Introduction to RL: A mathematically rigorous reading companion

Sebastian Scherer

September 24, 2018

1 Why another introduction?

The goal of this work is to provide a mathematically rigorous introduction to RL theory based on the excellent book "Introduction to reinforcement learning" by Sutton and Bartos (first edition). While I greatly enjoyed reading this book and appreciated its focused approach on developing an intuition for the Q- and V-functions, the algorithms and the general probabilistic framework introduced in the early chapters, I couldn't help but stumble at some points wondering how exactly a particular claim was justified. When I tried to bridge these gaps, further gaps unravelled, sometimes turning into chasms that I simply could not bridge using the theory presented in this book alone. In short, my inner mathematician wasn't satisfied with the inconsistent level of rigour applied throughout these sections. Queries on stack exchange as well as the various alternative resources applying even less rigour and, often times, introducing additional confusing notation, motivated me to try and remedy this myself. I therefore set out to try and rigorously formalize the theory presented, at least for the finite Markov Decicions Processes treated in the book, so that it may help let my inner mathematician sleep at night, as well as, and this is my sincere hope, provide a rigorous and helpful introduction for all those who are not only interested in the intuition but also appreciate a firm foundation on which to place it. The following manuscript can be used as an explanatory guide to the concepts presented in the book, or can be independently used as a rigourous introduction to value function theory in its own right.

2 Some notation

Like the reference book, we consider finite state, finite action markov decision processes ("finite MDPs"). As such, we denote by S the set of states achievable for a given finite MDP, and by A the set of executable actions a. We do not restrict ourselves to deterministic polices, and therefore treat a policy π as a conditional probability distribution over the executable action set A, conditioned on a given current state from S. In other words,

$$\begin{array}{rcl} \pi & : A \times S & \rightarrow & [0,1] \\ & (a,s) & \mapsto & \pi(a,s) \end{array}$$

where $\pi(a,s) = Pr_\pi(a|s)$ denotes the probability of choosing action $a \in A$ when in state $s \in S$ while acting according to policy π.

We encode our knowledge about the (reactionary) nature of our environment via the transition probabilities

$$P^a_{s,s'} = Pr(s_{t+1} = s' | s_t = s, a_t = a)$$

where $s, s' \in S$ and $a \in A$, denoting the probability of ending up in state s' at $t+1$ when coming from state s at t by executing a, and

$$R^a_{s,s'} = \mathbb{E}[r_t | s_t = s, s_{t+1} = s', a_t = a]$$

denoting the expected reward at time t due to ending up in state s' at $t+1$ when coming from state s at t by executing a.

Note that, in our notation, s_t and a_t denote the state and action at time t respectively, and thus r_t - NOT r_{t+1} as in the book - denotes the reward obtained AFTER being in s_t and executing a_t, thereby resulting in some (possibly the same) state s_{t+1}.

We use the same symbol $\gamma \in (0,1)$ to denote the reward discount factor.

Where deemed necessary, we will use expected values with policy indices, like $\mathbb{E}_\pi[\cdot]$, to clarify according to which distributions the specified expected value is to be viewed.

3 Value function and action-value function

As Sutton and Bartos point out, the standard approach to the analysis of optimal behaviour w.r.t a given MDP is by closely examining the value function V^π and action value function Q^π associated to a given policy π. They are an essential tool for quantitative analysis of policy driven behaviour, and thus, unsurprisingly, we will make heavy use of them throughout this guide. For a given policy π on a finite MDP, we call

$$\begin{array}{rcl} V^\pi : \ S & \rightarrow & \mathbb{R} \\ a & \mapsto & \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s] \end{array}$$

the *value function* of π, and

$$\begin{array}{rcl} Q^\pi : \ A \times S & \rightarrow & \mathbb{R} \\ (a,s) & \mapsto & \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a] \end{array}$$

the *action-value* function of π. The general idea of V^π is the quantification of the value a certain state s possesses under π, by assigning it the expected cumulative reward obtained when starting in that state s and then acting (i.e. choosing and exeucting actions) according to π. Q^π does very much the same thing, except for pairs of states and actions (a,s), assigning expected cumulative rewards when starting from s via action a, and only *then* acting according to π.

Let us have a closer look at the recursive nature of these functions - a feature we will exploit throughout the rest of this transcript.

Proposition 1. *(Parametrized bellman equations) Let π be an arbitrary policy, and let V^π and Q^π its associated (action-)value functions. Then the following identities hold:*

1. $V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} P^a_{s,s'}(R^a_{s,s'} + \gamma V^\pi(s'))$
2. $Q^\pi(s,a) = \sum P^a_{s,s'}[R^a_{s,s'} + \gamma \sum_{a'} \pi(s',a') Q^\pi(s',a')]$

Proof. To prove the first identity consider

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+k} | s_t = s] \\
&= \mathbb{E}_\pi[r_t + \gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s] \\
&= \mathbb{E}_\pi[r_t | s_t = s] + \mathbb{E}_\pi[\gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s] \\
&= \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s] \\
&= \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_{t+1} = s'] \\
&= \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} (R^a_{s,s'} + \gamma \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t'+k} | s_{t'} = s']) \\
&= \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} (R^a_{s,s'} + \gamma V^\pi(s')).
\end{aligned}$$

Note how we used the transitional probabilities and expected rewards to explicitly write out the expected value of the cumulative reward $\sum_{k=0}^\infty \gamma^k r_{t+k}$. Note also how we wrote $\mathbb{E}_\pi$ to indicate that the expected value is w.r.t to a sequence of actions $a_t, a_{t+1}, \dots$ and states $s_t, s_{t+1}, \dots$ resulting from acting according to π, made explicit in subsequent steps including the term $\pi(s,a)$.

