A formal introduction to RL theory

Sebastian Scherer

November 12, 2019

Contents

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. 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,

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

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

$$P^a_{s,s'} = Pr(s_{t+1} = s' | s_t = s, a_t = a) \tag{2}$$

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

$$R^a_{s,s'} = \mathbb{E}[r_t | s_t = s, s_{t+1} = s', a_t = a] \tag{3}$$

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. For the remainder of this script, we will assume a uniform bound on rewards throughout all time steps, i.e. that

$$-M \leqslant r_t \leqslant M \tag{4}$$

for some $M \in \mathbb{R}$ and each $t \in 0, 1, 2, \ldots$.

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

$$s_t \xrightarrow{\pi} a_t \overset{P^{a_t}_{s_t,\cdot}}{\rightarrow} s_{t+1} \xrightarrow{\pi} a_{t+1} \overset{P^{a_{t+1}}_{s_{t+1},\cdot}}{\rightarrow} s_{t+1} \xrightarrow{\pi} \ldots, \tag{5}$$

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

$$\begin{array}{rccl} V^\pi : & S & \to & \mathbb{R} \\ & s & \mapsto & \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s] \end{array} \tag{6}$$

the *value function* of π, and

$$\begin{array}{rccl} Q^\pi: & A \times S & \to & \mathbb{R} \\ & (a,s) & \mapsto & \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a] \end{array} \tag{7}$$

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.

3.1 Existential crisis

In line with our rigorous approach so far, we dedicate this subsection to proving the well definedness of the above quantities. That is, we get assurances for:

- the well definedness of the random variable defined by the series $G_t := \sum_{k=0}^{\infty} \gamma^k r_{t+k}$
- the well definedness of its expected value, i.e. the well definedness of V^π
- the well definedness of Q^π.
- some handy derivative results from the above three

We start our quest with guaranteeing the well definedness of G_t. As a random variable that is the discounted infinite sum of random variables r_{t+k}, $k = 0, 1, \dots$, it maps sequences $\omega = (s_t, a_t, s_{t+1}, a_{t+1}, \dots)$, with $s_t, s_{t+1}, \dots S$, $a_t, a_{t+1}, \dots A$ onto the set of real numbers $\mathbb{R}$. In other words, by denoting Ω as the set of all such sequences ω, we consider G_t to be

$$\begin{array}{rccl} G_t: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & \sum_{k=0}^{\infty} \gamma^k r_{t+k}(\omega) \end{array} \tag{8}$$

where $\sum_{k=0}^{\infty} \gamma^k r_{t+k}(\omega) \overset{!}{\in} \mathbb{R}$ remains to be shown to guarantee well definedness of this quantity as a (random variable) mapping. We will do so in the following

Proposition 1. *(Cumulative rewards exist) The random variable G_t as defined in 8 is well defined. In other words, the series $G_t(\omega) = \sum_{k=0}^{\infty} \gamma^k r_{t+k}(\omega)$ is convergent for all $\omega \in \Omega$.*

Proof. We will show that for any trajectory $\omega \in \Omega$ the sequence $G_t^n(\omega) := \sum_{k=0}^{n} \gamma^k r_{t+k}(\omega) \in \mathbb{R}$ is a cauchy sequence. Since $\mathbb{R}$ is complete, this will imply convergence of $G_t^n(\omega)$ in $\mathbb{R}$.

Let w.l.o.g $n > m$ to see that

$$|G_t^n(\omega) - G_t^n(\omega)| = \left| \sum_{k=m}^{n} \gamma^k r_{t+k}(\omega) \right| \overset{4}{\leqslant} M \sum_{k=m}^{n} \gamma^k. \tag{9}$$

Since $s_n := \sum_{k=0}^{n} \gamma^k \overset{n\to\infty}{\to} \frac{1}{1-\gamma}$ and therefore cauchy, there exists for every $\hat{\epsilon}$ an $N_{\hat{\epsilon}}$ such that, if, $n, m \geqslant N_{\hat{\epsilon}}$, $\sum_{k=m}^{n} \gamma^k = |s_n - s_m| < \hat{\epsilon}$. If now $\epsilon > 0$ is arbitrary but fixed, choose $\hat{\epsilon} = \frac{\epsilon}{M}$. Thus, for any $n, m > N_{\hat{\epsilon}}$, we have

$$|G_t^n(\omega) - G_t^n(\omega)| \overset{9}{\leqslant} M|s_n - s_m| < M\frac{\epsilon}{M} = \epsilon. \tag{10}$$

Thus for any $\omega \in \Omega$, the sequence $G_t^n(\omega)$ is cauchy, and thus convergent on $\mathbb{R}$. This proves the claim.

□

Having established $G_t(\omega) \in \mathbb{R}$, we can immediately see that it must be uniformly bounded on Ω.

Corollary 1. *(Cumulative rewards are bounded) The cumulative reward G_t as defined in 8 is uniformly bounded on Ω. In particular,*

$$|G_t(\omega)| < \frac{M}{1-\gamma} \tag{11}$$

for all $\omega \in \Omega$.

Proof. Thanks to Prop 1 we can justify writing

$$|G_t(\omega)| = \left|\sum_{k=0}^{\infty} \gamma^k r_{t+k}(\omega)\right| \leqslant \sum_{k=0}^{\infty} \gamma^k |r_{t+k}(\omega)| \overset{4}{\leqslant} M \sum_{k=0}^{\infty} \gamma^k = \frac{M}{1-\gamma}. \tag{12}$$

□

Now that the infinite discounted reward is a well defined random variable, we can sensibly pose the question regarding its expected value. Before we do so, however, let us have a look at the trajectory space Ω and consider why it is a countable, discrete set.

Let us denote by $c_S = |S|, c_A = |A|$ the cardinality, i.e. size, of the state space S and action space A, respectively. We can then order all possible trajectories in

$$\Omega = \{\omega = (s_1, a_2, s_2, a_2, \dots) | s_i \in S, a_i \in A, i = 1, 2, \dots\}. \tag{13}$$

To see such an ordering, refer to the appendix (? - still have to write this part). In particular, whenever we write ω_i for $i = 1, 2, \dots$ or similar we mean the ordered $\omega \in \Omega$.

We now establish the existence of the expected value of G_t.

