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

June 4, 2020

Abstract

This thesis mainly focuses on 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. Decision processes are formalized as a kind of stochastic processes and can model diverse problems, from playing games like chess or poker to control problems such as balancing a pole and more. We present the background theory for these algorithms, including basic theory on stochastic processes, existence of optimal polices and dynamic programming in the form of value iteration. Then Q-learning is analysed in a variety of settings, establishing existence of optimal Q-functions and policies and proving convergence bounds of various Q-learning algorithms. Finally we present and prove a convergence bound of the deep fitted Q-iteration algorithm considered in the preprint [7, 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 [22] to a result obtained in the preprint [7, Fan et al. (2020+)]. The introduction includes fundamental theory of the underlying field of *dynamic programming* and *reinforcement learning* (RL) and related topics such as *value iteration*.

1.1 Motivation

The topic was inspired by the performance of the algorithms implemented by [16, Mnih et al. (2015)]. In [16] 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 [16] is based on a Q-learning algorithm called the *deep Q-network* (DQN) algorithm. The success of DQN learning to play Atari was only the beginning. In the past 5 years variations of DQN combined with other techniques, has solved highly challenging and before unsolved problems, such as beating the best human players at the ancient board game *go*, and the previously best algorithms for playing chess, with an approach quite novel to ideas previously employed to such environments. Deep Q-learning is now a very active field of research and is being applied in numerous fields, including robotics.

Despite the great empirical success, it seems[1] that the theoretical understanding is lacking. The purpose of this thesis is to investigate what has been proven about the convergence of algorithms similar to the DQN algorithm and what mathematical theory is relevant to establish such proofs.

In the course of reading [7, Fan et al. (2020+)] which claims to establish theoretical justification for the convergence of DQN to *the optimal Q-function*, it became clear that the background theory of dynamic programming was essential in order to understand the results of [7] and related papers and be able to compare such results. Even questions as to in which settings optimal policies exist turns out to be a non-trivial question. Therefore in the end this thesis is partly about presenting the results of [7] and similar papers. And partly to build the background theory necessary to understand and compare these results.

This provides an introduction to the field of RL and Q-learning and sheds light on the original question of what can be said about convergence of Q-learning algorithms. Finally we will discuss some of the many questions that still remain.

[1] This is for example the opinion in [7]

1.2 How to read this thesis

After this introductory chapter, in chapter 2 we will become more rigorous and present the background theory of decision processes and dynamic programming, which will give language and notation necessary for describing RL and Q-learning mathematically. The first part for chapter 2 (sections 2.1 and 2.2) is mainly based on two sources, namely [19, Schäl (1975)] and [3, Bertsekas and Shreve (2007)]. [19] establishes general results of the existence and properties of optimal policies in a general history dependent (non-markovian) decision processes (HDPs). The more recent source [3] focus on the more standard setting of *Markov decision processes* (MDPs), but in a way that is somewhat different from the present standard which is used in [7]. In this thesis we use a combination of [19] and [3] to build a framework which is similar to that of [3], but many proofs are changed and relies instead on results of [19]. This way of building the theory is original to this thesis, and is done with the purpose of better suiting the later discussions of [7] in particular. Section 2.1 og 2.2 will be used in all subsequent chapters to guarantee the existence of optimal policies, value functions and well definedness of various operators.

The purpose of section 2.3 is to show how if one stays in the model-dependent setting instead of 'going model-free' as is traditional in RL, one may easily establish results on convergence of value-iteration with Q-functions. We do this drawing on two results from approximation theory, namely the universal approximation theorem for artificial neural networks, and the approximation properties of Bernstein polynomials.

Proceeding to chapter 3, the sections 3.1 and 3.2 is a survey a series of papers on convergence bounds for (model-free) RL algorithms. These two sections are based on the articles [23, Watkins and Dayan (1992)], [10, Jaakkola et al. (1994)], [20, Szepesvári (1997)], [14, Majeed and Hutter (2018)] and [15, Melo and Ribeiro (2007)]. In this part of chapter 3, all proofs are omitted and only the results are presented.

In section 3.3 the preprint [7, Fan et al. (2020+)] is introduced and its results are proved Also some minor mistakes are corrected leading to slightly different bounds. The mistakes will be pointed out in this section. In correspondance with the authors it was confirmed that the mistakes was indeed to be corrected in the next version.

In the comparison and conclusion we briefly compare results we have covered in the thesis and discuss to what degree we are able to answer original questions on the theoretical understanding of the success of Q-learning. Finally we discuss which in directions one could investigate further extending what was covered in this thesis.

1.3 What is 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. To understand RL, we will now briefly describe its roots in dynamic programming.

In Reinforcement Learning, as in dynamic programming, 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 (or dynamics) formalized as probability kernels $P_1, R_1, P_2, \dots$ specifying which states and rewards are likely to follow after some action is chosen. One can then specify rules π, called a *policy*, for how the agent should choose actions in every situation in the environment. Given an environment and a policy one obtains a 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 accumulated rewards called the *policy evaluations* V_π of the policy. The goal of reinforcement learning is to find an optimal policy $\pi*$, maximizing the value function.

V_π is viewed as function that evaluates for each *starting state* $s \in \mathcal{S}_1$ the expected accumulated 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 class of environments which are called Markov decision processes (MDPs). In an MDP the same state space $\mathcal{S}$, action space $\mathcal{A}$ and rules P, R are used 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 the maximum average values (according to V_1). Such policies are called *greedy policies.* We can then evaluate the policy π_1 yielding a new value function $V_2 = V_{\pi_1}$. The process of evaluating policies and picking greedy policies is formalised by so called *T-operators* T_π, T. One of these (T) is called the *Bellman optimality operator* and combines policy evaluation and greedy choices. This process of applying the T operators and picking greedy policies 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. 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 optimality operator: $TV^* = V^*$ This is called the *Bellman optimality equation* and is central to all problems in dynamic programming.

We have now described the roots of RL in dynamic programming. However RL usually refers to algorithms that are not merely value iterations, but instead work without directly using the transition and reward dynamics, and instead estimate value functions based only on sampling from the environment. Such algorithms are called *model-free* as opposed to *model-based* algorithms, that employ knowledge about the environment such as its transition distributions directly (such as pure dynamic programming). Thus most of the algorithms (e.g. DQN) we will discuss are model-free however we will also include a section on how *model-dependent* approximative Q-learning may work. Whether such algorithms lie in the field of RL can be debated.

1.4 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. This is especially a problem for model-free algorithms. To get around this problem *Q-functions* were 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. Also it turns out that Q-functions is 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 the *deep Q-network* (DQN) algorithm of [16] is based. Many direct approaching such as FQI and DQN can be seen as stochastic approximations of the Bellman optimality equation.

Q-learning is the category of algorithms that iteratively updates Q-functions in the attempt to improve the derived policy. *Deep* Q-learning is then the subcategory of algorithms which uses deep neural networks as approximators for the Q-functions. We will see in this thesis how Q-functions are used to find optimal policies (strategies) for decision processes and how they work as the underlying *knowledge* that drives the decisions of the agent. We will use a wide array of function classes in the attempt to approximate ideal Q-functions such as the policy evaluations and optimal Q-functions. All this will be made precise in chapter 2. Before starting on this we include a brief introduction to the basic concept and notation we are going to use throughout the thesis.

1.5 Basic concepts and notation

The real numbers $\mathbb{R}$ is endowed with the standard ordering with giving rise to the standard order topolog (definition A.2). This in turn give rise to the standard Borel σ-algebra (definition A.5) $\mathbb{B} = \sigma(\mathcal{O})$ generated by the open sets $\mathcal{O}$ of the standard topology on $\mathbb{R}$.

When considering a measurable space $\mathcal{X}$ we always denote its σ-algebra $\Sigma_{\mathcal{X}}$ when not ambiguous. We always consider the cartesian product of measurable spaces with the product σ-algebra (definition A.6) unless otherwise specified. We denote the set of measurable functions (definition A.11) $\mathcal{X} \to \mathcal{Y}$ between two measurable spaces by $\mathcal{M}(\Sigma_{\mathcal{X}}, \Sigma_{\mathcal{Y}})$ or $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ when the σ-algebras are not ambiguous or simply $\mathcal{M}(\mathcal{X})$ when $\mathcal{Y} = \mathbb{R}$.

The set of probability measures on $\mathcal{X}$ is denoted $\mathcal{P}(\Sigma_{\mathcal{X}})$ or $\mathcal{P}(\mathcal{X})$ when $\Sigma_{\mathcal{X}}$ is implicit (not to be confused with the powerset of $\mathcal{X}$ which we denote $2^{\mathcal{X}}$).

An $\mathcal{X}$-valued random variable $X : \Omega \to \mathcal{X}$ is a measurable function from *the background probability space* measure space $(\Omega, \Sigma_\Omega, \mathbb{P})$ into some measurable space $\mathcal{X}$. By abstract change of variable the distribution of the random variable X is the image probability measure $\mu = X(\mathbb{P})$ and we write $X \sim \mu$.

When talking about functions $f_1, f_2, \dots : \mathcal{X} \to \mathbb{R}$ limits are always understood pointwise, unless otherwise stated, meaning that $f_n \to f$ is to be read as $\forall x \in \mathcal{X} : f_n(x) \to f(x)$. The same goes for logical operators, e.g. $f > 0$ is to be understood as $f(x) > 0$ for all $x \in \mathcal{X}$.

Chapter 2

Decision models and value functions

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

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

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

is an **$\mathcal{X}$-probability kernel** on $\mathcal{Y}$ provided

1. $B \mapsto \kappa(B \mid x) \in \mathcal{P}(\mathcal{Y})$ that is $\kappa(\cdot \mid x)$ is a probability measure for any $x \in \mathcal{X}$.

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

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

Remark 2.2. Note that probability kernel κ can also be viewed as a mapping $\kappa : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$.

Probability kernels are easily obtained by integration over suitable measurable functions.

Example 2.3. If $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is a positive measurable function with the property that

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

for some measure μ on $\mathcal{Y}$ then $\kappa(B \mid x) = \int_B f(x,y) \, \mathrm{d}\mu(y)$ defines a $\mathcal{X}$-probability kernel on $\mathcal{Y}$. This follows by basic measure theory and we omit these details.

A handy property of kernels is

Proposition 2.4. Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a probability kernel and $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be measurable satisfying that $f(x, \cdot)$ is $\kappa(\cdot \mid x)$-integrable for every $x \in \mathcal{X}$. Then the map

$$x \mapsto \int f(x, \cdot) \, \mathrm{d}\kappa(\cdot \mid x)$$

is measurable into $(\mathbb{R}, \mathbb{B})$.

Proof. This is a matter of going through the standard construction of the integral, noting that indicator functions 1_A on a measurable set $A \in \Sigma_{\mathcal{X}}$ are measurable by definition 2.1.2 since κ is a kernel. Then extend by sums and limits. □

We can now state the definition of a decision process

Definition 2.5 (History dependent decision process)**.** A (countable) **history dependent decision process** (HDP) is determined by

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

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

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

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

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

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. $\mathfrak{A}_n(h_n) \subseteq \mathcal{A}_n$ 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 2.6 (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, and some distribution $\mu \in \mathcal{P}(\mathcal{S}_1)$ of the *starting state* $S_1 \sim \mu$ and some policy π intuitively we should be able to obtain a history by sampling

- an action $A_1 \in \mathfrak{A}_1(S_1)$ from $\pi_1(\cdot \mid H_1)$
- a state $S_2 \in \mathcal{S}_2$ from $P(\cdot \mid S_1, A_1)$,
- an action $A_2 \in \mathfrak{A}_2(S_1, A_1, S_2)$ from $\pi_2(\cdot \mid S_1, A_1, S_2)$
- and so on.

where $S_1, A_1, S_2, A_2, \dots$ And in fact we will now see that it is possible to use the transition and reward kernels to obtain a measure on $\mathcal{H}_\infty$. For this we need some additional measure theory on probability kernels.

Theorem 2.7 (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 $\lambda \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that

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

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

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

Immedially we see that $\lambda(\mathcal{X} \times \mathcal{Y}) = 1$. Let $G_1, G_2, \dots \in \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}}$ be mutually disjoint. Then $G_1^x, G_2^x, \dots$ are mutually disjoint aswell. So by monotone convergence

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

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

For an idea how to actually compute integrals over kernel derived measures we here include

Theorem 2.9 (Extended Tonelli and Fubini)**.** Let $\mu \in \mathcal{P}(\mathcal{X})$, $f \in \mathcal{M}(\Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}}, \mathbb{B})$ be a measurable function and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a probability kernel. Then

$$\int |f| \, \mathrm{d}\kappa\mu = \int \int |f| \, \mathrm{d}\kappa(\cdot \mid x) \, \mathrm{d}\mu(x)$$

Furthermore if this is finite, i.e. $f \in \mathcal{L}_1(\kappa(\cdot, \mu))$ then $A_0 := \left\{x \in \mathcal{X} \,\middle|\, \int f \, \mathrm{d}\kappa(\cdot \mid x) < \infty\right\} \in \Sigma_{\mathcal{X}}$ with $\mu(A_0) = 1$,

$$x \mapsto \begin{cases} \int f \, \mathrm{d}\kappa(\cdot \mid x) & x \in A_0 \\ 0 & x \notin A_0 \end{cases}$$

is $\Sigma_{\mathcal{X}}$-$\mathbb{B}$ measurable and

$$\int f \, \mathrm{d}\kappa\mu = \int_{A_0} \int f \, \mathrm{d}\kappa(\cdot \mid x) \, \mathrm{d}\mu(x)$$

Proof. We refer to [17] thm. 1.3.2 and 1.3.3. □

Proposition 2.10 (Composition of kernels)**.** Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\phi : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ be probability kernels. Then there exists a unique probability kernel $\phi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ satisfying

$$\phi\kappa(B \times C \mid x) = \int 1_B(y) \phi(C \mid x, y) \, \mathrm{d}\kappa(y \mid x), \quad B \in \Sigma_{\mathcal{Y}},\ C \in \Sigma_{\mathcal{Z}}$$

called the **composition** of ϕ and κ. The composition is associative, that is if $\psi : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \rightsquigarrow \mathcal{W}$ is another probability kernel, then $(\psi\phi)\kappa = \psi(\phi\kappa)$. Associativity extends to measures, that is if $\mu \in \mathcal{P}(\mathcal{X})$ is a probability measure then $\phi(\kappa\mu) = (\phi\kappa)\mu$. Furthermore if $f : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is a measurable function such that $f(x, \cdot, \cdot)$ is $\phi\kappa(\cdot \mid x)$-integrable then

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

Proof. See section A.2. □

Remark 2.11. A kernel $\varphi : \mathcal{Y} \rightsquigarrow \mathcal{Z}$ can also be considered a kernel $\varphi' : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ by defining $\varphi'(\cdot \mid x, y) = \kappa(\cdot \mid y)$ that is ignoring the first input. Therefore we can use proposition 2.10 on φ' to get $\varphi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$. Projection $\rho_{\mathcal{Z}}$ onto $\mathcal{Z}$ also preserves the kernel property, so $\rho_{\mathcal{Z}}(\varphi'\kappa) : \mathcal{X} \rightsquigarrow \mathcal{Z}$ is yet another probability kernel. We denote this $\varphi \circ \kappa = \rho_{\mathcal{Z}}(\varphi'\kappa)$ and this is also called *composition of kernels* by some authors. We can extend this to the kernel-measure composition (of theorem 2.7): If $\mu \in \mathcal{P}(\mathcal{X})$ is a probability measure on $\mathcal{X}$ then $\rho_{\mathcal{Y}}(\kappa\mu) \in \mathcal{P}(\mathcal{Y})$ and we denote this $\kappa \circ \mu = \rho_{\mathcal{Y}}(\kappa\mu)$. In fact $\circ$ makes the class of measurable spaces into a category (see [13, Lawvere (1962)]), with identity $\mathrm{id}_{\mathcal{X}}(\cdot \mid x) = \delta_x$ (for a proof of this last statement see proposition A.16).

From kernels to processes

Throughout this section let $(\mathcal{X}_n)_{n\in\mathbb{N}}$ be a sequence of measurable spaces. For each $n \in \mathbb{N}$ let $\kappa_n : \mathcal{X}^{\underline{n}} \rightsquigarrow \mathcal{X}_{n+1}$ be a probability kernel.

Proposition 2.12. For all $n \in \mathbb{N}$ the composition $\kappa_n \dots \kappa_1$ yields a $\mathcal{X}_1$-probability kernel on $\mathcal{X}_2 \times \dots \times \mathcal{X}_{n+1}$.

Proof. This is by induction using proposition 2.10. □

Proposition 2.12 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 $(\pi_1, \pi_2, \dots) \in R\Pi$ we can get a distribution of the nth history $H_n \in \mathcal{H}_n$ by the composition of kernels

$$P_{n-1}\pi_{n-1} \dots P_2\pi_2 P_1\pi_1\mu \in \mathcal{P}(\mathcal{H}_n)$$

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

Theorem 2.13 (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

$$\kappa^{\underline{n-1}}\mu(A_1 \times A_2 \times \dots \times A_n) = \rho\left(A_1 \times A_2 \times \dots \times A_n \times \prod_{k=n+1}^{\infty} \mathcal{X}_k\right), \qquad \forall A \in \Sigma_{\mathcal{X}^{\underline{n}}}, n \in \mathbb{N}$$

Proof. We refer to [11, Kallenberg (2002)] thm. 5.17. □

In particular theorem 2.13 applied to the measure Dirac-measure δ_x can be interpreted as starting the process in $x \in \mathcal{X}_1$. Later we would like to consider function defined

Proposition 2.14 (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 $\kappa(A \mid x) = \mu_x(\mathcal{X}_1 \times A)$ defines a probability kernel $\kappa : \mathcal{X}_1 \rightsquigarrow \mathcal{X}_2 \times \mathcal{X}_3 \times \dots$.

Proof. Since we already know by theorem 2.13 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 $A \in \Sigma$ where $\Sigma = \bigotimes_{i=2}^{\infty} \Sigma_{\mathcal{X}_i}$. Let $\phi_A = x \mapsto \mu_x(A)$ for all $A \in \Sigma$ and define

$$\mathbb{G} = \left\{A \in \Sigma \;\middle|\; \phi_A \in \mathcal{M}(\mathcal{X}_1, [0, 1])\right\}$$

The cylinder sets

$$\mathbb{O} = \left\{A_2 \times \dots \times A_i \times \mathcal{X}_{i+1}, \dots \;\middle|\; A_i \in \Sigma_{\mathcal{X}_i}, i - 1 \in \mathbb{N}\right\}$$

is a generator for Σ stable under finite intersections. By contruction $\mathbb{O} \subseteq \mathbb{G}$ since

$$\phi_{A_2 \times \cdots \times A_i \times \mathcal{X}_{i+1} \times \ldots} = \kappa^{\underline{i-1}}(A_2 \times \cdots \times A_i \mid \cdot)$$

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)

$$\sigma(\mathbb{O}) = \Sigma \subseteq \mathbb{G}$$

implying that ϕ_A is measurable for all $A \in \Sigma$.

For showing that $\mathbb{G}$ is a Dynkin class, notice that clearly $\mathcal{X}_2 \times \mathcal{X}_3 \times \ldots$ and $\varnothing$ are in $\mathbb{G}$. 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 2.15. The Ionescu-Tulcea kernel satisfies $\prod_{i=1}^{\infty} \kappa_i = (\prod_{i=2}^{\infty} \kappa_i)\kappa_1$.

Proof. Let $x \in \mathcal{X}_1$. Since for the composition of finitely many kernels by associativity (proposition 2.10) it holds that $\kappa_n \ldots \kappa_1 = (\kappa_n \ldots \kappa_2)\kappa_1$. Therefore for any $n \in \mathbb{N}$ and $A \in \Sigma_{\mathcal{X}_2} \otimes \cdots \otimes \Sigma_{\mathcal{X}_n}$, using theorem 2.13 we have

$$\left(\prod_{i=1}^{\infty} \kappa_i\right)\left(A \times \prod_{k=n+1}^{\infty} \mathcal{X}_k \,\middle|\, x\right) = (\kappa_{n-1} \ldots \kappa_2)\kappa_1(A \mid x) = \left(\left(\prod_{i=2}^{\infty} \kappa_i\right)\kappa_1\right)\left(A \times \prod_{k=n+1}^{\infty} \mathcal{X}_k \,\middle|\, x\right)$$

for all $n \in \mathbb{N}$ and $A \in \Sigma_{\mathcal{X}^{\underline{n}}}$. By the uniqueness in theorem 2.13 we are done. □

Let $\mu \in \mathcal{P}(\mathcal{S}_1)$ be a measure on the first state space. By theorem 2.13 a HDP and a policy $\pi = (\pi_1, \pi_2, \ldots) \in R\Pi$ gives rise to a kernel $\kappa_\pi : \mathcal{P}(\mathcal{S}_1) \to \mathcal{P}(\mathcal{H}_\infty)$, namely

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

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 $\mu = \delta_s$ for some $s \in \mathcal{S}_1$ integration with respect to the measure $\kappa_\pi \delta_s$ and $\kappa_\pi(\cdot \mid s)$ is equivalent in the sense that

$$\int f(s, a_1, s_2, \ldots)\, \mathrm{d}\kappa_\pi(a_1, s_2, \cdots \mid s) = \int f(s_1, a_1, s_2, \ldots)\, \mathrm{d}\kappa_\pi \delta_s(s_1, a_1, \ldots)$$

Both of these measures can be interpreted as the stochastic process arising from starting in state s and following policy π. We will sometimes 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. This is where the reward kernels $R_1, R_2, \ldots$ come into play. However their stochastic properties will not be relevant for this section. For now we will only care about their expected values. Therefore we will need

Assumption 1 (Finite reward bound)**.** For each $n \in \mathbb{N}$ there exists a bound $R_{\max,n} > 0$ such that for all $h_n \in \mathcal{H}_n$ it holds that

$$R_n([-R_{\max,n}, R_{\max,n}] \mid h_n) = 1$$

This is assumed in the rest of this section and all following sections.

Remark 2.16. Assumption 1 implies that all $R_n(\cdot \mid h_n)$ has moments of all orders for any $h_n \in \mathcal{H}_n$.

Definition 2.17. We define for each $n \in \mathbb{N}$

$$r_n : \mathcal{H}_n \to \mathbb{R}, \quad r(h_n) = \int x \, \mathrm{d}R_n(x \mid h_n)$$

called the nth **expected reward function**.

Remark 2.18. The expected reward function r_n is measurable due to proposition 2.4.

We are now ready to talk about *value functions.* A value function is any function $V : \mathcal{S}_1 \to \mathbb{R}$ which assigns a real number to a starting state (a state in $\mathcal{S}_1$). An important class of value functions are the policy evaluations:

Definition 2.19 (Finite policy evaluation)**.** We define the function $V_{n,\pi} : \mathcal{S}_1 \to \mathbb{R}$ by

$$V_{n,\pi}(s_1) = \mathbb{E}^{\pi}_{s_1} \sum_{i=1}^{n} r_i = \int \sum_{i=1}^{n} r_i(s_1, a_1, \ldots, a_{i-1}, s_i) \, \mathrm{d}\kappa_\pi s_1$$

called the kth **finite policy evaluation**. When $n = 0$ we say $V_{0,\pi} = V_0 := 0$ for any π.

