A formal introduction to RL theory

Sebastian Scherer

September 30, 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. 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 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.

Finally, the most difficult notation to right *and* consistent: expected values. We will use slightly different notations to indicate the various different underlying distriutions that govern the behaviour of the random variables involved, and w.r.t which the expected value needs to be viewed.

If we are dealing with an implicit sequence of actions chosen according to one policy like

$$s_t \overset{\pi}{\rightarrow} a_t \overset{P^{a_t}_{s_t,\cdot}}{\rightarrow} s_{t+1} \overset{\pi}{\rightarrow} a_{t+1} \overset{P^{a_{t+1}}_{s_{t+1},\cdot}}{\rightarrow} s_{t+1} \overset{\pi}{\rightarrow} \dots,$$

we will express this by writing $\mathbb{E}_\pi[\cdot]$. The contribution of the environment's state distribution $P^{a_t}_{s_t,\cdot}$ is implicit since we usually deal with one finite MDP at a time, thereby keeping this particular distribution constant throughout all of the proofs. An example of this is

$$\mathbb{E}_\pi[r_t | s_t = s],$$

which implies action a_t was taken according to π having started at state s at time t.

Therefore, if the random variable in question is only dependent on the environment's distribution, such as in the expression

$$\mathbb{E}[r_t + \gamma f(s_{t+1}) | s_t = s, a_t = a]$$

(where f is some deterministic function) we omit any index. Note that in this case both the immediate reward r_t as well as the next time step's state s_{t+1} are entirely dependent on the environment parameters $R^a_{s,s'}$ and $P^a_{s,s'}$, since we fixed the action a_t, thereby cutting any potential policy out of the loop.

In some cases, we will need to indicate the involved random variables' distributions explicitly and individually. In those cases, we will make clear what exactly we mean.

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

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

the *value function* of π, and

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

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.

□

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

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

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

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 $Q^{\pi_1} \equiv Q^{\pi_2}$*

Proof. This is most easily seen in the original, parametrized formulation of Proposition 2, 3. . Since both $R_{s,s'}$ and $P^a_{s,s'}$ are dependent on the environment only, and not on the policy in question, we clearly have

$$\begin{aligned} Q^{\pi_1}(s, a) &= \textstyle\sum_{s'} P^a_{s,s'}[R^a_{s,s'} + \gamma V^{\pi_1}(s')] \\ &= \textstyle\sum_{s'} P^a_{s,s'}[R^a_{s,s'} + \gamma V^{\pi_2}(s')] \\ &= Q^{\pi_2}(s, a) \end{aligned}$$

for all $(s, a) \in S \times A$. □

Another useful result derived from the same identity is formalized in the below

Corollary 2. *Let π be a policy for a finite MDP, and let $(s, a) \sim P_{s,a}$ be a randomly distributed state-action pair. Then*

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

Proof. Since we are dealing with a finite MDP, both states and actions are drawn from a finite set S and A, respectively. We can therefore write

$$\begin{aligned} \mathbb{E}_{s,a\sim P_{s,a}}[Q^\pi(s, a)] &= \textstyle\sum_{s,a} P_{s,a} Q^\pi(s, a) \\ &= \textstyle\sum_{s,a} P_{s,a} \mathbb{E}[r_t + \gamma V^\pi(s_{t+1})|s_t = s, a_t = a] \\ &= \mathbb{E}_{s_t,a_t\sim P_{s,a}}[r_t + \gamma V^\pi(s_{t+1})]. \end{aligned}$$

□

Now that we are a bit more comfortable with the concept of an (action-) value function, we can use it as a tool to quantify the *quality* of a given policy. Intuitively, it makes sense to regard a policy π_1 that induces a higher expected reward when starting from a given state s than, say, another policy π_2 as 'better' - at least for that given state. In other words, it makes sense to regard π_1 as a better policy when starting from s than π_2, if and only if $V^{\pi_1}(s) > V^{\pi_2}(s)$. Expanding this intuitive measure of comparison beyond a single state s to *all* elements of S, we arrive at the following natural