Lemma 1. *(Expected cumulative rewards exist, too) Let P_Ω be any trajectory distribution on the trajectory space Ω. Then the expected value of the cumulative reward G_t under that distribution is well defined and exists. In other words $\mathbb{E}_{\omega\sim P_\Omega}[G_t] < \infty$.*

Note that this is a pretty general statement about sufficient conditions for the existence of G_t's expected value. It includes, for example, the cases were we act accoding to some policy π, starting from some random state s distributed according to some random distribution P_S on S. In particular, if that distribution is deterministic, i.e. if

$$P_{S,s'}(s) = \begin{cases} 1 & \text{for } s = s' \\ 0 & \text{else} \end{cases} \tag{14}$$

we will have established the well-definedness of V_π as defined in Definition 6, since that is what the value function, according to 6, measures. To see this, note that in such a case, $P_\Omega = P^\pi_{\Omega|s'} = P^\pi \circ P_{S,s'}$ is the distribution resulting from sampling a starting state s according to $P_{S,s'}$, and then, given such a starting state, sampling a trajectory ω according to π. In such a case we have

$$V^\pi(s') \overset{Def6}{:=} \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s \right] = \mathbb{E}_{P_\Omega}[G_t] \overset{Lem1}{<} \infty \tag{15}$$

The analogous argument guaranteeing the well definedness of our action-value function Q^π as defined in Definition 7, uses a deterministic distribution $P_{S\times A,(s',a')}$ such that, for a fixed but arbitrarty pair $(s', a') \in S \times A$,

$$P_{S\times A,(s',a')}(s,a) = \begin{cases} 1 & \text{for } (s,a) = (s',a') \\ 0 & \text{else.} \end{cases} \tag{16}$$

This leads to a composite trajectory distribution $P_\Omega = P^\pi_{\Omega|s',a'} = P^\pi \circ P_{S,(s',a')}$ resulting from sampling a starting state *and* action (s, a) according to $P_{S,(s',a')}$, and then, given such a starting state-action pair, sampling a trajectory ω according to π. This yields

$$Q^\pi(s',a') \overset{Def7}{:=} \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s', a_t = a' \right] = \mathbb{E}_{P_\Omega}[G_t] \overset{Lem1}{<} \infty. \tag{17}$$

For easier reference later on, let's summarize these arguments in the form of the following

Corollary 2. *((Action-)Value functions exist) Let π be probabilistic or deterministic policy on a finite MDP. Then the functions V^π as defined in 6 and Q^π as defined in 7 are well defined.*

Proof. For a proof of both claims, see the argumentats presented in the paragraph preceding this corollary. In particular, for the claim about V^π see the derivation of equations 14 and 15. Similarly, the well definedness of Q^π is justified by the arguments formalized in equations 16 and 17. □

Having established the usefulness of this Lemma, let us finally prove it.

Proof. (of Lemma 1) Given some trajectory distribution P_Ω on the (unordered) set Ω, we need to show convergence of the series

$$\mathbb{E}_{\omega\sim P_\Omega}[G_t] = \sum_{\omega\in\Omega} P_\Omega(\omega) G_t(\omega) < \infty \tag{18}$$

regardless of the particular ordering of the trajectories ω. We do this by showing the convergence of this series' absolute series counterpart, namely

$$\sum_{\omega\in\Omega} |P_\Omega(\omega)G_t(\omega)| \overset{P_\Omega(\omega)\geqslant 0}{=} \sum_{\omega\in\Omega} P_\Omega(\omega)|G_t(\omega)| = \sum_{i=1}^{\infty} P_\Omega(\omega_i)|G_t(\omega_i)| < \infty \tag{19}$$

for the *one particular ordering* of Ω referred to earlier. The convergence of the absolute series in 19 for one possible ordering of $\omega \in \Omega$ then guarantees the convergence of the series in 18 for *any* ordering of $\omega \in \Omega$ to the same limit according to the "Famous Reordering Series Convergence Theorem". To show that the absolute series converges, we show that the sequence

$$EG^n_{abs} := \sum_{i=1}^{n} |G_t(\omega_i)|P_\Omega(\omega_i) \overset{Prop1}{<} \infty \tag{20}$$

induced by its partial sum is a Cauchy sequence. W.lo.g, let $n > m$ to see that

$$\begin{array}{rcl} |EG^n_{abs} - EG^m_{abs}| & = & \sum_{i=m}^{n} |G_t(\omega_i)|P_\Omega(\omega_i) \\ & \overset{Cor1}{\leqslant} & \frac{M}{1-\gamma}\sum_{i=m}^{n} P_\Omega(\omega_i). \end{array} \tag{21}$$

Since $\sum_{i=0}^{\infty} P_\Omega(\omega_i) = 1$, the sequence of partial sums

$$\hat{s}_n := \sum_{i=0}^{n} P_\Omega(\omega_i) \tag{22}$$

is necessarily convergent and thus a Cauchy sequence. This means that for any $\hat{\epsilon} > 0$, we can find an $N_{\hat{\epsilon}} \in \mathbb{N}$ such that for all $n > m > N_{\hat{\epsilon}}$ we have

$$\sum_{i=m}^{n} P_\Omega(\omega_i) \overset{P_\Omega(\omega_i)\geqslant 0}{=} |\hat{s}_n - \hat{s}_m| < \hat{\epsilon}. \tag{23}$$

Let $\epsilon > 0$ be arbitrary but fixed. Choosing $\hat{\epsilon} = \epsilon\frac{1-\gamma}{M}$ and $n > m > N_{\hat{\epsilon}}$, we indeed see that

$$|EG^n_{abs} - EG^m_{abs}| \overset{21}{\leqslant} \frac{M}{1-\gamma}\sum_{i=m}^{n} P_\Omega(\omega_i) \overset{23}{\leqslant} \frac{M}{1-\gamma}\hat{\epsilon} = \epsilon, \tag{24}$$

proving that $(EG^n_{abs})_n$ is indeed Cauchy and thus convergent in $\mathbb{R}$. This proves 19 for the one ordering of Ω constructed, and thus 18 for *any* ordering of $\omega \in \Omega$. This proves the Lemma.

□

In a similar vein, the expectations of the individual discounted rewards also exist for any finite MDP.