The finite policy evaluation measures the expected total reward of starting in state $s_1 \in \mathcal{S}_1$ and then following the policy π for n steps.

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

Assumption 2 (Discounting)**.** There exist a bound $R_{\max} > 0$ and a $\gamma \in [0, 1)$ called the **discount** factor such that $R_i([-R_{\max}\gamma^i, R_{\max}\gamma^i] \mid h_{i+1}) = 1, \ \forall h_{i+1} \in \mathcal{H}_{i+1}, \ i \in \mathbb{N}$.

This assumption allows the following definition of the following value function

Definition 2.20. We define the (infinite) **policy evaluation** by

$$V_\pi(s) = \mathbb{E}^{\pi}_{s} \lim_{n\to\infty} \sum_{i=1}^{n} r_i = \int \sum_{i=1}^{\infty} r_i \, \mathrm{d}\kappa_\pi s \tag{2.2}$$

Remark 2.21. In eq. (2.2) it is implicit that for each $i \in \mathbb{N}$ the reward r_i in the sum is applied the h_{i+1}th history as a integration variable from the measure $\kappa_\pi s$.

The infinite policy evaluation V_π measures the expected total reward after following the policy π an infinite number of steps. Whenever we work with the infinite policy evaluation we will always make assumption 2. We now mention some immediate properties of the finite and infinite policy evaluations

Proposition 2.22. Let assumption 2 hold. Then

1. $V_{n,\pi}, V_\pi$ are integrable with respect to $\kappa_\pi(\cdot \mid s)$ for any $\pi \in R\Pi$ and any $s \in \mathcal{S}_1$.

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

3. For any $\pi \in R\Pi$
$$\left|V_{n,\pi}\right|, |V_\pi| \leqslant R_{\max}(1-\gamma) < \infty$$

Proof.

1. By remark 2.16 the expected reward functions are measurable. Therefore sums and limits of them are aswell se proposition 2.4. Integrability follows once we show item 3.

2. By monotone or dominated convergence.

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

□

Remark 2.23. As this bound will occur again and again we denote it

$$V_{\max} := R_{\max}/(1-\gamma) \tag{2.3}$$

2.1.1 The optimal value function

Definition 2.24 (Optimal value functions)**.**

$$V_n^*(s) := \sup_{\pi\in R\Pi} V_{n,\pi}(s) = \sup_{\pi\in R\Pi} \mathbb{E}_s^\pi \sum_{i=1}^n r_i \qquad V^*(s) := \sup_{\pi\in R\Pi} V_\pi(s) = \sup_{\pi\in R\Pi} \mathbb{E}_s^\pi \sum_{i=1}^\infty r_i$$

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.

Remark 2.25. Under assumption 2 we have $\left|V_k^*\right|$, $|V^*| \leqslant V_{\max}$ since by proposition 2.22 all terms in the supremum are within this bound.

Remark 2.26. It is known that the optimal value function might not be Borel measurable (see ex. 2 p. 233 [3]). 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. Can optimal policies be chosen to be deterministic?

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

We will wait with questions 3 and 4 until the next section. In a quite general setting, questions 1 and 2 is investigated in [19, Schäl (1975)]. Here some additional structure on our process is imposed.

Definition 2.27 (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. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})$ is standard Borel for each $n \in N$.
2. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})$ is standard Borel. for each $n \in N$.
3. The set of admissible actions $\mathfrak{A}_n(h_n)$ is compact for any $h_n \in \mathcal{H}_n,\ n \in \mathbb{N}$.
4. $|V_\pi| < \infty$ for all policies $\pi \in R\Pi$.
5. $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N \geqslant n} \sup_{\pi \in R\Pi} \sum_{t=n+1}^{N} \mathbb{E}_s^\pi r_t \to 0, \quad n \to \infty$

Remark 2.28. The last two points in setting 1 is readily implied by assumption 2 because of proposition 2.22 and the fact that under assumption 2 we have $\sum_{t=n+1}^{N} r_t \leqslant \gamma^n R_{\max}$.

To understand the results of [19] which we will present shortly, we will need some additional concepts, namely the weak topology and semicontinuity.

Definition 2.29 (Weak topology)**.** Let $\mathcal{X}$ be a metrizable space equipped with the Borel σ-algebra. Consider the family of sets of probability measures on $\mathcal{X}$, that is a family of subsets of the space of $\mathcal{X}$-probability measures $\mathcal{P}(\mathcal{X})$

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

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 $\mathcal{V}$.

To get a feel for the properties of the weak topology we state the following proposition:

Proposition 2.30 (Properties of the weak topology)**.** Let $\mathcal{X}$ be a seperable metric space with metric d. Denote by $C(\mathcal{X})$ the set of continuous real-valued functions on $\mathcal{X}$ and by $U_d(\mathcal{X})$ the set of uniformly continuous real-valued functions on $\mathcal{X}$. Let $p, p_1, p_2, \dots \in \mathcal{P}(\mathcal{X})$. Then the following is equivalent:

(a) $p_n \to p$ (b) $\forall f \in C(\mathcal{X}) : \int f \,\mathrm{d}p_n \to \int f \,\mathrm{d}p$ (c) $\forall g \in U_p(\mathcal{X}) : \int g \,\mathrm{d}p_n \to \int g \,\mathrm{d}p$

Proof. We refer to [3] prop.7.21. □

The following proposition shows that it is possible to view the space of probability measures $\mathcal{P}(\mathcal{X})$ on a standard Borel space $\mathcal{X}$ as a metric space as well. This will also be relevant in the later sections.

Proposition 2.31. Let $\mathcal{X}$ be a standard Borel measurable space. Then the space $\mathcal{P}(\mathcal{X})$ of probabilty measures on $\mathcal{X}$ equipped with the weak topology is standard Borel. If furthermore $\mathcal{X}$ is compact then $\mathcal{P}(\mathcal{X})$ is also compact.

Proof. We refer to [3] cor.7.25.1 and prop.7.22. □

Definition 2.32 (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 $x \in U$. We define f to by **lower** semicontinuous if $-f$ is upper semicontinuous.

To get a feel for semicontinuity we here give some simple properties of semicontinuous functions.

Proposition 2.33.

1. If $\mathcal{X}$ is a metrizable space then $f : \mathcal{X} \to \overline{\mathbb{R}}$ is upper semicontinuous if and only if for each sequence $x_1, x_2, \dots \in \mathcal{X}$ with $x_n \to x \in \mathcal{X}$ we have $\limsup f(x_n) \leqslant f(x)$ (analogously $\liminf f(x_n) \geqslant f(x)$ for lower semicontinuous f).
2. If $f,\ g : \mathcal{X} \to \overline{\mathbb{R}}$ are upper (lower) semicontinuous then $f + g$ is upper (lower) semicontinuous.
3. If furthermore g is continuous and non-negative then $f \cdot g$ is upper (lower) semicontinuous.
4. 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.

Proof. We refer to [3] p. 147. □

We are now ready to present the results of [19]. Under setting 1 Schäl introduced two sets of criteria for the existence of an optimal policy:

Condition S (Set-wise continuity)**.** For all $n \in \mathbb{N}$

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

Condition W (Weak continuity)**.** For all $n \in \mathbb{N}$

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 2.34 (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 [19]. □

Corollary 2.35. Under setting 1 when either (S) or (W) hold then V^* is (Borel) measurable.

Proof. Since by theorem 2.34 there exists an optimal policy π^* we have $V^* = V_{\pi^*}$ which is measurable due to proposition 2.22. □

Schäls theorem tells us that optimal policies exist and that the n-stage optimal value functions converge to the infinite optimal value function 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 2.36 (Markov decision process)**.** A **Markov decision process** (MDP) consists of

1. $\mathcal{S}$ a measurable space of states.
2. $\mathcal{A}$ a measurable space of actions.
3. $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ a transition kernel.
4. $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathbb{R}$ a reward kernel discounted by
5. a disount factor $\gamma \in [0, 1)$ (see assumption 2).
6. $\mathfrak{A}(s) \subseteq \mathcal{A}$ a set of admissable actions for each $s \in \mathcal{S}$.

Remark 2.37. This is a special case of the history dependent decision process (definition 2.5) 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. That is $P_n(\cdot \mid s_1, \ldots, s_n, a_n) = P(\cdot \mid s_n, a_n)$ for all $n \in \mathbb{N}$.
- R_n depends only on s_n and a_n and does not differ with n except for the discount factor. That is $R = R_n/\gamma^{n-1}$ for all $n \in \mathbb{N}$

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

Remark 2.38. One could ask if it is possible to embed a HDP into an MDP by taking unions or cartesian products of the state and action spaces:

$$\mathcal{S} := \bigcup_{i \in \mathbb{N}} \mathcal{S}_i, \quad \mathcal{A} := \bigcup_{i \in \mathbb{N}} \mathcal{A}_i, \quad \text{or} \quad \mathcal{S} := \prod_{i \in \mathbb{N}} \mathcal{S}_i, \quad \mathcal{A} := \prod_{i \in \mathbb{N}} \mathcal{A}_i$$

One attempt at this can be found in [3] chapter 10, but this will not be covered here. Note however that any 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.

Remark 2.39. In an MDP the policy evaluations can be written

$$V_{\pi,n}(s_1) = \mathbb{E}^{\pi}_{s_1} \sum_{i=1}^{n} r_i = \int \sum_{i=1}^{n} \gamma^{i-1} r(s_i, a_i) \, \mathrm{d}\kappa_\pi(a_1, s_2, \cdots \mid s_1),$$

$$V_{\pi}(s_1) = \mathbb{E}^{\pi}_{s_1} \sum_{i=1}^{\infty} r_i = \int \sum_{i=1}^{\infty} \gamma^{i-1} r(s_i, a_i) \, \mathrm{d}\kappa_\pi(a_1, s_2, \cdots \mid s_1)$$

Recalling that $r_i(s_1, a_1, s_2, \ldots) = \gamma^{i-1} r(s_i, a_i)$.

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 2.40 (Policy classes)**.** A policy $\pi = (\pi_1, \pi_2, \dots) \in R\Pi$ is called **Markov** if it only depends on the last state is the history. That is there exist $\pi_1, \pi_2, \dots : \mathcal{S} \rightsquigarrow \mathcal{A}$ such that $\pi_i(\cdot \mid s_1, \dots s_i) = \pi_i(\cdot \mid s_i)$. We denote the set of (random) Markov policies by $M\Pi$. If $\pi_1 = \pi_2 = \dots$ the Markov policy is called **stationary** and the set of them denote by $S\Pi$. Furthermore π is called **deterministic** if all π_i are degenerate, i.e. for all i we have $\pi_i(\{a_i\} \mid h_i) = 1$ for some $a_i \in \mathcal{A}_i$. We denote the deterministic version of the policy classes by the letter D.

Remark 2.41. We have the following inclusions of policy classes

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

Note that stationary policies might not exist in HDPs, but always exist for MDPs. A policy $(\pi_1, \pi_2, \dots) \in R\Pi$ is deterministic if and only if there exist measurable functions $\varphi_n : \mathcal{H}_n \to \mathcal{A}$ such that $\pi_n(\cdot \mid h_n) = \delta_{\varphi_n(h_n)}$. Therefore we shall sometimes write $\pi_n(h_n) = \varphi_n(h_n)$, viewing π_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. We now need to integrate over value functions so it is necessary to talk about integrability.

Definition 2.42. Let $\mathcal{L}_\infty(\mathcal{S})$ denote the set of measurable functions on $\mathcal{S}$ which are essentially bounded with respect to all measures arising as the distribution $\rho_{\mathcal{S}_i}(\kappa_\pi \mu)$ of the ith state space given some initial distribution $\mu \in \mathcal{P}(\mathcal{S})$ and some policy $\pi \in R\Pi$. Here $\rho_{\mathcal{S}_i} : \mathcal{H}_\infty \to \mathcal{S}_i$ is projection onto the ith state space. Similarly $\mathcal{L}_p$ for $p \in [1, \infty)$ is defined with respect to all such probability measures.

Remark 2.43. Note that $\mathcal{L}_\infty(\mathcal{S})$ contains all (everywhere) bounded functions, so 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.

Definition 2.44 (The T-operators)**.** For a stationary policy $\tau \in S\Pi$ and a value function $V : \mathcal{S} \to \mathbb{R} \in \mathcal{L}_\infty(\mathcal{S})$ we define the operators

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

$$\text{The Bellman optimality operator: } TV := s \mapsto \sup_{a \in \mathfrak{A}(s)} \left(r(s,a) + \gamma \int V(s') \, \mathrm{d}P(s' \mid s, a) \right)$$

Remark 2.45. We will sometimes write $T_a = T_{\delta_a}$ for $a \in \mathfrak{A}(s)$. Using this we can express T alternatively as $TV = s \mapsto \sup_{a \in \mathfrak{A}(s)} T_a V(s)$.

The Bellman optimality operator T 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 2.46 (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. T and T_τ are γ-contractive on $\mathcal{L}_\infty(\mathcal{S})$.
6. V_τ is the unique bounded fixed point of T_τ in $\mathcal{L}_\infty(\mathcal{S})$.

Proof.

1. Measurability is by proposition 2.4 the rest follows by dominated convergence.
2. This is an exercise in using the definitions and properties of probability kernels that we have developed:

$$\begin{aligned}
&T_{\tau_1} V_{k,(\tau_2,\dots)}(s) \\
&\overset{\text{def}}{=} \int r(s_1,a_1) + \gamma \int \sum_{i=1}^{k} \gamma^{i-1} r(s_{i+1},a_{i+1}) \, \mathrm{d}\kappa_{(\tau_2,\dots)}(a_2,s_3,a_3,\dots \mid s_2) \, \mathrm{d}P\tau_1(a_1,s_2 \mid s_1) \\
&\overset{2.10}{=} \int\int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i) \, \mathrm{d}\kappa_{(\tau_2,\dots)}(a_2,s_3,a_3,\dots \mid s_2) \, \mathrm{d}P(s_2 \mid s_1,a_1) \, \mathrm{d}\tau_1(a_1 \mid s_1) \\
&\overset{2.10}{=} \int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i) \, \mathrm{d}\dots P\tau_2 P(a_1,s_2,\dots \mid s_1,a_1) \, \mathrm{d}\tau_1(a_1 \mid s_1) \\
&\overset{2.10}{=} \int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i) \, \mathrm{d}\dots P\tau_2 P\tau_1(a_1,s_2,\dots \mid s_1) \\
&\overset{\text{def}}{=} \int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i) \, \mathrm{d}\kappa_\pi(a_1,s_2,\dots \mid s_1) \\
&\overset{\text{def}}{=} V_{k+1,\pi}(s_1)
\end{aligned}$$

   Now use this inductively.

3. This is by 2. and a monotone or dominated convergence.
4. By 3. $T_\tau V_\tau = T_\tau \lim_{k\to\infty} T_\tau^k V_0 = \lim_{k\to\infty} T_\tau^{k+1} V_0 = V_\tau$.
5. Let $V, V' \in \mathcal{L}_\infty(\mathcal{S})$ and let $K = \left\|V - V'\right\|_\infty$. Then since the rewards are bounded

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

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

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

□

Proposition 2.46 gives us a way of obtaining the finite policy evaluations by iteratively using the T_τ operator (for $\tau = \tau_1, \tau_2, \dots$). We emphasize this by writing it as our first algorithm, called the *policy evaluation algorithm.* We should say that this algorithm has been considered before and is used in other algorithms like the *policy iteration algorithm* (see e.g. [3]).

```
Algorithm 1: Simple theoretical policy evaluation
  Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$, a Markov policy
         $\pi = (\tau_1, \tau_2, \dots) \in M\Pi$ to evaluate.
1 Initialize the expected reward function $r \leftarrow \int x \, \mathrm{d}R(x \mid \cdot)$.
2 Initialize the starting value estimator $\widetilde{V}_0 = 0$.
3 for $k = 0, 1, 2, \dots, K-1$ do
4     Update the value estimator $\widetilde{V}_{k+1} \leftarrow T_{\tau_{K-k}} \widetilde{V}_k$.
  Output: The finite policy evaluation $V_{\pi,K}$.
```

The reason behind calling line 3 *theoretical* is that we have not talked about how to actually represent any of r, T_π or V in a computer. We will take up this discussion later (see e.g. example 2.60).

In many cases we are actually interested the evaluation of a stationary policy $\tau \in S\Pi$ on an infinite horizon, that is we are interested in obtaining the infinite policy evaluation V_τ. To this end line 3 can still be used as a method of approximation, since by proposition 2.46.(3,5 and 6) and Banach fixed point theorem we have that

Corollary 2.47. Let $\tau \in S\Pi$ be a stationary policy. Then the kth finite policy evaluation $V_{\pi,K}$, that is the output of line 3 with k iterations, satisfy

$$\left|V_\tau - V_{\tau,k}\right| \leqslant \left\|V_\tau - V_{\tau,k}\right\|_\infty \leqslant \gamma^k V_{\max}$$

It is in fact mostly due corollary 2.47 that line 3 is used.

2.2.1 Greedy policies

Greedy policies will be a crucial tool in our investigation of optimal policies. Indeed it turns out that the optimal policy is a greedy policy with respect to the optimal value function.

Definition 2.48. Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A} \in S\Pi$ be a stationary policy and let $V : \mathcal{S} \to \mathbb{R}$ be a measurable value-function. We define

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

as the set of **greedy** actions w.r.t. V. If 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.

Remark 2.49. For a function $f : \mathcal{X} \to \overline{\mathbb{R}}$ and $A \subseteq \mathcal{X}$ we denote

$$\operatorname*{argmax}_{x \in A} f(x) = \left\{ x \in A \;\middle|\; f(x) = \sup_{x' \in A} f(x') \right\}$$

In order to prove the existence of greedy policies we need some additional structure on our MDP. Recall (definition 2.27) that a measurable space is standard Borel if it is Polish and equipped with the Borel σ-algebra. Also recall that the space of probability measures $\mathcal{P}(\mathcal{X})$ over a standard Borel space $\mathcal{X}$ is also standard Borel when endowed with the weak topology (see definition 2.29 and proposition 2.31).

Definition 2.50 (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.

Setting 2 (Greedy MDP)**.**

1. $\mathcal{S}$ and $\mathcal{A}$ are standard Borel.
2. The set of admissable actions $\mathfrak{A}(s) \subseteq \mathcal{A}$ is compact for all $s \in \mathcal{S}$ and $\Gamma = \{(s, a) \in \mathcal{S} \times \mathcal{A} \mid a \in \mathfrak{A}(s)\}$ 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.

Proposition 2.51. Let $\mathcal{X}$ and $\mathcal{Y}$ be separable metrizable and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a continuous probability 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. We refer to [3] prop.7.31. □

Proposition 2.52. A upper (lower) semicontinuous function $f : \mathcal{X} \to \overline{\mathbb{R}}$ on a compact metrizable space $\mathcal{X}$ 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. Let $x_1, x_2, \dots \in \mathcal{X}$ be a sequence such that $f(x_n) \to \sup_{x \in \mathcal{X}} f(x)$ then since $\mathcal{X}$ is compact this sequence has at least one accumulation point $x^* \in \mathcal{X}$. Let $x_{k_1}, x_{k_2}, \dots$ be a subsequence such that $x_{k_n} \to x^*$. Since f is upper semicontinuous

$$\sup_{x \in \mathcal{X}} f(x) = \limsup_{n \to \infty} f(x_n) = \limsup_{n \to \infty} f(x_{k_n}) \leqslant f(x^*) \leqslant \sup_{x \in \mathcal{X}} f(x)$$

The statement for lower semicontinuous f is analogous. □

Proposition 2.53. 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. Let

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

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 $f(x, \varphi(x)) = f^*(x)$.

Remark 2.54. Here $\mathrm{Gr}(f) = \big\{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid y = f(x)\big\}$ denotes the graph of the function $f : \mathcal{X} \to \mathcal{Y}$.

Proof of proposition 2.53. We refer to [3] prop.7.33. □

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

Proposition 2.55 (Existence of greedy deterministic policies)**.** Under setting 2 let $V : \mathcal{S} \to \mathbb{R} \in \mathcal{L}_\infty(\mathcal{S})$ be upper semicontinuous. Then for any $s \in \mathcal{S}$ it holds that

1. $(s, a) \mapsto T_a V(s)$ is upper semicontinuous.
2. $G_V(s) \neq \varnothing$. I.e. the set of greedy actions at the state s is non-empty.
3. There exist a deterministic greedy policy τ_V for V.
4. $TV(s) = T_{\tau_V} V(s) = \sup_{\tau \in S\Pi} T_\tau V(s)$ and TV is upper semicontinuous.

Proof.

1. This is a consequence of proposition 2.33 and proposition 2.51 since r is upper semicontinuous.
2. Since by 1. $(s, a) \mapsto T_a V(s)$ is upper semicontinuous, this follows by proposition 2.52.
3. Recall that the set of admissable state-action pairs $\Gamma = \big\{(s, a) \in \mathcal{S} \times \mathcal{A} \mid a \in \mathfrak{A}(s)\big\}$ assumed to be closed in setting 2. Let $\varphi : \mathcal{S} \to \mathcal{A}$ be the Borel-measurable function obtained from proposition 2.53 with $\mathrm{Gr}(\varphi) \subseteq \Gamma$. Let $s \in \mathcal{S}$ be any state. Then

   $$T_{\varphi(s)} V(s) = \sup_{a \in \mathfrak{A}(s)} T_a V(s)$$

   thus $\varphi(s) \in \mathrm{argmax}_{a \in \mathfrak{A}(s)} T_a V(s) = G_V(s)$. Therefore the induced deterministic policy

   $$\tau_V(\cdot \mid s) = \delta_{\varphi(s)}$$

   is greedy with respect to V. Upper semicontinuity of TV follows from proposition 2.51 since $TV = T_{\tau_V}$. If $\tau \in S\Pi$ is a stationary policy then

   $$\begin{aligned} T_\tau V(s) &= \int\int V(s')\, \mathrm{d}P(s' \mid s, a)\, \mathrm{d}\tau(a \mid s) \\ &\leqslant \int \max_{a \in \mathfrak{A}(s)} \int V(s')\, \mathrm{d}P(s' \mid s, a)\, \mathrm{d}\tau(a \mid s) \\ &= \max_{a \in \mathfrak{A}(s)} \int V(s')\, \mathrm{d}P(s' \mid s, a) \\ &= TV(s) \end{aligned}$$

   Therefore $\sup_{\tau \in S\Pi} V \leqslant TV$. On the other hand τ_V is in the supremum so $\sup_{\tau \in S\Pi} V \geqslant TV$.
4. By definition $T_{\tau_V} V(s) = T_{\mathrm{argmax}_{a \in \mathfrak{A}(s)} T_a V(s)} V(s) = \sup_{a \in \mathfrak{A}(s)} T_a V(s) = TV(s)$.