Definition 1. *(Policy ranking) Let π_1, π_2 be policies for a finite MDP. We say that*

$$\pi_1 \geq_V \pi_2$$

if and only if $V^{\pi_1}(s) \geq V^{\pi_2}(s)$ for all $s \in S$. We say that

$$\pi_1 >_V \pi_2$$

if and only if $\pi_1 \geq_V \pi_2$ and there exists at least one $s \in S$ such that $V^{\pi_1}(s) > V^{\pi_2}(s)$. Finally, we say that

$$\pi_1 =_V \pi_2$$

if and only if $V^{\pi_u}(s) = V^{\pi_l}(s)$ for all $s \in S$.

This ranking of policies w.r.t $\geq_V$ induces a partial ordering on the set of policies Π. Note that it is possible that neither $\pi_1 \geq_V \pi_2$ nor $\pi_1 \leq_V \pi_2$ for a given pair of policies π_1, π_2, since we demand that one value function exceeds the other for *all* $s \in S$. In other words, $\geq_V$ really only is a *partial* ordering on the set of policies Π.

We sometimes informally refer to the relation $\pi_1 \geq_V \pi_2$ as 'π is *at least as good* as π_2', to $\pi_1 >_V \pi_2$ as 'π_1 is *better than / an actual improvement over* π_2' and to $\pi_1 =_V \pi_2$ as 'π_1 is *as good as / equally good as* π_2'.

We have, even at this early stage, established enough theory to characterize some cases where a direct comparison of policies w.r.t $\geq_V$ is possible.

Theorem 1. *(Policy improvement theorem) Let π_u, π_l be two different policies for a finite MDP such that*

$$\mathbb{E}_{a \sim \pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] \geq V^{\pi_l}(s)$$

for all $s \in S$. Then

$$\pi_u \geq_V \pi_l.$$

Before we begin the proof, let us formulate the above statement in a slightly less formal way. Our condition on π_u and π_l can be paraphrased as follows: If the expected reward generated by following π_u for *one* time step (note the expected value is indexed with π_u, indicating that the one remaining free random variable a_t is chosen according to π_u *conditioned*, i.e. fixed in its state variable, on the value of the state s) and then following π_l for all subsequent time steps is *always* (i.e. for every starting state s) greater or equal to the expected reward generated by following π_l *from the start*, then the policy π_u must be at least as good as π_l overall. In other words, if 'prepending' your actions with one action from a specified policy does not deteriorate rewards, the policy generating that one inserted action at the start of your journeys is at least as good as the policy being prepended. We will actually use this idea in an induction approach to show that, as we iteratively increase the number of time steps in which the actions are being chosen according to π_u before switching back to π_l, the expected reward does not decrease as well as converges to V^{π_u}.

Another thing to note is that, if the policy π_u is deterministic, our condition in the theorem reduces to

$$Q^{\pi_l}(s, \pi_u(s)) \geq V^{\pi_l}(s)$$

as the expected value of a constant random variable reduces to that constant value.

Proof. We first need to extend our assumption to the case where the state s appearing on both sides is not fixed, but more generally a random variable distributed according to, say, some distribution P_s. Since P_s is a distribution over finite states, we can see that indeed

$$\begin{aligned}\mathbb{E}_{s\sim P_s}[V^{\pi_l}(s)] &= \textstyle\sum_s P_s V^{\pi_l}(s)\\ &\leq \textstyle\sum_s P_s \mathbb{E}_{a\sim\pi_u(s,\cdot)}[Q^{\pi_l}(s,a)]\\ &= \textstyle\sum_s P_s \sum_a \pi_u(s,a) Q^{\pi_l}(s,a)\\ &= \mathbb{E}_{s_t,a_t\sim(P_{s_t},\pi_u(s_t,\cdot))}[Q^{\pi_l}(s_t,a_t)].\end{aligned}$$

