in \mathcal{S}$$

- **Belleman optimality for** q_*

$$q_*(s,a) = \sum_{s',r} p(s', r|s,a)[r + \gamma v_*(s'), s \in \mathcal{S}, a \in \mathcal{A}(\int)$$

Optimal Policies and Value functions [?]

- For finite MDPs, the Bellman optimality equation has a unique solution.
- Really, Bellman equations are (nonlinear) systems on for each state or state-action pair. *That means we can employ solution methods for such equations (fixed point iteration, gradient descent, etc,* **if the dynamics are known**

Optimal Policy and Value function for Grid World, Example 3.8

- In a future homework, we will compute similar value functions for gridworld

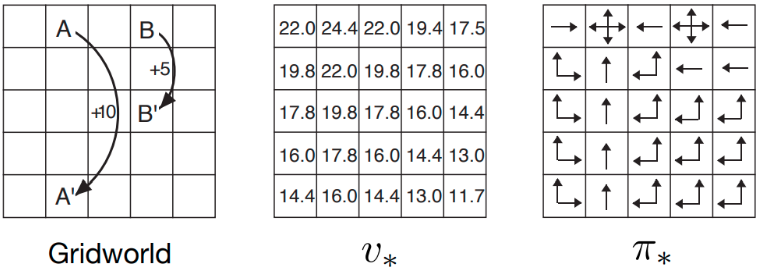

Figure: A gridworld with optimal policy and value function, Example 3.8 [?]

Spinning up's Taxonomy

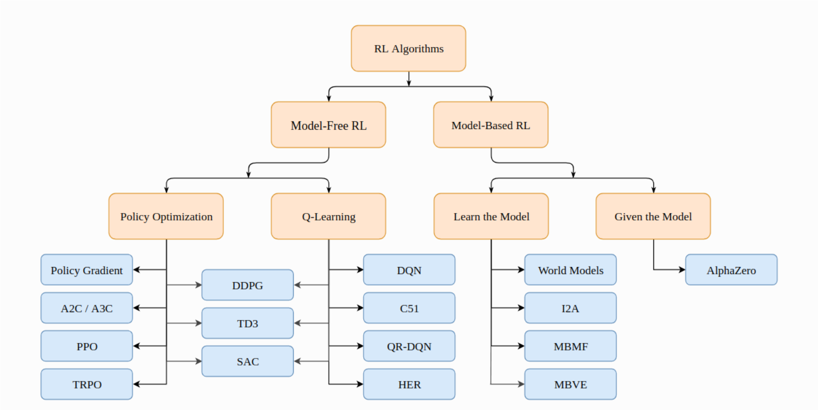

Figure: Non-exhaustive, but nice starting Taxonomy of (deep) RL methods. See also: citations.