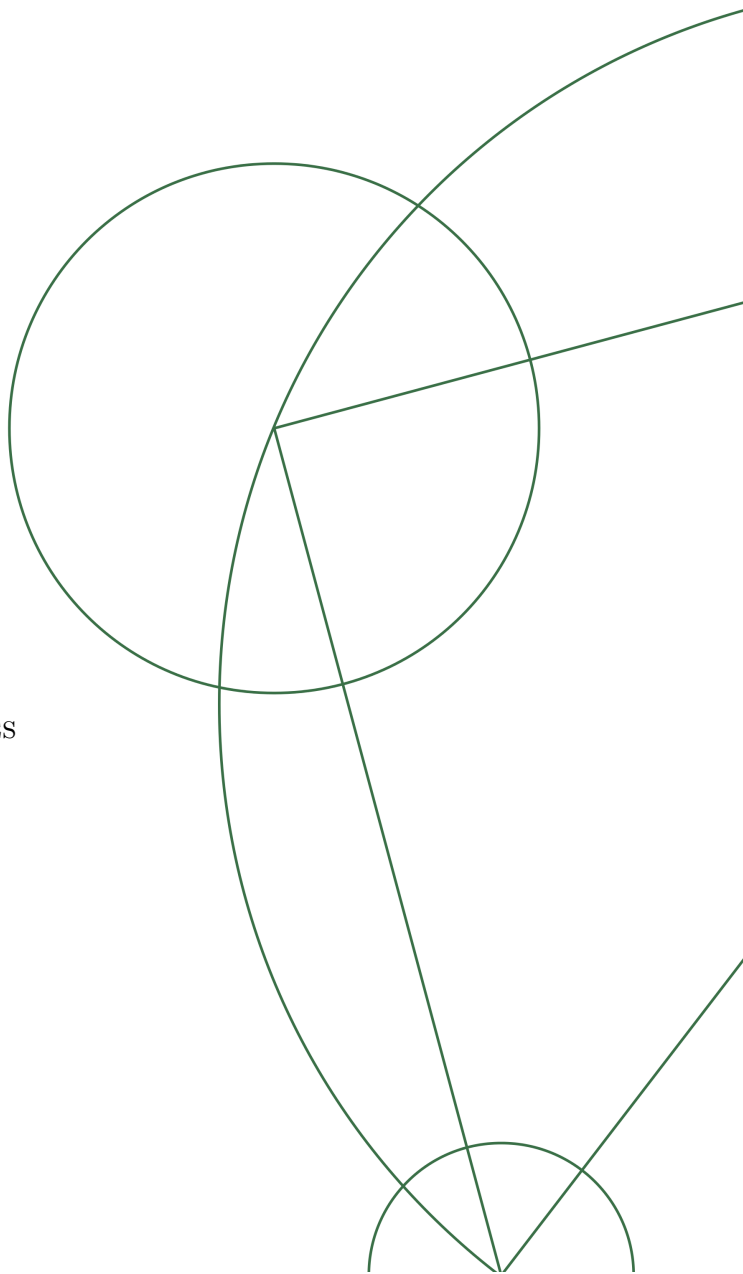
Department of Mathematical Sciences
UNIVERSITY OF COPENHAGEN

Jacob Harder

Theoretical aspects of Q-learning



Master thesis in Mathematics
Department of Mathematical Sciences
University of Copenhagen

Advisors
Stefan Horst Sommer

May 22, 2020

Abstract

This paper is mainly about the part of reinforcement learning that is called Q-learning, which is a category of algorithms which can *learn* from interaction with a decision process. We present the background theory for these algorithms, including basic theory on stochastic (decision) processes, existence of optimal polices and dynamic programming in the form of value iteration. Then Q-learning is analysed in a variety of settings. Among results other results we present and prove a convergence bound of the FQI algorithm obtained in the preprint [5, Fan et al. (2020+)], which guarantees convergence of Q-learning with deep neural networks for a broad class of continuous state-space Markov decision processes. In the course of this we discuss the relations between the various settings and their results.

Contents

Chapter 1

Introduction

In this thesis we give an introduction to Q-learning and discuss convergence results of *Q-learning* algorithms from its beginning in 1989 [18] to a result obtained in the preprint [5, Fan et al. (2020+)]. The introduction includes fundamental theory of the underlying field of *dynamic programming* and *reinforcement learning* and related topics such as *value iteration*.

1.1 Motivation

The topic was inspired by the performance of the algorithms implemented by [13, Mnih et al. (2015)]. In [13] it was shown how a single algorithm was able to achieve super human performances in a variety of problems, namely playing Atari 2600 video games, only using raw pixels and a reward (score) as input and a large number of interactions with the environment. The algorithm used in [13] is based on a Q-learning algorithm called the *deep Q-network* (DQN) algorithm.

The purpose of this thesis was initially to investigate what has been proven about the convergence of the DQN algorithm and what mathematical theory is relevant to establish such proofs.

The work began by considering the preprint [5, Fan et al. (2020+)] which claims to establish theoretical justification for the convergence of DQN to *the optimal Q-function*.

In the course of reading [5] it became clear that the background context of dynamic programming, reinforcement learning and value iteration was essential to understand the results of [5] and similar papers and also compare such results. Also questions as to in which settings optimal policies exists turned out to be a non-trivial question. Therefore in the end this thesis is partly about presenting the results of [5] and similar papers. And partly to build the background theory and compare these results.

1.2 What is Reinforcement Learning?

Reinforcement learning (RL) is a broad topic and a main branch of *machine learning* alongside *supervised* and *unsupervised learning*. Because of its broadness it overlaps with other disciplines such as *control theory* and dynamic programming.

In Reinforcement Learning we are concerned with finding an optimal policy for an agent in some environment. This environment is described by a sequence of state and action spaces $\mathcal{S}_1, \mathcal{A}_1, \mathcal{S}_2, \dots$ and rules $P_1, R_1, P_2, \dots$ specifying which states and rewards and likely to follow after some action is chosen. One can then specify rules π, called a *policy*, for how the agent should act in every

situation is the environment. Given an environment and a policy one obtains stochastic process, that is, a distribution on sequences of states, actions and rewards. One can then measure the performance of the policy by looking at the expected sum of rewards called the *value function* V_π of the policy. The goal of reinforcement learning is to find an optimal policy π^*, maximizing the value function.

V_π is viewed as function that evaluates for each *starting state* $s \in \mathcal{S}_1$ the expected total rewards when starting in state s and following policy π. There might therefore be different optimal policies for each such starting state. Traditionally one defines an optimal value function $V^*(s)$ by taking supremum over all policies $\sup_\pi V_\pi(s)$ for every state $s \in \mathcal{S}_1$. Then an optimal policy π^* should satisfy $V_{\pi^*} = V^*$, i.e. it should be optimal uniformly across all starting states $\mathcal{S}_1$. The existence of optimal policies defined in this way is a non-trivial question and we will devote some time on this.

A particular kind class of environments which are called Markov decision processes (MDPs), and work with the same state space $\mathcal{S}$, action space $\mathcal{A}$ and rules P, R throughout the process. They are by far the most well-studied environments. With an MDP and a value function V_1 satisfying certain assumptions one can obtain a policy π_1 by choosing actions leading to states with high values (according to V_1). Such policies are called *greedy policies*. We can then evaluate value of π_1 yielding a new value function V_2. This leads to an operator on the space of value functions called the *Bellman operator* and should satisfy $TV_1 = V_2$. This process of applying the Bellman operator can be continued indefinitely yielding a sequence of value functions and policies. Variations of this idea are called *value iteration* and *policy iteration*, and is derived from dynamic programming. The relations We show that value iteration converges to the optimal value functions given mild assumptions on the MDP. Furthermore we show that the optimal value functions is a fixed point of the Bellman operator: $TV^* = V^*$. This is called the *Bellman equation* and is the fundamental property that describes the concept of dynamic programming.

A further categorization of RL-algorithms can be made into *off-policy* and *on-policy* classes. This is simply whether the algorithm learns from data (states, actions and rewards) arising from following its own policy (on-policy) or it can learn from more arbitrary data (off-policy). This *more arbitrary data* could for example be the trajectory of another algorithm when interacting with a decision process, or simply state-action-reward pairs drawn from some distribution. In this thesis we will put emphasis on off-policy algorithms, to which both FQI and DQN belong.

1.3 What is Q-learning?

A problem with value functions defined on the set of states $\mathcal{S}$ is that picking optimal actions require knowledge of the transition dynamics P. Often we want to design algorithms that do not require such knowledge of P, so called *model-free* algorithms. To meet this requirement *Q-functions* are introduced, which evaluates the value of a state-action pair, instead of only a state.

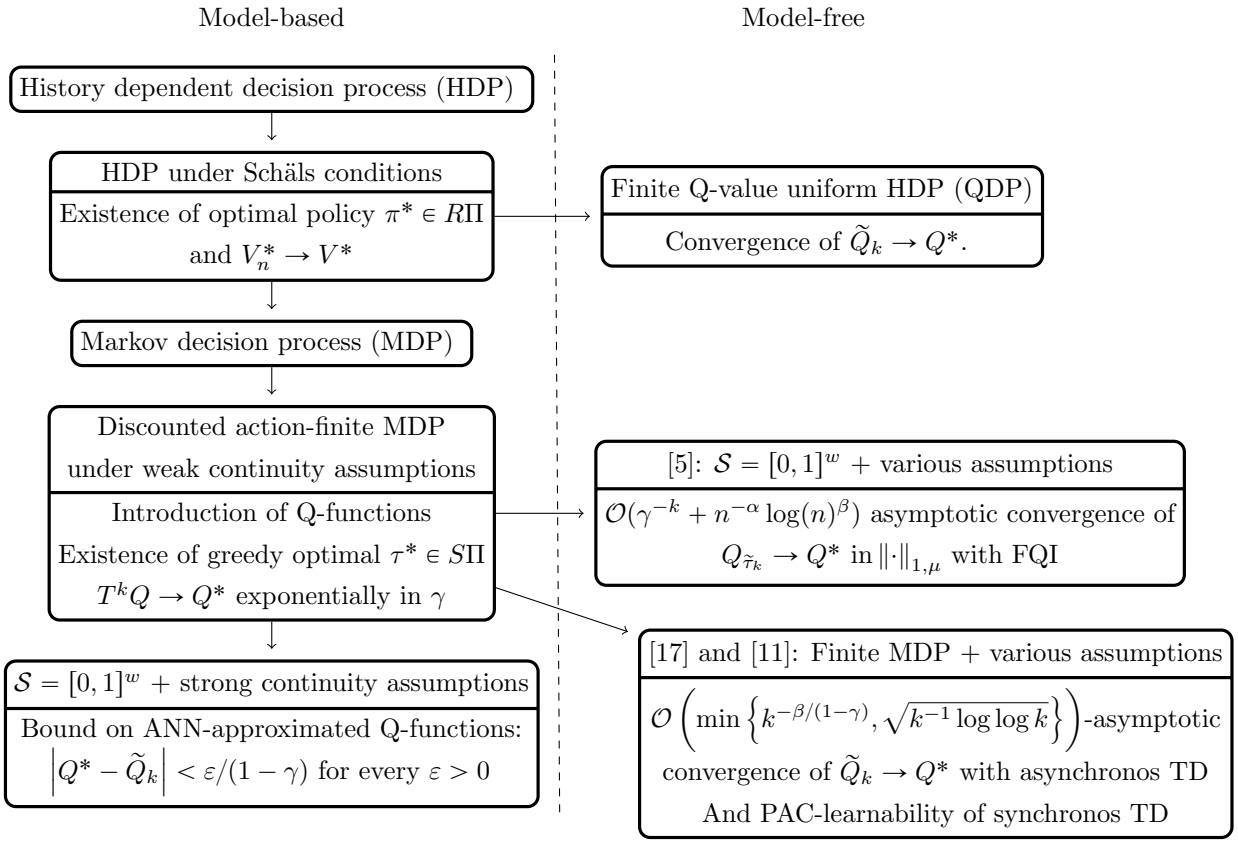
Given a Q-function Q, picking best actions according to Q now merely require maximization over Q itself. This is obviously an advantage if we are in situations where P is unknown. However it turns out also to be more convenient to work with computationally. In this thesis we show that value and policy iteration can be done for Q-functions in a virtually identical manner, when the process dynamics are known.

When the process dynamics are hidden designing algorithms becomes trickier. In such settings approaches to the problem fall in two categories. In the *indirect* approaches one attempts to

estimate the process dynamics first and then afterwards methods for the known-dynamics are applied. The *direct* approaches basically covers *the rest*. In the direct category we find the popular *temporal difference* algorithms on which *fitted Q-iteration* (FQI) and *deep Q-network* (DQN) as used in [13] is based. Many direct approaching such as FQI and DQN can be seen as stochastic approximations of the Bellman equation.

Q-learning is the category of algorithms that iteratively updates Q-functions in the attempt to improve the derived policy. The *deep network* in DQN comes from the concept of (*deep*) *artificial neural networks* (ANNs) which is a class of function approximators. The central idea of DQN is then simply to use deep neural networks to as approximators for the Q-functions.

1.4 Overview



Chapter 2

Decision models and value functions

To get started with talking about reinforcement learning, we need to define the most basic concept, the *environment* for the decision taking *agent*. This environment is formalized so called *decision process*. In order to define this we need the concept of a *probability kernel*

Definition 2.1 (Probability kernel)**.** Let $(\mathcal{X}, \Sigma_{\mathcal{X}}), (Y, \Sigma_{\mathcal{Y}})$ be measurable spaces. A function

$$\kappa(\cdot \mid \cdot) : \Sigma_{\mathcal{Y}} \times \mathcal{X} \to [0, 1]$$

is a $(\mathcal{X}, \Sigma_{\mathcal{X}})$-**probability kernel** on $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ provided

1. $B \mapsto \kappa(B \mid x) \in \mathcal{P}(\Sigma_{\mathcal{Y}})$ that is $\kappa(\cdot \mid x)$ is a probability measure for any $x \in \mathcal{X}$.
2. $x \mapsto \kappa(B \mid x) \in \mathcal{M}(\Sigma_{\mathcal{X}}, \Sigma_{\mathcal{Y}})$ that is $\kappa(B \mid \cdot)$ is $(\Sigma_{\mathcal{X}}\text{-}\Sigma_{\mathcal{Y}})$ measurable for any $B \in \Sigma_{\mathcal{Y}}$.

We then write $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$.

The following example shows how probability kernels are easily constructed.

Example 2.2. If $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is positive a measurable function with the property that

$$\forall x \in \mathcal{X} : \int f(x, y) \, \mathrm{d}\mu(y) = 1$$

then $\kappa(B \mid x) = \int_B f(x, y) \, \mathrm{d}\mu(y)$ defines a $\mathcal{X}$-probability kernel on $\mathcal{Y}$.

A handy property of kernels is

Proposition 2.3. Let $\kappa : \mathcal{X} \to \mathcal{Y}$ be a probability kernel and $f : \mathcal{X} \times \mathcal{Y} \to \overline{\underline{\mathbb{R}}}$ be measurable satisfying that $f(x, \cdot)$ is $\kappa(\cdot \mid x)$-integrable for every $x \in \mathcal{X}$. Then $x \mapsto \int f \, \mathrm{d}\kappa(\cdot \mid x)$ is measurable into $(\overline{\underline{\mathbb{R}}}, \overline{\underline{\mathbb{B}}})$.

Proof of proposition 2.3. Simple functions are measurable since κ is a kernel. Now extend by sums and limits. □

Here we denote $\overline{\underline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$ and $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$. These are endowed with the natural order topology (**??**) and the Borel σ-algebra (**??**) obtained from this.

We can now state the definition of a decision process

Definition 2.4 (History dependent decision process)**.** A (countable) **history dependent decision process** (HDP) is determined by

1. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})_{n\in\mathbb{N}}$ a measurable space of **states** for each timestep.
2. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})_{n\in\mathbb{N}}$ a measurable space of **actions** for each timestep.

   for each $n \in \mathbb{N} \cup \{\infty\}$ define the **history** spaces

$$\mathcal{H}_1 = \mathcal{S}_1, \quad \mathcal{H}_2 = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2, \quad \mathcal{H}_3 = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2 \times \mathbb{R} \times \mathcal{A}_2 \times \mathcal{S}_3$$

$$\mathcal{H}_n = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2 \times \mathbb{R} \times \mathcal{A}_2 \times \mathcal{S}_3 \times \mathbb{R} \times \cdots \times \mathcal{S}_n$$

$$\mathcal{H}_\infty = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2 \times \mathbb{R} \times \dots$$

3. $(P_n)_{n\in\mathbb{N}}$ a sequence of $\mathcal{H}_n \times \mathcal{A}_n \rightsquigarrow \mathcal{S}_{n+1}$ probability kernels called the **transition** kernels.
4. $(R_n)_{n\in\mathbb{N}}$ a sequence of $\mathcal{H}_{n+1} \rightsquigarrow \mathbb{R}$ probability kernels called the **reward** kernels.
5. $A_n(h_n) \subseteq \mathcal{A}_n$ a set of admissable actions for each $h_n \in \mathcal{H}_n$ and $n \in \mathbb{N}$.

With a HDP and an a way of choosing actions for each new state we can obtain sequence of states, actions and rewards, that is a history, by sampling from the kernels. To make precise what it means to choose actions we introduce the notion of a *policy*.

Definition 2.5 (Policy)**.** A (randomized) **policy** $\pi = (\pi_n)_{n\in\mathbb{N}}$ for a HDP is a sequence of probability kernels $\pi_n : \mathcal{H}_n \rightsquigarrow \mathcal{A}_n$, such that $\pi_n(A(h_i) \mid h_i) = 1$ for alle $h_i \in \mathcal{H}_i$, i.e. the policy chooses only admissable actions (with probability 1). The set of all policies we denote $R\Pi$.

With a HDP, a starting state S_1 and a policy π intuitively we should be able to obtain a history by sampling

- an action $A_1 \in A(H_1)$ from $\pi_1(\cdot \mid H_1)$ (where $H_1 = S_1$),
- a state $S_2 \in P(H_2)$ from $P(\cdot \mid H_1, A_1)$,
- a reward $R_1 \in \mathbb{R}$ from $R_1(\cdot \mid H_2)$,
- an action $A_2 \in A(H_2)$ from $\pi_2(\cdot \mid H_2)$
- and so on.

To make this precise we need some additional measure theory on probability kernels.

Theorem 2.6 (Integration of a kernel)**.** Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$. Then there exists a uniquely determined probability measure $\lambda \in \mathcal{P}(\Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y})$ such that

$$\lambda(A \times B) = \int_A \kappa(B, x)\, \mathrm{d}\mu(x)$$

We denote this measure $\lambda = \kappa\mu$.

Proof. For $G \in \Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y}$ and $x \in \mathcal{X}$ define $G^x := \{y \in \mathcal{Y} \mid (x, y) \in G\}$. It is easy to check that the map $x \mapsto \kappa(G^x \mid x)$ is measurable, using a Dynkin class argument. Thus we can define

$$\lambda(G) = \int \kappa(G^x \mid x)\, \mathrm{d}\mu(x)$$

Using this definition we see that $\lambda(\mathcal{X}\times\mathcal{Y}) = 1$ and by monotone convergence for disjoint $G_1, G_2, \ldots$

$$\lambda\left(\bigcup_{i\in\mathbb{N}} G_i\right) = \int \sum_{i=1}^{\infty} P_x(G_n^x)\,\mathrm{d}\mu(x) = \sum_{i=1}^{\infty} \lambda(G_i)$$

Uniqueness follows because the property

$$\lambda(A \times B) = \int_A \kappa(B, x)\,\mathrm{d}\mu(x)$$

should hold on the all product sets, which form an intersection-stable generating collection for $\Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}}$. □

Remark 2.7. In light of theorem 2.6 we can view a probability kernel as a mapping $\kappa : \mathcal{P}(X) \rightsquigarrow \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ defined by $\mu \mapsto \kappa\mu$.

Definition 2.8 (Composition of kernels)**.** Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\phi : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ be probability kernels. We define the composition $\phi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$ by

$$\phi\kappa(C \mid x) = \int \phi(C \mid x, y)\,\mathrm{d}\kappa(y \mid x)$$

.

Remark 2.9. Following remark 2.7 $\phi\kappa$ can be viewed as a mapping a from from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{X}\times\mathcal{Z})$. This is somewhat unsatisfactory. We are missing the intermediary space $\mathcal{Y}$. However writing $\phi(\kappa\mu)$ we obtain a measure on $\mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ as wanted. We will therefore use a slight abuse of notation and interpret compositions of kernels as including all intermediary spaces, when viewed as maps of measures, that is we will write $\phi\kappa\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. It is a trivial exercise to verify that composition when viewed this way is associative. That is if $\psi : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \rightsquigarrow \mathcal{W}$ is another probability kernel, then $((\psi\phi)\kappa)\mu = (\psi(\phi\kappa))\mu$.

Remark 2.10. When one has $\varphi : \mathcal{Y} \to \mathcal{Z}$ we can also use definition 2.8 since φ can be viewed as a $\mathcal{X} \times \mathcal{Y}$-kernel which does not depend on its input from $\mathcal{X}$. We write $\varphi \circ \kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$. This is often also referred to a *composition of kernels*. In fact it makes the class of measurable spaces into a category [10, Lawvere (1962)], with identity $\mathrm{id}_{\mathcal{X}}(\cdot \mid x) = \delta_x$.

Remark 2.11. Let $(\mathcal{X}_n, \Sigma_{\mathcal{X}_n})_{n\in\mathbb{N}}$ be a sequence of measurable spaces. For each $n \in \mathbb{N}$ define $\mathcal{X}^{\underline{n}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $\Sigma_{\mathcal{X}^{\underline{n}}} := \Sigma_{\mathcal{X}_1} \otimes \cdots \otimes \Sigma_{\mathcal{X}_n}$ and let $\kappa_n : \mathcal{X}^{\underline{n}} \rightsquigarrow \mathcal{X}_{n+1}$ be a probability kernel. Then by remark 2.9 $\kappa^{\underline{n}} := \kappa_n \ldots \kappa_1$ defines a map from $\mathcal{P}(\mathcal{X}_1)$ to $\mathcal{P}(\mathcal{X}^{\underline{n+1}})$.

Remark 2.9 allows us to make sense to finite decision processes. That is for any $n \in \mathbb{N}$, distribution $\mu \in \mathcal{P}(\mathcal{S}_1)$ of S_1 and policy $(\pi_1, \pi_2, \ldots) \in R\Pi$ we can get a distribution of the nth history $H_n \in \mathcal{H}_n$ by the composition of kernels

$$P_{n-1}\pi_{n-1}R_{n-2}P_{n-2}\pi_{n-2}\ldots R_2P_2\pi_2R_1P_1\pi_1\mu \in \mathcal{P}(\mathcal{H}_n)$$

We would like to extend this to a distribution on $\mathcal{H}_\infty$. To do this we will need

Theorem 2.12 (Ionescu-Tulcea extension theorem)**.** For every $\mu \in \mathcal{P}(\mathcal{X}_1)$ there exists a unique probability measure $\rho \in \mathcal{P}(\mathcal{X}^{\underline{\infty}})$ such that

$$\kappa^{\underline{n-1}}\mu(A) = \rho\left(A \times \prod_{k=n+1}^{\infty} \mathcal{X}_k\right), \qquad \forall A \in \Sigma_{\mathcal{X}^{\underline{n}}}, n \in \mathbb{N}$$

Proof. We refer to [8, Kallenberg (2002)] thm. 5.17. □

It is even possible to view the Ionescu-Tulcea construction as a kernel

Proposition 2.13 (Ionescu-Tulcea kernel)**.** Let μ_x denote the Ionescu-Tulcea measure of a sequence of probability kernels $\kappa_i : \mathcal{X}^{\underline{i}} \to \mathcal{X}_{i+1}$ with starting measure δ_x on $\mathcal{X}_1$ for any $x \in \mathcal{X}_1$. Then $\kappa(A \mid x) = \mu_x(A)$ defines a probability kernel $\kappa : \mathcal{X}_1 \to \mathcal{X}^{\underline{\infty}}$.

Proof. Since we already know that μ_x is a probability measure for any $x \in \mathcal{X}_1$, we just have to show that $\kappa(A \mid x) = \mu_x(A)$ is measurable as a function of x for all $A \in \Sigma_{\mathcal{X}^{\underline{\infty}}} = \bigotimes_{i=1}^{\infty} \Sigma_{\mathcal{X}_i}$. Let $\phi_A = x \mapsto \mu_x(A)$ for all $A \in \Sigma_{\mathcal{X}^{\underline{\infty}}}$ and define

$$\mathbb{G} = \left\{ A \in \bigotimes_{i=1}^{\infty} \Sigma_{\mathcal{X}_i} \;\middle|\; \phi_A \in \mathcal{M}(\Sigma_{\mathcal{X}_1}, \mathbb{B}_{[0,1]}) \right\}$$

The cylinder algebra

$$\mathbb{O} = \left\{ A_1 \times \cdots \times A_i \times \mathcal{X}_{i+1}, \ldots \;\middle|\; A_i \in \Sigma_{\mathcal{X}_i}, i \in \mathbb{N} \right\}$$

is a generator for $\Sigma_{\mathcal{X}^{\underline{\infty}}}$ stable under finite intersections. By contruction $\mathbb{O} \subseteq \mathbb{G}$ since

$$\phi_{A_1 \times \cdots \times A_i \times \mathcal{X}_{i+1} \times \ldots} = \kappa^{\underline{i-1}}(A_1 \times \cdots \times A_i \mid \cdot)$$

and any $\kappa^{\underline{i-1}}$ is a kernel making that function measurable. We will show that $\mathbb{G}$ is a Dynkin class. Then by Dynkins π-λ theorem (see theorem A.8)

$$\sigma(\mathbb{O}) = \Sigma_{\mathcal{X}^{\underline{\infty}}} \subseteq \mathbb{G}$$

implying that ϕ_A is measurable for all $A \in \Sigma_{\mathcal{X}^{\underline{\infty}}}$.