□

2.2.2 Existence of optimal policies

In the light of proposition 2.55 (and induction) since $V_0 = 0$ is upper semicontinuous for any $k \in \mathbb{N}$ we have that $T^k V_0$ is upper semicontinuous and thus has an associated greedy policy $\tau_{T^k V_0} \in DS\Pi$ which we will denote τ_k^*.

Proposition 2.56 (Existence of n-stage optimal policies)**.** 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 where $\tau_k^* = \tau_{T^k V_0}$ is any deterministic greedy policy for $T^k V_0$ for any $k \in \mathbb{N}$.

Proof. We begin by stating some preliminary facts.

Firstly with a (randomized history dependent) policy $\pi = (\pi_1, \pi_2, \dots) \in R\Pi$ one can obtain another policy $\pi^{s_1,a_1} \in R\Pi$ by taking a state action pair $(s_1, a_1) \in \mathcal{S} \times \mathcal{A}$ and consider the kernels $\pi_1^{s_1,a_1} = \pi_2(\cdot \mid s_1, a_1, \cdot), \pi_2^{s_1,a_1} = \pi_3(\cdot \mid s_1, a_1, \cdot, \cdot, \cdot), \dots$.

Secondly we have for any $V \in \mathcal{L}_\infty(\mathcal{S})$ that $TV(s) \overset{\text{def}}{=} \sup_{a \in \mathfrak{A}(s)} T_a V(s) = \sup_{\tau \in S\Pi} T_\tau V(s)$.

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. Let $s_1 \in \mathcal{S}$.

$$V_k^*(s_1) \overset{\text{def}}{=} \sup_{\pi \in R\Pi} \int \sum_{i=1}^{k} \gamma^{i-1} r(s_i, a_i) \, \mathrm{d}\kappa_\pi(a_1, s_2, \dots \mid s_1) \tag{2.4}$$

$$\overset{2.9}{=} \sup_{\pi \in R\Pi} \int r(s_1, a_1) + \gamma \int \left(\int \sum_{i=1}^{k-1} \gamma^{i-1} r(s_{i+1}, a_{i+1}) \, \mathrm{d}\kappa_{\pi^{s_1,a_1}}(a_2, s_3, \cdots \mid s_2) \right) \tag{2.5}$$

$$\mathrm{d}P(s_2 \mid s_1, a_1) \, \mathrm{d}\pi_1(a_1 \mid s_1) \tag{2.6}$$

$$\overset{\text{hyp.}}{\leqslant} \sup_{\pi \in R\Pi} \int r(s_1, a_1) + \gamma \int V_{k-1}^*(s_2) \, \mathrm{d}P(s_2 \mid s_1, a_1) \, \mathrm{d}\pi_1(a_1 \mid s_1) \tag{2.7}$$

$$= \sup_{\pi_1 \in S\Pi} T_{\pi_1} V_{k-1}^*(s_1) \tag{2.8}$$

$$\overset{2.55.4}{=} T V_{k-1}^*(s_1) \tag{2.9}$$

The integral makes sense in the eq. (2.7) because of the induction hypothesis, since $V_{k-1}^* = V_{k,(\tau_{k-1}^*,\dots,\tau_0^*)}$ and this is integrable like all policy evaluations (see proposition 2.22). In eq. (2.8) the supremum changes because only the first step in the policy is used (this first step is a stationary policy). Equation (2.9) is by proposition 2.55.4. Since s_1 was arbitrary we must have $V_k^* \leqslant T V_{k-1}^*$. On the other hand by proposition 2.46 and induction hypothesis we have

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

But since $(\tau_{k-1}^*, \dots, \tau_0^*)$ 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 preserves this property (see proposition 2.55.4). □

Proposition 2.57. Under setting 2 it holds that $V^* = \lim_{k \to \infty} T^k V_0^*$. Furthermore under the greedy policy τ_{V^*} exists and is a deterministic stationary optimal policy.

Proof. Since setting 2 implies condition (S) (see above theorem 2.34) we have by theorem 2.34 that $T^k V_0^* = V_k^* \to V^*$ (as always we mean pointwise convergence here). We know by proposition 2.56 that V_k^* is semi uppercontinuous for all $k \in \mathbb{N}$. Also we have that

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

Here $\downarrow$ denotes downwards monotone (pointwise) convergence. So by proposition 2.33 the infimum $\inf_k \widehat{V}_k = V^* - V_{\max}$ is upper semicontinuous and thus V^* is upper semicontinuous. Therefore by proposition 2.55 there exists a deterministic greedy policy τ_{V^*} which satisfies

$$T_{\tau_{V^*}} V^* = TV^* \tag{2.10}$$

By proposition 2.46 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.24) $V^* = \lim_{k\to\infty} T^k V_0^*$ is the unique fixed point of T in B. Again by Banach fixed point theorem and eq. (2.10) V^* is the fixed point of $T_{\tau_{V^*}}$, which by proposition 2.46 also has $V_{\tau_{V^*}}$ as fixed point. By uniqueness $V_{\tau_{V^*}} = V^*$ and thus τ_{V^*} is optimal. □

Remark 2.58. The property that $TV^* = V^*$ is often referred to as *Bellman's optimality equation.*

Corollary 2.59. Under setting 2 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 2.57. The bound on $\left|T^k V - V^*\right|$ is by the Banach fixed point theorem. Upper semicontinuity of the policy evaluations and optimal value functions followed from the proofs of proposition 2.55 and proposition 2.57. □

Comparison to results of [3, Bertsekas and Shreve (2007)]

In [3] (prop. 8.6 and cor. 9.7.2) results very similar to proposition 2.56 and proposition 2.57 are also established with in a slightly different setup. Besides having a state and action space, [3] 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 [3] 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 [3] 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

In the previous section we saw that under setting 2 the Bellman optimality operator T is equivalent to the T_τ operator when τ is a greedy policy with respect to the input. Also T applied repeatedly on $V_0 = 0$ creates the sequence of k-optimal policy evaluations which convergences to the optimal value function.

Value iteration is a broad notion that can refer to many algorithms in dynamic programming, that somehow updates value functions, but perhaps the simplest is just iterative application of the T-operator.

```
Algorithm 2: Simple theoretical value iteration
    Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$.
  1 Initialize the expected reward function $r \leftarrow \int x \ \mathrm{d}R(x \mid \cdot)$.
  2 Initialize the first estimated value function $\widetilde{V}_0 = 0$.
  3 for $k = 0, 1, 2, \dots, K-1$ do
  4     Update the value function $\widetilde{V}_{k+1} \leftarrow T\widetilde{V}_k$.
    Output: The $K$th optimal value function $V_K^* = \widetilde{V}_K$.
```

To actually use line 3 it becomes relevant if the environment can be *computed* efficiently (in a computer). An easy example such an easily computed environment is an MDP where both the state and the action space are finite i.e. $|\mathcal{S}| \cdot |\mathcal{A}| < N < \infty$ for a not too big number $N \in \mathbb{N}$. Indeed 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 (an MDP under setting 2), and therefore line 3 could be applied in other cases if one has a practical way of representing the iterations $TV_0, T^2V_0, \dots$.

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.

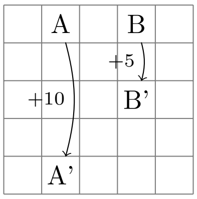


Figure 2.1: The simple gridworld Markov decision process.

Example 2.60 (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.

Finally $\gamma = 0.9$ is the standard value of the discount factor in this example. Note that the discount factor is part of the definition of the environment, and thus the concept of optimality for the environment depends on the value of γ.

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 $\mathfrak{A}(s)$ is equal to the full action space $\mathcal{A}$ for all $s \in \mathcal{S}$, which is trivially compact. The rewards are bounded by $R_{\max} = 10$ and therefore $V_{\max} = 10/(1-0.9) = 100$. Thus we can apply corollary 2.59 and for $\widetilde{V}_0 = 0$ get that

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

By proposition 2.46 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

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

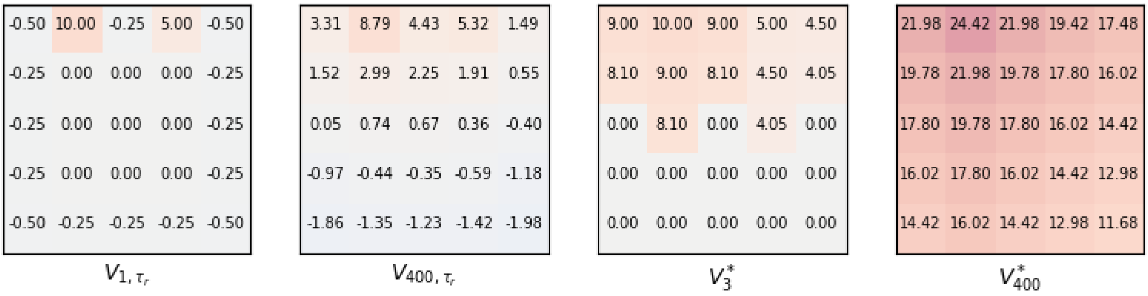
To use line 3 and line 3 on the gridworld example we need to understand how the kernels R, P, value functions, policies and operators T_τ, T are represented in a computer. Since $\mathcal{S}$ and $\mathcal{A}$ are finite and small we can simply treat value functions V as a vector $\widetilde{V} \in \mathbb{R}^{\mathcal{S}}$ and policies as a matrix of point probabilities $\widetilde{\tau} \in \mathbb{R}^{\mathcal{A}\times\mathcal{S}}$ so that $\widetilde{\tau}(a,s) = \tau(\{a\} \mid s)$. Since the rewards are deterministic the step $r \leftarrow \int x \, \mathrm{d}R(x \mid \cdot)$ is irrelevant. P is also deterministic so there exists a function $\widetilde{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ such that $P(\{\widetilde{P}(s,a)\} \mid s,a) = 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore the update step $\widetilde{V}_{k+1} \leftarrow T_\tau \widetilde{V}_k$ in line 3 becomes

$$\text{For each } s \in \mathcal{S} : \widetilde{V}_{k+1}(s) \leftarrow \sum_{a\in\mathcal{A}} \left(r(s,a) + \gamma \widetilde{V}_k(\widetilde{P}(s,a))\right) \widetilde{\tau}(a,s)$$

Similarly in line 3 the update step $\widetilde{V}_{k+1} \leftarrow T\widetilde{V}_k$ in line 3 becomes

$$\text{For each } s \in \mathcal{S} : \widetilde{V}_{k+1}(s) \leftarrow \max_{a\in\mathcal{A}} \left(r(s,a) + \gamma \widetilde{V}_k(\widetilde{P}(s,a))\right)$$

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



-0.50	10.00	-0.25	5.00	-0.50
-0.25	0.00	0.00	0.00	-0.25
-0.25	0.00	0.00	0.00	-0.25
-0.25	0.00	0.00	0.00	-0.25
-0.50	-0.25	-0.25	-0.25	-0.50

V_{1,τ_r}

3.31	8.79	4.43	5.32	1.49
1.52	2.99	2.25	1.91	0.55
0.05	0.74	0.67	0.36	-0.40
-0.97	-0.44	-0.35	-0.59	-1.18
-1.86	-1.35	-1.23	-1.42	-1.98

V_{400,τ_r}

9.00	10.00	9.00	5.00	4.50
8.10	9.00	8.10	4.50	4.05
0.00	8.10	0.00	4.05	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00

V_3^*

21.98	24.42	21.98	19.42	17.48
19.78	21.98	19.78	17.80	16.02
17.80	19.78	17.80	16.02	14.42
16.02	17.80	16.02	14.42	12.98
14.42	16.02	14.42	12.98	11.68

V_{400}^*

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).

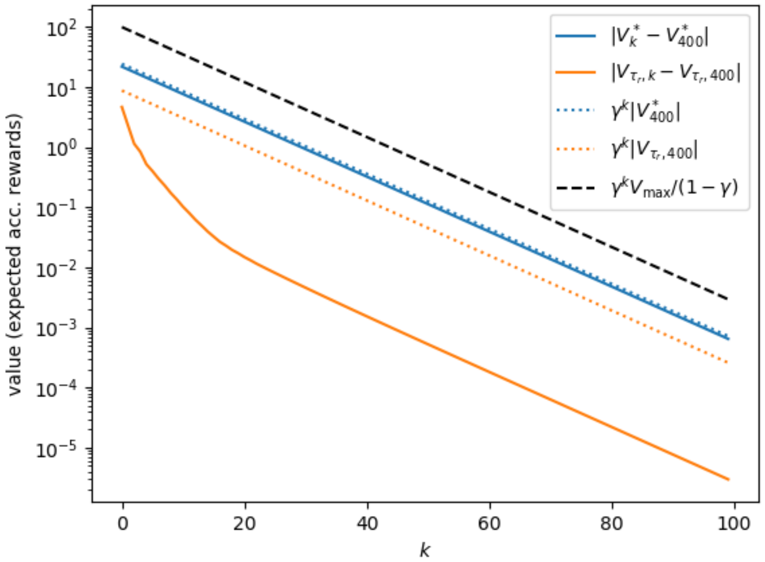


Figure 2.3: Convergence of gridworld value functions compared with the theoretical bounds. The black dashed line is the general theoretical bound for both T and T_τ by Banachs fixed point theorem and the maximum value $V_{\max} = R_{\max}/(1-\gamma)$. The dotted blue and orange uses $|V_k^*|$ and $|V_{\tau,k}|$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

2.2.4 Q-functions

A **Q-function** is any function that assigns a real number to every state-action pair, that is any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Q-functions are also called *action-value* functions, to distinguish them from the *value* functions we have discussed in the previous sections. The idea of Q-functions (and the letter Q) originates to [22, Watkins (1989)]. Upon the definition he notes

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

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 $a^* \in \mathfrak{A}(s)$ at state s requires only a maximization over the Q-function itself: $a^* = \operatorname{argmax}_{a \in \mathfrak{A}(s)} Q(s,a)$. This should be compared to finding an optimal action according to a value function V: $a^* = \operatorname{argmax}_{a \in \mathfrak{A}(s)} r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V$. 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 can only sample from P and R 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. 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.

Assumption 3 (Finite admissable actions)**.** $\mathfrak{A}(s)$ is finite for every $s \in \mathcal{S}$.

Throughout this section we will work under setting 2 and assumption 3.

Remark 2.61. We make assumption 3 to ensure that the supremum $\sup_{a\in\mathfrak{A}(s)} Q(s,a)$ is attained for any $Q \in \mathcal{L}_\infty(\mathcal{S}\times\mathcal{A})$. One might be able discard assumption 3 and instead demand that all Q-functions be upper semicontinuous generalizing the discussion in this section. We have not pursued this generalization.

Definition 2.62 (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 2.63 (Operators for Q-functions)**.** For any stationary policy $\tau \in S\Pi$ and integrable Q-function $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R} \in \mathcal{L}_\infty(\mathcal{S}\times\mathcal{A})$ we define

$$\text{Next-step operator: } P_\tau Q(s,a) = \int Q(s',a')\, \mathrm{d}\tau P(s',a' \mid s,a)$$

$$\text{Policy evaluation operator: } T_\tau Q(s,a) = r(s,a) + \gamma \int Q(s',a')\, \mathrm{d}\tau P(s',a' \mid s,a)$$

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

where $T_a = T_{\delta_a}$.

Remark 2.64. The next-step operator P_τ is defined for simplications in proofs, especially in the analysis of [7, Fan et al. (2020+)] in the later sections. Using P_τ we can write T_τ alternatively as $T_\tau Q(s,a) = r(s,a) + \gamma P_\tau Q(s,a)$.

Definition 2.65 (Greedy policies for Q-functions)**.** Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A}$ be a stationary policy. Define $G_Q(s) = \operatorname{argmax}_{a\in\mathfrak{A}(s)} Q(s,a)$. 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.

Proposition 2.66 (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
   (a) τ is greedy for $Q_{k,\pi}$ if and only if τ is greedy for $V_{k,\pi}$.
   (b) τ is greedy for Q_π if and only if τ is greedy for V_π.
4. Optimal policies are related by $\max_{a\in\mathfrak{A}(s)} Q^*(s,a) = V^*(s)$ 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 2.67 (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 2.66 and proposition 2.67. Measurability of $Q_{k,\pi}$ and Q_π follow from measurability of $V_{k,\pi}$, V_π and proposition 2.4. Upper semicontinuity of $Q_{k,\pi}$ and Q_π follows from proposition 2.51 because $V_{k,\pi}$ and V_π are upper semicontinuous (see corollary 2.59).

For proposition 2.66.1 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 2.66.2 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 2.66.2 iteratively starting with $\tau = \tau_1, \pi = (\tau_2, \tau_3, \dots)$.

The $\tau(Q_{k,\pi}) = \tau(V_{k,\pi})$ part of proposition 2.66.3 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 2.22.3.

For proposition 2.66.4 $Q_k^* = \sup_{\pi\in R\Pi}(r + \gamma\mathbb{E}V_{k,\pi}) \leqslant r + \gamma\mathbb{E}V_k^* = r + \gamma\mathbb{E}V_{\pi_k^*} \leqslant Q_k^*$. The same argument works for the second part.

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

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

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

Therefore by proposition 2.67.1

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

since by proposition 2.66.3 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 2.67.3.

Since now Q^* and Q_{τ^*} are fixed points for T they must be equal, concluding the last point, namely proposition 2.67.4. □

Q-iteration

Similar to the value iteration algorithm (line 3) we can define the corresponding for Q-iteration.

```
Algorithm 3: Simple theoretical Q-iteration
  Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$
1 Initialize the expected reward function: $r \leftarrow \int x \, \mathrm{d}R(x \mid \cdot)$.
2 Initialize the starting Q-estimator: $\widetilde{Q}_0 \leftarrow r$.
3 for $k = 0, 1, 2, \ldots, K-1$ do
4     Update the Q-estimator $\widetilde{Q}_{k+1} = T\widetilde{Q}_k$
  Output: The $K$-optimal Q-function $Q_K^* = \widetilde{Q}_K$.
```

By proposition 2.67.3 we thus have convergence of the line 3:

Corollary 2.68. $\left|Q_k^* - Q^*\right| \leqslant \gamma^k \|Q^*\|_\infty \leqslant \gamma^k V_{\max}$.

2.2.5 Why are we not done?

So far we have shown that value iteration under setting 2 and Q-iteration under additionally assumption 3, can solve all such discounted MDPs with exponential convergence in γ. This is a broad class of problems! We name a few examples:

1. Gridworld (example 2.60).

2. Board games like *chess* and *go* against a fixed (possibly randomized) opponent policy can be modelled accurately as such finite MDPs (putting γ close to 1, so that winning late in the game is still considered worth the effort).

3. *Pole balancing* in a 2D physical simulation environment (see e.g. the famous *cartpole* example from [2, Barto et al. (1983)]). One may even add random effects (such as *wind* effects) to full use of our stochastic setup.

The problem is of course that the value functions and operators which are used in line 3 are not computable in practice. For example the state space of chess is very large (roughly $|\mathcal{S}_{\text{chess}}| \geqslant 10^{43}$). This means that if we were to use line 3 naively (with finite implementation as example 2.60) then we would have to store a vector of roughly $N \cdot 10^{43}$ real numbers for each Q-function we define, where N is the average number of admissable actions at each state $\mathfrak{A}(s), s \in \mathcal{S}$ which has been estimated to around 35 for chess. This requires roughly $1.4 \cdot 10^{45}$ bytes, if each number is stored as a single precision floating point number (4 bytes). For comparison the entire digital data capacity in the world is estimated less than 10^{23} bytes as of 2020. Needless to say this is beyond any practical relevance.

The rest of this thesis is therefore about the situations where we have to use approximations and estimations for some or all of P, R and the Q-functions.

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.

It seems to this author, that this setting is not very well-studied in the case of a continuous state space. 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. [7], together with some standard results from approximation theory on artificial neural networks and Bernstein polynomials.

Let us consider any norm $\|\cdot\|$ on the set of Q-functions $\mathcal{Q} = \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$. 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

$$\left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\| \leqslant \varepsilon_k, \forall k \in \mathbb{N} \tag{2.11}$$

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

$$\left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \leqslant \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i \tag{2.12}$$

This is sometimes called the **approximation error** and we denote it

$$\varepsilon_{\text{approx}}(k) := \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i \tag{2.13}$$

Thus we get

Theorem 2.69. 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

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

Proof. By the discussion above and

$$\begin{aligned}\left\|Q^* - \widetilde{Q}_k\right\| &\leqslant \left\|Q^* - T^k\widetilde{Q}_0\right\| + \left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \\ &\overset{2.12}{\leqslant} \gamma^k\left\|Q^* - \widetilde{Q}_0\right\| + \varepsilon_{\text{approx}}(k)\end{aligned}$$

□

The first term in theorem 2.69 is sometimes called the **algorithmic** error[1]. The algorithmic error converges exponentially, so one is usually 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

$$\varepsilon_{\text{approx}}(k) = \varepsilon \frac{1-\gamma^k}{1-\gamma} \leqslant \frac{\varepsilon}{1-\gamma} \tag{2.14}$$

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 2.70. $\sum_{i-1}^k \gamma^{k-i}\varepsilon_i \to 0$ 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.

2.3.2 Using artifical neural networks

Setting 3 (Continuous MDP). An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0,1]^w$, $\mathcal{A}$ finite and a continuous expected reward function r.

Definition 2.71. 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 $(\sigma_i)_{i=1}^L$, weights $(W_i)_{i=1}^{L+1} \in M^{d_i \times d_{i-1}}$ and biases $(v_i)_{i=1}^{L+1} \in \mathbb{R}^{d_i}$ is the function $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$ defined by

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

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 components $\sigma_{ij} : \mathbb{R} \to \mathbb{R}$. We denote the class of these networks (or functions)

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

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

We shall often consider networks with only one type of activation functions, i.e. all activation functions are equal to one function $\sigma : \mathbb{R} \to \mathbb{R}$. We then write $f \in \mathcal{DN}\left(\sigma, (d_i)_{i=0}^{L+1}\right)$ as a shorthand.

Remark 2.72. 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 $W_{i+1}(k,j) \neq 0$. This is because they 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.* Indeed for every suitable collection of activation functions and every $L+1$-partite weighted graph G satisfying

> The ith partition is only connected to the neighboring $(i-1)$th and $(i+1)$th partition. (2.15)

[1] For example in [7].

