

Department of Mathematical Sciences  
UNIVERSITY OF COPENHAGEN



© 2020 Jacob Harder

JACOB HARDER

THEORETICAL ASPECTS OF Q-LEARNING

Thesis submitted to the University of Copenhagen for the degree of Master of Science

in Mathematics

in 2020

by

Jacob Harder

MASTER THESIS IN MATHEMATICS

DEPARTMENT OF MATHEMATICAL SCIENCES

UNIVERSITY OF COPENHAGEN

Thesis supervisor: Professor Stefan Horst Sommer

ADVISORS

STEFAN HORST SOMMER

MAY 29, 2020



## Abstract

This paper is mainly about the part of reinforcement learning that is called Q-learning, which is a category of algorithms which can *learn* from interaction with a decision process. We present the background theory for these algorithms, including basic theory on stochastic (decision) processes, existence of optimal policies and dynamic programming in the form of value iteration. Then Q-learning is analysed in a variety of settings. Among results other results we present and prove a convergence bound of the FQI algorithm obtained in the preprint [6, 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

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Motivation . . . . .	3
1.2	What is Reinforcement Learning? . . . . .	3
1.3	What is Q-learning? . . . . .	5
1.4	Basic concepts and notation . . . . .	5
<b>2</b>	<b>Decision models and value functions</b>	<b>6</b>
2.1	Policy evaluation and value functions . . . . .	10
2.1.1	The optimal value function . . . . .	12
2.2	Markov decision processes . . . . .	15
2.2.1	Greedy policies . . . . .	18
2.2.2	Existence of optimal policies . . . . .	21
2.2.3	Value iteration . . . . .	22
2.2.4	Q-functions . . . . .	25
2.2.5	Why are we not done? . . . . .	28
2.3	Model-based Q-learning . . . . .	28
2.3.1	Algorithmic and approximation errors . . . . .	29
2.3.2	Using artificial neural networks . . . . .	30
2.3.3	Using Bernstein polynomials . . . . .	33
<b>3</b>	<b>Model-free algorithms</b>	<b>36</b>
3.1	Finite case . . . . .	36
3.1.1	History dependent setting . . . . .	39
3.2	Linear function approximation . . . . .	41
3.3	Deep Q-learning . . . . .	43
3.3.1	Introduction . . . . .	43
3.3.2	Assumptions . . . . .	45
3.3.3	The main theorem . . . . .	46
3.3.4	Relation to DQN . . . . .	47
3.3.5	Differences in notation . . . . .	49
3.3.6	Proofs . . . . .	49
<b>4</b>	<b>Comparison and conclusion</b>	<b>63</b>
4.1	Comparison . . . . .	63
4.2	Conclusion . . . . .	64
4.3	Further directions . . . . .	64
4.4	Notes on references . . . . .	65
4.5	Credits . . . . .	65
<b>A</b>	<b>Appendices</b>	<b>66</b>
A.1	Basic definitions . . . . .	66
A.2	Proofs of auxiliary results . . . . .	68
A.3	Lemmas for [6, Fan et al. (2020+)] . . . . .	69
A.4	Disambiguation . . . . .	70

# 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 [20] to a result obtained in the preprint [6, 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 [15, Mnih et al. (2015)]. In [15] 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 [15] is based on a Q-learning algorithm called the *deep Q-network* (DQN) algorithm.

The purpose of this thesis is to investigate what has been proven about the convergence of the DQN algorithm and what mathematical theory is relevant to establish such proofs. This is however not an easy task, as we will only cover a small fraction of the topic.

The work began by considering the preprint [6, Fan et al. (2020+)] which claims to establish theoretical justification for the convergence of DQN to *the optimal Q-function*. In the course of reading [6] it became clear that the background context of dynamic programming, reinforcement learning and value iteration was essential to understand the results of [6] and similar papers and also compare such results. Also questions as to in which settings optimal policies exists turned out to be a non-trivial question. Therefore in the end this thesis is partly about presenting the results of [6] and similar papers. And partly to build the background theory necessary to understand and compare these results.

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

### 1.2 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  $\pi$ , 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 sum of rewards called the *value function*  $V_\pi$  of the policy. The goal of reinforcement learning is to find an optimal policy  $\pi^*$ , maximizing the value function.

$V_\pi$  is viewed as a function that evaluates for each *starting state*  $s \in \mathcal{S}_1$  the expected total rewards when starting in state  $s$  and following policy  $\pi$ . There might therefore be different optimal policies for each such starting state. Traditionally one defines an optimal value function  $V^*(s)$  by taking the supremum over all policies  $\sup_\pi V_\pi(s)$  for every state  $s \in \mathcal{S}_1$ . Then an optimal policy  $\pi^*$  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 to 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  $\pi_1$  by choosing actions leading to the maximum average values (according to  $V_1$ ). Such policies are called *greedy policies*. We can then evaluate the value of  $\pi_1$  yielding a new value function  $V_2$ . This leads to an operator on the space of value functions called the *Bellman operator* and should satisfy  $TV_1 = V_2$ . This process of applying the Bellman operator can be continued indefinitely yielding a sequence of value functions and policies. Variations of this idea are called *value iteration* and *policy iteration*, and is derived from dynamic programming. 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 RL as dynamic programming. To go beyond this, RL commonly 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).

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

### 1.3 What is Q-learning?

A problem with value functions defined on the set of states  $\mathcal{S}$  is that picking optimal actions require knowledge of the transition dynamics  $P$ . 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 [15] 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.

### 1.4 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.1). This in turn give rise to the standard Borel  $\sigma$ -algebra (definition A.4)  $\mathbb{B} = \sigma(\mathcal{O})$  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  $\sigma$ -algebra  $\Sigma_{\mathcal{X}}$  when not ambiguous. We always consider the cartesian product of measurable spaces with the product  $\sigma$ -algebra (definition A.5) unless otherwise specified. We denote the set of measurable functions (definition A.10)  $\mathcal{X} \rightarrow \mathcal{Y}$  between two measurable spaces by  $\mathcal{M}(\Sigma_{\mathcal{X}}, \Sigma_{\mathcal{Y}})$  or  $\mathcal{M}(\mathcal{X}, \mathcal{Y})$  when the  $\sigma$ -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  $X : \Omega \rightarrow \mathcal{X}$  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  $f_1, f_2, \dots : \mathcal{X} \rightarrow \mathbb{R}$  limits are always understood pointwise, unless otherwise stated, meaning that  $f_n \rightarrow f$  is to be read as  $\forall x \in \mathcal{X} : f_n(x) \rightarrow f(x)$ . 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

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

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  $\Sigma_{\mathcal{X}}\text{-}\Sigma_{\mathcal{Y}}$  measurable for any  $B \in \Sigma_{\mathcal{Y}}$ .

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

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

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

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

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

for some measure  $\mu$  on  $\mathcal{Y}$  then  $\kappa(B \mid x) = \int_B f(x, y) \, d\mu(y)$  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  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  be measurable satisfying that  $f(x, \cdot)$  is  $\kappa(\cdot \mid x)$ -integrable for every  $x \in \mathcal{X}$ . Then the map

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

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  $A \in \Sigma_{\mathcal{Y}}$  are measurable by definition 2.1.2 since  $\kappa$  is a kernel. Then extend by sums and limits.  $\square$

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

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

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

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

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 admissible 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 admissible actions (with probability 1). The set of all policies we denote  $\Pi$ .

With a HDP, and some distribution  $\mu \in \mathcal{P}(\mathcal{S}_1)$  of the *starting state*  $S_1 \sim \mu$  and some policy  $\pi$  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

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

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

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

Immediately 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

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

Uniqueness follows because the property

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

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

**Remark 2.8.** In light of theorem 2.7 we can view a probability kernel as a mapping  $\kappa : \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X} \times \mathcal{Y})$  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

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

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

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

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

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

*Proof.* We refer to [16] thm. 1.3.2 and 1.3.3.  $\square$

**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  $\phi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$  satisfying

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

called the **composition** of  $\phi$  and  $\kappa$ . 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  $f : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}$  is a measurable function such that  $f(x, \cdot, \cdot)$  is  $\phi\kappa(\cdot \mid x)$ -integrable then

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

*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  $\varphi'$  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. In fact  $\circ$  makes the class of measurable spaces into a category [12, Lawvere (1962)], with identity  $\text{id}_{\mathcal{X}}(\cdot \mid x) = \delta_x$ .

### 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  $\kappa_n : \mathcal{X}^n \rightsquigarrow \mathcal{X}_{n+1}$  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  $n$ th 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  $\rho \in \mathcal{P}(\mathcal{X}^{\infty})$  such that

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

*Proof.* We refer to [10, Kallenberg (2002)] thm. 5.17. □

In particular theorem 2.13 applied to the measure Dirac-measure  $\delta_x$  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  $\mu_x$  denote the Ionescu-Tulcea measure of a sequence of probability kernels  $\kappa_i : \mathcal{X}^i \rightarrow \mathcal{X}_{i+1}$  with starting measure  $\delta_x$  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  $\mu_x$  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

$$\mathbb{G} = \{A \in \Sigma \mid \phi_A \in \mathcal{M}(\mathcal{X}_1, [0, 1])\}$$

The cylinder sets

$$\mathbb{O} = \{A_2 \times \dots \times A_i \times \mathcal{X}_{i+1}, \dots \mid A_i \in \Sigma_{\mathcal{X}_i}, i - 1 \in \mathbb{N}\}$$

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

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

and any  $\kappa^{i-1}$  is a kernel making that function measurable. We will show that  $\mathbb{G}$  is a Dynkin class. Then by Dynkins  $\pi$ - $\lambda$  theorem (see theorem A.7)

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

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

For showing that  $\mathbb{G}$  is a Dynkin class, notice that clearly  $\mathcal{X}_2 \times \mathcal{X}_3 \times \dots$  and  $\emptyset$  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  $\phi_{\bigcup_{n=1}^{\infty} B_n} = \lim_{n \rightarrow \infty} \phi_{B_n}$  is again measurable as it is a limit of measurable functions, showing that  $\mathbb{G}$  is a Dynkin class.  $\square$

We will denote the Ionescu-Tulcea kernel  $\dots \kappa_2 \kappa_1$  or  $\prod_{i=1}^{\infty} \kappa_i$  or simply  $\kappa^{\infty}$ . 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  $\kappa_n \dots \kappa_1 = (\kappa_n \dots \kappa_2) \kappa_1$ . Therefore for any  $n \in \mathbb{N}$  and  $A \in \Sigma_{\mathcal{X}_2} \otimes \dots \otimes \Sigma_{\mathcal{X}_n}$ , using theorem 2.13 we have

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

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

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  $\pi = (\pi_1, \pi_2, \dots) \in R\Pi$  gives rise to a kernel  $\kappa_{\pi} : \mathcal{P}(\mathcal{S}_1) \rightarrow \mathcal{P}(\mathcal{H}_{\infty})$ , namely

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

In particular  $\kappa_{\pi} \mu$  can be interpreted as the stochastic process arising from sampling the first state from  $\mu$  and then follow  $\pi$  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

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

Both of these measures can be interpreted as the stochastic process arising from starting in state  $s$  and following policy  $\pi$ . 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  $R_1, R_2, \dots$  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  $n \in \mathbb{N}$

$$r_n : \mathcal{H}_n \rightarrow \mathbb{R}, \quad r(h_n) = \int x \, dR_n(x \mid h_n)$$

called the  $n$ th **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  $V : \mathcal{S}_1 \rightarrow \mathbb{R}$  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  $V_{n, \pi} : \mathcal{S}_1 \rightarrow \mathbb{R}$  by

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

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

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

We would like to extend this to an infinite policy evaluation i.e. letting  $n$  tend to  $\infty$ . 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

$$V_{\pi}(s) = \mathbb{E}_s^{\pi} \lim_{n \rightarrow \infty} \sum_{i=1}^n r_i = \int \sum_{i=1}^{\infty} r_i \, d\kappa_{\pi} s \quad (2.2)$$

**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  $V_{\pi}$  measures the expected total reward after following the policy  $\pi$  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.  $\lim_{n \rightarrow \infty} V_{n,\pi} = V_\pi$  for all  $\pi \in R\Pi$ .

