An Elementary Proof that Q-learning Converges Almost Surely

Matthew T. Regehr *
University of Alberta
mregehr@ualberta.ca

Alex Ayoub
University of Alberta
aayoub@ualberta.ca

August 9, 2021

1 Introduction

Watkins' and Dayan's Q-learning is a model-free reinforcement learning algorithm that iteratively refines an estimate for the optimal action-value function of an MDP by stochastically "visiting" many state-ation pairs [Watkins and Dayan, 1992]. Variants of the algorithm lie at the heart of numerous recent state-of-the-art achievements in reinforcement learning, including the superhuman Atari-playing deep Q-network [Mnih et al., 2015].

The goal of this paper is to reproduce a precise and (nearly) self-contained proof that Q-learning converges. Much of the available literature leverages powerful theory to obtain highly generalizable results in this vein. However, this approach requires the reader to be familiar with and make many deep connections to different research areas. A student seeking to deepen their understand of Q-learning risks becoming caught in a vicious cycle of "RL-learning Hell". For this reason, we give a complete proof from start to finish using only one external result from the field of stochastic approximation, despite the fact that this minimal dependence on other results comes at the expense of some "shininess".

2 Related Works

The first proof that Q-learning converges with probability 1 is outlined in [Watkins, 1989] and given more fully in [Watkins and Dayan, 1992]. The proof of [Tsitsiklis, 1994] applies the theory of stochastic approximation to allow a far more general asynchronous structure. [Even-Dar et al., 2003] builds upon this work to derive more precise rates of convergence. Another approach by [Borkar and Meyn, 2000] leverages the Lyapunov theory of ordinary differential equations to analyze a swath of stochastic approximation algorithms. Lastly, [Szepesvári and Littman, 1996] analyzes Q-learning in the setting of generalized MDPs and focuses on the contractivity properties of dynamic programming operators.

3 Background

We make frequent use of standard measure theoretic and linear analytic notation and thus invite the reader to read Section A upon encountering any unfamiliar symbols or terms.

*An early version of this work was submitted as author's CMPUT 653 course project in Winter 2021.

3.1 Markov Decision Processes

A typical formalization of environment in reinforcement learning—and the one we study here—is the Markov decision process (MDP). A reader familiar with the fundamentals of reinforcement learning may skip this subsection without issue.

Definition 1. *A countable (finite) discounted MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ where $\mathcal{S}$ and $\mathcal{A}$ are countable (finite) sets of "states" and "actions" respectively, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a "transition kernel", $r \in \ell^\infty(\mathcal{S} \times \mathcal{A})$ represents "rewards", and $\gamma \in [0, 1)$ is a "discount rate".*

In order to design agents that make "good" decisions when interacting with an MDP, we would like to somehow measure the value of making certain decisions in certain states. A convenient approach to measuring value relies on the fixed point theory of so-called "dynamic programming" operators. The following class of operators will serve our purposes nicely.

Definition 2. *The "Bellman optimality operator" of an MDP $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ is*

$$T_M^* : \ell^\infty(\mathcal{S} \times \mathcal{A}) \to \ell^\infty(\mathcal{S} \times \mathcal{A}), q \mapsto (s, a) \mapsto r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) \sup_{a' \in \mathcal{A}} q(s', a').$$

Incidentally, exact or even approximate knowledge of the fixed point[1] of the Bellman optimality operator is sufficient to act optimally or near-optimally[2]. For now, however, it is enough that a unique fixed point exists. The proof is a routine application of the well-known Banach fixed point theorem and can be found in Section C.

Theorem 1. *For any MDP M, T_M^* admits a unique fixed point q_M^*, which we refer to as the "optimal action-value function" for M.*

The following bound will serve a useful purpose in proving our main theorem. As before, a proof can be found in Section C.

Lemma 1. *For any MDP M with rewards r and discount rate γ,*

$$\|q_M^*\|_\infty \leq \frac{\|r\|_\infty}{1 - \gamma}.$$

3.2 Sampling Trajectories from an MDP

In order to compute Q-learning iterates, we would like to sample trajectories from a distribution that respects the dynamics of a given countable discounted MDP $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$. To that end, we require some statistical apparatus. Once again, the reader is referred to Section A if any notation is unfamiliar.

Definition 3. *The "trajectory space" of M is the measurable space*

$$(\Omega_M, \mathcal{F}_M) := \left((\mathcal{S} \times \mathcal{A} \times \mathcal{S})^{\mathbb{N}_0}, \bigotimes_{t \in \mathbb{N}_0} \mathcal{P}(\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \right).$$

[1] A fixed point of a map $f : \mathcal{X} \to \mathcal{X}$ is a point $x^* \in \mathcal{X}$ such that $f(x^*) = x^*$.

[2] See Lemma I at `https://rltheory.github.io/lecture-notes/planning-in-mdps/lec6/`.

Definition 4. *The "trajectory process" of M is the sequence $(S_0, A_0, S_0', S_1, A_1, S_1', \dots)$ of $\mathcal{F}_M/\mathcal{P}(\mathcal{S})$ and $\mathcal{F}_M/\mathcal{P}(\mathcal{A})$-measurable projections defined by*[3]

$$((S_0, A_0, S_0'), (S_1, A_1, S_1'), \dots) := \mathrm{id}_{\Omega_M}.$$

Definition 5. *The set of "trajectory measures" on M, denoted $\Delta_T(M)$, is the set of probability measures $\mathbb{P} \in \Delta(\Omega_M, \mathcal{F}_M)$ satisfying*

$$\mathbb{P}(S_t' = s_t' | S_0, A_0, S_0', \dots, S_t, A_t) = P(s' | S_t, A_t)$$

almost surely (a.s.) for any $s_t' \in \mathcal{S}$ and $t \in \mathbb{N}_0$.

Definition 6. *The "occurences" of $(s, a) \in \mathcal{S} \times \mathcal{A}$ along a "trajectory" $\omega \in \Omega_M$ constitute*

$$\mathcal{T}_{(s,a)}(\omega) := \{t \in \mathbb{N}_0 : (S_t, A_t)(\omega) = (s, a)\}.$$

4 The Q-learning Algorithm

Our overall goal is to design a reinforcement learning agent that makes good decisions in a given environment. To that end, we seek to develop an algorithm that closely approximates the optimal action-value function for a given MDP. Furthermore, we would like to do this without explicitly accessing an environment's transition kernel as these are frequently unavailable in real-world applications. On the other hand, many real-world environments permit the sampling of transitions and in fact we will use sampling to develop the Q-learning algorithm. In particular, by stochastically "visiting" many state-action pairs, we iteratively refine an estimate for q_M^*. The details of how we visit states and choose actions should not matter as long as our samples cover the state-action space sufficiently well. Altogether, these ideas form the basis of Watkins' and Dayan's Q-learning [Watkins and Dayan, 1992].