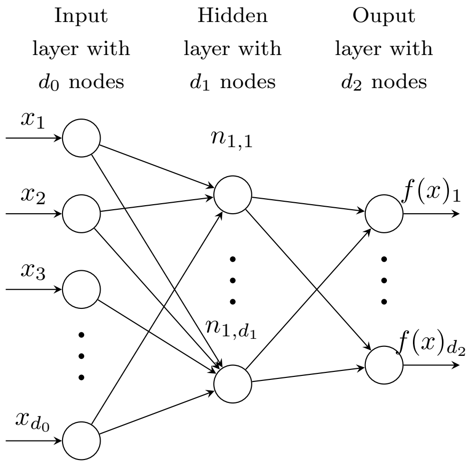

Input layer with d_0 nodes
Hidden layer with d_1 nodes
Ouput layer with d_2 nodes
x_1
x_2
x_3
x_{d_0}
$n_{1,1}$
n_{1,d_1}
$f(x)_1$
$f(x)_{d_2}$

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

Then there exists a corresponding ANN corresponding to G. 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 $f \in \mathcal{DN}\left((\sigma_i)_{i=1}^{L},\ (d_i)_{i=0}^{L+1}\right)$ and $g \in \mathcal{DN}\left((\sigma_i')_{i=1}^{L'},\ (d_i')_{i=0}^{L'+1}\right)$ are two ANNs and $d_{L+1} = d_0'$ (i.e. that the dimensions match $\operatorname{im}(f) \subseteq \operatorname{dom}(g)$) then $g \circ f \in \mathcal{DN}\left((\sigma_1, \ldots, \sigma_L, \sigma_1', \ldots, \sigma_{L'}'), (d_1, \ldots, d_L, d_1', \ldots, d_{L'}')\right)$. By side-by-side alignment we mean the situation where $L = L'$ and one creates the function $h : \mathbb{R}^{d_0+d_0'} \to \mathbb{R}^{d_L+d_L'}$ by defining $h(x_1, \ldots, x_{d_0+d_0'}) = (f(x_1, \ldots, x_{d_0}), g(x_1, \ldots, x_{d_0'}))$. With this way of defining h we have that h is an ANN with structure $(d_0 + d_0', \ldots, d_{L+1} + d_{L+1}')$. Generally any way of stitching together graphs into n-partite graphs satisfying eq. (2.15) will gives ways of producing new ANNs.

Theorem 2.73 (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, unbiased final layer (that is $v_2 = 0$) 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 of theorem 2.73. The original proof in [5, Cybenko (1989)] is very short and elegant, but non-constructive, using the Riesz Representation and Hahn-Banach theorems to obtain a contractiction 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 [4, Chen et al. (1990)] to the statement as presented above and their proof is constructive. □

We will now show how ANNs can be used to approximate the optimal value function to arbitrary precision, and look at a particular class of ANNs called *ReLU networks*, which are defined by their use of the *ReLU activation function* $\sigma_r(x) = \max(0, x)$.

Definition 2.74. We define the class of ReLU networks as the ANNs (see definition 2.71) with all ReLU activation functions, and write $\mathcal{RN}\left((d_i)_{i=0}^{L+1}\right) := \mathcal{DN}\left(\sigma_r, (d_i)_{i=0}^{L+1}\right)$.

Proposition 2.75. Under setting 3 let $\varepsilon > 0$. Assume that either

1. P is deterministic with $P(\cdot \mid s, a) = \delta_{p(s,a)}$. For some continuous $p : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$.

   or

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 $(\widetilde{Q}_i)_{i=1}^{k} \subseteq \mathcal{RN}(w|\mathcal{A}|, N, 1)$ 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.

c. It is possible to join a finite number of ReLU networks $f_{a_1}, \dots, f_{a_a} : \mathcal{S} \to \mathbb{R}$ into a bigger ReLU network $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $f(s, a) = f_a(s)$.

Using these facts we can use the universal approximation theorem (theorem 2.73) 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

$$\left| f_{a,k} - T\widetilde{Q}_{k-1}(\cdot, a) \right| < \varepsilon \tag{2.16}$$

Here $\widetilde{Q}_0 = r$ and $\widetilde{Q}_k$ is obtained recursively by joining for each a the components $f_{a,k}$ into a single network $\widetilde{Q}_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $\widetilde{Q}_k(s, a) = f_{a,k}(s)$. By eq. (2.16) on each of its components approximates $T\widetilde{Q}_{k-1}$ to ε precision. $\widetilde{Q}_0$ is continuous by setting 3 and by a. and b. $T\widetilde{Q}_k$ is as well.

We will now establish points a., b. and c.

a. follows by the fact that that composition of continuous functions are continuous.

b. Let $Q : \mathcal{S} \times \mathcal{A} \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ be continuous and let $x_1, x_2, \dots \in \mathcal{S} \times \mathcal{A}$ with $x_\ell \to x \in \mathcal{S} \times \mathcal{A}$. We will show that under 1. or 2. we have $TQ(x_\ell) \to TQ(x)$.

   1. In this case $TQ(x) = r(x) + \gamma \max_{a' \in \mathcal{A}} Q(p(x), a')$ can be seen as the composition of the continuous functions p, Q, max, + and r.

   2. We have by dominated convergence and the assumption that r is continuous (see setting 3) that

$$\begin{aligned} TQ(x_\ell) &= r(x_\ell) + \gamma \int \max_{a' \in \mathcal{A}} Q(s', a') p(s' \mid x_\ell) \, \mathrm{d}v(s') \\ &\to r(x) + \gamma \int \max_{a' \in \mathcal{A}} Q(s', a') p(s' \mid x) \, \mathrm{d}v(s') \\ &= TQ(x) \end{aligned}$$

c. Let $k \in \mathbb{N}$. To join the components $f_{a,k}$ into a single network we embed $\mathcal{A}$ into $[0,1]^{\mathfrak{a}}$ (where $|\mathcal{A}| = \mathfrak{a}$) by enumerating the actions $\mathcal{A} = \{a_1, a_2, \ldots, a_{\mathfrak{a}}\}$ and putting $a_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ where the 1 is on the ith spot (this is called the *one-hot embedding*). Let L_a, $(d_{a,i})_{i=0}^{L_a+1}$ denote the number of hidden layers and structure of $f_{a,k}$. We can now define $\widetilde{Q}_k : [0,1]^{w+\mathfrak{a}} \to \mathbb{R}$ as the ReLU network with $L = 2 + \max_{a \in \mathcal{A}} L_a$ hidden layers and structure $d_0 = w + \mathfrak{a}$, $d_1 = w \cdot \mathfrak{a}$ and $d_i = \sum_{a \in \mathcal{A}} d'_{a,i-1}$ for $i = 2, \ldots L-1$ putting $d'_{a,i} = d_{a,i}$ for $1 \leqslant i \leqslant L_a - 1$ and $d'_{a,i} = d_{a,L_a}$ for $L_a \leqslant i \leqslant L-1$ then $d_L = \mathfrak{a}$ and finally $d_{L+1} = 1$. The first layer consist of the affine map $w_1 : \mathbb{R}^{w+\mathfrak{a}} \to \mathbb{R}^{w \cdot \mathfrak{a}}$ defined by

$$w_1(s, 0, \ldots, 1, \ldots, 0) = (s, \ldots, s+1, \ldots, s) - 1$$

where $s = (s_1, \ldots, s_w) \in \mathbb{R}^w$ and where we use the notation $1 = (1, \ldots, 1) \in \mathbb{R}^k$ for any $k \in \mathbb{N}$. Applying the ReLU activation $\sigma_r = \max(0, \cdot)$ coordinate-wise we get

$$\sigma_r(w_1(s, 0, \ldots, 1, \ldots, 0)) = (0, \ldots, s, \ldots, 0)$$

since all $s_i \in [0,1]$, so $\max(0, s_i - 1) = 0$ and $\max(0, s_i) = s_i$. We now use the component networks middle part of the network of $\widetilde{Q}_k$. For $2 \leqslant i \leqslant L$ put $w_i = (w_{1,i-1}, \ldots, w_{\mathfrak{a},i-1}) : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ where we define

$$w_{j,i} = \begin{cases} \text{the } i\text{th affine map of } f_{a_j,k} & 1 \leqslant i \leqslant L_{a_j} \\ \text{the identity map: id} : \mathbb{R}^{d_{a_j,i-1}} \to \mathbb{R}^{d_{a_j,i}} & L_{a_j} < i < L \\ \text{the } i\text{th affine map of } f_{a_j,L_{a_j}+1} & i = L \\ \text{summation: } (x_1, \ldots, x_{\mathfrak{a}}) \mapsto \sum_{\ell=1}^{\mathfrak{a}} x_\ell & i = L+1 \end{cases}$$

With this construction we have that $\widetilde{Q}_k(s,a) = f_{a,k}(s)$ for all $a \in \mathcal{A}$. And that $\widetilde{Q}_k(s,a) \in \mathcal{RN}(w + \mathfrak{a}, w \cdot \mathfrak{a}, d_2, \ldots, d_{L-1}, d_{\mathfrak{a}}, 1)$.

□

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. However it is still not clear if this method is feasible computationally. To investigate this and indeed for any chance to implement the method in practice one would need to go through the construction in [4]. 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 we need a stronger form of continuity, namely Lipschitz continuity (see definition A.17), 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 $\|\cdot\|$-Lipschitz with constants L_r, L_p respectively for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ for some norm $\|\cdot\|$.

Definition 2.76 (Bernstein polynomial)**.** The (multivariate) Bernstein polynomial $B_{f,n}$ of degree $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ approximating the function $f : [0,1]^w \to \mathbb{R}$ is defined by

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

Notice that this a polynomial of (multivariate) degree $\|n\|_1 = n_1 + \cdots + n_w$.

Theorem 2.77 (Approximation with Bernstein polynomials)**.** Let $f : [0,1]^w \to \mathbb{R}$ be Lipschitz w.r.t. the standard euclidean 2-norm induced metrics on $[0,1]^w$ and $\mathbb{R}$ with constant L. Let $n = (n_1, \dots, n_w) \in \mathbb{N}^w$. The Bernstein polynomial $B_{f,n} : [0,1]^w \to \mathbb{R}$ satisfies

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

Proof. We refer to [9, Heitzinger (2002)] thm. B..7. □

Lemma 2.78. 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 measurable $Q : \mathcal{S} \times \mathcal{A} \to [-V_{\max}, V_{\max}]$.

Proof. Because of the Lipschitz property of r and p we have for any measurable $Q : \mathcal{S} \times \mathcal{A} \to [-V_{\max}, V_{\max}]$ and $s \neq s' \in \mathcal{S}$ that

$$\begin{aligned}
\left|TQ(s,a) - TQ(s',a)\right| \leqslant{}& \left|r(s,a) - r(s',a)\right| \\
&+ \gamma \int \left| \max_{a' \in \mathcal{A}} Q(s'', a') p(s'' \mid s, a) - \max_{a'' \in \mathcal{A}} Q(s', a'') p(s'' \mid s', a) \right| \, \mathrm{d}\mu(s'') \\
&\leqslant L_r \|s - s'\| + \gamma \int \left| \max_{a' \in \mathcal{A}} Q(s', a') \right| \left| p(s'' \mid s, a) - p(s'' \mid s', a) \right| \, \mathrm{d}\mu(s'') \\
&\leqslant L_r \|s - s'\| + \gamma \int V_{\max} L_p \|s - s'\| \, \mathrm{d}\mu(s'') \\
&= (L_r + \gamma V_{\max} L_p) \|s - s'\|
\end{aligned}$$

□

Now we can bound

Proposition 2.79. Given an MDP satisfying setting 4 and using $\|\cdot\|_\infty$ we can bound

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

Proof. Following the procedure from leading to eq. (2.12), starting with $\widetilde{Q}_0 = 0$ and using the $n = (n_1, \dots, n_w)$ degree Bernstein polynomium $\widetilde{Q}_k = B_{T\widetilde{Q}_{k-1}, n}$ as approximation for $T\widetilde{Q}_{k-1}$ we know by induction and the results lemma 2.78 and theorem 2.77.2 that $T\widetilde{Q}_k$ is L_T-Lipschitz for any $k \in \mathbb{N}$. Now by choosing the euclidean norm $\|\cdot\| = \|\cdot\|_2$ we have by theorem 2.77.1 that

$$\varepsilon_i = \left\|\widetilde{Q}_k - T\widetilde{Q}_{k-1}\right\|_\infty \leqslant \frac{L_T}{2} \sqrt{\sum_{j=1}^{w} \frac{1}{n_j}} = \varepsilon \tag{2.17}$$

where ε is the one-step error defined in eq. (2.11). Now by eq. (2.14) we have that

$$\varepsilon_{\text{approx}} \leqslant \frac{\varepsilon}{1-\gamma} \tag{2.18}$$

Combining eq. (2.17) and eq. (2.18) and noting that $L_T = L_r + \gamma V_{\max} L_p$ finishes the proof. □

To make more clear what are the implications of proposition 2.79 we give a corollary where we put $n_j = m$ for all $j \in [w]$:

Corollary 2.80. Under setting 4 and using Bernstein polynomials of degree $n = (m, \ldots, m) \in \mathbb{N}^w$ for $m \in \mathbb{N}$ we have the following bound

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

In particular $\left\| Q^* - \widetilde{Q}_k \right\|_\infty = \mathcal{O}(\gamma^{-k} + \frac{1}{\sqrt{m}})$ when using k iterations.

Proof. Use proposition 2.79. □

This gives a very concrete way of constructing an arbitrarily good approximation to Q^* using polynomials. A major drawback is the restriction on the transition dynamics P. For example we cannot use corollary 2.80 on deterministic decision processes, since if P is deterministic then there are no measure $\mu \in \mathcal{P}(\mathcal{S})$ which allows for a density $p(\cdot \mid s, a)$ (i.e. $p \cdot \mu = P(\cdot \mid s, a)$), unless $P(\cdot \mid s, a) = \delta_{s'}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, which would lead to a quite boring environment. Generally the processes with fast convergence bounds according to corollary 2.80 must be very stochastic.

This concludes our investigation of model-based Q-learning and we will now look at the much more well-studied field of model-free Q-learning.

Chapter 3

Model-free algorithms

In this section we will look at what can be done when the process dynamics are not directly accesible, but also has to be learned from sampling. This means that we do not have access to the distributions of the transition kernel $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ and the reward kernel $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathbb{R}$, Therefore several functions that were directly available previously, such as the expected reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and the Q-function operators on P_τ, T_τ and T, now has to be estimated from samples.

Motivation

Model-free algorithms present very active field of research and cutting-edge research groups such as *Google DeepMind* are notoriously interested in removing model dependencies from their algorithms. This interest means that many newly developed empirically succesful algorithms such as DQN are model-free.

Why put such a restriction on the algorithms? There are multiple reasons. Firstly one might be in a situation where the process dynamics are actually unknown, such as the environment of the stock market where the agent is an investor. Here probabilities of next states depend on real world events and decisions of a high number of people and other algorithms. It seems hard to model such an environment in an exact way, and therefore the idea of designing an algorithm that works without a precise description of its environment is attractive. Secondly it is practical to have algorithms that works *out of the box* without the need for adaptions to each different model. With model-free algorithms one only needs access to sampling methods from the environment, while it may require more *tingering* to make model-dependent processes work. Thirdly there is a philosophical aspect to model-free algorithms of approaching human-like intelligent behavior, in the sense of being able to cope with every environment with the same *algorithm* (not saying that any algorithm has come very far in achieving such a human-like behavior yet).

A final reason for basing algorithms on sampling is the general advantage of Monte Carlo methods that they can lead to faster computation and are more easily applied to complex systems than analytical approaches.

Problems with applying model dependent solutions

It is clear that in the model-free setting line 3 will not work without modification, because we have not access to the distributions of R and P, except through sampling. To make the scheme work anyway we could simply avoid taking expectations and use the random outcomes of the kernels by using the update step

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

where $r' \sim R(\cdot \mid s,a)$ and $s' \sim P(\cdot \mid s,a)$ are sampled. This can be viewed as a stochastic version of the T-operator update step of the Q-iteration. Used naively this leads to the following algorithm

```
Algorithm 4: Random theoretical Q-iteration (example of thought)
  Input: MDP $(\mathcal{S},\mathcal{A},P,R,\gamma)$, number of iterations $K$
1 For all $(s,a) \in \mathcal{S} \times \mathcal{A}$ let $\widetilde{Q}_0(s,a)$ be sampled from $R(\cdot \mid s,a)$.
2 for $k = 0,1,2,\dots,K-1$ do
3     For all $(s,a) \in \mathcal{S} \times \mathcal{A}$ sample a reward $r'$ from $R(\cdot \mid s,a)$, a next-state $s'$ from $P(\cdot \mid s,a)$
        and let
          $\widetilde{Q}_{k+1}(s,a) \leftarrow r' + \gamma \sup_{a' \in \mathcal{A}} \widetilde{Q}_k(s',a')$
4 Define $\pi_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
  Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\pi_K$
```

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 with 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 thesis.

Generally about this section

Sections 3.1 and 3.2 are mostly surveying different results of model-free algorithms without any proofs, while section 3.3 contains a detailed description of the recent result on deep fitted Q-iteration ([7, Fan et al. (2020+)]) including proofs. This way the purpose of sections 3.1 and 3.2 is to provide context for section 3.3 and a general idea of the field of theoretical reinforcement learning.

We start by looking at the finite case, where most and strongest results are available.

3.1 Finite case

Setting 5 (Finite, discounted, action-unrestricted MDP)**.** A (discounted) 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. Also we assume unrestricted actions, i.e. that the set of admissable actions $\mathfrak{A}(s) = \mathcal{A}$ is the same for all $s \in \mathcal{S}$.

Remark 3.1. As discussed in example 2.60 the strong results from model-based Q-learning hold under setting 5, including

1. fulfillment for the Bellman optimality equation $TQ^* = Q^*$ (proposition 2.67.3),

[1] This classification is discussed in [12, Kearns and Singh (1999)].

2. the existence of an deterministic stationary optimal $\pi^* \in DS\Pi$ which is greedy for $Q^* = Q_{\pi^*}$ (proposition 2.57 and proposition 2.67.4),

3. convergence of $T^k Q \to Q^*$ exponentially in γ for any bounded measurable Q-functions $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (proposition 2.67.3).

Of course being in the model-free setting this does not provide any solutions yet.

A popular indirect approach, and the basis of most of the model-free algorithms we will discuss, is called *temporal difference* (TD) learning and was invented for the setting of finite MDPs. The idea has since been used in continuous settings as well, as we will we in the next sections. 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 called the **learning rate** (of the kth step). Equation (3.1) can be viewed as the linear interpolation moving $\widetilde{Q}_k(s,a)$ toward

$$y = r' + \gamma \max_{a' \in \mathcal{A}} \widetilde{Q}_k(s',a')$$

with the weight α_k. We name the term y the (sampled) **T-value** of the pair (s,a). This is due to the similarity between y and $T\widetilde{Q}_k(s,a) = r(s,a) + \gamma \int \max_{a' \in \mathcal{A}} \widetilde{Q}_k(s',a')\, \mathrm{d}P(s' \mid s,a)$. Another important term in eq. (3.1) is the **temporal difference** $\alpha_k(r' + \gamma \cdot \max_{a \in \mathcal{A}} \widetilde{Q}_k(s',a') - \widetilde{Q}_k(s,a))$ occuring from a simple rearrangement.

TD learning addresses the problem of unstable updates due to stochastic rewards, by its use of interpolation combined with a learning rate which is usually fixed before running the algoritm and is set to decay from 1 to 0 in some fashion as $k \to \infty$.

Algorithm

There are many variations of TD learning algorithms. We will here look at the *finite asynchronos Q-learning* algorithm which is based on updating the Q-function estimators one state-action pair at a time using the TD update step.

Algorithm 5: Finite asynchronos Q-learning

```
Input: Finite MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$, state-action pairs
       $(s_1, a_1, \dots, s_K, a_K)$, learning rates $(\alpha_1, \dots, \alpha_K)$, initial $\widetilde{Q}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
for $k = 1, 2, \dots, K$ do
    Sample $r' \sim R(\cdot \mid s_k, a_k)$, $s' \sim P(\cdot \mid s_k, a_k)$.
    Update action-value function: For all $(s,a) \in \mathcal{S} \times \mathcal{A}$ let
        $\widetilde{Q}_k(s,a) \leftarrow \begin{cases} \widetilde{Q}_{k-1}(s,a) & (s,a) \neq (s_k,a_k) \\ (1-\alpha_k)\widetilde{Q}_{k-1} + \alpha_k(r' + \gamma \max_{a' \in \mathcal{A}} \widetilde{Q}_{k-1}(s',a')) & (s,a) = (s_k,a_k) \end{cases}$
Define $\widetilde{\pi}_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\widetilde{\pi}_K$
```

Results

The convergence result for the finite asynchronos Q-learning algorithm which we will now present was originally obtained in [23, Watkins and Dayan (1992)] of a TD algorithm using Q-functions. The result was extended slightly in [10, Jaakkola et al. (1994)] and is here presented more in the style of [10].

Theorem 3.2 (Watkins, Dayan 1992)**.** Let $s_1, a_1, s_2, a_2, \dots \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A} \times \dots$ be random variables and $\alpha_1, \alpha_2, \dots \in [0,1]$ be a sequence of learning rates. The output $\widetilde{Q}_K$ of the finite asynchronos Q-learning algorithm converges to Q^* provided that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ it holds that

1. $\mathbb{P}\left(\sum_{i=1}^{\infty} \alpha_i[(s_i, a_i) = (s,a)] = \infty\right) = 1$ and $\mathbb{P}\left(\sum_{i=1}^{\infty} \alpha_i^2[(s_i, a_i) = (s,a)] < \infty\right) = 1$.
2. $\mathbb{V}(R(\cdot \mid s,a)) < \infty$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Here $[(s_i, a_i) = (s,a)]$ is the Bernoulli random variable with $\mathbb{P}([(s_i,a_i) = (s,a)] = 1) = \mathbb{P}(s_i = s \text{ and } a_i = a)$.

Remark 3.3. In the formulation in [10] the sums of learning rates were supposed to converge *uniformly* (see definition A.13). However this is equivalent to this formulation because of the fact that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ we have that $\mathbb{P}(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f_n(s,a)| \to 0) = 1 \iff \mathbb{P}\left(|f_n(s,a)| \to 0\right) = 1$. 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 since $\mathbb{V}(R(\cdot \mid s,a)) \leqslant \mathbb{E}(2R_{\max})^2 = 4R_{\max}^2$.

Theorem 3.2 establishes our first convergence guarantee of Q-learning in a model-free setting. In a special case of the same setup, asymptotical almost sure convergence rates where established by [20, Szepesvári (1997)]:

Theorem 3.4 (Szepesvári)**.** Under setting 5 using the finite asynchronos algorithm let $K \in \mathbb{N}$ and $(s_1, a_1), (s_2, a_2) \dots, (s_K, a_K)$ be sampled i.i.d. from a distribution $p \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ with $\operatorname{supp}(p) = \mathcal{S} \times \mathcal{A}$, i.e. all state-action pairs have non-zero probability of occuring. Set the learning rates such that $\alpha_k = |\{i \in [k] \mid (s_i, a_i) = (s_k, a_k)\}|^{-1}$, i.e. they are 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[2]

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

and

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

For the output $\widetilde{Q}_K$ of the finite asynchronos Q-learning algorithm with K-steps.

Remark 3.5. In theorem 3.4 eq. (3.2) is tightest when $\gamma > 1 - \beta/2$ otherwise eq. (3.3) is tighter.

This concludes our section on finite model-free MDPs. We note that we have only covered a tiny fraction of the litterature on this topic. One source that was also considered, but did not make it into this thesis is [6, Even-dar and Mansour (2001)] which establishes PAC-learnability of a closely related *synchronos* finite Q-learning algorithm, and provides some theoretical justification for picking learning rates decreasing as $n^{-0.85}$, which has since become a popular choice.