3. For any  $\pi \in R\Pi$

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

*Proof.*

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

2. By monotone or dominated convergence.

3. For any  $\pi \in R\Pi$

$$|V_\pi(s)| \leq \mathbb{E}_s^\pi \sum_{i \in \mathbb{N}} |r_i| \leq \sum_{i \in \mathbb{N}} \gamma^{i-1} R_{\max} = R_{\max}/(1 - \gamma)$$

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).

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

This is called the **optimal value function** (and the  $n$ th 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  $|V_k^*|, |V^*| \leq V_{\max}$  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 [2]). Perhaps this is not surprising 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 extent does an optimal policy  $\pi^*$  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  $\pi^*$ ?

We will wait with questions 3 and 4 until the next section. In a quite general setting, questions 1 and 2 is investigated in [18, 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 separable completely metrizable space, and  $\Sigma_{\mathcal{X}}$  is the Borel  $\sigma$ -algebra of  $\mathcal{X}$ , that is the  $\sigma$ -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.  $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N \geq n} \sup_{\pi \in R\Pi} \sum_{t=n+1}^N \mathbb{E}_s^{\pi} r_t \rightarrow 0, \quad n \rightarrow \infty$

**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  $\sum_{t=n+1}^N r_t \leq \gamma^n R_{\max}$ .

To understand the results of [18] 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  $\sigma$ -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  $\mathcal{P}(\mathcal{X})$

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

and where  $C(\mathcal{X})$  denote the set of continuous functions  $\mathcal{X} \rightarrow \mathbb{R}$ . 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 separable 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:

$$(a) \, p_n \rightarrow p \quad (b) \, \forall f \in C(\mathcal{X}) : \int f \, dp_n \rightarrow \int f \, dp \quad (c) \, \forall g \in U_d(\mathcal{X}) : \int g \, dp_n \rightarrow \int g \, dp$$

*Proof.* We refer to [2] 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 probability 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 [2] cor.7.25.1 and prop.7.22. □

**Definition 2.32** (Semicontinuity). Let  $\mathcal{X}$  be a topological space and  $f : \mathcal{X} \rightarrow \overline{\mathbb{R}}$  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 be **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  $f : \mathcal{X} \rightarrow \overline{\mathbb{R}}$  is upper semicontinuous if and only if for each sequence  $x_1, x_2, \dots \in \mathcal{X}$  with  $x_n \rightarrow x \in \mathcal{X}$  we have  $\limsup f(x_n) \leq f(x)$  (analogously  $\liminf f(x_n) \geq f(x)$  for lower semicontinuous  $f$ ).
2. If  $f, g : \mathcal{X} \rightarrow \overline{\mathbb{R}}$  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 [2] p. 147. □

We are now ready to present the results of [18]. 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.  $V_n^* \rightarrow V^*$  as  $n \rightarrow \infty$ .

*Proof.* We refer to [18]. □

**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  $\pi^*$  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 is 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 discount factor  $\gamma \in [0, 1)$  (see assumption 2).
6.  $\mathfrak{A}(s) \subseteq \mathcal{A}$  a set of admissible 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

- $\mathcal{S}_1 = \mathcal{S}_2 = \dots = \mathcal{S}$ ,  $\mathcal{A}_1 = \mathcal{A}_2 = \dots = \mathcal{A}$ .
- $P_n$  depends only on  $s_n$  and  $a_n$  and does not differ with  $n$ . That is  $P_n(\cdot \mid s_1, \dots, s_n, a_n) = P(\cdot \mid s_n, a_n)$  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 [2] 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  $\mathcal{S}_1, \mathcal{A}_1, \dots$ , 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

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

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

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



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  $\pi$  is called **deterministic** if all  $\pi_i$  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

$$\begin{array}{ccccc} S\Pi & \subseteq & M\Pi & \subseteq & R\Pi \\ \cup & & \cup & & \cup \\ DS\Pi & \subseteq & DM\Pi & \subseteq & DP\Pi \end{array}$$

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  $\varphi_n : \mathcal{H}_n \rightarrow \mathcal{A}$  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  $\pi_n$  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  $i$ th state space given some initial distribution  $\mu \in \mathcal{P}(\mathcal{S})$  and some policy  $\pi \in R\Pi$ . Here  $\rho_{\mathcal{S}_i} : \mathcal{H}_\infty \rightarrow \mathcal{S}_i$  is projection onto the  $i$ th state space.

**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  $\nu$  then it is enough to ensure boundedness  $\nu$ -almost everywhere.

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

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

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

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

The Bellman optimality operator  $T$  is harder to work with than  $T_\tau$  because it involves a supremum. Therefore we will first take a closer look at properties of  $T_\tau$ .

**Proposition 2.46** (Properties of the  $T_\tau$ -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.  $T_\tau$  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.  $V_\pi = \lim_{k \rightarrow \infty} T_{\tau_1} \dots T_{\tau_k} V_0$
4. For the stationary policy  $\tau$  we have  $T_\tau V_\tau = V_\tau$ .
5.  $T$  and  $T_\tau$  are  $\gamma$ -contractive on  $\mathcal{L}_\infty(\mathcal{S})$ .
6.  $V_\tau$  is the unique bounded fixed point of  $T_\tau$  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:

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

Now use this inductively.

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

$$|T^\tau V - T^\tau V'| = \gamma \left| \int V(s') - V'(s') \, dP_\tau(a, s' \mid \cdot) \right| \leq \gamma K$$

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

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  $T_\tau$  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. [2]).

---

**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.

Initialize the expected reward function  $r \leftarrow \int x \, dR(x \mid \cdot)$ .

Initialize the starting value estimator  $\tilde{V}_0 = 0$ .

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

    Update the value estimator  $\tilde{V}_{k+1} \leftarrow T_{\tau_{K-k}} \tilde{V}_k$ .

**Output:** The finite policy evaluation  $V_{\pi, K}$ .

---

The reason behind calling algorithm 1 *theoretical* is that we have not talked about how to actually represent any of  $r$ ,  $T_\pi$  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  $V_\tau$ . To this end algorithm 1 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  $k$ th finite policy evaluation  $V_{\pi, K}$ , that is the output of algorithm 1 with  $k$  iterations, satisfy

$$|V_\tau - V_{\tau, k}| \leq \|V_\tau - V_{\tau, k}\|_\infty \leq \gamma^k V_{\max}$$

It is in fact mostly due corollary 2.47 that algorithm 1 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  $V : \mathcal{S} \rightarrow \mathbb{R}$  be a measurable value-function. We define

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

as the set of **greedy** actions w.r.t.  $V$ . If 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  $\tau$  is called greedy w.r.t.  $V$ . We will often denote a  $V$ -greedy policy by  $\tau_V$ .

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

$$\operatorname{argmax}_{x \in A} f(x) = \left\{ x \in A \mid f(x) = \sup_{x' \in A} f(x') \right\}$$

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  $\sigma$ -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

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

is continuous.

**Setting 2** (Greedy MDP).

1.  $\mathcal{S}$  and  $\mathcal{A}$  are standard Borel.
2. The set of admissible actions  $A(s) \subseteq \mathcal{A}$  is compact for all  $s \in \mathcal{S}$  and  $\Gamma = \{(s, a) \in \mathcal{S} \times \mathcal{A} \mid a \in A(s)\}$  is a closed subset of  $\mathcal{S} \times \mathcal{A}$ .
3. The transition kernel  $P$  is continuous.
4. The expected reward function  $r = \int r' \, dR(r' \mid \cdot)$  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  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \overline{\mathbb{R}}$  be Borel measurable, bounded from below or above. Define

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

Then

- $f$  upper semicontinuous and bounded from above implies that  $\lambda$  is upper semicontinuous and bounded from above.
- $f$  lower semicontinuous and bounded from below implies that  $\lambda$  is lower semicontinuous and bounded from below.

*Proof.* We refer to [2] prop.7.31. □

**Proposition 2.52.** A upper (lower) semicontinuous function  $f : \mathcal{X} \rightarrow \overline{\mathbb{R}}$  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  $f(x_n) \rightarrow \sup_{x \in \mathcal{X}} f(x)$  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  $x_{k_n} \rightarrow x^*$ . Since  $f$  is upper semicontinuous

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

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  $f : \Gamma \rightarrow \overline{\mathbb{R}}$  be upper semicontinuous. Let

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

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

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

*Proof of proposition 2.53.* We refer to [2] 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  $V : \mathcal{S} \rightarrow \mathbb{R} \in \mathcal{L}_{\infty}(\mathcal{S})$  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.  $G_V(s) \neq \emptyset$ . I.e. the set of greedy actions at the state  $s$  is non-empty.
3. There exist a deterministic greedy policy  $\tau_V$  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 admissible state-action pairs  $\Gamma = \{(s, a) \in \mathcal{S} \times \mathcal{A} \mid a \in A(s)\}$  assumed to be closed in setting 2. Let  $\varphi : \mathcal{S} \rightarrow \mathcal{A}$  be the Borel-measurable function obtained from proposition 2.53 with  $\text{Gr}(\varphi) \subseteq \Gamma$ . Let  $s \in \mathcal{S}$  be any state. Then

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

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

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

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

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

Therefore  $\sup_{\tau \in S\Pi} V \leq TV$ . On the other hand  $\tau_V$  is in the supremum so  $\sup_{\tau \in S\Pi} V \geq TV$ .

4. By definition  $T_{\tau_V} V(s) = T_{\text{argmax}_{a \in A(s)} T_a V(s)} V(s) = \sup_{a \in A(s)} T_a V(s) = TV(s)$ . □

### 2.2.2 Existence of optimal policies

In the light of proposition 2.55 (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  $\tau_{T^k V_0} \in D\Pi$  which we will denote  $\tau_k^*$ .

**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  $\pi_1^{s_1, a_1} = \pi_2(\cdot \mid s_1, a_1, \cdot), \pi_2^{s_1, a_1} = \pi_3(\cdot \mid s_1, a_1, \cdot, \cdot), \dots$ .

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

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  $s_1 \in \mathcal{S}$ .

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

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

$$dP(s_2 \mid s_1, a_1) d\pi_1(a_1 \mid s_1) \quad (2.6)$$

$$\stackrel{\text{hyp.}}{\leq} \sup_{\pi \in R\Pi} \int r(s_1, a_1) + \gamma \int V_{k-1}^*(s_2) dP(s_2 \mid s_1, a_1) d\pi_1(a_1 \mid s_1) \quad (2.7)$$

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

$$\stackrel{2.55.4}{=} TV_{k-1}^*(s_1) \quad (2.9)$$

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  $V_k^* \leq TV_{k-1}^*$ . On the other hand by proposition 2.46 and induction hypothesis we have

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

But since  $(\tau_{k-1}^*, \dots, \tau_0^*)$  occur in the supremum we must then also have  $TV_{k-1}^* \leq V_k^*$ . Note that upper semicontinuity of  $V_k^*$  follows since  $T$  preserves this property (see proposition 2.55.4).  $\square$

**Proposition 2.57.** Under setting 2 it holds that  $V^* = \lim_{k \rightarrow \infty} T^k V_0^*$ . Furthermore under the greedy policy  $\tau_{V^*}$  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  $T^k V_0^* = V_k^* \rightarrow V^*$  (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

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

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  $\tau_{V^*}$  which satisfies

$$T_{\tau_{V^*}} V^* = TV^* \quad (2.10)$$

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.15)  $V^* = \lim_{k \rightarrow \infty} T^k V_0^*$  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  $\tau_{V^*}$  is optimal.  $\square$

**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  $\tau_{V^*}$  is optimal and for any  $V \in \mathcal{L}(\mathcal{S} \times \mathcal{A})$  it holds that

$$|T^k V - V^*| \leq \gamma^k |V - V^*|$$

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  $|T^k V - V^*|$  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.  $\square$

### Comparison to results of [2, Bertsekas and Shreve (2007)]

In [2] (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, [2] also considers a non-empty Borel space called the *disturbance space*  $W$ , a *disturbance kernel*  $p : \mathcal{S} \times \mathcal{A} \rightarrow W$ , instead of a transition kernel which on the other hand is a deterministic *system function*  $f : \mathcal{S} \times \mathcal{A} \times W \rightarrow \mathcal{S}$  which should be Borel measurable. Also the rewards are interpreted as negative costs, and thus  $g$  is required to be semi *lower*continuous. In [2] are also found much theory assuming *semianalytic* functions instead of semicontinuous ones.