Definition 7 (Q-learning)**.** *The "Q-learning iterates" on a finite MDP M with discount rate γ induced by a "stepsize" sequence $\alpha = (\alpha_t)_{t \in \mathbb{N}_0}$ in $\mathbb{R}$ and a trajectory $\omega \in \Omega_M$ form the sequence*[4] *$(Q_t^\alpha(\omega))_{t \in \mathbb{N}_0}$ in $\ell^\infty(\mathcal{S} \times \mathcal{A})$ defined recursively by $Q_0^\alpha(\omega) \equiv \mathbf{0}$ and*[5]

$$Q_{t+1}^\alpha := (s, a; \omega) \mapsto \begin{cases} (1 - \alpha_t) Q_t^\alpha(s, a; \omega) + \alpha_t (r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q_t^\alpha(S_t'(\omega), a'; \omega)) & \text{if } t \in \mathcal{T}_{(s,a)}(\omega) \\ Q_t^\alpha(s, a; \omega) & \text{otherwise} \end{cases}$$

for $t \in \mathbb{N}_0$.

Remark 1. *While the construction of the Q-learning iterates depends explicitly on the states, actions, rewards, and discount rate of an MDP, it does not depend directly on the transition kernel of an MDP. This increases the flexibility of Q-learning and, as we will see later, does not preclude convergence as long as the trajectories are sampled from an appropriate distribution.*

[3] For convenience, we suppress M from the notation of the trajectory process as the correct meaning should always be deducible via "type inference".

[4] Similarly, we omit M from the notation of the Q-learning iterates and rely instead upon context and prepositional phrases to make the underlying MDP unambiguous.

[5] We adopt the function "currying" convention $f(y; x) := f(x)(y)$ for $f : \mathcal{X} \to \mathcal{Y} \to \mathcal{Z}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$.

5 Convergence of Q-learning

Q-learning iterates in hand, we are ready to state the assumptions that lead to convergence.

Definition 8. *Let M be an MDP. A trajectory measure $\mathbb{P} \in \Delta_T(M)$ (see Definition 5) and a sequence $(\alpha_t)_{t\in\mathbb{N}_0}$ in $[0,1]$ are said to satisfy the Robbins–Monro condition when*

$$\sum_{t\in\mathcal{T}_{(s,a)}(\omega)} \alpha_t = \infty \quad \text{and} \quad \sum_{t\in\mathcal{T}_{(s,a)}(\omega)} \alpha_t^2 < \infty.$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $\mathbb{P}$-almost all $\omega \in \Omega_M$. The set of all such trajectory measure-stepsize sequence pairs is denoted $\nu(M)$.

Remark 2. *The condition that $\sum_{t\in\mathcal{T}_{(s,a)}(\omega)} \alpha_t = \infty$ requires that $\mathcal{T}_{(s,a)}(\omega)$ be infinite for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $\mathbb{P}$-almost all $\omega \in \Omega_M$, i.e. the sampling strategy that produces the measure $\mathbb{P}$ must visit all state-action pairs infinitely often.*

At last, we have arrived at our main result. The proof is delayed until Subsection 5.2 as only then will we be adequately equipped for the task.

Theorem 2. *Let M be a finite MDP and let $(\mathbb{P},\alpha) \in \nu(M)$ be a Robbins–Monro trajectory measure-stepsize sequence pair for M. Then the Q-learning iterates $(Q_t^\alpha(\omega))_{t\in\mathbb{N}_0}$ on M converge uniformly to q_M^* for $\mathbb{P}$-almost all $\omega \in \Omega_M$.*

5.1 The Action-Replay Processes

We begin our journey toward convergence by showing that an MDP M can be recovered by a certain limiting process from a trajectory-dependent MDP whose whose optimal action-value functions track the Q-learning iterates on M. We will see that this construction serves as the primary proof device for proving the convergence of Q-learning.

Definition 9. *The "action-replay process" of an MDP $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma\rangle$ induced by a stepsize sequence $\alpha = (\alpha_t)_{t\in\mathbb{N}_0}$ and a trajectory $\omega \in \Omega_M$ is the MDP $\hat{M}^\alpha(\omega) := \langle \hat{\mathcal{S}}, \mathcal{A}, \hat{P}, \hat{r}, \gamma\rangle$ where $\hat{\mathcal{S}} := \mathcal{S}\times\mathbb{N}_0 \cup \{s_{absorb}\}$;*

$$\begin{aligned}
\hat{P}((S'_{t'}(\omega),t')|(s,t),a) &:= \alpha_{t'} \prod_{\tau\in\mathcal{T}_{(s,a)}(\omega)\cap(t',t)} (1-\alpha_\tau),\\
\hat{P}(s_{absorb}|(s,t),a) &:= \prod_{\tau\in\mathcal{T}_{(s,a)}(\omega)\cap[0,t)} (1-\alpha_\tau), \text{ and}\\
\hat{P}(s_{absorb}|s_{absorb},a) &:= 1
\end{aligned}$$

for $(s,a) \in \mathcal{S}\times\mathcal{A}$, $t \in \mathbb{N}_0$, and $t' \in \mathcal{T}_{(s,a)}(\omega)\cap[0,t)$ as well as $\hat{P}(\cdot|\cdot,\cdot) \equiv \mathbf{0}$ everwhere else; and, finally,

$$\hat{r}((s,t),a) := r(s,a) \sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,t)} \hat{P}((S'_{t'}(\omega),t')|(s,t),a)$$

for $(s,t) \in \mathcal{S}\times\mathbb{N}_0$ and $a \in \mathcal{A}$ as well as $\hat{r}(\cdot,\cdot) \equiv \mathbf{0}$ everwhere else.

Our next theorem reduces the analysis of Q-learning iterates to analysis of the optimal action-value function of an action-replay process.

