

# A Theoretical Analysis of Fitted Q-Iteration

Jacob Harder

University of Copenhagen

April 26, 2020

## Abstract

This paper is mainly about Q-learning in the generality of continuous state and action spaces. We will present the theory necessary to discuss Q-learning in this generality, a variety of settings in which Q-learning has been analysed and the results of such analyses. In the course of this we discuss the relations between the various settings and their results. Finally we will present and prove a yet unpublished result by [Fan et. al] about the convergence of fitted Q-iteration in a quite general setting.

## 1 Introduction

### 1.1 Foreword

I came upon the idea to write about this subject because I was fascinated by the performance of Q-learning algorithms implemented by [Mnih et al.].

Coming from a background of mixed pure mathematics and probability theory the main purpose of this master thesis was initially to investigate what has been proven about the convergence of Q-learning algorithms. In particular Q-learning algorithms using (deep) ANNs.

In the course of this I discovered that the frameworks and settings in which various Q-learning algorithms are analysed varies greatly across literature. Also questions as to in which degree optimal strategies exist in these various frameworks turns out to be non-trivial when the state and action spaces are uncountable.

Therefore this paper is partially about building a framework for analysing Q-learning algorithms in a variety of settings. And partially to present the results that occur in each setting and discuss their importance and generality.

### 1.2 Reinforcement learning in general

In Reinforcement Learning (RL) in general we are concerned with finding an optimal policy for an agent in some environment. This environment is described by a so-called decision process consisting of a sequence of state and action spaces  $\mathcal{S}_1, \mathcal{A}_1, \mathcal{S}_2, \dots$  and rules  $P_1, R_1, P_2, \dots$  specifying which states and rewards and likely to follow after some action is chosen. In the general case  $\mathcal{S}_i, \mathcal{A}_i$  are measurable spaces and  $P_i, R_i$  are probability kernels on  $\mathcal{S}_{i+1}$  and  $\overline{\mathbb{R}}$ , respectively. One then attempts to find a *policy* (behavior or strategy) that maximizes the rewards returned from the environment. A policy and a distribution over starting states  $\mu \in \mathcal{P}(\mathcal{S}_1)$  give rise to a countable stochastic process,  $(X_i)_{i \in \mathbb{N}} = (S_i, A_i, R_i)_{i \in \mathbb{N}}$  that is a probability measure  $P_\mu^\pi$  on  $\mathcal{S}_1 \times \mathcal{A}_1 \times \dots$ .

Intuitively  $S_1$  is drawn from  $\mu$ , then for all  $i \in \mathbb{N}$   $A_i$  is drawn from  $\pi(\cdot \mid S_i)$ , a reward is then drawn from  $R(\cdot \mid S_i, A_i)$ , then  $S_{i+1}$  is drawn from  $P(\cdot \mid S_i, A_i)$  and so on. We devote a considerable part of this paper discussing how this is constructed in concrete term, in the chapters on measure theory and decision processes.

In RL there is a categorization of algorithms 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 paper we exclusively consider off-policy algorithms.

### 1.3 Measure theory

In this section we will present some background theory which is necessary for the constructions and derivations in the following sections. This is mainly standard theory of stochastic processes with some slight extensions. The reader may skip this if already familiar with such theory and use this section as a reference when going into the technical details in the following sections.

#### 1.3.1 Background and notation

We work with a background probability space  $(\Omega, \Sigma_\Omega, \mathbb{P})$ . For a measurable space  $(\mathcal{X}, \Sigma_\mathcal{X})$  we denote the set of probability measures on this space  $\mathcal{P}(\Sigma_\mathcal{X})$  or simply  $\mathcal{P}(\mathcal{X})$  when the  $\sigma$ -algebra is unambiguous. When taking cartesian products  $\mathcal{X} \times \mathcal{Y}$  of measurable spaces  $(\mathcal{X}, \Sigma_\mathcal{X}), (\mathcal{Y}, \Sigma_\mathcal{Y})$  we always endow such with the product  $\sigma$ -algebra  $\Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y}$ , unless otherwise specified. A map  $f : \mathcal{X} \rightarrow \mathcal{Y}$  is called  $\Sigma_\mathcal{X}$ - $\Sigma_\mathcal{Y}$  measurable provided  $f^{-1}(\Sigma_\mathcal{Y}) \subseteq \Sigma_\mathcal{X}$  and we denote the set of such functions  $\mathcal{M}(\Sigma_\mathcal{X}, \Sigma_\mathcal{Y})$ . By a random variable  $X$  on  $(\mathcal{X}, \Sigma_\mathcal{X})$  mean a  $\Sigma_\Omega$ - $\Sigma_\mathcal{X}$  measurable map.

#### 1.3.2 Kernels

**Definition 1** (Probability kernel). Let  $(\mathcal{X}, \Sigma_\mathcal{X}), (\mathcal{Y}, \Sigma_\mathcal{Y})$  be measurable spaces. A function

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

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

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

When the  $\sigma$ -algebras are unambiguous we shall simply say an  $\mathcal{X} \rightsquigarrow \mathcal{Y}$  kernel. For any  $x \in \mathcal{X}$  and  $f \in \mathcal{L}_1(\kappa(\cdot \mid x))$  we write the integral of  $f$  over  $\kappa(\cdot \mid x)$  as  $\int f(y) d\kappa(y \mid x)$ .

We now state some fundamental results on probability kernels

**Theorem 1** (Integration of a kernel). Let  $\mu \in \mathcal{P}(\mathcal{X})$  and  $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ . Then there exists a uniquely determined probability measure  $\lambda \in \mathcal{P}(\Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y})$  such that

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

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

*Proof.* We refer to [ref to EH markov, thm. 1.2.1]. □

Notice that by theorem 1 besides getting a probability measure on  $\mathcal{X} \times \mathcal{Y}$  we get an induced probability measure on  $\mathcal{Y}$  defined by  $B \mapsto (\kappa\mu)(\mathcal{X} \times B)$ . We will denote this measure by  $\kappa \circ \mu$ . This way  $\kappa$  can also be seen as a mapping from  $\mathcal{P}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{Y})$ . Also note that  $\kappa \circ \delta_x = \kappa(\cdot | x)$ .

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

**Theorem 2** (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

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

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

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

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

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

*Proof.* We refer to [ref to EH markov, thm. 1.3.2 + 1.3.3] □

**Proposition 1** (Composition of kernels). Let  $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}, \psi : \mathcal{Y} \rightsquigarrow \mathcal{Z}$  be probability kernels. Then

$$(\psi \circ \kappa)(A | x) := \int \psi(A | y) d\kappa(y | x), \quad \forall A \in \Sigma_{\mathcal{Z}}, x \in \mathcal{X}$$

is a  $\mathcal{X} \rightsquigarrow \mathcal{Z}$  probability kernel called the composition of  $\kappa$  and  $\psi$ . The composition operator  $\circ$  is associative, i.e. if  $\phi : \mathcal{Z} \rightsquigarrow \mathcal{W}$  is a third probability kernel then  $(\phi \circ \psi) \circ \kappa = \phi \circ (\psi \circ \kappa)$ . The associativity also extends to measures, i.e.  $\forall \mu \in \mathcal{X} : (\psi \circ \kappa) \circ \mu = \psi \circ (\kappa \circ \mu)$  and this is uniquely determined by  $\psi, \kappa$  and  $\mu$ .

*Proof.* The first assertion is a trivial verification of the two conditions in definition 1 and left as an exercise. For the associativity we refer to [todo ref to EH markov, lem. 4.5.4]. □

Proposition 1 actually makes the class of measurable spaces into a category [todo ref: see Lawvere, The Category of Probabilistic Mappings], with identity  $\text{id}_{\mathcal{X}}(\cdot | x) = \delta_x$ . Notice that the mapping  $(A, x) \mapsto \delta_x(A) \kappa(A | x)$  defines a probability kernel  $\mathcal{X} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$  which we could denote  $\text{id}_{\mathcal{X}} \times \kappa$ . Now if  $\psi : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$  is a kernel then by proposition 1 the composition  $(\text{id}_{\mathcal{X} \times \mathcal{Y}} \times \psi) \circ (\text{id}_{\mathcal{X}} \times \kappa)$  is a kernel  $\mathcal{X} \rightarrow \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$  which we will denote  $\psi\kappa$ . It inherits associativity from  $\circ$  and again this associativity extends to application on measures: if  $\mu$  is a measure on  $\mathcal{X}$  then  $\psi(\kappa\mu) = (\psi\kappa)\mu$ .

**Proposition 2.** Let  $\kappa : \mathcal{X} \rightarrow \mathcal{Y}$  be a probability kernel and  $f : \mathcal{Y} \rightarrow \overline{\mathbb{R}}$  be integrabel. Then  $x \mapsto \int f d\kappa(\cdot | x)$  is measurable into  $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$ .

*Proof.* Simple functions are measurable since  $\kappa$  is a kernel. Now extend by sums and limits. □

### 1.3.3 Kernel derived processes

Let  $(\mathcal{X}_n, \Sigma_{\mathcal{X}_n})_{n \in \mathbb{N}}$  be a sequence of measurable spaces. For each  $n \in \mathbb{N}$  define  $\mathcal{X}^n := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ ,  $\Sigma_{\mathcal{X}^n} := \Sigma_{\mathcal{X}_1} \otimes \cdots \otimes \Sigma_{\mathcal{X}_n}$  and let  $\kappa_n : \mathcal{X}^n \rightsquigarrow \mathcal{X}_{n+1}$  be a probability kernel. Then  $\kappa^n := \kappa_n \cdots \kappa_1$  is a kernel from  $\mathcal{X}_1$  to  $\mathcal{X}^n$ . So for any probability measure  $\rho_1 \in \mathcal{P}(\mathcal{X}_1)$  there exists a unique probability measure  $\rho_n$  on  $\mathcal{X}^n$  defined by  $\kappa^n \rho_1$ . This does however not imply existence of a measure on

$$(\mathcal{X}^\infty, \Sigma_{\mathcal{X}^\infty})$$

where  $\mathcal{X}^\infty := \prod_{n \in \mathbb{N}} \mathcal{X}_n$  and  $\Sigma_{\mathcal{X}^\infty} := \bigotimes_{n \in \mathbb{N}} \Sigma_{\mathcal{X}_n}$ . We will need this to talk about countable stochastic processes arising from decision models, so we include here a result by C. Ionescu-Tulcea (1949):

**Theorem 3** (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

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

*Proof.* We refer to [Kallenberg, Foundations, thm. 5.17]. □

We will also need to integrate with respect to starting conditions on such a stochastic process. To this end it is convenient to extend the Ionescu-Tulcea construction from a measure to a kernel. As I was not able to find a proof of this elsewhere an original proof is included here.