It is possible to recover setting 2 from the semicontinuous setting in [2] 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  $T_\tau$  operator when  $\tau$  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  $K$ .

Initialize the expected reward function  $r \leftarrow \int x \, dR(x \mid \cdot)$ .

Initialize the first estimated value function  $\tilde{V}_0 = 0$ .

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

    Update the value function  $\tilde{V}_{k+1} \leftarrow T\tilde{V}_k$ .

**Output:** The  $K$ th optimal value function  $V_K^* = \tilde{V}_K$ .

---

To actually use algorithm 2 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 algorithm 2 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.

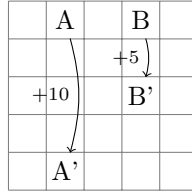


Figure 2.1: The simple gridworld Markov decision process.

**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  $\gamma$ .

Finite spaces are trivially standard Borel with the discrete topology, which also makes every map  $(s, a) \mapsto \mathcal{X}$  into some topological space continuous. In particular  $P$  is continuous and  $r$  is (upper semi)continuous. The set of admissible actions  $A(s)$  is equal to the full action space  $\mathcal{A}$  for all  $s \in \mathcal{S}$ , which is trivially compact. The rewards are bounded by  $R_{\max} = 10$  and therefore  $V_{\max} = 10/(1 - 0.9) = 100$ . Thus we can apply corollary 2.59 and for  $\tilde{V}_0 = 0$  get that

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



By proposition 2.46 for any stationary policy  $\tau \in \text{SII}$  we have that  $T_\tau$  is also  $\gamma$ -contractive and we easily get the same bound on the policy evaluation

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

To use algorithm 1 and algorithm 2 on the gridworld example we need to understand how the kernels  $R, P$ , value functions, policies and operators  $T_\tau, T$  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  $\tilde{V} \in \mathbb{R}^{\mathcal{S}}$  and policies as a matrix of point probabilities  $\tilde{\tau} \in \mathbb{R}^{\mathcal{A} \times \mathcal{S}}$  so that  $\tilde{\tau}(a, s) = \tau(\{a\} \mid s)$ . Since the rewards are deterministic the step  $r \leftarrow \int x \, dR(x \mid \cdot)$  is irrelevant.  $P$  is also deterministic so there exists a function  $\tilde{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  such that  $P(\{\tilde{P}(s, a)\} \mid s, a) = 1$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . Therefore the update step  $\tilde{V}_{k+1} \leftarrow T_\tau \tilde{V}_k$  in algorithm 1 becomes

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

Similarly in algorithm 2 the update step  $\tilde{V}_{k+1} \leftarrow T \tilde{V}_k$  in algorithm 2 becomes

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

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

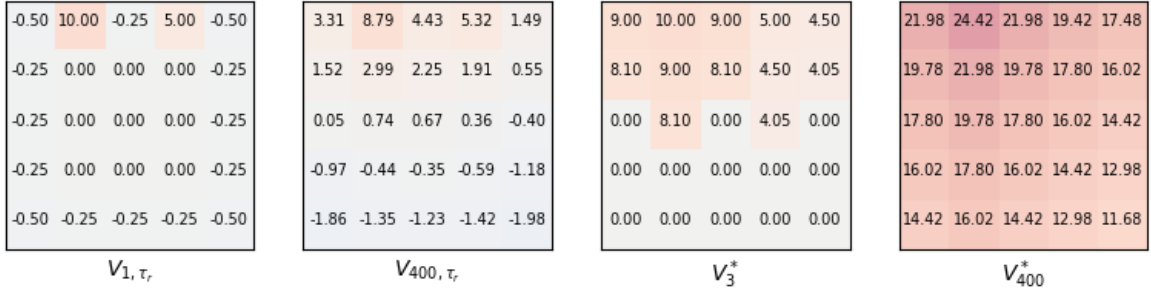


Figure 2.2: Value functions of the gridworld environment. Note that  $V_{\max} \cdot \gamma^{400} = 100 \cdot (0.9)^{400} \approx 4.97 \cdot 10^{-17}$  so  $V_{400}^*$  and  $V_{\tau_r, 400}$  are very close to the true infinite horizon value functions  $V^*$  and  $V_{\tau_r}$  (providing numerical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

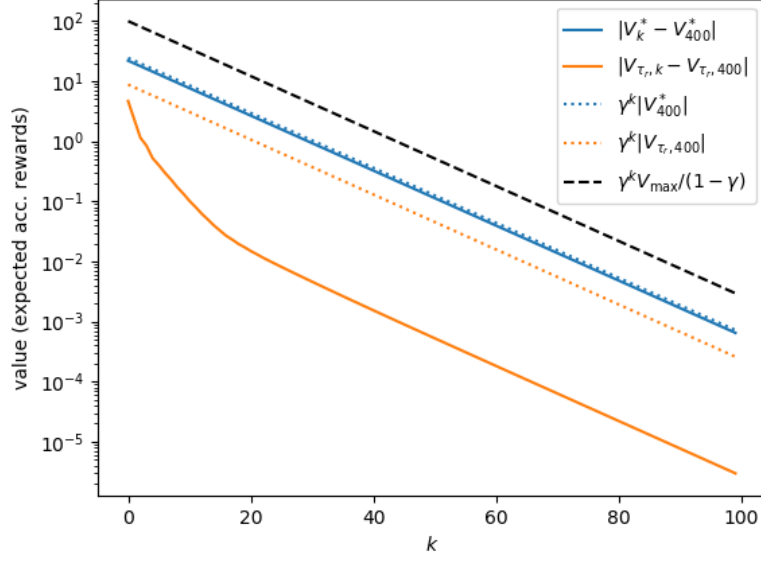


Figure 2.3: Convergence of gridworld value functions compared with the theoretical bounds. The black dashed line is the general theoretical bound for both  $T$  and  $T_\tau$  by Banachs fixed point theorem and the maximum value  $V_{\max} = R_{\max}/(1 - \gamma)$ . The dotted blue and orange uses  $|V_k^*|$  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  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ . 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 [20, Watkins (1989)]. Upon the definition he notes

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

A clear advantage of working with Q-function  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  rather than a value function  $V : \mathcal{S} \rightarrow \mathbb{R}$ , is that finding the optimal action  $a^* \in A(s)$  at state  $s$  requires only a maximization over the Q-function itself:  $a^* = \operatorname{argmax}_{a \in A(s)} Q(s, a)$ . This should be compared to finding an optimal action according to a value function  $V$ :  $a^* = \operatorname{argmax}_{a \in A(s)} r(s, a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V$ . Besides being less simple, this requires taking an expectation with respect to both the reward and transition kernel. Later we will study settings where we 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 admissible actions).  $A(s)$  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  $\sup_{a \in A(s)} Q(s, a)$  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

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

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

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  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$  we define

$$\begin{aligned} \text{Next-step operator: } P_\tau Q(s, a) &= \int Q(s', a') d\tau P(s', a' | s, a) \\ \text{Policy evaluation operator: } T_\tau Q(s, a) &= r(s, a) + \gamma \int Q(s', a') d\tau P(s', a' | s, a) \\ \text{Bellman optimality operator: } TQ(s, a) &= r(s, a) + \gamma \int \max_{a' \in \mathcal{A}} Q(s', a') dP(s' | s, a) \end{aligned}$$

where  $T_a = T_{\delta_a}$ .

**Remark 2.64.** The next-step operator  $P_\tau$  is defined for simplications in proofs, especially in the analysis of [6, Fan et al. (2020+)] in the later sections. Using  $P_\tau$  we can write  $T_\tau$  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  $G_Q(s) = \operatorname{argmax}_{a \in A(s)} Q(s, a)$ . If there exist a measurable set  $G_Q^\tau(s) \subseteq G_Q(s)$  for every  $s \in \mathcal{S}$  such that

$$\tau \left( G_Q^\tau(s) \mid s \right) = 1$$

then  $\tau$  is said to be **greedy** with respect to  $Q$  and is denoted  $\tau_Q$ .

**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.  $T_\tau$ -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)  $\tau$  is greedy for  $Q_{k,\pi}$  if and only if  $\tau$  is greedy for  $V_{k,\pi}$ .
  - (b)  $\tau$  is greedy for  $Q_\pi$  if and only if  $\tau$  is greedy for  $V_\pi$ .
4. Optimal policies are related by  $\max_{a \in A(s)} Q^*(s, a) = V^*(s)$  and

$$Q_k^*(s, a) = r(s, a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V_k^*, \quad Q^*(s, a) = r(s, a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V^*$$

**Proposition 2.67** (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.  $Q_\pi = \lim_{k \rightarrow \infty} Q_{k,\pi}$  and  $Q^* = \lim_{k \rightarrow \infty} Q_k^*$ .
3.  $T, T_\tau$  are  $\gamma$ -contractive on  $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$  and  $Q^*, Q_\tau$  are their unique fixed points.
4.  $Q^* = Q_{\tau^*}$  and  $Q_{k,\pi}, Q_\pi, Q_k^*, Q^*$  are all upper semicontinuous and bounded by  $V_{\max}$ .

*Proof of proposition 2.66 and proposition 2.67.* Measurability of  $Q_{k,\pi}$  and  $Q_\pi$  follow from measurability of  $V_{k,\pi}, V_\pi$  and proposition 2.4. Upper semicontinuity of  $Q_{k,\pi}$  and  $Q_\pi$  follows from proposition 2.51 because  $V_{k,\pi}$  and  $V_\pi$  are upper semicontinuous (see corollary 2.59).

For proposition 2.66.1 we have

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

For proposition 2.66.2 we sketch the idea by

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

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  $Q_\pi = \lim_{k \rightarrow \infty} Q_{k,\pi}$  follows from dominated convergence and proposition 2.22.3.

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

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

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

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

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

$$T_{\tau_Q}(s, a) = r(s, a) + \gamma \int \max_{a' \in A(s')} Q(s', a') \, dP(s' | s, a) = TQ(s, a)$$

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  $\tau_{k-1}^*$  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  $T_\pi$  follows from the same argument as for value functions. Banach fixed point theorem now concludes proposition 2.67.3.

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

## Q-iteration

Similar to the value iteration algorithm (algorithm 2) 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  $K$

Initialize the expected reward function:  $r \leftarrow \int x \, dR(x \mid \cdot)$ .

Initialize the starting Q-estimator:  $\tilde{Q}_0 \leftarrow r$ .