Clearly $\mathcal{X}^{\underline{\infty}}, \varnothing \in \mathbb{G}$ and if $A, B \in \mathbb{G}$ with $A \subseteq B$ then $\phi_{B \setminus A} = \phi_B - \phi_A \in \mathbb{G}$. Finally if $(B_n)_{n \in \mathbb{N}}$ is an ($\subseteq$-) increasing sequence in $\mathbb{G}$ then $\phi_{\bigcup_{n=1}^{\infty} B_n} = \lim_{n \to \infty} \phi_{B_n}$ is again measurable as it is a limit of measurable functions, showing that $\mathbb{G}$ is a Dynkin class. □

We will denote the Ionescu-Tulcea kernel $\ldots \kappa_2 \kappa_1$ or $\prod_{i=1}^{\infty} \kappa_i$ or simply $\kappa^{\underline{\infty}}$. The next lemma will come in handy when manipulating with integrals over kernel derived measures.

Lemma 2.14. The Ionescu-Tulcea kernel satisfies $\prod_{i=1}^{\infty} \kappa_i = (\prod_{i=2}^{\infty} \kappa_i)\kappa_1$.

Proof. Let $x \in \mathcal{X}_1$. Notice that by associativity of the composition of finitely many kernels $\kappa_n \ldots \kappa_1 \mu = (\kappa_n \ldots \kappa_2)(\kappa_1 \mu)$. This implies that

$$\left(\prod_{i=1}^{\infty} \kappa_i \mu \right) \left(A \times \prod_{k=n+1}^{\infty} \mathcal{X}_k \right) = \left(\left(\prod_{i=2}^{\infty} \kappa_i \right) \kappa_1 \mu \right) \left(A \times \prod_{k=n+1}^{\infty} \mathcal{X}_k \right)$$

for all $n \in \mathbb{N}$ and $A \in \Sigma_{\mathcal{X}^{\underline{n}}}$. By the uniqueness in theorem 2.12 we are done. □

Let $\mu \in \mathcal{P}(\mathcal{S}_1)$ be a measure on the first state space. By theorem 2.12 a HDP and a policy π gives rise to a kernel $\kappa_\pi : \mathcal{P}(\mathcal{S}_1) \to \mathcal{P}(\mathcal{H}_\infty)$, namely

$$\kappa_\pi = \ldots R_2 P_2 \pi_2 R_1 P_1 \pi_1 \mu \tag{2.1}$$

In particular $\kappa_\pi \mu$ can be interpreted as the stochastic process arising from sampling the first state from μ and then follow π for a countable number of steps. We will denote expectation with respect to $\kappa_\pi \mu$ by $\mathbb{E}_\mu^\pi$. In the case where $\kappa_\pi \delta_s$ can be interpreted as the stochastic process arise from starting in state s and following policy π. We will abuse notation slightly, writing $\kappa_\pi \delta_s = \kappa_\pi s$ and $\mathbb{E}_{\delta_s}^\pi = \mathbb{E}_s^\pi$.

2.1 Policy evaluation and value functions

The next step is to evaluate how *good* a policy is. To this end we introduce *value functions*. In order for the sum of finitely many rewards to have a meaningful expected value we will need one of the following conditions:

Condition F^- (Reward finity from above)**.** $\int_{[0,\infty]} x \, \mathrm{d}R_i(x \mid h) < \infty$ for all $h \in \mathcal{H}_{i+1}$ and $i \in \mathbb{N}$

Condition F^+ (Reward finity from below)**.** $\int_{[-\infty,0]} x \, \mathrm{d}R_i(x \mid h) > -\infty$ for all $h \in \mathcal{H}_{i+1}$ and $i \in \mathbb{N}$

Under (F^+) or (F^-) following definition makes sense

Definition 2.15 (Finite horizon value function)**.** Let $\underline{R}_i : \mathcal{H}_\infty \to \overline{\mathbb{R}}$ be the projection map onto the ith reward. We define the function $V_{n,\pi} : \mathcal{S}_1 \to \overline{\mathbb{R}}$ by

$$V_{n,\pi}(s_1) = \mathbb{E}_s^\pi \sum_{i=1}^{n} \underline{R}_i$$

called the kth finite horizon value function. When $n = 0$ we say $V_{0,\pi} = V_0 := 0$ for any π.

The finite horizon value function measures the expected total reward of starting in state s and then follow the policy π for n steps. This way it measures the *value* of that particular state given a policy and *horizon* (number of steps).

We would like to extend this to an infinite horizon value function, i.e. letting n tend to ∞. To ensure that the integral is well-defined we introduce the following conditions

Condition P (Reward non-negativity)**.** $R_i([0,\infty] \mid h) = 1$, $\forall h \in \mathcal{H}_{i+1}, i \in \mathbb{N}$

Condition N (Reward non-positivity)**.** $R_i([-\infty,0] \mid h) = 1 \ \forall h \in \mathcal{H}_{i+1}, i \in \mathbb{N}$

Condition D (Discounting)**.** There exist a bound $R_{\max} > 0$ and a $\gamma \in [0,1)$ called the **discount** factor such that $R_i([-R_{\max}\gamma^i, R_{\max}\gamma^i]) = 1 \ \forall h \in \mathcal{H}_{i+1}, i \in \mathbb{N}$

Remark 2.16. The letters F^+, F^-, P, N and D are adopted from [1].

Definition 2.17. We define the infinite horizon value function by

$$V_\pi(s) = \mathbb{E}_s^\pi \lim_{n\to\infty} \sum_{i=1}^{n} \underline{R}_i$$

The infinite horizon value function V_π measures the expected total reward after following the policy π an infinite number of steps.

Remark 2.18. Whenever we are working with the finite horizon value function we will always assume that either (F^+) or (F^-) holds without stating this explicitly. If a result only holds under e.g. (F^+) we will of course be explicit about this be marking it accordingly with a (F^+).

Similarly whenever we work with the infinite horizon value function we will always assume that at least one of (P), (N) or (D) holds. We will mark propositions and theorems by e.g. (D) (P) when the result only holds for if discounting *or* reward non-negativity is assumed. Note that obviously (P) implies F^+ and (N) implies F^-.

We mention some immediate properties of the finite and infinite horizon value functions

Proposition 2.19.

1. $V_{n,\pi}, V_\pi$ are measurable into $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$ and under (D) they are integrable with respect to $\kappa_\pi(\cdot \mid s)$ for any $\pi \in R\Pi$.

2. $\lim_{n\to\infty} V_{n,\pi} = V_\pi$ for all $\pi \in R\Pi$.

3. Under (D) for any $\pi \in R\Pi$

$$\left|V_{n,\pi}\right|, |V_\pi| \leqslant R_{\max}(1-\gamma) < \infty$$

Proof.

1. Use proposition 2.3.

2. By monotone or dominated convergence.

3. For any $\pi \in R\Pi$
$$|V_\pi(s)| \leqslant \mathbb{E}_s^\pi \sum_{i\in\mathbb{N}} |\underline{R}_i| \leqslant \sum_{i\in\mathbb{N}} \gamma^{i-1} R_{\max} = R_{\max}/(1-\gamma)$$
This also covers $V_{n,\pi}$.

□

Remark 2.20. As this bound will occur again and again we denote it

$$V_{\max} := R_{\max}/(1-\gamma) \tag{2.2}$$

2.1.1 The optimal value function

Definition 2.21 (Optimal value functions)**.**

$$V_n^*(s) := \sup_{\pi\in R\Pi} V_{n,\pi}(s) \qquad\qquad V^*(s) := \sup_{\pi\in R\Pi} V_\pi(s)$$

This is called the **optimal value function** (and the nth optimal value function). A policy $\pi^* \in R\Pi$ for which $V_{\pi^*} = V^*$ is called an **optimal policy**. If $V_{n,\pi^*} = V_n^*$ it is called n-optimal.

Proposition 2.22. Under (D) we have $\left|V_k^*\right|, |V^*| \leqslant V_{\max}$.

Proof. All terms in the suprema are within this bound. □

Remark 2.23. It is known that the optimal value function might not be Borel measurable (see ex. 2 p. 233 [1]). Perhaps this is not suprising since we are taking a supremum over sets of policies which might have cardinality of at least the continuum.

At this point some relevant questions can be asked.

1. To which extend does an optimal policy π^* exist?

2. Does V_n^* converge to V^*?

3. When can optimal policies be chosen to be Markov, deterministic, etc.?

4. Can an algorithm be designed to efficiently find V^* and π^*?

In a quite general setting, questions 1 and 2 is investigated in [16, Schäl (1975)]. Here some additional structure on our process is imposed.

Definition 2.24 (Standard Borel measurable space)**.** A measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ is called **standard Borel** if $\mathcal{X}$ is Polish space, that is a seperable completely metrizable space, and $\Sigma_{\mathcal{X}}$ is the Borel σ-algebra of $\mathcal{X}$, that is the σ-algebra generated by all open sets.

Setting 1 (Schäl)**.**

1. $V_\pi < \infty$ for all policies $\pi \in R\Pi$.
2. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})$ are all standard Borel.
3. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})$ are all standard Borel.
4. The set of admissible actions $A_n(h_n)$ is compact for any $h_n \in \mathcal{H}_n,\ n \in \mathbb{N}$.
5. The kernels $(P_n, R_n)_{n\in\mathbb{N}}$ are independent of rewards in the process.
6. $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N\geqslant n} \sup_{\pi\in R\Pi} \sum_{t=n+1}^{N} \mathbb{E}_s^\pi \underline{R}_n \to 0, \quad n \to \infty$

In this setting Schäl introduced two sets of criteria for the existence of an optimal policy:

Condition S.

1. The function
$$(a_1, a_2, \ldots, a_n) \mapsto P_n(\cdot \mid s_1, a_1, s_2, a_2, \ldots, s_n, a_n)$$
is set-wise continuous (hence the name **S**) for all $s_1, \ldots, s_n \in \mathcal{S}^{\underline{n}}$.
2. r_n is upper semi-continuous.

Condition W.

1. The function
$$(h_n, a_n) \mapsto P_n(\cdot \mid h_n, a_n)$$
is weakly continuous (hence the name **W**).
2. r_n is continuous.

Theorem 2.25 (Schäl)**.** Under setting 1 when either (S) or (W) hold then

1. There exist an optimal policy $\pi^* \in R\Pi$.
2. $V_n^* \to V^*$ as $n \to \infty$.

Proof. We refer to [16]. □

Corollary 2.26. Under setting 1 when either (S) or (W) hold then V^* is (Borel) measurable.

Proof. Since by theorem 2.25 there exists an optimal policy π^* we have $V^* = V_{\pi^*}$ which is measurable due to proposition 2.19. □

Schäls theorem tells us that optimal policies exist in a wide class of decision processes. In many cases we are looking at processes in which the next state in independent of the history, that is *Markov*. In such cases it makes sense to ask if optimal policies can be chosen to also be Markov. Such questions will be addressed in the next section.

2.2 Markov decision processes

Definition 2.27 (Markov decision process)**.** A **Markov decision process** (MDP) consists of

1. $(\mathcal{S}, \Sigma_{\mathcal{S}})$ a measurable space of states.
2. $(\mathcal{A}, \Sigma_{\mathcal{A}})$ a measurable space of actions.
3. $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ a transition kernel.
4. $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow \overline{\underline{\mathbb{R}}}$ a reward kernel.
5. An optional disount factor $\gamma \in [0, 1]$ (when not discounting put $\gamma = 1$).
6. $A(s) \subseteq \mathcal{A}$ a set of admissable actions for each $s \in \mathcal{S}$.

This is a special case of the history dependent decision process (definition 2.4) with

- $\mathcal{S}_1 = \mathcal{S}_2 = \cdots = \mathcal{S}, \quad \mathcal{A}_1 = \mathcal{A}_2 = \cdots = \mathcal{A}.$
- P_n depends only on s_n and a_n and does not differ with n, i.e. $P_n(\cdot \mid s_1, \ldots, s_n, a_n) = P(\cdot \mid s_n, a_n)$ for all $n \in \mathbb{N}$.
- R_n depends only on s_n and a_n and does not differ with n except for a potential discount. I.e. $R = R_n/\gamma^{n-1}$ for all $n \in \mathbb{N}$

We will write P instead of P_n understanding kernel compositions as if using P_n.

Remark 2.28. Expectations over any reward R_i occuring in an MDP can be computed from a function $r : \mathcal{S} \times \mathcal{A} \to \overline{\underline{\mathbb{R}}}$, defined by $r(s, a) = \int r' \, \mathrm{d}R(r' \mid s, a)$. To see this note that

$$\begin{aligned}
\mathbb{E}^{\pi}_{\mu} \underline{R}_i &= \int \underline{R}_i \, \mathrm{d}\kappa_{\pi}\mu \\
&= \int r_i \, \mathrm{d}R_i P \pi_i \ldots R_1 P \pi_1 \mu(s_1, a_1, \ldots, s_{i+1}, r_i) \\
&= \int \gamma^{i-1} r(s_i, a_i) \, \mathrm{d}\pi_i \ldots R_1 P \pi_1 \mu(s_1, a_1, \ldots, s_i, a_i) \\
&= \mathbb{E}^{\pi}_{\mu} \gamma^{i-1} r(\underline{S}_i, \underline{A}_i)
\end{aligned}$$

where $\underline{S}_i, \underline{A}_i$ are projection onto the ith state and action.

Remark 2.29. One could ask if it is possible to embed a HDP into an MDP by setting $\mathcal{S} := \bigcup_{i \in \mathbb{N}} \mathcal{S}_i$ and $\mathcal{A} := \bigcup_{i \in \mathbb{N}} \mathcal{A}_i$ or similar. One attempt at this can be found in [1] chapter 10, but this will not be covered here. Note however that whatever properties, such as those in setting 1, one assumes regarding the spaces $\mathcal{S}_1, \mathcal{A}_1, \ldots$, one must reconsider if each such property hold in the new constructed MDP.

Intuitively when the environment is a Markov decision process it should not be necessary that policies depend on the history. To talk about this topic we introduce

Definition 2.30 (Policy classes)**.** A policy $\pi = (\tau_1, \tau_2, \ldots) \in R\Pi$ is called **Markov** if it only depends on the last state is the history. That is there exist $\tau_1, \tau_2, \cdots : \mathcal{S} \rightsquigarrow \mathcal{A}$ such that $\tau_i(\cdot \mid s_1, \ldots s_i) = \tau_i(\cdot \mid s_i)$. We denote the set of (random) Markov policies by $M\Pi$. If $\tau_1 = \tau_2 = \ldots$ the Markov policy is called **stationary** and the set of them denote by $S\Pi$. Furthermore π is called **deterministic** if all τ_i are degenerate, i.e. for all i we have $\tau_i(\{a_i\} \mid h_i) = 1$ for some $a_i \in \mathcal{A}_i$. We denote the deterministic version of the policy classes by the letter D.

Remark 2.31. We have the following inclusions of policy classes

$$\begin{array}{ccc} S\Pi & \subseteq M\Pi & \subseteq R\Pi \\ \cup\!| & \cup\!| & \cup\!| \\ DS\Pi & \subseteq DM\Pi & \subseteq D\Pi \end{array}$$

Note that stationary policies might not exist in HDPs, but always exist for MDPs. A policy $(\tau_1, \tau_2, \dots) \in R\Pi$ is deterministic if and only if there exist measurable functions $\varphi_n : \mathcal{H}_n \to \mathcal{A}$ such that $\tau_n(\cdot \mid h_n) = \delta_{\varphi_n(h_n)}$. Therefore we shall sometimes write $\tau_n(h_n) = \varphi_n(h_n)$, viewing τ_n as a function. For convenience will view stationary policies $\tau \in S\Pi$ interchangeably as kernels $\tau : \mathcal{S} \rightsquigarrow \mathcal{A}$ and as the policy $(\tau, \tau, \dots)$.

We will prove that in MDPs under mild assumptions the optimal policy can be chosen to be deterministic, Markov, and even stationary. Before we do this we define some important tools for studying MDPs.

Definition 2.32 (The T-operators)**.** For a stationary policy $\tau \in S\Pi$ and measurable $V : \mathcal{S} \to \overline{\mathbb{R}}$ we define the operators

$$T_\tau V := s \mapsto \int r(s,a) + \gamma V(s') \, \mathrm{d}(P\tau)(a, s' \mid s)$$

$$TV := s \mapsto \sup_{a \in A(s)} T_a V(s)$$

where $T_a = T_{\delta_a}$ for $a \in A(s)$.

Remark 2.33. For the integral to make sense we assume under (D) that V is bounded, under (P) that $V \geqslant 0$ and under (N) that $V \leqslant 0$.

The T operator is sometimes called the *Bellman-operator*. It is harder to work with than T_τ because it envolves a supremum. Therefore we will first take a closer look at properties of T_τ.

Proposition 2.34 (Properties of the T_τ-operator)**.** Let $\pi = (\tau_1, \tau_2, \dots) \in M\Pi$ be a Markov policy, and $\tau \in S\Pi$ be a stationary policy.

1. T_τ is measurable and commutes with limits.
2. $V_{k,\pi} = T_{\tau_1} V_{k-1,(\tau_2,\dots)} = T_{\tau_1} \dots T_{\tau_k} V_0$.
3. $V_\pi = \lim_{k\to\infty} T_{\tau_1} \dots T_{\tau_k} V_0$
4. For the stationary policy τ we have $T_\tau V_\tau = V_\tau$.
5. (D) T and T_τ are γ-contractive on $\mathcal{L}_\infty(\mathcal{S})$.
6. (D) V_τ is the unique bounded fixed point of T_τ in $\mathcal{L}_\infty(\mathcal{S})$

Remark 2.35. Here $\mathcal{L}_\infty(\mathcal{S})$ denotes the set of essentially bounded measurable functions on $\mathcal{S}$. Since we interested in integration with respect to the variety of measures (arising as $(P \circ \tau)\mu$ for some μ) we must assume essential boundedness with respect to all such measures. Note that the set of every-bounded measurable functions is a subset of $\mathcal{L}_\infty(\mathcal{S})$ defined in this way, so it is non-empty. In the situation where one can prove that all measures in question are absolutely continuous w.r.t. some measure ν then it is enough to ensure boundedness ν-almost everywhere.

Proof of proposition 2.34.

1. Measurability is by proposition 2.3 the rest follows by monotone or dominated convergence.

2. We have

$$\begin{aligned}
&T_{\pi_1} V_{k,(\pi_2,\dots)}(s_1) \\
&= \int r(s_1,a_1) + \gamma \int \sum_{i=2}^{k+1} \gamma^{i-2} r(s_i,a_i)\ \mathrm{d}\kappa_{(\pi_2,\dots)}(a_2,s_3,a_3,\cdots \mid s_2)\ \mathrm{d}P\pi_1(a_1,s_2 \mid s_1) \\
&= \int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i)\ \mathrm{d}\dots P\pi_2 P\pi_1(a_1,s_2,\cdots \mid s_1) \\
&= \int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i)\ \mathrm{d}\kappa_\pi(a_1,s_2,\cdots \mid s_1) \\
&= V_{k+1,\pi}(s_1)
\end{aligned}$$

   Now use this inductively.

3. This is by 2. and a monotone or dominated convergence.

4. By 3. $T_\pi V_\pi = T_\pi \lim_{k\to\infty} T_\pi^k V_0 = \lim_{k\to\infty} T_\pi^{k+1} V_0 = V_\pi$.

5. Let $V, V' \in \mathcal{L}_\infty(\mathcal{S})$ and let $K = \left\|V - V'\right\|_\infty$. Then since the rewards are bounded

$$\left|T^\pi V - T^\pi V'\right| = \gamma \left| \int V(s') - V'(s')\ \mathrm{d}P\pi(s' \mid s) \right| \leqslant \gamma K$$

   For T use the same argument and the fact that $\left|\sup_x f(x) - \sup_y g(y)\right| \leqslant \left|\sup_x f(x) - g(x)\right|$ for any $f, g : X \to \underline{\mathbb{R}}$.

6. By 4., 5. and Banach fixed point theorem.

□

2.2.1 Greedy policies

Definition 2.36. Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A} \in S\Pi$ be a stationary policy and let $V : \mathcal{S} \to \overline{\mathbb{R}}$ be a measurable value-function. We define

$$G_V(s) = \operatorname*{argmax}_{a \in A(s)} T_a V(s) \subseteq A(s)$$

as the set of **greedy** actions w.r.t. V. If for which there exists a measurable $G_V^\tau(s) \subseteq G_V(s)$ such that

$$\tau(G_V^\tau(s) \mid s) = 1$$

for every $s \in \mathcal{S}$, then τ is called greedy w.r.t. V. We will often denote a V-greedy policy by τ_V.

In order to talk about existence of greedy policies we need some additional structure on our MDP.

Definition 2.37 (Borel σ-algebra)**.** For a topological space the **Borel** σ-algebra is the smallest σ-algebra containing all open sets.

Definition 2.38 (Weak topology)**.** Let $\mathcal{X}$ be a metrizable space equipped with the Borel σ-algebra. Consider the family of subsets of $\mathcal{P}(\mathcal{X})$

$$\mathcal{V} := \left\{ V_\varepsilon(p,f) \;\middle|\; \varepsilon > 0, p \in \mathcal{P}(\mathcal{X}), f \in C(\mathcal{X}) \right\}, \text{ where } V_\varepsilon(p,f) := \left\{ q \in \mathcal{P}(\mathcal{X}) \;\middle|\; \left| \int f\ \mathrm{d}q - \int f\ \mathrm{d}p \right| < \varepsilon \right\}$$

and where $C(\mathcal{X})$ denote the set of continuous functions $\mathcal{X} \to \mathbb{R}$. The **weak** topology on $\mathcal{P}(\mathcal{X})$ is the coarsest topology containing $\mathcal{V}$.

Recall (definition 2.24) that a measurable space is standard Borel if it is Polish and equipped with the Borel σ-algebra.

Proposition 2.39. Let $\mathcal{X}$ be a standard Borel measurable space. Consider the space $\mathcal{P}(\mathcal{X})$ of probabilty measures on $\mathcal{X}$ equipped with the weak topology. Then $\mathcal{P}(\mathcal{X})$ is standard Borel. If furthermore $\mathcal{X}$ is compact then $\mathcal{P}(\mathcal{X})$ is also compact.

Proof. We refer to [1] cor.7.25.1 and prop.7.22. □

Definition 2.40 (Continuous kernel)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be standard Borel measurable spaces. A probability kernel $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is **continuous** if the map

$$\gamma_\kappa : \mathcal{X} \to \mathcal{P}(\mathcal{Y}) = x \mapsto \kappa(\cdot \mid x)$$

is continuous.

Definition 2.41 (Semicontinuity)**.** Let $\mathcal{X}$ be a topological space and $f : \mathcal{X} \to \overline{\mathbb{R}}$ be a extended real-valued function. Then f is **upper** semicontinuous at $x_0 \in \mathcal{X}$ if for every $y > f(x_0)$ there exists a neighborhood U of x_0 such that $f(x) < y$ for all $x \in U$. If $-f$ is upper semicontinuous, then f is **lower** semicontinuous.

Proposition 2.42.

1. If $f,\ g : \mathcal{X} \to \overline{\mathbb{R}}$ are upper (lower) semicontinuous then $f + g$ is upper (lower) semicontinuous.
2. If furthermore g is continuous and non-negative then fg is upper (lower) semicontinuous.
3. If $(f_i)_{i \in I}$ are an arbitrary collection of upper (lower) semicontinuous functions then the infimum $\inf_{i \in I} f_i$ (supremum $\sup_{i \in I} f_i$) is again upper (lower) semicontinuous.

Proposition 2.43. Let $\mathcal{X}$ and $\mathcal{Y}$ be separable metrizable and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a continuous stochastic kernel. Let $f : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be Borel measurable, bounded from below or above. Define

$$\lambda(x) := \int f(x, y)\ \mathrm{d}\kappa(y \mid x)$$

Then

- f upper semicontinuous and bounded from above implies that λ is upper semicontinuous and bounded from above.
- f lower semicontinuous and bounded from below implies that λ is lower semicontinuous and bounded from below.

Proof. □

Proposition 2.44. A upper (lower) semicontinuous function $f : \mathcal{X} \to \overline{\mathbb{R}}$ on a compact set $\mathcal{X}$ attains its supremum (infimum). That is there exists an $x^* \in \mathcal{X}$ ($x_* \in \mathcal{X}$) such that $f(x^*) = \sup_{x \in \mathcal{X}} f(x)$ ($f(x_*) = \inf_{x \in \mathcal{X}} f(x)$).

Proof. □

Proposition 2.45. Let $\mathcal{X}$ be metrizable, $\mathcal{Y}$ compact metrizable, $\Gamma \subseteq \mathcal{X} \times \mathcal{Y}$ be closed with $\rho_{\mathcal{X}}(\Gamma) = \mathcal{X}$, where $\rho_{\mathcal{X}}$ is projection onto $\mathcal{X}$ and let $f : \Gamma \to \overline{\mathbb{R}}$ be upper semicontinuous.

$$f^* : \mathcal{X} \to \overline{\mathbb{R}} = x \mapsto \sup_{y \in \Gamma(x)} f(x, y)$$

where $\Gamma(x) = \{y \in \mathcal{Y} \mid (x, y) \in \Gamma\}$. Then f^* is upper semicontinuous and there exists a Borel-measurable function $\varphi : \mathcal{X} \to \mathcal{Y}$ such that $\mathrm{Gr}(\varphi) \subseteq \Gamma$ and $f(x, \varphi(x)) = f^*(x)$.