**Proposition 3** (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  $\kappa(A \mid x) = \mu_x(A)$  defines a probability kernel  $\kappa : \mathcal{X}_1 \rightarrow \mathcal{X}^\infty$ .

*Proof.* Since we already know 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  $A \in \Sigma_{\mathcal{X}^\infty} = \bigotimes_{i=1}^{\infty} \Sigma_{\mathcal{X}_i}$ . Let  $\phi_A = x \mapsto \mu_x(A)$  for all  $A \in \Sigma_{\mathcal{X}^\infty}$  and define

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

The cylinder algebra

$$\mathbb{C} = \{ A_1 \times \cdots \times A_i \times \mathcal{X}_{i+1} \times \cdots \mid A_i \in \Sigma_{\mathcal{X}_i}, i \in \mathbb{N} \}$$

Is a generator for  $\Sigma_{\mathcal{X}^\infty}$  stable under finite intersections. By construction  $\mathbb{C} \subseteq \mathbb{G}$  since

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

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

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

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

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

We will denote the Ionescu-Tulcea kernel  $\dots \kappa_2 \kappa_1$  or  $\prod_{i=1}^{\infty} \kappa_i$  or simply  $\kappa^{\infty}$ .

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

*Proof.* Let  $x \in \mathcal{X}_1$ . Notice that by associativity of the finitely induced measures  $\kappa_n \dots \kappa_1 \delta_x = (\kappa_n \dots \kappa_2)(\kappa_1 \delta_x)$ . This implies that

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

for all  $n \in \mathbb{N}$  and  $A \in \Sigma_{\mathcal{X}^n}$ . By the uniqueness in theorem 3 we are done.  $\square$

This lemma will come in handy when manipulating with integrals over kernel derived measures.

## 2 Decision models and value functions

In this section we will develop general theory about decision processes and value function (including Q-functions) that is used across all sources considered in this paper. We will also take up the question of optimal policy existence and prove this in different settings for reference in the latter sections.

### 2.1 History dependent decision process

We define in this section a quite general framework. We do this partly in the quest to have a united framework to talk about results from a variety of sources, and relate them to each other in generality. And partly to avoid defining various concepts such as value functions everytime a new context is considered. A source which uses a setup which is almost as general can be found in [ref. to Schal]. In this section recall that  $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$ ,  $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$  and  $\overline{\underline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$ .

**Definition 2** (History dependent decision process). A **history dependent decision process** (HDP) is determined by

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

for each  $n \in \mathbb{N}$  we define the so called **history** spaces

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

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

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

with associated product  $\sigma$ -algebras

3.  $(P_n)_{n \in \mathbb{N}}$  a sequence of  $\mathcal{H}_n \times \mathcal{A}_n \rightsquigarrow \mathcal{S}_{n+1}$  kernels called the **transition** kernels.
4.  $(R_n)_{n \in \mathbb{N}}$  a sequence of  $\mathcal{H}_{n+1} \rightsquigarrow \overline{\underline{\mathbb{R}}}$  kernels called the **reward** kernels.

The name *decision process* is used for many different processes across literature but many of them generalize to the above. Some authors use the name *dynamic programming model* to refer to such processes. Notice the slight irregularity in the beginning of the history spaces: We are missing a reward state after  $\mathcal{S}_1$ . We could avoid this by introducing some start reward we will do without.

**Assumption 1.** (Reward independence)  $P_n, R_n$  and policies are only allowed to depend on the past states and actions, and not the rewards.

In all sources known to this writer assumption 1 is assumed. This is a bit of a puzzle since it is obvious that one could want to define algorithms (policies) that take into account which rewards they received in the past. We will also do this but stick to the standard and never attempt to evaluate ideal value functions of policies that depend on rewards. Thus we will let assumption 1 hold from now on and throughout this paper.

The majority of sources considered in this paper also specialize with the following:

**Assumption 2** (One state and action space).  $\mathcal{S}_1 = \mathcal{S}_2 = \dots := \mathcal{S}$   $\mathcal{A}_1 = \mathcal{A}_2 = \dots := \mathcal{A}$

We will do without this for the rest of this section in order to present some results in the generality they deserve. And later we will look at settings which do not specialize this way. One could ask if it is possible to embed the general decision process into one with assumption 2 by setting  $\mathcal{S} := \bigcup_{i \in \mathbb{N}} \mathcal{S}_i$  and  $\mathcal{A} := \bigcup_{i \in \mathbb{N}} \mathcal{A}_i$  or similar. One attempt at this can be found in [BS SOC, chp. 10], but this will not be covered here.

Other ways to specialize include reducing one or both of the transition and reward kernels to functions defined on  $\mathcal{S} \times \mathcal{A}$ . These processes are often called *deterministic*, but the exact definitions vary across sources, and we will instead specify each setting individually.

For a decision process we can define

**Definition 3** (Policy). A (randomized) **policy**  $\pi = (\pi_n)_{n \in \mathbb{N}}$  is a sequence of  $\mathcal{H}_n \rightsquigarrow \mathcal{A}_n$  kernels. The set of all policies we denote  $R\Pi$ . The policy  $\pi$  is called **semi Markov** if each  $\pi_i$  only depends on the first and last state in the history and is called **Markov** if only the last. The sets are denoted  $sM\Pi$  and  $M\Pi$ . 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  $a_i \in \mathcal{A}_i$ . Under assumption 2 it makes sense to make a (Markov) policy  $(\pi, \pi, \dots)$ , where  $\pi$  only depends on the last state. Such a policy is called **stationary**, and the set of them denoted  $S\Pi$ . We denote the deterministic version of the policy classes by the letter  $D$ .

We have the following inclusions

$$\begin{array}{ccccccc} S\Pi & \subseteq & M\Pi & \subseteq & sM\Pi & \subseteq & R\Pi \\ \cup & & \cup & & \cup & & \cup \\ DS\Pi & \subseteq & DM\Pi & \subseteq & DsM\Pi & \subseteq & D\Pi \end{array}$$

**Proposition 4.** A dynamic programming model together with a policy  $\pi$  defines a probability kernel  $\kappa_\pi : \mathcal{S}_1 \rightarrow \mathcal{H}_\infty$ .

*Proof.* This is the Ionescu-Tulcea kernel generated by  $\dots R_2 P_2 \pi_2 R_1 P_1 \pi_1$ . □

This kernel yields a probability measure  $\kappa_\pi \mu$  on  $\mathcal{H}_\infty$  for every  $\mu \in \mathcal{S}_1$ . In particular for any  $s \in \mathcal{S}_1$   $\kappa_\pi \delta_s$  yields the measure  $\kappa(\cdot \mid s)$  and we shall occasionally write this  $\kappa_\pi s$  and integration with respect to it  $\mathbb{E}_s^\pi$ .

Across literature generally any function mapping a state space  $\mathcal{S}$  to  $\overline{\mathbb{R}}$  can be called a (state) **value** function. Similarly any  $\overline{\mathbb{R}}$  valued function on pairs of states and actions can be called (state) **action value** or **Q-** function. The idea behind such functions are commonly to estimate the cumulative rewards associated with a state or state-action pair and the trajectory of states it can lead to. In order to define some standard value functions we will need one of the following conditions:

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

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

The letter  $F$  comes from [todo ref Bertekas SOC]. When assuming either of  $(F^+)$  or  $(F^-)$  we ensure that the summation of finitely many rewards has a well defined mean in  $\overline{\mathbb{R}}$ , and then the following definition makes sense

**Definition 4** (Finite horizon value function). Let  $\underline{R}_i : \mathcal{H}_\infty \rightarrow \overline{\mathbb{R}}$  be the projection onto the  $i$ th reward. Define

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

called the  $k$ th finite horizon value function. When  $n = 0 \forall \pi : V_{0,\pi} = V_0 := 0$ .

The finite horizon value function measures the expected total reward of starting in state  $s$  and then follow the policy  $\pi$  for  $n$  steps. This way it measures the *value* of that particular state given a policy and *horizon* (number of steps). We would like to extend this to an infinite horizon value function, i.e. letting  $n$  tend to  $\infty$ . To ensure that the integral is well-defined we need one of the following conditions

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

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

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

Again the letters P, N and D are adopted from [todo ref Bertsekas SOC].

**Definition 5.** We define the infinite horizon value function by

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

The infinite horizon value function  $V_\pi$  measures the expected total reward after following the policy  $\pi$  an infinite number of steps.

**Remark 1.** Whenever we are working with the finite horizon value function we will always assume that either  $(F^+)$  or  $(F^-)$  holds without stating this explicitly. If a result only holds under e.g.  $(F^+)$  we will of course be explicit about this by marking it accordingly with a  $(F^+)$ .

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

**Remark 2.** Since we are under assumption 1, when talking about the finite or infinite value functions, we can actually reduce the reward kernels to functions  $r_i : \mathcal{H}_{i+1} \rightarrow \overline{\mathbb{R}} = h \mapsto \int r dR_i(r | h)$  (note that  $r_i$  is measurable due to proposition 2). Another way of stating this is that the value functions are indifferent to whether we use deterministic or stochastic rewards. This however does not mean that we can dispose completely of stochastic rewards, as they still make a difference to model-free algorithms that do not know the reward kernel, and therefore cannot simply integrate it.

For use later we mention some properties of these value functions.

**Proposition 5.** When well-defined the value functions  $V_{n,\pi}, V_\pi$  are measurable into  $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$ .