Lemma 2. *Let $k \in \mathbb{N}$ be fixed but arbitrary. Then the expectation of the discounted reward at time step $t+k$, where t is the starting point in time for all trajectories considered, exist and are bounded by M. The same holds for the absolute value of the discounted reward at time step $t+k$. In other words, we have both $\mathbb{E}_{\omega\sim P_\Omega}[\gamma^k r_{t+k}] \leqslant \gamma^k M$ and $\mathbb{E}_{\omega\sim P_\Omega}[|\gamma^k r_{t+k}|] \leqslant \gamma^k M$.*

Proof. The proof for either claim is nearly identical to arguments presented in the proof of Lemma 1. To avoid repetition we point out the two differences: instead of $G_t(\omega)$, we use either $\gamma^k r_{t+k}$ or $|\gamma^k r_{t+k}| = \gamma^k |r_{t+k}|$ depending on which of the two claims we want to prove. Secondly, the inequality in 21 is now justified by the assumption of universal boundedness 4, and the bound $\frac{M}{1-\gamma}$ is to be replaced by $\gamma^k M$. Virtually every other argument can be copy pasted to prove both $\mathbb{E}_{\omega\sim P_\Omega}[\gamma^k r_{t+k}] \leqslant \gamma^k M$ and $\mathbb{E}_{\omega\sim P_\Omega}[|\gamma^k r_{t+k}|] \leqslant \gamma^k M$. □

Before concluding this existential episode, let us present one more result and its application to our value function setting.

Lemma 3. *(Infinite linearity) Let P_Ω be a distribution on the discrete trajectory space Ω. Then*

$$\mathbb{E}_{\omega\sim P_\Omega}\left[G_t(\omega)\right] = \lim_{n\to\infty} \mathbb{E}_{\omega\sim P_\Omega}\left[\sum_{k=0}^{n} \gamma^k r_{t+k}\right] \tag{25}$$

Proof. The first step of the proof is to reduce the claim to that of linearity of the expectation for infinite random series, i.e. to reduce our claim to

$$\mathbb{E}_{\omega\sim P_\Omega}\left[\sum_{k=0}^{\infty} X_k\right] \stackrel{!}{=} \sum_{k=0}^{\infty} \mathbb{E}_{\omega\sim P_\Omega}\left[X_k\right] \tag{26}$$

and then refer to a standard argument in probability theory. To see that claim 25 can be rephrased in the form of 26, set $X_k := \gamma^k r_{t+k}$ for $k = 0, 1, \ldots,$ giving equality of the respective left hand sides. Further, note that with these X_k and the fact that expectations are *always* linear for *finite* sums of random variables, we can see that the right hand side of 25 can be rewritten like

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}_{\omega\sim P_\Omega}\left[\sum_{k=0}^{n} \gamma^k r_{t+k}\right] &= \lim_{n\to\infty} \mathbb{E}_{\omega\sim P_\Omega}\left[\sum_{k=0}^{n} X_k\right] \\ &= \lim_{n\to\infty} \sum_{k=0}^{n} \mathbb{E}_{\omega\sim P_\Omega}\left[X_k\right] \\ &= \sum_{k=0}^{\infty} \mathbb{E}_{\omega\sim P_\Omega}\left[X_k\right], \end{aligned} \tag{27}$$

showing the claimed equivalence of both statements. To prove our modified claim 26, a standard probability theory argument states that all we require is for

$$\sum_{k=0}^{\infty} \mathbb{E}_{\omega\sim P_\Omega}\left[|X_i|\right] \stackrel{!}{<} \infty \tag{28}$$

to hold. This of course is equivalent to showing

$$\sum_{k=0}^{\infty} \mathbb{E}_{\omega\sim P_\Omega}\left[\gamma^k |r_{t+k}|\right] \stackrel{!}{<} \infty, \tag{29}$$

where we already exploited $\gamma > 0$. We can quickly verify that indeed

$$\begin{array}{rcl} \sum_{k=0}^{n} \mathbb{E}_{\omega \sim P_\Omega} \left[\gamma^k |r_{t+k}| \right] & \overset{Lem2}{\leqslant} & \sum_{k=0}^{n} \gamma^k M \\ & < & \sum_{k=0}^{\infty} \gamma^k M \\ & = & \frac{M}{1-\gamma}. \end{array} \tag{30}$$

Letting $n \to \infty$ on either side, we clearly have 29, thus proving the required condition and therefore proving the original claim 25. □

A useful consequence of this result is an alternative expression for our (action-) value functions V^π and Q^π for some policy π on a finite MDP.

Corollary 3. *(Limit your expectations) Let π be a policy on a finite MDP. Then*

1.

$$V^\pi(s) = \lim_{n \to \infty} \mathbb{E}_\pi \left[\sum_{k=0}^{n} \gamma^k r_{t+k} | s_t = s \right] \tag{31}$$

2.

$$Q^\pi(s,a) = \lim_{n \to \infty} \mathbb{E}_\pi \left[\sum_{k=0}^{n} \gamma^k r_{t+k} | s_t = s, a_t = a \right] \tag{32}$$

Proof. The proof for either identity goes all the way back (see the section surrounding equations 14, 15, 16 and 17) to the realisation that fixing the starting state s_t and then following a given policy π is inducing a distribution $P_\Omega = P^\pi_{\Omega|s'} = P^\pi \circ P_{S,s'}$ on the trajectory space Ω. We can write

$$\begin{array}{rcl} V^\pi(s) & = & \mathbb{E}_\pi \left[G_t | s_t = s \right] \\ & = & \mathbb{E}_{\omega \sim P_\Omega} \left[G_t \right] \\ & \overset{Lem3}{=} & \lim_{n \to \infty} \mathbb{E}_{\omega \sim P_\Omega} \left[\sum_{k=0}^{n} \gamma^k r_{t+k} \right] \\ & = & \lim_{n \to \infty} \mathbb{E}_\pi \left[\sum_{k=0}^{n} \gamma^k r_{t+k} | s_t = s \right], \end{array} \tag{33}$$

proving 31. Equivalently, fixing a starting state s_t *and* starting action a_t and then following a given policy π induces a distribution $P_\Omega = P^\pi_{\Omega|s',a'} = P^\pi \circ P_{S,(s',a')}$. This allows us to write

$$\begin{array}{rcl} Q^\pi(s,a) & = & \mathbb{E}_\pi \left[G_t | s_t = s, a_t = a \right] \\ & = & \mathbb{E}_{\omega \sim P_\Omega} \left[G_t \right] \\ & \overset{Lem3}{=} & \lim_{n \to \infty} \mathbb{E}_{\omega \sim P_\Omega} \left[\sum_{k=0}^{n} \gamma^k r_{t+k} \right] \\ & = & \lim_{n \to \infty} \mathbb{E}_\pi \left[\sum_{k=0}^{n} \gamma^k r_{t+k} | s_t = s, a_t = a \right], \end{array} \tag{34}$$

proving 32.