Theorem 3. *Let M be a finite MDP, let α be a stepsize sequence, and let $(Q_t^\alpha(\omega))_{t\in\mathbb{N}_0}$ be the induced Q-learning iterates on M. For every $\omega \in \Omega_M$, $t \in \mathbb{N}_0$, and $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$q^*_{\hat{M}^\alpha(\omega)}((s,t),a) = Q_t^\alpha(s,a;\omega).$$

Before we prove the theorem, we strongly encourage the reader to prove the following lemma that shows that, while the dynamics of the action-replay processes may look intimidating at a first glance, their recursive form is much more pleasant to work with.

Lemma 2. *With all terms as in Definition 9, $(s,a) \in \mathcal{S}\times\mathcal{A}$, and $\omega \in \Omega_M$, we have*

$$\hat{P}((S'_{t'}(\omega),t')|(s,t+1),a) = \hat{P}((S'_{t'}(\omega),t')|(s,t),a)$$

for any $t \notin \mathcal{T}_{(s,a)}(\omega)$ and $t' \in \mathcal{T}_{(s,a)}(\omega)\cap[0,t+1)$ as well as

$$\hat{P}((S'_t(\omega),t)|(s,t+1),a) = \alpha_t$$

and

$$\hat{P}((S'_{t'}(\omega),t')|(s,t+1),a) = (1-\alpha_t)\hat{P}((S'_{t'}(\omega),t')|(s,t),a)$$

for any $t \in \mathcal{T}_{(s,a)}(\omega)$ and $t' \in \mathcal{T}_{(s,a)}(\omega)\cap[0,t)$.

Proof of Theorem 3. Fix $\omega\in\Omega_M$ and let $\hat{M}^\alpha(\omega) = \langle \hat{\mathcal{S}}, \mathcal{A}, \hat{P}, \hat{r}, \gamma\rangle$.

We begin by establishing an extremely useful form for the optimal action-values of $\hat{M}^\alpha(\omega)$. To that end, notice that, for any $a\in\mathcal{A}$,

$$\begin{aligned}
q^*_{\hat{M}^\alpha(\omega)}(s_{\text{absorb}},a) &= T^*_{\hat{M}^\alpha(\omega)} q^*_{\hat{M}^\alpha(\omega)}(s_{\text{absorb}},a)\\
&= \hat{r}(s_{\text{absorb}},a) + \gamma \sum_{\sigma'\in\mathcal{S}_{\hat{M}^\alpha(\omega)}} \hat{P}(\sigma'|s_{\text{absorb}},a)\max_{a'\in\mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}(\sigma',a')\\
&= \gamma \max_{a'\in\mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}(s_{\text{absorb}},a'),
\end{aligned}$$

so, taking a maximum over $a\in\mathcal{A}$, we must have $\max_{a'\in\mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}(s_{\text{absorb}},a') = 0$ and hence

$$\begin{aligned}
q^*_{\hat{M}^\alpha(\omega)}((s,k),a) &= T^*_{\hat{M}^\alpha(\omega)} q^*_{\hat{M}^\alpha(\omega)}((s,k),a)\\
&= \hat{r}((s,k),a) + \gamma \sum_{\sigma'\in\mathcal{S}_{\hat{M}^\alpha(\omega)}} \hat{P}(\sigma'|(s,k),a)\max_{a'\in\mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}(\sigma',a')\\
&= r(s,a) \sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,k)} \hat{P}((S'_{t'}(\omega),t')|(s,k),a) +\\
&\quad \gamma\hat{P}(s_{\text{absorb}}|(s,k),a)\cancelto{0}{\max_{a'\in\mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}(s_{\text{absorb}},a')} + \qquad (1)\\
&\quad \gamma \sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,k)} \hat{P}((S'_{t'}(\omega),t')|(s,k),a)\max_{a'\in\mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_{t'}(\omega),t'),a')\\
&= \sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,k)} \hat{P}((S'_{t'}(\omega),t')|(s,k),a)\left(r(s,a) + \gamma\max_{a'\in\mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_{t'}(\omega),t'),a')\right)
\end{aligned}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $k\in\mathbb{N}_0$.

With this in mind, we now prove the theorem by induction on t. Since $[0,0) = \varnothing$, Equation (1) yields $q^*_{\hat{M}^\alpha(\omega)}((s,0),a) = 0 = Q^\alpha_0(s,a;\omega)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and hence the base case holds. As for the inductive step, let $t \in \mathbb{N}_0$, assume the claim holds for t, and let $(s,a) \in \mathcal{S} \times \mathcal{A}$. We consider two cases.

If $t \notin \mathcal{T}_{(s,a)}(\omega)$, then, by Equation (1) and Lemma 2, we have

$$\begin{aligned}
q^*_{\hat{M}^\alpha(\omega)}((s,t+1),a) &= \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t+1)} \hat{P}((S'_{t'}(\omega),t')|(s,t+1),a) \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_{t'}(\omega),t'),a') \right) \\
&= \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((S'_{t'}(\omega),t')|(s,t),a) \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_{t'}(\omega),t'),a') \right) \\
&= q^*_{\hat{M}^\alpha(\omega)}((s,t),a) \\
&= Q^\alpha_t(s,a;\omega) \\
&= Q^\alpha_{t+1}(s,a;\omega).
\end{aligned}$$

Likewise, if $t \in \mathcal{T}_{(s,a)}(\omega)$, then, by Equation (1) and Lemma 2,

$$\begin{aligned}
q^*_{\hat{M}^\alpha(\omega)}((s,t+1),a) &= \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t+1)} \hat{P}((S'_{t'}(\omega),t')|(s,t+1),a) \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_{t'}(\omega),t'),a') \right) \\
&= (1-\alpha_t) \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((S'_{t'}(\omega),t')|(s,t),a) \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_{t'}(\omega),t'),a') \right) \\
&\quad + \alpha_t \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_t(\omega),t),a') \right) \\
&= (1-\alpha_t) q^*_{\hat{M}^\alpha(\omega)}((s,t),a) + \alpha_t \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*_{\hat{M}^\alpha(\omega)}((S'_t(\omega),t),a') \right) \\
&= (1-\alpha_t) Q^\alpha_t(s,a;\omega) + \alpha_t \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q^\alpha_t(S'_t(\omega),a';\omega) \right) \\
&= Q^\alpha_{t+1}(s,a;\omega)
\end{aligned}$$

and hence the inductive step holds as well. □

At the beginning of Subsection 5.1, we promised that an MDP can be recovered from its action-replay process via a limiting procedure; we now make good on that promise.