*Proof.* Use proposition 2. □

**Proposition 6.**  $\lim_{n \rightarrow \infty} V_{n,\pi} = V_\pi$  for all  $\pi \in R\Pi$ .

*Proof.* By monotone or dominated convergence. □

**Proposition 7.** Under (D) for any  $\pi \in R\Pi$  we have

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

*Proof.* For any  $\pi \in R\Pi$

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

This also covers  $V_{n,\pi}$ . □

As this bound will occur again and again we denote it

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

### 2.1.1 Optimal policies

Let  $(\mathcal{S}_n, \mathcal{A}_n, P_n, R_n)_{n \in \mathbb{N}}$  be a decision process.

**Definition 6** (Optimal value functions).

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

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.

**Proposition 8.** (D)

$$|V_k^*|, |V^*| \leq V_{\max}.$$

*Proof.* By proposition 7 all terms in the suprema are within this bound. □

**Remark 3.** An interesting fact about the optimal value functions is that they might not be Borel measurable [todo ref to counterexample] even in the finite case. After all we are taking a supremum over sets of policies which have cardinality of at least the continuum. However it is sometimes possible to show that they are. We will take these discussions as they occur in various settings.

At this point some central questions can be asked.

1. To which extend does an optimal policy  $\pi^*$  exist?
2. Does  $V_n^*$  converge to  $V^*$ ?
3. When can optimal policies be chosen to be Markov, deterministic, etc.?
4. Can an algorithm be designed to efficiently find  $V^*$  and  $\pi^*$ ?

These questions has been answered in a variety of settings. We will try to address them in order by strength of assumptions they require.



### 2.1.2 Schäl's theorem

In a quite general setting, questions 1 and 2 was investigated by M. Schäl in 1974 [todo ref. to On Dynamic Programming: Compactness of the space of policies, 1974]. Here some additional structure on our process is imposed:

**Setting 1** (Schäl). 1.  $V_\pi < \infty$  for all policies  $\pi \in R\Pi$ .

2.  $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})$  is assumed to be standard Borel. I.e.  $\mathcal{S}_n$  is a non-empty Borel subset of a Polish space and  $\Sigma_{\mathcal{S}_n}$  is the Borel subsets of  $\mathcal{S}_n$ .
3.  $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})$  is similarly assumed to be standard Borel.
4.  $\mathcal{A}_n$  is compact.
5.  $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N \geq n} \sup_{\pi \in R\Pi} \sum_{t=n+1}^N \mathbb{E}_s^\pi r_t \rightarrow 0$  as  $n \rightarrow \infty$ .

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

**Condition S.** 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  $s_1, \dots, s_n \in \mathcal{S}^n$ .

2.  $r_n$  is upper semi-continuous.

**Condition W.** 1. The function

$$(h_n, a_n) \mapsto P_n(\cdot \mid h_n, a_n)$$

is weakly continuous (hence the name **W**).

2.  $r_n$  is continuous.

**Theorem 4** (Schäl). 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 [todo ref: On Dynamic Programming: Compactness of the space of policies, M. Schäl 1974]. □

Schäl's theorem tells us that optimal policies exist in a wide class of decision processes. However in many cases we are looking at processes in which the next state is independent of the history. In such cases it makes sense to ask if optimal policies can be chosen within the system of policy subclasses. Such questions will be addressed in the next section.

## 2.2 The Markov decision process and its operators

**Definition 7** (Markov decision process). A **Markov decision process** (MDP) consists of

1.  $(\mathcal{S}, \Sigma_{\mathcal{S}})$  a measurable space of states.
2.  $(\mathcal{A}, \Sigma_{\mathcal{A}})$  a measurable space of actions.

3.  $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$  a transition kernel.
4.  $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow \overline{\mathbb{R}}$  a reward kernel.
5. An optional discount factor  $\gamma \in [0, 1]$  (when not discounting put  $\gamma = 1$ ).

This is a special case of the general decision process (??) with

- Assumption 2 is satisfied i.e.  $\mathcal{S}_1 = \mathcal{S}_2 = \dots, \mathcal{A}_1 = \dots = \mathcal{A}$ .
- $P_n$  depends only on  $s_n$  and  $a_n$  and does not differ with  $n$ . I.e.  $P_n(\cdot \mid s_1, \dots, s_n, a_n) = P(\cdot \mid s_n, a_n)$  for all  $n \in \mathbb{N}$ .
- $R_n$  depends only on  $s_n$  and  $a_n$  and does not differ with  $n$  except for a potential discount. I.e.  $R = R_n/\gamma^{n-1}$  for all  $n \in \mathbb{N}$

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

At this point it makes sense to define

**Definition 8** (The  $T$ -operators). For a stationary policy  $\pi$  and measurable  $V : \mathcal{S} \rightarrow \overline{\mathbb{R}}$  with  $V \geq 0$ ,  $V \leq 0$  or  $|V| < \infty$  we define the operators

$$\begin{aligned}
 P_\pi V &:= s \mapsto \int V(s') dP\pi(s' \mid s) \\
 T_\pi V &:= s \mapsto \int r(s, a) + \gamma V(s') d(P\pi)(a, s' \mid s) \\
 TV &:= s \mapsto \sup_{a \in \mathcal{A}} T_a V(s)
 \end{aligned}$$

where  $T_a = T_{\delta_a}$ .

We want to define

**Proposition 9** (Properties of the  $T$ -operators). Let  $\pi = (\pi_1, \pi_2, \dots)$  be a Markov policy.

1. The operators  $P_\pi, T_\pi$  and  $T$  commutes with limits.
2.  $V_{k,\pi} = T_{\pi_1} V_{k-1,(\pi_2,\dots)} = T_{\pi_1} \dots T_{\pi_k} V_0$ .
3.  $V_\pi = \lim_{k \rightarrow \infty} T_{\pi_1} \dots T_{\pi_k} V_0$
4. If  $\pi$  is stationary  $T_\pi V_\pi = V_\pi$ .
5. (D)  $T$  and  $T_\pi$  are  $\gamma$ -contractive on  $\mathcal{L}_\infty(\mathcal{S})$ .
6. (D)  $V_\pi$  is the unique bounded fixed point of  $T_\pi$  in  $\mathcal{L}_\infty(\mathcal{S})$

*Proof.*

1. By monotone or dominated convergence theorems.

2.

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

Now use this inductively.

3. This is by 2. and proposition 6.

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

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

$$|T^{\pi} V - T^{\pi} V'| = \gamma \left| \int V(s') - V'(s') dP_{\pi}(s' \mid s) \right| \leq \gamma K$$

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.

□

## 2.3 Q-functions

**Definition 9.** Let  $\pi \in R\Pi$ . Define

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

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

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

The letter Q originates to a PhD thesis by C. Watkins from 1989 [todo ref C. Watkins, 1989]. Upon his definition he noted

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

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

kernel will remain known and we will in this section see how the results of state-value functions translate to Q-functions. The results in this section are original in the generality here presented, as I was unable to find them elsewhere.

**Proposition 10.** Let  $\pi = (\pi_1, \pi_2, \dots) \in R\Pi$  then  $\lim_{k \rightarrow \infty} Q_{k,\pi} = Q_\pi$ . Furthermore it holds that  $|Q_{k,\pi}|, |Q_\pi|, |Q_k^*|, |Q^*| \leq V_{\max}$ .

*Proof.* By dominated convergence or monotone convergence and proposition 7.  $\square$

In parallel to the operators for state-value functions we define

**Definition 10** ( $T$  operators for Q-functions). For any stationary policy  $\pi \in S\Pi$  and measurable  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  with  $Q \geq 0, Q \leq 0$  or  $|Q| < \infty$  we define

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

where  $T_a = T_{\delta_a}$ .

**Proposition 11** (Properties of T-operators for Q-functions). Let  $\pi = (\pi_1, \pi_2, \dots) \in R\Pi$ .

1. If  $\pi$  is Markov and  $\mu$  is a stationary policy then  $T_\mu Q_{k,\pi} = r + \gamma \mathbb{E} T_\mu V_{k,\pi}$
2.  $Q_{k,\pi} = T_{\pi_1} \dots T_{\pi_k} Q_0$ .
3. For stationary  $\pi \in S\Pi$  we have  $T_\pi Q_\pi = Q_\pi$ .
4. (D)  $T_\pi$  is  $\gamma$ -contractive on  $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$  and  $Q_\pi$  is the unique fixed point of  $T_\pi$  in  $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ .

*Proof.*

1. This is essentially due to properties of the kernels. The idea is sketched here

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

2. Use 1. iteratively starting with  $\mu = \pi_1, \pi = (\pi_2, \pi_3, \dots)$ .
3. By 2.  $T_\pi Q_\pi = T_\pi(r + \gamma \mathbb{E} \lim_{k \rightarrow \infty} T_\pi^k V_0) = \lim_{k \rightarrow \infty} T_\pi(r + \gamma \mathbb{E} T_\pi^k V_0) = \lim_{k \rightarrow \infty} (r + \gamma \mathbb{E} T_\pi^{k+1} V_0) = r + \gamma \mathbb{E} \lim_{k \rightarrow \infty} T_\pi^{k+1} V_0 = r + \gamma \mathbb{E} V_\pi = Q_\pi$ .
4. The contrativeness of  $T_\pi$  follows from the same argument as for value functions. 2. and Banach fixed point theorem does the rest.

$\square$

**Definition 11.** Let  $\pi : \mathcal{S} \rightsquigarrow \mathcal{A}$  be a stationary policy. Define  $A_s = \operatorname{argmax}_{a \in \mathcal{A}} Q(s, a)$ . If there exist a measurable subset  $B_s \subseteq A_s$  for every  $s \in \mathcal{S}$  such that

$$\pi(B_s \mid s) = 1$$

then  $\pi$  is said to be **greedy** with respect to  $Q$  and is denoted  $\pi_Q$ .

**Proposition 12** (Properties of greedy policies). For any integrable  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \overline{\mathbb{R}}$  if  $\pi_Q$  is greedy with respect to  $Q$  then  $T_{\pi_Q}Q = TQ$ .