Setting 2 (Greedy MDP)**.**

1. $\mathcal{S}$ and $\mathcal{A}$ are standard Borel.
2. The set of admissable actions $A(s) \subseteq \mathcal{A}$ is compact for all $s \in \mathcal{S}$ and $\Gamma = \{(s, a) \in \mathcal{S} \times \mathcal{A} \mid a \in A(s)\}$ is a closed subset of $\mathcal{S} \times \mathcal{A}$.
3. The transition kernel P is continuous.
4. The expected reward function $r = \int r' \, \mathrm{d}R(r' \mid \cdot)$ is upper semicontinuous and bounded from above.

Since $\mathcal{S}$ is now a topological space the property of semicontinuity makes sense for value functions.

Proposition 2.46. Under setting 2 let $V : \mathcal{S} \to \overline{\mathbb{R}}$ be upper semicontinuous and bounded from above. Then the following holds

1. For every $s \in \mathcal{S}$ we have that $(s, a) \mapsto T_a V(s)$ is upper semicontinuous and bounded from above.
2. For every $s \in \mathcal{S}$ we have that $G_V(s)$ is non-empty.
3. There exist a deterministic greedy policy τ_V for V.
4. $T_{\tau_V} V = TV$ and TV is upper semicontinuous.

Proof.

1. This is a consequence of proposition 2.42 and proposition 2.43 since r is upper semicontinuous.
2. Since by 1. $(s, a) \mapsto T_a V(s)$ is upper semicontinuous, this follows by proposition 2.44.
3. By proposition 2.45 there exists a Borel-measurable function $\varphi : \mathcal{S} \to \mathcal{A}$ with $\mathrm{Gr}(\varphi) \subseteq \Gamma$ such that
$$T_{\varphi(s)} V(s) = \sup_{a \in A(s)} T_a V(s)$$
thus $\varphi(s) \in \mathrm{argmax}_{a \in A(s)} T_a V(s) = G_V(s)$. Therefore the induced deterministic policy
$$\tau_V(\cdot \mid s) = \delta_{\varphi(s)}$$
is greedy with respect to V.
4. By definition $T_{\tau_V} V(s) = T_{\mathrm{argmax}_{a \in A(s)} T_a V(s)} V(s) = \sup_{a \in A(s)} T_a V(s) = TV(s)$.

□

2.2.2 Existence of optimal policies

In the light of proposition 2.46 (and induction) since $V_0^* = 0$ is upper semicontinuous for any $k \in \mathbb{N}$ $T^k V_0^*$ is upper semicontinuous and thus has an associated greedy policy $\tau_{T^k V_0^*}$ which we will denote τ_k^*.

Proposition 2.47. Under setting 2 we have that

$$V_k^* = T^k V_0^* = T_{\tau_{k-1}^*} \dots T_{\tau_0^*} V_0^* = V_{k,(\tau_{k-1}^*,\dots,\tau_0^*)}$$

and this is an upper semicontinuous function. Thus $(\tau_{k-1}^*, \dots, \tau_0^*)$ is a deterministic k-optimal policy for any $k \in \mathbb{N}$.

Proof. As induction basis observe that $0 = V_0 = V_0^*$ is upper semicontinuous. Assume that $T^{k-1} V_0 = V_{k-1}^*$ is upper semicontinuous.

$$\begin{aligned}
V_k^*(s) &= \sup_{\pi \in R\Pi} \int \sum_{i=1}^{k} \gamma^{i-1} \underline{R}_i \, \mathrm{d}\kappa_\pi(\cdot \mid s) \\
&= \sup_{\pi \in R\Pi} \int r(s,a) + \gamma \left(\sum_{i=1}^{k-1} \gamma^{i-1} \underline{R}_i \, \mathrm{d}\kappa_{(\pi_2,\pi_3,\dots)}(\cdot \mid s,a,s') \, \mathrm{d}P(s' \mid s,a) \right) \mathrm{d}\pi_1(a \mid s) \\
&\leqslant \sup_{\pi_1 \in S\Pi} \int r(s,a) + \gamma \int V_{k-1}^* \, \mathrm{d}P(s' \mid s,a) \, \mathrm{d}\pi_1(a \mid s) \\
&= \sup_{\pi_1 \in S\Pi} T_{\pi_1} V_{k-1}^*(s) = T V_{k-1}^*(s)
\end{aligned}$$

Since s was arbitrary we must have $V_k^* \leqslant T V_{k-1}^*$. On the other hand by proposition 2.34 and induction hypothesis we have

$$T V_{k-1}^*(s) = T_{\tau_{V_{k-1}^*}} V_{k-1}^*(s) = T_{\tau_{V_{k-1}^*}} \dots T_{\tau_{V_0^*}} V_0 = V_{k,(\tau_{V_{k-1}^*},\dots,\tau_{V_0^*})}$$

But since $(\tau_{V_{k-1}^*}, \dots, \tau_{V_0^*})$ occur in the supremum we must then also have $T V_{k-1}^* \leqslant V_k^*$. Note that upper semicontinuity of V_k^* follows since T_a preverses this property (see proposition 2.46). □

Proposition 2.48. Under setting 2 and the last point in setting 1, that is the condition

$$\forall s \in \mathcal{S} : \sup_{N \geqslant n} \sup_{\pi \in R\Pi} \sum_{i=n+1}^{N} \mathbb{E}_s^\pi \underline{R}_i \to 0, \quad n \to \infty$$

it holds that $V^* = \lim_{k\to\infty} T^k V_0^*$. Furthermore under (D) the greedy policy τ_{V*} exists and is a deterministic stationary optimal policy.

Proof. Since setting 2 and the last point in setting 1 implies the rest of setting 1 and condition (S) we have by theorem 2.25 that $T^k V_0^* = V_k^* \to V^*$. We know by proposition 2.47 that V_k^* is semi uppercontinuous for all $k \in \mathbb{N}$. Under (D) we have that

$$\widehat{V}_k := V_k^* - V_{\max}(1 - \gamma^k) \downarrow V^* - V_{\max}$$

So by proposition 2.42 the infimum $\inf_k \widehat{V}_k = V^* - V_{\max}$ is upper semicontinuous and thus V^* is upper semicontinuous. Therefore by proposition 2.46 there exists a deterministic greedy policy τ_{V*} which satisfies

$$T_{\tau_{V*}} V^* = T V^* \tag{2.3}$$

By proposition 2.34 (under (D)) T and $T_{\tau_{V*}}$ is contractive on the Banach space $B = \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A}) \ni V_0^*$. Therefore by Banach fixed point theorem (see theorem A.6) $V^* = \lim_{k\to\infty} T^k V_0^*$ is the unique

fixed point of T in B. Again by Banach fixed point theorem and eq. (2.3) V^* is the fixed point of $T_{\tau_{V^*}}$, which by proposition 2.34 also has $V_{\tau_{V^*}}$ as fixed point. By uniqueness $V_{\tau_{V^*}} = V^*$ and thus τ_{V^*} is optimal. □

Remark 2.49. The property that $TV^* = V^*$ is often referred to as *Bellman's optimality equation.*

Corollary 2.50. Under setting 2 and (D) we have that τ_{V^*} is optimal and for any $V \in \mathcal{L}(\mathcal{S} \times \mathcal{A})$ it holds that

$$\left|T^k V - V^*\right| \leqslant \gamma^k \left|V - V^*\right|$$

Furthermore $V_{k,\pi}, V_\pi, V_k^*, V^*$ are all upper semicontinuous.

Proof. Since (D) implies the last point in setting 1 we can apply proposition 2.48. The bound on $\left|T^k V - V^*\right|$ is by the Banach fixed point theorem. Upper semicontinuity of the value functions followed from the proofs of proposition 2.46 and proposition 2.48. □

Comparison to results of [1, Bertsekas and Shreve (2007)]

In [1] (prop. 8.6 and cor. 9.7.2) results very similar to proposition 2.47 and proposition 2.48 are also established with in a slightly different setup. Besides having a state and action space, [1] also considers a non-empty Borel space called the *disturbance space* W, a *disturbance kernel* $p : \mathcal{S} \times \mathcal{A} \to W$, instead of a transition kernel which on the other hand is a deterministic *system function* $f : \mathcal{S} \times \mathcal{A} \times W \to \mathcal{S}$ which should be Borel measurable. Also the rewards are interpreted as negative costs, and thus g is required to be semi *lower*continuous. In [1] are also found much theory assuming semi*analytic* functions instead of semicontinuous ones.

It is possible to recover setting 2 from the semicontinuous setting in [1] by the following procedure. Set $P(\cdot \mid s, a) = f(s, a, p(\cdot \mid s, a))$ and maximize rewards of upper semicontinuous instead of minimizing lower semicontinuous ones.

2.2.3 Value iteration

Value iteration is a broad notion that can refer to many algorithms in dynamic programming, that somehow updates value functions. We here present perhaps to most basic of such algorithms, which is simply an iterative application of the T operator.

Algorithm 1: Simple theoretical value iteration

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations K, initial value function $\widetilde{V}_0$

$r \leftarrow \int x \, \mathrm{d}R(x \mid \cdot)$.

for $k = 0, 1, 2, \ldots, K-1$ **do**

  $\widetilde{V}_{k+1} \leftarrow \sup_{a \in \mathcal{A}} r(\cdot, a) + \gamma \int \widetilde{V}_k(s') \, \mathrm{d}P(s' \mid \cdot, a)$

Output: An estimator $\widetilde{V}_K$ of V^*

The results of the previous section, in particular corollary 2.50 is the theoretical foundation for *value iteration.* In particular by corollary 2.50 we know under setting 2 and (D) that algorithm 1 convergences exponentially in γ to the optimal value function. Furthermore by proposition 2.47 if we set $\widetilde{V}_0 = 0$ and define $\widetilde{\tau}_1, \ldots, \widetilde{\tau}_k$ as the greedy policies with respect to the outputs $\widetilde{V}_1, \ldots, \widetilde{V}_k$ of algorithm 1. Then $\widetilde{\pi} = (\widetilde{\tau}_k, \ldots, \widetilde{\tau}_1)$ is k-optimal.

Value iteration was invented for finite state and action spaces, but as we have shown, exponential convergence to the optimal infinite horizon value function is guaranteed in far more general case

(setting 2 and (D)), and therefore algorithm 1 could be applied in other cases if one has a practical way of representing the iterations $TV_0, T^2V_0, \ldots$.

Value iteration is a widely used as an example of simple reinforcement learning. We will now look at a classic example from a 2015 course in RL by David Silver where the environment is a finite MDP.

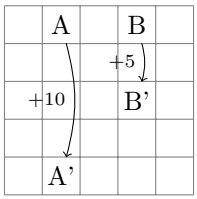


Figure 2.1: The simple gridworld Markov decision process.

Example 2.51 (Gridworld)**.** The *gridworld* MDP consist of 25 states $\mathcal{S} = [5]^2$ and 4 actions $\mathcal{A} = \{U, D, L, R\}$ for *up*, *down*, *left* and *right*. The transition P and reward R kernels are deterministic: the agent moves 1 square up, down, left or right according to the chosen action and receives a reward of 0, except:

- Any move that would move the agent out of the grid results in no movement and a reward of -1.
- Any action in the state $A = (2, 1)$ results in $A' = (2, 5)$ as the next state and a reward of 10.
- Any action in the state $B = (4, 1)$ results in $B' = (4, 3)$ as the next state and a reward of 5.

Finite spaces are trivially standard Borel with the discrete topology, which also makes every map $(s, a) \mapsto \mathcal{X}$ into some topological space continuous. In particular P is continuous and r is (upper semi)continuous. The set of admissible actions $A(s)$ is equal to the full action space $\mathcal{A}$ for all $s \in \mathcal{S}$, which is trivially compact. The rewards are bounded by $R_{\max} = 10$. Setting $\gamma = 0.9$ makes (D) hold, and puts $V_{\max} = 10/(1 - 0.9) = 100$. Thus we can apply corollary 2.50 and for $\widetilde{V}_0 = 0$ get that

$$\left|\widetilde{V}_K - V^*\right| \leqslant \gamma^K \left|V^*\right| \leqslant \gamma^K V_{\max} = 100 \cdot 0.9^{-K}$$

By proposition 2.34 for any stationary policy $\tau \in S\Pi$ we have that T_τ is also γ-contractive and we easily get the same bound on the policy evaluation

$$\left|T_\tau^k V_0 - V_\tau\right| \leqslant \gamma^K V_{\max}$$

Measures over finite spaces can be represented as real vectors, and kernels as matrices. For example the transition kernel $P : \mathcal{P}(\mathcal{S} \times \mathcal{A}) \to \mathcal{P}(\mathcal{S})$ is a $|\mathcal{S}| \times (|\mathcal{S}| \cdot |\mathcal{A}|)$-matrix. Calculating TV and $T_\pi V$ is then a matter of matrix multiplication and taking maxima over vectors.

Define the stationary policy $\tau_r(\cdot \mid \cdot) = \frac{1}{4}$ which chooses actions uniformly at random at every state. Below is shown some value functions of the gridworld environment (correct up to errors due to machine precision).

2.2.4 Q-functions

Throughout this section we assume setting 2, (D) and furthermore that $A(s)$ is finite for every $s \in \mathcal{S}$.

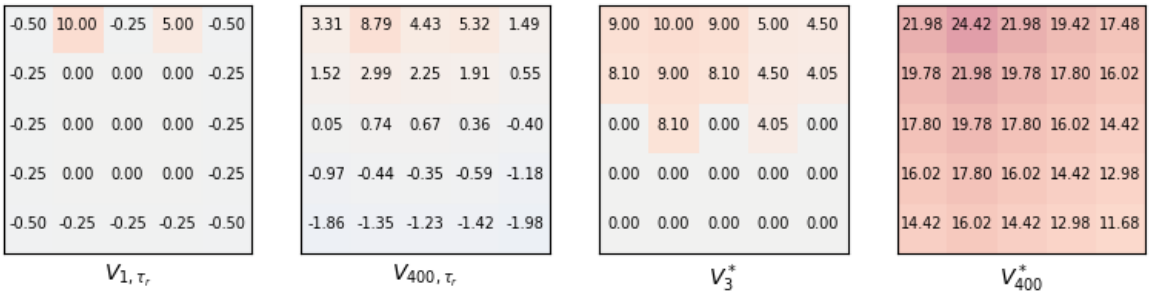


Figure 2.2: Value functions of the gridworld environment. Note that $V_{\max} \cdot \gamma^{400} = 100 \cdot (0.9)^{400} \approx 4.97 \cdot 10^{-17}$ so V_{400}^* and $V_{\tau_r,400}$ are very close to the true infinite horizon value functions V^* and V_{τ_r} (providing numerical errors are insignificant).

A **Q-function** is any function that assigns a (extended) real number to every state-action pair, that is any function $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$. Q-function are also called *action-value* functions, to distinguish them from the *value* functions we have discussed in the previous sections. Because of the similar role Q-functions play compared to value function, many concepts such as T-operators and the finite, infinite horizon policy evaluations and greedy policies, can be defined analogously.

Definition 2.52 (Policy evaluation for Q-functions)**.** Let $\pi \in R\Pi$. Define

$$Q_{k,\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V_{k,\pi}, \qquad Q_\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V_\pi$$

$$Q_k^* = \sup_{\pi \in R\Pi} Q_{k,\pi}, \qquad Q^* = \sup_{\pi \in R\Pi} Q_\pi$$

Define $Q_0 = r$ then we make the convention that $Q_0^* = Q_{0,\pi} = Q_0 = r$.

Definition 2.53 (Operators for Q-functions)**.** For any stationary policy $\tau \in S\Pi$ and measurable $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ with $Q \geqslant 0, Q \leqslant 0$ or $|Q| < \infty$ we define

$$P_\tau Q(s,a) = \int Q(s',a') \, \mathrm{d}\tau P(s',a' \mid s,a)$$

$$T_\tau Q = r + \gamma P_\tau Q$$

$$TQ(s,a) = r(s,a) + \gamma \int \max_{a' \in \mathcal{A}} Q(s',a') \, \mathrm{d}P(s' \mid s,a)$$

where $T_a = T_{\delta_a}$.

Remark 2.54. The P_τ operator is defined for simplications in proofs, especially in the analysis of [5, Fan et al. (2020+)] in the later sections.

Definition 2.55 (Greedy policies for Q-functions)**.** Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A}$ be a stationary policy. Define $G_Q(s) = \operatorname{argmax}_{a \in A(s)} Q(s,a)$. If there exist a measurable set $G_Q^\tau(s) \subseteq G_Q(s)$ for every $s \in \mathcal{S}$ such that

$$\tau\left(G_Q^\tau(s) \,\middle|\, s\right) = 1$$

then τ is said to be **greedy** with respect to Q and is denoted τ_Q.

The idea of Q-functions (and the letter Q) originates to [18, Watkins (1989)]. Upon the definition he notes

> "This is much simpler to calculate than $[V_\pi]$ for to calculate $[Q_\pi]$ it is only necessary to look one step ahead [...]"

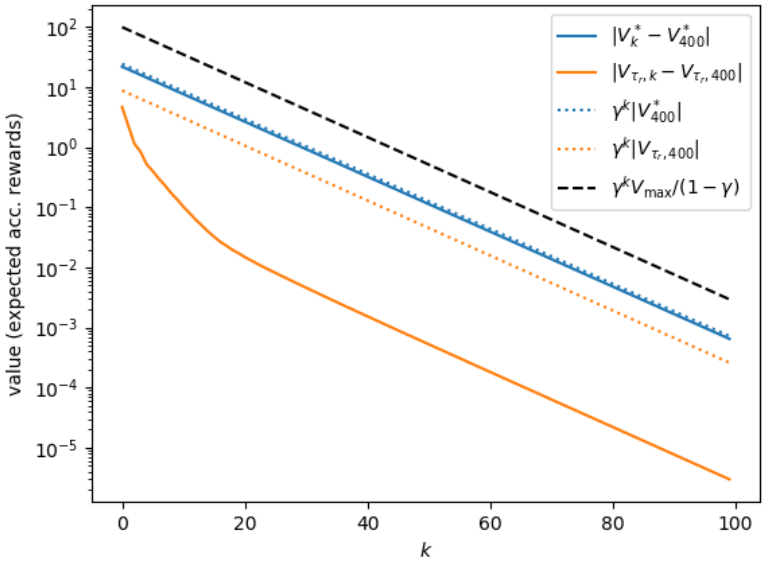


Figure 2.3: Convergence of gridworld value functions compared with the theoretical bounds. The black dashed line is the general theoretical bound for both T and T_τ by Banachs fixed point theorem and the maximum value $V_{\max} = R_{\max}/(1-\gamma)$. The dotted blue and orange uses $\left|V_k^*\right|$ and $\left|V_{\tau,k}\right|$ respectively, which might not be available. ($\gamma = 0.9$).

A clear advantage of working with Q-function $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ rather than a value function $V : \mathcal{S} \to \overline{\mathbb{R}}$, is that finding the optimal action in state s requires only a maximization over the Q-function itself: $a = \operatorname{argmax}_{a \in A(s)} Q(s,a)$. This should be compared to finding a best action according to a value function V: $a = \operatorname{argmax}_{a \in A(s)} r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V$. Besides being less simple, this requires taking an expectation with respect to both the reward and transition kernel. Later we will study settings where we are not allowed to know the process kernels when attempting to find the optimal strategy. In these situations the advantage of Q-functions is clear. For now however the transition kernel will remain known and we will in this section see how the results of state-value functions translate to Q-functions.

Proposition 2.56 (Relations between Q- and value functions)**.** Let $\pi = (\tau_1, \tau_2, \dots) \in M\Pi$ be a Markov policy and $\tau \in S\Pi$ stationary. Then

1. Policy evaluations are related by $\mathbb{E}_{\tau(\cdot|s)} Q_{k,\pi} = V_{k+1,(\tau,\pi)}(s)$.
2. T_τ-operators are related by $T_\tau Q_{k,\pi}(s,a) = r + \gamma \mathbb{E}_{P(\cdot|s,a)} T_\tau V_{k,\pi}$.
3. Greedy policies for policy evaluations are the same, that is
$$\tau(Q_{k,\pi}) = \tau(V_{k,\pi}), \text{ and } \tau(Q_\pi) = \tau(V_\pi)$$
4. Optimal policies are related by $\max_{a \in A(s)} Q^*(s,a) = V^*(s)$ and
$$Q_k^*(s,a) = r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V_k^*, \quad Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V^*$$

Proposition 2.57 (Properties of Q-functions)**.** Let $\pi = (\tau_1, \tau_2, \dots) \in M\Pi$ be a Markov policy and $\tau \in S\Pi$ stationary. Then

1. $Q_{k,\pi} = T_{\tau_1} \dots T_{\tau_k} Q_0$ and $Q_k^* = T_{\tau_{k-1}^*} \dots T_{\tau_0^*} Q_0^* = T^k Q_0^*$.
2. $Q_\pi = \lim_{k\to\infty} Q_{k,\pi}$ and $Q^* = \lim_{k\to\infty} Q_k^*$.
3. T, T_τ are γ-contractive on $\mathcal{L}_\infty(\mathcal{S}\times\mathcal{A})$ and Q^*, Q_τ are their unique fixed points.
4. $Q^* = Q_{\tau^*}$ and $Q_{k,\pi}$, Q_π, Q_k^*, Q^* are all upper semicontinuous and bounded by $V_{\max}$.

Proof of proposition 2.56 and proposition 2.57. Measurability of $Q_{k,\pi}$ and Q_π follow from measurability of $V_{k,\pi}$, V_π and proposition 2.3. Upper semicontinuity of $Q_{k,\pi}$ and Q_π follows from proposition 2.43 because $V_{k,\pi}$ and V_π are upper semicontinuous (see corollary 2.50).

For proposition 2.56.1 we have

$$\begin{aligned}\mathbb{E}_{\tau(\cdot|s)} Q_{k,\pi} &= \int r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V_{k,\pi} \,\mathrm{d}\tau(a \mid s) \\ &= \int r(s,a) + \gamma \sum_{i=1}^{k} \gamma^{i-1} r(s_i,a_i) \,\mathrm{d}P\tau_k \dots P\tau_1 P\tau(a, s_1, a_1, \dots, s_k \mid s) \\ &= V_{k+1,(\tau,\pi)}\end{aligned}$$

For proposition 2.56.2 we sketch the idea by

$$T_\tau Q_{k,\pi} = r + \gamma \int r + \gamma V_{k,\pi} \,\mathrm{d}P \,\mathrm{d}\tau P = r + \gamma \int r + \gamma V_{k,\pi} \,\mathrm{d}P\tau \,\mathrm{d}P = r + \gamma \int T_\tau V_{k,\pi} \,\mathrm{d}P$$

For $Q_{k,\pi} = T_{\tau_1} \dots T_{\tau_k} Q_0$ use proposition 2.56.2 iteratively starting with $\tau = \tau_1, \pi = (\tau_2, \tau_3, \dots)$.

The $\tau(Q_{k,\pi}) = \tau(V_{k,\pi})$ part of proposition 2.56.3 is by definition of the two concepts of greedy functions.

That $Q_\pi = \lim_{k\to\infty} Q_{k,\pi}$ follows from dominated convergence and proposition 2.19.3.

For proposition 2.56.4 $Q_k^* = \sup_{\pi\in R\Pi}(r + \gamma\mathbb{E}V_{k,\pi}) \leqslant r + \gamma\mathbb{E}V_k^* = r + \gamma\mathbb{E}V_{\pi_k^*} \leqslant Q_k^*$. The same argument works for the second part.

Let $s \in \mathcal{S}$ then $\sup_{a\in A(s)} Q^*(s,a) = \sup_{a\in A(s)}(r(s,a) + \gamma\mathbb{E}_{P(\cdot|s,a)}V^*) = TV^*(s) = V^*(s)$.

By the definition of Q_{τ^*} we have $Q^* = r + \gamma\mathbb{E}V^* = r + \gamma\mathbb{E}V_{\pi^*} = Q_{\tau^*}$.

$T_\tau Q_\tau = T_\tau(r + \gamma\mathbb{E}\lim_{k\to\infty} T_\tau^k V_0) = \lim_{k\to\infty} T_\tau(r + \gamma\mathbb{E}T_\tau^k V_0) = \lim_{k\to\infty}(r + \gamma\mathbb{E}T_\tau^{k+1}V_0) = r + \gamma\mathbb{E}\lim_{k\to\infty} T_\tau^{k+1}V_0 = r + \gamma\mathbb{E}V_\tau = Q_\tau$.

We $T_{\tau_Q} Q = TQ$ for any measurable Q because

$$T_{\tau_Q}(s,a) = r(s,a) + \gamma \int \max_{a'\in A(s')} Q(s',a') \,\mathrm{d}P(s' \mid s,a) = TQ(s,a)$$

Therefore by proposition 2.57.1

$$T_{\tau_{k-1}^*} Q_{k-1,(\tau_{k-2}^*,\dots,\tau_0^*)} = TQ_{k-1}^*$$

since by proposition 2.56.3 τ_{k-1}^* is greedy for Q_{k-1}^*. Now use induction to get $Q_{k-1}^* = T^k Q_0^*$.