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

    Update the Q-estimator  $\tilde{Q}_{k+1} = T\tilde{Q}_k$

**Output:** The  $K$ -optimal Q-function  $Q_K^* = \tilde{Q}_K$ .

---

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

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

### 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  $\gamma$ . 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  $\gamma$  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 [1, Barto et al. (1983)]). One may add random effects such and *wind* to make such an example

The problem is of course that the value functions and operators which are used in algorithm 3 are not computable in practice. For example the state space of chess is very large (roughly  $|\mathcal{S}_{\text{chess}}| \geq 10^{43}$ ). This means that if we were to use algorithm 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 admissible 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. [6], 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  $\mathcal{Q} = \{f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}\}$ . Let  $\mathcal{F} \subseteq \mathcal{Q}$  be some subclass of Q-functions. Let  $\tilde{Q}_0 \in \mathcal{F}$  be bounded in  $\|\cdot\|$ . Suppose we can approximate  $T\tilde{Q}_0$  by some  $\tilde{Q}_1 \in \mathcal{F}$  to  $\varepsilon_1 > 0$  precision and then approximate  $T\tilde{Q}_1$  by  $\tilde{Q}_2 \in \mathcal{F}$  and so on. This way we get a sequence of Q-functions satisfying

$$\|T\tilde{Q}_{k-1} - \tilde{Q}_k\| \leq \varepsilon_k, \forall k \in \mathbb{N} \quad (2.11)$$

First observe that

$$\begin{aligned} \|T^k \tilde{Q}_0 - \tilde{Q}_k\| &\leq \|T^k \tilde{Q}_0 - T\tilde{Q}_{k-1}\| + \|T\tilde{Q}_{k-1} - \tilde{Q}_k\| \\ &\leq \gamma \|T^{k-1} \tilde{Q}_0 - \tilde{Q}_{k-1}\| + \|T\tilde{Q}_{k-1} - \tilde{Q}_k\| \end{aligned}$$

Using this iteratively we get

$$\|T^k \tilde{Q}_0 - \tilde{Q}_k\| \leq \sum_{i=1}^k \gamma^{k-i} \varepsilon_i \quad (2.12)$$

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

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

Thus we get

**Theorem 2.69.** Let  $\|\cdot\|$  be a norm on the space of functions  $\mathcal{S} \times \mathcal{A} \rightarrow \bar{\mathbb{R}}$ . Let  $\tilde{Q}_k$  be obtained from a function class  $\mathcal{F}$  such that  $\|\tilde{Q}_k - T\tilde{Q}_{k-1}\| \leq \varepsilon_k$  for any  $k \in \mathbb{N}$ . Then

$$\|Q^* - \tilde{Q}_k\| \leq \gamma^k \|Q^* - \tilde{Q}_0\| + \varepsilon_{\text{approx}}(k)$$

*Proof.* By the discussion above and

$$\begin{aligned} \|Q^* - \tilde{Q}_k\| &\leq \|Q^* - T^k \tilde{Q}_0\| + \|T^k \tilde{Q}_0 - \tilde{Q}_k\| \\ &\stackrel{2.12}{\leq} \gamma^k \|Q^* - \tilde{Q}_0\| + \varepsilon_{\text{approx}}(k) \end{aligned}$$

□

The first term in theorem 2.69 is sometimes called the **algorithmic** error<sup>1</sup>. 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

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

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

**Proposition 2.70.**  $\sum_{i=1}^k \gamma^{k-i} \varepsilon_i \rightarrow 0$  whenever  $\varepsilon_k \rightarrow 0$  as  $k \rightarrow \infty$ .

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

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

□

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

### 2.3.2 Using artificial 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  $f : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{L+1}}$  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  $\sigma_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i}$  is coordinate-wise application of components  $\sigma_{ij} : \mathbb{R} \rightarrow \mathbb{R}$ . 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  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ . 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  $i$ th 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

$$\begin{aligned} &\text{The } i\text{th partition is only connected to the neighboring } (i-1)\text{th and } (i+1)\text{th} \\ &\text{partition.} \end{aligned} \quad (2.15)$$

---

<sup>1</sup>For example in [6].

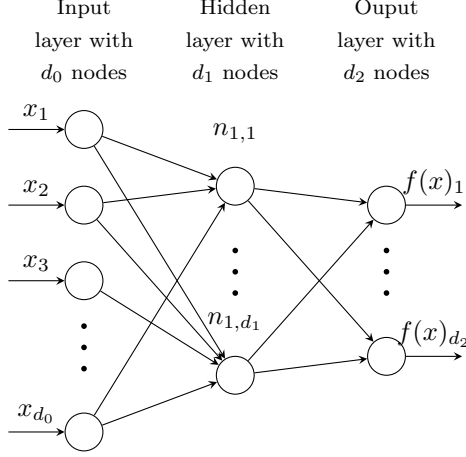


Figure 2.4: An ANN with one hidden layer ( $L = 1$ ). Notice that there is no edge from  $n_{0,3}$  to  $n_{1,1}$  which means that  $W_1(1,3) = 0$ .

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  $\text{im}(f) \subseteq \text{dom}(g)$ ) then  $g \circ f \in \mathcal{DN}\left((\sigma_1, \dots, \sigma_L, \sigma'_1, \dots, \sigma'_{L'}), (d_1, \dots, d_L, d'_1, \dots, d'_{L'})\right)$ . By side-by-side alignment we mean the situation where  $L = L'$  and one creates the function  $h : \mathbb{R}^{d_0+d'_0} \rightarrow \mathbb{R}^{d_L+d'_L}$  by defining  $h(x_1, \dots, x_{d_0+d'_0}) = (f(x_1, \dots, x_{d_0}), g(x_1, \dots, x_{d'_0}))$ . With this way of defining  $h$  we have that  $h$  is an ANN with structure  $(d_0 + d'_0, \dots, d_{L+1} + d'_{L+1})$ . 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  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  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  $\sigma$  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 [4, 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  $\sigma$  should satisfy

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

This was extended in [3, Chen et al. (1990)] to the statement as presented above and their proof is constructive.  $\square$

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*  $\sigma_r(x) = \max(0, x)$ . We define the class of ReLU networks as the ANNs 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 2.74.** 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  $p : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ .

or

2.  $P(\cdot \mid s, a)$  is absolutely continuous with respect to the same measure  $\nu$  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  $(\tilde{Q}_i)_{i=1}^k \subseteq \mathcal{RN}(w|\mathcal{A}|, N, 1)$  such that

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

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

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

*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  $f_{a_1}, \dots, f_{a_n} : \mathcal{S} \rightarrow \mathbb{R}$  into a bigger ReLU network  $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  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

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

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

$$|f_{a,k} - T\tilde{Q}_{k-1}(\cdot, a)| < \varepsilon \quad (2.16)$$

Here  $\tilde{Q}_0 = r$  and  $\tilde{Q}_k$  is obtained recursively by joining for each  $a$  the components  $f_{a,k}$  into a single network  $\tilde{Q}_k : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  such that  $\tilde{Q}_k(s, a) = f_{a,k}(s)$ . By eq. (2.16) on each of its components approximates  $T\tilde{Q}_{k-1}$  to  $\varepsilon$  precision.  $\tilde{Q}_0$  is continuous by setting 3 and by a. and b.  $T\tilde{Q}_k$  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  $x_{\ell} \rightarrow x \in \mathcal{S} \times \mathcal{A}$ . We will show that under 1. or 2. we have  $TQ(x_{\ell}) \rightarrow TQ(x)$ .

- 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

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

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  $\mathcal{A} = \{a_1, a_2, \dots, a_{\mathfrak{a}}\}$  and putting  $a_i = (0, \dots, 0, 1, 0, \dots, 0)$  where the 1 is on the  $i$ th 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  $\tilde{Q}_k : [0, 1]^{w+\mathfrak{a}} \rightarrow \mathbb{R}$  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  $i = 2, \dots, L-1$  putting  $d'_{a,i} = d_{a,i}$  for  $1 \leq i \leq L_a - 1$  and  $d'_{a,i} = d_{a,L_a}$  for  $L_a \leq i \leq L-1$  then  $d_L = \mathfrak{a}$  and finally  $d_{L+1} = 1$ . The first layer consist of the affine map  $w_1 : \mathbb{R}^{w+\mathfrak{a}} \rightarrow \mathbb{R}^{w \cdot \mathfrak{a}}$  defined by

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

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

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

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  $\tilde{Q}_k$ . For  $2 \leq i \leq L$  put  $w_i = (w_{1,i-1}, \dots, w_{\mathfrak{a},i-1}) : \mathbb{R}^{d_{i-1}} \rightarrow \mathbb{R}^{d_i}$  where we define

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

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

□

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 [3]. 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.11), 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  $p(\cdot \mid s, a) : \mathcal{S} \rightarrow \mathbb{R}$  with respect to  $\mu$  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 2.75** (Bernstein polynomial). The (multivariate) Bernstein polynomial  $B_{f,n}$  of degree  $n = (n_1, \dots, n_w) \in \mathbb{N}^w$  approximating the function  $f : [0, 1]^w \rightarrow \mathbb{R}$  is defined by

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

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

**Theorem 2.76** (Approximation with Bernstein polynomials). Let  $f : [0, 1]^w \rightarrow \mathbb{R}$  be Lipschitz w.r.t. the standard euclidean 2-norm induced metrics on  $[0, 1]^w$  and  $\mathbb{R}$  with constant  $L$ . Then for any  $n = (n_1, \dots, n_w) \in \mathbb{N}^w$  there exists a polynomial  $B_{f,n} : [0, 1]^w \rightarrow \mathbb{R}$  of degree  $\leq \|n\|_1$  such that

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

*Proof.* We refer to [8, Heitzinger (2002)] thm. B.7. □

**Lemma 2.77.** Under setting 4  $TQ(\cdot, a)$  is Lipschitz with constant  $L_T = L_r + \gamma V_{\max} L_p$  for all  $a \in \mathcal{A}$  and measurable  $Q : \mathcal{S} \times \mathcal{A} \rightarrow [-V_{\max}, V_{\max}]$ .

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

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

□

Now we can bound

**Proposition 2.78.** Given an MDP satisfying setting 4 and using the 2-norm as metric we can bound

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

*Proof.* Following the procedure from leading to eq. (2.12), starting with  $\tilde{Q}_0 = 0$  and using the  $n = (n_1, \dots, n_w)$  degree Bernstein polynomial  $\tilde{Q}_k = B_{T\tilde{Q}_{k-1}, n}$  as approximation for  $T\tilde{Q}_{k-1}$  we know by induction and the results lemma 2.77 and theorem 2.76.2 that  $T\tilde{Q}_k$  is  $L_T$ -Lipschitz for any  $k \in \mathbb{N}$ . Now by choosing the euclidean norm  $\|\cdot\| = \|\cdot\|_2$  we have by theorem 2.76.1 that

$$\varepsilon_i = \|\tilde{Q}_k - T\tilde{Q}_{k-1}\|_2 \leq \frac{L_T}{2} \sqrt{\sum_{j=1}^w \frac{1}{n_j}} = \varepsilon \quad (2.17)$$

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

$$\varepsilon_{\text{approx}} \leq \frac{\varepsilon}{1 - \gamma} \quad (2.18)$$

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 2.78 we give a corollary where we put  $n_j = m$  for all  $j \in [w]$ :

**Corollary 2.79.**

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

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

*Proof.* Use proposition 2.78. □

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 2.79 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 2.79 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 unknown. In this case we cannot calculate directly neither  $r$ ,  $T_\pi Q$  nor  $TQ$  because the transition and reward kernels  $P, R$  are unknown.

It is clear that algorithm 3 will not work without modification in this case. Simply because  $R$  and  $P$  are not available. To make the scheme work anyway we could simply avoid taking expectations and use the random outcomes of the kernels. Leading to

---

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

---

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

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

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

$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \tilde{Q}_{k+1}(s, a) \leftarrow r' + \gamma \sup_{a' \in \mathcal{A}} \tilde{Q}_k(s', a')$   
 where  $r' \sim R(\cdot \mid s, a), s' \sim P(\cdot \mid s, a)$ .