Let $s \in S$ be arbitrary but fixed. We then see that, by our assumption, the definition of the value function V^π, and Proposition 2, 3., we have

$$\begin{aligned}V^{\pi_l}(s) &\leq \mathbb{E}_{\pi_u}[Q^{\pi_l}(s_t,a_t)|s_t=s]\\ &= \textstyle\sum_a \pi_u(s,a) Q^{\pi_l}(s,a)\\ &= \textstyle\sum_a \pi_u(s,a)\mathbb{E}[r_t+\gamma V^{\pi_l}(s_{t+1})|s_t=s,a_t=a]\\ &= \mathbb{E}_{\pi_u}[r_t+\gamma V^{\pi_l}(s_{t+1})|s_t=s]\\ &= \mathbb{E}_{\pi_u}[r_t|s_t=s]+\gamma\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+1})|s_t=s]\end{aligned}$$

Remeber that the index π_u denotes that any impliicit intermediate action a was taken according to π_u. Let further $s_{t+k} \sim P_s^{k*\pi_u}$ denote the distribution for the state at $t+k$ given that the state at time t was s and the subsequent k action(s) $a_t, \ldots, a_{t+k-1}$ were chosen according to π_u. Then we can apply our expected value version of the initial assumption to see that

$$\begin{aligned}&\mathbb{E}_{\pi_u}[r_t|s_t=s]+\gamma\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+1})|s_t=s]\\ =\ &\mathbb{E}_{\pi_u}[r_t|s_t=s]+\gamma\mathbb{E}_{s_{t+1}\sim P_s^{1*\pi_u}}[V^{\pi_l}(s_{t+1})]\\ \leq\ &\mathbb{E}_{\pi_u}[r_t|s_t=s]+\gamma\mathbb{E}_{s_{t+1},a_{t+1}\sim(P_s^{1*\pi_u},\pi_u(s_{t+1},\cdot))}[Q^{\pi_l}(s_{t+1},a_{t+1})]\end{aligned}$$

Applying Corollary 2 with $P_{s',a'} = (P_s^{1*\pi_u}, \pi_u(s',\cdot))$ we arrive at

$$\begin{aligned}&\mathbb{E}_{\pi_u}[r_t|s_t=s]+\gamma\mathbb{E}_{s_{t+1},a_{t+1}\sim(P_s^{1*\pi_u},\pi_u(s_{t+1},\cdot))}[Q^{\pi_l}(s_{t+1},a_{t+1})]\\ =\ &\mathbb{E}_{\pi_u}[r_t|s_t=s]+\gamma\mathbb{E}_{s_{t+1},a_{t+1}\sim(P_s^{1*\pi_u},\pi_u(s_{t+1},\cdot))}[r_{t+1}+\gamma V^{\pi_l}(s_{t+2})]\\ =\ &\mathbb{E}_{\pi}[r_t|s_t=s]+\gamma\mathbb{E}_{\pi_u}[r_{t+1}+\gamma V^{\pi_l}(s_{t+2})|s_t=s]\\ =\ &\mathbb{E}_{\pi_u}[r_t+\gamma r_{t+1}|s_t=s]+\gamma^2\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+2})|s_t=s].\end{aligned}$$