Theorem 4. *Let $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ be an MDP and let $(\mathbb{P}, \alpha) \in \nu(M)$ (recall Definition 8). Then, for any $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $\mathbb{P}$-almost all $\omega \in \Omega_M$,*

$$\hat{r}((s,t),a;\omega) \xrightarrow{t \to \infty} r(s,a)$$

and

$$\sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((s',t')|(s,t),a;\omega) \xrightarrow{t \to \infty} P(s'|s,a)$$

where $\hat{M}^\alpha(\omega) = \langle \hat{\mathcal{S}}, \mathcal{A}, \hat{P}(\omega), \hat{r}(\omega), \gamma \rangle$.

The proof rests on a classic result from the theory of stochastic approximation.

Theorem 5 (The Robbins–Monro Theorem)**.** *For any familes of random variables* $(\beta_t)_{t\in\mathbb{N}_0}$, $(\xi_t)_{t\in\mathbb{N}_0}$, *and* $(X_t)_{t\in\mathbb{N}_0}$ *such that* $(\beta_t)_{t\in\mathbb{N}_0}$ *is non-negative and satisfies* $\sum_{t\in\mathcal{T}_{(s,a)}(\omega)}\beta_t=\infty$ *as well as* $\sum_{t\in\mathcal{T}_{(s,a)}(\omega)}\beta_t^2<\infty$ *a.s.,* $\mathbb{E}[\xi_t]=\Xi$ *for all* $t\in\mathbb{N}_0$, $(\xi_t)_{t\in\mathbb{N}_0}$ *is bounded a.s., and*

$$X_{t+1}=(1-\beta_t)X_t+\beta_t\xi_t$$

for all $t\in\mathbb{N}_0$, *we have that* $X_t\to\Xi$ *a.s.*

A statement and proof of the theorem can be found under Theorem 2.3.1 in [Kushner and Clark, 1978] and its original, weaker variant (quadratic mean convergence rather than almost sure convergence) is stated and proved in [Robbins and Monro, 1951].

Proof of Theorem 4. Fix $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and discard a $\mathbb{P}$-null set from Ω_M so that $\sum_{t\in\mathcal{T}_{(s,a)}(\omega)}\alpha_t=\infty$ and $\sum_{t\in\mathcal{T}_{(s,a)}(\omega)}\alpha_t^2<\infty$ for $\omega\in\Omega_M$. Furthermore, for any $k\in\mathbb{N}_0$ and $\omega\in\Omega_M$, let $T_k(\omega)$ be the k^{th} smallest element of $\mathcal{T}_{(s,a)}(\omega)$ (where $T_0(\omega):=\min\mathcal{T}_{(s,a)}(\omega)$), which is well-defined by Remark 2.

We now show that the reward limit holds. To that end, for $t\in\mathbb{N}_0$ and $\omega\in\Omega_M$, define

$$X_t(\omega):=\hat{r}((s,t),a;\omega)=r(s,a)\sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,t)}\hat{P}((S'_{t'}(\omega),t')|(s,t),a;\omega).$$

Then, for any $t\in\mathbb{N}_0$ and $\omega\in\Omega_M$, $t\notin\mathcal{T}_{(s,a)}(\omega)$ implies

$$\begin{aligned}X_{t+1}(\omega)&=r(s,a)\sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,t+1)}\hat{P}((S'_{t'}(\omega),t')|(s,t+1),a;\omega)\\&=r(s,a)\sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,t)}\hat{P}((S'_{t'}(\omega),t')|(s,t),a;\omega)\\&=X_t(\omega)\end{aligned}$$

by Lemma 2, whereas $t\in\mathcal{T}_{(s,a)}(\omega)$ implies that

$$\begin{aligned}X_{t+1}(\omega)&=r(s,a)\sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,t+1)}\hat{P}((S'_{t'}(\omega),t')|(s,t+1),a;\omega)\\&=r(s,a)\left((1-\alpha_t)\sum_{t'\in\mathcal{T}_{(s,a)}(\omega)\cap[0,t)}\hat{P}((S'_{t'}(\omega),t')|(s,t),a;\omega)+\alpha_t\right)\\&=(1-\alpha_t)X_t(\omega)+\alpha_t r(s,a)\end{aligned}$$

by Lemma 2. In particular, we have

$$X_{T_{k+1}}=(1-\alpha_{T_k})X_{T_k}+\alpha_{T_k}r(s,a)$$

for all $k\in\mathbb{N}_0$. By Theorem 5, $X_{T_k}(\omega)\xrightarrow{k\to\infty}r(s,a)$ for $\mathbb{P}$-almost all $\omega\in\Omega_M$. Finally, since $(X_t)_{t\in\mathbb{N}_0}$ is constant between the terms of the subsequence $(X_{T_k})_{k\in\mathbb{N}_0}$, we have

$$\hat{r}((s,t),a;\omega)=X_t(\omega)\xrightarrow{t\to\infty}r(s,a)$$

for $\mathbb{P}$-almost all $\omega\in\Omega_M$ as well.

Next, we show that the dynamics limit holds in an analogous fashion. To that end, for $t \in \mathbb{N}_0$ and $\omega \in \Omega_M$, define

$$Y_t(\omega) := \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((s',t')|(s,t),a;\omega).$$

Then, for any $t \in \mathbb{N}_0$ and $\omega \in \Omega_M$, $t \notin \mathcal{T}_{(s,a)}(\omega)$ implies $Y_{t+1}(\omega) = Y_t(\omega)$ by Lemma 2, whereas $t \in \mathcal{T}_{(s,a)}(\omega)$ implies that

$$\begin{aligned}
Y_{t+1}(\omega) &= \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t+1)} \hat{P}((s',t')|(s,t+1),a) \\
&= \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t+1)} \mathbb{1}(S'_{t'}(\omega) = s') \hat{P}((S'_{t'}(\omega),t')|(s,t+1),a) \\
&= (1-\alpha_t) \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \mathbb{1}(S'_{t'}(\omega) = s') \hat{P}((S'_{t'}(\omega),t')|(s,t),a) + \alpha_t \mathbb{1}(S'_t(\omega) = s') \\
&= (1-\alpha_t) \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((s',t')|(s,t),a) + \alpha_t \mathbb{1}(S'_t(\omega) = s') \\
&= (1-\alpha_t) Y_t(\omega) + \alpha_t \mathbb{1}(S'_t(\omega) = s')
\end{aligned}$$