Define  $\pi_K$  as the greedy policy w.r.t.  $\tilde{Q}_K$

**Output:** An estimator  $\tilde{Q}_K$  of  $Q^*$  and policy  $\pi_K$

---

We immediately 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  $\tilde{Q}_k$  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<sup>1</sup>. In the *indirect* approaches one tries to first estimate  $P$  and  $R$  by sampling. We can then use the model-dependent methods of the last chapter using estimated kernels  $\tilde{P}$  and  $\tilde{R}$ . We will not go further into indirect methods in this thesis. The *direct* approaches covers *the rest*, and it is mainly these we are going to look at throughout this paper. A popular indirect approach is called *temporal difference* (TD) learning and was initially invented for finite MDPs.

### 3.1 Finite case

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

---

<sup>1</sup>This classification is discussed in [11, Kearns and Singh (1999)].

TD learning is based on the following update scheme

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

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

We will now look at a convergence result originally obtained in [21, Watkins and Dayan (1992)] of a TD algorithm using Q-functions. The result was extended slightly in [9, Jaakkola et al. (1994)] and is here presented more in the style of [9].

---

**Algorithm 5:** Finite asynchronos Q-learning

---

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

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

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

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

    Update action-value function:

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

Define  $\pi_K$  as the greedy policy w.r.t.  $\tilde{Q}_K$

**Output:** An estimator  $\tilde{Q}_K$  of  $Q^*$  and policy  $\pi_K$

---

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

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

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

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

In a special case of the same setup, convergence rates where established by [19, Szepesvári (1997)].

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

$$|\tilde{Q}_t - Q^*| \leq B \frac{1}{t^{\beta(1-\gamma)}} \quad (3.2)$$

and

$$|\tilde{Q}_t - Q^*| \leq B \sqrt{\frac{\log \log t}{t}} \quad (3.3)$$

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

A paper [13, Majeed and Hutter (2018)] proves that  $Q$ -learning is PAC-learnable given some additional assumptions.

---

**Algorithm 6:** Finite synchronos  $Q$ -learning

---

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

**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:

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

Define  $\pi_K$  as the greedy policy w.r.t.  $\tilde{Q}_K$

**Output:** An estimator  $\tilde{Q}_K$  of  $Q^*$  and policy  $\pi_K$

---

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

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

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

If the synchronos algorithm (algorithm 6) is run with

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

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

If the asynchronos algorithm (algorithm 5) with

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

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

In [13]  $L$  is called the *covering rate*.

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

learning theoretically optimal values for  $\omega$ . For example if we want to optimize for the bound on  $K$  for  $\gamma \rightarrow 1$  using the synchronos algorithm, we get the following rate (treating other variables as constant)  $K \geq C'(1/(1-\gamma)^{4/\omega} + 1/(1-\gamma)^{1/(1-\omega)})$  for some  $C' > 0$ . Thus picking  $\omega = 4/5$  is optimal. As another example: running the asynchronos algorithm and wanting to minimize for large covering rates  $L$ . We get  $K \geq C''(L^{2+1/\omega} + L^{1/(1-\omega)})$  for some  $C'' > 0$ . This is optimized for  $\omega \approx 0.77$ . Then in [13] experiments points to  $\omega = 0.85$  as being optimal using a scheme generating random finite MDPs. Other authors have since used this number as a standard value for the learning rate (see e.g. [5, Devraj and Meyn (2017)])

### 3.1.1 History dependent setting

**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.
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  $r : \mathcal{H}^* \rightarrow \mathbb{R}$ . This is discounted by  $\gamma \in [0, 1)$  in accordance to condition (D). That is  $r(h_n)$  is bounded in the interval  $[-\gamma^{n-2}R_{\max}, \gamma^{n-2}R_{\max}]$  for any  $h_n \in \mathcal{H}_n$ . Furthermore  $r$  depends only on  $s_1 a_1 \dots s_k r_{k-1} a_k$  when evaluated on  $h_{k+1} = s_1 a_1 \dots r_{k-1} a_k s_{k+1} \in \mathcal{H}_{k+1}$ .

**Remark 3.5.** Note that the finite setting 6 is a special case of setting 1 considered by [Schäl, 1974], because Polishness and compactness of  $\mathcal{S}, \mathcal{A}$  is readily implied by using the discrete topology in the finite state and action spaces, and the fact that (D) implies pt. 5 in setting 1. Further the conditions (S) and (W) of Schal are also both implied by the discreteness. This implies by theorem 2.34 the existence of an optimal  $\pi^* \in R\Pi$  and that  $V_n^* \rightarrow V^*$ .

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

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

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

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

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

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

$$p_h : \{\phi(h)\} \times \mathcal{A} \rightsquigarrow \mathcal{X}$$



$$p_h(x' | xa) = \sum_{s: \phi(hr'as)=x'} P(s | ha), \quad r' = r(h, a)$$

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

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

**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 as in setting 6 together with the state-uniformity condition is by [13] called a *Q-Value uniform decision process* (QDP).

**Theorem 3.6** (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:

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

**Theorem 3.7.** Within setting 6 assume

1. State-uniformity (assumption 4).
2. Any state is reached eventually under any policy (called *state-process ergodicity* in [13]).
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  $q_0 : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$  the update step eq. (3.6) yields a sequence  $(q_t)_{t \in \mathbb{N}}$  which converges to the optimal  $q^* = Q^*$ .

It seems relevant to ask how restrictive the state-uniformity assumption is. [13] answers this by an array of examples showing the following relations of the classes of decision processes:

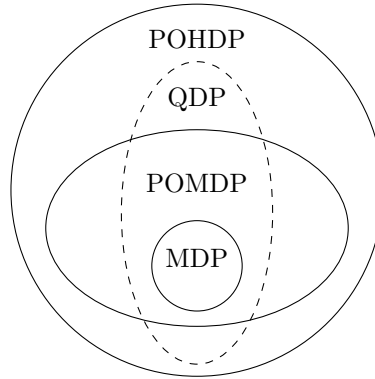


Figure 3.1: Classes of finite decision processes considered in [13], under setting 6.

Recall that QDP is a partially observable HDP under state-uniformity (assumption 4).

One point remains unclear after reading [13]: Why is  $q^*$  and  $Q^*$  well defined as the solution to their functional equations (eq. (3.4) and eq. (3.5)) and how are they related to the optimal value function  $V^*(s) = \sup_{\pi \in R\Pi} \mathbb{E}_s^\pi \sum_{i=1}^\infty \gamma^{i-1} r_i$  (see definition 2.24) of a general HDP? A sensible thing to ask would be that  $Q^*(h, a) = r(h, a) + \gamma \mathbb{E}_{P(\cdot|ha)} V^*$ . However we will not go further into these details.

### 3.2 Linear function approximation

This section is based on [14, Melo and Ribeiro (2007)].

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

1. An MDP  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$  (see definition 2.36).
2. Discounted, i.e. (D) holds with  $\gamma \in [0, 1)$ .
3.  $\mathcal{S} \subseteq \mathbb{R}^w$  is a compact subset of a euclidean space.
4.  $\mathcal{A}$  is finite and not varying for each  $s \in \mathcal{S}$ , that is the set of admissible actions satisfy  $A(s) = \mathcal{A}$  for all  $s \in \mathcal{S}$ .
5.  $r_i$  is upper semicontinuous .

**Remark 3.8.** Item 5 was actually not part of the assumptions in [14]. We include it here in order to ensure the existence of an optimal policy and thus measurability of  $V^*$ .

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

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

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

The gradient of  $Q_\theta$  over  $\theta$  is

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

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

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

---

**Algorithm 7:** Q-learning with linear approximation

---

**Input:** MDP  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ , policy  $\pi$ , 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  $\tilde{\pi}_K$  as the greedy policy w.r.t.  $\tilde{Q}_K := Q_{\theta_{K+1}}$ .

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

---

In order to understand the results of the analysis of algorithm 7 found in [14], 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  $\rho_i : \mathcal{X}^\infty \rightarrow \mathcal{X}$  be projection on the  $i$ th space. Define for any  $A \in \Sigma_{\mathcal{X}}$  the function  $\tau_A : \mathcal{X}^\infty \rightarrow \bar{\mathbb{N}} = \inf\{i \in \mathbb{N} \mid \rho_i \in A\}$ . Intuitively this function records the earliest time where the process enter the set  $A \subseteq \mathcal{X}$ . Define the function  $\eta_A : \mathcal{X}^\infty \rightarrow \bar{\mathbb{N}} = \sum_{i \in \mathbb{N}} 1_A \circ \rho_i$ . This function records the total number of times in which the process is inside the set  $A$ . Let  $\varphi \in \mathcal{P}(\mathcal{X})$  be a probability measure on  $\mathcal{X}$ .

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

**Definition 3.10** (Positivity).

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

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

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

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

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

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

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

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

we have that

$$\theta_k \rightarrow \theta^*$$

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

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

Furthermore the orthogonal projection is expressible as

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

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

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

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

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

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

## 3.3 Deep Q-learning

### 3.3.1 Introduction

This section is mostly based on the paper [6, Fan et al. (2020+)].

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

Though [6] claims to investigate the deep Q-network algorithm (see algorithm 9), instead of analysing DQN directly, a related algorithm called *fitted Q-iteration* (FQI) is analysed instead and bounds on its convergence is established.

The algorithm analysed by [Fan et al] is

---

**Algorithm 8:** Fitted Q-Iteration Algorithm

---

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

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

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

$S'_i \sim P(S_i, A_i)$

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

    Update action-value function:

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

Define  $\pi_K$  as the greedy policy w.r.t.  $\tilde{Q}_K$

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

---

We will now look at the function class used in [6].

### ReLU Networks

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

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

$$\bullet \max_{\ell \in [L+1]} \|\tilde{W}_\ell\|_\infty \leq 1 \quad \bullet \sum_{\ell=1}^{L+1} \|\tilde{W}_\ell\|_0 \leq s \quad \bullet \max_{j \in [d_{L+1}]} \|f_j\|_\infty \leq V$$

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

**Remark 3.18.** Following the graph interpretation of ANNs (see remark 2.72) the condition that  $\sum_{i=1}^{L+1} \|\tilde{W}_i\|_0 \leq s$  corresponds to graph-theoretical sparsity of the graph derived from the ANN.

The idea to work with this particular subclass of neural networks come from [17] (p. 22), which establishes the following lemma

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

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

In the course of establishing the results in [6] we will not go more into this result or other properties of ReLU networks in particular, instead putting emphasis on how to use this result to obtain the main theorem, which we will present shortly.

### 3.3.2 Assumptions

#### Hölder Smoothness

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

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

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

$$C_w(\mathcal{D}, \beta, H) := \left\{ f : \mathcal{D} \rightarrow \mathbb{R} \mid \|f\|_{C_w} \leq H \right\}$$

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

$$f = g_q \circ \dots \circ g_1$$

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

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

**Definition 3.22.** Define

$$\mathcal{F}_0 = \{f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \mid f(\cdot, a) \in \mathcal{F}(s, V) \forall a \in \mathcal{A}\}$$

and

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

We here recall the definition of the operators for Q-functions (definition 2.63). For any stationary policy  $\tau \in \text{SII}$  we define

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

matching the definitions in [6].

**Assumption 5.**  $T\mathcal{F}_0 \subseteq \mathcal{G}_0$ . I.e. it 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 Hölder 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: Setting  $\mathcal{D} = [0, 1]^r$ ,  $q = 1$  and taking both the expected reward function and transition kernel to be Hölder smooth.