[2] For the definition of asymptotical almost certainty see definition A.14 and example A.15.

3.1.1 History dependent setting

Staying in setting of decision processes with finite states and actions, we will now turn to the setting of history dependent processes, and present a result by [14, Majeed and Hutter (2018)].

Setting 6 (Finite HDP)**.**

1. A history dependent decision process (see definition 2.5), 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.

Under this setting we will write elements $h_k = (s_1, a_1, s_2, \ldots, a_{k-1}, s_k) \in \mathcal{H}_k$ as strings $h_k = s_1a_1s_2 \ldots a_{k-1}s_k$ for convenience.

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 $r : \mathcal{H}^* \times \mathcal{A} \to \mathbb{R}$. This is discounted by $\gamma \in [0, 1)$, that is $r(h_n, a) \in [-\gamma^{n-2}R_{\max}, \gamma^{n-2}R_{\max}]$ for any $h_n \in \mathcal{H}_n$ and $a \in \mathcal{A}$.

4. Actions are unrestricted, so that $\mathfrak{A}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$.

Remark 3.6. This setting can be analysed with the tools we developed in chapter 2: 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. Discounting with $\gamma \in [0, 1)$ implies setting 1.5. Further the conditions (S) and (W) of Schäl are also both implied by the discreteness. This implies by theorem 2.34 the existence of an optimal $\pi^* \in R\Pi$ and that $V_n^* \to V^*$. These remarks was not discussed in [14].

Within setting 6 we introduce some additional concepts extending what we have previously defined in the context of decision processes.

History based Q-functions

So far we have only discussed Q-functions for MDPs, where they are functions from $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Within setting 6 Q-functions are generalized so that they are taking values in $\mathcal{H}^* \times \mathcal{A}$. Likewise the T-operator is generalized by defining

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

The optimal Q-function Q^* is defined in [14] as the fixed point of the T operator in $\mathcal{L}_\infty(\mathcal{H}^* \times \mathcal{A})$. It is not discussed in [14] why this is well-defined.

Partial observability

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 $x_n = \phi(h_{n-1}a_{n-1}s_n)$ is the state s_n as it is perceived by the agent. This is called **partial observability**. ϕ is a assumed to be surjective. In applications this could be a partially observable environment or a latent space. Using ϕ 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 $\phi = \rho_{\mathcal{S}}$ where $\rho_{\mathcal{S}}$ is projection onto the last state space.

Now kernel for the observable process is defined for each $h \in \mathcal{H}^*$

$$\begin{aligned} p_h &: \mathcal{A} \rightsquigarrow \mathcal{X} \\ p_h(x' \mid a) &= \sum_{s:\phi(has)=x'} P(s \mid ha) \end{aligned}$$

Remark 3.7. Another way of stating this which was not considered in [14] is that p_h can be expressed as the image measure $p_h(\cdot \mid a) = \phi_{ha}(P(\cdot \mid ha))$, where we define $\phi_{ha}(s) = \phi(has)$.

We can now consider yet another kind of Q-functions defined on $\mathcal{X} \times \mathcal{A} \to \mathbb{R}$, which we will call *partial* to avoid confusion. The optimal partial Q-function q_h^*, naturally dependent on a history $h \in \mathcal{H}^*$ and is defined by the equation

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

Again in [14] it is left as an exercise to the reader to ponder if this is well defined. Now a central assumption is introduced:

Assumption 4 (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 under setting 6 together with the state-uniformity condition is by [14] called a *Q-Value uniform decision process* (QDP). The justification for this is the following theorem by Hutter:

Theorem 3.8 (Hutter, 2016)**.** Under assumption 4 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 3.9. Within setting 6 assume

1. State-uniformity (assumption 4).
2. Any state is reached eventually under any policy (called *state-process ergodicity* in [14]).
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 Q^*.

It seems relevant to ask how restrictive the state-uniformity assumption is. [14] answers this by an array of examples showing the following relations of the classes of decision processes:

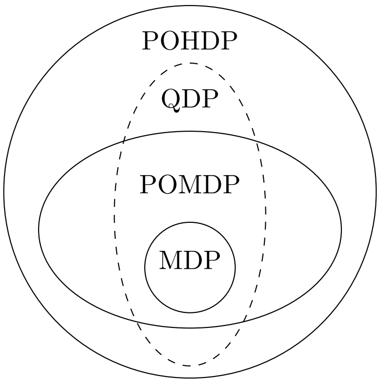


Figure 3.1: Classes of finite decision processes considered in [14], under setting 6 (recall that QDP is a partially observable HDP under state-uniformity assumption 4).

Remark 3.10. It remains unclear after reading [14] why q^* and Q^* are 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 2.24) of a general HDP? A sensible thing to ask would be that $Q^*(h,a) = r(h,a) + \gamma \mathbb{E}_{P(\cdot|ha)} V^*$. An analysis of this could have been made by generalizing the results on Q-functions of section 2. The main discussion of Q-functions in this thesis have been aimed at MDPs since most results are in this setting, and we will not go further into these details. Alternatively a further study into the articles of Hutter (and associated authors) might also gives such insights.

With this we conclude our section on finite decision processes and turn to processes with continuous state spaces.

3.2 Linear function approximation

In this section we will look at a general method of approximation of Q-functions, namely approximation by a linear span from a set of basis functions. This was investigated by [15, Melo and Ribeiro (2007)] on which is section is based.

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

1. An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (see definition 2.36).

2. $\mathcal{S} \subseteq \mathbb{R}^w$ is a compact subset of a euclidean space.

3. $\mathcal{A}$ is finite and unrestricted, that is $\mathfrak{A}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$.

4. r_i is upper semicontinuous .

Remark 3.11. Item 4 was actually not part of the assumptions in [15]. We include it here in order to ensure the existence of an optimal policy and thus measurability of V^*.

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

$$Q_\theta(s,a) = \sum_{i=1}^M \theta_i \xi_i(s,a) = \xi^T \theta \tag{3.7}$$

where we view $\xi = (\xi_1, \dots, x_M) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^M$ as the combined *vector* function, and ξ^T is transposition so that $\xi^T \theta$ is another way of writing the standard inner (dot) product.

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 $q^* = \rho_{\mathcal{F}} Q^*$ where $\rho_{\mathcal{F}}$ is the orthogonal projection on $\mathcal{F}$. Denote by θ^* the coordinates of this projection, i.e. $q^* = Q_{\theta*} = \rho_{\mathcal{F}} Q^*$.

It is easily seen that the gradient of Q_θ over θ is $\nabla_\theta Q_\theta = \xi$. This gives the idea for a temporal difference term with approximation from $\mathcal{F}$ using the update step. Let $(s_1, a_1, s_2, a_2, \dots) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \dots$ be states and actions sampled from an decision process when we could use the following update on our parameters:

$$\theta_{k+1} = \theta_k + \alpha_k \xi(s_k, a_k) \left(r_k + \gamma \max_{b \in \mathcal{A}} Q_{\theta_k}(s_{k+1}, b) - Q_{\theta_k}(s_k, a_k) \right) \tag{3.8}$$

Algorithm 6: Q-learning with linear approximation

```
Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, policy $\pi \in R\Pi$, 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 $\widetilde{\pi}_K$
```

The policy given as input to this algorithm, may depend on the history and should be seen as a way of sampling from $\mathcal{A}$, rather than a effective strategy. The main point in this section is to show convergence, so we are interested in a policy providing sufficient support of the distributions of the samples (s_i, a_i) across $\mathcal{S} \times \mathcal{A}$. This is made precise by the assumptions of *ergodicity* on the process.

We can view an MDP as a stationary process $\mathfrak{P}$ on $\mathcal{S}$ generated by kernel $P\pi$ for a policy $\pi \in S\Pi$. This makes sense to the property of ergodicity for the process which can be viewed as the continuous state-space equivalent of the requirement that every state and action is visited infinitely often in the finite MDPs. For a full definition of *geometric ergodicity*, which we will need in the following main result from [15], see section A.1.1.

Theorem 3.12 (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[3] 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 line 4 is run with learning rates from a sequence $\{\alpha_k\}_{k \in \mathbb{N}}$ satisfying $\alpha_k \in [0, 1]$ and

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

[3] See section A.1.1.

we have that

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

with probability 1, and Q_{θ^*} satisfies

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

Furthermore the orthogonal projection is expressible as

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

Remark 3.13. Recall the definition of the kernel-derived measure $\pi\mu(S \times A) = \int_S \pi(A \mid s) \, \mathrm{d}\mu(s)$ (see theorem 2.7).

The theorem 3.12 by Melo and Ribeiro shows that Q-learning still works when using a gradient step version of the temporal difference update in a continuous state space setting, and it guarantees convergence to optimality within the approximating function class. However there is still room for improvement since theorem 3.12 does not tell us:

1. How fast is the convergence? 2. How far is Q_{θ^*} from Q^*? 3. How far is $Q_{\widetilde{\pi}_K}$ from Q^*?

Remark 3.14. 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 3.26).

3.3 Deep fitted Q-iteration

3.3.1 Introduction

This section is about the results of [7, Fan et al. (2020+)], which we will present, discuss and prove. 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 [7] claims to investigate the deep Q-network algorithm, instead of analysing DQN, another called *deep fitted Q-iteration* (DQI) algorithm is analysed instead and bounds on its convergence is established.

We begin by presenting the general *fitted Q-iteration* (FQI) algorithm on which DQI is based:

```
Algorithm 7: Fitted Q-Iteration Algorithm
Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, function class $\mathcal{F}$, sampling distribution $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$,
       number of iterations $K$, batch-size $n$, initial estimator $\widetilde{Q}_0$
for $k = 0, 1, 2, \dots, K-1$ do
    Sample $n$ times independently from the distribution $\nu$ to get the batch $(S_i, A_i)_{i \in [n]}$.
    For each $i \in [n]$ sample a reward $R_i \sim R(S_i, A_i)$ and a next-state $S'_i \sim P(S_i, A_i)$.
    From this define the T-values $Y_i \leftarrow R_i + \gamma \cdot \max_{a \in \mathcal{A}} \widetilde{Q}_k(S'_i, a)$
    Update action-value function by solving the least squares optimization problem
    $\widetilde{Q}_{k+1} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n (Y_i - f(S_i, A_i))^2$ over the function class $\mathcal{F}$.
Define $\widetilde{\pi}_K$ 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 $\widetilde{\pi}_K$
```

Remark 3.15. Rather than a single algorithm we may view FQI as a class of concrete, because (1.) The function class is not specified. (2.) It is not specified how to solve the optimization problem in line 5. These two points are linked in these that the optimization problem is probably better handled separately for each function class.

The deep fitted Q-iteration algorithm is an FQI-algorithm where the function class $\mathcal{F}$ is a particular class of artificial neural networks which we will define now.

ReLU Networks

Let $f \in \mathcal{RN}\left((d_i)_{i=1}^{L+1}\right)$ be a ReLU network (see definition 2.74) with weights $(W_i, v_i)_{i=1}^{L+1}$ and define $\widetilde{W}_i = (W_i, v_i)$, as the vector containing all weight and bias parameters of the ith layer of f, that is, all entries in the matrix W_i and vector v_i. We can then consider the magnitude of the maximum parameter $\left\|\widetilde{W}_i\right\|_\infty$, and the number of non-zero parameter $\left\|\widetilde{W}_i\right\|_0$ in the ith layer of the network. Also denote by $(f_j)_{j \in d_{L+1}} = f$ the components (coordinates) of the network f.

Definition 3.16 (Sparse ReLU networks)**.** For $s, V \in \mathbb{R}$ the ReLU network f is called (s, V)-**sparse** if

$$1.\ \max_{\ell \in [L+1]} \left\|\widetilde{W}_\ell\right\|_\infty \leqslant 1 \qquad 2.\ \sum_{\ell=1}^{L+1} \left\|\widetilde{W}_\ell\right\|_0 \leqslant s \qquad 3.\ \max_{j \in [d_{L+1}]} \left\|f_j\right\|_\infty \leqslant V$$

The set of them we denote $\mathcal{SRN}\left(s, V, (d_i)_{i=0}^{L+1}, L\right)$ and by $\mathcal{SRN}(s, V)$ we mean the set of (s, V)-sparse ReLU networks with any (finite) structure. We may leave out L when clear from the structure writing $\mathcal{SRN}\left(s, V, (d_i)_{i=0}^{L+1}\right)$.

Remark 3.17. Following the graph interpretation of ANNs (see remark 2.72) 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.

Definition 3.18 (Deep fitted Q-iteration)**.** A **deep fitted Q-iteration** (DQI) algorithm, is the fitted Q-iteration algorithm when applied with a function class of sparse ReLU networks $\mathcal{SRN}$.

The reason for working with this particular subclass of neural networks is due to the following lemma found in [18, Schmidt-Hieber (2017)] p. 22 (we have not yet defined Hölder smooth functions. For this see definition 3.20).

Lemma 3.19 (Approximation of Hölder Smooth Functions by ReLU networks)**.** Let $m, M \in \mathbb{Z}_+, \beta > 0$ and $H > 0$ 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 $f \in \mathcal{SRN}\left(s, \infty, (d_j)_{j=0}^{L+1}\right)$ 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 [7] 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.

3.3.2 Assumptions

Before we present the main result of [7] we will first properly state the rather intricate assumptions that it requires.

Hölder Smoothness

According to [7] (def. 2.2) the following property is *widely used as to characterize regularity of functions.*

Definition 3.20 (Hölder smoothness)**.** Let $\mathcal{S}$ be subset of $\mathbb{R}^w$ with non-empty interior $\mathcal{S}^\circ \neq \varnothing$ (see definition A.1), $\beta > 0$ be a real number, $k = \lfloor \beta \rfloor \in \mathbb{N}_0$ and $f : \mathcal{S} \to \mathbb{R} \in C^k$ be a k times continuously differentiable function (see definition A.21). Define the **Hölder smooth norm** of f by

$$\|f\|_{C_w} := \sum_{|\alpha| < \beta} \|\partial^\alpha f\|_\infty + \sum_{\|\alpha\|_1 = \lfloor \beta \rfloor} \sup_{\substack{x \neq y \\ x,y \in \mathcal{S}^\circ}} \frac{|\partial^\alpha (f(x) - f(y))|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \tag{3.9}$$

where $\alpha = (\alpha_1, \ldots, \alpha_w) \in \mathbb{N}_0^w$. 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

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

With this we can define the criteria we are actually interested in using

Definition 3.21. For any $j \in [q]$ let $t_j, p_j \in \mathbb{N}$, $t_j \leqslant p_j$ and $H_j, \beta_j > 0$. We say that $f : [a_1, b_1]^{p_1} \to \mathbb{R}$ is a **composition of Hölder smooth functions** when

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

for some functions $g_j : [a_j, b_j]^{p_j} \to [a_{j+1}, b_{j+1}]^{p_{j+1}}$ (where $p_{q+1} = 1$) that only depend on t_j of their inputs for each of their components $g_{jk} : [a_j, b_j]^{p_j} \to [a_{j+1}, b_{j+1}]$, 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

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

Example 3.22. We have for example $\mathcal{G}(w, w, \beta, H) = C_w([a_1, b_1]^w, \beta, H)$.

Definition 3.23. Define

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

and

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

Here $\mathcal{SRN}(s, V)$ denotes the set of (s, V)-sparse ReLU networks with any possible network structure.

The class $\mathcal{F}_0$ is the function class we are going to use in the version of the DQI algorithm, for which we will soon establish convergence bounds.

In order to make sense of the first assumption (assumption 5) which we are going to present shortly we recall here the definition of the operators for Q-functions: (definition 2.63). 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 [7].

Assumption 5. It is assumed $T\mathcal{F}_0 \subseteq \mathcal{G}_0$. 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 5 holds: Seting $\mathcal{D} = [0,1]^r$, $q = 1$ and taking both the expected reward function and transition kernel to be Hölder smooth.

Example 3.24. 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 $C_w([0,1]^w, \beta, H)$. Then

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

when $q = 1, p_1 = w, t_1 = w, \beta_1 = \beta$ and $H_1 = (1+\gamma V_{\max})H$. 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

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

Concentration coefficients

In analysing DQI we will work with two distributions (measures) on $\mathcal{S}\times\mathcal{A}$. The first measure ν is the batch sampling distribution used in the FQI algorithm (line 2). The other μ is used to measure the distance to the optimal Q-function Q^* from the algorithm output $\widetilde{Q}_K$. The next assumption has to do with the difference between these two measures.

Since we are in the setting of an MDP $(\mathcal{S},\mathcal{A},P,R,\gamma)$ recall from chapter 2 that for a sequence of stationary policies $\pi_1, \pi_2, \dots \in S\Pi$ and a measure $\mu \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$ we can get a probability measure $\dots\pi_2 P \pi_1 P \in \mathcal{P}(\mathcal{H}_\infty)$ where $\mathcal{H}_\infty = (\mathcal{S}\times\mathcal{A})^\infty$. Let $\rho_m : \mathcal{H}_\infty \to \mathcal{S}$ denote projection onto the mth state-action pair in $\mathcal{H}_\infty$. Then using the alternative kernel composition (see remark 2.11) we can write the distribution of the mth state action pair as

$$\rho_m(\dots\pi_2 P \pi_1 P \mu) = (\pi_{m-1}P) \circ \cdots \circ (\pi_1 P) \in \mathcal{P}(\mathcal{S})$$

Definition 3.25 (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

$$\kappa(m, \nu_1, \nu_2) = \sup_{\pi_1,\ldots,\pi_m \in M\Pi} \left[\mathbb{E}_{v_2} \left(\frac{\mathrm{d}((P\pi_m) \circ \cdots \circ (P\pi_1)\nu_1)}{\mathrm{d}\nu_2} \right)^2 \right]^{1/2}$$

where $\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}$ are the Radon-Nikodym derivative of the measures μ_1, μ_2 (see theorem A.23).

Assumption 6. For two probability measure $\nu, \mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ on $\mathcal{S} \times \mathcal{A}$ it is assumed that there exists a finite constant $\phi_{\mu,\nu} > 0$ such that

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

We are not going further into examples of when this assumption holds or the size of the constant $\phi_{\mu,\nu}$. Below assumption 4.3 in [7] are found references to detailed discussions of assumption 6.

3.3.3 The main theorem

The main result of this section and one of the main results of [7] is

Theorem 3.26 (Fan, Yang, Xie, Wang)**.** Let setting 7 hold and let $\nu, \mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be probability measures on $\mathcal{S} \times \mathcal{A}$ such that assumption 6 hold with constant $\phi_{\mu,\nu} > 0$. We will use the FQI with the following inputs:

1. ν as the batch sampling distribution.
2. $n \in \mathbb{N}$ a sufficiently large[4] number serving as the batch size.
3. approximating function class $\mathcal{F}_0$ (see definition 3.23).

We assume that there exists constants $q \in \mathbb{N}$ and $\{p_j, t_j, \beta_j, H_j\}_{j \in [q]}$ such that assumption 5 hold and that furthermore there exists 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

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

with structure satisfying

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

such that when running the FQI algorithm, its output satisfy

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

here $C_\varepsilon, C_{\widetilde{L}}, C_{\widetilde{s}} > 0$ are constants not depending on n or K. In other words

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

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

[4] We will elaborate on this in the proof.

This bound on the convergence of the FQI algorithm is quite remarkable in terms of class of environments 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 [16, Mnih et al. (2015)]. Note that we denote by $Q(\theta)$ the network with parameters θ, i.e. $Q(\theta)$ is a function.

```
Algorithm 8: Deep Q-Network
Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$, batch size $n$, exploration factor $\epsilon$,
       function class $\mathcal{F}$ of deep neural networks parametrized by some $\Theta \subseteq \mathbb{R}^D$, $D \in \mathbb{N}$,
       target update frequency $T_{\text{target}}$, 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 $k = 0, 1, 2, \dots, K-1$ do
    With probability $\epsilon$ 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
        $$\theta_{k+1} \leftarrow \theta_k - \alpha_k \frac{1}{n} \sum_{i=1}^{n} (Y_i - Q(\theta_k)(s_i, a_i)) \cdot \nabla_\theta Q(\theta)(s_i, a_i)$$
    For every $T_{\text{target}}$ steps update the target network by setting $\theta^*_{\ell(k+1)} \leftarrow \theta_{k+1}$ 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 $\tau^*$.
```

DQN 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 $\widetilde{Q}_k$.

[7] 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 [16] it holds $\sim 10^6$ transitions. Because of this it is argued in [7] 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.

3.3.5 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.

Differences in notation

Because σ is used ambigously in theorem 3.26 we denote the probability distribution σ from [7] p. 20 by ν instead. I avoid the shorthand defined in [7] 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 [7] to denote the euclidean dimension of the state space, while here we use w. We use $\mathcal{SRN}(s, V, (d_j)_{j=0}^{L+1}, L)$ to denote the class of sparse ReLU networks while [7] use $\mathcal{F}(L, \{d_j\}_{j=0}^{L+1}, s, V)$.

3.3.6 Proofs

The proof of theorem 3.26 combines two results. The first on the error propagation and the second on the error ocurring in a single step.

Theorem 3.27 (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

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

It is in the second theorem (3.28) on the one-step approximation error that we deviate slightly from [7] by correcting some mistakes. We will note these deviations as they occur in the proof (see remark 3.42, and remark 3.46).

Theorem 3.28 (One-step Approximation Error)**.** Let

1. $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be an MDP.
2. $\mathcal{F} \subseteq \left\{f : \mathcal{S} \times \mathcal{A} \to [-V_{\max}, V_{\max}] \mid f \in \mathcal{M}\right\}$ be any set of measurable functions on $\mathcal{S} \times \mathcal{A}$ bound above and below by $V_{\max}$.

3. $\mathcal{G} = T(\mathcal{F})$ the class of functions obtainable by applying T to some function in $\mathcal{F}$.

4. $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be a probability measure.

5. $(S_i, A_i)_{i\in[n]}$ be n i.i.d. samples with distribution $(S_i, A_i)(\mathbb{P}) = \nu$.

6. $(R_i, S_i')_{i\in[n]}$ be the rewards and next states sampled corresponding to the samples. That is $(S_i, A_i, R_i)(\mathbb{P}) = R\nu$ and $(S_i, A_i, S_i')(\mathbb{P}) = P\nu$.

7. $Q \in \mathcal{F}$ be a fixed Q-function approximator.

8. $Y_i = R_i + \gamma \max_{a\in\mathcal{A}} Q(S_i', a)$.

9. $\widehat{Q} = \operatorname{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(S_i, A_i) - Y_i)^2$.

10. $\kappa \in (0, 1]$, $\delta > 0$ be fixed.

11. $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ be a minimal δ-covering of $\mathcal{F}$ w.r.t. $\|\cdot\|_\infty$.

12. $N_\delta = |\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)|$ be the number of elements in this covering.