□

3.2 Relationships

This subsection is dedicated to highlighting important and useful relationships between value and action-value functions. We will have a look at expressing one via the other, as well as recursive identities across time steps t.

Due to the heavy lifting done in the previous subsection as well as the inherently local (w.r.t time t, that is) behaviour of (finite) MDP, all probabilistic considerations in this section reduce to finite event spaces. In particular, instead of dealing with infinite but discrete trajectories $\omega = (s_t, a_t, s_{t+1}, a_{t+1}, \dots)$, we can focus on a finite set of state or action random variables $(s_t, a_t), \dots, (s_{t+k}, a_{t+k})$, where k will depend on the given context, but most of the times will not exceed 1. All expected values are therefore to be understood with respect to the finite distributions on these finite state-action space tuples resulting from following some specified policy π, sometimes with additional constraints on starting states or actions s_t, a_t. While the underlying trajectory distributions P_Ω as outlined in the section surrounding equations **??**, **??**, **??** and **??** technically still apply, the properties defining a (finite) MDP as well as our previous work allow us to break free from this slightly technical "underground" world and instead work in the "above-ground" and easy-to-handle environment of finite state-action distributions.

The connection between these two worlds, in particular the transformation from trajectories ω to individual states s_{t+k} and actions a_{t+k}, can be done by effectively grouping trajectories' probabilities that share, say, a state s at a fixed point in time $t + k$. To be precise, for some distribution P_Ω on Ω, we can generate a distributions $P_{P_\Omega,S}$ for s_{t+k} on S by putting

$$P_{P_\Omega,S}(s_{t+k} = s) := \sum_{\omega \in \Omega_s^{t+k}} P_\Omega(\omega), \tag{35}$$

where $\Omega_s^{t+k} := \{\omega = (s_t, a_t, s_{t+1}, a_{t+1}, \dots) \in \Omega | s_{t+k} = s\}$. Similar constructions can be made to generate distributions over the action space A for the random variable a_{t+k} representing the action chosen at time $t + k$, and so on. In this sense, any policy π (with optional constraints on states and/or actions at some fixed point(s) in time) generates some P_Ω on Ω, which in turn generates a host of distributions on S and A, for individual states and actions, respectively. It is in this way that the following results can be connected to the results in the previous section. Personally, I found it helpful to keep in mind the concept of an "original source" distribution P_Ω when dealing with different policies' behaviour and relative performance.

Having said all that, let us have a closer look at the recursive nature of both V^π and Q^π - a feature we will exploit throughout the rest of this transcript.

Proposition 2. *(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

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+k} | s_t = s] \\
&= \mathbb{E}_\pi[r_t + \gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s] \\
&= \mathbb{E}_\pi[r_t | s_t = s] + \mathbb{E}_\pi[\gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s] \\
&\overset{2,3}{=} \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s] \\
&= \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_{t+1} = s'] \\
&= \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} (R^a_{s,s'} + \gamma \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t'+k} | s_{t'} = s']) \\
&= \textstyle\sum_a \pi(s,a) \sum_{s'} P^a_{s,s'} (R^a_{s,s'} + \gamma V^\pi(s')).
\end{aligned}$$

Note how we used the transitional probabilities and expected rewards to explicitly write out the expected value of the cumulative reward $\sum_{k=0}^\infty \gamma^k r_{t+k}$ established in Lemma 1. Note also how we wrote $\mathbb{E}_\pi$ to indicate that the expected value is w.r.t to a sequence of actions $a_t, a_{t+1}, \ldots$ and states $s_t (= s$, as per explicit condition "$|s_t = s$"$), s_{t+1}, \ldots$ 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

$$\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+k} | s_t = s, a_t = a] \\
&= \mathbb{E}_\pi[r_t + \gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s, a_t = a] \\
&= \mathbb{E}_\pi[r_t | s_t = s, a_t = a] + \gamma \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s, a_t = a] \\
&\overset{2,3}{=} \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} \\
&+ \gamma \textstyle\sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_t = s, a_t = a, s_{t+1} = s', a_{t+1} = a'] \\
&= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t+1+k} | s_{t+1} = s', a_{t+1} = a'] \\
&= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') \mathbb{E}_\pi[\textstyle\sum_{k=0}^\infty \gamma^k r_{t'+k} | s_{t'} = s', a_{t'} = a'] \\
&= \textstyle\sum_{s'} P^a_{s,s'} R^a_{s,s'} + \gamma \sum_{s'} P^a_{s,s'} \sum_{a'} \pi(s,a') Q^\pi(s',a') \\
&= \textstyle\sum_{s'} P^a_{s,s'} (R^a_{s,s'} + \gamma \sum_{a'} \pi(s,a') Q^\pi(s',a'))
\end{aligned}$$

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:

$$V^\pi(s) = \mathbb{E}_\pi[r_t + \gamma V^\pi(s_{t+1}) | s_t = s] \tag{36}$$

and

$$Q^\pi(s) = \mathbb{E}_\pi[r_t + \gamma Q^\pi(s_{t+1}, a_{t+1}) | s_t = s, a_t = a]. \tag{37}$$

These are the 'standard' bellman equations.

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

Proposition 3. *(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 4. *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 5. *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

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

□

3.3 Comparing and improving

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*

$$\pi_1 \geqslant_V \pi_2$$

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

$$\pi_1 >_V \pi_2$$

if and only if $\pi_1 \geqslant_V \pi_2$ *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 $\geqslant_V$ induces a partial ordering on the set of policies Π. Note that it is possible that neither $\pi_1 \geqslant_V \pi_2$ nor $\pi_1 \leqslant_V \pi_2$ 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, $\geqslant_V$ really only is a *partial* ordering on the set of policies Π.