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

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

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

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

similarly

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

Thus since  $p$  and  $r$  are Hölder smooth

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

### Concentration coefficients

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

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

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

**Assumption 6.** Let  $\nu$  be the sampling distribution from the algorithm, and  $\mu$  the distribution over which we measure the error in the main theorem, then we assume

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

### 3.3.3 The main theorem

**Theorem 3.25** (Fan, Yang, Xie, Wang). Under ?? let  $\mu$  be any distribution over  $\mathcal{S} \times \mathcal{A}$ . Make assumption 5 and assumption 6 with the constants  $\phi_{\mu, \nu} > 0$ ,  $q \in \mathbb{N}$  and  $\{p_j, t_j, \beta_j, H_j\}_{j \in [q]}$ . Furthermore assume that there exists a constant  $\xi > 0$  such that

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

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

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

with structure satisfying

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

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

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

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

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

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

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

### 3.3.4 Relation to DQN

The following is famous *DQN*-algorithm proposed by [15, Mnih et al. (2015)].



---

**Algorithm 9:** Deep Q-Network

---

**Input:** MDP  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ , number of iterations  $K$ , batch size  $n$ , exploration factor  $\epsilon$ , 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  $\{\alpha_t\}_{t \geq 0}$

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

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

Initialize target network  $Q_{\text{target},0} = \tilde{Q}_0$  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  $\tilde{Q}_k$ , that is  $A_k$  is picked from  $\text{argmax}_{a \in \mathcal{A}} \tilde{Q}_k(S_k, a)$ .

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

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

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

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

    Update the Q-network by performing a gradient descent step

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

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

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

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

---

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

[6] 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  $T_{\text{target}}$  steps.

In practice the size of the replay memory buffer is very large, for example in [15] it holds  $\sim 10^6$  transitions. Because of this it is argued in [6] 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  $\tilde{Q}_{k-1}$  to update  $\tilde{Q}_k$ , the role of  $\tilde{Q}_{k-1}$  becomes similar to the target network  $Q_{\text{target},\ell(k-1)}$  of DQN.

### Critique of this relation

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

### 3.3.5 Differences in notation

Because  $\sigma$  is used ambiguously in theorem 3.25 we denote the probability distribution  $\sigma$  from [6] p. 20 by  $\nu$  instead. I avoid the shorthand defined in [6] 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 [6] to denote the euclidean dimension of the state space, while here we use  $w$ .

### 3.3.6 Proofs

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

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

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

Where

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

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

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

Then

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

Where

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

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

*Proof of main theorem.* Using theorem 3.26 we get

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

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

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

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

**Step 1.** We want to employ the lemma by [17] (lemma 3.19) to each Hölder smooth part of  $g = Tf$  where  $f$  is an arbitrary network in  $\mathcal{F}_0$  and then piece it together somehow, using that ReLU networks are easily stitched together into bigger ReLU networks. Therefore the first step is to refit our Hölder Smooth compositions in  $\mathcal{G}_0$  to be defined on a hyper-cube instead. This is a relatively simple procedure:

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

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

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

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

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

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

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

### Step 1.1

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

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

By lemma 3.19 there exists a ReLU network

$$\hat{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \tilde{d}_j p_{j+1}, \dots, \tilde{d}_j p_{j+1}, p_{j+1}\right\}, (\tilde{s}_j + 4) \cdot p_{j+1}\right) \quad (3.14)$$

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

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

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

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

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

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

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

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

$$\left\| \tilde{h}_{jk} - h_{jk} \right\|_{\infty} = \left\| \tau \circ \hat{h}_{jk} - \tau \circ h_{jk} \right\|_{\infty} \quad (3.17)$$

$$\leq \left\| \hat{h}_{jk} - h_{jk} \right\|_{\infty} \quad (3.18)$$

$$\leq 2W3^{-\beta_j} N^{-\beta_j/t_j} \quad (3.19)$$

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

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

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

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

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

and for  $\tilde{s}$

$$\begin{aligned} \tilde{s} &= \sum_{j=1}^q (\tilde{s}_j + 4)p_{j+1} \\ &\leq \sum_{j=1}^q (141N(m + 6)(t_j + \beta_j + 1)^{3+t_j} + 4)p_{j+1} \end{aligned}$$

$$\begin{aligned}
&\leq 142N(\log n)^\xi(2\eta + 6)\log_2(n)\sum_{j=1}^q(t_j + \beta_j + 1)^{3+t_j} \\
&\leq 142N(\log n)^\xi(2\eta + 6)\log_2(e)\log(n)(\log n)^\xi \\
&= 142N(2\eta + 6)\log_2(e)(\log n)^{1+2\xi} \\
&= C_{\tilde{s}}n^{a^*}(\log n)^{\xi^*}
\end{aligned}$$

where  $C_{\tilde{s}} = 142(2\eta + 6)\log_2(e)$ . Now we bound  $\left\|\tilde{f} - f(\cdot, a)\right\|_\infty$ . Define  $G_j = h_j \circ \dots \circ h_1$ ,  $\tilde{G}_j = \tilde{h}_j \circ \dots \circ \tilde{h}_1$  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

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

so by induction and eq. (3.15)

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

and therefore

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

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  $\delta$ -covering of  $\mathcal{F}(\tilde{L}, \{\tilde{d}_j\}_{j=1}^{\tilde{L}+1}, \tilde{s})$  by

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

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

$$|\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)| \leq |\mathcal{N}_\delta|^{|\mathcal{A}|}$$

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

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

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

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

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

*Proof.* We refer to theorem 14.5 in [todo: ref to Anthony Bartlett, thm. 14.5].  $\square$

With lemma 3.29 and  $n$  sufficiently large we can bound

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

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

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

where

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

only depends on the constants in assumption 6 finishing the proof.  $\square$

Now for theorem 3.26.

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

*Proof.*

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

$\square$

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

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

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

*Proof.* Recall that

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

Thus

$$\mathbb{E}_\mu[(P^{\tau_m} \dots P^{\tau_1})(f)] = \int (P^{\tau_m} \dots P^{\tau_1})(f) d\mu \quad (3.22)$$

$$= \int f d(P^{\tau_m} \dots P^{\tau_1} \mu) \quad (3.23)$$

$$= \int f \frac{d(P^{\tau_m} \dots P^{\tau_1} \mu)}{d\nu} d\nu \quad (3.24)$$

$$\leq \left\| \frac{d(P^{\tau_m} \dots P^{\tau_1} \mu)}{d\nu} \right\|_{2, \nu} \cdot \|f\|_{2, \nu} \quad (3.25)$$

$$\leq \kappa(m, \mu, \nu) \|f\|_{2, \nu} \quad (3.26)$$

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

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

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

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

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

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

The proof consists of four steps.

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

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

This implies

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

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

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

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

$$Q^* - Q^{\pi_K} \leq \gamma U^{-1}(P^{\pi^*}(Q^* - \tilde{Q}_K) - P^{\pi_K}(Q^* - \tilde{Q}_K)) \quad (3.28)$$

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

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

and a lower bound

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

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

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

and

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

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

$$\begin{aligned} Q^* - Q^{\pi_K} &\leq U^{-1} \left( \gamma^{K+1} ((P^{\pi^*})^{K+1} - P^{\pi_K} \dots P^{\pi_0}) (Q^* - \tilde{Q}_0) \right. \\ &\quad \left. + \sum_{i=0}^{K-1} \gamma^{K-i} ((P^*)^{K-i} - P^{\pi_K} \dots P^{\pi_{i+1}}) \varrho_{i+1} \right) \end{aligned} \quad (3.33)$$

For shorthand define constants

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

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

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

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

Then by eq. (3.33)

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



So by linearity of expectation

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} = \mathbb{E}_\mu |Q^* - Q^{\pi_K}| \quad (3.38)$$

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

With the bound on rewards we (crudely) estimate

$$\mathbb{E}_\mu O_K | Q^* - \tilde{Q}_0 \leq 2V_{\max} = 2R_{\max}/(1-\gamma) \quad (3.40)$$

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

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

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

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

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

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

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

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

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

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

Focusing on the first term on RHS of eq. (3.46), if we then we can take the norm out of the sum as a constant. We are left with

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

Where the last inequality is due to assumption 6. Combining eq. (3.46) and eq. (3.47) we arrive at

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

□

Finally we show theorem 3.27.

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

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

*Proof.* See [Vershynin 2010, p. 12].

□

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

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

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

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

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$$

*Proof.* See [Vershynin 2010, p. 14]

□

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

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

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

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

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

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

□

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

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

where  $\sigma^2 = \sum_{i=1}^n \mathbb{E}U_i^2$ .

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

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

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

and  $X_i := (S_i, A_i)$ . Notice that  $\tilde{Q}$  differs from  $\hat{Q}$  in that  $\tilde{Q}$  approximates  $TQ$  w.r.t.  $\|\cdot\|_\nu^2$  while  $\hat{Q}$  approximates  $Y = (Y_1, \dots, Y_n)$  in mean squared error over  $X = (X_1, \dots, X_n)$ . We shall be loose about applying functions to vectors (of random variables) in the sense that they are applied entry-wise. We use  $\|\cdot\|_p$  to denote the (finite dimensional)  $p$ -norm ( $p$  omitted 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 definition (of  $\hat{Q}$ ) for all  $f \in \mathcal{F}$  we have  $\|\hat{Q}(X) - Y\|^2 \leq \|f(X) - Y\|^2$ , leading to

$$\|Y\|^2 + \|\hat{Q}\|^2 - 2Y \cdot \hat{Q} \leq \|Y\|^2 + \|f\|^2 - 2Y \cdot f \quad (3.49)$$

$$\iff \|\hat{Q}\|^2 + \|TQ\|^2 - 2\hat{Q} \cdot TQ \leq \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \hat{Q} - 2Y \cdot f - 2\hat{Q} \cdot TQ + 2f \cdot TQ \quad (3.50)$$

$$\iff \|\hat{Q} - TQ\|^2 \leq \|f - TQ\|^2 + 2(Y - TQ) \cdot (\hat{Q} - f) \quad (3.51)$$

$$\iff \|\hat{Q} - TQ\|^2 \leq \|f - TQ\|^2 + 2\xi \cdot (\hat{Q} - f) \quad (3.52)$$

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

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

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

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

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

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

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

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

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

By this lemma we can deduce

$$\mathbb{E}(\xi \cdot (\hat{Q} - f)) = \mathbb{E}(\xi \cdot (\hat{Q} - TQ)) \quad (3.53)$$

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

$$\left| \mathbb{E}(\xi \cdot (\hat{Q} - TQ)) \right| \leq \left| \mathbb{E}(\xi \cdot (\hat{Q} - f_{k^*})) \right| + \left| \mathbb{E}(\xi \cdot (f_{k^*} - TQ)) \right| \quad (3.54)$$

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

$$\left| \mathbb{E} \xi \cdot (\hat{Q} - f_{k*}) \right| \leq \mathbb{E} \left( \|\xi\| \|\hat{Q} - f_{k*}\| \right) \leq \mathbb{E}(\|\xi\|) \sqrt{n} \delta \leq 2n V_{\max} \delta \quad (3.55)$$

where we have used that  $\|\hat{Q} - f_{k*}\|_{\infty} \leq \delta$  so

$$\|\hat{Q} - f_{k*}\|^2 = \sum_{i=1}^n (\hat{Q}(X_i) - f_{k*}(X_i))^2 \leq \sum_{i=1}^n \delta^2 = n \delta^2 \quad (3.56)$$

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

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

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

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

Note that since  $\xi_i$  are centered  $Z_j$ . For a sub- $\sigma$ -algebra  $\Sigma$  define the *sub-gaussian* norm by

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

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

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

$$\|\xi_i(f_j(X_i) - TQ(X_i))\|_{\psi_2, \Sigma} \leq 2V_{\max} |f_j(X_i) - TQ(X_i)| \quad (3.59)$$