by Lemma 2. In particular, we have

$$Y_{T_{k+1}} = (1-\alpha_{T_k}) Y_{T_k} + \alpha_{T_k} \mathbb{1}(S'_{T_k} = s')$$

for all $k \in \mathbb{N}_0$. But, for any $k \in \mathbb{N}_0$,

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(S'_{T_k} = s')] &= \mathbb{P}(S'_{T_k} = s') \\
&= \sum_{t=0}^{\infty} \mathbb{P}(T_k = t, S'_t = s') \\
&= \sum_{t=0}^{\infty} \mathbb{P}(\left|\mathcal{T}_{(s,a)} \cap [0,t)\right| = k-1, S_t = s, A_t = a, S'_t = s') \\
&= \sum_{t=0}^{\infty} \mathbb{P}(\left|\mathcal{T}_{(s,a)} \cap [0,t)\right| = k-1, S_t = s, A_t = a) P(s'|s,a) \\
&= P(s'|s,a) \sum_{t=0}^{\infty} \mathbb{P}(T_k = t) \\
&= P(s'|s,a)
\end{aligned}$$

since $\left|\mathcal{T}_{(s,a)} \cap [0,t)\right|$ is a $\sigma(S_0, A_0, S'_0, \ldots, S_{t-1}, A_{t-1})$-measurable random variable and since $\mathbb{P}$ is a trajectory measure on M. By Theorem 5, $Y_{T_k}(\omega) \xrightarrow{k\to\infty} P(s'|s,a)$ for $\mathbb{P}$-almost all $\omega \in \Omega_M$. As $(Y_t)_{t\in\mathbb{N}_0}$ is constant between the terms of the subsequence $(Y_{T_k})_{k\in\mathbb{N}_0}$,

$$\sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((s',t')|(s,t),a;\omega) = Y_t(\omega) \xrightarrow{t\to\infty} P(s'|s,a)$$

for $\mathbb{P}$-almost all $\omega \in \Omega_M$ as well. □

5.2 Proof of Theorem 2

Having tamed the action-replay processes, all of the conceptual pieces are now in place to prove the convergence of Q-learning. For the sake of digestibility, we have factored out some of the technical heavy lifting into the following two lemmas.

Lemma 3. *Let $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ be an MDP, $\alpha = (\alpha_t)_{t \in \mathbb{N}_0}$ a stepsize sequence in $[0,1]$, $\omega \in \Omega_M$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$. For any $\tilde{t}, t \in \mathbb{N}_0$ with $\tilde{t} \leq t$,*

$$\sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,\tilde{t})} \hat{P}((S'_{t'}(\omega), t') | (s,t), a) \leq e^{-\sum_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} \alpha_\tau}$$

where $\hat{M}^\alpha(\omega) = \langle \hat{\mathcal{S}}, \mathcal{A}, \hat{P}, \hat{r}, \gamma \rangle$.

Proof. Since $1 - \alpha \leq e^{-\alpha}$ for all $\alpha \in \mathbb{R}$,

$$\begin{aligned}
\sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,\tilde{t})} \hat{P}((S'_{t'}(\omega), t') | (s,t), a) &= \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,\tilde{t})} \alpha_{t'} \prod_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap (t',t)} (1 - \alpha_\tau) \\
&\leq \prod_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} (1 - \alpha_\tau) + \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,\tilde{t})} \alpha_{t'} \prod_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap (t',t)} (1 - \alpha_\tau) \\
&= \prod_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} (1 - \alpha_\tau) \\
&\leq \prod_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} e^{-\alpha_\tau} \\
&= e^{-\sum_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} \alpha_\tau}
\end{aligned}$$

where the second equality follows by induction on $\tilde{t}$ (we encourage the reader to check). □

Lemma 4. *Let $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ be a finite MDP, let $\alpha = (\alpha_t)_{t \in \mathbb{N}_0}$ be a stepsize sequence, let $\omega \in \Omega_M$, let $(Q_t := Q_t^\alpha(\omega))_{t \in \mathbb{N}_0}$ be the induced Q-learning iterates on M, and let $\tilde{t}, t \in \mathbb{N}_0$ with $\tilde{t} \leq t$. Then, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $|Q_t(s,a) - q^*_M(s,a)|$ is at most*

$$\gamma \max_{t' \in [\tilde{t},t)} \|Q_{t'} - q^*_M\|_\infty + \|\hat{r}_t - r\|_\infty + \left(\frac{\gamma \|r\|_\infty}{1-\gamma}\right) \left(|\mathcal{S}| \left\| \hat{P}_t - P \right\|_\infty + 2e^{-\sum_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} \alpha_\tau} \right)$$

where $\hat{M} := \hat{M}^\alpha(\omega) = \langle \hat{\mathcal{S}}, \mathcal{A}, \hat{P}, \hat{r}, \gamma \rangle$, $\hat{P}_t(s'|s,a) := \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((s',t')|(s,t),a)$, and $\hat{r}_t(s,a) := \hat{r}((s,t),a)$.

Proof. Fix $(s,a) \in \mathcal{S} \times \mathcal{A}$. By Theorem 3 and the triangle inequality,

$$\begin{aligned}
|Q_t(s,a) - q^*_M(s,a)| &= \left| T^*_{\hat{M}} q^*_{\hat{M}}((s,t),a) - T^*_M q^*_M(s,a) \right| \\
&\leq |\hat{r}((s,t),a) - r(s,a)| + \\
&\quad \gamma \Bigg| \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((S'_{t'}(\omega), t') | (s,t), a) \max_{a' \in \mathcal{A}} q^*_{\hat{M}}((S'_{t'}(\omega), t'), a') \\
&\qquad - \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} q^*_M(s', a') \Bigg|.
\end{aligned}$$