*Proof.*

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

□

## 2.4 Bertsekas-Shreve framework

The theory described here is largely based on [ref to Bertsekas-Shreve, Stochastic Optimal Control]. Their framework is cost-based as opposed to the this paper reward-based outset. This means that positive and negative, upper and lower, supremum and infimum, ect. are opposite to the source.

**Setting 2** (BS). • We consider an MDP  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$  (see definition 7).

- $\mathcal{S}$  and  $\mathcal{A}$  are Borel spaces.
- $\mathcal{A}$  is compact.
- $P(S \mid \cdot)$  is continuous for any  $S \in \Sigma_{\mathcal{S}}$ .
- $r(s, a) = \gamma^{1-i} \int x dR(x \mid s, a)$  is upper semicontinuous and uniformly bounded from above (least upper bound denoted  $0 < R_{\max} < \infty$ ).
- The policies must consist of universally measurable probability kernels.
- One of (P), (N) or (D) is always assumed.

The original setup in [ref to Bertsekas-Shreve, Stochastic Optimal Control] is slightly different than the setup here presented. Besides having a state and action space, it also features 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. Moreover it allows for constrains on the action space for each state. This is made precise by a function  $U : \mathcal{S} \rightarrow \Sigma_{\mathcal{A}}$  and a restriction on  $R\Pi$  that all policies  $\pi$  should satisfy  $\pi(U(s) \mid s) = 1$ . Lastly the rewards are interpreted as negative costs, and thus  $g$  is required to be semi *lower*continuous.

By setting  $P(\cdot \mid s, a) = f(s, a, p(\cdot \mid s, a))$  and maximizing rewards of upper semicontinuous instead of minimizing lower semicontinuous ones, we fully capture all aspects of the original process and its results, except the for the action constrains.

Notice that setting 2 implies  $(F^+)$ . Throughout this section are always assumed.

**Proposition 13.** Let  $\mathcal{X}, \mathcal{Y}$  be separable and metrizable,  $\kappa : \mathcal{X} \rightarrow \mathcal{Y}$  be a continuous probability kernel and  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \overline{\mathbb{R}}$  be Borel-measurable satisfying one of  $f \leq 0, f \geq 0, |f| < \infty$ . If  $f$  is bounded from above (below) and upper (lower) semicontinuous then

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

is bounded from above (below) and upper (lower) semicontinuous.

*Proof.* We refer to [BS SOC, prop. 7.31].  $\square$

**Proposition 14** (Prop. 8.6 in BS).  $V_k^* = T^k V_0$  and is upper semicontinuous. Furthermore for any  $k \in \mathbb{N}$  there exists a deterministic, Markov, Borel-measurable  $k$ -optimal policy  $\pi_k^* = (\pi_{k,1}^*, \pi_{k,2}^*, \dots, \pi_{k,k}^*, \dots) \in DM\Pi$ . These policies satisfy  $\pi_i^* = (\pi_{k,k-i}^*, \dots, \pi_{k,k}^*, \dots)$  for any  $i < k$ .

**Theorem 5** (Cor. 9.17.2 in BS). Under (N) or (D)  $V^* = \lim_{k \rightarrow \infty} V_k^*$  and is upper semicontinuous. Furthermore there exist a deterministic stationary, Borel-measurable policy  $\pi^*$ .

#### 2.4.1 Analytic setting

**Setting 3** (BS Analytic). The same as setting 2 except:  $P$  is not necessarily continuous.  $r$  is upper semianalytic.  $\mathcal{A}$  is not necessarily compact, but there exists a  $k \in \mathbb{N}$  such that  $\forall \lambda \in \mathbb{R}, n \geq k, s \in \mathcal{S}$

$$A_n^\lambda(s) = \left\{ a \in \mathcal{A} \mid r(s, a) + \gamma \int V_n^* P(\cdot \mid s, a) \geq \lambda \right\}$$

is a compact subset of  $\mathcal{A}$ .

**Theorem 6** (Prop. 9.17 BS). Under setting 3 we have  $V^* = \lim_{n \rightarrow \infty} V_n^*$  for all  $s \in \mathcal{S}$  and there exists a optimal policy  $\pi^*$  which is stationary and deterministic.

*Proof.* We refer to [todo ref to Bertsekas and Schreve, Stochastic Optimal Control: The Discrete-Time Case, prop. 9.17].  $\square$

#### 2.4.2 Implications for value-functions

Let setting 2 hold.

**Proposition 15.**  $V^* = V_{\pi^*} = T_{\pi^*} V^* = TV^*$

(D)  $V^*$  is the unique fixed point of  $T$  in  $\mathcal{L}_\infty(\mathcal{S})$ .

*Proof.* Since  $\pi^*$  is optimal  $V^* = V_{\pi^*}$  which by proposition 9 equals  $T_{\pi^*} V_{\pi^*}$ . By theorem 5 and proposition 14  $TV^* = T \lim_{k \rightarrow \infty} T^k V_0 = \lim_{k \rightarrow \infty} T^{k+1} V_0 = V^*$ . If (D) holds  $V^* \in \mathcal{L}_\infty(\mathcal{S})$  so by proposition 9 5. and 6. we are done.  $\square$

**Proposition 16.**

1.  $Q_k^* = r + \gamma \mathbb{E} V_k^*$  and is upper semicontinuous.
2. (N) (D)  $Q^* = r + \gamma \mathbb{E} V^*$  and is upper semicontinuous.
3. (N) (D)  $\sup_{a \in \mathcal{A}} Q^*(s, a) = V^*(s)$ .
4. (N) (D)  $Q^* = \lim_{k \rightarrow \infty} Q_k^*$ .
5. (N) (D)  $Q^* = Q_{\pi^*}$ .

*Proof.*

1. Since  $V_k^*$  is measurable due to proposition 14 we see that  $Q_k^* = \sup_{\pi \in R\Pi} (r + \gamma \mathbb{E} V_{k,\pi}) \leq r + \gamma \mathbb{E} V_k^* = r + \gamma \mathbb{E} V_{\pi_k^*} \leq Q_k^*$ . Proposition 13 gives upper semicontinuity.
2. Since  $V^*$  is measurable due to theorem 5. Now follow the argument for 1.

3. Let  $s \in \mathcal{S}$  then  $\sup_{a \in \mathcal{A}} Q^*(s, a) = \sup_{a \in \mathcal{A}} (r(s, a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V^*) = TV^*(s) = V^*(s)$ .
4. By monotone or dominated convergence and theorem 5.
5. By proposition 15 and 2.  $Q^* = r + \gamma \mathbb{E} V^* = r + \gamma \mathbb{E} V_{\pi^*} = Q_{\pi^*}$ .

□

**Proposition 17.**

1.  $TQ_k^* = r + \gamma \mathbb{E} TV_k^*$  and if  $\pi^* = (\pi_1^*, \pi_2^* \dots)$  is  $k$ -optimal then  $Q_k^* = T_{\pi_1^*} \dots T_{\pi_k^*} r = T^k r$ .
2.  $TQ^* = r + \gamma \mathbb{E} TV^*$  and  $TQ^* = Q^*$ .
3. (D)  $T$  is  $\gamma$ -contractive on  $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$  and  $Q^*$  is the unique fixed point of  $T$  in  $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ .

*Proof.*

- 1.

$$\begin{aligned}
TQ_k^*(s, a) &= T(r + \gamma \mathbb{E} V_k^*)(s, a) \\
&= r(s, a) + \gamma \int \sup_{a' \in \mathcal{A}} (r(s', a') + \gamma \mathbb{E}_{P(\cdot|s',a')} V_k^*) dP(s' | s, a) \\
&= r(s, a) + \gamma \int \sup_{a' \in \mathcal{A}} \left( r(s', a') + \gamma \int V_k^*(s'') dP(s'' | s', a') \right) dP(s' | s, a) \\
&= r(s, a) + \gamma \int TV_k^*(s') dP(s' | s, a)
\end{aligned}$$

To get  $Q_k^* = T^k r$  use this inductively  $Q_k^* = r + \gamma \mathbb{E} V_k^* = r + \gamma TV_{k-1}^* = TQ_{k-1}^* = \dots$ . The statement  $Q_k^* = T_{\pi_1^*} \dots T_{\pi_k^*} r$  is from proposition 11.

2. The argument from 1. also implies this first statement in 2. Now  $TQ^* = r + \gamma \mathbb{E} TV^* = r + \gamma \mathbb{E} V^* = Q^*$  by proposition 15.
3. The argument is similar to proposition 9 pt. 5.

□

**Corollary 1.** (D)

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

$$\|T^k Q - Q^*\|_\infty \leq \gamma^k \|Q - Q^*\|_\infty$$

*Proof.* This is directly from proposition 17 pt. 3.

□

**Proposition 18.**

1. Let  $\pi_i$  be greedy w.r.t.  $Q_{i-1}^*$  then  $(\pi_i, \pi_{i-1}, \dots, \pi_1)$  is  $i$ -optimal for any  $i \in \mathbb{N}$ .
2. (N) (D) Any greedy strategy for  $Q^*$  is optimal and such exist.

*Proof.* 1. Such greedy policies exist because  $Q_{k,\pi}$  is upper semicontinuous by proposition 14. For induction base observe that  $Q_{1,\pi_1} = T_{\pi_1} Q_0 = TQ_0 = Q_1^*$ . Now assume  $Q_{i-1,\pi_{i-1},\dots,\pi_1} = Q_{i-1}^*$ . Then  $Q_{i,(\pi_i,\dots,\pi_1)} = T_{\pi_i} Q_{i-1,(\pi_{i-1},\dots,\pi_1)} = T_{\pi_i} Q_{i-1}^* = TQ_{i-1}^* = Q_i^*$ .

2. Since  $Q$  is upper semicontinuous in the second entry the set  $A_s = \operatorname{argmax}_{a \in \mathcal{A}} Q(s, a)$  is non-empty and measurable for all  $s$ . Pick (by axiom of choice) an  $a_s \in A_s$  for every  $s \in \mathcal{S}$ . Then  $\pi(\cdot | s) = \delta_{a_s}$  is greedy with respect to  $Q$ . □

**Remark 4.** Most of the results of this section hold also under setting 3 with the addition that 'semicontinuous' is replaced by 'semianalytic'.