We do one more step to reiterate the feasability of this induction, and see that

$$\begin{aligned}&\mathbb{E}_{\pi_u}[r_t+\gamma r_{t+1}|s_t=s]+\gamma^2\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+2})|s_t=s]\\ &\mathbb{E}_{\pi_u}[r_t+\gamma r_{t+1}|s_t=s]+\gamma^2\mathbb{E}_{s_{t+2}\sim P_s^{2*\pi_u}}[V^{\pi_l}(s_{t+2})]\\ \leq\ &\mathbb{E}_{\pi_u}[r_t+\gamma r_{t+1}|s_t=s]+\gamma^2\mathbb{E}_{s_{t+2},a_{t+2}\sim(P_s^{2*\pi_u},\pi_u(s_{t+2},\cdot))}[Q^{\pi_l}(s_{t+2},a_{t+2})]\\ =\ &\mathbb{E}_{\pi_u}[r_t+\gamma r_{t+1}|s_t=s]+\gamma^2\mathbb{E}_{s_{t+2},a_{t+2}\sim(P_s^{2*\pi_u},\pi_u(s_{t+2},\cdot))}[r_{t+2}+\gamma V^{\pi_l}(s_{t+3})]\\ =\ &\mathbb{E}_{\pi_u}[r_t+\gamma r_{t+1}|s_t=s]+\gamma^2\mathbb{E}_{\pi_u}[r_{t+2}+\gamma V^{\pi_l}(s_{t+3})|s_t=s]\\ =\ &\mathbb{E}_{\pi_u}[r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}|s_t=s]+\gamma^3\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+3})|s_t=s]\end{aligned}$$

We have now established that, for any starting state s, it is more rewarding to take the first $n_a = 3$ actions according to π_u before switching to π_l, than it is to immediately

follow π_l from the starting state s. Repeating our argument infintely many times, i.e. letting $n_a \to \infty$, we see that

$$\begin{array}{lll} V^{\pi_l}(s) & \underset{n_a \overset{\to}{\to} \infty}{\leq} & \mathbb{E}_{\pi_u}[\sum_{k=0}^{n_a-1} \gamma^k r_{t+k} | s_t = s] \quad + \quad \gamma^{n_a} \mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+n_a}) | s_t = s] \\ & & V^{\pi_u}(s) \quad + \quad 0, \end{array}$$

where we have implicitly used that

$$0 \leq \gamma^{n_a} \mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+n_a}) | s_t = s] \leq \gamma^{n_a} \max_{s \in S} V^{\pi_l}(s) \overset{n_a \overset{\to}{\to} \infty}{\to} 0.$$

Informally speaking, taking an infinite number of actions according to π_u before switching to π_l essentially means simply following π_u and generates, independent of the starting state s, an expected cumulative reward that is at least as high as the one generated by simply following π_l. This proves the claim.

□

The above theorem states that $\pi_u \geq_V \pi_l$ holds whenever $\mathbb{E}_{a \sim \pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] \geq V^{\pi_l}(s)$ for all $s \in S$. Then $\pi_u \geq_V \pi_l$. Can we tweak these assumptions to guarantee actual improvement over π_l, i.e. $\pi_u >_V \pi_l$?

Theorem 2. *(Improved policy improvement theorem) Let π_u, π_l be two different policies for a finite MDP.*

1. *If*
$$\mathbb{E}_{a \sim \pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] = V^{\pi_l}(s)$$
for all $s \in S$, then
$$\pi_u =_V \pi_l.$$

2. *If*
$$\mathbb{E}_{a \sim \pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] \geq V^{\pi_l}(s)$$
for all $s \in S$, then
$$\pi_u \geq_V \pi_l.$$

3. *If*
$$\mathbb{E}_{a \sim \pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] \geq V^{\pi_l}(s)$$
for all $s \in S$, then
$$\pi_u >_V \pi_l$$
and there is at least one $s \in S$ such that
$$\mathbb{E}_{a \sim \pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] > V^{\pi_l}(s),$$
then
$$\pi_u >_V \pi_l.$$

Proof. Claim 2. is just the original Policy Improvement Theorem.

To see claim 1., simply replace the $\geq$ in the proof of the Policy Improvement Theorem with $=$. This equality persists through the induction step for all starting states s and yields equality of the respective value functions V^{π_u} and V^{π_l}.