But $|\hat{r}((s,t),a) - r(s,a)| = |\hat{r}_t(s,a) - r(s,a)| \leq \|\hat{r}_t - r\|_\infty$ and, applying the triangle inequality once more,

$$\left| \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((S'_{t'}(\omega), t')|(s,t),a) \max_{a' \in \mathcal{A}} q^*_{\hat{M}}((S'_{t'}(\omega), t'), a') - \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} q^*_M(s', a') \right|$$

is bounded by the sum of (2) and (3) where

$$\begin{aligned}
&\left| \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((S'_{t'}(\omega), t')|(s,t),a) \left(\max_{a' \in \mathcal{A}} q^*_{\hat{M}}((S'_{t'}(\omega), t'), a') - \max_{a' \in \mathcal{A}} q^*_M(S'_{t'}(\omega), a') \right) \right| && (2) \\
&\quad \leq \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} \hat{P}((S'_{t'}(\omega), t')|(s,t),a) \max_{a' \in \mathcal{A}} \left| q^*_{\hat{M}}((S'_{t'}(\omega), t'), a') - q^*_M(S'_{t'}(\omega), a') \right| + \\
&\qquad \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,\tilde{t})} \hat{P}((S'_{t'}(\omega), t')|(s,t),a) \left(\left\| q^*_{\hat{M}} \right\|_\infty + \|q^*_M\|_\infty \right) \\
&\quad \leq \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} \hat{P}((S'_{t'}(\omega), t')|(s,t),a) \max_{a' \in \mathcal{A}} |Q_{t'}(S'_{t'}(\omega), a') - q^*_M(S'_{t'}(\omega), a')| + \\
&\qquad && \text{(Theorem 3)} \\
&\qquad \left(\frac{\|\hat{r}\|_\infty + \|r\|_\infty}{1-\gamma} \right) \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,\tilde{t})} \hat{P}((S'_{t'}(\omega), t')|(s,t),a) && \text{(Lemma 1)} \\
&\quad \leq \max_{t' \in [\tilde{t},t)} \|Q_{t'} - q^*_M\|_\infty + \left(\frac{2\|r\|_\infty}{1-\gamma} \right) e^{-\sum_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [\tilde{t},t)} \alpha_\tau} && \text{(Lemma 3)}
\end{aligned}$$

and

$$\begin{aligned}
&\left| \sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((S'_{t'}(\omega), t')|(s,t),a) \max_{a' \in \mathcal{A}} q^*_M(S'_{t'}(\omega), a') - \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} q^*_M(s', a') \right| && (3) \\
&\quad = \left| \sum_{s' \in \mathcal{S}} \left(\sum_{t' \in \mathcal{T}_{(s,a)}(\omega) \cap [0,t)} \hat{P}((s', t')|(s,t),a) - P(s'|s,a) \right) \max_{a' \in \mathcal{A}} q^*_M(s', a') \right| \\
&\quad \leq \|q^*_M\|_\infty \sum_{s' \in \mathcal{S}} \left| \hat{P}_t(s'|s,a) - P(s'|s,a) \right| \\
&\quad \leq \left(\frac{|\mathcal{S}| \, \|r\|_\infty}{1-\gamma} \right) \left\| \hat{P}_t - P \right\|_\infty && \text{(Lemma 1)}
\end{aligned}$$

(where the equality follows from the fact that $s' \neq S'_{t'}(\omega)$ implies $\hat{P}((s', t')|(s,t),a) = 0$), which yields the desired bound. □

It is time to finish the job. While most of the error terms provided by Lemma 4 can be controlled in a straightforward manner via Theorem 4, it is not immediately clear how to control $\max_{t' \in [\tilde{t},t)} \|Q_{t'} - q^*_M\|_\infty$. However, we will see that it may be subdued by repeatedly applying Lemma 4 until a sufficiently small exponential coefficient is obtained.

Proof Theorem 2. Taking finite unions of null sets as needed, discard a $\mathbb{P}$-null set from Ω_M so that, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\omega \in \Omega_M$, $\sum_{t \in \mathcal{T}_{(s,a)}(\omega)} \alpha_t = \infty$ holds in addition to the conclusion of Theorem 4. With this in mind, fix $\omega \in \Omega_M$, put $(Q_t)_{t\in\mathbb{N}_0} := (Q_t^\alpha(\omega))_{t\in\mathbb{N}_0}$, and let $(\hat{P}_t)_{t\in\mathbb{N}_0}$ as well as $(\hat{r}_t)_{t\in\mathbb{N}_0}$ be as in Lemma 4.

Now, let $\epsilon > 0$ and choose $k \in \mathbb{N}$ sufficiently large so that

$$\gamma^{k+1} \leq \frac{\epsilon(1-\gamma)}{8\left\|r\right\|_\infty}$$

(where $\cdot/0 := \infty$). Furthermore, by Theorem 4, we may find $t_0 \in \mathbb{N}_0$ such that

$$\left\|\hat{r}_t - r\right\|_\infty \leq \frac{\epsilon(1-\gamma)}{4}$$

and

$$\left\|\hat{P}_t - P\right\|_\infty \leq \frac{\epsilon(1-\gamma)^2}{4\gamma\left|\mathcal{S}\right|\left\|r\right\|_\infty}$$

for $t \geq t_0$. Finally, as $\mathcal{S} \times \mathcal{A}$ is finite and as $\sum_{t \in \mathcal{T}_{(s,a)}(\omega)} \alpha_t = \infty$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we may choose $t_k \geq \cdots \geq t_1 \geq t_0$ sufficiently far apart such that

$$e^{-\sum_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [t_{i-1}, t_i)} \alpha_\tau} \leq \frac{\epsilon(1-\gamma)^2}{8\gamma\left\|r\right\|_\infty}$$

for all $i \in \{1,\ldots,k\}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.

In particular, for any $i \in \{1,\ldots,k\}$, $t \geq t_i$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
&\left\|\hat{r}_t - r\right\|_\infty + \left(\frac{\gamma\left\|r\right\|_\infty}{1-\gamma}\right)\left(\left|\mathcal{S}\right|\left\|\hat{P}_t - P\right\|_\infty + 2e^{-\sum_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [t_{i-1}, t)} \alpha_\tau}\right) \\
&\qquad \leq \left\|\hat{r}_t - r\right\|_\infty + \left(\frac{\gamma\left\|r\right\|_\infty}{1-\gamma}\right)\left(\left|\mathcal{S}\right|\left\|\hat{P}_t - P\right\|_\infty + 2e^{-\sum_{\tau \in \mathcal{T}_{(s,a)}(\omega) \cap [t_{i-1}, t_i)} \alpha_\tau}\right) \\
&\qquad \leq \frac{3}{4}\epsilon(1-\gamma)
\end{aligned}$$

since $(\alpha_t)_{t\in\mathbb{N}_0}$ is non-negative, so it follows by inductive application of Lemma 4 that