Then

$$\left\|\widehat{Q} - TQ\right\|_\nu^2 \leqslant C_5 \log(N_{(1/n)})/n + 2\omega(\mathcal{F}) + C_3\sqrt{\log(N_{(1/n)})/n} + C_6 n^{-1}$$

for absolute constants $C_3, C_5, C_6 > 0$ not depending on n, and where we define

$$\omega(\mathcal{F}) = \sup_{g\in\mathcal{G}} \inf_{f\in\mathcal{F}} \frac{1}{n}\mathbb{E}\|f - g\|_\nu^2 \tag{3.10}$$

Remark 3.29. This is a slightly different result than the one stated in [7], where they had (translated into notation used here)

$$\left\|\widehat{Q} - TQ\right\|_\nu^2 \leqslant (1+\kappa)^2\omega(\mathcal{F}) + CV_{\max}^2/(n\kappa) \cdot \log(N_\delta) + C' \cdot V_{\max}\delta \tag{3.11}$$

for some constants $C, C' > 0$ (also not depending on n) and where in the application of the theorem they set $\kappa = 1$ and $\delta = 1/n$. This is mainly due to corrections in their proof.

The proofs of theorem 3.27 and theorem 3.28 are found below, but first we will show how to combine them to obtain theorem 3.26.

Proof of main theorem (3.26). Using theorem 3.27 we get

$$\|Q^* - Q_{\pi_K}\|_{1,\mu} \leqslant \frac{2\phi_{\mu,\nu}\gamma}{(1-\gamma)^2}\varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2}R_{\max} \tag{3.12}$$

where $\varepsilon_{\max} = \max_{k\in[K]}\left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\|_{2,\nu}$. Using theorem 3.28 with $Q = \widetilde{Q}_{k-1}$ and $\mathcal{F} = \mathcal{F}_0$, we get

$$\varepsilon_{\max} \leqslant C_5 \log(N_{(1/n)})/n + 2\omega(\mathcal{F}) + C_3\sqrt{\log(N_{(1/n)})/n} + C_6 n^{-1} \tag{3.13}$$

where $N_0 = N_{(1/n)} = |\mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty)|$. It remains only to bound $\omega(\mathcal{F}_0)$ and N_0. We start with $\omega(\mathcal{F}_0)$.

Step 1. We want to employ the lemma by [18] (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 (see remark 2.72). Therefore the first step is to refit our Hölder Smooth compositions in $\mathcal{G}_0$ to be defined on the unit cube in the respective dimensions 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

$$g_{jk} \in C_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j), \qquad j \in [q], k \in [p_{j+1}] \tag{3.14}$$

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

$$\begin{aligned} h_1 &= g_1/(2H_1) + 1/2 \\ h_j(u) &= g_j(2H_{j-1}u - H_{j-1})/(2H_j) + 1/2, && j \in \{2, \ldots, q-1\} \\ h_q(u) &= g_q(2H_{q-1}u - H_{q-1}) \end{aligned} \tag{3.15}$$

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

$$\begin{aligned} h_{1k} &\in C_{t_1}([0,1]^{t_1}, \beta_1, 1) \\ h_{jk} &\in C_{t_j}([0,1]^{t_j}, \beta_j, (2H_{j-1})^{\beta_j}), && j \in \{2, \ldots, q-1\} \\ h_q &\in C_{t_q}([0,1]^{t_q}, \beta_q, H_q(2H_{q-1})^{\beta_q}) \end{aligned} \tag{3.16}$$

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

Step $\frac{3}{2}$

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 $N \geqslant \max\{(\beta_j+1)^{t_j}, (H_j+1)e^{t_j} \mid j \in [q]\}$.

$$W := \max\left(\left\{(2H_{j-1})^{\beta_j} \mid 1 \leqslant j \leqslant q-1\right\} \cup \left\{H_q(2H_{q-1})^{\beta_q}, 1\right\}\right) \tag{3.17}$$

By lemma 3.19 there exists a ReLU network

$$\widehat{h}_{jk} \in \mathcal{SRN}\left((\widetilde{s}_j + 4) \cdot p_{j+1}, V_{\max},\ (t_j, \widetilde{d}_j p_{j+1}, \ldots, \widetilde{d}_j p_{j+1}, p_{j+1}), L_j + 2\right) \tag{3.18}$$

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

$$\left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty \leqslant (2W+1)6^{t_j}N2^{-m} + W3^{\beta_j}N^{-\beta_j/t_j} \leqslant 2W3^{\beta_j}N^{-\beta_j/t_j} \tag{3.19}$$

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

Proof. This is the simple network $\min(0, \max(1, u)) = \sigma_r(1 - \sigma_r(1-u))$ where $\sigma_r(x) = \max(0, x)$ is the ReLU activation function. The weights are $w_1 = w_2 = -1$ and $v_1 = v_2 - 1$. □

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

$$\widetilde{h}_{jk} \in \mathcal{SRN}\left((\widetilde{s}_j + 4)p_{j+1}, V_{\max}, (t_j, \widetilde{d}_j, \ldots, \widetilde{d}_j, 1), L_j + 2\right) \tag{3.20}$$

and since $h_{jk}([0,1]^{t_j}) \in [0,1]$ by eq. (3.19) we have

$$\left\|\widetilde{h}_{jk} - h_{jk}\right\|_\infty = \left\|\tau \circ \widehat{h}_{jk} - \tau \circ h_{jk}\right\|_\infty \leqslant \left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty \leqslant 2W3^{-\beta_j}N^{-\beta_j/t_j} \tag{3.21}$$

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 $\widetilde{f} \in \mathcal{SRN}\left(\widetilde{s}, V_{\max}, (w, \widetilde{d}, \ldots, \widetilde{d}, 1), \widetilde{L}\right)$. 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. (3.19)

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

and therefore

$$\omega(\mathcal{F}_0)\leqslant\ C_N n^{-2\alpha^*\min_{j\in[q]}\beta_j^*/t_j}\ \leqslant\ C_N n^{-2\alpha^*\kappa^*} \tag{3.22}$$

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 $\mathcal{SRN}\left(\widetilde{s},V_{\max},(\widetilde{d}_j)_{j=1}^{\widetilde{L}+1},\widetilde{L}\right)$ by

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

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

$$\left|\mathcal{N}(\delta,\mathcal{F}_0,\|\cdot\|_\infty)\right|\leqslant|\mathcal{N}_\delta|^{|\mathcal{A}|}$$

We now utilize a lemma found in [1, Anthony and Bartlett (2002)]

Lemma 3.31 (Covering number of ReLU networks)**.** Consider the family of ReLU networks

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

where $\mathcal{SRN}$ is defined in definition 3.16. Let $D:=\prod_{\ell=1}^{L+1}(d_\ell+1))$. Then for any $\delta>0$

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

Proof. We refer to theorem 14.5 in [1]. □

With lemma 3.31 and n sufficiently large we can bound

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

$$
\begin{aligned}
&\leqslant 2|\mathcal{A}| \left(c_{\tilde{s}} n^{\alpha^*} \log(n)^{\xi^*} + 1\right) \log\left(2(c_{\tilde{L}} \log(n)^{\xi^*} + 1)(6n^{\alpha^*} \log(n)^{\xi^*} + 1)^{\tilde{L}+1}\right) \\
&\leqslant 4|\mathcal{A}|\, c_{\tilde{s}} n^{\alpha^*} \log(n)^{\xi^*} (\tilde{L}+1) \log\left(24 c_{\tilde{L}} \log(n)^{\xi^*} n^{\alpha^*} \log(n)^{\xi^*}\right) \\
&\leqslant 8|\mathcal{A}|\, c_{\tilde{s}} n^{\alpha^*} \log(n)^{\xi^*} c_{\tilde{L}} \log(n)^{\xi^*} (\alpha^* + 2) \log(n) \\
&= 8 c_{\tilde{s}} c_{\tilde{L}} (\alpha^* + 2) n^{\alpha^*} \log(n)^{1+2\xi^*} \\
&= c_{N_0} (\alpha^* + 2) n^{\alpha^*} \log(n)^{1+2\xi^*} \qquad (3.23)
\end{aligned}
$$

Where we define $c_{N_0} = 8c_{\tilde{s}}c_{\tilde{L}}$. Using eq. (3.23), eq. (3.22) and eq. (3.13) we can bound

$$
\varepsilon_{\max} \leqslant C_5 C_{N_0} n^{\alpha^*-1}/2 \log(n)^{1+2\xi^*} + \sqrt{C_5 C_{N_0} n^{\alpha^*-1} \log(n)^{1+2\xi^*}} + 2c_N n^{-2\alpha^*\kappa^*} + C_6 n^{-1} \quad (3.24)
$$

Since we are interested in convergence (to 0) we may assume that n is sufficiently large such that only the largest exponent in the above expression is significant. This leads to the simplification

$$
\begin{aligned}
\varepsilon_{\max} &\leqslant (2C_5 C_{N_0} + C_3 + 2c_N + C_6) n^{\max\{(\alpha^*-1)/2, -2\alpha^*\kappa^*\}} \log(n)^{1+2\xi^*} \qquad (3.25) \\
&= C_7 n^{\max\{(\alpha^*-1)/2, -2\alpha^*\kappa^*\}} \log(n)^{1+2\xi^*} \qquad (3.26)
\end{aligned}
$$

where $C_7 := (2C_5 C_{N_0} + C_3 + 2c_N + C_6)$. Now using eq. (3.12) and eq. (3.26)

$$
\left\|Q^* - Q_{\pi_K}\right\|_{1,\mu} \leqslant C_7 \frac{\phi_{\mu,\nu}\gamma}{(1-\gamma)^2} V_{\max}^2 n^{\max\{-2\alpha^*\kappa^*, (\alpha^*-1)/2\}} \log(n)^{1+2\xi^*} + \frac{4\gamma}{(1-\gamma)^2} R_{\max}\gamma^K
$$

where C_7 only depends on the constants in assumption 6 finishing the proof. □

Before we proceed to prove theorem 3.27, we will establish a couple of lemmas.

Lemma 3.32. *$TQ \geqslant T_\pi Q$ 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. This is an easy consequence of the definitions (definition 2.63)

$$
\begin{aligned}
(TQ)(s,a) &= r(s,a) + \gamma \int \max_{a'} Q(s',a') \, \mathrm{d}P(s' \mid s,a) \\
&\geqslant r(s,a) + \gamma \int\int Q(s',a'') \, \mathrm{d}\pi(a'' \mid s') \, \mathrm{d}P(s' \mid s,a) \\
&= T_\pi Q(s,a)
\end{aligned}
$$

since $\max_{a'} Q(s',a') \geqslant Q(s',a'')$ for any $a'' \in \mathcal{A}$. □

The next lemma (last before proof of 3.28) is about the relation between the next-step operator P_τ and the concentration coefficients.

We recall here some details regarding composition of kernels and measures discussed in remark 2.11. A stationary policy $\tau : \mathcal{S} \rightsquigarrow \mathcal{A} \in S\Pi$ composed with the transition kernel $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ yields a kernel $\tau P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S} \times \mathcal{A}$. The $\circ$-composition of kernels is forgets histories and so if $\tau' \in S\Pi$ we have that $(\tau' P) \circ (\tau P) : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S} \times \mathcal{A}$. Lastly the kernel-measure $\circ$-composition by a probability measure $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ we have that $(\tau P) \circ \mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$.

Lemma 3.33. *Let $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be an action-value function, $\tau_1, \ldots, \tau_m$ be policies and $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be a probability measure. Then*

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

For any measure $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ which is absolutely continuous w.r.t. $(\tau_m P) \circ \cdots \circ (\tau_1 P) \circ \mu$. Here κ is the concentration coefficients defined in definition 3.25.

Proof. Recall that

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

Now

$$\begin{aligned}\mathbb{E}_\mu[P_{\tau_m}\ldots P_{\tau_1}f] &= \int P_{\tau_{m-1}}\ldots P_{\tau_2}f\ \mathrm{d}\tau_m P\ \mathrm{d}\mu &(3.27)\\ &= \int f\ \mathrm{d}(\tau_1 P)\circ\cdots\circ(\tau_m P)\circ\mu &(3.28)\\ &= \int f\frac{\mathrm{d}(\tau_1 P)\circ\cdots\circ(\tau_m P)\circ\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu &(3.29)\\ &\leqslant \left\|\frac{(\tau_1 P)\circ\cdots\circ(\tau_m P)\circ\mu}{\mathrm{d}\nu}\right\|_{2,\nu}\cdot\|f\|_{2,\nu} &(3.30)\\ &\leqslant \kappa(m,\mu,\nu)\|f\|_{2,\nu} &(3.31)\end{aligned}$$

Where eq. (3.29) is due to the Radon-Nikodym theorem (theorem A.23) and eq. (3.30) is Cauchy-Schwarz. □

We now turn to the proof of theorem 3.27.

Proof of theorem 3.27. 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

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

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^*. Also if $f, f' : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ are measurable we have

$$f\geqslant f' \implies P_\pi f\geqslant P_\pi f' \tag{3.32}$$

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 3.32 we can make the upper bound

$$\begin{aligned}Q^* - Q_{\pi_K} &= T_{\pi*}Q^* - T_{\pi_K}Q_{\pi_K}\\ &= T_{\pi*}Q^* + (T_{\pi*}\widetilde{Q}_K - T_{\pi*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T\widetilde{Q}_K) - T_{\pi_K}Q_{\pi_K}\\ &= (T_{\pi*}\widetilde{Q}_K - T\widetilde{Q}_K) + (T_{\pi*}Q^* - T_{\pi*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T_{\pi_K}Q_{\pi_K})\\ &\leqslant (T_{\pi*}Q^* - T_{\pi*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T_{\pi_K}Q_{\pi_K})\\ &= (T_{\pi*}Q^* - T_{\pi*}\widetilde{Q}_K) + (T_{\pi_K}\widetilde{Q}_K - T_{\pi_K}Q_{\pi_K})\\ &= \gamma P_{\pi*}(Q^* - \widetilde{Q}_K) + \gamma P_{\pi_K}(\widetilde{Q}_K - Q_{\pi_K})\\ &= \gamma(P_{\pi*} - P_{\pi_K})(Q^* - \widetilde{Q}_K) + \gamma P_{\pi_K}(Q^* - Q_{\pi_K}) &(3.33)\end{aligned}$$

This implies

$$(I-\gamma P_{\pi_K})(Q^* - Q_{\pi_K}) \leqslant \gamma(P_{\pi*} - P_{\pi_K})(Q^* - \widetilde{Q}_K) \tag{3.34}$$

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

$$U = \sum_{i=0}^{\infty} \gamma^i (P_{\pi_K})^i \tag{3.35}$$

From eq. (3.35) and eq. (3.32) 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. (3.34) to obtain

$$Q^* - Q_{\pi_K} \leqslant \gamma U (P_{\pi*}(Q^* - \widetilde{Q}_K) - P_{\pi_K}(Q^* - \widetilde{Q}_K)) \tag{3.36}$$

Step 2 Using lemma 3.32 for any $i \in [K]$ we can get an upper bound

$$\begin{aligned}
Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T_{\pi*}\widetilde{Q}_i - T_{\pi*}\widetilde{Q}_i) \\
&= (Q^* - T_{\pi*}\widetilde{Q}_i) + (T\widetilde{Q}_i - \widetilde{Q}_{i+1}) + (T_{\pi*}\widetilde{Q}_i - T\widetilde{Q}_i) \\
&= (T_{\pi*}Q^* - T_{\pi*}\widetilde{Q}_i) + \varrho_{i+1} + (T_{\pi*}\widetilde{Q}_i - T\widetilde{Q}_i) \\
&\leqslant T_{\pi*}Q^* - T_{\pi*}\widetilde{Q}_i + \varrho_{i+1} \\
&= \gamma P_{\pi*}(Q^* - \widetilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{3.37}$$

and a lower bound

$$\begin{aligned}
Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T_{\pi_i}Q^* - T_{\pi_i}Q^*) \\
&= (T_{\pi_i}Q^* - T_{\pi_i}\widetilde{Q}_i) + \varrho_{i+1} + (TQ^* - T_{\pi_i}Q^*) \\
&\geqslant T_{\pi_i}Q^* - T_{\pi_i}\widetilde{Q}_i + \varrho_{i+1} \\
&= \gamma P_{\pi_i}(Q^* - \widetilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{3.38}$$

Applying eq. (3.37) and eq. (3.38) iteratively we get

$$Q^* - \widetilde{Q}_K \leqslant \gamma^K (P_{\pi*})^K (Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P_{\pi*})^{K-1-i} \varrho_{i+1} \tag{3.39}$$

and

$$Q^* - \widetilde{Q}_K \geqslant \gamma^K (P_{\pi_{K-1}} \dots P_{\pi_0})(Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P_{\pi_{K-1}} \dots P_{\pi_{i+1}}) \varrho_{i+1} \tag{3.40}$$

Step 3 Combining eq. (3.39) and eq. (3.40) with eq. (3.36) we get

$$\begin{aligned}
Q^* - Q_{\pi_K} \leqslant U^{-1} \Bigg(&\gamma^{K+1}((P_{\pi*})^{K+1} - P_{\pi_K} \dots P_{\pi_0})(Q^* - \widetilde{Q}_0) \\
&+ \sum_{i=0}^{K-1} \gamma^{K-i}((P_{\pi*})^{K-i} - P_{\pi_K} \dots P_{\pi_{i+1}}) \varrho_{i+1} \Bigg)
\end{aligned} \tag{3.41}$$

For shorthand define constants

$$\alpha_i = \frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}} \text{ for } 0 \leqslant i \leqslant K-1 \text{ and } \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} \tag{3.42}$$

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

$$O_i = (1-\gamma)/2U^{-1}[(P_{\pi*})^{K-i} + (P_{\pi_K} \dots P_{\pi_{i+1}})] \tag{3.43}$$

$$O_K = (1-\gamma)/2U^{-1}[(P_{\pi*})^{K+1} + (P_{\pi_K} \dots P_{\pi_0})] \tag{3.44}$$

Then by eq. (3.41)

$$\left|Q^* - Q_{\pi_K}\right| \leqslant \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{i=0}^{K-1}\alpha_i O_i|\varrho_{i+1}| + \alpha_K O_K\left|Q^* - \widetilde{Q}_0\right|\right] \tag{3.45}$$

So by linearity of expectation

$$\begin{aligned}\left\|Q^* - Q_{\pi_K}\right\|_{1,\mu} &= \mathbb{E}_\mu\left|Q^* - Q_{\pi_K}\right| &(3.46)\\ &\leqslant \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{i=0}^{K-1}\alpha_i\mathbb{E}_\mu(O_i|\varrho_{i+1}|) + \alpha_K\mathbb{E}_\mu(O_K\left|Q^* - \widetilde{Q}_0\right|)\right] &(3.47)\end{aligned}$$

With the bound on rewards we (crudely) estimate

$$\mathbb{E}_\mu O_K\left|Q^* - \widetilde{Q}_0\right| \leqslant 2V_{\max} = 2R_{\max}/(1-\gamma) \tag{3.48}$$

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

Step 4 Using the sum expansion of U we get

$$\begin{aligned}\mathbb{E}_\mu(O_i|\varrho_{i+1}|) &= \frac{1-\gamma}{2}\mathbb{E}_\mu\left(U^{-1}[(P_{\pi_K})^{K-i} + P_{\pi_K}\dots P_{\pi_{i+1}}]|\varrho_{i+1}|\right) &(3.49)\\ &= \frac{1-\gamma}{2}\mathbb{E}_\mu\left(\sum_{j=0}^{\infty}[(P_{\pi_K})^j(P_{\pi_K})^{K-i} + (P_{\pi_K})^{j+1}P_{\pi_{K-1}}\dots P_{\pi_{i+1}}]|\varrho_{i+1}|\right) &(3.50)\\ &= \frac{1-\gamma}{2}\sum_{j=0}^{\infty}\mathbb{E}_\mu\left([(P_{\pi_K})^j(P_{\pi_K})^{K-i} + (P_{\pi_K})^{j+1}P_{\pi_{K-1}}\dots P_{\pi_{i+1}}]|\varrho_{i+1}|\right) &(3.51)\end{aligned}$$

Notice that there are $K-i+j$ P-operators on both terms in the sum. Therefore were can employ lemma 3.33 twice. Moreover define $\varepsilon_{\max} = \max_{i\in[K]}\|\varrho_i\|_{2,\nu}$. Then

$$\begin{aligned}\mathbb{E}_\mu(O_i|\varrho_{i+1}|) &\leqslant (1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu)\|\varrho_{i+1}\|_{2,\nu}\\ &\leqslant \varepsilon_{\max}(1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu) &(3.52)\end{aligned}$$

Using eq. (3.47), eq. (3.48) and eq. (3.52)

$$\begin{aligned}\left\|Q^* - Q_{\pi_K}\right\|_{1,\mu} \leqslant \frac{2\gamma(1-\gamma^{K+1})}{1-\gamma}&\left[\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\right]\varepsilon_{\max}\\ &+\frac{4\gamma(1-\gamma^{K+1})}{(1-\gamma)^3}\alpha_K R_{\max}\end{aligned} \tag{3.53}$$

Focusing on the first term on RHS of eq. (3.53), if we then we can take the norm out of the sum as a constant. We are left with

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

$$\leqslant \frac{1}{1-\gamma^{K+1}(1-\gamma)}\phi_{\mu,\nu} \tag{3.54}$$

Where the last inequality is due to assumption 6.

Combining eq. (3.53) and eq. (3.54) we arrive at

$$\left\|Q^* - Q_{\pi_K}\right\|_{1,\mu} \leqslant \frac{2\gamma \cdot \phi_{\mu,\nu}}{(1-\gamma)^2} \cdot \varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} \cdot R_{\max} \tag{3.55}$$

□

Before turning to the proof of theorem 3.28 we introduce again some lemmas for use in the proof.

Definition 3.34 (Sub-gaussian random variables)**.** Let $X \in \mathcal{X}$ be a random variable. The **sub-gaussian norm** of X is defined as

$$\|X\|_{\psi_2} = \sup_{p \geqslant 1} p^{-1/2} \|X\|_p \tag{3.56}$$

(where $\|X\|_p = \left(\int |\mathcal{X}|^p \, \mathrm{d}\mathbb{P}\right)^{1/p}$). When $\|X\|_{\psi_2} < \infty$ then X is called a **sub-gaussian random variable**.

Lemma 3.35 (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

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

(by centered we mean $\mathbb{E}X_i = 0$).

Proof. See [21, Vershynin (2010)] p. 12. □

Definition 3.36 (Sub-exponential norm)**.** For a random variable define

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

called the **sub-exponential norm**. In case $\|X\|_{\psi_1} < \infty$ we say X is a **sub-exponential random variable**.

Lemma 3.37 (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 [21] p. 14. □

Proposition 3.38. 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 $v = (v_1, \ldots, v_n)$. Cauchy-Schwarz applied to some vector w and $(1, \ldots, 1)$ yields

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

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

$$\mathbb{E}\|v\|_1 = \|w\|_1 \leqslant \sqrt{n}\sqrt{\sum_{i=1}^{n}(\mathbb{E}v_i)^2} \leqslant \sqrt{n}\sqrt{\mathbb{E}\sum_{i=1}^{n} v_i^2} = \sqrt{n}\sqrt{\mathbb{E}\|v\|_2^2}$$

□

Theorem 3.39 (Bernstein's inequality)**.** Suppose $U_1, \ldots, U_n$ 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. See e.g. [1, Anthony and Bartlett (2002)] p. 363. □

We can now begin the proof of theorem 3.28.