To see the third claim, we again use the Policy Improvement Theorem to see that $\pi_u \geq_V \pi_l$. We now have to find at least one $s \in S$ satisfying $V^{\pi_u}(s) > V^{\pi_l}(s)$. Repeating the argument displayed in the proof of the Policy Comparison Theorem, we see that, applying it to the one state s guaranteed by our starting assumption to achieve $\mathbb{E}_{a\sim\pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] > V^{\pi_l}(s)$, the strict inequality persists throughout the induction step and really yields $V^{\pi_u}(s) > V^{\pi_l}(s)$. This shows $\pi_l u >_V \pi_l$ and completes the proof. □

We have now established some degree of comparability based on the policies' in question's V and Q functions. Before we can use this result to iteratively construct better and better policies, let us formalize the concept of a greedy policy.

Definition 2. *Let π be a policy for a finite MDP and let Q be some (possibly but not necessarily π's) action-value function . For any state $s \in S$, denote by*

$$maxact^Q(s) := \{a \in A | Q(s,a) = \max Q(s,a)\},$$

the set of actions at which $Q(s,\cdot)$ achieves its maximum. We say that π is Q-greedy if and only if for every $s \in S$ we have

$$\sum_{a\in A} \pi(s,a) = \sum_{a\in maxact^Q(s)} \pi(s,a) = 1$$

.

If a policy is greedy w.r.t its own Q function, we simply call it *greedy*. In other words, a *greedy* policy π only chooses among the actions that maximise its own Q^π-function for the given state. The only thing that matters for these policies when making a decision in state s on what to do next is whether that immediate action achieves a maximal 'immediate' pay-off of $\max_{a\in A} Q^\pi(s,a)$.

At first glance, such a policy might be somewhat short-sighted, seemingly ignoring potential future consquences of its immediate actions for the sake of instantaneous profit. However, the notion of the action-value function is to encode the future cumulative reward's expected value - in other words, it is a function that very much 'looks ahead' and considers future consequences; namely, all of them.

This means that a *greedy* policy might not be as shortsighted, and therefore not such a bad thing, after all. Indeed, we will later see that the best policies are exactly the ones that are *greedy*.

Let us therefore have a closer look at a greedy policy's value function.

Corollary 3. *(Greedy policy values) Let π_g be a greedy policy for a finite MDP. Then*

$$V^{\pi_g} \equiv max_{a\in A} Q^{\pi_g}(\cdot, a).$$

Proof. We write for any fixed but arbitrary $s \in S$

$$\begin{aligned} V^{\pi_g}(s) &= \sum_a \pi_g(s,a) Q^{\pi_g}(s,a) \\ &= \sum_{a \in maxact^{Q^{\pi_g}}} \pi_g(s,a) Q^{\pi_g}(s,a) \\ &= \max_{a \in A} Q^{\pi_g}(s,a). \end{aligned}$$

□

In other words, a greedy policy's value function is just the maximum of its action-value function at the present state, taken over all possible actions. That is not surprising, seeing as maximising its own Q function is what is driving a greedy policy as per definition.

The following result shows that any group of policies greedy w.r.t some Q function are of the same quality. Unsurprisingly, the identical strategy of local maximisation does not lead to great deal of variability in policies' performances. It also shows that any policy that is not *greedy* can be improved by making it *greedy*.

Lemma 1. *(Greedy policy improvement) Let π_g and π_c be policies for finite MDP, and let π_g be Q^{π_c}-greedy. If π_c is Q^{π_c}-greedy, too, then $\pi_g =_V \pi_c$. Otherwise, $\pi_g >_V \pi_c$.*

Proof. Let $s \in S$ be fixed but arbitrary. We write

$$\begin{aligned} \mathbb{E}_{a \sim \pi_g(s,\cdot)}[Q^{\pi_c}(s,a)] &= \sum_a \pi_g(s,a) Q^{\pi_c}(s,a) \\ &= \sum_{a \in maxact^{\pi_c}(s)} \pi_g(s,a) \max_{a \in A} Q^{\pi_c}(s,a) \\ &= \max_{a \in A} Q^{\pi_c}(s,a) \\ &\geq \sum_{a \in A} \pi_c(s,a) Q^{\pi_c}(s,a) \\ &= V^{\pi_c}(s). \end{aligned}$$