Based on the results established so far we can as a non-practical example design the following algorithm:

---

**Algorithm 1:** Simple theoretical Q-iteration

---

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

$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : r(s, a) \leftarrow \int x dR(x | s, a).$

$\tilde{Q}_0 \leftarrow r$

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

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

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

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

---

**Proposition 19.** (D)

The output  $\tilde{Q}_K$  of algorithm 1 converges to the optimal Q-function  $Q^*$  with rate  $\gamma^K$  concretely  $\|\tilde{Q}_K - Q^*\|_\infty \leq \gamma^K \|Q^*\|_\infty$ .

*Proof.* This is by corollary 1. □

### 2.4.3 Finite Q-iteration

Concluding on the results so far we have showed how if one knows the dynamics of a stationary decision process satisfying rather broad criteria, such as continuity and compactness, the optimal policy and state-value function can be found simply by iteration over the  $T$ -operator and picking a greedy strategy (see proposition 19). Of course this is practical computationally, only if the resulting  $Q$  functions can be represented and computed in finite space and time. This is trivially the case when

**Assumption 3.**  $\mathcal{S} \times \mathcal{A}$  is finite.

Say  $|\mathcal{S}| = k$  and  $|\mathcal{A}| = \ell$ . In this case the transition operator  $P$  can be represented as a matrix of *transition probabilities*

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

then the algorithm becomes



---

**Algorithm 2:** Simple finite Q-iteration

---

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

Set  $r \leftarrow (\int \text{rd}R(\cdot \mid s_1, a_1), \dots, \int \text{rd}R(\cdot \mid s_k, a_\ell))^T$

and  $\tilde{Q}_0 \leftarrow r$ .

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

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

    Update action-value function:

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

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

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

---

**Proposition 20.** The output  $\tilde{Q}_K$  from algorithm 2 is  $K$ -optimal and  $\|\tilde{Q}_K - Q^*\|_\infty \leq \gamma^K \|Q^*\|_\infty$ .

*Proof.* See proposition 19.  $\square$

## 2.5 Approximation

In this section we will look at what happens if we instead use approximations of the  $Q$ -functions and  $T$  operator. We first look at a naive approach using  $Q$ -functions.

Let  $\tilde{Q}_0$  be any bounded  $Q$ -function. Suppose we approximate  $T\tilde{Q}_0$  by a  $Q$ -function  $\tilde{Q}_1$  to  $\varepsilon_1 > 0$  precision and then approximate  $T\tilde{Q}_1$  and so on getting a sequence of  $Q$ -functions satisfying

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

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

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

Then we can bound

$$\begin{aligned} |Q^* - \tilde{Q}_k| &\leq |Q^* - T^k \tilde{Q}_0| + |T^k \tilde{Q}_0 - \tilde{Q}_k| \\ &\leq \gamma^k |Q^* - \tilde{Q}_0| + \varepsilon_a(k) \end{aligned}$$

These terms are sometimes called the *algorithmic* and *approximation* errors.

The algorithmic error quickly while the other depends on our step-wise approximations. For example  $\varepsilon_i(k) = \varepsilon$  we easily get the bound  $\varepsilon_a(k) = \varepsilon \frac{1-\gamma^k}{1-\gamma} \leq \frac{\varepsilon}{1-\gamma}$ . Or if  $\varepsilon_i \leq c\gamma^i$  we get  $\varepsilon_a(k) \leq ck\gamma^k \rightarrow 0$  as  $k \rightarrow \infty$ . Generally if one can show that  $\varepsilon_i \rightarrow 0$  we have

**Proposition 21.**  $\sum_{i=1}^k \gamma^{k-i} \varepsilon_i \rightarrow 0$  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  $m > M$

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

$\square$

Let  $\mathcal{F}$  be a class of functions

**Theorem 7** (Universal Approximation Theorem for ANNs). Let  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  be non-constant, bounded and continuous function. Let  $\varepsilon > 0$  and  $f \in C(I_m)$ . Then there exists a ANN  $F$  with one hidden layer and activation function  $\sigma$  such that

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

### 3 Hidden dynamics

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 1 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 3:** 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  $\tilde{Q}_k$ s are well defined, they cannot converge if  $R$  is not deterministic. Therefore this approach is not attractive in a continuous or stochastic setting.

There are broadly two ways of dealing with such

#### 3.1 Finite case

A common way to overcome the problem of convergence is called *temporal difference* (TD) learning and is based on the following update scheme

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

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  $\alpha_k (r' + \gamma \cdot \max_{a' \in \mathcal{A}} \tilde{Q}_k(s', a') - \tilde{Q}_k(s, a))$  occurring from rearranging eq. (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 by Watkins and Dayan [todo ref Watkins and Dayan, 1992] of a TD algorithm using Q-functions, which was extended slightly by [Jaakola, Jordan, Singh, 1993].

---

**Algorithm 4:** Finite asynchronos Q-learning

---

**Input:** MDP  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$  such that  $|\mathcal{S}||\mathcal{A}| < \infty$ , 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 8** (Watkins, Dayan). 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 ?? 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 must 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 were established by [Szepesvári 1996].

**Theorem 9** (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

$$|\tilde{Q}_t - Q^*| \leq B \frac{1}{t^{\beta(1-\gamma)}} \quad (2)$$

and

$$|\tilde{Q}_t - Q^*| \leq B \sqrt{\frac{\log \log t}{t}} \quad (3)$$

Here eq. (2) is tightest when  $\gamma > 1 - \beta/2$  otherwise eq. (3) is tighter.

A paper [MH, 2003] proves that  $Q$ -learning is PAC-learnable given some additional assumptions.

---

**Algorithm 5:** 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 10** (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 5) 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 4) 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 [MH]  $L$  is called the *covering rate*.

An interesting side note to theorem 10 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)  $K \geq C'(1/(1-\gamma)^{4/\omega} + 1/(1-\gamma)^{1/(1-\omega)})$ . 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  $K \geq C''(L^{2+1/\omega} + L^{1/(1-\omega)})$  which is optimized for  $\omega \approx 0.77$ . It is then shown (in [MH]) experimentally that  $\omega = 0.85$  is optimal using a way of generating random finite MDPs. It seems this value has since been adopted as a kind of standard value for the learning rate (see e.g. [Devraj Meyn, 2018])

### 3.1.1 History dependent setting

**Setting 4** (Finite HDP).

1. A history dependent decision process (see definition 2), 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 5.** Note that the finite setting 4 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 4 the existence of an optimal  $\pi^* \in R\Pi$  and that  $V_n^* \rightarrow V^*$ .

Within setting 4 Q-functions are generalized so that they are taking values in  $\mathcal{H}^* \times \mathcal{A}$ . We likewise generalize the  $T$  function by

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

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 4 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

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

**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 4 together with the state-uniformity condition is by [MH] called a *Q-Value uniform decision process* (QDP).

**Theorem 11** (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:

$$q_{t+1}(x, a) = q_t(x, a) + \alpha_t(x, a) \left( r' + \gamma \max_{a' \in \mathcal{A}} q_t(x', a') - q_t(x, a) \right), \quad x = \phi(h), r' = r(h, a) \quad (4)$$

**Theorem 12.** Within setting 4 assume

1. State-uniformity (assumption 4).
2. Any state is reached eventually under any policy (called *state-process ergodicity* in [MH]).
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. (4) 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. [MH] answers this by an array of examples showing the following relations of the classes of decision processes:

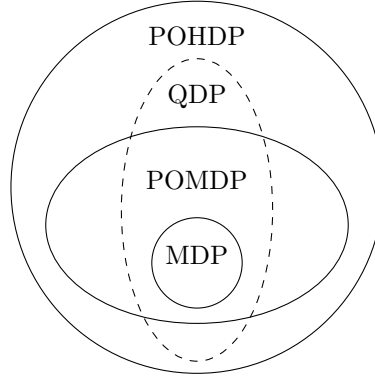


Figure 1: Classes of finite decision processes considered in [MH].

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

One point remains unclear after reading [MH]: Why is  $q^*$  and  $Q^*$  well defined by their recursive definition 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 6) of a general HDP? A sensible thing to ask would be that  $Q^*(h, a) = r(h, a) + \gamma \mathbb{E}_{P(\cdot|h,a)} V^*$ . However we will not go further into these details.

## 3.2 Results for continuous settings

### 3.2.1 Linear function approximation

This section is based on the 2007 article by [Melo, Ribeiro].

**Setting 5** (Melo, Rebeiro).

1. An MDP  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$  (see definition 7).
2. Discounted, i.e. (D) holds with  $\gamma \in [0, 1)$ .

3.  $\mathcal{S} \subseteq \mathbb{R}^w$  is compact.
4.  $\mathcal{A}$  is finite.
5.  $r_i$  is upper semicontinuous .

**Remark 6.** Item 5 was actually not part of the assumptions in (author?) [1]. 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 action value functions,  $\xi_i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ ,  $\forall i \in [M]$ . Denote  $\mathcal{Q} := \text{span}\{\xi_i \mid i \in [M]\}$  and for  $\theta \in \mathbb{R}^M$

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

We would now like to find the best approximation  $q^* \in \mathcal{Q}$  to  $Q^*$  within the span. If we measure distance by the  $\mathcal{L}_2$ -norm this  $\rho_{\mathcal{Q}} Q^*$  where  $\rho_{\mathcal{Q}}$  is the orthogonal projection on  $\mathcal{Q}$ .

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  $\mathcal{Q}$  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 6:** 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 6 found in [MR], 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  $\tau_A : \mathcal{X}^\infty \rightarrow \overline{\mathbb{N}} = \inf\{i \in \mathbb{N} \mid \rho_i \in A\}$ . Intuitively this function records the earliest time where the process enter the set  $A \subseteq \mathcal{X}$ . Define the function  $\eta_A : \mathcal{X}^\infty \rightarrow \overline{\mathbb{N}} = \sum_{i \in \mathbb{N}} 1_A \circ \rho_i$ . This function records the total number of times in which the process is inside the set  $A$ . Let  $\varphi \in \mathcal{P}(\mathcal{X})$  be a probability measure on  $\mathcal{X}$ .