Proof of theorem 3.28. First some introductory fixing of notation and variables. Fix a minimal δ-covering of $\mathcal{F}$ with centers $f_1, \ldots, f_{N_\delta}$. Define

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

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

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 $Y = (Y_1, \ldots, Y_n)$ in mean squared error over $X = (X_1, \ldots, X_n)$. 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

$$\|Y\|^2 + \left\|\widehat{Q}\right\|^2 - 2Y \cdot \widehat{Q} \leqslant \|Y\|^2 + \|f\|^2 - 2Y \cdot f \tag{3.57}$$

$$\iff \left\|\widehat{Q}\right\|^2 + \|TQ\|^2 - 2\widehat{Q} \cdot TQ \leqslant \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \widehat{Q} - 2Y \cdot f - 2\widehat{Q} \cdot TQ + 2f \cdot TQ \tag{3.58}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leqslant \|f - TQ\|^2 + 2(Y - TQ) \cdot (\widehat{Q} - f) \tag{3.59}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leqslant \|f - TQ\|^2 + 2\xi \cdot (\widehat{Q} - f) \tag{3.60}$$

Where $\xi_i := Y_i - TQ(X_i)$ and $\xi := (\xi_1, \ldots, \xi_n)$. Let $\Sigma = (X_1, \ldots, X_n)^{-1}(\mathbb{B}_n) \in \mathcal{H}$ be the σ-algebra generated by the samples. We denote by $\mathbb{E}_\Sigma$ the conditional expectation with respect to the subalgebra $\Sigma \subseteq \Sigma_\Omega$, which is also called the conditional expectation with respect to the samples $X_1, \ldots, X_n$. For example we have $\mathbb{E}_\Sigma X_i = X_i$ for all $i \in [n]$. Now we proof a minor lemma

Proposition 3.40. $\mathbb{E}_\Sigma \xi_i = 0$ and thus $\mathbb{E}(\xi_i g(X_i)) = 0$ for any function $g : \mathbb{R} \to \mathbb{R}$.

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

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

where $S_i' \sim P(\cdot \mid X_i)$, $R_i \sim R(\cdot \mid X_i)$. Therefore

$$\begin{aligned} \mathbb{E}_\Sigma Y_i &= \int x \,\mathrm{d}R(x \mid X_i) + \gamma \int \max_{a \in \mathcal{A}} Q(s,a) \,\mathrm{d}P(s \mid X_i) \\ &= r(X_i) + \gamma \int \max_{a \in \mathcal{A}} Q(s,a) \,\mathrm{d}P(s \mid X_i) = TQ(X_i) \end{aligned}$$

therefore

$$\mathbb{E}_\Sigma \xi_i = \mathbb{E}_\Sigma \left(Y_i - TQ(X_i) \right) = \mathbb{E}_\Sigma Y_i - TQ(X_i) = 0$$

□

By this lemma we can deduce

$$\mathbb{E}\left(\xi \cdot (\widehat{Q} - f)\right) = \mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ)\right) \tag{3.61}$$

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

$$\left|\mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ)\right)\right| \leqslant \left|\mathbb{E}\left(\xi \cdot (\widehat{Q} - f_{k*})\right)\right| + \left|\mathbb{E}\left(\xi \cdot (f_{k*} - TQ)\right)\right| \tag{3.62}$$

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

$$\left|\mathbb{E}\xi \cdot (\widehat{Q} - f_{k*})\right| \leqslant \mathbb{E}\left(\|\xi\| \left\|\widehat{Q} - f_{k*}\right\|\right) \leqslant \mathbb{E}(\|\xi\|)\sqrt{n}\delta \leqslant 2nV_{\max}\delta \tag{3.63}$$

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

$$\left\|\widehat{Q} - f_{k*}\right\|^2 = \sum_{i=1}^n (\widehat{Q}(X_i) - f_{k*}(X_i))^2 \leqslant \sum_{i=1}^n \delta^2 = n\delta^2 \tag{3.64}$$

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

$$\|\xi\|^2 = \sum_{i=1}^n (Y_i - TQ(X_i))^2 \leqslant \sum_{i=1}^n (2V_{\max})^2 = 4V_{\max}^2 n \tag{3.65}$$

To bound the second term in eq. (3.62) define

$$Z_j := \xi \cdot (f_j - TQ) \left\| f_j - TQ \right\|^{-1} \tag{3.66}$$

Note that since ξ_i is centered Z_i is also centered for any $i \in [n]$. For a sub-σ-algebra $\Sigma' \subseteq \Sigma_\Omega$ define the the conditional *sub-gaussian* norm by

Definition 3.41 (Conditional sub-gaussian norm)**.**

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

Recall that Σ is the subalgebra generated by $(X_1, \ldots, X_n)$. Because of proposition 3.40 $\xi_i(f_j(X_i) - TQ(X_i))$ is centered for any $i \in [n]$ and

$$\left\|\xi_i(f_j(X_i) - TQ(X_i))\right\|_{\psi_2, \Sigma} \leqslant 2V_{\max} \left| f_j(X_i) - TQ(X_i) \right| \tag{3.67}$$

Therefore by lemma 3.35 ([21, Vershynin (2010)] lemma 5.9)

$$\|Z_j\|_{\psi_2,\Sigma}^2 \leqslant \|f_j - TQ\|^{-2} \left\| \sum_{i=1}^n \xi_i (f_j(X_i) - TQ(X_i)) \right\|_{\psi_2,\Sigma}^2 \tag{3.68}$$

$$\overset{3.35}{\leqslant} \|f_j - TQ\|^{-2} C_1 \sum_{i=1}^n \|\xi_i (f_j(X_i) - TQ(X_i))\|_{\psi_2,\Sigma}^2 \tag{3.69}$$

$$\leqslant \|f_j - TQ\|^{-2} C_1 \sum_{i=1}^n 4V_{\max}^2 |f_j(X_i) - TQ(X_i)|^2 \tag{3.70}$$

$$= 4V_{\max}^2 C_1 \tag{3.71}$$

Remark 3.42 (Mistake in [7])**.** It is at this point there was a mistake in the proof of [7] p. 45 eq. (C.40). Here a bound on $\mathbb{E}(\exp(t \cdot Z_j^2))$ is attempted using lemma 5.15 [21, Vershynin (2010)] which states

> **Lemma 3.43.** Let X be a centered sub-exponential random variable. Then for t such that $|t| \leqslant c/\|X\|_{\psi_1}$ one has
>
> $$\mathbb{E}\exp(tX) \leqslant \exp(Ct^2 \|X\|_{\psi_1}^2)$$
>
> where $C, c > 0$ are absolute constants.

The mistake here is that while Z_j^2 is sub-exponential enough, it is not clear why it should be centered. Indeed as $Z_j^2 \geqslant 0$ a.s. it is centered if and only if it is equal to 0 a.s!

We therefore proceed differently, but luckily end up (at eq. (3.78)) with a bound very similar to that of [7] (C.42), only different by a absolute constant factor.

Our argument is based on a lemma in the same section of [21]:

Lemma 3.44. $\mathbb{E}\exp(cX^2/\|X\|_{\psi_2}^2) \leqslant \exp(1) = e$

Proof. See [21] p. 11 between eqs. (5.11) and (5.12). □

By this lemma we have

$$\mathbb{E}\exp\left(cZ_j^2/\|Z_j\|_{\psi_2}^2\right) \leqslant e \tag{3.72}$$

so

$$\mathbb{E}\max_{j\in N_\delta} Z_j^2 = \frac{\max_{j\in[N_\delta]}\|Z_j\|_{\psi_2}^2}{c}\mathbb{E}\left(\max_{j\in[N_\delta]} \frac{cZ_j^2}{\max_{k\in[N_\delta]}\|Z_k\|_{\psi_2}}\right) \tag{3.73}$$

$$\leqslant \frac{4V_{\max}^2 C_1}{c}\mathbb{E}\left(\max_{j\in N_\delta}\frac{cZ_j^2}{\|Z_j\|_{\psi_2}}\right) \tag{3.74}$$

$$\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(\mathbb{E}\max_{j\in N_\delta}\exp\left(\frac{cZ_j^2}{\|Z_j\|_{\psi_2}}\right)\right) \tag{3.75}$$

$$\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(\sum_{j\in[N_\delta]}\mathbb{E}\exp\left(\frac{cZ_j^2}{\|Z_j\|_{\psi_2}}\right)\right) \tag{3.76}$$

$$\leqslant \frac{4V_{\max}^2 C_1}{c}\log(eN_\delta) \tag{3.77}$$

$$\leqslant C_2^2 V_{\max}^2 \log(N_\delta) \tag{3.78}$$

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

$$\begin{aligned}\mathbb{E}\left(\xi \cdot (f_{k*} - TQ)\right) &= \mathbb{E}\left(\|f_{k*} - TQ\||Z_{k*}|\right) &(3.79)\\ &\leqslant \mathbb{E}\left(\left(\left\|\widehat{Q} - TQ\right\| + \left\|\widehat{Q} - f_{k*}\right\|\right)|Z_{k*}|\right) &(3.80)\\ &\leqslant \mathbb{E}\left(\left(\left\|\widehat{Q} - TQ\right\| + n\delta\right)|Z_{k*}|\right) &(3.81)\\ &\leqslant \left(\mathbb{E}\left(\left\|\widehat{Q} - TQ\right\| + n\delta\right)^2\right)^{1/2} \left(\mathbb{E}Z_{k*}^2\right)^{1/2} &(3.82)\\ &\leqslant \mathbb{E}\left(\left\|\widehat{Q} - TQ\right\| + n\delta\right) \left(\mathbb{E}Z_{k*}^2\right)^{1/2} &(3.83)\\ &\leqslant \left(\sqrt{\mathbb{E}\left\|\widehat{Q} - TQ\right\|_2^2} + n\delta\right) C_2 V_{\max}\sqrt{\log(N_\delta)} &(3.84)\end{aligned}$$

Where eq. (3.79) to eq. (3.80) is by the triangle inequality, and eq. (3.83) to eq. (3.84) is by Jensens inequality. Combining eq. (3.60), eq. (3.62), eq. (3.63) and eq. (3.84) we get

$$\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 \leqslant \mathbb{E}\|f - TQ\|^2 + 4nV_{\max}\delta + 2C_2V_{\max}\sqrt{\log(N_\delta)}\left(\sqrt{\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2} + \delta n\right) \quad (3.85)$$

$$\begin{aligned}&= 2C_2V_{\max}\sqrt{\log(N_\delta)}\sqrt{\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2} + \mathbb{E}\|f - TQ\|^2 + 4nV_{\max}\delta\\ &\quad + 2C_2V_{\max}\sqrt{\log(N_\delta)}\delta n \qquad (3.86)\end{aligned}$$

Lemma 3.45. Let $a, b > 0,\ \kappa \in (0, 1]$ then $a^2 \leqslant 2ab + c \implies a^2 \leqslant (1 + \kappa)^2 b^2/\kappa + (1 + \kappa)c$

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 3.45 applied to eq. (3.86) we get

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 \leqslant& \frac{(1+\kappa)^2}{\kappa}C_2^2V_{\max}^2\frac{1}{n}\log(N_\delta)\\ &+ (1+\kappa)\left(2C_2V_{\max}\sqrt{\log(N_\delta)}\delta + \frac{1}{n}\mathbb{E}\|f - TQ\|^2\right) \qquad (3.87)\end{aligned}$$

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. (3.87) holds for any $f \in \mathcal{F}$ we can further say

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

$$+ (1+\kappa)\left(2C_2 V_{\max}\sqrt{\log(N_\delta)}\delta + \sup_{g\in\mathcal{G}}\inf_{f\in\mathcal{F}}\|f-g\|_\nu^2\right) \tag{3.88}$$

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

$$\left|\left(\widehat{Q}(x)-TQ(x)\right)^2 - \left(f_{k*}(x)-TQ(x)\right)^2\right| = \left|\widehat{Q}(x)-f_{k*}(x)\right| \cdot \left|\widehat{Q}(x)+f_{k*}(x)-2TQ(x)\right| \tag{3.89}$$

$$\leqslant 4V_{\max}\delta \tag{3.90}$$

Using this twice we can say

$$(\widehat{Q}(\widehat{X}_i)-TQ(\widehat{X}_i))^2 \tag{3.91}$$

$$\leqslant (\widehat{Q}(\widetilde{X}_i)-TQ(\widetilde{X}_i))^2 - (f_{k*}(\widetilde{X}_i)-TQ(\widetilde{X}_i))^2 + (f_{k*}(\widetilde{X}_i)-TQ(\widetilde{X}_i))^2 \tag{3.92}$$

$$\begin{aligned}\leqslant\; & (f_{k*}(\widetilde{X}_i)-TQ(\widetilde{X}_i))^2 + (\widehat{Q}(X_i)-TQ(X_i))^2 - (\widehat{Q}(X_i)-TQ(X_i))^2 \\ & + (f_{k*}(X_i)-TQ(X_i))^2 - (f_{k*}(X_i)-TQ(X_i))^2 + 4V_{\max}\delta\end{aligned} \tag{3.93}$$

$$\leqslant (\widehat{Q}(X_i)-TQ(X_i))^2 + (f_{k*}(\widetilde{X}_i)-TQ(\widetilde{X}_i))^2 - (f_{k*}(X_i)-TQ(X_i))^2 + 8V_{\max}\delta \tag{3.94}$$

Thus we get

$$\left\|\widehat{Q}-TQ\right\|_\nu^2 \tag{3.95}$$

$$= \mathbb{E}\frac{1}{n}\sum_{i=1}^n (\widehat{Q}(\widetilde{X}_i)-TQ(\widetilde{X}_i))^2 \tag{3.96}$$

$$\leqslant \mathbb{E}\frac{1}{n}\sum_{i=1}^n \left((\widehat{Q}(X_i)-TQ(X_i))^2 + (f_{k*}(\widetilde{X}_i)-TQ(\widetilde{X}_i))^2 - (f_{k*}(X_i)-TQ(X_i))^2\right) + 8V_{\max}\delta \tag{3.97}$$

$$= \frac{1}{n}\mathbb{E}\left\|\widehat{Q}-TQ\right\|^2 + \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n h_{k*}(X_i,\widetilde{X}_i)\right) + 8V_{\max}\delta \tag{3.98}$$

Where we define

$$h_j(x,y) := \left(f_j(y)-TQ(y)\right)^2 - \left(f_j(x)-TQ(x)\right)^2 \tag{3.99}$$

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

$$T := \max_{j\in[N_\delta]}\left|\sum_{i=1}^n h_j(X_i,\widetilde{X}_i)/\Upsilon\right| \tag{3.100}$$

Then we can bound the middle term in eq. (3.98)

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n h_{k*}(X_i,\widetilde{X}_i)\right) \leqslant \Upsilon/n\mathbb{E}\max_{j\in[N_\delta]}\left(\left|\sum_{i=1}^n h_j(X_i,\widetilde{X}_i)/\Upsilon\right|\right) \tag{3.101}$$

$$\leqslant \Upsilon/n\mathbb{E}T \tag{3.102}$$

We want to use Bernsteins inequality (theorem 3.39) with $U_i = h_j(X_i,\widetilde{X}_i)$. Therefore notice that $\left|h_j\right| \leqslant \Upsilon^2$ and

$$\mathbb{V}h_j(X_i,\widetilde{X}_i) = 2\mathbb{V}\left(f_j(X_i)-TQ(X_i)\right)^2 \tag{3.103}$$

$$\leqslant 2\mathbb{E}\left(f_j(X_i)-TQ(X_i)\right)^4 \tag{3.104}$$

$$\leqslant 2\Upsilon^4 \tag{3.105}$$

(here $\mathbb{V}$ denotes the variance operator on random variables). so by union bounding for any $u < 6n\Upsilon$ we have

$$\mathbb{E}T = \int_0^\infty \mathbb{P}(T \geqslant t) \tag{3.106}$$

$$\leqslant u + \int_u^\infty \mathbb{P}(T \geqslant t)\ \mathrm{d}t \tag{3.107}$$

$$\leqslant u + \int_u^\infty 2N_\delta \exp\left(\frac{-t^2}{2\Upsilon t/3 + 4n\Upsilon^2}\right)\ \mathrm{d}t \tag{3.108}$$

$$\leqslant u + 2N_\delta \int_u^\infty \exp\left(\frac{-t^2}{2\Upsilon^2(t/(3\Upsilon) + 2n)}\right)\ \mathrm{d}t \tag{3.109}$$

$$\leqslant u + 2N_\delta \left(\int_u^{6n\Upsilon} \exp\left(\frac{-t^2}{8n\Upsilon^2}\right)\ \mathrm{d}t + \int_{6n\Upsilon}^\infty \exp\left(\frac{-t}{4/3\Upsilon}\right)\ \mathrm{d}t\right) \tag{3.110}$$

$$\leqslant u + 2N_\delta \left(\frac{8n\Upsilon}{2u} \exp\left(\frac{-u^2}{8n\Upsilon}\right) + \frac{4\Upsilon}{3} \exp\left(\frac{-24n\Upsilon}{3\Upsilon}\right)\right) \tag{3.111}$$

Remark 3.46. Here is another small mistake in [7] eq. (C.53) which corresponds to eq. (3.108) above. The mistake occurs in the application of Bernsteins inequality, where a factor of 2 is missing in front of n. This is the reason why [7] splits the integral at $t = 3\Upsilon n$ rather than at $t = 6\Upsilon n$ as we do here.

where we use lemma A.30 from eq. (3.110) to eq. (3.111). Now set $u = \Upsilon\sqrt{8n \log N_\delta}$. Continuing from eq. (3.111) we get

$$\cdots = \Upsilon\sqrt{8n \log N_\delta} + \frac{\Upsilon^2 8nN_\delta}{\Upsilon\sqrt{8n \log N_\delta}} \exp(-\log N_\delta) + 8/3N_\delta\Upsilon \exp(-9/2n) \tag{3.112}$$

$$= \Upsilon 2\sqrt{2n}\left(\log N_\delta + \frac{1}{\log N_\delta}\right) + 8/3N_\delta e^{-9/2n} \tag{3.113}$$

$$\leqslant 4\sqrt{2}\Upsilon\sqrt{n \log N_\delta} + 8/3\Upsilon \tag{3.114}$$

Combining eq. (3.102), eq. (3.98) and eq. (3.88) and recalling the definition of $\omega(\mathcal{F})$ (see eq. (3.10)) we get

$$\left\|\widehat{Q} - TQ\right\|_\nu^2 \leqslant \frac{1}{n}\left\|\widehat{Q} - TQ\right\|^2 + \frac{\Upsilon}{n}\mathbb{E}T + 8V_{\max}\delta \tag{3.115}$$

$$\leqslant \frac{1}{n}\left\|\widehat{Q} - TQ\right\|^2 + \frac{1}{n}\left(4\sqrt{2}\Upsilon^2\sqrt{n \log(N_\delta)} + 8/3\Upsilon^2\right) + 8V_{\max}\delta \tag{3.116}$$

$$\leqslant \frac{1}{n}\left\|\widehat{Q} - TQ\right\|^2 + C_3\sqrt{\log(N_\delta)/n} + C_4 n^{-1} + 8V_{\max}\delta \tag{3.117}$$

$$\leqslant \frac{(1+\kappa)^2}{\kappa} C_2^2 V_{\max}^2 \log(N_\delta)/n + (1+\kappa)\omega(\mathcal{F}) + (1+\kappa)2C_2V_{\max}\sqrt{\log(N_\delta)}\delta$$
$$+ C_3\sqrt{\log(N_\delta)/n} + C_4 n^{-1} + 8V_{\max}\delta \tag{3.118}$$

where $C_3 = 4\sqrt{2}\Upsilon^2$ and $C_4 = 8/3\Upsilon^2$. Now we set $\kappa = 1$ and $\delta = 1/n$ to det

$$\left\|\widehat{Q} - TQ\right\|_\nu^2 \leqslant 4C_2^2V_{\max}^2 \log(N_{(1/n)})/n + 2\omega(\mathcal{F}) + 4C_2V_{\max}\sqrt{\log(N_{(1/n)})}/n$$
$$+ C_3\sqrt{\log(N_{(1/n)})/n} + C_4n^{-1} + 8V_{\max}n^{-1} \tag{3.119}$$

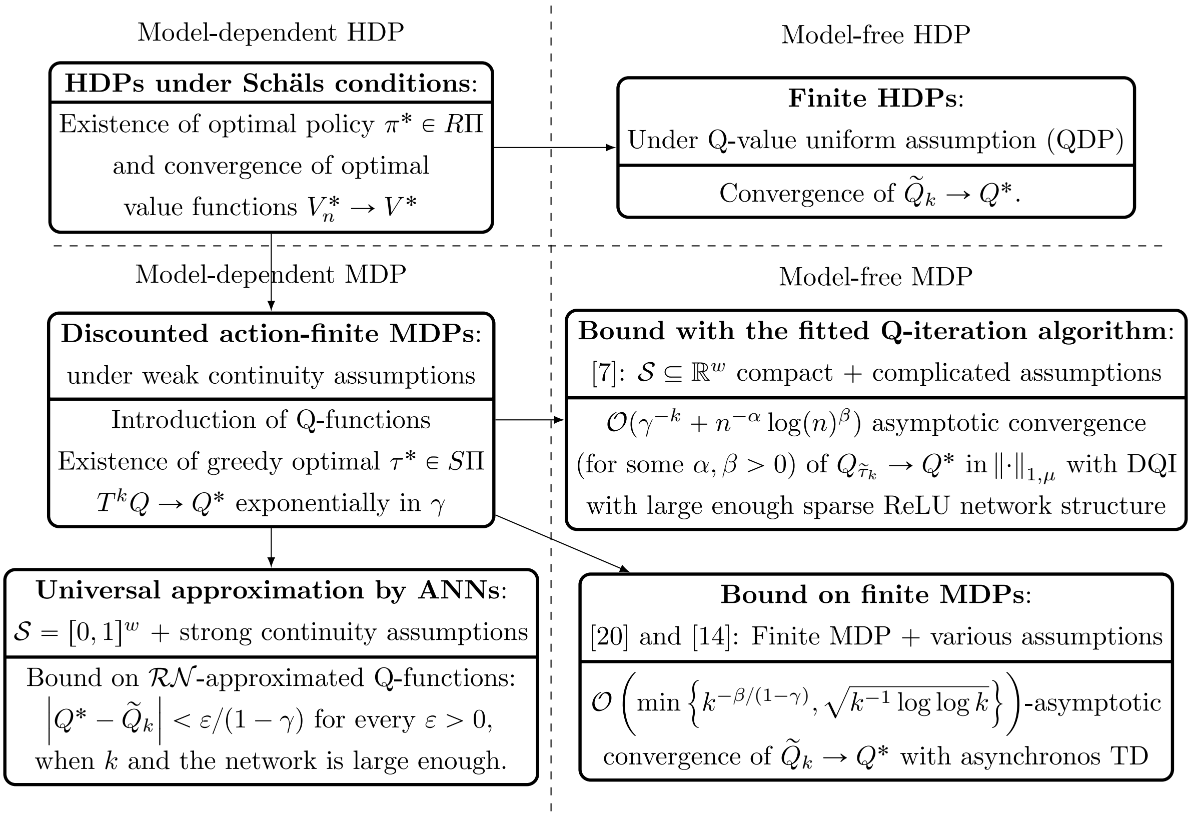
$$\leqslant C_5 \log(N_\delta)/n + 2\omega(\mathcal{F}) + C_3\sqrt{\log(N_{(1/n)})/n} + C_6n^{-1} \tag{3.120}$$

where $C_5 = 4C_2^2V_{\max}^2$ and $C_6 = C_4 + 8V_{\max}$, finishing the proof. □

Chapter 4

Comparison and conclusion

4.1 Comparison



Model-dependent HDP
Model-free HDP
HDPs under Schäls conditions:
Existence of optimal policy $\pi^* \in R\Pi$ and convergence of optimal value functions $V_n^* \to V^*$
Finite HDPs:
Under Q-value uniform assumption (QDP)
Convergence of $\widetilde{Q}_k \to Q^*$.
Model-dependent MDP
Model-free MDP
Discounted action-finite MDPs: under weak continuity assumptions
Introduction of Q-functions
Existence of greedy optimal $\tau^* \in S\Pi$
$T^k Q \to Q^*$ exponentially in γ
Bound with the fitted Q-iteration algorithm:
[7]: $\mathcal{S} \subseteq \mathbb{R}^w$ compact + complicated assumptions
$\mathcal{O}(\gamma^{-k} + n^{-\alpha}\log(n)^{\beta})$ asymptotic convergence (for some $\alpha, \beta > 0$) of $Q_{\widetilde{\tau}_k} \to Q^*$ in $\|\cdot\|_{1,\mu}$ with DQI with large enough sparse ReLU network structure
Universal approximation by ANNs:
$\mathcal{S} = [0,1]^w$ + strong continuity assumptions
Bound on $\mathcal{RN}$-approximated Q-functions:
$\left|Q^* - \widetilde{Q}_k\right| < \varepsilon/(1-\gamma)$ for every $\varepsilon > 0$,
when k and the network is large enough.
Bound on finite MDPs:
[20] and [14]: Finite MDP + various assumptions
$\mathcal{O}\left(\min\left\{k^{-\beta/(1-\gamma)}, \sqrt{k^{-1}\log\log k}\right\}\right)$-asymptotic convergence of $\widetilde{Q}_k \to Q^*$ with asynchronos TD

Figure 4.1: Some of the convergence bound presented in this thesis. Arrows $A \to B$ implies that setting B is a special case of setting A.