Because $V^* = V_{\tau^*}$ we have

$$TQ^* = T_{\tau^*} = r + \gamma\mathbb{E}T_{\tau^*}V_{\tau^*} = r + \gamma\mathbb{E}V^* = Q^*$$

The contrativeness of T and T_π follows from the same argument as for value functions. Banach fixed point theorem now concludes proposition 2.57.3.

Since now Q^* and Q_{τ^*} are fixed points for T they must be equal, concluding the last point, namely proposition 2.57.4. □

Corollary 2.58. For any $Q \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ $T^k Q$ converges to Q^* with rate γ^k. That is

$$\left|T^k Q - Q^*\right| \leqslant \gamma^k |Q - Q^*|$$

Proof. By proposition 2.57.3. □

2.2.5 Q-iteration

Similar to the value iteration algorithm (algorithm 1) we can define the corresponding for Q-iteration.

Algorithm 2: Simple theoretical Q-iteration

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations K

$r \leftarrow \int x \, \mathrm{d}R(x \mid \cdot)$

$\widetilde{Q}_0 \leftarrow r$

for $k = 0, 1, 2, \ldots, K-1$ **do**

$\quad \widetilde{Q}_{k+1} \leftarrow r + \gamma \int \sup_{a' \in \mathcal{A}} \widetilde{Q}_k(s', a') \, \mathrm{d}P(s' \mid \cdot)$

Output: An estimator $\widetilde{Q}_K$ of Q^*

Proposition 2.59. (D)

The output $\widetilde{Q}_K$ of algorithm 2 converges to the optimal Q-function Q^* with rate γ^K concretely $\left\|\widetilde{Q}_K - Q^*\right\|_\infty \leqslant \gamma^K \|Q^*\|_\infty$.

Proof. This is by corollary 2.58. □

Finite Q-iteration

We have shown how if one knows the dynamics of a stationary decision process satisfying rather broad criteria, such as continuity and compactness, the optimal policy and state-value function can be found simply by iteration over the T-operator and picking a greedy strategy (see proposition 2.59). Of course this is practical computationally, only if the resulting Q functions can be represented and computed in finite space and time. An obvious situation in which such a representation and computation is possible, is the finite case. Say $|\mathcal{S}| = k$ and $|\mathcal{A}| = \ell$. In this case the transition operator P can be represented as a matrix of *transition probabilities*

$$P := \begin{pmatrix} P(s_1 \mid s_1, a_1) & \ldots & P(s_k \mid s_1, a_1) \\ \vdots & \vdots & \vdots \\ P(s_1 \mid s_k, a_\ell) & \ldots & P(s_k \mid s_k, a_\ell) \end{pmatrix}$$

then the algorithm becomes

Algorithm 3: Simple finite Q-iteration

Input: MPD $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations K

Set $r \leftarrow \left(\int r \, \mathrm{d}R(r \mid s_1, a_1), \ldots, \int r \, \mathrm{d}R(r \mid s_k, a_\ell)\right)^T$ and $\widetilde{Q}_0 \leftarrow r$.

for $k = 0, 1, 2, \ldots, K-1$ **do**

Set $m(\widetilde{Q}_k) \leftarrow (\max_{a \in \mathcal{A}} Q(s_1, a), \ldots, \max_{a \in \mathcal{A}} Q(s_k, a))^T$

Update action-value function:

$$\widetilde{Q}_{k+1} \leftarrow r + \gamma P m(\widetilde{Q}_k)$$

Output: An estimator $\widetilde{Q}_K$ of Q^*

Proposition 2.60. The output $\widetilde{Q}_K$ from algorithm 3 is K-optimal and $\left\|\widetilde{Q}_K - Q^*\right\|_\infty \leqslant \gamma^K \|Q^*\|_\infty$.

Proof. See corollary 2.58. □

2.3 Model-based Q-learning

In this section we will look at what happens if we instead use approximations of the Q-functions and T operator. This means that we are in a setting where we can somehow calculate r and TQ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, but it is hard or infeasible to represent them (or at least one of them) directly. The purpose of this is to show how results about the convergence of Q-learning is rather easily obtained if one has direct access to the transition and reward kernels P and R.

This setting is not very well-studied in the case of a continuous state space (at least in the sources known to this writer). This is perhaps because it is considered solved by the results of theoretical Q-learning presented in the previous section. However as we have argued, this only have practical relevance when it is feasible to represent TQ. Therefore we find it relevant to consider this setting in more detail. What *is* very well-studied is a further generalized setting where T and r are assumed to be unknown, that is, one has only access to their distributions via sampling from them. Solutions for that setting are called *model-free*. We will deal with that setting in the next section.

2.3.1 Algorithmic and approximation errors

In the following we present some rather simple bounding techniques which is inspired by arguments found in e.g. [5], together with some standard results from approximation theory on artificial neural networks and Bernstein polynomials. Throughout this section we assume (D) i.e. that we are discounting with some $\gamma \in [0, 1)$.

Let us consider any norm $\|\cdot\|$ on the set of Q-functions $\mathcal{Q} = \{f : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}\}$. Let $\mathcal{F} \subseteq \mathcal{Q}$ be some subclass of Q-functions Let $\widetilde{Q}_0 \in \mathcal{F}$ be bounded in $\|\cdot\|$. Suppose we can approximate $T\widetilde{Q}_0$ by some $\widetilde{Q}_1 \in \mathcal{F}$ to $\varepsilon_1 > 0$ precision and then approximate $T\widetilde{Q}_1$ by $\widetilde{Q}_2 \in \mathcal{F}$ and so on. This way we get a sequence of Q-functions satisfying

$$\left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\| \leqslant \varepsilon_k, \forall k \in \mathbb{N}$$

First observe that

$$\begin{aligned}\left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| &\leqslant \left\|T^k\widetilde{Q}_0 - T\widetilde{Q}_{k-1}\right\| + \left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\| \\ &\leqslant \gamma\left\|T^{k-1}\widetilde{Q}_0 - \widetilde{Q}_{k-1}\right\| + \left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\|\end{aligned}$$

Using this iteratively we get

$$\left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \leqslant \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i$$

Thus we get

Theorem 2.61. Let $\|\cdot\|$ be a norm on the space of functions $\mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$. Let $\widetilde{Q}_k$ be obtained from a function class $\mathcal{F}$ such that $\left\|\widetilde{Q}_k - T\widetilde{Q}_{k-1}\right\| \leqslant \varepsilon_k$ for any $k \in \mathbb{N}$. Then

$$\left\|Q^* - \widetilde{Q}_k\right\| \leqslant \gamma^k\left\|Q^* - \widetilde{Q}_0\right\| + \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i$$

Proof. By the discussion above and

$$\begin{aligned}\left\|Q^* - \widetilde{Q}_k\right\| &\leqslant \left\|Q^* - T^k\widetilde{Q}_0\right\| + \left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \\ &\leqslant \gamma^k\left\|Q^* - \widetilde{Q}_0\right\| + \varepsilon_{\text{approx}}(k)\end{aligned}$$

□

The first term in theorem 2.61 is sometimes called the *algorithmic* error[1] while

$$\varepsilon_{\text{approx}} := \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i$$

is called the *approximation* error.

The algorithmic error converges exponentially, so one is often happy with this part not spending time trying to bound this tighter. The approximation error depends on our step-wise approximations. For example if $\varepsilon_i(k) = \varepsilon$ for some $\varepsilon > 0$ we easily get the bound

$$\varepsilon_{\text{approx}}(k) = \varepsilon\frac{1-\gamma^k}{1-\gamma} \leqslant \frac{\varepsilon}{1-\gamma} \tag{2.4}$$

If $\varepsilon_i \leqslant c\gamma^i$ we get $\varepsilon_{\text{approx}}(k) \leqslant ck\gamma^k \to 0$ as $k \to \infty$. Generally if one can show that $\varepsilon_i \to 0$ we have

Proposition 2.62. $\sum_{i-1}^{k} \gamma^{k-i}\varepsilon_i \to 0$ whenever $\varepsilon_k \to 0$ as $k \to \infty$.

Proof. Let $\varepsilon > 0$. Find N such that $\varepsilon_n \leqslant \varepsilon(1-\gamma)/2$ for all $n > N$ and find $M > N$ such that $\gamma^M \leqslant \varepsilon\gamma^N\left(\sum_{i=1}^{N} \gamma^{N-i}\varepsilon_i\right)^{-1}$. Then for all $m > M$

$$\sum_{i=1}^{m} \gamma^{m-i}\varepsilon_i \leqslant \gamma^{m-N}\sum_{i=1}^{N} \gamma^{N-i}\varepsilon_i + \sum_{i=N+1}^{m} \gamma^{m-i}\varepsilon(1-\gamma)/2 \leqslant \varepsilon/2 + \varepsilon/2 \leqslant \varepsilon$$

□

We will now explore two different ways of obtaining bounds on the approximation error.

[1] For example in [5].

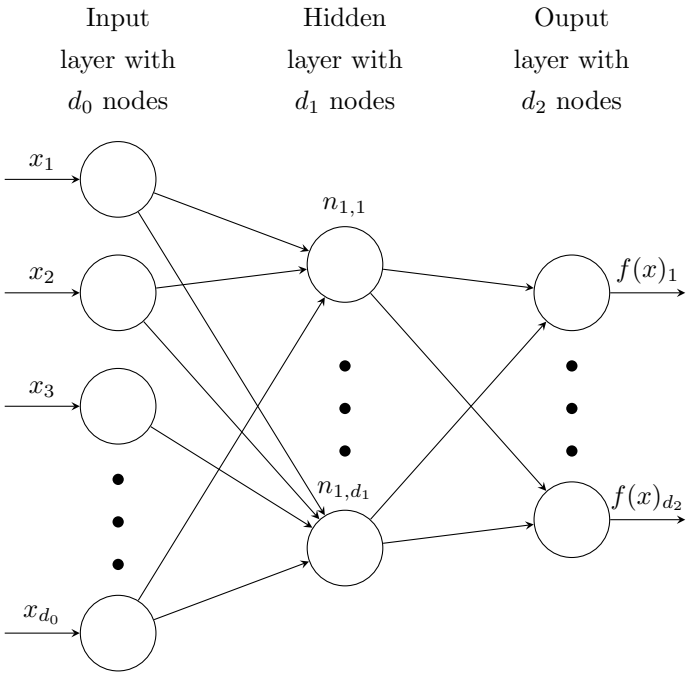


Figure 2.4: An ANN with one hidden layer. Notice that there is no edge from $n_{0,3}$ to $n_{1,1}$ which means that $W_1(1,3) = 0$.

2.3.2 Using artifical neural networks

Setting 3 (Continuous MDP)**.** An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0,1]^w$ and $\mathcal{A}$ finite. Assume that r is continuous.

Definition 2.63. An **artificial neural network** (ANN) with $L \in \mathbb{N}_0$ hidden layers, structure $(d_i)_{i=0}^{L+1} \subseteq \mathbb{N}$, activation functions $\sigma_i = (\sigma_{ij})_{j=1}^{d_i}$, where $\sigma_{ij} : \mathbb{R} \to \mathbb{R}$ are real-valued functions on $\mathbb{R}$, and weights $W_i \in M^{d_i \times d_{i-1}},\ v_i \in \mathbb{R}^{d_i},\ i \in [L+1]$ is the function $F : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$

$$F = w_{L+1} \circ \sigma_L \circ w_L \circ \sigma_{L-1} \circ \cdots \circ w_1$$

where w_i is the affine function $x \mapsto W_i x + v_i$, and $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ is coordinate-wise application of the components σ_{ij} for $i \in [L+1]$.

We denote the class of these networks (or functions)

$$\mathcal{DN}\left((\sigma_i, d_i)_{i=0}^{L+1}\right)$$

Remark 2.64. Artificial neural networks are often illustrated as $L+1$-partite graphs with d_i nodes in the ith partition. A node $n_{i,j}$ in partition i is then connected to a node $n_{i+1,k}$ if $W_{i+1}(k,j) \neq 0$. Artifical neural networks were inspired by the structure of *neurons* in nerve tissue (e.g. the brain) of living organisms, with the graph nodes corresponding to neurons and edges to *axons*.

From the graph view-point it easy to see that one may join neural nets to form larger ones, by either function composition or side-by-side alignment. That this if $f \in \mathcal{DN}\left((\sigma_i, d_i)_{i=1}^{L_f+1}\right)$ and ...

An ANN is called *deep* if there are two or more hidden layers.

Theorem 2.65 (Universal Approximation Theorem for ANNs)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be non-constant, bounded and continuous activation function. Let $\varepsilon > 0$ and $f \in C([0,1]^w)$. Then there exists an

$N \in \mathbb{N}$ and a network $F \in \mathcal{DN}(\sigma, (w, N, 1))$ with one hidden layer and activation function σ such that

$$\|F - f\|_\infty < \varepsilon$$

In other words $\bigcup_{N\in\mathbb{N}} \mathcal{DN}(\sigma, (w, N, 1))$ is dense in $C([0, 1]^w)$.

Discussion of proofs. The original proof in [3, Cybenko (1989)] is very short and elegant, but non-constructive, using the Riesz Representation and Hahn-Banach theorems to obtain a contracticton to the statement that $\bigcup_{N\in\mathbb{N}} \mathcal{DN}(\sigma, (w, N, 1))$ is dense in $C([0, 1]^w)$. Furthermore it considered only *sigmoidal* activations functions, meaning that σ should satisfy

$$\sigma(x) \to \begin{cases} 0 & x \to -\infty \\ 1 & x \to \infty \end{cases}$$

This was extended in [2, Chen et al. (1990)] to the statement as presented above and their proof is constructive. □

Proposition 2.66. Under setting 3 let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-constant, bounded, continuous activation function. Let $\varepsilon > 0$. Assume that either

1. P is deterministic with $P(\cdot \mid s, a) = \delta_{p(s,a)}$. For some continuous $p : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$.
2. $P(\cdot \mid s, a)$ is absolutely continuous with respect to the same measure ν on $\mathcal{S}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ with density $p(\cdot \mid s, a)$ which is continuous in s.

Then for every $k \in \mathbb{N}$ there exists a $N \in \mathbb{N}$ and a sequence of Q-networks $(\widetilde{Q}_i)_{i=1}^k \subseteq \mathcal{DN}(\sigma, \{w|\mathcal{A}|, N, 1\})$ such that

$$\varepsilon_{\text{approx}}(i) = \left\| T\widetilde{Q}_{i-1} - \widetilde{Q}_i \right\|_\infty < \varepsilon$$

for all $i \in [k]$. In particular

$$\left| Q^* - \widetilde{Q}_k \right| < \varepsilon/(1 - \gamma)$$

Proof. The key points are that

a. Any ANN with continuous activation functions is continuous.

b. Under assumptions 1. or 2. the Bellman operator T preserves continuity.

Using these facts we can use the universal approximation theorem (theorem 2.65) to get a series of networks

$$f_{a,k} : \mathcal{S} \to \mathbb{R}$$

for each $a \in \mathcal{A}$ and $k \in \mathbb{N}$ satisfying

$$\left| f_{a,k} - T\widetilde{Q}_{k-1}(\cdot, a) \right| < \varepsilon \tag{2.5}$$

Here $\widetilde{Q}_0 = 0$ and $\widetilde{Q}_k$ is obtained recursively by combining $f_{a,k}$ into a single network $\widetilde{Q}_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ which by eq. (2.5) on each of its components approximates $T\widetilde{Q}_{k-1}$ to ε precision. $\widetilde{Q}_0$ is obviously continuous and because of a. and b. $T\widetilde{Q}_k$ is as well. □

This gives us the first method of how to approximate Q^* arbitrarily closely on continuous state spaces, in the case where it is infeasible to represent TQ directly. To use this method in practice one would need to go through the construction in [2]. We will not go further into this, and instead focus on another approximation method using *Bernstein polynomials*.

2.3.3 Using Bernstein polynomials

In this case the need a stronger form of continuity, namely Lipschitz continuity, to establish the bounds.

Setting 4 (Bernstein approximable MDP)**.** An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0,1]^w$ and $\mathcal{A}$ finite. Assume that there exists a probability measure $\mu \in \mathcal{S}$, such that $P(\cdot \mid s,a)$ has density $p(\cdot \mid s,a) : \mathcal{S} \to \mathbb{R}$ with respect to μ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Furthermore assume that $r(\cdot, a),\ p(s \mid \cdot, a)$ are Lipschitz with constants $L_r,\ L_p$ respectively for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Definition 2.67 (Bernstein polynomial)**.** The multivariate Bernstein polynomial $B_{f,n}$ with exponents $n = (n_1, \dots, n_w) \in \mathbb{N}^w$ approximating the function $f : [0,1]^w \to \mathbb{R}$ is defined by

$$B_{f,n}(x_1, \dots, x_w) = \sum_{j=1}^{w} \sum_{k_j=0}^{n_j} f\left(\frac{k_1}{n_1}, \dots, \frac{k_w}{n_w}\right) \prod_{\ell=1}^{w} \left(\binom{n_\ell}{k_\ell} x_\ell^{k_\ell} (1-x_\ell)^{n_\ell - k_\ell} \right)$$

Notice that this a polynomial of (multivariate) degree $n_1 + \cdots + n_w$.

Theorem 2.68. Let $f : [0,1]^w \to \mathbb{R}$ be Lipschitz (see definition A.2) w.r.t. the standard euclidean 2-norm induced metrics on $[0,1]^w$ and $\mathbb{R}$ with constant L. Then for any $n = (n_1, \dots, n_w) \in \mathbb{N}^w$ there exists a polynomial $B_{f,n} : [0,1]^w \to \mathbb{R}$ of degree $\leqslant \|n\|_1$ such that

1. $\left\|f - B_{f,n}\right\|_2 \leqslant \frac{L}{2}\sqrt{\sum_{j=1}^{w} \frac{1}{n_j}}$
2. $\left\|B_{f,n}\right\|_\infty \leqslant \|f\|_\infty$

Lemma 2.69. Under setting 4 $TQ(\cdot, a)$ is Lipschitz in $\|\cdot\|_2$ with constant $L_T = (L_r + \gamma V_{\max} L_p)$ for all $a \in \mathcal{A}$ and $Q : \mathcal{S} \times \mathcal{A} \to [-V_{\max}, V_{\max}]$.

Now we can bound

Proposition 2.70.

$$\varepsilon_{\text{approx}} \leqslant \frac{L_r + \gamma V_{\max} L_p}{2(1-\gamma)} \sqrt{\sum_{j=1}^{w} \frac{1}{n_j}}$$

For example if we put $n_j = m$ for all j we get

Proposition 2.71.

$$\left\|Q^* - \widetilde{Q}_k\right\| \leqslant \gamma^{-k} \left\|Q^* - \widetilde{Q}_0\right\| + \frac{L_r + \gamma V_{\max} L_p}{2(1-\gamma)} \sqrt{w} m^{-1/2}$$

In particular $\left\|Q^* - \widetilde{Q}_k\right\|_\infty = \mathcal{O}(\gamma^{-k} + \frac{1}{\sqrt{m}})$ when using k iterations and approximating with multivariate polynomials of maximum degree $w \cdot m$.

This gives a very concrete way of constructing an arbitrarily good approximation to Q^* using polynomials.

Chapter 3

Model-free algorithms

In this section we will look at what can be done when the process dynamics are unknown. In this case we cannot calculate directly neither r, $T_\pi Q$ nor TQ because the transition and reward kernels P, R are unknown.

It is clear that algorithm 2 will not work without modification in this case. Simply because R and P are not available. To make the scheme work anyway we could simply avoid taking expectations and use the random outcomes of the kernels. Leading to

Algorithm 4: Random theoretical Q-iteration (example of thought)

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations K

$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \widetilde{Q}_0(s,a) \leftarrow X \sim R(\cdot \mid s,a)$.

for $k = 0, 1, 2, \ldots, K-1$ **do**

$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \widetilde{Q}_{k+1}(s,a) \leftarrow r' + \gamma \sup_{a' \in \mathcal{A}} \widetilde{Q}_k(s', a')$

where $r' \sim R(\cdot \mid s,a), s' \sim P(\cdot \mid s,a)$.

Define π_K as the greedy policy w.r.t. $\widetilde{Q}_K$

Output: An estimator $\widetilde{Q}_K$ of Q^* and policy π_K

We immedially run into problems in the uncountable case, because drawing uncountably many times from a distribution is not easily defined in a sensible way. Even in the finite case where the functions $\widetilde{Q}_k$ are well defined, they cannot converge if R is not deterministic. Therefore this approach does not work in a continuous or stochastic setting.

There are broadly two ways of dealing with these problems[1]. In the *indirect* approaches one tries to first estimate P and R by sampling. We can then use the model-dependent methods of the last chapter using estimated kernels $\widetilde{P}$ and $\widetilde{R}$. We will not go further into indirect methods in this thesis. The *direct* approaches covers *the rest*, and it is mainly these were are going to look at throughout this paper. A popular indirect approach is called *temporal difference* (TD) learning and was initially invented for finite MDPs.

3.1 Finite case

Setting 5 (Finite MDP)**.** An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where $|\mathcal{S}| = \mathfrak{s} \in \mathbb{N}$ and $|\mathcal{A}| = \mathfrak{a} \in \mathbb{N}$ are finite.

[1] This classification is discussed in [9, Kearns and Singh (1999)].

TD learning is based on the following update scheme

$$\widetilde{Q}_{k+1}(s,a) \leftarrow (1-\alpha_k)\widetilde{Q}_k(s,a) + \alpha_k(r' + \gamma \cdot \max_{a' \in \mathcal{A}} \widetilde{Q}_k(s',a')) \tag{3.1}$$

Here r' and s' are the reward and next-state drawn from the reward and transition kernels, and $\alpha_k \in [0,1]$ is the so-called **learning rate** (of the kth step). The 'temporal difference' is also the name of term $\alpha_k(r' + \gamma \cdot \max_{a \in \mathcal{A}} \widetilde{Q}_k(s',a') - \widetilde{Q}_k(s,a))$ occuring from rearranging eq. (3.1). Usually the learning rate is fixed before running the algoritm (does not depend on the history) and is set to decay from 1 to 0 in some fashion as $k \to \infty$.

We will now look at a convergence result originally obtained in [19, Watkins and Dayan (1992)] of a TD algorithm using Q-functions. The result was extended slightly in [7, Jaakkola et al. (1994)] and is here presented more in the style of [7].

Algorithm 5: Finite asynchronos Q-learning

Input: Finite MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, i.e. $|\mathcal{S}||\mathcal{A}| < \infty$, number of iterations K, state-action pairs $(s_1, a_1, \ldots, s_K, a_K)$, learning rates $(\alpha'_1, \ldots, \alpha'_K)$, initial $\widetilde{Q}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

Put $\alpha_k(s,a) \leftarrow \begin{cases} \alpha'_k & (s,a) = (s_k, a_k) \\ 0 & (s,a) \neq (s_k, a_k) \end{cases}$.

for $k = 1, 2, \ldots, K$ **do**

Sample $r' \sim R(\cdot \mid s_k, a_k), s' \sim P(\cdot \mid s_k, a_k)$

Update action-value function:

$$\widetilde{Q}_k \leftarrow \widetilde{Q}_{k-1} + \alpha_k(r' + \max_{a' \in \mathcal{A}} \widetilde{Q}_{k-1}(s',a'))$$

Define π_K as the greedy policy w.r.t. $\widetilde{Q}_K$

Output: An estimator $\widetilde{Q}_K$ of Q^* and policy π_K

Note that only the value of the pair (s_k, a_k) are updated in each step of the algorithm (since $\alpha_k(s,a) = 0$ for all $(s,a) \neq (s_k, a_k)$).

Theorem 3.1 (Watkins, Dayan 1992)**.** Let $s_1, a_1, s_2, a_2, \cdots \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A} \times \ldots$ be random variables, and $\alpha_1, \alpha_2, \cdots \in [0,1]$. The output $\widetilde{Q}_K$ of algorithm 5 converges to Q^* provided

1. $\mathbb{P}\left(\sum_{i=1}^{\infty} \alpha_i(s,a) = \infty\right) = 1, \mathbb{P}\left(\sum_{i=1}^{\infty} \alpha_i^2(s,a) < \infty\right) = 1.$
2. $\mathrm{Var}(R(\cdot \mid s,a)) < \infty$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
3. If $\gamma = 1$ all policies lead to a reward-free terminal state almost surely.

In the original formulation the sums of learning rates were supposed to converge *uniformly*. However this is equivalent to this formulation because of the fact that $\mathbb{P}(\sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |f_n(s,a)| \to 0) = 1 \iff \mathbb{P}\left(|f_n(s,a)| \to 0\right) = 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ whenever $\mathcal{S}, \mathcal{A}$ is finite. Notice that the first condition implies that all state-action pairs occur infinitely often almost surely. Also notice that the second condition is automatically fulfilled under (D) since then $\mathrm{Var}(R(\cdot \mid s,a)) \leqslant \mathbb{E}(2R_{\max})^2 = 4R_{\max}^2$.

In a special case of the same setup, convergence rates where established by [17, Szepesvári (1997)].

Theorem 3.2 (Szepesvári)**.** Let $t \in \mathbb{N}$ and $s_1, a_1, s_2, \ldots, s_t, a_t$ be sampled i.i.d. from $p \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$. Set the learning rates such that $\alpha'_k = |\{i \in [k-1] \mid (s_i, a_i) = (s_k, a_k)\}|^{-1}$, i.e. the reciprocal of the frequency of (s_k, a_k) at step k. Let $\beta = \max_{x \in \mathcal{S} \times \mathcal{A}} p(x) / \min_{x \in \mathcal{S} \times \mathcal{A}} p(x)$. Then for some $B > 0$ the following holds asymptotically almost surely

$$\left| \widetilde{Q}_t - Q^* \right| \leqslant B \frac{1}{t^{\beta(1-\gamma)}} \tag{3.2}$$

and

$$\left| \widetilde{Q}_t - Q^* \right| \leqslant B \sqrt{\frac{\log \log t}{t}} \tag{3.3}$$

Here eq. (3.2) is tightest when $\gamma > 1 - \beta/2$ otherwise eq. (3.3) is tighter.