**Definition 12** (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 13** (Positivity).

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

**Definition 14** (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 15** (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 16** (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  $\pi \in \Pi$ .

**Theorem 13** (Melo, Rebeiro). Let  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$  be an MDP as of setting 5. Let  $\pi \in \Pi$  be a stationary process and  $\mathfrak{P}$  the process kernel derived by  $P\pi$ . Assume that  $\mathfrak{P}$  is geometrically ergodic with invariant measure  $\mu$  and that  $\pi(a | s) > 0$  for all  $a \in \mathcal{A}$  and  $\mu_{\pi}$ -almost all  $s \in \mathcal{S}$ . Assume that  $\sum_{i=1}^M |\xi_i| \leq 1$ . Then if algorithm 6 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

$$Q_{\theta^*} = \rho_Q T Q_{\theta^*}$$

Furthermore the orthogonal projection is expressible as

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

(recall that  $\pi\mu(S \times A) = \int_S \pi(A | s) d\mu(s)$ ).

## 4 Deep fitted Q-iteration

### 4.1 Introduction

#### 4.1.1 The decision model

**Setting 6** (Fan et al.).

1. We're considering an MDP (definition 7). That is a state and action space  $(\mathcal{S}, \mathcal{A})$  and a transition and reward kernel  $P, R$  which only depends on the previous state-action pair.
2.  $S \subseteq \mathbb{R}^w$  is a compact subset of a euclidean space.
3.  $\mathcal{A}$  is finite.
4. Discounted factor satisfy  $0 < \gamma < 1$ .



### 4.1.2 Artificial neural networks

**Definition 17.** An ANN (Artificial Neural Network) with structure  $\{d_i\}_{i=0}^{L+1} \subseteq \mathbb{N}$ , activation functions  $\sigma_i = (\sigma_{ij} : \mathbb{R} \rightarrow \mathbb{R})_{j=1}^{d_i}$  and weights  $\{W_i \in M^{d_i \times d_{i-1}}, v_i \in \mathbb{R}^{d_i}\}_{i=1}^{L+1}$  is the function  $F : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{L+1}}$

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

where  $w_i$  is the affine function  $x \mapsto W_i x + v_i$  for all  $i$ .

Here  $\sigma_i(x_1, \dots, x_{d_i}) = (\sigma_{i1}(x_1), \dots, \sigma_{id_i}(x_{d_i}))$ .

$L \in \mathbb{N}_0$  is called the number of hidden layers.

$d_i$  is the number of neurons or nodes in layer  $i$ .

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

**Definition 18** (Sparse ReLU Networks). For  $s, V \in \mathbb{R}$  a **(s,V)-Sparse ReLU Network** is an ANN  $f$  with all activation functions being *ReLU* i.e.  $\sigma_{ij} = \max(\cdot, 0)$  and with weights  $(W_\ell, v_\ell)$  satisfying

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

Here  $\widetilde{W}_\ell = (W_\ell, v_\ell)$ . The set of them we denote  $\mathcal{F}(L, \{d_i\}_{i=0}^{L+1}, s, V)$ .

### 4.1.3 Fitted Q-Iteration

The algorithm analysed by [Fan et al] is

---

#### Algorithm 7: 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:

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

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

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

---

## 4.2 Notational deviations from [TODO ref YangXieWang]

Because  $\sigma$  is used ambiguously in theorem 14 we denote the probability distribution  $\sigma$  from [YangXieWang, thm. 6.2, p. 20] by  $\nu$  instead.

I avoid the shorthand defined in [YangXieWang, 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 my notation thus becomes  $\|f\|_n \rightsquigarrow \|f\|/n$ .

### 4.3 Assumptions

#### 4.3.1 Hölder Smoothness

**Definition 19** (Hölder smoothness). For  $f : \mathcal{S} \rightarrow \mathbb{R}$  we define

$$\|f\|_{C_w} := \sum_{|\alpha| < \beta} \|\partial^\alpha f\|_\infty + \sum_{\|\alpha\|_1 = \lfloor \beta \rfloor} \sup_{x \neq y} \frac{|\partial^\alpha(f(x) - f(y))|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \quad (5)$$

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 20.** 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 21.** 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

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

**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.

**Proposition 22.** Assume for all  $a \in \mathcal{A}$  that  $P(\cdot \mid s, a)$  is absolutely continuous w.r.t.  $\lambda^k$  with density  $p(\cdot \mid s, a)$ , Assume for all  $s' \in \mathcal{S}$  that  $s \mapsto p(s' \mid s, a)$  and  $s \mapsto r(s, a)$  are both in  $C_w([0, 1]^r, \beta, H)$ . Then  $T\mathcal{F}_0 \subseteq C_w([0, 1]^r, \beta, (1 + \gamma V_{\max})H)$ .

*Proof.* 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

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

□

#### 4.3.2 Concentration coefficients

**Definition 22** (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

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

**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

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

#### 4.4 Main theorem

**Theorem 14** (Yang, Xie, Wang). 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 7 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  $K$ .

## 4.5 Proofs

The proof of theorem 14 combines two results. The first on the error propagation and the second on the error occurring in a single step.

**Theorem 15** (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 16** (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 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 15 and theorem 16 are found below, but first we will show how to combine them to obtain theorem 14.

*Proof of main theorem.* Using theorem 15 we get

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq \frac{2\phi_{\mu,\nu}\gamma}{(1-\gamma)^2} \varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max} \quad (6)$$

where  $\varepsilon_{\max} = \max_{k \in [K]} \|T\tilde{Q}_{k-1} - \tilde{Q}_k\|_{2,\nu}$ . Using theorem 16 with  $Q = \tilde{Q}_{k-1}$ ,  $\mathcal{F} = \mathcal{F}_0$ ,  $\epsilon = 1$  and  $\delta = 1/n$ , we get

$$\varepsilon_{\max} \leq 6n^{-1} C_2^2 V_{\max}^2 \log(N_0) + 2\omega(\mathcal{F}_0) + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_0} + 16V_{\max}n^{-1} \quad (7)$$

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 following lemma by [Schmidt-Hieber 2019, thm. 5, p. 22]

**Lemma 2** (Approximation of Hölder Smooth Functions). 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  $f \in \mathcal{F}(L, \{d_j\}_{j=0}^{L+1}, s, \infty)$  with  $s \leq 141(r + \beta + 1)^{3+r}N(m + 6)$  such that

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

to each Hölder smooth part of  $g$  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

$$g_{jk} \in C_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j), \quad j \in [q], k \in [p_{j+1}] \quad (8)$$

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

$$\begin{aligned} h_1 &= g_1/(2H_1) + 1/2 \\ h_j(u) &= g_j(2H_{j-1}u - H_{j-1})/(2H_j) + 1/2, & j \in \{2, \dots, q-1\} \\ h_q(u) &= g_q(2H_{q-1}u - H_{q-1}) \end{aligned} \quad (9)$$

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

$$\begin{aligned} h_{1k} &\in C_{t_1}([0, 1]^{t_1}, \beta_1, 1) \\ h_{jk} &\in C_{t_j}([0, 1]^{t_j}, \beta_j, (2H_{j-1})^{\beta_j}), & j \in \{2, \dots, q-1\} \\ h_q &\in C_{t_q}([0, 1]^{t_q}, \beta_q, H_q(2H_{q-1})^{\beta_q}) \end{aligned} \quad (10)$$

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

$$W := \max\left(\left\{(2H_{j-1})^{\beta_j} \mid 1 \leq j \leq q-1\right\} \cup \left\{H_q(2H_{q-1})^{\beta_q}, 1\right\}\right) \quad (11)$$

By lemma 2 there exists a ReLU network

$$\hat{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \tilde{d}_j p_{j+1}, \dots, \tilde{d}_j p_{j+1}, p_{j+1}\right\}, (\tilde{s}_j + 4) \cdot p_{j+1}\right) \quad (12)$$

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

$$\left\|\hat{h}_{jk} - h_{jk}\right\|_\infty \leq (2W + 1)6^{t_j}N2^{-m} + W3^{\beta_j}N^{-\beta_j/t_j} \leq 2W3^{\beta_j}N^{-\beta_j/t_j} \quad (13)$$

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

**Lemma 3.** 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

$$\tilde{h}_{jk} \in \mathcal{F} \left( L_j + 2, \{t_j, \tilde{d}_j, \dots, \tilde{d}_j, 1\}, (\tilde{s}_j + 4)p_{j+1} \right) \quad (14)$$

and since  $h_{jk}([0, 1]^{t_j}) \in [0, 1]$  by eq. (13)

$$\left\| \tilde{h}_{jk} - h_{jk} \right\|_{\infty} = \left\| \tau \circ \hat{h}_{jk} - \tau \circ h_{jk} \right\|_{\infty} \quad (15)$$

$$\leq \left\| \hat{h}_{jk} - h_{jk} \right\|_{\infty} \quad (16)$$

$$\leq 2W3^{-\beta_j} N^{-\beta_j/t_j} \quad (17)$$

**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  $\tilde{f} \in \mathcal{F} \left( \tilde{L}, \{r, \tilde{d}, \dots, \tilde{d}, 1\}, \tilde{s} \right)$ . 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  $\tilde{s}$

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

where  $C_s = 142(2\eta + 6)\log_2(e)$ . Now we bound  $\|\tilde{f} - f(\cdot, a)\|_\infty$ . 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

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

so by induction and eq. (13)

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

and therefore

$$\begin{aligned} \omega(\mathcal{F}_0) &= \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} \\ &\leq C_N n^{-2\alpha^* \min_{j \in [q]} \beta_j^*/t_j} \\ &\leq C_N n^{-2\alpha^* \kappa^*} \end{aligned} \tag{18}$$

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 4** (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 18. 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 4 and  $n$  sufficiently large we can bound

$$\begin{aligned}
\log N_0 &= \log |\mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty)| \\
&\leq |\mathcal{A}| \cdot \log |\mathcal{N}_{1/n}| \\
&\leq |\mathcal{A}| (\tilde{s} + 1) \log(2(\tilde{L} + 1)\tilde{D}^2 n) \\
&\leq |\mathcal{A}| (c_{\tilde{s}} n^{\alpha^*} \log(n)^{\xi^*} + 1) 2 \log \left( 2(c_{\tilde{L}} \log(n)^{\xi^*} + 1) \prod_{\ell=1}^{\tilde{L}+1} (\tilde{d} + 1) \right) \\
&\leq 2|\mathcal{A}| (c_{\tilde{s}} n^{\alpha^*} \log(n)^{\xi^*} + 1) \log \left( 2(c_{\tilde{L}} \log(n)^{\xi^*} + 1)(6n^{\alpha^*} \log(n)^{\xi^*} + 1)^{\tilde{L}+1} \right) \\
&\leq 4|\mathcal{A}| c_{\tilde{s}} n^{\alpha^*} \log(n)^{\xi^*} (\tilde{L} + 1) \log \left( 24c_{\tilde{L}} \log(n)^{\xi^*} n^{\alpha^*} \log(n)^{\xi^*} \right) \\
&\leq 8|\mathcal{A}| c_{\tilde{s}} n^{\alpha^*} \log(n)^{\xi^*} c_{\tilde{L}} \log(n)^{\xi^*} (\alpha^* + 2) \log(n) \\
&= 8c_{\tilde{s}} c_{\tilde{L}} (\alpha^* + 2) n^{\alpha^*} \log(n)^{1+2\xi^*}
\end{aligned} \tag{19}$$

Finally, combining eq. (6), eq. (7), eq. (18), eq. (19) 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 15.