We sometimes informally refer to the relation $\pi_1 \geqslant_V \pi_2$ 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 $\geqslant_V$ is possible.

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

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

for all $s \in S$. *Then*

$$\pi_u \geqslant_V \pi_l.$$

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

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

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

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

The second and main part of the proof will consist of showing the aforementioned policy improvement via prepending π_l with π_u. Formally, this means that for any $s \in S$, $n = 1, 2, ...$ the inequality

$$V^{\pi_l}(s) \overset{!}{\leqslant} \mathbb{E}_{\pi_u}\left[\sum_{k=0}^{n-1} \gamma^k r_{t+k} | s_t = s\right] + \gamma^n \mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+n})|s_t = s\Big]$$

holds. We will show this claim by induction over n.

For the induction start, 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

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

proving the claim for $n = 1$. Remember that the index π_u denotes that any implicit intermediate action a was taken according to π_u.

For the induction step, let us introduce some additional notation. Let $s_{t+k} \sim P_s^{k*\pi_u}$ 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

$$\begin{aligned}V^{\pi_l}(s) &\overset{I.S.}{\leqslant} \mathbb{E}_{\pi_u}[\textstyle\sum_{k=0}^{n-2} \gamma^k r_{t+k}|s_t = s] + \gamma^{n-1}\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+n-1})|s_t = s]\\ &= \mathbb{E}_{\pi_u}[\textstyle\sum_{k=0}^{n-2} \gamma^k r_{t+k}|s_t = s] + \gamma^{n-1}\mathbb{E}_{s_{t+n-1}\sim P_s^{(n-1)*\pi_u}}[V^{\pi_l}(s_{t+n-1})]\\ &\leqslant \mathbb{E}_{\pi_u}[\textstyle\sum_{k=0}^{n-2} \gamma^k r_{t+k}|s_t = s]\\ &\quad +\gamma^{n-1}\mathbb{E}_{(s_{t+n-1},a_{t+n-1})\sim(P_s^{(n-1)*\pi_u},\pi_u(s_{t+n-1},\cdot))}[Q^{\pi_l}(s_{t+n-1},a_{t+n-1})]\end{aligned}$$

Applying Corollary 2 with $P_{s',a'} = (P_s^{1*\pi_u}, \pi_u(s',\cdot))$ to the last expression in the above sequence, we arrive at

$$
\begin{aligned}
V^{\pi_l}(s) &\leqslant \mathbb{E}_{\pi_u}[\textstyle\sum_{k=0}^{n-2}\gamma^k r_{t+k}|s_t=s] \\
&\quad +\gamma^{n-1}\mathbb{E}_{(s_{t+n-1},a_{t+n-1})\sim(P_s^{(n-1)*\pi_u},\pi_u(s_{t+n-1},\cdot))}[r_{t+n-1}+\gamma V^{\pi_l}(s_{t+n})] \\
&= \mathbb{E}_{\pi_u}[\textstyle\sum_{k=0}^{n-2}\gamma^k r_{t+k}|s_t=s] \\
&\quad +\gamma^{n-1}\mathbb{E}_{\pi_u}[r_{t+n-1}+\gamma V^{\pi_l}(s_{t+n})|s_t=s] \\
&= \mathbb{E}_{\pi_u}[\textstyle\sum_{k=0}^{n-2}\gamma^k r_{t+k}|s_t=s]+\mathbb{E}_{\pi_u}[\gamma^{n-1}r_{t+n-1}|s_t=s] \\
&\quad +\gamma^{n}\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+n})|s_t=s] \\
&= \mathbb{E}_{\pi_u}[\textstyle\sum_{k=0}^{n-1}\gamma^k r_{t+k}|s_t=s]+\gamma^{n}\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+n})|s_t=s],
\end{aligned}
$$

completing the induction step and proving the claim. Letting $n \to \infty$, we see that

$$
\begin{array}{rcccl}
V^{\pi_l}(s) & \underset{n\to\infty}{\leqslant} & \mathbb{E}_{\pi_u}[\sum_{k=0}^{n}\gamma^k r_{t+k}|s_t=s] & + & \gamma^n\mathbb{E}_{\pi_u}[V^{\pi_l}(s_{t+n})|s_t=s] \\
 & \underset{\to}{} & V^{\pi_u}(s) & + & 0.
\end{array}
$$

For the second term, we have implicitly used that

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

The convergence claim regarding the first term was proven in Corollary 3, 31.

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 theorem.

□

The above theorem states that $\pi_u \geqslant_V \pi_l$ holds whenever $\mathbb{E}_{a\sim\pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] \geqslant V^{\pi_l}(s)$ for all $s \in S$, then $\pi_u \geqslant_V \pi_l$. 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*
$$\mathbb{E}_{a\sim\pi_u(s,\cdot)}[Q^{\pi_l}(s,a)] \geqslant V^{\pi_l}(s)$$
for all $s \in S$, then
$$\pi_u \geqslant_V \pi_l.$$

3. If

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

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 $\geqslant$ 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 $\pi_u \geqslant_V \pi_l$. 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 6. *(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 $s \in S$

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

□

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 4. *(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

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

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 $\geqslant$ 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 useful tools from functional analysis.

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?

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

4.1 Existential crisis

Let us begin this subsection by formalizing the concept of an optimal 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*

$$\pi^* \geqslant_V \pi$$

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 not at all clear, or even intuitive, that such an optimal policy need exist for a given finite MDP. As we will see later in this section, an optimal policy indeed does exist under the conditions treated in this article. Somewhat unintuitively, will first analyse the behaviour of surmised optimal policies, and use these properties to show to then prove their existence. This approach brings us to the next subsection.

4.2 Behavioural issues

This subsection is dedicated to deriving some useful characteristics of the mysterious (as of yet) optimal policy of a given finite MDP. These results will turn out to be essential for the contruction of explicit algorithms aiming to converge on optimal policies. We start of with an easy corollary that states that, while there might be more than one optimal policy, they all share the same value and action value function. Since it is the value function that we use to rank and disambiguate policies from one another, the following statement essentially says that there is no need to tell one optimal policy apart from any other optimal policy, they are - literally for what it's worth - the same thing.