A paper [11, Majeed and Hutter (2018)] proves that Q-learning is PAC-learnable given some additional assumptions.

Algorithm 6: Finite synchronos Q-learning

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ such that $|\mathcal{S}||\mathcal{A}| < \infty$, number of iterations K, learning rates $(\alpha_1, \ldots, \alpha_K)$, initial $\widetilde{Q}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

for $k = 1, 2, \ldots, K$ **do**

Sample $r' \sim R(\cdot \mid s_k, a_k), s' \sim P(\cdot \mid s_k, a_k)$

Update action-value function:

$$\widetilde{Q}_k \leftarrow \widetilde{Q}_{k-1} + \alpha_k (r' + \max_{a' \in \mathcal{A}} \widetilde{Q}_{k-1}(s', a'))$$

Define π_K as the greedy policy w.r.t. $\widetilde{Q}_K$

Output: An estimator $\widetilde{Q}_K$ of Q^* and policy π_K

Theorem 3.3 (Mansour 2003)**.** Assume (P) and (D). Let $\alpha_k = 1/(k+1)^\omega$ where $\omega \in (1/2, 1]$. Fix $C > 0$, a sufficiently large constant. Let $\varepsilon, \delta > 0$ and define

$$A = \frac{4V_{\max}^2 \log(2|\mathcal{S}||\mathcal{A}|\, V_{\max}/\delta(1-\gamma)\varepsilon)}{(1-\gamma)^2 \varepsilon^2}, \quad B = 2\log(V_{\max}/\varepsilon)/(1-\gamma)$$

The following hold for any $\psi > 0$.

If the synchronos algorithm (algorithm 6) is run with

$$K \geqslant C \begin{cases} A^{1/\omega} + B^{1/(1-\omega)} & \omega \in (1/2, 1) \\ \frac{(2+\psi)^B}{\psi^2} \left(A + \frac{4V_{\max}^2 \log(1/\psi)}{(1-\gamma)^2 \varepsilon^2} \right) & \omega = 1 \end{cases}$$

then with probability at least $1 - \delta$ we have $\left\| \widetilde{Q}_K - Q^* \right\|_\infty < \varepsilon$.

If the asynchronos algorithm (algorithm 5) with

$$K \geqslant C \begin{cases} (L^{1+3\omega} A)^{1/\omega} + (LB)^{1/(1-\omega)} & \omega \in (1/2, 1) \\ \frac{(L+\psi L+1)^B}{\psi^2} \left(A + \frac{4V_{\max}^2 \log(1/\psi)}{(1-\gamma)^2 \varepsilon^2} \right) & \omega = 1 \end{cases}$$

and the state-action pairs $(s_1, a_1, \ldots, s_K, a_K)$ are drawn from a distribution such that every pair in $\mathcal{S} \times \mathcal{A}$ appears in every sequence of length at least $L > 0$, then with probability at least $1 - \delta$ we have $\left\| \widetilde{Q}_K - Q^* \right\|_\infty < \varepsilon$.

In [11] L is called the *covering rate*.

Remark 3.4. An interesting side note to theorem 3.3 is that one can use the bounds to give hints at how to tune the learning rate by changing ω. Optimizing for different scenarios yield different

learning theoretically optimal values for ω. For example if we want to optimize for the bound on K for $\gamma \to 1$ using the synchronos algorithm, we get the following rate (treating other variables as constant) $K \geqslant C'(1/(1-\gamma)^{4/\omega} + 1/(1-\gamma)^{1/(1-\omega)})$ for some $C' > 0$. Thus picking $\omega = 4/5$ is optimal. As another example: running the asynchronos algorithm and wanting to minimize for large covering rates L. We get $K \geqslant C''(L^{2+1/\omega} + L^{1/(1-\omega)})$ for some $C'' > 0$. This is optimized for $\omega \approx 0.77$. Then in [11] experiments points to $\omega = 0.85$ as being optimal using a scheme generating random finite MDPs. Other authors have since used this number as a standard value for the learning rate (see e.g. [4, Devraj and Meyn (2017)])

3.1.1 History dependent setting

Setting 6 (Finite HDP)**.**

1. A history dependent decision process (see definition 2.4), with a single *finite* state space, a single finite action space $(\mathcal{S}, \mathcal{A})$, and transition and reward kernels $(P_n, R_n)_{n\in\mathbb{N}}$. Define $\mathcal{H}^* := \bigcup_{i\in\mathbb{N}} \mathcal{H}_n$, the space of finite histories.
2. $(P_n)_{n\in\mathbb{N}}$ is viewed as a single kernel $P : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{S}$.
3. $(R_n)_{n\in\mathbb{N}}$ is deterministic and viewed as a single function $r : \mathcal{H}^* \to \mathbb{R}$. This is discounted by $\gamma \in [0, 1)$ in accordance to condition (D). That is $r(h_n)$ is bounded in the interval $[-\gamma^{n-2}R_{\max}, \gamma^{n-2}R_{\max}]$ for any $h_n \in \mathcal{H}_n$. Furthermore r depends only on $s_1a_1 \dots s_k r_{k-1}a_k$ when evaluated on $h_{k+1} = s_1a_1 \dots r_{k-1}a_k s_{k+1} \in \mathcal{H}_{k+1}$.

Remark 3.5. Note that the finite setting 6 is a special case of setting 1 considered by [Schäl, 1974], because Polishness and compactness of $\mathcal{S}, \mathcal{A}$ is readily implied by using the discrete topology in the finite state and action spaces, and the fact that (D) implies pt. 5 in setting 1. Further the conditions (S) and (W) of Schal are also both implied by the discreteness. This implies by theorem 2.25 the existence of an optimal $\pi^* \in R\Pi$ and that $V_n^* \to V^*$.

Within setting 6 Q-functions are generalized so that they are taking values in $\mathcal{H}^* \times \mathcal{A}$. We likewise generalize the T function by

$$TQ(h, a) := r' + \gamma \sum_{s\in\mathcal{S}} \max_{a'\in\mathcal{A}} Q(hr'as, a')P(s \mid ha), \qquad r' = r(h, a) \tag{3.4}$$

The optimal Q-function Q^* is defined in [Majeed, Hutter] as the fixed point of the T operator in $\mathcal{L}_\infty(\mathcal{H}^* \times \mathcal{A})$.

Now a function $\phi : \mathcal{H}^* \to \mathcal{X}$ is introduced which maps a history to a new finite space $\mathcal{X}$. The intuition here is that $x_n = \phi(h_{n-1}r_{n-2}a_{n-1}s_n)$ is the state s_n as it is perceived by the agent. This is called **partial observability**. ϕ is a assumed to be surjective so $\mathcal{X}$ is a finite space of reduced size in comparison to $\mathcal{S}$. In applications this could be a partially observable environment or a latent space.

This way we are now considering a class of problems which is wider than a history dependent decision process (HDP). Namely a partially observable HDP or shortened: POHDP. A HDP under setting 6 is the subclass of POHDP where $\mathcal{S} = \mathcal{X}$ and $\phi = \mathrm{id}_{\mathcal{S}}$.

Let $\phi_{hra}(s) = \phi(hras)$. Then we can define a kernel

$$\begin{aligned} &p_h : \{\phi(h)\} \times \mathcal{A} \rightsquigarrow \mathcal{X} \\ &p_h(x' \mid xa) = \sum_{s:\phi(hr'as)=x'} P(s \mid ha), \quad r' = r(h, a) \end{aligned}$$

or expressed as an image measure $p_h(\cdot \mid xa) = \phi_{hr'a}(P(\cdot \mid ha))$. and further function q_h^* by the equation

$$q_h^*(x,a) = r' + \gamma \sum_{x' \in \mathcal{X}} \max_{a' \in \mathcal{A}} q_h^*(x',a') p_h(x' \mid xa), \quad r' = r(h,a) \tag{3.5}$$

Assumption 1 (State-uniformity condition)**.** For any $h, h' \in \mathcal{H}^*$ we have

$$\phi(h) = \phi(h') \implies Q^*(h,\cdot) = Q^*(h',\cdot)$$

A process as in setting 6 together with the state-uniformity condition is by [11] called a *Q-Value uniform decision process* (QDP).

Theorem 3.6 (Hutter, 2016)**.** Under assumption 1 we have $q_{h'}^*(\phi(h),a) = Q^*(h,a)$ for any $h' \in \mathcal{H}^*$.

With this as a motivation we will try to use the standard TD update step as for an MDP environment:

$$q_{t+1}(x,a) = q_t(x,a) + \alpha_t(x,a)\left(r' + \gamma \max_{a \in \mathcal{A}} q_t(x',a') - q_t(x,a)\right), \quad x = \phi(h), r' = r(h,a) \tag{3.6}$$

Theorem 3.7. Within setting 6 assume

1. State-uniformity (assumption 1).
2. Any state is reached eventually under any policy (called *state-process ergodicity* in [11]).
3. Learning rate satisfies
$$\sum_{t=0}^{\infty} \alpha_t(x,a) = \infty, \quad \sum_{t=0}^{\infty} \alpha_t(x,a)^2 < \infty$$

Then starting with any $q_0 : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ the update step eq. (3.6) yields a sequence $(q_t)_{t \in \mathbb{N}}$ which converges to the optimal $q^* = Q^*$.

It seems relevant to ask how restrictive the state-uniformity assumption is. [11] answers this by an array of examples showing the following relations of the classes of decision processes:

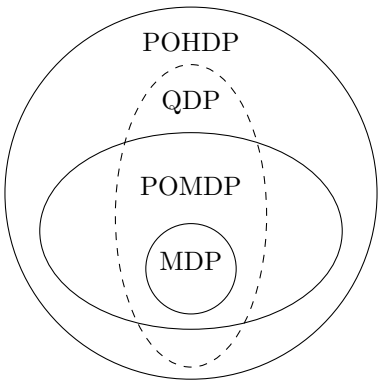


Figure 3.1: Classes of finite decision processes considered in [11], under setting 6.

Recall that QDP is a partially observable HDP under state-uniformity (assumption 1).

One point remains unclear after reading [11]: Why is q^* and Q^* well defined as the solution to their functional equations (eq. (3.4) and eq. (3.5)) and how are they related to the optimal value function $V^*(s) = \sup_{\pi \in R\Pi} \mathbb{E}_s^\pi \sum_{i=1}^{\infty} \gamma^{i-1} r_i$ (see definition 2.21) of a general HDP? A sensible thing to ask would be that $Q^*(h,a) = r(h,a) + \gamma \mathbb{E}_{P(\cdot|ha)} V^*$. However we will not go further into these details.

3.2 Linear function approximation

This section is based on [12, Melo and Ribeiro (2007)].

Setting 7 (Continuous state, finite action, discounted MDP)**.**

1. An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (see definition 2.27).
2. Discounted, i.e. (D) holds with $\gamma \in [0, 1)$.
3. $\mathcal{S} \subseteq \mathbb{R}^w$ is a compact subset of a euclidean space.
4. $\mathcal{A}$ is finite and not varying for each $s \in \mathcal{S}$, that is the set of admissible actions satisfy $A(s) = \mathcal{A}$ for all $s \in \mathcal{S}$.
5. r_i is upper semicontinuous .

Remark 3.8. Item 5 was actually not part of the assumptions in [12]. We include it here in order to ensure the existence of an optimal policy and thus measurability of V^*.

Let $\{\xi_1, \dots, \xi_M\}$ be a finite set of linearly independent, measurable and bounded action value functions, $\xi_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\forall i \in [M]$. Denote $\mathcal{F} := \operatorname{span}\{\xi_i \mid i \in [M]\}$ and for $\theta \in \mathbb{R}^M$

$$Q_\theta(s, a) = \sum_{i=1}^{M} \theta_i \xi_i(s, a) = \xi^T \theta$$

Note that $\mathcal{F} \subseteq \mathcal{L}_2(\mathcal{S} \times \mathcal{A})$ since any Q_θ is bounded and $\mathcal{S}$ is compact (so closed and bounded). We would now like to find the best approximation $q^* \in \mathcal{F}$ to Q^* within the span. If we measure distance by the $\mathcal{L}_2$-norm this is simply $\rho_{\mathcal{F}} Q^*$ where $\rho_{\mathcal{F}}$ is the orthogonal projection on $\mathcal{F}$. Denote by θ^* the coordinates of this projection, i.e. $\mathcal{F}_{\theta^*} = \rho_{\mathcal{F}} Q^*$.

The gradient of Q_θ over θ is

$$\nabla_\theta Q_\theta(s, a) = \xi(s, a)$$

This gives the idea for a temporal difference with approximation from $\mathcal{F}$ using the update step

$$\theta_{k+1} = \theta_k + \alpha_k \xi(s_k, a_k) \left(r_k + \gamma \max_{b \in \mathcal{A}} Q_{\theta_k}(s_{k+1}, b) - Q_{\theta_k}(s_k, a_k) \right)$$

Algorithm 7: Q-learning with linear approximation

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, policy π, number of iterations K, learning rates $(\alpha_1, \dots, \alpha_K)$, initial $\theta_1 \in \mathbb{R}^M$

for $k = 1, 2, \dots, K$ **do**

Sample $a_k \sim \pi(\cdot \mid s_k)$, $s_{k+1} \sim P(\cdot \mid s_k, a_k)$, $r_k \sim R(\cdot \mid s_k, a_k)$.

Update action-value parameter:

$$\theta_{k+1} = \theta_k + \alpha_k \xi(s_k, a_k) \left(r_k + \gamma \max_{b \in \mathcal{A}} Q_{\theta_k}(s_{k+1}, b) - Q_{\theta_k}(s_k, a_k) \right)$$

Define $\widetilde{\pi}_K$ as the greedy policy w.r.t. $\widetilde{Q}_K := Q_{\theta_{K+1}}$.

Output: An estimator $\widetilde{Q}_K$ of Q^* and policy $\widetilde{\pi}_K$

In order to understand the results of the analysis of algorithm 7 found in [12], we need to define some concepts from ergodic theory.

Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{X}$ be a transition kernel. Let $\mathfrak{P} = \kappa^\infty : \mathcal{X} \rightsquigarrow \mathcal{X}^\infty$. And denote by $\mathfrak{P}_x = \mathfrak{P}\delta_x \in \mathcal{P}(\mathcal{X}^\infty)$ the probability measure for the process starting at $x \in \mathcal{X}$. Let $\rho_i : \mathcal{X}^\infty \to \mathcal{X}$ be projection on the ith space. Define for any $A \in \Sigma_\mathcal{X}$ the function $\tau_A : \mathcal{X}^\infty \to \overline{\mathbb{N}} = \inf\{i \in \mathbb{N} \mid \rho_i \in A\}$. Intuitively this function records the earliest time where the process enter the set $A \subseteq \mathcal{X}$. Define the function $\eta_A : \mathcal{X}^\infty \to \overline{\mathbb{N}} = \sum_{i\in\mathbb{N}} 1_A \circ \rho_i$. This function records the total number of times in which the process is inside the set A. Let $\varphi \in \mathcal{P}(\mathcal{X})$ be a probability measure on $\mathcal{X}$.

Definition 3.9 (Invariant measure)**.** A countably additive measure $\mu \in \mathcal{P}(\mathcal{X})$ is said to be **invariant** w.r.t κ if $\kappa \circ \mu = \mu$.

Definition 3.10 (Positivity)**.**

$\mathfrak{P}$ is called **positive** if it admits an κ-invariant probability measure μ.

Definition 3.11 (Irreducibility)**.** $\mathfrak{P}$ is called φ-irreducible $\mathfrak{P}_x(\tau_A < \infty) > 0$ for all $A \in \Sigma_\mathcal{X}$ with $\varphi(A) > 0$ and all $x \in \mathcal{X}$.

Definition 3.12 (Harris recurrency)**.** $\mathfrak{P}$ is called φ-Harris recurrent if it it φ-irreducible and $\mathfrak{P}_x(\eta_A = \infty) = 1$ for all $A \in \Sigma_\mathcal{X}$ with $\varphi(A) > 0$ and all $x \in \mathcal{X}$.

Definition 3.13 (Geometric ergodicity)**.** A Markov process $\mathfrak{P}$ is called **geometrically ergodic** if it is positive with invariant measure μ, φ-Harris recurrent for some $\varphi \in \mathcal{P}(\mathcal{X})$ and $\exists t > 1$ such that

$$\sum_{i=1}^{\infty} t^i \|P_x^n - \mu\|_{TV} < \infty, \quad \forall x \in \mathcal{X}$$

Since the P, R of our MDP is reward independent we can view the MDP as a stationary process $\mathfrak{P}$ on $\mathcal{S}$ generated by kernel $P\pi$ for a policy $\pi \in S\Pi$.

Theorem 3.14 (Melo, Ribeiro)**.** Let $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be an MDP as of setting 7. Let $\pi \in S\Pi$ be a stationary process and $\mathfrak{P}$ the process kernel derived by $P\pi$. Assume that $\mathfrak{P}$ is geometrically ergodic with invariant measure μ and that $\pi(a \mid s) > 0$ for all $a \in \mathcal{A}$ and μ-almost all $s \in \mathcal{S}$. Assume that $\sum_{i=1}^{M} |\xi_i| \leqslant 1$. Then if algorithm 7 is run with learning rates from a sequence $\{\alpha_k\}_{k\in\mathbb{N}}$ satisfying $\alpha_k \in [0, 1]$ and

$$\sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty$$

we have that

$$\theta_k \to \theta^*$$

with probability 1, and $Q_{\theta *}$ satisfies

$$Q_{\theta *} = \rho_\mathcal{F} T Q_{\theta *}$$

Furthermore the orthogonal projection is expressible as

$$\rho_\mathcal{F} Q = \xi^T \frac{\mathbb{E}_{\pi\mu}(\xi Q)}{\mathbb{E}_{\pi\mu}(\xi\xi^T)}$$

(recall the definition of the kernel-derived measure $\pi\mu(S \times A) = \int_S \pi(A \mid s)\, \mathrm{d}\mu(s)$).

This gives us our first sort of convergence garantee for Q-learning in continuous state space setting. However there is still room for improvement since theorem 3.14 does not tell us:

1. How fast is the convergence?
2. How far is $Q_{\theta *}$ from Q^*?

3. How far is $Q_{\widetilde{\pi}_K}$ from Q^*?

Remark 3.15. Question 2. is probably best handled seperately for each function class $\mathcal{F}$.

In a quite similar setting these questions are answered for the fitted Q-iteration algorithm in the next section (theorem 3.25).

3.3 Deep Q-learning

3.3.1 Introduction

This section is mostly based on the paper [5, Fan et al. (2020+)].

Similar to the linear function approximation (see section 3.2), in deep Q-learning we use a class of functions parametrized by some set $\Theta \subseteq \mathbb{R}^D$. This time the function class is not linear combinations of a set of basis functions, but a class of artificial neural networks. Also we use the same setting (setting 7) of a continuous state space, finite action space discounted MDP.

Though [5] claims to investigate the deep Q-network algorithm (see algorithm 9), instead of analysing DQN directly, a related algorithm called *fitted Q-iteration* (FQI) is analysed instead and bounds on its convergence is established.

The algorithm analysed by [Fan et al] is

Algorithm 8: Fitted Q-Iteration Algorithm

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, function class $\mathcal{F}$, sampling distribution ν, number of iterations K, number of samples n, initial estimator $\widetilde{Q}_0$

for $k = 0, 1, 2, \ldots, K-1$ **do**

Sample i.i.d. observations $\{(S_i, A_i), i \in [n]\}$ from ν obtain $R_i \sim R(S_i, A_i)$ and $S_i' \sim P(S_i, A_i)$

Let $Y_i = R_i + \gamma \cdot \max_{a \in \mathcal{A}} \widetilde{Q}_k(S_i', a)$

Update action-value function:

$$\widetilde{Q}_{k+1} \leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(S_i, A_i))^2$$

Define π_K as the greedy policy w.r.t. $\widetilde{Q}_K$

Output: An estimator $\widetilde{Q}_K$ of the optimal value function Q^* and an estimator of the optimal policy $\widetilde{\tau}_K$

We will now look at the function class used in [5].

ReLU Networks

Remark 3.16. For a artificial neural network $f \in \mathcal{DN}\left((\sigma_i, d_i)_{i=1}^{L+1})\right)$ with weights $(\widetilde{W}_i)_{i=1}^{L+1} = (W_i, v_i)_{i=1}^{L+1}$ we can consider the size of the maximum weight of the ith weight pair $\left\|\widetilde{W}_i\right\|_\infty$, and its number of non-zero $\left\|\widetilde{W}_i\right\|_0$.

Definition 3.17 (Sparse ReLU Networks)**.** For $s, V \in \mathbb{R}$ a (s,V)-**Sparse ReLU Network** is an artificial neural network f with all activation functions being *ReLU* i.e. $\sigma_{ij} = \max(\cdot, 0)$ and with weights (W_ℓ, v_ℓ) satisfying

- $\max_{\ell\in[L+1]}\left\|\widetilde{W}_\ell\right\|_\infty \leqslant 1$
- $\sum_{\ell=1}^{L+1}\left\|\widetilde{W}_\ell\right\|_0 \leqslant s$
- $\max_{j\in[d_{L+1}]}\|f_j\|_\infty \leqslant V$

The set of them we denote $\mathcal{F}\left(L, \{d_i\}_{i=0}^{L+1}, s, V\right)$.

Remark 3.18. Following the graph interpretation of ANNs (see remark 2.64) the condition that $\sum_{i=1}^{L+1}\left\|\widetilde{W}_i\right\|_0 \leqslant s$ corresponds to graph-theoretical sparsity of the graph derived from the ANN.

The idea to work with this particular subclass of neural networks come from [15] (p. 22), which establishes the following lemma

Lemma 3.19 (Approximation of Hölder Smooth Functions by ReLU networks)**.** Let $m, M \in \mathbb{Z}_+$ with $N \geqslant \max\{(\beta+1)^r, (H+1)e^r\}$, $L = 8 + (m+5)(1+\lceil\log_2(r+\beta)\rceil)$, $d_0 = r, d_j = 6(r+\lceil\beta\rceil)N, d_{L+1} = 1$. Then for any $g \in \mathcal{C}_r\left([0,1]^r, \beta, H\right)$ there exists a ReLU network $f \in \mathcal{F}\left(L, \{d_j\}_{j=0}^{L+1}, s, \infty\right)$ with $s \leqslant 141(r+\beta+1)^{3+r}N(m+6)$ such that

$$\|f-g\|_\infty \leqslant (2H+1)6^r N(1+r^2+\beta^2)2^{-m} + H3^\beta N^{-\beta/r}$$

In the course of establishing the results in [5] we will not go more into this result or other properties of ReLU networks in particular, instead putting emphasis on how to use this result to obtain the main theorem, which we will present shortly.

3.3.2 Assumptions

Hölder Smoothness

Definition 3.20 (Hölder smoothness)**.** For $f : \mathcal{S} \to \mathbb{R}$ we define

$$\|f\|_{C_w} := \sum_{|\alpha|<\beta}\|\partial^\alpha f\|_\infty + \sum_{\|\alpha\|_1=\lfloor\beta\rfloor}\sup_{x\neq y}\frac{|\partial^\alpha(f(x)-f(y))|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} \tag{3.7}$$

Where $\alpha = (\alpha_1, \ldots, \alpha_w) \in \mathbb{N}_0^w$. And ∂^k is the partial derivative w.r.t. the kth variable. If $\|f\|_{C_w} < \infty$ then f is **Hölder smooth**. Given a compact subset $\mathcal{D} \subseteq \mathbb{R}^w$ the space of Hölder smooth functions on $\mathcal{D}$ with norm bounded by $H > 0$ is denoted

$$C_w(\mathcal{D}, \beta, H) := \left\{f : \mathcal{D} \to \mathbb{R} \;\middle|\; \|f\|_{C_w} \leqslant H\right\}$$

Definition 3.21. Let $t_j, p_j \in \mathbb{N}$, $t_j \leqslant p_j$ and $H_j, \beta_j > 0$ for $j \in [q]$. We say that f is a **composition of Hölder smooth functions** when

$$f = g_q \circ \cdots \circ g_1$$

for some functions $g_j : [a_j, b_j]^{p_j} \to [a_{j+1}, b_{j+1}]^{p_{j+1}}$ that only depend on t_j of their inputs for each of their components g_{jk}, and satisfies $g_{jk} \in C_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j)$, i.e. they are Holder smooth. We denote the class of these functions

$$\mathcal{G}(\{p_j, t_j, \beta_j, H_j\}_{j\in[q]})$$

Definition 3.22. Define

$$\mathcal{F}_0 = \left\{f : \mathcal{S}\times\mathcal{A} \to \mathbb{R} \;\middle|\; f(\cdot, a) \in \mathcal{F}(s, V)\ \forall a \in \mathcal{A}\right\}$$

and

$$\mathcal{G}_0 = \left\{f : \mathcal{S}\times\mathcal{A} \to \mathbb{R} \;\middle|\; f(\cdot, a) = \mathcal{G}(\{p_j, t_j, \beta_t, H_j\}_{j\in[q]})\ \forall a \in \mathcal{A}\right\}$$

We here recall the definition of the operators for Q-functions (definition 2.53). For any stationary policy $\tau \in S\Pi$ we define

$$P_\tau Q(s,a) = \int Q(s',a') \, \mathrm{d}\tau P(s',a' \mid s,a)$$
$$T_\tau Q = r + \gamma P_\tau Q$$
$$TQ(s,a) = r(s,a) + \gamma \int \max_{a' \in \mathcal{A}} Q(s',a') \, \mathrm{d}P(s' \mid s,a)$$

matching the definitions in [5].