A similar line of thinking shows us that

$$\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+k} | s_t = s, a_t = a] \\
&= \mathbb{E}_\pi[r_t + \gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s, a_t = a] \\
&= \mathbb{E}_\pi[r_t | s_t = s, a_t = a] + \gamma \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s, a_t = a] \\
&= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s, a_t = a, s_{t+1} = s', a_{t+1} = a'] \\
&= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_{t+1} = s', a_{t+1} = a'] \\
&= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t'+k} | s_{t'} = s', a_{t'} = a'] \\
&= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') Q^\pi(s',a') \\
&= \textstyle\sum_{s'} P^a_{s,s'} (R^a_{s,s'} + \gamma \sum_{a'} \pi(s,a') Q^\pi(s',a'))
\end{aligned}$$

Note that we used the markov property which allowed us to drop past states and actions when going from line 4 to line 5.

□

Note that, remembering the definitions of $P^a_{s,s'}$ and $R^a_{s,s'}$, these parametrizations can be rewritten in a slightly more compact way:

$$V^\pi(s) = \mathbb{E}_\pi[r_t + \gamma V^\pi(s_{t+1}) | s_t = s]$$

and

$$Q^\pi(s) = \mathbb{E}_\pi[r_t + \gamma Q^\pi(s_{t+1}, a_{t+1})|s_t = s, a_t = a].$$

These are the 'standard' bellman equations.

We now characterize the relationship between these two functions in the following

Proposition 2. *(QV Relationships) Let π be an arbitrary policy. Then the following identities hold:*

1. $V^\pi(s) = \sum_a \pi(s,a) Q_\pi(s,a)$

2. $V^\pi(s) = \sum_{a,s'} \pi(s,a) P^a_{s,s'}[R^a_{s,s'} + \pi(s',a')\gamma Q^\pi(s',a')]$

3. $Q^\pi(s,a) = \sum_{s'} \pi(s,a) P^a_{s,s'}[R^a_{s,s'} + \gamma V^\pi(s')]$

Proof. To see that the first claims holds, we use the explicit distribution of taking an action a when in state s and following π to see that indeed

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+k}|s_t = s] \\ &= \textstyle\sum_a \pi(s,a)\mathbb{E}_\pi[\sum_{k=0}^\infty \gamma^k r_{t+k}|s_t = s, a_t = a] \\ &= \textstyle\sum_a \pi(s,a) Q^\pi(s,a). \end{aligned}$$

For the second claim, following a similar line of argument as we have done for Proposition 1, we see that

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi[r_t|s_t = s] + \gamma\mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k}|s_t = s] \\ &= \textstyle\sum_a \sum_{s'} \pi(s,a) P^a_{s,s'} R^a_{s,s'} \\ &+ \textstyle\sum_a \sum_{s'} \pi(s,a) P^a_{s,s'} \gamma \sum_{a'} \pi(s,a')\mathbb{E}_\pi[\sum_{k=0}^\infty \gamma^k r_{t+1+k}|s_t = s, s_{t+1} = s', a_{t+1} = a'] \\ &= \textstyle\sum_a \sum_{s'} \pi(s,a) P^a_{s,s'} R^a_{s,s'} \\ &+ \textstyle\sum_a \sum_{s'} \pi(s,a) P^a_{s,s'} \gamma \sum_{a'} \pi(s,a')\mathbb{E}_\pi[\sum_{k=0}^\infty \gamma^k r_{t'+k}|s_{t'} = s', a_{t'} = a'] \\ &= \textstyle\sum_a \sum_{s'} \pi(s,a) P^a_{s,s'}(R^a_{s,s'} + \gamma \sum_{a'} \pi(s',a') Q^\pi(s',a')). \end{aligned}$$

The third equality uses the markov property. Lastly, we verify that

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi[r_t|s_t = s, a_t = a] + \gamma\mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k}|s_t = s, a_t = a] \\ &= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \mathbb{E}_\pi[\sum_{k=0}^\infty \gamma^k r_{r+1+k}|s_t = s, a_t = a, s_{t+1} = s'] \\ &= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \mathbb{E}_\pi[\sum_{k=0}^\infty \gamma^k r_{r'+k}|s'_t = s'] \\ &= \textstyle\sum_{s'} P^a_{s,s'}(R_{s,s'} + \gamma V^\pi(s')) \end{aligned}$$

completing the proof.

□

As before we give the more compact versions of these identities:

$$V^\pi(s) = \mathbb{E}_\pi[Q^\pi(s_t, a_t)|s_t = s],$$
$$V^\pi(s) = \mathbb{E}_\pi[r_t + \gamma Q^\pi(s_{t+1}, a_{t+1})|s_t = s]$$

and

$$Q^\pi(s, a) = \mathbb{E}[r_t + \gamma V^\pi(s_{t+1})|s_t = s, a_t = a].$$

Note that the last identity's expected value is *not* w.r.t π - that's because there is nothing random left to be determined according to π. The state s_t is given, the action a_t specified and fixed, and the expected value of the next state s_{t+1} is entirely dependent on how the environment reacts to this combination; and the term V^π is a deterministic function. This implies that policies sharing the same V also share the same Q. The reverse is not necessarily true. We formalize this realisation in the subsequent

Corollary 1. *Let π_1, π_2 be two arbitrary policies such that $V^{\pi_1} \equiv V^{\pi_2}$. Then $Q^{\pi_1} \equiv Q^{\pi_2}$*