Therefore by lemma 3.32

$$\|Z_j\|_{\psi_2, \Sigma}^2 \leq \|f_j - TQ\|^{-2} \left\| \sum_{i=1}^n \xi_i(f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma}^2 \quad (3.60)$$

$$\leq \|f_j - TQ\|^{-2} C_1 \sum_{i=1}^n \|\xi_i(f_j(X_i) - TQ(X_i))\|_{\psi_2, \Sigma}^2 \quad (3.61)$$

$$\leq \|f_j - TQ\|^{-2} C_1 \sum_{i=1}^n 4V_{\max}^2 |f_j(X_i) - TQ(X_i)|^2 \quad (3.62)$$

$$= 4V_{\max}^2 C_1 \quad (3.63)$$

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

$$\mathbb{E} \exp \left( c Z_j^2 / \|Z_j\|_{\psi_2}^2 \right) \leq e \quad (3.64)$$

so

$$\mathbb{E} \max_{j \in N_{\delta}} Z_j^2 = \frac{\max_{j \in [N_{\delta}]} \|Z_j\|_{\psi_2}^2}{c} \mathbb{E} \left( \max_{j \in [N_{\delta}]} \frac{c Z_j^2}{\max_{k \in [N_{\delta}]} \|Z_k\|_{\psi_2}} \right) \quad (3.65)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \mathbb{E} \left( \max_{j \in N_{\delta}} \frac{c Z_j^2}{\|Z_j\|_{\psi_2}} \right) \quad (3.66)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \log \left( \mathbb{E} \max_{j \in N_{\delta}} \exp \left( \frac{c Z_j^2}{\|Z_j\|_{\psi_2}} \right) \right) \quad (3.67)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \log \left( \sum_{j \in [N_\delta]} \mathbb{E} \exp \left( \frac{cZ_j^2}{\|Z_j\|_{\psi_2}} \right) \right) \quad (3.68)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \log(eN_\delta) \quad (3.69)$$

$$\leq C_2^2 V_{\max}^2 \log(N_\delta) \quad (3.70)$$

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

$$\mathbb{E}(\xi \cdot (f_{k*} - TQ)) = \mathbb{E}(\|f_{k*} - TQ\| |Z_{k*}|) \quad (3.71)$$

$$\leq \mathbb{E} \left( \left( \|\hat{Q} - TQ\| + \|\hat{Q} - f_{k*}\| \right) |Z_{k*}| \right) \quad (3.72)$$

$$\leq \mathbb{E} \left( \left( \|\hat{Q} - TQ\| + n\delta \right) |Z_{k*}| \right) \quad (3.73)$$

$$\leq \left( \mathbb{E} \left( \|\hat{Q} - TQ\| + n\delta \right)^2 \right)^{1/2} \left( \mathbb{E} Z_{k*}^2 \right)^{1/2} \quad (3.74)$$

$$\leq \mathbb{E} \left( \|\hat{Q} - TQ\| + n\delta \right) \left( \mathbb{E} Z_{k*}^2 \right)^{1/2} \quad (3.75)$$

$$\leq \left( \sqrt{\mathbb{E} \|\hat{Q} - TQ\|_2^2} + n\delta \right) \left( \mathbb{E} Z_{k*}^2 \right)^{1/2} \quad (3.76)$$

$$\leq \left( \sqrt{\mathbb{E} \|\hat{Q} - TQ\|_2^2} + n\delta \right) C_2^2 V_{\max}^2 \log(N_\delta) \quad (3.77)$$

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

$$\mathbb{E} \|\hat{Q} - TQ\|^2 \leq \mathbb{E} \|f - TQ\|^2 + 4nV_{\max}\delta + \left( \sqrt{\mathbb{E} \|\hat{Q} - TQ\|^2} + \sqrt{n}\delta \right) C_2 V_{\max} \sqrt{\log(N_\delta)} \quad (3.78)$$

$$= C_2 V_{\max} \sqrt{n \log(N_\delta)} \sqrt{\mathbb{E} \|\hat{Q} - TQ\|^2} + nC_2^2 \delta V_{\max}^2 \log(N_\delta) + \mathbb{E} \|f - TQ\|^2 \quad (3.79)$$

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

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

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

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

□

By lemma 3.39 applied to eq. (3.79)

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

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

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

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

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

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

**Step 2** Here we link up  $\|\hat{Q} - TQ\|_\sigma^2$  with  $\mathbb{E}_n^1 \|\hat{Q} - TQ\|^2$ . First note that

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

$$\leq 4V_{\max}\delta \quad (3.83)$$

Using this twice we can say

$$(\hat{Q}(\hat{X}_i) - TQ(\hat{X}_i))^2 \quad (3.84)$$

$$\leq (\hat{Q}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 - (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 + (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 \quad (3.85)$$

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

$$\leq (\hat{Q}(X_i) - TQ(X_i))^2 + (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 + 8V_{\max}\delta \quad (3.87)$$

Thus we get

$$\|\hat{Q} - TQ\|_\sigma^2 \quad (3.88)$$

$$= \mathbb{E}\frac{1}{n}\sum_{i=1}^n (\hat{Q}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 \quad (3.89)$$

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

$$= \frac{1}{n}\|\hat{Q} - TQ\|^2 + \frac{1}{n}\sum_{i=1}^n h_{k*}(X_i, \tilde{X}_i) + 8V_{\max}\delta \quad (3.91)$$

Where we define

$$h_j(x, y) := (f_j(y) - TQ(y))^2 - (f_j(x) - TQ(x))^2 \quad (3.92)$$

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

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

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

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

$$\leq \Upsilon/n \mathbb{E} T \quad (3.95)$$

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

$$\text{Var} h_j(X_i, \tilde{X}_i) = 2 \text{Var} (f_j(X_i) - TQ(X_i))^2 \quad (3.96)$$

$$\leq 2 \mathbb{E} (f_j(X_i) - TQ(X_i))^4 \quad (3.97)$$

$$\leq 2 \Upsilon^4 \quad (3.98)$$

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

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

$$\leq u + \int_u^\infty \mathbb{P}(T \geq t) dt \quad (3.100)$$

$$\leq u + \int_u^\infty 2N_\delta \exp \left( \frac{-t^2}{2\Upsilon t/3 + 4n\Upsilon^2} \right) dt \quad (3.101)$$

$$\leq u + 2N_\delta \int_u^\infty \exp \left( \frac{-t^2}{2\Upsilon^2(t/(3\Upsilon) + 2n)} \right) dt \quad (3.102)$$

$$\leq u + 2N_\delta \left( \int_u^{6n\Upsilon} \exp \left( \frac{-t^2}{8n\Upsilon^2} \right) dt + \int_{6n\Upsilon}^\infty \exp \left( \frac{-t}{4/3\Upsilon} \right) dt \right) \quad (3.103)$$

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

where we use lemma A.16 from eq. (3.103) to eq. (3.104). Now set  $u = \Upsilon \sqrt{8n \log N_\delta}$ . We continuing from eq. (3.104) getting

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

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

$$\leq 4\sqrt{2}\Upsilon \sqrt{n \log N_\delta} + 8/3 \Upsilon \quad (3.107)$$

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

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

$$\begin{aligned} &\leq \frac{(1+\kappa)^2}{\kappa} \frac{1}{n} C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa) \left( \delta C_2^2 V_{\max}^2 \log(N_\delta) + \omega(\mathcal{F}) \right) \\ &\quad + 8\sqrt{2} V_{\max} n^{-1/2} \sqrt{\log N_\delta} + 8V_{\max}(n^{-1} + \delta) \end{aligned} \quad (3.109)$$

finishing the proof.  $\square$

# Chapter 4

## Comparison and conclusion

### 4.1 Comparison

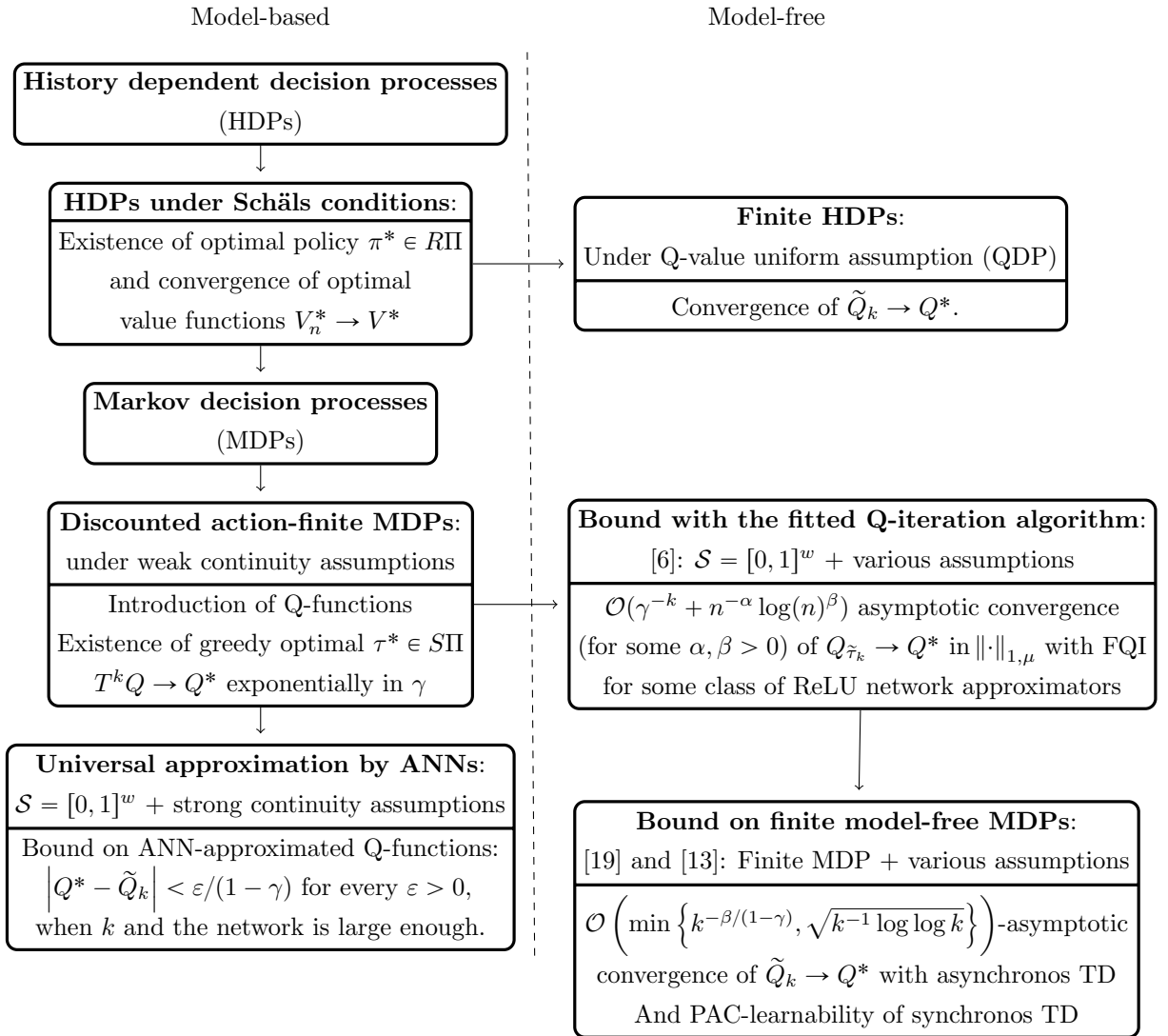


Figure 4.1: Some of the results presented in this thesis ordered by generality.



## 4.2 Conclusion

In this paper we have build up the theory behind Q-learning, covering decision models, optimality of policies, value functions and their iteration methods. This gave an introduction to Q-learning and a general framework from which to understand and compare results within the field. We then turned to model-free algorithms and presented convergence results for such in a variety of settings with state space being both finite and infinite and dynamics being allowed to depend on history or not. Finally we presented and proved convergence of the fitted Q-iteration algorithm as obtained in [6]. 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 optimality at present. Theoretically you could say that Q-learning is solved in many situations, since, as we have established, there is convergence guarantees for broad classes of problems. However as to how these convergence results relate to practical aspects of Q-learning we can still say little and as to the succes of the DQN of [15] we are not much further in understanding. The major reason is that the computational aspects are so important to their succes, and this part is mostly ignored in the results we have covered. Even though we establish results of the related FQI algorithm in [6], it is unclear if it captures the critical aspects of DQN, such as experience replay. In [6] convergence of FQI is guaranteed given corresponding increases in iterations, batch size and function space complexity. It is hard to interpret exactly how large these increases must be or whether it is practical.