Assumption 2. $T\mathcal{F}_0 \subseteq \mathcal{G}_0$. I.e. t is assumed that $Tf \in \mathcal{G}_0$ for any $f \in \mathcal{F}_0$, so when using the Bellman optimality operator on our sparse ReLU networks, we should stay in the class of compositions of Holder smooth functions.

If also $\mathcal{G}_0$ is well approximated by functions in $\mathcal{F}_0$ then this assumption implies that $\mathcal{F}_0$ is approximately closed under the Bellman operator T and thus that Q^* is close to $\mathcal{F}_0$.

We now look at a simple example where assumption 2 holds: Seting $\mathcal{D} = [0,1]^r$, $q = 1$ and taking both the expected reward function and transition kernel to be Hölder smooth.

Example 3.23. Assume for all $a \in \mathcal{A}$ that $P(\cdot \mid s,a)$ is absolutely continuous w.r.t. λ^k (the k dimensional Lebesgue measure) with density $p(\cdot \mid s,a)$, that for all $s' \in \mathcal{S}$ we have $s \mapsto p\left(s' \mid s,a\right)$ and $s \mapsto r(s,a)$ are both Hölder smooth in the class $C_w([0,1]^r, \beta, H)$. Then

$$T\mathcal{F}_0 \subseteq C_w([0,1]^r, \beta, (1+\gamma V_{\max})H)$$

To see this let Let $f \in \mathcal{F}_0$ and $\alpha \in \mathbb{N}_0^w$. Observe that

$$\begin{aligned}\partial^\alpha (Tf)(s,a) &= \partial_s^\alpha \left(r(s,a)\right) + \gamma \int_{\mathcal{S}} \partial_s^\alpha \left[\max_{a' \in \mathcal{A}} f(s',a') p\left(s' \mid s,a\right)\right] \mathrm{d}s' \\ &\leqslant \partial_s^\alpha \left(r(s,a)\right) + \gamma V_{\max} \sup_{s' \in \mathcal{S}} \partial_s^\alpha p\left(s' \mid s,a\right)\end{aligned}$$

similarly

$$\begin{aligned}\partial^\alpha (Tf)(s,a) - \partial^\alpha (Tf)(s',a) \leqslant{}& \partial_s^\alpha \left(r(s,a)\right) - \partial_s^\alpha \left(r(s',a)\right) \\ &+ \gamma V_{\max} \sup_{s'' \in \mathcal{S}} \left(\partial_s^\alpha p(s'' \mid s,a) - \partial_s^\alpha p(s'' \mid s',a)\right)\end{aligned}$$

Thus since p and r are Hölder smooth

$$\begin{aligned}\|Tf\|_{C_w} \leqslant{}& \sum_{|\alpha|<\beta} \left(\|\partial^\alpha r(\cdot,a)\|_\infty + \gamma V_{\max} \sup_{s \in \mathcal{S}} \|\partial^\alpha p(s \mid \cdot, a)\|_\infty\right) \\ &+ \sum_{\|\alpha\|_1 = \lfloor\beta\rfloor} \sup_{x \neq y} \left(\frac{|\partial^\alpha (r(x,a) - r(y,a))|}{\|x-y\|_\infty^{\beta - \lfloor\beta\rfloor}} + \gamma V_{\max} \sup_{s \in \mathcal{S}} \frac{|\partial^\alpha (p(s \mid x,a) - p(s \mid y,a))|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}}\right) \\ \leqslant{}& H + \gamma V_{\max} H = (1+\gamma V_{\max})H\end{aligned}$$

Concentration coefficients

Definition 3.24 (Concentration coefficients)**.** Let $\nu_1, \nu_2 \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be probability measures, absolutely continuous w.r.t. $\lambda^w \otimes \mu_{\mathcal{A}}$ (the product of the w-dimensional Lebesgue measure and the counting measure on $\mathcal{A}$). Define

$$\kappa(m, \nu_1, \nu_2) = \sup_{\pi_1, \ldots, \pi_m} \left[\mathbb{E}_{\nu_2} \left(\frac{\mathrm{d}(P_{\pi_m} \ldots P_{\pi_1} \nu_1)}{\mathrm{d}\nu_2}\right)^2\right]^{1/2}$$

where $\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}$ are the Radon-Nikodym derivative of the measures μ_1, μ_2 (see Todo reference).

Assumption 3. Let ν be the sampling distribution from the algorithm, and μ the distribution over which we measure the error in the main theorem, then we assume

$$(1-\gamma)^2 \sum_{m \geqslant 1} \gamma^{m-1} m\kappa(m,\mu,\nu) = \phi_{\mu,\nu} < \infty$$

3.3.3 The main theorem

Theorem 3.25 (Fan, Yang, Xie, Wang)**.** Under **??** let μ be any distribution over $\mathcal{S} \times \mathcal{A}$. Make assumption 2 and assumption 3 with the constants $\phi_{\mu,\nu} > 0$, $q \in \mathbb{N}$ and $\{p_j, t_j, \beta_j, H_j\}_{j\in[q]}$. Furthermore assume that there exists a constant $\xi > 0$ such that

$$\max\left\{\sum_{j=1}^{q}(t_j+\beta_j+1)^{3+t_k}, \sum_{j=1}^{q}\log(t_j+\beta_j), \max_{j\in[q]} p_j\right\} \leqslant (\log n)^{\xi}$$

Set $\beta_j^* = \beta_j \prod_{\ell=j+1}^{q} \min(\beta_\ell, 1)$ for $j \in [q-1]$, $\beta_q^* = 1$, $\alpha^* = \max_{j\in[q]} t_j/(2\beta_j^* + t_j)$, $\xi^* = 1 + 2\xi$ and $\kappa^* = \min_{j\in[q]} \beta_j^*/t_j$. Then there exists a class of ReLU networks

$$\mathcal{F}_0 = \{f : \mathcal{S}\times\mathcal{A} \to \mathbb{R} : f(\cdot, a) \in \mathcal{F}(\widetilde{L}, \{\widetilde{d}_j\}_{j=0}^{\widetilde{L}+1}, \widetilde{s}) \mid a \in \mathcal{A}\}$$

with structure satisfying

$$\widetilde{L} \lesssim (\log n)^{\xi^*}, \widetilde{d}_0 = r, \widetilde{d}_j \leqslant 6n^{\alpha^*}(\log n)^{\xi^*}, d_{L+1} = 1, \widetilde{s} \lesssim n^{\alpha^*} \cdot (\log n)^{\xi^*}$$

such that when running algorithm 8 with $\mathcal{F}_0$ and n is sufficiently large

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} \leqslant C_\varepsilon \frac{\phi_{\mu,\nu}\gamma}{(1-\gamma)^2} V_{\max}^2 n^{\max\{-2\alpha^*\kappa^*, (\alpha^*-1)/2\}} \log(n)^{1+2\xi^*} + \frac{4\gamma}{(1-\gamma)^2} R_{\max}\gamma^K$$

where $C_\varepsilon > 0$ is a constant not depending on n or K. Thus

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} = \mathcal{O}\left(n^{-\varepsilon^*}\log(n)^{c^*} + \gamma^K\right)$$

for some $\varepsilon^*, c^* > 0$.

This bound on the convergence of the FQI algorithm is quite remarkable in terms of the broad class of environments (problems) that it shows can be solved approximatively by using sampling from the environment to update a ANN-represented Q-function. In particular it is the most general result on convergence rates for model-free and continuous state space algorithms, among the sources we survey in this thesis.

3.3.4 Relation to DQN

The following is famous *DQN*-algorithm proposed by [13, Mnih et al. (2015)].

Algorithm 9: Deep Q-Network

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations K, batch size n, exploration factor ϵ, function class $\mathcal{F}$ of deep neural networks parametrized by some $\Theta \subseteq \mathbb{R}^D$, $D \in \mathbb{N}$, target update frequency T_{target}, learning rates $\{\alpha_t\}_{t \geqslant 0}$

Initialize replay memory $\mathcal{M} \leftarrow \varnothing$ as empty.

Pick a initial Q-network $\widetilde{Q}_0 = Q(\theta_0)$ by sampling $\theta_0 \in \Theta$ from some distribution.

Initialize target network $Q_{\text{target},0} = \widetilde{Q}_0$ by picking the target parameters $\theta_0^* = \theta_0$ and setting $Q_{\text{target},0} = Q(\theta_0^*)$.

for $k = 0, 1, 2, \ldots, K-1$ **do**

With probability ϵ sample A_k uniformly from $\mathcal{A}$, and with probability $1 - \epsilon$ choose A_k greedily with respect to $\widetilde{Q}_k$, that is A_k is picked from $\operatorname{argmax}_{a \in \mathcal{A}} \widetilde{Q}_k(S_k, a)$.

Sample (observe) S_{k+1} and R_k (from $P(\cdot \mid S_k, A_k)$ and $R(\cdot \mid S_k, A_k)$).

Store the transition (S_k, A_k, R_k, S_{k+1}) in the replay memory $\mathcal{M}$, potentially replacing an old (random) transition if the memory is *full*.

Experience replay: Sample batch of transitions $(s_i, a_i, r_i, s_i')_{i \in [n]}$ from the replay memory $\mathcal{M}$.

For each $i \in [n]$ let $Y_i = r_i + \gamma \max_{a \in \mathcal{A}} Q_{\text{target},\ell(k)}(s_i', a)$.

Update the Q-network by performing a gradient descent step

$$\theta_{k+1} \leftarrow \alpha_k \frac{1}{n} \sum_{i=1}^{n} (Y_i - Q(\theta_k)(s_i, a_i)) \cdot \nabla_\theta Q(\theta)(s_i, a_i)$$

For every T_{target} steps update the target network by setting $Q(\theta^*_{\ell(k+1)}) \leftarrow Q(\theta^*_{\ell(k)})$ where $\ell(k)$ is the number of updates of the target network at step k.

Put $\widetilde{Q}_K = Q(\theta_K)$ and pick a greedy policy $\widetilde{\tau}$ with respect to $\widetilde{Q}_K$.

Output: An estimator $\widetilde{Q}_K$ of the optimal value function Q^* and $\widetilde{\tau}_K$ an estimator of the optimal policy τ^*.

Algorithm 9 is an off-policy algorithm because it updates the parameter θ_k based on picking the greedy action of the target network $Q_{\text{target},\ell(k)}$, while the policy being followed is an ϵ-greedy policy where the greedy part is with respect to $\widetilde{Q}_k$.

[5] stresses two *tricks* that drives the succes of DQN, which is the use of

1. *experience replay*
2. *target network*

Experience replay is the basic method of keeping a replay memory set (or *buffer*) from which samples (or *mini-batches*) are drawn which then are used in each gradient descent update of the Q-network.

The target network is a past version of the Q-network that is used in the gradient step update as the goal after using the Bellman operator, It is then updated every T_{target} steps.

In practice the size of the replay memory buffer is very large, for example in [13] it holds $\sim 10^6$ transitions. Because of this it is argued in [5] that

> "experience replay is close to sampling independent transitions from a given distribution $\sigma \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$"

For the target network it is argued that when having a large enough batch (n) and using $\widetilde{Q}_{k-1}$ to update $\widetilde{Q}_k$, the role of $\widetilde{Q}_{k-1}$ becomes similar to the target network $Q_{\text{target},\ell(k-1)}$ of DQN.

Critique of this relation

While the arguments for the similarity between FQI and DQN are intuitively reasonable, the rigiorous proofs are missing and it is unclear if a convergence result about FQI has implications for DQN.

3.3.5 Differences in notation

Because σ is used ambigously in theorem 3.25 we denote the probability distribution σ from [5] p. 20 by ν instead. I avoid the shorthand defined in [5] p. 26 bottom: $\|f\|_n^2 = 1/n \cdot \sum_{i=1}^n f(X_i)^2$. and use p-norms instead. The conversion to the notation used here becomes $\|f\|_n \rightsquigarrow \|f\|/n$. The letter r is used in [5] to denote the euclidean dimension of the state space, while here we use w.

3.3.6 Proofs

The proof of theorem 3.25 combines two results. The first on the error propagation and the second on the error ocurring in a single step.

Theorem 3.26 (Error Propagation)**.** Let $\{\widetilde{Q}_i\}_{0 \leqslant i \leqslant K}$ be the iterates of the fitted Q-iteration algorithm. Then

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} \leqslant \frac{2\phi_{\mu,\nu}\gamma}{(1-\gamma)^2} \cdot \varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} \cdot R_{\max}$$

Where

$$\varepsilon_{\max} = \max_{k \in [K]} \left\| T\widetilde{Q}_{k-1} - \widetilde{Q}_k \right\|_{2,\nu}$$

Theorem 3.27 (One-step Approximation Error)**.** Let

- $\mathcal{F} \subseteq \mathcal{B}(\mathcal{S} \times \mathcal{A}, V_{\max})$ be a class of bounded measurable functions
- $\mathcal{G} = T(\mathcal{F})$ the class of functions obtainable by applying T to some function in $\mathcal{F}$.
- $\nu \in \mathcal{P}(\mathcal{S}, \mathcal{A})$ be a probability measure
- $(S_i, A_i)_{i \in [n]}$ be n i.i.d. samples following ν
- $(R_i, S_i')_{i \in [n]}$ be the rewards and next states sampled corresponding to the samples
- $Q \in \mathcal{F}$ be fixed
- $Y_i = R_i + \gamma \max_{a \in \mathcal{A}} Q(S_i', a)$
- $\widehat{Q} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(S_i, A_i) - Y_i)^2$
- $\kappa \in (0, 1],\ \delta > 0$ be fixed
- $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ a minimal δ-covering of $\mathcal{F}$ w.r.t. $\|\cdot\|_\infty$
- $N_\delta = \left|\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)\right|$ the number of elements in this covering

Then

$$\begin{aligned}\left\|\widehat{Q} - TQ\right\|_\nu^2 \leqslant & \frac{(1+\kappa)^2}{\kappa} \frac{1}{n} C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa)\left(\delta C_2^2 V_{\max}^2 \log(N_\delta) + \omega(\mathcal{F})\right) \\ & + 8\sqrt{2} V_{\max} n^{-1/2} \sqrt{\log N_\delta} + 8 V_{\max}(n^{-1} + \delta)\end{aligned}$$

Where

$$\omega(\mathcal{F}) = \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} \frac{1}{n} \mathbb{E}\|f - g\|_\nu^2$$

The proofs of theorem 3.26 and theorem 3.27 are found below, but first we will show how to combine them to obtain theorem 3.25.

Proof of main theorem. Using theorem 3.26 we get

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} \leqslant \frac{2\phi_{\mu,\nu}\gamma}{(1-\gamma)^2}\varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2}R_{\max} \tag{3.8}$$

where $\varepsilon_{\max} = \max_{k\in[K]}\left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\|_{2,\nu}$. Using theorem 3.27 with $Q = \widetilde{Q}_{k-1}$, $\mathcal{F} = \mathcal{F}_0$, $\epsilon = 1$ and $\delta = 1/n$, we get

$$\varepsilon_{\max} \leqslant 6n^{-1}C_2^2V_{\max}^2\log(N_0) + 2\omega(\mathcal{F}_0) + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_0} + 16V_{\max}n^{-1} \tag{3.9}$$

where $N_0 = \left|\mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty)\right|$. The remains only to wound $\omega(\mathcal{F}_0)$ and N_0, starting with $\omega(\mathcal{F}_0)$.

Step 1. We want to employ the lemma by [15] (lemma 3.19) to each Hölder smooth part of $g = Tf$ where f is an arbitrary network in $\mathcal{F}_0$ and then piece it together somehow, using that ReLU networks are easily stitched together into bigger ReLU networks. Therefore the first step is to refit our Hölder Smooth compositions in $\mathcal{G}_0$ to be defined on a hyper-cube instead. This is a relatively simple procedure:

Let $f \in \mathcal{G}_0$ then $f(\cdot, a) \in \mathcal{G}(\{p_j, t_j, \beta_j, H_j\})$ for all $a \in \mathcal{A}$. Therefore $f(\cdot, a) = g_q \circ \cdots \circ g_1$ where the (sub-)components $(g_{jk})_{k=1}^{p_{j+1}} = g_j$ satisfy

$$g_{jk} \in C_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j), \qquad j \in [q], k \in [p_{j+1}] \tag{3.10}$$

Here $a_1 = 0, b_1 = 1$ and, $a_j < b_j \in \mathbb{R}$ are some real numbers for $2 \leqslant j \leqslant q$. Notice that the Hölder smooth condition implies that $g_{jk}([a_j, b_j]^{t_j}) \subseteq [-H_j, H_j]$. Define

$$\begin{aligned} h_1 =& g_1/(2H_1) + 1/2 \\ h_j(u) =& g_j(2H_{j-1}u - H_{j-1})/(2H_j) + 1/2, \qquad j \in \{2, \ldots, q-1\} \\ h_q(u) =& g_q(2H_{q-1}u - H_{q-1}) \end{aligned} \tag{3.11}$$

Then $g_q \circ \cdots \circ g_1 = h_q \circ \cdots \circ h_1$ and

$$\begin{aligned} h_{1k} &\in C_{t_1}([0,1]^{t_1}, \beta_1, 1) \\ h_{jk} &\in C_{t_j}([0,1]^{t_j}, \beta_j, (2H_{j-1})^{\beta_j}), \qquad j \in \{2, \ldots, q-1\} \\ h_q &\in C_{t_q}([0,1]^{t_q}, \beta_q, H_q(2H_{q-1})^{\beta_q}) \end{aligned} \tag{3.12}$$

This concludes our construction of the refit of the components of g to unit intervals.

Step 1.1

Define $N := \max_{j\in[q]} n^{t_j/(2\beta_j^* + t_j)}$ $\eta := \log\left((2W+1)6^{t_j}N/(W3^{\beta_j}N^{-\beta_j/t_j})\right)$, and $m := \eta\lceil\log_2 n\rceil$, and assume n is sufficiently large such that $N \geqslant \max\left\{(\beta_j+1)^{t_j}, (H_j+1)e^{t_j} \mid j \in [q]\right\}$.

$$W := \max\left(\left\{(2H_{j-1})^{\beta_j} \mid 1 \leqslant j \leqslant q-1\right\} \cup \left\{H_q(2H_{q-1})^{\beta_q}, 1\right\}\right) \tag{3.13}$$

By lemma 3.19 there exists a ReLU network

$$\widehat{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \widetilde{d}_j p_{j+1}, \ldots, \widetilde{d}_j p_{j+1}, p_{j+1}\right\}, (\widetilde{s}_j + 4)\cdot p_{j+1}\right) \tag{3.14}$$

where $\widetilde{d}_j = 6(t_j + \lceil\beta_j\rceil)N$ and $\widetilde{s}_j \leqslant 141(t_j + \beta_j + 1)^{3+t_j}N(m+6)$ such that

$$\left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty \leqslant (2W+1)6^{t_j}N2^{-m} + W3^{\beta_j}N^{-\beta_j/t_j} \leqslant 2W3^{\beta_j}N^{-\beta_j/t_j} \tag{3.15}$$

Since h_{j+1} is defined on $[0,1]^{t_{j+1}}$ but $\widetilde{h}_j$ takes values in $\mathbb{R}$ we need to restrict $\widetilde{h}_j$ somehow to stitch the two together (by function composition). This is easily done by

Lemma 3.28. Restriction to $[0,1]$ is expressible as a two-layer ReLU network with 4 non-zero weights.

Proof. Namely $\tau(u) = 1 - (1-u)_+ = \min\left\{\max\left\{u, 0\right\}, 1\right\}$. □

Now define $\widetilde{h}_{jk} = \tau \circ \widehat{h}_{jk}$ (and $\widetilde{h}_j = (\widetilde{h}_{jk})_{k\in[p_{j+1}]}$). Then

$$\widetilde{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \widetilde{d}_j, \ldots, \widetilde{d}_j, 1\right\}, (\widetilde{s}_j + 4)p_{j+1}\right) \tag{3.16}$$

and since $h_{jk}([0,1]^{t_j}) \in [0,1]$ by eq. (3.15)

$$\begin{aligned}
\left\|\widetilde{h}_{jk} - h_{jk}\right\|_\infty &= \left\|\tau \circ \widehat{h}_{jk} - \tau \circ h_{jk}\right\|_\infty && (3.17)\\
&\leqslant \left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty && (3.18)\\
&\leqslant 2W3^{-\beta_j} N^{-\beta_j/t_j} && (3.19)
\end{aligned}$$

Having employed lemma 3.19 we now need to stitch it back together:

Step 2. Now define $\widetilde{f} : \mathcal{S} \to \mathbb{R}$ as $\widetilde{f} = \widetilde{h}_1 \circ \cdots \circ \widehat{h}_1$. If we set $\widetilde{L} := \sum_{j=1}^q (L_j + 2)$, $\widetilde{d} := \max_{j\in[q]} \widetilde{d}_j p_{j+1}$ and $\widetilde{s} := \sum_{j=1}^q (\widetilde{s}_j + 4)p_{j+1}$. Then $\widetilde{f} \in \mathcal{F}\left(\widetilde{L}, \left\{r, \widetilde{d}, \ldots, \widetilde{d}, 1\right\}, \widetilde{s}\right)$. We now take a moment to verify the size of the constants involved in the network. Starting with $\widetilde{L}$.

$$\begin{aligned}
\widetilde{L} &\leqslant \sum_{j=1}^q (L_j + 2)\\
&= \sum_{j=1}^q (8 + (\eta\lceil\log_2 n\rceil + 5)(1 + \lceil\log_2(\beta_j + t_j)\rceil))\\
&\leqslant \sum_{j=1}^q (8 + (\eta\log_2 n + \eta + 5)(2 + \log_2(\beta_j + t_j)))\\
&\leqslant 8q + (2\eta + 5)\log_2(n)\sum_{j=1}^q (2 + \log_2(\beta_j + t_j))\\
&\leqslant 8q + (2\eta + 5)\log_2(n)(2q + \log(n)^\xi)\\
&\leqslant (10q + 1)(2\eta + 5)\log_2(e)\log(n)^{1+\xi}\\
&\leqslant C_{\widetilde{L}}\log(n)^{1+2\xi}
\end{aligned}$$

where $C_{\widetilde{L}} = (10q + 1)(2\eta + 5)\log_2(e)$. For $\widetilde{d}$ we have

$$\begin{aligned}
\widetilde{d} &= \max_{j\in[q]} \widetilde{d}_j p_{j+1}\\
&= \max_{j\in[q]} 6(t_j + \beta_j + 1)Np_{j+1}\\
&\leqslant 6N(\max_{j\in[q]} p_j)(\max_{j\in[q]}(t_j + \beta_j + 1))\\
&\leqslant 6N(\log n)^{2\xi}\\
&\leqslant 6n^{\alpha^*}(\log n)^{\xi^*}
\end{aligned}$$

and for $\widetilde{s}$

$$\begin{aligned}
\widetilde{s} &= \sum_{j=1}^{q}(\widetilde{s}_j+4)p_{j+1} \\
&\leqslant \sum_{j=1}^{q}(141N(m+6)(t_j+\beta_j+1)^{3+t_j}+4)p_{j+1} \\
&\leqslant 142N(\log n)^{\xi}(2\eta+6)\log_2(n)\sum_{j=1}^{q}(t_j+\beta_j+1)^{3+t_j} \\
&\leqslant 142N(\log n)^{\xi}(2\eta+6)\log_2(e)\log(n)(\log n)^{\xi} \\
&= 142N(2\eta+6)\log_2(e)(\log n)^{1+2\xi} \\
&= C_{\widetilde{s}}n^{a^*}(\log n)^{\xi^*}
\end{aligned}$$

where $C_{\widetilde{s}} = 142(2\eta+6)\log_2(e)$. Now we bound $\left\|\widetilde{f}-f(\cdot,a)\right\|_\infty$. Define $G_j = h_j\circ\cdots\circ h_1$, $\widetilde{G}_j = \widetilde{h}_j\circ\cdots\circ\widetilde{h}_1$ for $j\in[q]$, $\lambda_j = \prod_{\ell=j+1}^{q}(\beta_\ell\wedge 1)$ for all $j\in[q-1]$ and $\lambda_q = 1$. We have

$$\begin{aligned}
\left\|G_j-\widetilde{G}_j\right\|_\infty &= \left\|h_j\circ G_{j-1}-h_j\circ\widetilde{G}_{j-1}+h_j\circ\widetilde{G}_{j-1}-\widetilde{h}_j\circ\widetilde{G}_{j-1}\right\| \\
&\leqslant \left\|h_j\circ\widetilde{G}_{j-1}-h_j\circ G_{j-1}\right\|_\infty+\left\|h_j\circ\widetilde{G}_{j-1}-h_j\circ G_{j-1}\right\|_\infty \\
&\leqslant W\left\|G_{j-1}-\widetilde{G}_{j-1}\right\|_\infty^{\beta_j\wedge 1}+\left\|\widetilde{h}_j-h_j\right\|_\infty^{\lambda_j}
\end{aligned}$$

so by induction and eq. (3.15)

$$\begin{aligned}
\left\|f(\cdot,a)-\widetilde{f}\right\|_\infty &= \left\|G_q-\widetilde{G}_q\right\|_\infty \\
&\leqslant W^q\sum_{j-1}^{q}\left\|\widetilde{h}_j-h_j\right\|_\infty^{\lambda_j} \\
&\leqslant W^q\sum_{j-1}^{q}\left(2W3^{\beta_j}N^{-\beta_j/t_j}\right)^{\lambda_j} \\
&\leqslant 2q3^{\max_{j\in[q]}\beta_j^*}W^{q+1}\max_{j\in[q]}N^{-\beta_j^*/t_j} \\
&\leqslant c_N^{1/2}\max_{j\in[q]}n^{-\alpha^*\beta_j^*/t_j} \\
&\leqslant c_N^{1/2}n^{-\alpha^*\min_{j\in[q]}\beta_j^*/t_j}
\end{aligned}$$

and therefore

$$\begin{aligned}
\omega(\mathcal{F}_0) &= \sup_{g\in\mathcal{G}}\inf_{f\in\mathcal{F}} \\
&\leqslant C_Nn^{-2\alpha^*\min_{j\in[q]}\beta_j^*/t_j} \\
&\leqslant C_Nn^{-2\alpha^*\kappa^*} \qquad (3.20)
\end{aligned}$$

where we define $\kappa^* = \min_{j\in[q]}\beta_j^*/t_j$.