**Lemma 5.**  $TQ \geq T^\pi Q$  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.*

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

$\square$

**Lemma 6.** 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 22.

*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

$$\mathbb{E}_\mu[(P^{\tau_m} \dots P^{\tau_1})(f)] = \int (P^{\tau_m} \dots P^{\tau_1})(f) d\mu \quad (20)$$

$$= \int f d(P^{\tau_m} \dots P^{\tau_1} \mu) \quad (21)$$

$$= \int f \frac{d(P^{\tau_m} \dots P^{\tau_1} \mu)}{d\nu} d\nu \quad (22)$$

$$\leq \left\| \frac{d(P^{\tau_m} \dots P^{\tau_1} \mu)}{d\nu} \right\|_{2,\nu} \cdot \|f\|_{2,\nu} \quad (23)$$

$$\leq \kappa(m, \mu, \nu) \|f\|_{2,\nu} \quad (24)$$

Where eq. (22) is due to the Radon-Nikodym theorem and eq. (23) is Cauchy-Schwarz.  $\square$

*Proof of theorem 15.* 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 5 we can make the upper bound

$$\begin{aligned} Q^* - Q^{\pi_K} &= T^{\pi^*} Q^* - T^{\pi_K} Q^{\pi_K} \\ &= T^{\pi^*} Q^* + (T^{\pi^*} \tilde{Q}_K - T^{\pi^*} \tilde{Q}_K) + (T\tilde{Q}_K - T\tilde{Q}_K) - T^{\pi_K} Q^{\pi_K} \\ &= (T^{\pi^*} \tilde{Q}_K - T\tilde{Q}_K) + (T^{\pi^*} Q^* - T^{\pi^*} \tilde{Q}_K) + (T\tilde{Q}_K - T^{\pi_K} Q^{\pi_K}) \\ &\leq (T^{\pi^*} Q^* - T^{\pi^*} \tilde{Q}_K) + (T\tilde{Q}_K - T^{\pi_K} Q^{\pi_K}) \\ &= (T^{\pi^*} Q^* - T^{\pi^*} \tilde{Q}_K) + (T^{\pi_K} \tilde{Q}_K - T^{\pi_K} Q^{\pi_K}) \\ &= \gamma P^{\pi^*} (Q^* - \tilde{Q}_K) + \gamma P^{\pi_K} (\tilde{Q}_K - Q^{\pi_K}) \\ &= \gamma (P^{\pi^*} - P^{\pi_K}) (Q^* - \tilde{Q}_K) + \gamma P^{\pi_K} (Q^* - Q^{\pi_K}) \end{aligned} \quad (25)$$

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. (25) to obtain

$$Q^* - Q^{\pi_K} \leq \gamma U^{-1} (P^{\pi^*} (Q^* - \tilde{Q}_K) - P^{\pi_K} (Q^* - \tilde{Q}_K)) \quad (26)$$

**Step 2** Using lemma 5 for any  $i \in [K]$  we can get an upper bound

$$\begin{aligned}
Q^* - \tilde{Q}_{i+1} &= Q^* + (T\tilde{Q}_i - T\tilde{Q}_i) - \tilde{Q}_{i+1} + (T^{\pi^*}\tilde{Q}_i - T^{\pi^*}\tilde{Q}_i) \\
&= (Q^* - T^{\pi^*}\tilde{Q}_i) + (T\tilde{Q}_i - \tilde{Q}_{i+1}) + (T^{\pi^*}\tilde{Q}_i - T\tilde{Q}_i) \\
&= (T^{\pi^*}Q^* - T^{\pi^*}\tilde{Q}_i) + \varrho_{i+1} + (T^{\pi^*}\tilde{Q}_i - T\tilde{Q}_i) \\
&\leq T^{\pi^*}Q^* - T^{\pi^*}\tilde{Q}_i + \varrho_{i+1} \\
&= \gamma P^{\pi^*}(Q^* - \tilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{27}$$

and a lower bound

$$\begin{aligned}
Q^* - \tilde{Q}_{i+1} &= Q^* + (T\tilde{Q}_i - T\tilde{Q}_i) - \tilde{Q}_{i+1} + (T^{\pi_i}Q^* - T^{\pi_i}Q^*) \\
&= (T^{\pi_i}Q^* - T^{\pi_i}\tilde{Q}_i) + \varrho_{i+1} + (TQ^* - T^{\pi_i}Q^*) \\
&\geq T^{\pi_i}Q^* - T^{\pi_i}\tilde{Q}_i + \varrho_{i+1} \\
&= \gamma P^{\pi_i}(Q^* - \tilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{28}$$

Applying eq. (27) and eq. (28) iteratively we get

$$Q^* - \tilde{Q}_K \leq \gamma^K (P^{\pi^*})^K (Q^* - \tilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi^*})^{K-1-i} \varrho_{i+1} \tag{29}$$

and

$$Q^* - \tilde{Q}_K \geq \gamma^K (P^{\pi_{K-1}} \dots P^{\pi_0}) (Q^* - \tilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi_{K-1}} \dots P^{\pi_{i+1}}) \varrho_{i+1} \tag{30}$$

**Step 3** Combining eq. (29) and eq. (30) with eq. (26) we get

$$\begin{aligned}
Q^* - Q^{\pi_K} &\leq U^{-1} \left( \gamma^{K+1} ((P^{\pi^*})^{K+1} - P^{\pi_K} \dots P^{\pi_0}) (Q^* - \tilde{Q}_0) \right. \\
&\quad \left. + \sum_{i=0}^{K-1} \gamma^{K-i} ((P^*)^{K-i} - P^{\pi_K} \dots P^{\pi_{i+1}}) \varrho_{i+1} \right)
\end{aligned} \tag{31}$$

For shorthand define constants

$$\alpha_i = \frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}} \text{ for } 0 \leq i \leq K-1 \text{ and } \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} \tag{32}$$

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

$$O_i = (1-\gamma)/2U^{-1} [(P^{\pi^*})^{K-i} + (P^{\pi_K} \dots P^{\pi_{i+1}})] \tag{33}$$

$$O_K = (1-\gamma)/2U^{-1} [(P^{\pi^*})^{K+1} + (P^{\pi_K} \dots P^{\pi_0})] \tag{34}$$

Then by eq. (31)

$$|Q^* - Q^{\pi_K}| \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[ \sum_{i=0}^{K-1} \alpha_i O_i |\varrho_{i+1}| + \alpha_K O_K |Q^* - \tilde{Q}_0| \right] \tag{35}$$

So by linearity of expectation

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} = \mathbb{E}_\mu |Q^* - Q^{\pi_K}| \tag{36}$$

$$\leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[ \sum_{i=0}^{K-1} \alpha_i \mathbb{E}_\mu (O_i |\varrho_{i+1}|) + \alpha_K \mathbb{E}_\mu (O_K |Q^* - \tilde{Q}_0|) \right] \tag{37}$$

With the bound on rewards we (crudely) estimate

$$\mathbb{E}_\mu O_K \left| Q^* - \tilde{Q}_0 \right| \leq 2V_{\max} = 2R_{\max}/(1 - \gamma) \quad (38)$$

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

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

$$\mathbb{E}_\mu(O_i | \varrho_{i+1}) \quad (39)$$

$$= \frac{1-\gamma}{2} \mathbb{E}_\mu \left( U^{-1} [(P^{\pi_K})^{K-i} + P^{\pi_K} \dots P^{\pi_{i+1}}] | \varrho_{i+1} \right) \quad (40)$$

$$= \frac{1-\gamma}{2} \mathbb{E}_\mu \left( \sum_{j=0}^{\infty} [(P^{\pi_K})^j (P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1} P^{\pi_{K-1}} \dots P^{\pi_{i+1}}] | \varrho_{i+1} \right) \quad (41)$$

$$= \frac{1-\gamma}{2} \sum_{j=0}^{\infty} \mathbb{E}_\mu \left( [(P^{\pi_K})^j (P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1} P^{\pi_{K-1}} \dots P^{\pi_{i+1}}] | \varrho_{i+1} \right) \quad (42)$$

Notice that there are  $K - i + j$   $P$ -operators on both terms in the sum. Therefore we can employ lemma 6 twice. Moreover define  $\varepsilon_{\max} = \max_{i \in [K]} \|\varrho_i\|_{2,\nu}$ . Then

$$\begin{aligned} \mathbb{E}_\mu(O_i | \varrho_{i+1}) &\leq (1 - \gamma) \sum_{j=0}^{\infty} \gamma^j \kappa(K - i + j; \mu, \nu) \|\varrho_{i+1}\|_{2,\nu} \\ &\leq \varepsilon_{\max} (1 - \gamma) \sum_{j=0}^{\infty} \gamma^j \kappa(K - i + j; \mu, \nu) \end{aligned} \quad (43)$$