Corollary 7. *(It's all the same thing) Let π_1, π_2 be two optimal policies for a finite MDP. Then*

1.
$$V^{\pi_1} \equiv V^{\pi_2} \tag{38}$$

2.
$$Q^{\pi_1} \equiv Q^{\pi_2} \tag{39}$$

Proof. Since both π_1 and π_2 are optimal policies, we have both

$$\pi_1 \geqslant_V \pi_2 \Leftrightarrow V^{\pi_1}(s) \geqslant V^{\pi_2}(s) \ \forall s \in S \tag{40}$$

as well as

$$\pi_2 \leqslant_V \pi_1 \Leftrightarrow V^{\pi_2}(s) \geqslant V^{\pi_1}(s) \ \forall s \in S, \tag{41}$$

clearly implying the equality stated in 38. Claim 39 follows from 38 and Corollary 4. □

Knowing that all optimal policies share the same value and action value function, we can define the optimal versions of these functions independently of an attached optimal policy π^*.

Definition 4. *(Optimal (action-)values) Let π^* be any optimal policy for a given finite MDP. We call $V^* = V^{\pi^*}$ the optimal value function, and $Q^* = Q^{\pi^*}$ the optimal action-value function.*

Corollary 7 guarantees the welldefinedness of the above terms. We go on to show a useful property of the optimal value function V^*.

Lemma 5. *(Different kind of optimum) For each $s \in S$,$a \in A$, the following hold:*

1.
$$V^*(s) = \max_{\pi \in \Pi} V^{\pi}(s) \tag{42}$$

2.

$$Q^*(s,a) = \max_{\pi \in \Pi} Q^\pi(s,a) \tag{43}$$

In other words, both the optimal value function and the optimal action-value function can be seen as the results of pointwise optimization over the state space S.

Proof. We start with the first claim. Let $s \in S$ be fixed but arbitrary. By the very definition of V^* as the value function of an optimal policy π^* which satisfies **??**, we have

$$\max_{\pi \in \Pi} V^\pi(s) = \max_{\pi^* \in \Pi, \pi \text{optimal}} V^{\pi^*}(s). \tag{44}$$

By Corollary 7 we can *arbitrarily* choose any optimal policy π_1^* to obtain

$$\max_{\pi^* \in \Pi, \pi \text{optimal}} V^{\pi^*}(s) = V^{\pi_1^*}(s) = V^*(s), \tag{45}$$

proving the identity 42. As for the second claim, let $(s,a) \in S \times A$ be arbitrary but fixed. Then

$$\begin{array}{rcl} \max_{\pi \in \Pi} Q^\pi(s,a) & \overset{Prop42}{=} & \max_{\pi \in \Pi} \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^\pi(s') \right] \\ & = & \sum_{s'} P^a_{ss'} R^a_{ss'} + \max_{\pi \in \Pi} \sum_{s'} P^a_{ss'} \gamma V^\pi(s'). \end{array} \tag{46}$$

where we have used that the term quantifying the expectation of the immediate reward $\mathbb{E}[r_t | s_t = s, a_t = a] = \sum_{s'} P^a_{ss'} R^a_{ss'}$ is independent of the particular policy π if both s_t and a_t are fixed at s and a, respectively. We see that

$$\begin{array}{rl} & \sum_{s'} P^a_{ss'} R^a_{ss'} + \max_{\pi \in \Pi} \sum_{s'} P^a_{ss'} \gamma V^\pi(s') \\ \leqslant & \sum_{s'} P^a_{ss'} R^a_{ss'} + \sum_{s'} P^a_{ss'} \gamma \max_{\pi \in \Pi} V^\pi(s') \\ \overset{42}{=} & \sum_{s'} P^a_{ss'} R^a_{ss'} + \sum_{s'} P^a_{ss'} \gamma V^*(s') \\ = & \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^*(s') \right]. \end{array} \tag{47}$$

Arbitrarily choosing an optimal policy π^* we finally see that

$$\begin{array}{rl} & \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^*(s') \right] \\ \overset{Def4}{=} & \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^{\pi^*}(s') \right] \\ \overset{Prop3}{=} & Q^{\pi^*}(s,a) \\ \overset{Def4}{=} & Q^*(s,a), \end{array} \tag{48}$$

proving "$\geqslant$" in 43. To see that "$\leqslant$" holds, consider, again using some arbitrary optimal policy π^*, that, since the optimal policy π^*, too, is an element of the set of *all* policies Π, we also have

$$Q^*(s,a) \overset{Def4}{=} Q^{\pi^*}(s,a) \leqslant \max_{\pi \in \Pi} Q^\pi(s,a). \tag{49}$$

Since $(s,a) \in S \times A$ was arbitrary, $\geqslant$ and $\leqslant$ together yield 43. □

It is worth reminding ourselves that this property assumes the existence of an optimal policy $\pi*$ with corresponding value and action value functions $V^{\pi^*} = V^*$, $Q^{\pi^*} = Q^*$, respectively. In other words, the right hand sides of both identities are well defined, but there is no guarantee that the policy counterpart, as implied by the left hand sides, exists.

We next show that greedyness is a necessary condition for a policy to be optimal.

Lemma 6. *(Necessary conditions for optimal policies) Let π^* be an optimal policy. Then π^* is greedy.*

Proof. Assume that π^* isn't greedy. By setting

$$\pi_{imp}(s,a) := \begin{cases} \frac{1}{|maxact^{Q^{\pi^*}}(s)|}, & a \in maxact^{Q^{\pi^*}} \\ 0, & a \notin maxact^{Q^{\pi^*}}, \end{cases} \tag{50}$$

we can see that π_{imp} and π^* satisfy the conditions of the Improved Policy improvement Theorem 2, implying that π_{imp} is Q^{π^*} greedy. Applying Lemma 4 therefore yields

$$\pi_{imp} >_V \pi^*, \tag{51}$$

clearly a violation of our optimality assumption on π^* requiring

$$V^{\pi^*}(s) \geqslant V^{\pi}(s) \tag{52}$$

for all $s \in S$, $\pi \in \Pi$. Hence, π^* must be greedy.