Step 3. Finally what is left is to bound the covering number of $\mathcal{F}_0$. Denote by $\mathcal{N}_\delta$ the δ-covering of $\mathcal{F}\left(\widetilde{L},\{\widetilde{d}_j\}_{j=1}^{\widetilde{L}+1},\widetilde{s}\right)$ by

$$\mathcal{N}_\delta := \mathcal{N}\left(\delta,\mathcal{F}\left(\widetilde{L},\{\widetilde{d}_j\}_{j=1}^{\widetilde{L}+1},\widetilde{s}\right),\|\cdot\|_\infty\right)$$

Since $\mathcal{N}_\delta$ is a covering, for any $f\in\mathcal{F}_0$ and $a\in\mathcal{A}$ you can find a $g_a\in\mathcal{N}_\delta$ such that $\|f(\cdot,a)-g_a\|_\infty<\delta$. Now let $g:\mathcal{S}\times\mathcal{A}\to\mathbb{R} = (s,a)\mapsto g_a(s)$. Then $\|f-g\|_\infty<\delta$, so we can bound the covering number of $\mathcal{F}_0$ by

$$\left|\mathcal{N}(\delta,\mathcal{F}_l,\|\cdot\|_\infty)\right|\leqslant|\mathcal{N}_\delta|^{|\mathcal{A}|}$$

We now utilize a lemma found in [todo: ref to Anthony Bartlett 2009]

Lemma 3.29 (Covering Number of ReLU Networks)**.** Consider the family of ReLU networks

$$\mathcal{F}\left(L, \{d_j\}_{j=0}^{L+1}, s, V_{\max}\right)$$

where $\mathcal{F}$ is defined in definition 3.17. Let $D := \prod_{\ell=1}^{L+1}(d_\ell + 1))$. Then for any $\delta > 0$

$$\mathcal{N}\left(\delta, \mathcal{F}\left(L, \{d_j\}_{j=0}^{L+1}, s, V_{\max}\right), \|\cdot\|_\infty\right) \leqslant (2(L+1)D^2/\delta)^{s+1}$$

Proof. We refer to theorem 14.5 in [todo: ref to Anthony Bartlett, thm. 14.5]. □

With lemma 3.29 and n sufficiently large we can bound

$$\begin{aligned}
\log N_0 &= \log\left|\mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty)\right| \\
&\leqslant |\mathcal{A}| \cdot \log\left|\mathcal{N}_{1/n}\right| \\
&\leqslant |\mathcal{A}|\,(\widetilde{s}+1)\log(2(\widetilde{L}+1)\widetilde{D}^2 n) \\
&\leqslant |\mathcal{A}|\,(c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*}+1)2\log\left(2(c_{\widetilde{L}}\log(n)^{\xi^*}+1)\prod_{\ell=1}^{\widetilde{L}+1}(\widetilde{d}+1)\right) \\
&\leqslant 2|\mathcal{A}|\,(c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*}+1)\log\left(2(c_{\widetilde{L}}\log(n)^{\xi^*}+1)(6n^{\alpha^*}\log(n)^{\xi^*}+1)^{\widetilde{L}+1}\right) \\
&\leqslant 4|\mathcal{A}|\,c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*}(\widetilde{L}+1)\log\left(24 c_{\widetilde{L}}\log(n)^{\xi^*} n^{\alpha^*}\log(n)^{\xi^*}\right) \\
&\leqslant 8|\mathcal{A}|\,c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*} c_{\widetilde{L}}\log(n)^{\xi^*}(\alpha^*+2)\log(n) \\
&= 8 c_{\widetilde{s}} c_{\widetilde{L}}(\alpha^*+2) n^{\alpha^*}\log(n)^{1+2\xi^*}
\end{aligned} \tag{3.21}$$

Finally, combining eq. (3.8), eq. (3.9), eq. (3.20), eq. (3.21) and fiddling around with constants one obtains

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} \leqslant C_\varepsilon \frac{\phi_{\mu,\nu}\gamma}{(1-\gamma)^2} V_{\max}^2 n^{\max\{-2\alpha^*\kappa^*, (\alpha^*-1)/2\}}\log(n)^{1+2\xi^*} + \frac{4\gamma}{(1-\gamma)^2} R_{\max}\gamma^K$$

where

$$C_\varepsilon = 160 C_2^2 C_{\widetilde{s}} C_{\widetilde{L}}(\alpha^*+2) + 4C_N + 32$$

only depends on the constants in assumption 3 finishing the proof. □

Now for theorem 3.26.

Lemma 3.30. $TQ \geqslant T^\pi Q$ for any policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ and any action value function $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$.

Proof.

$$\begin{aligned}
(TQ)(s,a) &= \mathbb{E}\left(R(s,a) + \gamma \max_{a'} Q(S', a') \mid S' \sim P(\cdot \mid s, a)\right) \\
&\geqslant \mathbb{E}\left(R(s,a) + \gamma Q(S', A') \mid S' \sim P(\cdot \mid s, a), A' \sim \pi(\cdot \mid S')\right) \\
&= T^\pi Q(s,a)
\end{aligned}$$

□

Lemma 3.31. Let $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be an action-value function, $\tau_1, \dots, \tau_m$ be policies and $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be a probability measure. Then

$$\mathbb{E}_\mu[(P^{\tau_m} \dots P^{\tau_1})(f)] \leqslant \kappa(k - i + j; \mu, \nu)\|f\|_{2,\nu}$$

For any measure $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ which is absolutely continuous w.r.t. $(P^{\tau_m} \dots P^{\tau_1})(\mu)$. Here κ is the concentration coefficients defined in definition 3.24.

Proof. Recall that

$$\begin{aligned}\kappa(m; \mu, \nu) &:= \sup_{\pi_1,\dots,\pi_m} \left[\mathbb{E}_\nu \left| \frac{\mathrm{d}(P^{\pi_m} \dots P^{\pi_1}\mu)}{\mathrm{d}\nu} \right|^2 \right]^{1/2} \\ &= \sup_{\pi_1,\dots,\pi_m} \left\| \frac{\mathrm{d}(P^{\pi_m} \dots P^{\pi_1}\mu)}{\mathrm{d}\nu} \right\|_{2,\nu}\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}_\mu[(P^{\tau_m} \dots P^{\tau_1})(f)] &= \int (P^{\tau_m} \dots P^{\tau_1})(f)\, \mathrm{d}\mu && (3.22)\\ &= \int f\, \mathrm{d}(P^{\tau_m} \dots P^{\tau_1}\mu) && (3.23)\\ &= \int f \frac{\mathrm{d}(P^{\tau_m} \dots P^{\tau_1}\mu)}{\mathrm{d}\nu}\, \mathrm{d}\nu && (3.24)\\ &\leqslant \left\| \frac{\mathrm{d}(P^{\tau_m} \dots P^{\tau_1}\mu)}{\mathrm{d}\nu} \right\|_{2,\nu} \cdot \|f\|_{2,\nu} && (3.25)\\ &\leqslant \kappa(m, \mu, \nu)\|f\|_{2,\nu} && (3.26)\end{aligned}$$

Where eq. (3.24) is due to the Radon-Nikodym theorem and eq. (3.25) is Cauchy-Schwarz. □

Proof of theorem 3.26. First some things to keep in mind during the proof. Recall that $V_{\max} = R_{\max}/(1-\gamma)$ and that π_Q is the greedy policy w.r.t. Q. Denote

$$\pi_i = \pi_{\widetilde{Q}_i},\ Q_{i+1} = T\widetilde{Q}_i,\ \varrho_i = Q_i - \widetilde{Q}_i, \quad \text{for } i \in \{0, \dots, K+1\}$$

Note that for any policy π, P^π is linear and 1-contrative on $\mathcal{L}^\infty(\mathcal{S} \times \mathcal{A})$. Also

$$T^\pi Q^\pi = Q^\pi,\ TQ = T^{\pi_Q} Q,\ TQ^* = Q^* = Q^{\pi^*}$$

where π^* is greedy w.r.t. Q^*. If $f > f'$ for $f, f' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ then $P^\pi f \geqslant P^\pi f'$.

The proof consists of four steps.

Step 1 We start by relating $Q^* - Q^{\pi_K}$, the quantity of interest, to $Q^* - \widetilde{Q}_K$, which is more related to the output of the algorithm. Using lemma 3.30 we can make the upper bound

$$\begin{aligned}Q^* - Q^{\pi_K} &= T^{\pi^*}Q^* - T^{\pi_K}Q^{\pi_K} \\ &= T^{\pi^*}Q^* + (T^{\pi^*}\widetilde{Q}_K - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T\widetilde{Q}_K) - T^{\pi_K}Q^{\pi_K} \\ &= (T^{\pi^*}\widetilde{Q}_K - T\widetilde{Q}_K) + (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K}) \\ &\leqslant (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K}) \\ &= (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T^{\pi_K}\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K}) \\ &= \gamma P^{\pi^*}(Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K}(\widetilde{Q}_K - Q^{\pi_K}) \\ &= \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K}(Q^* - Q^{\pi_K}) && (3.27)\end{aligned}$$

This implies

$$(I - \gamma P^{\pi_K})(Q^* - Q^{\pi_K}) \leqslant \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}_K)$$

Since γP^{π_K} is γ-contractive, $U = (I - \gamma P^{\pi_K})^{-1}$ exists as a bounded operator on $\mathcal{L}^\infty(\mathcal{S} \times \mathcal{A})$ and equals

$$U = \sum_{i=0}^{\infty} \gamma^i (P^{\pi_K})^i$$

From this we also see that $f \geqslant f' \implies Uf \geqslant Uf'$ for any $f, f' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Therefore we can apply U on both sides of eq. (3.27) to obtain

$$Q^* - Q^{\pi_K} \leqslant \gamma U^{-1}(P^{\pi^*}(Q^* - \widetilde{Q}_K) - P^{\pi_K}(Q^* - \widetilde{Q}_K)) \tag{3.28}$$

Step 2 Using lemma 3.30 for any $i \in [K]$ we can get an upper bound

$$\begin{aligned} Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi^*}\widetilde{Q}_i - T^{\pi^*}\widetilde{Q}_i) \\ &= (Q^* - T^{\pi^*}\widetilde{Q}_i) + (T\widetilde{Q}_i - \widetilde{Q}_{i+1}) + (T^{\pi^*}\widetilde{Q}_i - T\widetilde{Q}_i) \\ &= (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_i) + \varrho_{i+1} + (T^{\pi^*}\widetilde{Q}_i - T\widetilde{Q}_i) \\ &\leqslant T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_i + \varrho_{i+1} \\ &= \gamma P^{\pi^*}(Q^* - \widetilde{Q}_i) + \varrho_{i+1} \end{aligned} \tag{3.29}$$

and a lower bound

$$\begin{aligned} Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi_i}Q^* - T^{\pi_i}Q^*) \\ &= (T^{\pi_i}Q^* - T^{\pi_i}\widetilde{Q}_i) + \varrho_{i+1} + (TQ^* - T^{\pi_i}Q^*) \\ &\geqslant T^{\pi_i}Q^* - T^{\pi_i}\widetilde{Q}_i + \varrho_{i+1} \\ &= \gamma P^{\pi_i}(Q^* - \widetilde{Q}_i) + \varrho_{i+1} \end{aligned} \tag{3.30}$$

Applying eq. (3.29) and eq. (3.30) iteratively we get

$$Q^* - \widetilde{Q}_K \leqslant \gamma^K (P^{\pi^*})^K (Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi^*})^{K-1-i} \varrho_{i+1} \tag{3.31}$$

and

$$Q^* - \widetilde{Q}_K \geqslant \gamma^K (P^{\pi_{K-1}} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi_{K-1}} \dots P^{\pi_{i+1}}) \varrho_{i+1} \tag{3.32}$$

Step 3 Combining eq. (3.31) and eq. (3.32) with eq. (3.28) we get

$$\begin{aligned} Q^* - Q^{\pi_K} \leqslant U^{-1} \Bigg(& \gamma^{K+1}((P^{\pi^*})^{K+1} - P^{\pi_K} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) \\ & + \sum_{i=0}^{K-1} \gamma^{K-i}((P^*)^{K-i} - P^{\pi_K} \dots P^{\pi_{i+1}}) \varrho_{i+1} \Bigg) \end{aligned} \tag{3.33}$$

For shorthand define constants

$$\alpha_i = \frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}} \quad \text{for } 0 \leqslant i \leqslant K-1 \text{ and } \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} \tag{3.34}$$

(note that $\sum_{i=0}^{K} \alpha_i = 1$) and operators

$$O_i = (1-\gamma)/2U^{-1}[(P^{\pi^*})^{K-i} + (P^{\pi_K} \dots P^{\pi_{i+1}})] \tag{3.35}$$

$$O_K = (1-\gamma)/2U^{-1}[(P^{\pi^*})^{K+1} + (P^{\pi_K} \dots P^{\pi_0})] \tag{3.36}$$

Then by eq. (3.33)

$$\left|Q^* - Q^{\pi_K}\right| \leqslant \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{i=0}^{K-1}\alpha_i O_i|\varrho_{i+1}| + \alpha_K O_K\left|Q^* - \widetilde{Q}_0\right|\right] \tag{3.37}$$

So by linearity of expectation

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} = \mathbb{E}_\mu\left|Q^* - Q^{\pi_K}\right| \tag{3.38}$$

$$\leqslant \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{i=0}^{K-1}\alpha_i \mathbb{E}_\mu(O_i|\varrho_{i+1}|) + \alpha_K \mathbb{E}_\mu(O_K\left|Q^* - \widetilde{Q}_0\right|)\right] \tag{3.39}$$

With the bound on rewards we (crudely) estimate

$$\mathbb{E}_\mu O_K\left|Q^* - \widetilde{Q}_0\right| \leqslant 2V_{\max} = 2R_{\max}/(1-\gamma) \tag{3.40}$$

The remaining difficulty lies in $\mathbb{E}_\mu(O_i|\varrho_{i+1}|)$.

Step 4 Using the sum expansion of U^{-1} we get

$$\mathbb{E}_\mu(O_i|\varrho_{i+1}|) \tag{3.41}$$

$$= \frac{1-\gamma}{2}\mathbb{E}_\mu\left(U^{-1}[(P^{\pi_K})^{K-i} + P^{\pi_K}\dots P^{\pi_{i+1}}]|\varrho_{i+1}|\right) \tag{3.42}$$

$$= \frac{1-\gamma}{2}\mathbb{E}_\mu\left(\sum_{j=0}^{\infty}[(P^{\pi_K})^j(P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1}P^{\pi_{K-1}}\dots P^{\pi_{i+1}}]|\varrho_{i+1}|\right) \tag{3.43}$$

$$= \frac{1-\gamma}{2}\sum_{j=0}^{\infty}\mathbb{E}_\mu\left([(P^{\pi_K})^j(P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1}P^{\pi_{K-1}}\dots P^{\pi_{i+1}}]|\varrho_{i+1}|\right) \tag{3.44}$$

Notice that there are $K-i+j$ P-operators on both terms in the sum. Therefore were can employ lemma 3.31 twice. Moreover define $\varepsilon_{\max} = \max_{i\in[K]}\|\varrho_i\|_{2,\nu}$. Then

$$\begin{aligned}\mathbb{E}_\mu(O_i|\varrho_{i+1}|) &\leqslant (1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu)\|\varrho_{i+1}\|_{2,\nu} \\ &\leqslant \varepsilon_{\max}(1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu)\end{aligned} \tag{3.45}$$

Using eq. (3.39), eq. (3.40) and eq. (3.45)

$$\begin{aligned}\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} \leqslant \frac{2\gamma(1-\gamma^{K+1})}{1-\gamma}&\left[\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\right]\varepsilon_{\max} \\ &+\frac{4\gamma(1-\gamma^{K+1})}{(1-\gamma)^3}\alpha_K R_{\max}\end{aligned} \tag{3.46}$$

Focusing on the first term on RHS of eq. (3.46), if we then we can take the norm out of the sum

as a constant. We are left with

$$\begin{aligned}
&\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\\
&=\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\frac{(1-\gamma)\gamma^{K-i+j-1}}{1-\gamma^{K+1}}\kappa(K-i+j;\mu,\nu)\\
&=\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{j=0}^{\infty}\sum_{i=0}^{K-1}\gamma^{K-i+j-1}\kappa(K-i+j;\mu,\nu)\\
&\leqslant\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{m=0}^{\infty}\gamma^{m-1}\cdot m\cdot\kappa(m;\mu,\nu)\\
&\leqslant\frac{1}{1-\gamma^{K+1}(1-\gamma)}\phi_{\mu,\nu}
\end{aligned}\tag{3.47}$$

Where the last inequality is due to assumption 3. Combining eq. (3.46) and eq. (3.47) we arrive at

$$\left\|Q^*-Q^{\pi_K}\right\|_{1,\mu}\leqslant\frac{2\gamma\cdot\phi_{\mu,\nu}}{(1-\gamma)^2}\cdot\varepsilon_{\max}+\frac{4\gamma^{K+1}}{(1-\gamma)^2}\cdot R_{\max}\tag{3.48}$$

□

Finally we show theorem 3.27.

Lemma 3.32 (Rotation invariance)**.** Let $(X_i)_{i=1}^n$ be independent, centered and sub-gaussian. Then $\sum_{i=1}^n X_i$ is centered and sub-gaussian with

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2\leqslant C\sum_{i=1}^n\|X_i\|_{\psi_2}^2$$

Proof. See [Vershynin 2010, p. 12]. □

Definition 3.33 (Sub-exponential norm)**.** For a random variable define

$$\|X\|_{\psi_1}=\sup_{p\geqslant 1}p^{-1}\|X\|_p$$

called the sub-exponential norm, said to 'exist' if finite. In that case X is said to be 'sub-exponential'.

Lemma 3.34 (Sub-gaussian squared is sub-exponential)**.** A random variable X is sub-gaussian if and only if X^2 is sub-exponential and

$$\|X\|_{\psi_2}^2\leqslant\left\|X^2\right\|_{\psi_1}\leqslant 2\|X\|_{\psi_2}^2$$

Proof. See [Vershynin 2010, p. 14] □

Proposition 3.35. Let v be a random vector in $\mathbb{R}^n$ then

$$\mathbb{E}\|v\|_1\leqslant\sqrt{n}\sqrt{\mathbb{E}\|v\|_2^2}$$

Proof. Denote v's coordinates $v = (v_1, \dots, v_n)$. Cauchy-Schwarz applied to some vector w and $(1, \dots, 1)$ yields

$$\|w\|_1 \leqslant \sqrt{n}\|w\|_2$$

Now let $w = (\mathbb{E}v_1, \dots, \mathbb{E}v_n)$. Then by linearity of expectation and Jensens inequality

$$\mathbb{E}\|v\| = \|w\| \leqslant \sqrt{n}\sqrt{\sum_{i=1}^{n}(\mathbb{E}v_i)^2} \leqslant \sqrt{n}\sqrt{\mathbb{E}\sum_{i=1}^{n} v_i^2} = \sqrt{n}\sqrt{\mathbb{E}\|v\|_2^2}$$

□

Theorem 3.36 (Bernstein's inequality)**.** Suppose $U_1, \dots, U_n$ are independent with $\mathbb{E}U_i = 0, |U_i| \leqslant M$ for all $i \in [n]$. Then for all $t > 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} U_i\right| \geqslant t\right) \leqslant \exp\left(\frac{-t^2}{2/3Mt + 2\sigma^2}\right)$$

where $\sigma^2 = \sum_{i=1}^{n} V(U_i)$.

Proof of theorem 3.27. First some introductory fixing of notation and variables. Fix a minimal δ-covering of $\mathcal{F}$ with centers $f_1, \dots, f_{N_\delta}$. Define

$$\widetilde{Q} := \operatorname*{argmin}_{f \in \mathcal{F}} \|f - TQ\|_\nu^2$$

$$k^* := \operatorname*{argmin}_{k \in [N_\delta]} \left\|f_k - \widehat{Q}\right\|_\infty$$

and $X_i := (S_i, A_i)$. Notice that $\widetilde{Q}$ differs from $\widehat{Q}$ in that $\widetilde{Q}$ approximates TQ w.r.t. $\|\cdot\|_\nu^2$ while $\widehat{Q}$ approximates $Y = (Y_1, \dots, Y_n)$ in mean squared error over $X = (X_1, \dots, X_n)$. We shall be loose about applying functions to vectors (of random variables) in the sense that they are applied entry-wise. We use $\|\cdot\|_p$ to denote the (finite dimensional) p-norm (p ommitted when $p = 2$). When talking about p-norms on the random variables we always specify the distribution (e.g. $\|\cdot\|_\nu$). When the sample (e.g. X) is clear from context we omit it writing $\|f\| = \|f(X)\|$.

Step 1 By definion (of $\widehat{Q}$) for all $f \in \mathcal{F}$ we have $\left\|\widehat{Q}(X) - Y\right\|^2 \leqslant \|f(X) - Y\|^2$, leading to

$$\|Y\|^2 + \left\|\widehat{Q}\right\|^2 - 2Y \cdot \widehat{Q} \leqslant \|Y\|^2 + \|f\|^2 - 2Y \cdot f \tag{3.49}$$

$$\iff \left\|\widehat{Q}\right\|^2 + \|TQ\|^2 - 2\widehat{Q} \cdot TQ \leqslant \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \widehat{Q} - 2Y \cdot f - 2\widehat{Q} \cdot TQ + 2f \cdot TQ \tag{3.50}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leqslant \|f - TQ\|^2 + 2(Y - TQ) \cdot (\widehat{Q} - f) \tag{3.51}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leqslant \|f - TQ\|^2 + 2\xi \cdot (\widehat{Q} - f) \tag{3.52}$$

Where $\xi_i := Y_i - TQ(X_i)$ and $\xi := (\xi_1, \dots, \xi_n)$. Let $\Sigma = (X_1, \dots, X_n)^{-1}(\mathbb{B}_n) \in \mathcal{H}$ be the σ-algebra generated by the samples. Now we proof a minor lemma

Proposition 3.37. $\mathbb{E}(\xi_i \mid \Sigma) = 0$ and thus $\mathbb{E}(\xi_i g(X_i)) = 0$ for any function $g : \mathbb{R} \to \mathbb{R}$.