If π_c is Q^{π_c}-*greedy*, then we have equality in the above chain for all $s \in S$ and the Improved Policy Improvement Theorem, 1., implies $\pi_g =_V \pi_c$. If π_c is not Q^{π_c}-*greedy*, then there is at least one $s \in S$ such that $>$ holds (and still $\geq$ for all other s). In that case the Improved Policy Improvement Theorem, 3., yields the desired claim. □

Inspecting the above proof we notice an important detail: Any deterministic policy that chooses an action from $maxact^{\pi_c}(s)$ given any state s satisfies the condition of the corollary and thus is at least as good as (the potentially probabilistic) π_c. This means that for every probabilistic policy there is a deterministic policy that is at least as good, and it can be constructed explicitly by letting it choose any one action that maximises the probabilistic policy's action-value function.

4 Optimal policies and how to find them

In this section we will use our concept of policy comparison to formalize and analyse the concept of a *best* policy. We will investigate existence, uniqueness and characterizations of such policies. We will mainly build on the previous sections results, but will at a later point be forced to introduce some usefull tools from functional analysis.

With this road map in mind, let us start our journey ny clarifying what we mean by an *optimal* or *best* policy.

Definition 3. *(Optimal policy) Let π^* be a policy for a finite MDP. We call π^* an optimal policy if and only if we have*

$$\pi^* \geq_V \pi$$

for all policies $\pi \in \Pi$.

In other words, a policy π^* whose value function V^{π^*} dominates the value functions V^π of all other policies is optimal for the given finite MDP, and what we consider 'best'. This approach seems sensible, since optimality of a policy according to the above definition maximises the expected cumulative reward (that is what a policy's value function represents).

It is fair to say that most, if not all, of the subsequent results presented in this script deal with the analysis of optimal policies. In particular, we will formally answer the following questions:

- Is there always a (unique) optimal policy for a finite MDP?
- Are there any characteristic traits that all optimal policies share, and if so, how can we make use of them to find these policies?
- Can we give a constructive way of finding or at least approximating these optimal policies?

We first turn to the question of existence. Using a result that is equivalent to the infamous *Axiom of Choice*, we are able to derive the following

Proposition 3. *(Optimal existence) For each finite MDP there exists at least one optimal (deterministic) policy.*

Proof. We will use Zorn's Lemma to prove the existence of an optimal deterministic policy, and then use results derived in the previous section to see that no probabilistic policy can outperform such a policy.

Zorn's Lemma states that any partially ordered set $(S, \leq_S)$ in which every chain (i.e. set of elements $C \subset S$ that can be ordered such that $e_1 \leq_S e_2 \leq_S ...$) has an upper bound in S (i.e. an $e_{upper} \in S$ such that $e \leq_S e_{upper}$ for all $e \in S$) has a maximal element (i.e. an $e_{max_S} \in S$ such that if $e_{max_S} \leq e$ for some $e \in S$ then $e = e_{max_S}$ must be true).

We will define our partially ordered set as follows. Consider all value functions derived from deterministic policies, with the partial order

$$V^{\pi_1} \leq V^{\pi_2} \Leftrightarrow \pi_1 \leq_V \pi_2.$$

Since for any finite MDP, there is only a finite number of deterministic policies, any chain in this partially ordered set is finite, and therefore has an upper bound - namely the last element of the chain. Zorn's Lemma now implies the existence of a deterministic policy π^* and corresponding value function such that for any other deterministic policy π with value function V^π achieving $\pi^* \leq_V \pi$ then $V^\pi \equiv V^{\pi^*}$. $\square$