□

For some $\pi \in \Pi$, consider the original *Bellman optimality equation* given by

$$Q^{\pi}(s,a) \stackrel{!}{=} \mathbb{E}[r_t + \gamma \max_{a'} Q(s_{t+1}, a') | s_t = s, a_t = a]. \tag{53}$$

We next show that a policy π satisifies this identity holds true for greedy policies.

Lemma 7. *(Bell optimality equations) Let $\pi \in \Pi$ be an arbitrary policy. If π is greedy then it satifies the Bellman optimality equation 53.*

Proof. Let π be greedy. By one of our early results, namely Proposition 3 and Corollary 6, we clearly have

$$\begin{aligned} Q^{\pi}(s,a) &= \mathbb{E}[r_t + \gamma V^{\pi}(s_{t+1}) | s_t = s, a_t = a] \\ &= \mathbb{E}[r_t + \gamma \max_{a' \in A} Q^{\pi}(s_{t+1}, a') | s_t = s, a_t = a]. \end{aligned} \tag{54}$$

□

4.3 The Bellman operator

In this section, we will develop a slightly different view point on the action-value function of a policy - in particular, the action value function of greedy policies. We ask the reader to bear with us on this seemingly sudden deviation from our path in the hope that these results' benefits will become apparent soon enough. To this end, consider the set of non-negative valued functions mapping into $\mathbb{R}^+$, that is, functions of the form

$$\mathbb{Q}_{S,A} := \{Q : (S, A) \mapsto \mathbb{R}_0^+\} \tag{55}$$

equipped with the maximum norm

$$|Q|_\infty := \max_{(s,a)\in S\times A} |Q(s, a)|. \tag{56}$$

Since both state and action states are finite, respectively, we can choose an arbitrary ordering of actions and states $A = \{a_1, \ldots, a_{|A|}\}$ and $S = \{s_1, \ldots, s_{|S|}\}$. This in turn allows us to put an ordering on the state-action pairs $(s, a)_1 = (s_1, a_1), (s, a)_2 = (s_1, a_2), \ldots, (s, a)_{n_{S,A}} = (s_{|S|}, a_{|A|})$, where $n_{S,A} := |S| \cdot |A|$. Using this ordering, we can embedd our set of action-value functions $\mathbb{Q}_{S,A}$ into the normed space $(\mathbb{R}^{n_{S,A}}, |\cdot|_\infty)$ via the canonical mapping

$$\begin{array}{rccc} E: & \mathbb{Q}_{S,A} & \hookrightarrow & \mathbb{R}^{\mathbb{K}\mathbb{S},\mathbb{A}} \\ & Q & \mapsto & (Q((s, a)_1), \ldots, Q((s, a)_{n_{S,A}}). \end{array} \tag{57}$$

Thus, we can interpret our action-value functions as a subset of the Banach space $(\mathbb{R}^{n_{S,A}}, |\cdot|_\infty)$. It can easily be shown that $\mathbb{Q}_{S,A} \subset \mathbb{R}^{n_{S,A}}$ is a closed subset w.r.t $|\cdot|_\infty$. Since closed subspaces of complete spaces are complete themselves w.r.t the parent norm, we have obtained the following

Lemma 8. *Let $\mathbb{Q}_{S,A}$ be defined as in equation 55. Then $(\mathbb{Q}_{S,A}, |\cdot|_\infty)$ is a Banach space.*

The above result provides the perfect frame work to revisit Bellman's optimality equation from an operator theory point of view. Consider the non-linear operator on $\mathbb{Q}_{S,A}$ defined by

$$\begin{array}{rccc} B_\mathbb{Q}: & \mathbb{Q}_{S,A} & \to & \mathbb{Q}_{S,A} \\ & Q & \mapsto & B_\mathbb{Q} \end{array} \tag{58}$$

where the image $B_\mathbb{Q}(Q)$ of Q under $B_\mathbb{Q}$ is given by

$$B_\mathbb{Q}(Q)(s, a) := \mathbb{E}[r_t + \gamma \max_{a'} Q(s_{t+1}, a')|s_t = s, a_t = a], \tag{59}$$

$s \in S$, $a \in A$. Note that $B_\mathbb{Q}$ is well-defined due to $r_t \geqslant 0$ fr any sensible t. It is also worth noting that for $B_\mathbb{Q}$ to be applied, no policy π needs to be specified. Indeed, Q does not even need to be an actual action-value function derived from some policy π (although it usully will be - see below sections). All it requires is a function mapping state-action pairs to the non-negative part of the real numbers.

What does any of this have to do with our quest of finding and constructing (and, not to forget, proving the existence of) some optimal policy π^*? Well, we know that greedyness is a necessary condition for optimality (Lemma 6), and that satisfying the Bellman optimality equation is a necessary condition for greedyness (Lemma 7). To extend this reasoning by one more step, we can see that satisfying the Bellman optimality equation **??** is, by construction of $B_{\mathbb{Q}}$, equivalent to $Q^\pi = B_{\mathbb{Q}}(Q^\pi)$ in $\mathbb{Q}_{S,A}$. Thus we see the following implications hold true:

$$\pi^* \text{ optimal} \implies \pi^* \text{ greedy} \implies Q^{\pi^*} \text{ is a fixed point of } B_{\mathbb{Q}}. \tag{60}$$

A direct consequence of this is: if there is no fixed point of $B_{\mathbb{Q}}$, there can't be any optimal policy π^*. The good news is that not only does such a fixed point always exist, but it is also unique.

Proposition 4. *Let S and A be finite state and action spaces, respectively. Then the operator $B_{\mathbb{Q}}$ has a unique fixed point Q^*.*

A word of warning: Our notation Q^* heavily implies that we are dealing with the optimal policy's action value function. *We have not shown this yet.* The implications in 60 are strictly one-way, and there is still some work to be done to link the fixed point of $B_{\mathbb{Q}}$ back to some optimal policy π^*.