Proof. Recall that $X_i = (S_i, A_i)$,

$$Y_i = R_i + \gamma \max_{a \in \mathcal{A}} Q(S_{i+1}, a)$$

where $S_{i+1} \sim P(X_i)$, $R_i \sim R(X_i)$ and

$$TQ(X_i) = \mathbb{E}_\Sigma R_i' + \gamma \mathbb{E}_\Sigma Q(S', \operatorname*{argmax}_{a \in \mathcal{A}} Q(S', a))$$

where $S' \sim P(X_i)$, $R_i' \sim R(X_i)$. Since S' and S_{i+1} are i.i.d.

$$\begin{aligned}
\mathbb{E}_\Sigma \xi_i &= \mathbb{E}_\Sigma \left(Y_i - TQ(X_i) \right) \\
&= \mathbb{E}_\Sigma R_i - \mathbb{E}_\Sigma R_i' + \gamma \left(\mathbb{E}_\Sigma \left(\max_{a \in \mathcal{A}} Q(S_{i+1}, a) \right) - \mathbb{E}_\Sigma \operatorname*{argmax}_{a \in \mathcal{A}} \left(Q(S', a) \right) \right) \\
&= 0
\end{aligned}$$

Therefore $\mathbb{E}(\xi_i \mid \Sigma) = 0$. □

By this lemma we can deduce

$$\mathbb{E}\left(\xi \cdot (\widehat{Q} - f) \right) = \mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ) \right) \tag{3.53}$$

To bound this we insert f_{k*} by the triangle inequality

$$\left| \mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ) \right) \right| \leqslant \left| \mathbb{E}\left(\xi \cdot (\widehat{Q} - f_{k*}) \right) \right| + \left| \mathbb{E}\left(\xi \cdot (f_{k*} - TQ) \right) \right| \tag{3.54}$$

We now bound these two terms. The first by Cauchy-Schwarz

$$\left| \mathbb{E}\xi \cdot (\widehat{Q} - f_{k*}) \right| \leqslant \mathbb{E}\left(\|\xi\| \left\| \widehat{Q} - f_{k*} \right\| \right) \leqslant \mathbb{E}(\|\xi\|) \sqrt{n} \delta \leqslant 2nV_{\max}\delta \tag{3.55}$$

where we have used that $\left\| \widehat{Q} - f_{k*} \right\|_\infty \leqslant \delta$ so

$$\left\| \widehat{Q} - f_{k*} \right\|^2 = \sum_{i=1}^n (\widehat{Q}(X_i) - f_{k*}(X_i))^2 \leqslant \sum_{i=1}^n \delta^2 = n\delta^2 \tag{3.56}$$

and that $|Y_i|, TQ(X_i) \leqslant V_{\max}$ so

$$\|\xi\|^2 = \sum_{i=1}^n (Y_i - TQ(X_i))^2 \leqslant \sum_{i=1}^n (2V_{\max})^2 = 4V_{\max}^2 n \tag{3.57}$$

To bound the second term in eq. (3.54) define

$$Z_j := \xi \cdot (f_j - TQ) \left\| f_j - TQ \right\|^{-1} \tag{3.58}$$

Note that since ξ_i are centered Z_j. For a sub-σ-algebra Σ define the *sub-gaussian* norm by

Definition 3.38 (Sub-gaussian norm)**.**

$$\|W\|_{\psi_2, \Sigma} := \sup_{p \geqslant 1} p^{-1/2} \left(\mathbb{E}_\Sigma |W|^p \right)^{1/p}$$

Because of proposition 3.37 $\xi_i(f_j(X_i) - TQ(X_i))$ is centered for any $i \in [n]$ and

$$\left\| \xi_i (f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma} \leqslant 2V_{\max} \left| f_j(X_i) - TQ(X_i) \right| \tag{3.59}$$

Therefore by lemma 3.32

$$\begin{aligned}
\left\| Z_j \right\|_{\psi_2, \Sigma}^2 &\leqslant \left\| f_j - TQ \right\|^{-2} \left\| \sum_{i=1}^n \xi_i (f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma}^2 && (3.60) \\
&\leqslant \left\| f_j - TQ \right\|^{-2} C_1 \sum_{i=1}^n \left\| \xi_i (f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma}^2 && (3.61) \\
&\leqslant \left\| f_j - TQ \right\|^{-2} C_1 \sum_{i=1}^n 4V_{\max} \left| f_j(X_i) - TQ(X_i) \right|^2 && (3.62) \\
&= 4V_{\max}^2 C_1 && (3.63)
\end{aligned}$$

Observe that $\|X\|_p \leqslant \sqrt{p}\ \sup_{p\geqslant 1}\|X\|_{\psi_2}$. Thus by [Vershynin 2010, p. 11 and Lemma 5.5]

$$\mathbb{E}\exp\left(cZ_j^2/\left\|Z_j\right\|_{\psi_2}^2\right) \leqslant e \tag{3.64}$$

so

$$\begin{aligned}
\mathbb{E}\max_{j\in N_\delta} Z_j^2 &= \frac{\max_{j\in[N_\delta]}\left\|Z_j\right\|_{\psi_2}^2}{c}\mathbb{E}\left(\max_{j\in[N_\delta]}\frac{cZ_j^2}{\max_{k\in[N_\delta]}\|Z_k\|_{\psi_2}}\right) && (3.65)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\mathbb{E}\left(\max_{j\in N_\delta}\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right) && (3.66)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(\mathbb{E}\max_{j\in N_\delta}\exp\left(\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right)\right) && (3.67)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(\sum_{j\in[N_\delta]}\mathbb{E}\exp\left(\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right)\right) && (3.68)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(eN_\delta\right) && (3.69)\\
&\leqslant C_2^2 V_{\max}^2\log(N_\delta) && (3.70)
\end{aligned}$$

Where $C_2 := \sqrt{8C_1/c}$. Now we can bound

$$\begin{aligned}
\mathbb{E}\left(\xi\cdot(f_{k*}-TQ)\right) &= \mathbb{E}\left(\|f_{k*}-TQ\||Z_{k*}|\right) && (3.71)\\
&\leqslant \mathbb{E}\left(\left(\left\|\widehat{Q}-TQ\right\|+\left\|\widehat{Q}-f_{k*}\right\|\right)|Z_{k*}|\right) && (3.72)\\
&\leqslant \mathbb{E}\left(\left(\left\|\widehat{Q}-TQ\right\|+n\delta\right)|Z_{k*}|\right) && (3.73)\\
&\leqslant \left(\mathbb{E}\left(\left\|\widehat{Q}-TQ\right\|+n\delta\right)^2\right)^{1/2}\left(\mathbb{E}Z_{k*}^2\right)^{1/2} && (3.74)\\
&\leqslant \mathbb{E}\left(\left\|\widehat{Q}-TQ\right\|+n\delta\right)\left(\mathbb{E}Z_{k*}^2\right)^{1/2} && (3.75)\\
&\leqslant \left(\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|_2^2}+n\delta\right)\left(\mathbb{E}Z_{k*}^2\right)^{1/2} && (3.76)\\
&\leqslant \left(\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|_2^2}+n\delta\right)C_2^2V_{\max}^2\log(N_\delta) && (3.77)
\end{aligned}$$

Where eq. (3.71) to eq. (3.72) is by the triangle inequality and eq. (3.75) to eq. (3.76) is proposition 3.35. Combining eq. (3.52), eq. (3.54), eq. (3.55) and eq. (3.77)

$$\begin{aligned}
\mathbb{E}\left\|\widehat{Q}-TQ\right\|^2 &\leqslant \mathbb{E}\|f-TQ\|^2 + 4nV_{\max}\delta + \left(\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|^2}+\sqrt{n}\delta\right)C_2V_{\max}\sqrt{\log(N_\delta)} && (3.78)\\
&= C_2V_{\max}\sqrt{n\log(N_\delta)}\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|^2} + nC_2^2\delta V_{\max}^2\log(N_\delta) + \mathbb{E}\|f-TQ\|^2 && (3.79)
\end{aligned}$$

Lemma 3.39. Let $a, b > 0, \kappa \in (0,1]$ then

$$a^2 \leqslant 2ab + c \implies a^2 \leqslant (1+\kappa)^2b^2/\kappa + (1+\kappa)c$$

Proof. $0 \leqslant (x-y)^2 = x^2 + y^2 - 2xy \implies 2xy \leqslant x^2 + y^2$ for any $x, y \in \mathbb{R}$ so

$$\begin{aligned} 2ab &= 2\sqrt{\frac{\kappa}{1+\kappa}}a\sqrt{\frac{1+\kappa}{\kappa}}b \\ &\leqslant \frac{\kappa}{1+\kappa}a^2 + \frac{1+\kappa}{\kappa}b^2 \end{aligned}$$

□

By lemma 3.39 applied to eq. (3.79)

$$\frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 \leqslant \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa)\left(\delta C_2^2 V_{\max}^2 \log(N_\delta) + \frac{1}{n}\mathbb{E}\|f - TQ\|^2\right) \tag{3.80}$$

We now take a closer look at the last term. Since f and TQ doesn't depend on the X_i's we have

$$\begin{aligned} \frac{1}{n}\mathbb{E}\|f - TQ\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(f(X_i) - TQ(X_i))^2 \\ &= \mathbb{E}(f(X_i) - TQ(X_i))^2 \\ &= \|f - TQ\|_\nu^2 \end{aligned}$$

Now since eq. (3.80) holds for any $f \in \mathcal{F}$ we can further say

$$\begin{aligned} \frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 &\leqslant \frac{(1+\kappa)^2}{\kappa}\delta C_2^2 V_{\max}^2 \log(N_\delta) \\ &\quad + (1+\kappa)\left(C_2^2 V_{\max}^2 \log(N_\delta) + \sup_{g\in\mathcal{G}}\inf_{f\in\mathcal{F}}\|f - g\|_\nu^2\right) \\ &\leqslant \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa)\left(C_2^2 V_{\max}^2 \log(N_\delta) + \omega(\mathcal{F})\right) \end{aligned} \tag{3.81}$$

where we take the supremum over $\mathcal{G}$ (recall $TQ \in \mathcal{G}$).

Step 2 Here we link up $\left\|\widehat{Q} - TQ\right\|_\sigma^2$ with $\mathbb{E}\frac{1}{n}\left\|\widehat{Q} - TQ\right\|^2$. First note that

$$\left|\left(\widehat{Q}(x) - TQ(x)\right)^2 - \left(f_{k*}(x) - TQ(x)\right)^2\right| = \left|\widehat{Q}(x) - f_{k*}(x)\right| \cdot \left|\widehat{Q}(x) + f_{k*}(x) - 2TQ(x)\right| \tag{3.82}$$

$$\leqslant 4V_{\max}\delta \tag{3.83}$$

Using this twice we can say

$$(\widehat{Q}(\widehat{X}_i) - TQ(\widehat{X}_i)^2 \tag{3.84}$$

$$\leqslant (\widehat{Q}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 + (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 \tag{3.85}$$

$$\begin{aligned} &\leqslant (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 + (\widehat{Q}(X_i) - TQ(X_i))^2 - (\widehat{Q}(X_i) - TQ(X_i))^2 \\ &\quad + (f_{k*}(X_i) - TQ(X_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 + 4V_{\max}\delta \end{aligned} \tag{3.86}$$

$$\leqslant (\widehat{Q}(X_i) - TQ(X_i))^2 + (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 + 8V_{\max}\delta \tag{3.87}$$

Thus we get

$$\left\|\widehat{Q} - TQ\right\|_\sigma^2 \tag{3.88}$$

$$= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(\widehat{Q}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 \tag{3.89}$$

$$\leqslant \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\left((\widehat{Q}(X_i) - TQ(X_i))^2 + (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2\right) + 8V_{\max}\delta \tag{3.90}$$

$$= \frac{1}{n}\left\|\widehat{Q} - TQ\right\|^2 + \frac{1}{n}\sum_{i=1}^{n} h_{k*}(X_i, \widetilde{X}_i) + 8V_{\max}\delta \tag{3.91}$$

Where we define

$$h_j(x,y) := \big(f_j(y) - TQ(y)\big)^2 - \big(f_j(x) - TQ(x)\big)^2 \tag{3.92}$$

For any $j \in [N_\delta]$. Define $\Upsilon = 2V_{\max}$ and

$$T := \max_{j\in[N_\delta]}\left|\sum_{i=1}^{n} h_j(X_i, \widetilde{X}_i)/\Upsilon\right| \tag{3.93}$$

Then we can bound the middle term in eq. (3.91)

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} h_{k*}(X_i, \widetilde{X}_i)\right) \leqslant \Upsilon/n\mathbb{E}\max_{j\in[N_\delta]}\left(\left|\sum_{i=1}^{n} h_j(X_i, \widetilde{X}_i)/\Upsilon\right|\right) \tag{3.94}$$

$$\leqslant \Upsilon/n\mathbb{E}T \tag{3.95}$$

We want to use Bernsteins inequality (theorem 3.36) with $U_i = h_j(X_i, \widetilde{X}_i)$. Therefore notice that $|h_j| \leqslant \Upsilon^2$ and

$$\mathrm{Var}h_j(X_i, \widetilde{X}_i) = 2\mathrm{Var}\left(f_j(X_i) - TQ(X_i)\right)^2 \tag{3.96}$$

$$\leqslant 2\mathbb{E}\left(f_j(X_i) - TQ(X_i)\right)^4 \tag{3.97}$$

$$\leqslant 2\Upsilon^4 \tag{3.98}$$

so by union bounding for any $u < 6n\Upsilon$ we have

$$\mathbb{E}T = \int_0^\infty \mathbb{P}(T \geqslant t) \tag{3.99}$$

$$\leqslant u + \int_u^\infty \mathbb{P}(T \geqslant t)\,\mathrm{d}t \tag{3.100}$$

$$\leqslant u + \int_u^\infty 2N_\delta \exp\left(\frac{-t^2}{2\Upsilon t/3 + 4n\Upsilon^2}\right)\,\mathrm{d}t \tag{3.101}$$

$$\leqslant u + 2N_\delta \int_u^\infty \exp\left(\frac{-t^2}{2\Upsilon^2(t/(3\Upsilon) + 2n)}\right)\,\mathrm{d}t \tag{3.102}$$

$$\leqslant u + 2N_\delta\left(\int_u^{6n\Upsilon} \exp\left(\frac{-t^2}{8n\Upsilon^2}\right)\,\mathrm{d}t + \int_{6n\Upsilon}^\infty \exp\left(\frac{-t}{4/3\Upsilon}\right)\,\mathrm{d}t\right) \tag{3.103}$$

$$\leqslant u + 2N_\delta\left(\frac{8n\Upsilon}{2u}\exp\left(\frac{-u^2}{8n\Upsilon}\right) + \frac{4\Upsilon}{3}\exp\left(\frac{-24n\Upsilon}{3\Upsilon}\right)\right) \tag{3.104}$$

where we use lemma A.1 from eq. (3.103) to eq. (3.104). Now set $u = \Upsilon\sqrt{8n\log N_\delta}$. We continuing from eq. (3.104) getting

$$\cdots = \Upsilon\sqrt{8n\log N_\delta} + \frac{\Upsilon^2 8nN_\delta}{\Upsilon\sqrt{8n\log N_\delta}}\exp(-\log N_\delta) + 8/3N_\delta\Upsilon\exp(-9/2n) \tag{3.105}$$

$$= \Upsilon 2\sqrt{2n}\left(\log N_\delta + \frac{1}{\log N_\delta}\right) + 8/3N_\delta e^{-9/2n} \tag{3.106}$$

$$\leqslant 4\sqrt{2}\Upsilon\sqrt{n\log N_\delta} + 8/3\Upsilon \tag{3.107}$$

Now inserting eq. (3.107) and eq. (3.81) into eq. (3.91) we get

$$\left\|\widehat{Q} - TQ\right\|_\nu^2 \leqslant \frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1}+\delta) \tag{3.108}$$

$$\leqslant \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2V_{\max}^2\log(N_\delta) + (1+\kappa)\left(\delta C_2^2V_{\max}^2\log(N_\delta) + \omega(\mathcal{F})\right) + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1}+\delta) \tag{3.109}$$

finishing the proof. □

Chapter 4

Comparison and conclusion

4.1 Comparison

4.2 Conclusion

In this paper we have build up the theory behind Q-learning, covering decision models, optimality of policies, value functions and their iteration methods. This gave an introduction to Q-learning and a general framework from which to understand and compare results within the field. We then turned to model-free algorithms and presented convergence results for such in a variety of settings with state space being both finite and infinite and dynamics being allowed to depend on history or not. Finally we presented and proved convergence of the fitted Q-iteration algorithm as obtained in [5]. All together this paints a picture of what Q-learning is, how it was developed, which topics it is related to, what its challenges are and what it is possible to say theoretically about its convergence to optimaliy at present. Theoretically you could say that Q-learning is solved in many situations, since, as we have established, there is convergence guarrantees for broad classes of problems. However as to how these convergence results relate to practical aspects of Q-learning we can still say little and as to the succes of the DQN of [13] we are not much further in understanding. The major reason is that the computational aspects are so important to their succes, and this part is mostly ignored in the results we have covered. Even though we establish results of the related FQI algorithm in [5], it is unclear if it captures the critical aspects of DQN, such as experience replay. In [5] convergence of FQI is guaranteed given corresponding increases in iterations, batch size and function space complexity. It is hard to interpret exactly how large these increases must be or whether it is practical.

4.3 Further directions

The litterature on Q-learning algorithms and relating topics such as function approximation, dynamic programming and artificial neural networks is vast, and only very little made it into this thesis. An obvious direction to go is to review more of the most recent results in order to give a more complete picture of the field.

Relation between FQI and DQN

Find a way to prove or disprove the conjecture in [5] that results about the convergence of FQI can have implications for the DQN algorithm.

Suboptimality of policies

This is relating to decision processes and value functions. Through out the paper we discuss a wide array of approximations of Q^*. The default strategy is then to accept some close-enough approximation $\widetilde{Q}$ and then pick the greedy policy $\widetilde{\pi}$ with respect to $\widetilde{Q}$. We then measure our deviation from optimality in terms of the distance $\left\|Q^* - \widetilde{Q}\right\|_\infty$. However in most cases we do not estimate the deviation of $Q_{\widetilde{\pi}}$ from Q^* which from a theorical point of view should be a better measure of the sub-optimality of $\widetilde{\pi}$ compared to π^*. Some sources like [5] succeed in bounding $\|Q^* - Q_{\widetilde{\pi}}\|_\infty$, while many others make do with a bound on $\left\|Q^* - \widetilde{Q}\right\|_\infty$. To this end it could be interesting to establish relations between $\|Q^* - Q_{\widetilde{\pi}}\|_\infty$ and $\left\|Q^* - \widetilde{Q}\right\|_\infty$.

Bernstein polynomials vs. orthogonal projection

A Bernstein polynomial B_f approximating a function f are constructed by evaluating the functions at a finite number of points (see definition 2.67). Since we in this setting are concerned with approximation in the 2-norm, another approach would be to simply take the orthogonal projection of TQ onto the span of polynomials of degree less than n. One should keep in mind that this requires integration of $|TQ(\cdot, a)f_i|$ for every basis polynomial f_i, which is potentially hard to compute. On the other hand, as the orthogonal projection is distance minimizing, it should provide the best approximation with polynomials. The relation between the performances of the Berstein polynomial and the orthogonal projection, both in terms of accuracy and computational complexity, could be interesting analyse.

4.4 Notes on references

The proofs on basic measure theory are inspired by ones found in [14, Rønn-Nielsen and Hansen (2014)] and [8, Kallenberg (2002)]. A good survey on results on optimal policy existence in the special case of Markov decision processes can be found in [6, Feinberg (2012)], however proofs in this source is either missing or sketched (as one must expect in a survey).

4.5 Credits

I would like to thank PhD-student Jonas Rysgaard Jensen for helping me out with a proof on the Ionescu Tulcea kernel, my cousin Rune Harder Bak for reading the mess I've made, my dormmates at the P.C.Petersens dorm for good company and very necessary recreational breaks from writing, my aunt Susanne for letting me stay at her house during the covid-19 and my the rest of my family for love and support.

Chapter A

Appendices

A.1 Lemmas for Fan et al.

Lemma A.1. For $x > 0$.

$$\int_x^\infty e^{-t^2/2}\,\mathrm{d}t \leqslant \frac{1}{x}e^{-x^2/2}$$

Proof. Observe that for $t \geqslant x > 0$ we have $1 \leqslant t/x$ so

$$\begin{aligned}\int_x^\infty e^{-t^2/2}\,\mathrm{d}t &\leqslant \int_x^\infty \frac{t}{x}e^{-t^2/2}\,\mathrm{d}t \\ &\leqslant \frac{1}{x}e^{-x^2/2}\end{aligned}$$

□

A.2 Other notes

Definition A.2 (Lipschitz continuity)**.** Let $(\mathcal{X}, d_\mathcal{X})$, $(\mathcal{Y}, d_\mathcal{Y})$ be metric spaces. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to **Lipschitz** with constant $L > 0$ if

$$d_\mathcal{Y}(f(x), f(y)) \leqslant L d_\mathcal{X}(x, y)$$

Definition A.3 (Almost sure uniform convergence of random processes)**.** A sequence of random processes $X_n : \mathcal{X} \times \Omega \to \mathbb{R}$ is said to converge **almost surely uniformly** to $X : \mathcal{X} \times \Omega \to \mathbb{R}$ if and only if

$$\mathbb{P}(\sup_{x\in\mathcal{X}}|X_n(x) - X(x)| \to 0) = 1$$

Definition A.4 (Uniform convergence in probability of random processes)**.** A sequence of random processes $X_n : \mathcal{X} \times \Omega \to \mathbb{R}$ is said to converge **uniformly in probability** to $X : \mathcal{X} \times \Omega \to \mathbb{R}$ if and only if

$$\sup_{x\in\mathcal{X}}|X_n(x) - X(x)| \xrightarrow{P} 0$$

Proposition A.5. $id_{\mathcal{P}(X)} = \mu \mapsto \kappa \circ \mu$ where $\kappa(\cdot \mid x) = \delta_x(\cdot)$. Thus κ can be seen as an identity mapping on $\mathcal{P}(X)$.

Proof.

$$\kappa\mu A = \int \delta_x(A)\, \mathrm{d}\mu(x) = \mu A$$

□

Theorem A.6 (Banach fixed point theorem)**.** Let $(\mathcal{X}, d)$ be a complete metric space and $T : \mathcal{X} \to \mathcal{X}$ be a contraction, i.e. $d(Tx, Ty) < \gamma d(x, y)$ for some $0 < \gamma < 1$ and all $x, y \in \mathcal{X}$. Then T has a unique fixed point x^* and for every $x \in \mathcal{X}$ it holds that $T^k x \to x^*$ as $k \to \infty$, with rate $d(T^k x, x^*) < \gamma^k d(x, x^*)$.

Definition A.7 (Dynkin class)**.** Let D be a pavement of X, that is a collection of subsets of X. D is called a **Dynkin class** if

1. $X \in D$,
2. If $A, B \in D$ and $A \subseteq B$ then $B \backslash A \in D$,
3. If $A_1, A_2, \cdots \in D$ with $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$ then $\bigcup_{n=1}^{\infty} A_n \in D$.

Theorem A.8 (Dynkins π-λ theorem)**.** Let P be a pavement of X which is stable under finite intersections (such are called π-systems) and D a Dynkin class (see definition A.7). If $P \subseteq D$ then $\sigma(P) \subseteq D$ where $\sigma(P)$ is the smallest σ-algebra containing P.

A.3 Disambiguation

- $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, $\overline{\underline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$.
- $\mathrm{id}_X := x \mapsto x$ the identity function on X.
- $[\phi] := \begin{cases} 1 & \phi \\ 0 & \neg\phi \end{cases}$: 0-1 indicator for logical formulas.
- $[q] := \{1, \ldots, q\}$ for $q \in \mathbb{N}$.
- $1_A(a) := [a \in A]$: the indicator function.
- $C_{\mathbb{K}}(X) := \{f : X \to \mathbb{K} \mid f \text{ continuous}\}$, for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. $C(X) = C_{\mathbb{R}}(X)$
- $C_b(X) := \{f \in C(X) \mid \exists K : \|f\|_\infty < K\}$. I.e. the space of bounded continuous functions.
- ANN: abbreviation for artificial neural network see definition 2.63.
- δ_a: Dirac-measure of point a, i.e. $\delta_a(A) = [a \in A] = 1_A(a)$.
- $(\Omega, \mathcal{F}, \mathbb{P})$: background probability space, that is source space for random variables.
- $\mathbb{B}_n$: the n-dimensional Borel σ-algebra.
- λ^n: the n-dimension Lebesgue measure.
- $\mathcal{P}(A)$: the set of all probability measures on a measurable space (A, Σ_A).
- 2^A: the powerset of A.
- $\mathcal{M}(A, B)$: set of Σ_A-Σ_B measurable functions.
- $\mathcal{L}_p$: set of functions f with $\|f\|_p < \infty$ $p \in [1, \infty]$.
- $\mathbb{E}, \mathbb{E}_\mu$: expectation, that is integration w.r.t. the measure $\mathbb{P}$ or μ respectively.
- Var: variance operator.

Bibliography

[1] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 2007. ISBN 1886529035.

[2] Tianping Chen, Hong Chen, and Reuy-wen Liu. A constructive proof and an extension of cybenko's approximation theorem. 03 1990. doi: 10.1007/978-1-4612-2856-1.

[3] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

[4] Adithya M. Devraj and Sean P. Meyn. Fastest convergence for q-learning. *CoRR*, abs/1707.03770, 2017. URL `http://arxiv.org/abs/1707.03770`.

[5] Jianqing Fan, Zhuoran Yang, Yuchen Xie, and Zhaoran Wang. A theoretical analysis of deep q-learning. *CoRR*, abs/1901.00137, 2020+. URL `http://arxiv.org/abs/1901.00137`.

[6] Eugene Feinberg. Total expected discounted reward mdps: Existence of optimal policies. 05 2012.

[7] Tommi Jaakkola, Michael Jordan, and Satinder Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201, 11 1994. doi: 10.1162/neco.1994.6.6.1185.

[8] Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002. ISBN 0-387-95313-2. doi: 10.1007/978-1-4757-4015-8. URL `http://dx.doi.org/10.1007/978-1-4757-4015-8`.

[9] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in Neural Information Processing*, 11, 04 1999.

[10] F. William Lawvere. The category of probabilistic mappings. 1962.

[11] Sultan Javed Majeed and Marcus Hutter. On q-learning convergence for non-markov decision processes. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2546–2552. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/353. URL `https://doi.org/10.24963/ijcai.2018/353`.

[12] F. S. Melo and M. I. Ribeiro. Convergence of q-learning with linear function approximation. In *2007 European Control Conference (ECC)*, pages 2671–2678, 2007.

[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

[14] Anders Rønn-Nielsen and Ernst Hansen. *Conditioning and Markov properties*. 2014. ISBN 978-87-7078-980-6.

[15] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *ArXiv*, abs/1708.06633, 2017. URL https://arxiv.org/abs/1708.06633v4.

[16] Manfred Schäl. On dynamic programming: Compactness of the space of policies. *Stochastic Processes and their Applications*, 3(4):345 – 364, 1975. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(75)90031-9. URL http://www.sciencedirect.com/science/article/pii/0304414975900319.

[17] Csaba Szepesvári. The asymptotic convergence-rate of q-learning. 01 1997.

[18] Christopher Watkins. Learning from delayed rewards. 01 1989.

[19] Christopher Watkins and Peter Dayan. Technical note: Q-learning. *Machine Learning*, 8: 279–292, 05 1992. doi: 10.1007/BF00992698.