## 4.3 Further directions

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

### Relation between FQI and DQN

Find a way to prove or disprove the conjecture in [6] that results about the convergence of FQI can have implications for the DQN algorithm.

### Suboptimality of policies

This is relating to decision processes and value functions. Through out the paper we discuss a wide array of approximations of  $Q^*$ . The default strategy is then to accept some close-enough approximation  $\tilde{Q}$  and then pick the greedy policy  $\tilde{\pi}$  with respect to  $\tilde{Q}$ . We then measure our deviation from optimality in terms of the distance  $\|Q^* - \tilde{Q}\|_\infty$ . However in most cases we do not estimate the deviation of  $Q_{\tilde{\pi}}$  from  $Q^*$  which from a theoretical point of view should be a better measure of the sub-optimality of  $\tilde{\pi}$  compared to  $\pi^*$ . Some sources like [6] succeed in bounding  $\|Q^* - Q_{\tilde{\pi}}\|_\infty$ , while many others make do with a bound on  $\|Q^* - \tilde{Q}\|_\infty$ . To this end it could be interesting to establish relations between  $\|Q^* - Q_{\tilde{\pi}}\|_\infty$  and  $\|Q^* - \tilde{Q}\|_\infty$ .

### 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 2.75). 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 Bernstein 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 [16, Rønn-Nielsen and Hansen (2014)] and [10, Kallenberg (2002)]. A good survey on results on optimal policy existence in the special case of Markov decision processes can be found in [7, Feinberg (2012)], however proofs in this source is either missing or sketched (as one must expect in a survey).

## 4.5 Credits

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

# Chapter A

## Appendices

### A.1 Basic definitions

**Definition A.1** (Order Topology). Given a totally ordered set  $(\mathcal{X}, <)$  the **order topology** is the topology generated by the subbase of sets on the form

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

**Definition A.2** ( $\sigma$ -algebra). A  **$\sigma$ -algebra**  $\Sigma$  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

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

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

**Theorem A.3.** For any pavement  $\Gamma \subseteq 2^{\mathcal{X}}$  of a set  $\mathcal{X}$  there exists a *smallest*  $\sigma$ -algebra  $\Sigma \subseteq 2^{\mathcal{X}}$  on  $\mathcal{X}$  satisfying

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

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

**Definition A.4** (Borel  $\sigma$ -algebra). For a topological space the **Borel**  $\sigma$ -algebra is the smallest  $\sigma$ -algebra containing all open sets.

**Definition A.5** (Product  $\sigma$ -algebra). Let  $(\mathcal{X}_i, \mathcal{A}_i)_{i \in I}$  be a collection of measurable spaces. the product  $\sigma$ -algebra

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

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

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

**Definition A.6** (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  $B \setminus A \in D$ ,
3. If  $A_1, A_2, \dots \in D$  with  $A_n \subseteq A_{n+1}$  for all  $n \in \mathbb{N}$  then  $\bigcup_{n=1}^{\infty} A_n \in D$ .

**Theorem A.7** (Dynkin's  $\pi$ - $\lambda$  theorem). Let  $P$  be a pavement of  $X$  which is stable under finite intersections (such are called  $\pi$ -systems) and  $D$  a Dynkin class (see definition A.6). If  $P \subseteq D$  then  $\sigma(P) \subseteq D$  where  $\sigma(P)$  is the smallest  $\sigma$ -algebra containing  $P$ .

**Definition A.8** (Measure). Given a measurable space  $(\mathcal{X}, \Sigma)$  a **measure** is a function  $\mu : \Sigma \rightarrow [0, \infty]$  satisfying

1.  $\mu(\emptyset) = 0$
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  $A_1, A_2, \dots \in \Sigma$ .

If furthermore  $\mu(\mathcal{X}) = 1$  then  $\mu$  is called a **probability measure**.

**Theorem A.9** (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.  $\emptyset \in \mathcal{S}$
2.  $S, T \in \mathcal{S} \implies S \cap T \in \mathcal{S}$
3. For  $S, T \in \mathcal{S}$  there exists finitely many disjoint subsets  $S_1, S_2, \dots, S_n \in \mathcal{S}$  so that  $S \setminus T = \bigcup_{i=1}^n S_i$ .

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

- i.  $\mu(\emptyset) = 0$
- ii. For a countable mutually disjoint collection of subsets  $S_1, S_2, \dots \in \mathcal{S}$  it holds that  $\mu\left(\bigcup_{i \in \mathbb{N}} S_i\right) = \sum_{i \in \mathbb{N}} \mu(S_i)$ .

Then  $\mu$  has an extension to a measure  $\mu$  on  $\sigma(\mathcal{S})$ . Furthermore if there exists an increasing sequence of subsets  $S_1 \subseteq S_2 \subseteq \dots \in \mathcal{S}$  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  $\mu$  extends uniquely to a probability measure on  $(\mathcal{X}, \sigma(\mathcal{S}))$ .

**Definition A.10** (Measurable function). A functions  $f : \mathcal{X} \rightarrow \mathcal{Y}$  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.11** (Lipschitz continuity). Let  $(\mathcal{X}, d_{\mathcal{X}})$ ,  $(\mathcal{Y}, d_{\mathcal{Y}})$  be metric spaces. A function  $f : \mathcal{X} \rightarrow \mathcal{Y}$  is said to **Lipschitz** with constant  $L > 0$  if

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

**Definition A.12** (Almost sure uniform convergence of random processes). A sequence of random processes  $X_n : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$  is said to converge **almost surely uniformly** to  $X : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$  if and only if

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

**Definition A.13** (Uniform convergence in probability of random processes). A sequence of random processes  $X_n : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$  is said to converge **uniformly in probability** to  $X : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$  if and only if

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

**Proposition A.14.**  $id_{\mathcal{P}(X)} = \mu \mapsto \kappa \circ \mu$  where  $\kappa(\cdot | x) = \delta_x(\cdot)$ . Thus  $\kappa$  can be seen as an identity mapping on  $\mathcal{P}(X)$ .

*Proof.*

$$\kappa \mu A = \int \delta_x(A) d\mu(x) = \mu A$$

□

**Theorem A.15** (Banach fixed point theorem). Let  $(\mathcal{X}, d)$  be a complete metric space and  $T : \mathcal{X} \rightarrow \mathcal{X}$  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  $T^k x \rightarrow x^*$  as  $k \rightarrow \infty$ , with rate  $d(T^k x, x^*) < \gamma^k d(x, x^*)$ .

## 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  $g_{x, B \times C} = y \mapsto 1_B(y) \phi(C | x, y)$  is measurable since it is the product (multiplication is measurable) of the measurable functions  $1_B$  and  $\phi(C | x, y)$  (measurability of  $\phi$  comes from definition 2.1.2 since  $\phi$  is a probability kernel). Further  $g_{x, B \times C} \in [0, 1]$  so it is  $\kappa(\cdot | x)$ -integrable. We also have  $g_{x, \mathcal{Y} \times \mathcal{Z}} = 1$  and  $g_{x, \emptyset} = 0$ . Since  $\{B \times C | B \in \Sigma_{\mathcal{Y}}, C \in \Sigma_{\mathcal{Z}}\}$  is a semi-ring generating  $\Sigma_{\mathcal{Y}} \otimes \Sigma_{\mathcal{Z}}$  by theorem A.9 we have that  $\phi \kappa(\cdot | x)$  extends uniquely to a probability measure on  $\mathcal{Y} \times \mathcal{Z}$ . Since  $(x, y) \mapsto 1_B(y) \phi(C | x, y)$  is measurable by proposition 2.4 we have that  $\phi \kappa(B \times C | x)$  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

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

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)

$$\int f(x', y) \kappa \delta_x(x', y) = \int \int f(x', y) d\kappa(y | x') d\delta_x(x') = \int f(x, y) \kappa(y | x)$$

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

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

□

### A.3 Lemmas for [6, Fan et al. (2020+)]

**Lemma A.16.** For  $x > 0$ .

$$\int_x^\infty e^{-t^2/2} \, dt \leq \frac{1}{x} e^{-x^2/2}$$

*Proof.* Observe that for  $t \geq x > 0$  we have  $1 \leq t/x$  so

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

□

## A.4 Disambiguation

- $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$ ,  $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ ,  $\overline{\underline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$ .
- $\text{id}_X := x \mapsto x$  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.
- $C_{\mathbb{K}}(X) := \{f : X \rightarrow \mathbb{K} \mid f \text{ continuous}\}$ , for  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ .  $C(X) = C_{\mathbb{R}}(X)$
- $C_b(X) := \{f \in C(X) \mid \exists K : \|f\|_{\infty} < K\}$ . I.e. the space of bounded continuous functions.
- ANN: abbreviation for artificial neural network see definition 2.71.
- $\delta_a$ : Dirac-measure of point  $a$ , i.e.  $\delta_a(A) = [a \in A] = 1_A(a)$ .
- $(\Omega, \mathcal{F}, \mathbb{P})$ : background probability space, that is source space for random variables.
- $\mathbb{B}_n$ : the  $n$ -dimensional Borel  $\sigma$ -algebra.
- $\lambda^n$ : the  $n$ -dimension Lebesgue measure.
- $\mathcal{P}(A)$ : the set of all probability measures on a measurable space  $(A, \Sigma_A)$ .
- $2^A$ : the powerset of  $A$ .
- $\mathcal{M}(A, B)$ : set of  $\Sigma_A$ - $\Sigma_B$  measurable functions.
- $\mathcal{L}_p$ : set of functions  $f$  with  $\|f\|_p < \infty$   $p \in [1, \infty]$ .
- $\mathbb{E}, \mathbb{E}_{\mu}$ : expectation, that is integration w.r.t. the measure  $\mathbb{P}$  or  $\mu$  respectively.
- Var: variance operator.

# Bibliography

- [1] 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.
- [2] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 2007. ISBN 1886529035.
- [3] 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.
- [4] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.
- [5] Adithya M. Devraj and Sean P. Meyn. Fastest convergence for q-learning. *CoRR*, abs/1707.03770, 2017. URL <http://arxiv.org/abs/1707.03770>.
- [6] 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>.
- [7] Eugene Feinberg. Total expected discounted reward mdps: Existence of optimal policies. 05 2012.
- [8] C. Heitzinger. *Simulation and Inverse Modeling of Semiconductor Manufacturing Processes*. 2002. URL <https://books.google.dk/books?id=LpmxcQAACAAJ>.
- [9] 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.
- [10] 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>.
- [11] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in Neural Information Processing*, 11, 04 1999.
- [12] F. William Lawvere. The category of probabilistic mappings. 1962.
- [13] 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>.



- [14] 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.
- [15] 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, Dhharshan 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>.
- [16] Anders Rønn-Nielsen and Ernst Hansen. *Conditioning and Markov properties*. 2014. ISBN 978-87-7078-980-6.
- [17] 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>.
- [18] 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](https://doi.org/10.1016/0304-4149(75)90031-9). URL <http://www.sciencedirect.com/science/article/pii/0304414975900319>.
- [19] Csaba Szepesvári. The asymptotic convergence-rate of q-learning. 01 1997.
- [20] Christopher Watkins. Learning from delayed rewards. 01 1989.
- [21] Christopher Watkins and Peter Dayan. Technical note: Q-learning. *Machine Learning*, 8: 279–292, 05 1992. doi: 10.1007/BF00992698.