$$\begin{aligned}
\left\|Q_t - q_M^*\right\|_\infty &\leq \gamma \max_{t' \in [t_k, t)} \left\|Q_{t'} - q_M^*\right\|_\infty + \frac{3}{4}\epsilon(1-\gamma) \\
&\leq \gamma \max_{t' \in [t_k, t)} \left(\gamma \max_{t'' \in [t_{k-1}, t')} \left\|Q_{t''} - q_M^*\right\|_\infty + \frac{3}{4}\epsilon(1-\gamma)\right) + \frac{3}{4}\epsilon(1-\gamma) \\
&= \gamma^2 \max_{t' \in [t_{k-1}, t)} \left\|Q_{t'} - q_M^*\right\|_\infty + \frac{3}{4}\epsilon(1-\gamma)(1+\gamma) \\
&\leq \cdots \\
&= \gamma^{k+1} \max_{t' \in [t_0, t)} \left\|Q_{t'} - q_M^*\right\|_\infty + \frac{3}{4}\epsilon(1-\gamma)(1+\gamma+\cdots+\gamma^k) \\
&\leq \left(\frac{\epsilon(1-\gamma)}{8\left\|r\right\|_\infty}\right)\left(\frac{2\left\|r\right\|_\infty}{1-\gamma}\right) + \frac{\frac{3}{4}\epsilon(1-\gamma)}{1-\gamma} \qquad \text{(Lemma 1)} \\
&= \epsilon
\end{aligned}$$

for all $t \geq t_k$ and, with that, the beast has been slain. □

References

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2): 447–469, 2000.

Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for q-learning. *Journal of machine learning Research*, 5(1), 2003.

Harold J. Kushner and Dean S. Clark. *Stochastic approximation methods for constrained and unconstrained systems / Harold J. Kushner, Dean S. Clark*. Springer-Verlag New York, 1978. ISBN 0387903410.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Csaba Szepesvári and Michael L Littman. Generalized markov decision processes: Dynamic-programming and reinforcement-learning algorithms. In *Proceedings of International Conference of Machine Learning*, volume 96, 1996.

John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

A Notation

A.1 Measure Theory

Definition 10. *A σ-algebra on a non-empty set $\mathcal{X}$ is collection of subsets $\mathcal{F}$ of $\mathcal{X}$ satisfying*

(i) $\varnothing \in \mathcal{F}$;

(ii) $\forall A \in \mathcal{F},\ \mathcal{X} \setminus A \in \mathcal{F}$; *and*

(iii) $\forall A_1, A_2, \cdots \in \mathcal{F},\ \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

In this case, we call $(\mathcal{X}, \mathcal{F})$ a measurable space.

Definition 11. *Given measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ as well as a function $A : \mathcal{X} \to \mathcal{Y}$, the σ-algebra induced by A is*

$$\sigma_{\mathcal{G}}(A) := \{A^{-1}(G) : G \in \mathcal{G}\}$$

and is usually denoted $\sigma(A)$ (when $\mathcal{G}$ is clear from context). Moreover, if $\sigma_{\mathcal{G}}(A) \subseteq \mathcal{F}$, we say that A is $\mathcal{F}/\mathcal{G}$-measurable, $\mathcal{F}$-measurable, or just measurable for short.

Remark 3. *Every non-empty set $\mathcal{X}$ admits at least one σ-algebra—namely $\mathcal{P}(\mathcal{X})$—and if $\{\mathcal{F}_i : i \in \mathcal{I}\}$ is a non-empty family of σ-algebras on $\mathcal{X}$, then $\bigcap_{i\in\mathcal{I}} \mathcal{F}_i$ is a σ-algebra on $\mathcal{X}$.*

Definition 12. *Given measurable spaces $(\mathcal{X}_i, \mathcal{F}_i)_{i\in\mathcal{I}}$, the product σ-algebra on $\bigtimes_{i\in\mathcal{I}} \mathcal{X}_i$*

$$\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i := \bigcap \{\mathcal{F} \text{ a } \sigma\text{-algebra on } \bigtimes_{i\in\mathcal{I}} \mathcal{X}_i : \forall i \in \mathcal{I},\ \pi_i \text{ is } \mathcal{F}/\mathcal{F}_i\text{-measurable}\}$$

is the smallest σ-algebra with respect to which each projection $\pi_i : \bigtimes_{j\in\mathcal{I}} \mathcal{X}_j \to \mathcal{X}_i$ is measurable.

Definition 13. *A probability measure on a measurable space $(\mathcal{X}, \mathcal{F})$ is $\mathbb{P} : \mathcal{F} \to [0, \infty]$ s.t.*

(i) $\mathbb{P}(\mathcal{X}) = 1$; *and*

(ii) $\forall A_1, A_2, \cdots \in \mathcal{F},\ (A_n)_{n\in\mathbb{N}}$ *pairwise disjoint* $\implies \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$.

Altogether, we call $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ a probability space and real-valued measurable functions on $\mathcal{X}$ are called random variables.

Definition 14. *The set of probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$ is denoted $\Delta(\mathcal{X}, \mathcal{F})$. If $\mathcal{X}$ is countable, we write $\Delta(\mathcal{X}) := \Delta(\mathcal{X}, \mathcal{P}(\mathcal{X}))$ for short.*

A.2 Function Spaces

Definition 15. *Let $\mathcal{I}$ be a non-empty set. The supremum norm on $\mathcal{I} \to \mathbb{R}$ ($\mathbb{R}^{\mathcal{I}}$ for short) is*

$$\|\cdot\|_{\mathcal{I},\infty} : \mathbb{R}^{\mathcal{I}} \to [0, \infty], x \mapsto \sup_{i\in\mathcal{I}} |x(i)|$$

and the set of bounded real-valued functions on $\mathcal{I}$ is

$$\ell^\infty(\mathcal{I}) := \{x \in \mathbb{R}^{\mathcal{I}} : \|x\|_{\mathcal{I},\infty} < \infty\}.$$

Frequently, $\mathcal{I}$ is clear from context, in which case we write $\|\cdot\|_\infty$ instead of $\|\cdot\|_{\mathcal{I},\infty}$.

B Banach's Fixed Point Theorem

In order to prove Theorem 1, we first need to know a little bit about metric spaces.

Definition 16. *Let E be non-empty. We call $d : E \times E \to [0, \infty)$ a metric on E when*

(i) $\forall x, y \in E,\ d(x, y) = 0 \iff x = y$;

(ii) $\forall x, y \in E,\ d(x, y) = d(y, x)$; *and*

(iii) $\forall x, y, z \in E,\ d(x, z) \leq d(x, y) + d(y, z)$.