Overall every result provide insights to some classes of decision processes which is not covered by any of the other. For the history dependent processes we have convergence guarantees in the finite Q-value uniform (QDP) setting, as proved in [14]. This specializes to finite model-free MDPs, but presents no new insight in this setting, as such a result was already covered in [23]. It does however provide evidence that results on Q-learning may be extended to a history dependent setting.

Turning to MDPs, as we have discussed, the exponential convergence of the model-dependent Q-iteration while impressive is often not practical and so should be seen more as a theoretical background for the other results. The model-dependent result using ReLU network approximated Q-functions in the continuous setting shows that model-dependent methods, when applicable, provide the same exponential convergence to a near-optimal approximated Q-function. This is comparable to the convergence result by [7], except that it is less complicated and requires less assumptions except for the critical assumption of model-dependency. However because of it being model dependent, it not clear how much information about the model-free setting it implies.

The result by [7] of DQI has the advantage of being model-free and working with a continuous state space, unlike the other results. Another advantage is that it succeeds in evaluating the distance to the actual policy evaluation $Q_{\widetilde{\tau}}$ which provides better evidence of the nearness to optimality of its output policy, compared to all the other results, which only measures the distance to the approximating Q-functions. A weak point is however that it measures distance in the $\|\cdot\|_{1,\mu}$ norm, which is a bit non-standard and weaker than the rest of the results, which measures distance in $\|\cdot\|_{\infty}$. Another weak point is that the result and its assumptions are quite complicated and it is not clear how big the it includes are.

Specializing to finite model-free MDPs we get much better convergence rates in [20] even using the $\|\cdot\|_{\infty}$ norm. This is however at the expense requiring finiteness and a full table of state-action values instead of an approximating scheme. Also distance is measured between Q^* and $\widetilde{Q}_K$ rather than $\widetilde{Q}_{\widetilde{\pi}_K}$ as in [7].

4.2 Conclusion

In this thesis 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 [7]. 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 success of the DQN of [16] we are not much further in understanding. The major reason is that the computational aspects are so important to their success, and this part is mostly ignored in the results we have covered. Even though we establish results of the related FQI algorithm in [7], it is unclear if it captures the critical aspects of DQN, such as experience replay. In [7] 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. However it is also possible that DQN can be seen largely as an instance of FQI and that the results we have covered from [7] does explain part of its success.

4.3 Further directions

Examples

In this thesis we talk little about concrete decision processes, only applying our theory on the rather trivial gridworld environment in example 2.60. Examples are important drivers much of both theoretical and more empirical research in RL. Applying the theory in this thesis on concrete cases would be a interesting next step.

Relation between FQI and DQN

Find a way to prove or disprove the conjecture in [7] that convergence bounds on the FQI algorithm can have implications for the DQN algorithm. Answering this question probably requires research beyond what can be found in litterature already, as otherwise [7] would have drawn on such sources.

Application of the results of [15, Melo and Ribeiro (2007)]

Though [15] does not make explicit bounds for the distance of an approximating Q-function to Q^*, its results provide strong theoretical results of model-free Q-learning given a finite set of basis functions. Such a finite set of basis functions is provided by e.g. polynomials. It could be interesting to combine the results of [15] with various results in approximation theory such as theory on Bernstein polynomials and compare the resulting bounds of convergence to the bounds in [7].

Suboptimality of policies

This is relating to decision processes and value functions. Through out this thesis 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 [7] succeed in bounding $\|Q^* - Q_{\widetilde{\pi}}\|_\infty$, while most others is satisfied 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 2.76). 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 [17, Rønn-Nielsen and Hansen (2014)] and [11, Kallenberg (2002)]. A good survey on results on optimal policy existence in the special case of Markov decision processes can be found in [8, 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 family for love and support.

Chapter A

Appendices

A.1 Basic definitions and results

Definition A.1 (Interior)**.** For a subset $A \subseteq \mathcal{X}$ of a topological space $(\mathcal{X}, \mathcal{O}_{\mathcal{X}})$ the **interior** $A^\circ \subseteq A$ of A is the union of all open sets $U \in \mathcal{O}_{\mathcal{X}}$ which are contained in A. That is

$$A^\circ = \bigcup_{U \in \mathcal{U}} U, \text{ where } \mathcal{U} = \{U \in \mathcal{O}_{\mathcal{X}} \mid U \subseteq A\}$$

Definition A.2 (Order Topology)**.** Given a totally ordered set $(\mathcal{X}, <)$ the **order topology** is the topology generated by the subbase of sets on the form

$$\left\{x \mid a < x\right\},\ a \in \mathcal{X} \text{ and } \left\{x \mid x < b\right\}, b \in \mathcal{X}$$

Definition A.3 (σ-algebra)**.** A **σ-algebra** Σ on a set $\mathcal{X}$ is a pavement (family of subsets of $\mathcal{X}$) $\Sigma \subseteq 2^{\mathcal{X}}$ (where $2^{\mathcal{X}}$ denotes the powerset of $\mathcal{X}$) satisfying

- $\emptyset,\ \mathcal{X} \in \Sigma$.
- $A \in \Sigma \implies \mathcal{X} \backslash A \in \Sigma$.
- If $A_1, A_2, \dots \in \Sigma$ are a countable collection of subsets of $\mathcal{X}$ in Σ then $\bigcup_{i \in \mathbb{N}} A_i \in \Sigma$.

The pair $(\mathcal{X}, \Sigma)$ of a set and a σ-algebra on it is called a **measurable space**.

Theorem A.4. For any pavement $\Gamma \subseteq 2^{\mathcal{X}}$ of a set $\mathcal{X}$ there exists a *smallest* σ-algebra $\Sigma \subseteq 2^{\mathcal{X}}$ on $\mathcal{X}$ satisfying

1. $\Gamma \subseteq \Sigma$.
2. For any σ-algebra Σ' for which $\Gamma \subseteq \Sigma'$ it holds that $\Sigma \subseteq \Sigma'$.

This smallest σ-algebra is denoted $\sigma(\Gamma)$.

Definition A.5 (Borel σ-algebra)**.** For a topological space the **Borel** σ-algebra is the smallest σ-algebra containing all open sets.

Definition A.6 (Product σ-algebra)**.** Let $(\mathcal{X}_i, \mathcal{A}_i)_{i\in I}$ be a collection of measurable spaces. the product σ-algebra

$$\bigotimes_{i\in I} \mathcal{A}_i$$

is the smallest σ-algebra making all coordinate projections $\rho_i : \prod_{j\in I} \mathcal{X}_j \to \mathcal{X}_i$ measurable. In particular if $|I| = 2$

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma\left(\left\{A_1 \times \mathcal{X}_2 \mid A_1 \in \mathcal{A}_1\right\} \cup \left\{\mathcal{X}_1 \times A_2 \mid A_2 \in \mathcal{A}_2\right\}\right)$$

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.

Definition A.9 (Measure)**.** Given a measurable space $(\mathcal{X}, \Sigma)$ a **measure** is a function $\mu : \Sigma \to [0, \infty]$ satisfying

1. $\mu(\varnothing) = 0$
2. $\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i)$ for any countable collection of mutually disjoint sets $A_1, A_2, \cdots \in \Sigma$.

If there exists a sequence of subsets $A_1 \subseteq A_2 \subseteq \cdots \subseteq \mathcal{X}$ with $\bigcup_{i\in\mathbb{N}} A_i = \mathcal{X}$ and $\mu(A_i) < \infty$ for all $i \in \mathbb{N}$ then μ is called **σ-finite**. If $\mu(\mathcal{X}) < \infty$ then μ is called **finite**, and if furthermore $\mu(\mathcal{X}) = 1$ then μ is called a **probability measure**.

Theorem A.10 (Carathéodory's extension theorem)**.** Let $\mathcal{X}$ be a set and $\mathcal{S} \subset 2^{\mathcal{X}}$ be a pavement of $\mathcal{X}$ satisfying

1. $\varnothing \in \mathcal{X}$
2. $S,\ T \in \mathcal{S} \implies S \cap T \in \mathcal{X}$
3. For $S,\ T \in \mathcal{S}$ there exists finitely many disjoint subsets $S_1, S_2, \ldots, S_n \in \mathcal{S}$ so that $S \backslash T = \bigcup_{i=1}^{n} S_i$.

($\mathcal{S}$ is then called a *semi-ring*). Let $\mu : \mathcal{S} \to [0, \infty]$ be a function satisfying

i. $\mu(\varnothing) = 0$

ii. For a countable mutually disjoint collection of subsets $S_1, S_2, \cdots \in \mathcal{S}$ it holds that $\mu\left(\bigcup_{i\in\mathbb{N}} S_i\right) = \sum_{i\in\mathbb{N}} \mu(S_i)$.

Then μ has an extension to a measure μ on $\sigma(\mathcal{S})$. Furthermore if there exists an increasing sequence of subsets $S_1 \subseteq S_2 \subseteq \cdots \in \mathcal{S}$ of $\mathcal{S}$ satisfying $\bigcup_{i\in\mathbb{N}} S_i = \mathcal{X}$ and $\mu(S_i) < \infty$ for all $i \in \mathbb{N}$ then the extension is unique. In particular if $\mathcal{X} \in \mathcal{S}$ and $\mu(\mathcal{X}) = 1$ then μ extends uniquely to a probability measure on $(\mathcal{X}, \sigma(\mathcal{S}))$.

Definition A.11 (Measurable function)**.** A functions $f : \mathcal{X} \to \mathcal{Y}$ between two measurable spaces are called **measurable** if

$$f^{-1}(\Sigma_{\mathcal{Y}}) = \left\{ f^{-1}(B) \mid B \in \Sigma_{\mathcal{Y}} \right\} \subseteq \Sigma_{\mathcal{X}}$$

The set of such functions we denote $\mathcal{M}(\Sigma_{\mathcal{X}}, \Sigma_{\mathcal{Y}})$ or $\mathcal{M}(\mathcal{X}, \mathcal{Y})$.

Definition A.12 (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.13 (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

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

Definition A.14. A sequence of events $A_1, A_2, \cdots \subseteq \Omega$ is said to be **asymptotically almost sure** if $\mathbb{P}(A_k) \to 1$ for $k \to \infty$.

Example A.15. For example if $U_1, U_2, \cdots \sim \text{Unif}(0,1)$ are i.i.d. random variables, $X_k = \max_{i\in[k]} U_i$ for $k \in \mathbb{N}$ and $\varepsilon > 0$ then the events $(A_k)_{k\in\mathbb{N}} = (X_k > 1 - \varepsilon)_{k\in\mathbb{N}}$ are asymptotically almost sure since $\mathbb{P}(A_k) \to 1$ as $k \to \infty$. The property $X_k > 1 - \varepsilon$ is then said to hold *asympotically almost surely.*

Proposition A.16. $\text{id}_{\mathcal{P}(X)} = \mu \mapsto \kappa \circ \mu$ where $\kappa(\cdot \mid x) = \delta_x(\cdot)$. Thus κ can be seen as an identity mapping on $\mathcal{P}(X)$.

Proof.

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

□

Definition A.17 (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.18 (Differentiability in one variable)**.** A function $f : A \to \mathbb{R}$ where $A \subseteq \mathbb{R}$ is an open subset of the real numbers is **differentiable** at $x \in \mathbb{R}$ if the **derivative**

$$f'(x) := \lim_{x_n \to x} \frac{f(x) - f(x_n)}{x - x_n}$$

exists, is finite and is the same for any sequence $(x_n)_{n\in\mathbb{N}} \subseteq A$ converging to x with $x_n \neq x$ for all $n \in \mathbb{N}$. If f is differentiable at A if it is differentiable for every $x \in A$. If $f' : A \to \mathbb{R}$ is continuous then we write $f \in C^1(A)$. If $f'' = (f')' : A \to \mathbb{R}$ exists and is continuous we write $f \in C^2(A)$. Like this for $k \in \mathbb{N}_0$ we say that C^k is the set of k times continuously differentiable functions, and we write $f^{(k)}$ for the kth derivative, when $k = 0$ we have $C^0(A) = C(A)$ the set of continuous functions and $f^{(0)} = f$. This extends to C^∞, called the set of **smooth** functions, for any element is continuously differentiable n times for any $n \in \mathbb{N}_0$.

Definition A.19 (Partial derivatives)**.** Let $f : U \to \mathbb{R}$ where $U \subseteq \mathbb{R}^n$ is open be a function satisfying for some $x = (x_1, \dots, x_n) \in U$ that $f_{x,i} = x_i \mapsto f(x_1, \dots, x_i, \dots, x_n) \in C^1(\rho_i(U))$ where $\rho_i : U \to \mathbb{R}$ is projection onto the ith coordinate. The partial derivative of f with respect to the ith variable at x is the function $\delta_i f(x) := f'_{x,i}(x_i)$. For $k \in \mathbb{N}_0$ if $f_{x,i} \in C^k$ then write $\delta_i^k f(x) := f^{(k)} f_{x,i}(x_i)$ whenever this exists. If $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$ we denote by $\delta^\alpha f(x) := \delta_1^{\alpha_1} \dots \delta_n^{\alpha_n} f(x)$.

Remark A.20. A standard result called *Schwartz's theorem* say that the order in which partial derivatives are taken does not matter when these such derivates are continuous.

Definition A.21 (Differentiability in $\mathbb{R}^n$)**.** A function $f : U \to \mathbb{R}$ defined on an open set $U \subseteq \mathbb{R}^n$ is said to be C^k for $k \in \mathbb{N}_0$ if the partial derivatives $\partial^\alpha f : U \to \mathbb{R}$ exists and is continuous for all $\alpha \in \mathbb{N}_0^n$ with $\|\alpha\|_1 = \alpha_1 + \cdots + \alpha_n \leqslant k$.

Definition A.22 (Absolutely continuity of measures)**.** Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ be σ-finite measures then μ is said to be **absolutely continuous** with respect to ν, written $\mu << \nu$ if for all $A \in \Sigma_{\mathcal{X}}$ we have $\nu(A) = 0 \implies \mu(A) = 0$.

Theorem A.23 (Radon-Nikodym)**.** Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ with $\mu << \nu$. Then there exists a positive measurable function $f : \mathcal{X} \to [0, \infty)$ such that $\mu(A) = \int_A f \, \mathrm{d}\nu$. This function is denoted $f = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$.

Theorem A.24 (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 $d(T^k x, x^*) < \gamma^k d(x, x^*)$.

A.1.1 Geometric ergodicity

To understand geometric ergodicity 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 A.25 (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 A.26 (Positivity)**.**

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

Definition A.27 (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 A.28 (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 A.29 (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

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

A.2 Proofs of auxiliary results

Proof of proposition 2.10. Let $x \in \mathcal{X}$, $B \in \Sigma_{\mathcal{Y}}$ and $C \in \Sigma_{\mathcal{Z}}$. First of all $g_{x,B\times C} = y \mapsto 1_B(y)\phi(C \mid x, y)$ is measurable since it is the product (multiplication is measurable) of the measurable functions 1_B and $\phi(C \mid x, y)$ (measurability of ϕ comes from definition 2.1.2 since ϕ is a probability kernel). Further $g_{x,B\times C} \in [0,1]$ so it is $\kappa(\cdot \mid x)$-integrable. We also have $g_{x,\mathcal{Y}\times\mathcal{Z}} = 1$ and $g_{x,\varnothing} = 0$. Since $\{B \times C \mid B \in \Sigma_{\mathcal{Y}}, C \in \Sigma_{\mathcal{Z}}\}$ is a semi-ring generating $\Sigma_{\mathcal{Y}} \otimes \Sigma_{\mathcal{Z}}$ by theorem A.10 we have that $\phi\kappa(\cdot \mid x)$ extends uniquely to a probability measure on $\mathcal{Y} \times \mathcal{Z}$. Since $(x, y) \mapsto 1_B(y)\phi(C \mid x, y)$ is measurable by proposition 2.4 we have that $\phi\kappa(B \times C \mid x)$ is measurable in x. We have now shown that $\phi\kappa$ is a probability kernel.

We now show associativity of composition with measures, i.e. that $(\phi\kappa)\mu = \phi(\kappa\mu)$ when $\mu \in \mathcal{P}(\mathcal{X})$. Let $A \in \Sigma_{\mathcal{X}}$ then

$$\begin{aligned}
(\phi\kappa)\mu(A \times (B \times C)) &\overset{2.7}{=} \int_A \phi\kappa(B \times C \mid x) \,\mathrm{d}\mu(x) \\
&= \int_A \int 1_B(y)\phi(C \mid x, y) \,\mathrm{d}\kappa(y \mid x) \,\mathrm{d}\mu(x) \\
&= \int \int 1_{A\times B}(x, y)\phi(C \mid x, y) \,\mathrm{d}\kappa(y \mid x) \,\mathrm{d}\mu(x) \\
&\overset{2.9}{=} \int \int 1_{A\times B}(x, y)\phi(C \mid x, y) \,\mathrm{d}\kappa\mu(x, y) \\
&= \phi(\kappa\mu)((A \times B) \times C)
\end{aligned}$$

The associativity of the product of three kernels is left as an exercise.

As a preliminary lemma to the last statement notice that by theorem 2.9 (Fubini)

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

Therefore again by Fubini and the property of integration over the Dirac measure

$$\begin{aligned}
\int f(x, y, z) \,\mathrm{d}\phi\kappa(y, z \mid x) &= \int f(x', y, z) \,\mathrm{d}\phi(\kappa\delta_x)(x', y, z) \\
&\overset{2.9}{=} \int f(x', y, z) \,\mathrm{d}\phi(z \mid x', y) \,\mathrm{d}\kappa\delta_x(x', y) \\
&\overset{2.9}{=} \int \int f(x', y, z) \,\mathrm{d}\phi(z \mid x', y) \,\mathrm{d}\kappa(y \mid x') \,\mathrm{d}\delta_x(x') \\
&= \int f(x, y, z) \,\mathrm{d}\phi(z \mid x, y) \,\mathrm{d}\kappa(y \mid x)
\end{aligned}$$

□

A.3 Lemmas for [7, Fan et al. (2020+)]

Lemma A.30. 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.4 Disambiguation

- $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ set of extended real numbers.
- $\mathrm{id}_X := x \mapsto x$ the identity function on X.
- $[\phi] := \begin{cases} 1 & \phi \\ 0 & \neg\phi \end{cases}$: 0-1 indicator for logical formulas.
- $[q] := \{1, \dots, q\}$ for $q \in \mathbb{N}$.
- $1_A(a) := [a \in A]$: the indicator function.
- $C(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{K} \mid f \text{ continuous}\}$ set of continuous real-valued functions on $\mathcal{X}$.
- ANN: abbreviation for artificial neural network see definition 2.71.
- δ_a: Dirac-measure of point a, i.e. $\delta_a(A) = [a \in A] = 1_A(a)$.
- $(\Omega, \Sigma_\Omega, \mathbb{P})$: 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.
- $\mathcal{P}(\mathcal{X})$: the set of all probability measures on a measurable space $(\mathcal{X}, \Sigma_\mathcal{X})$.
- $2^\mathcal{X}$: the powerset of the set $\mathcal{X}$.
- $\mathcal{M}(\mathcal{X}, \mathcal{Y})$: set of $\Sigma_\mathcal{X}$-$\Sigma_\mathcal{Y}$ measurable functions.
- $\mathcal{L}_p(\mathcal{X}) = \left\{f : \mathcal{X} \to \mathbb{R} \;\middle|\; \int |f|^p \ \mathrm{d}\mu < \infty,\ \forall \mu \in A\right\}$ the set of functions which are (p, μ)-integrable for all measures μ in a certain set of measures A, see definition 2.42.
- $\mathbb{E}, \mathbb{E}_\mu$: expectation, that is integration w.r.t. the measure $\mathbb{P}$ or μ respectively.
- $\mathbb{V}$: variance operator.

Bibliography

[1] Martin Anthony and Peter Bartlett. *Neural Network Learning: Theoretical Foundations.* 01 2002. ISBN 978-0-521-57353-5. doi: 10.1017/CBO9780511624216.

[2] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983.

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

[4] 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.

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

[6] Eyal Even-dar and Yishay Mansour. Learning rates for q-learning. volume 5, 04 2001. doi: 10.1007,3-540-44581-1-39.

[7] 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`.

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

[9] C. Heitzinger. *Simulation and Inverse Modeling of Semiconductor Manufacturing Processes.* 2002. URL `https://books.google.dk/books?id=LpmxcQAACAAJ`.

[10] 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.

[11] 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`.

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

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

[14] 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`.

[15] 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.

[16] 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`.

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

[18] 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`.

[19] 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`.

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

[21] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, 11 2010. doi: 10.1017/CBO9780511794308.006.

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

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