Proof. We will prove the claim by showing that $B_{\mathbb{Q}}$ is a contraction on $(\mathbb{Q}_{S,A}, |\cdot|_\infty)$ and then apply the Banach Fixed Point Theorem (see appendix). In other words, we need to show

$$|B_{\mathbb{Q}}(Q_1) - B_{\mathbb{Q}}(Q_2)|_\infty < c\,|Q_1 - Q_2|_\infty \tag{61}$$

for some $0 < c < 1$ and any pair $Q_1, Q_2 \in \mathbb{Q}_{S,A}$. Noting that we can rewrite

$$\begin{aligned} B_{\mathbb{Q}}(Q)(s,a) &= \mathbb{E}[r_t + \gamma \max_{a'} Q(s_{t+1}, a') | s_t = s, a_t = a] \\ &= \sum_{s' \in S} P^a_{s,s'} \big(R^a_{s,s'} + \gamma \max_{a' \in A} Q(s', a') \big), \end{aligned} \tag{62}$$

we see that

$$
\begin{aligned}
|B_{\mathbb{Q}}(Q_1) - B_{\mathbb{Q}}(Q_2)|_\infty &= \max_{(s,a)\in S\times A} \left| \sum_{s'\in S} P^a_{s,s'} \gamma \left(\max_{a'\in A} Q_1(s',a') - \max_{a'\in A} Q_2(s',a') \right) \right| \\
&\leqslant \gamma \max_{(s,a)\in S\times A} \left(\sum_{s'\in S} P^a_{s,s'} \left| \max_{a'\in A} Q_1(s',a') - \max_{a'\in A} Q_2(s',a') \right| \right) \\
&\leqslant \gamma \max_{(s,a)\in S\times A} \left(\sum_{s'\in S} P^a_{s,s'} \max_{a'\in A} \left| Q_1(s',a') - Q_2(s',a') \right| \right) \\
&\leqslant \gamma \max_{(s,a)\in S\times A} \left(\sum_{s'\in S} P^a_{s,s'} \max_{(s'',a')\in S\times A} \left| Q_1(s'',a') - Q_2(s'',a') \right| \right) \\
&= \gamma \max_{(s,a)\in S\times A} \left(\sum_{s'\in S} P^a_{s,s'} \left| Q_1 - Q_2 \right|_\infty \right) \\
&= \gamma \left| Q_1 - Q_2 \right|_\infty .
\end{aligned}
\tag{63}
$$

Since $0 < \gamma < 1$, there is a $\gamma < c < 1$ such that equation 61 holds, proving $B_{\mathbb{Q}}$ is a contraction on $(\mathbb{Q}_{S,A}, |\cdot|_\infty)$. By the Banach Fixed Point Theorem it has a unique fixed point $Q^* \in \mathbb{Q}_{S,A}$.

□

4.4 Building optimal policies

We are finally able to formulate this article's main result.

Theorem 3. *(Building optimal policies) Let $\pi_0 \in \Pi$ be an arbitrary (deterministic or probabilistic) policy for a given finite MDP. For $i = 1, \ldots,$ iteratively put*

$$
\pi_i(s,a) := \begin{cases} 1, & a = a_i, i = \min\{j | a_j \in maxact^{Q^{\pi_{i-1}}}(s)\} \\ 0, & otherwise. \end{cases}
\tag{64}
$$

Then there exists an $N \in \mathbb{N}$ such that the following hold:

1. *$\pi_{i-1} <_V \pi_i$ for $i = 1, \ldots, N$*
2. *$\pi_N =_V \pi_j$ for $j = N, \ldots,$*
3. *$\pi_N = \pi^*$, i.e. π_N is optimal*

The notation of the first case in 64 refers to the ordering of the states $(a_1, \ldots, a_{|A|}) = A$ already mentioned earlier. While the ordering is arbitrary and not required, we refer to it here to formulate our policy improvement algorithm in a closed, succinct form. What we mean to say is to really only chose the 'first' (w.r.t our abritrary ordering on A) action in $maxact^{Q^{\pi_{i-1}}}(s)$ and to ignore all other actions, even if $|maxact^{Q^{\pi_{i-1}}}(s)| > 1$. This way, we remain in the space of *deterministic* policies.

Proof. The first two points are a result of repeatedly applying Lemma 4. By construction 64, for all $i = 1, \cdots,$ the i-th policy π_i is $Q^{\pi_{i-1}}$-greedy, implying

$$\pi_i >_V \pi_{i-1} \tag{65}$$

unless π_{i-1} is $(Q^{\pi_{i-1}})$-greedy, too, in which case

$$\pi_i =_V \pi_{i-1}. \tag{66}$$

This implies that the above iteration stops generating improved policies *if and only if* a greedy policy is generated. Let us denote the first time this happens by index N. Since $n_{S,A} = |S| \cdot |A| < \infty$, there are only a finite number of deterministic policies at our disposal. Since the sequence of policies

$$\pi_0 <_V \pi_1 < \cdots < \pi_{N-1} < \pi_N \tag{67}$$

more generally implies

$$\pi_0 \neq \pi_1 \neq \cdots \neq \pi_{N-1} \neq \pi_N, \tag{68}$$

we are essentially drawing without replacement from a finite pot of policies until π_N is reached and the sequence stagnates. Thus, the greedy policy break criteria has to be reached in a finite number of steps $N \leqslant n_{S,A} < \infty$.

What is left to show is that π_N is actually optimal. We will do this by a proof of contradiction.

If π_N is not optimal, there must be at least one other policy $\pi_{imp} \in \Pi$ for which

$$\pi_{imp} >_V \pi_N \tag{69}$$

holds true. If π_{imp} itself is not greedy, we can apply the above iteration to $\pi_0 = \pi_{imp}$ to obtain a second greedy policy, $\pi_{imp,N_{imp}}$ satisfying

$$\pi_N <_V \pi_{imp,N_{imp}}. \tag{70}$$

By Corollary 6, this implies

$$\max_{a \in A} Q^{\pi_N}(s,a) < \max_{a \in A} Q^{\pi_{imp,N_{imp}}}(s,a) \tag{71}$$

for at least one $s \in S$, and, more generally,

$$Q^{\pi_N} \neq Q^{\pi_{imp,N_{imp}}}. \tag{72}$$

But both Q^{π_N} and $Q^{\pi_{imp,N_{imp}}}$ are fixed points of $B_{\mathbb{Q}}$, which only has one fixed point by Proposition 4. This contradiction implies that π_N must have been optimal, and the proof is complete.

□