In this case, the pair (E, d) is called a metric space.

Definition 17. *A metric space (E, d) is said to be complete when, for all sequences $(x_n)_{n\in\mathbb{N}_0}$ in E satisfying*

$$\forall \epsilon > 0,\ \exists n_0 \in \mathbb{N}_0,\ \forall n_1, n_2 \geq n_0,\ d(x_{n_1}, x_{n_2}) \leq \epsilon$$

(i.e. $(x_n)_{n\in\mathbb{N}_0}$ is a Cauchy sequence), we have that $d(x_n, x_\infty) \to 0$ for some $x_\infty \in E$.

Definition 18. *Let (E, d) be a metric space and let $\gamma \in [0, 1)$. We say that a map $T : E \to E$ is a γ-contraction on (E, d) when*

$$d(T(x), T(y)) \leq \gamma d(x, y)$$

holds for all $x, y \in E$.

Theorem 6 (Banach's Fixed Point Theorem)**.** *Let (E, d) be a complete metric space and let $T : E \to E$ be a γ-contraction for some $\gamma \in [0, 1)$. Then T admits a unique fixed point.*

The proof of Banach's fixed point theorem is a classic exercise in analysis. We omit it here but encourage the reader to try it on their own (hint: fix an arbitrary $x_0 \in E$ and show that $(T^n(x_0))_{n\in\mathbb{N}_0}$ is Cauchy by leveraging the fact that $\sum_{n=0}^{\infty} \gamma^n$ is a convergent series).

C Proofs of Results in Subsection 3.1

In any case, the latter fixed point theorem is all we need to show that optimal action-value functions exist and are unique in an MDP.

Proof of Theorem 1. Let $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$.

It is a straightforward exercise to verify that

$$d_\infty : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \times \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to [0, \infty], (q_1, q_1) \mapsto \|q_1 - q_2\|_\infty$$

is a metric on $\ell^\infty(\mathcal{S} \times \mathcal{A})$ and we omit the details.

As for completeness, let $(q_n)_{n\in\mathbb{N}_0}$ be a Cauchy sequence in $(\ell^\infty(\mathcal{S} \times \mathcal{A}), d_\infty)$ and let $\epsilon > 0$. For each $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $n_1, n_2 \in \mathbb{N}_0$, $|q_{n_1}(s, a) - q_{n_2}(s, a)| \leq d_\infty(q_{n_1}, q_{n_2})$, which implies that $(q_n(s, a))_{n\in\mathbb{N}_0}$ is a Cauchy sequence in $\mathbb{R}$ and hence, by completeness of $\mathbb{R}$, converges to

some $q_\infty(s,a) \in \mathbb{R}$; in particular, there is $n_{(s,a)} \in \mathbb{N}_0$ such that $|q_n(s,a) - q_\infty(s,a)| \leq \frac{\epsilon}{2}$ for $n \geq n_{(s,a)}$. Furthermore, there is $n_0 \in \mathbb{N}_0$ for which $d_\infty(q_{n_1}, q_{n_2}) \leq \frac{\epsilon}{2}$ for $n_1, n_2 \geq n_0$. Hence

$$\begin{aligned} d_\infty(q_n, q_\infty) &= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |q_n(s,a) - q_\infty(s,a)| \\ &\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_\infty(q_n, q_{\max\{n_0, n_{(s,a)}\}}) + \left| q_{\max\{n_0,n_{(s,a)}\}}(s,a) - q_\infty(s,a) \right| \right) \\ &\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(\frac{\epsilon}{2} + \frac{\epsilon}{2} \right) \\ &\leq \epsilon \end{aligned}$$

for $n \geq n_0$ and so $d_\infty(q_n, q_\infty) \to 0$. In particular, $d_\infty(q_{n_0}, q_\infty) < 1$ for some $n_0 \in \mathbb{N}_0$ and thus

$$\|q_\infty\|_\infty \leq \|q_{n_0}\|_\infty + d_\infty(q_{n_0}, q_\infty) < \infty,$$

i.e. $q_\infty \in \ell^\infty(\mathcal{S}\times\mathcal{A})$ so that the latter is complete with respect to d_∞ as claimed.

Finally, we claim that T_M^* is a γ-contraction on $(\ell^\infty(\mathcal{S}\times\mathcal{A}), d_\infty)$ as the conclusion will then follow immediately from Theorem 6. Indeed, for any $q_1, q_2 \in \ell^\infty(\mathcal{S}\times\mathcal{A})$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned} |T_M^* q_1(s,a) - T_M^* q_2(s,a)| &= \left| \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) \left(\max_{a'\in\mathcal{A}} q_1(s',a') - \max_{a'\in\mathcal{A}} q_2(s',a') \right) \right| \\ &\leq \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) \left| \max_{a'\in\mathcal{A}} q_1(s',a') - \max_{a'\in\mathcal{A}} q_2(s',a') \right| \\ &\leq \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) \max_{a'\in\mathcal{A}} |q_1(s',a') - q_2(s',a')| \\ &\leq \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) d_\infty(q_1,q_2) \\ &= \gamma d_\infty(q_1,q_2), \end{aligned}$$

which implies that $d_\infty(T_M^* q_1, T_M^* q_2) \leq \gamma d_\infty(q_1,q_2)$ as desired. □

Lastly, the proof of Lemma 1 follows from a straightforward calculation.

Proof of Lemma 1. Let $M = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$. Then, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned} |q_M^*(s,a)| &= |T_M^* q_M^*(s,a)| \\ &= \left| r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) \sup_{a'\in\mathcal{A}} q_M^*(s',a') \right| \\ &\leq |r(s,a)| + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) \left| \sup_{a'\in\mathcal{A}} q_M^*(s',a') \right| \\ &\leq \|r\|_\infty + \gamma \|q_M^*\|_\infty \sum_{s'\in\mathcal{S}} P(s'|s,a) \\ &= \|r\|_\infty + \gamma \|q_M^*\|_\infty . \end{aligned}$$

In particular, $\|q_M^*\|_\infty \leq \|r\|_\infty + \gamma \|q_M^*\|_\infty$ and hence $\|q_M^*\|_\infty \leq \frac{\|r\|_\infty}{1-\gamma}$. □