Using eq. (37), eq. (38) and eq. (43)

$$\begin{aligned} \|Q^* - Q^{\pi_K}\|_{1,\mu} &\leq \frac{2\gamma(1 - \gamma^{K+1})}{1 - \gamma} \left[ \sum_{i=0}^{K-1} \sum_{j=0}^{\infty} \alpha_i \gamma^j \kappa(K - i + j; \mu, \nu) \right] \varepsilon_{\max} \\ &\quad + \frac{4\gamma(1 - \gamma^{K+1})}{(1 - \gamma)^3} \alpha_K R_{\max} \end{aligned} \quad (44)$$

Focusing on the first term on RHS of eq. (44), if we then we can take the norm out of the sum as a constant. We are left with

$$\begin{aligned} &\sum_{i=0}^{K-1} \sum_{j=0}^{\infty} \alpha_i \gamma^j \kappa(K - i + j; \mu, \nu) \\ &= \sum_{i=0}^{K-1} \sum_{j=0}^{\infty} \frac{(1 - \gamma) \gamma^{K-i+j-1}}{1 - \gamma^{K+1}} \kappa(K - i + j; \mu, \nu) \\ &= \frac{1 - \gamma}{1 - \gamma^{K+1}} \sum_{j=0}^{\infty} \sum_{i=0}^{K-1} \gamma^{K-i+j-1} \kappa(K - i + j; \mu, \nu) \\ &\leq \frac{1 - \gamma}{1 - \gamma^{K+1}} \sum_{m=0}^{\infty} \gamma^{m-1} \cdot m \cdot \kappa(m; \mu, \nu) \\ &\leq \frac{1}{1 - \gamma^{K+1}(1 - \gamma)} \phi_{\mu,\nu} \end{aligned} \quad (45)$$

Where the last inequality is due to assumption 6. Combining eq. (44) and eq. (45) we arrive at

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq \frac{2\gamma \cdot \phi_{\mu,\nu}}{(1 - \gamma)^2} \cdot \varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1 - \gamma)^2} \cdot R_{\max} \quad (46)$$

□

Finally we show theorem 16.

**Lemma 7** (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 23** (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 8** (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 23.** 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 17** (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 16.* 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

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

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

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

$$\|Y\|^2 + \|\hat{Q}\|^2 - 2Y \cdot \hat{Q} \leq \|Y\|^2 + \|f\|^2 - 2Y \cdot f \quad (47)$$

$$\iff \|\hat{Q}\|^2 + \|TQ\|^2 - 2\hat{Q} \cdot TQ \leq \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \hat{Q} - 2Y \cdot f - 2\hat{Q} \cdot TQ + 2f \cdot TQ \quad (48)$$

$$\iff \|\hat{Q} - TQ\|^2 \leq \|f - TQ\|^2 + 2(Y - TQ) \cdot (\hat{Q} - f) \quad (49)$$

$$\iff \|\hat{Q} - TQ\|^2 \leq \|f - TQ\|^2 + 2\xi \cdot (\hat{Q} - f) \quad (50)$$

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

By this lemma we can deduce

$$\mathbb{E}(\xi \cdot (\hat{Q} - f)) = \mathbb{E}(\xi \cdot (\hat{Q} - TQ)) \quad (51)$$

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

$$\left| \mathbb{E}(\xi \cdot (\hat{Q} - TQ)) \right| \leq \left| \mathbb{E}(\xi \cdot (\hat{Q} - f_{k*})) \right| + \left| \mathbb{E}(\xi \cdot (f_{k*} - TQ)) \right| \quad (52)$$

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

$$\left| \mathbb{E} \xi \cdot (\hat{Q} - f_{k*}) \right| \leq \mathbb{E} \left( \|\xi\| \|\hat{Q} - f_{k*}\| \right) \leq \mathbb{E}(\|\xi\|) \sqrt{n} \delta \leq 2n V_{\max} \delta \quad (53)$$

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

$$\|\hat{Q} - f_{k*}\|^2 = \sum_{i=1}^n (\hat{Q}(X_i) - f_{k*}(X_i))^2 \leq \sum_{i=1}^n \delta^2 = n\delta^2 \quad (54)$$

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

$$\|\xi\|^2 = \sum_{i=1}^n (Y_i - TQ(X_i))^2 \leq \sum_{i=1}^n (2V_{\max})^2 = 4V_{\max}^2 n \quad (55)$$

To bound the second term in eq. (52) define

$$Z_j := \xi \cdot (f_j - TQ) \|f_j - TQ\|^{-1} \quad (56)$$

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

**Definition 24** (Sub-gaussian norm).

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

Because of proposition 24  $\xi_i(f_j(X_i) - TQ(X_i))$  is centered for any  $i \in [n]$  and

$$\|\xi_i(f_j(X_i) - TQ(X_i))\|_{\psi_2, \Sigma} \leq 2V_{\max} |f_j(X_i) - TQ(X_i)| \quad (57)$$

Therefore by lemma 7

$$\|Z_j\|_{\psi_2, \Sigma}^2 \leq \|f_j - TQ\|^{-2} \left\| \sum_{i=1}^n \xi_i(f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma}^2 \quad (58)$$

$$\leq \|f_j - TQ\|^{-2} C_1 \sum_{i=1}^n \|\xi_i(f_j(X_i) - TQ(X_i))\|_{\psi_2, \Sigma}^2 \quad (59)$$

$$\leq \|f_j - TQ\|^{-2} C_1 \sum_{i=1}^n 4V_{\max}^2 |f_j(X_i) - TQ(X_i)|^2 \quad (60)$$

$$= 4V_{\max}^2 C_1 \quad (61)$$

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

$$\mathbb{E} \exp \left( c Z_j^2 / \|Z_j\|_{\psi_2}^2 \right) \leq e \quad (62)$$

so

$$\mathbb{E} \max_{j \in N_\delta} Z_j^2 = \frac{\max_{j \in [N_\delta]} \|Z_j\|_{\psi_2}^2}{c} \mathbb{E} \left( \max_{j \in [N_\delta]} \frac{c Z_j^2}{\max_{k \in [N_\delta]} \|Z_k\|_{\psi_2}^2} \right) \quad (63)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \mathbb{E} \left( \max_{j \in N_\delta} \frac{c Z_j^2}{\|Z_j\|_{\psi_2}^2} \right) \quad (64)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \log \left( \mathbb{E} \max_{j \in N_\delta} \exp \left( \frac{c Z_j^2}{\|Z_j\|_{\psi_2}^2} \right) \right) \quad (65)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \log \left( \sum_{j \in [N_\delta]} \mathbb{E} \exp \left( \frac{c Z_j^2}{\|Z_j\|_{\psi_2}^2} \right) \right) \quad (66)$$

$$\leq \frac{4V_{\max}^2 C_1}{c} \log(e N_\delta) \quad (67)$$

$$\leq C_2^2 V_{\max}^2 \log(N_\delta) \quad (68)$$

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

$$\mathbb{E}(\xi \cdot (f_{k*} - TQ)) = \mathbb{E}(\|f_{k*} - TQ\| |Z_{k*}|) \quad (69)$$

$$\leq \mathbb{E}\left(\left(\|\hat{Q} - TQ\| + \|\hat{Q} - f_{k*}\|\right) |Z_{k*}| \right) \quad (70)$$

$$\leq \mathbb{E}\left(\left(\|\hat{Q} - TQ\| + n\delta\right) |Z_{k*}| \right) \quad (71)$$

$$\leq \left(\mathbb{E}\left(\|\hat{Q} - TQ\| + n\delta\right)^2\right)^{1/2} \left(\mathbb{E}Z_{k*}^2\right)^{1/2} \quad (72)$$

$$\leq \mathbb{E}\left(\|\hat{Q} - TQ\| + n\delta\right) \left(\mathbb{E}Z_{k*}^2\right)^{1/2} \quad (73)$$

$$\leq \left(\sqrt{\mathbb{E}\|\hat{Q} - TQ\|_2^2} + n\delta\right) \left(\mathbb{E}Z_{k*}^2\right)^{1/2} \quad (74)$$

$$\leq \left(\sqrt{\mathbb{E}\|\hat{Q} - TQ\|_2^2} + n\delta\right) C_2^2 V_{\max}^2 \log(N_\delta) \quad (75)$$

Where eq. (69) to eq. (70) is by the triangle inequality and eq. (73) to eq. (74) is proposition 23. Combining eq. (50), eq. (52), eq. (53) and eq. (75)

$$\mathbb{E}\|\hat{Q} - TQ\|^2 \leq \mathbb{E}\|f - TQ\|^2 + 4nV_{\max}\delta + \left(\sqrt{\mathbb{E}\|\hat{Q} - TQ\|^2} + \sqrt{n}\delta\right) C_2 V_{\max} \sqrt{\log(N_\delta)} \quad (76)$$

$$= C_2 V_{\max} \sqrt{n \log(N_\delta)} \sqrt{\mathbb{E}\|\hat{Q} - TQ\|^2} + nC_2^2 \delta V_{\max}^2 \log(N_\delta) + \mathbb{E}\|f - TQ\|^2 \quad (77)$$

**Lemma 9.** 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 9 applied to eq. (77)

$$\frac{1}{n} \mathbb{E}\|\hat{Q} - TQ\|^2 \leq \frac{(1+\kappa)^2}{\kappa} \frac{1}{n} C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa) \left( \delta C_2^2 V_{\max}^2 \log(N_\delta) + \frac{1}{n} \mathbb{E}\|f - TQ\|^2 \right) \quad (78)$$

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

$$\begin{aligned} \frac{1}{n} \mathbb{E} \left\| \hat{Q} - TQ \right\|^2 &\leq \frac{(1+\kappa)^2}{\kappa} \delta C_2^2 V_{\max}^2 \log(N_\delta) \\ &\quad + (1+\kappa) \left( C_2^2 V_{\max}^2 \log(N_\delta) + \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} \|f - g\|_\nu^2 \right) \\ &\leq \frac{(1+\kappa)^2}{\kappa} \frac{1}{n} C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa) \left( C_2^2 V_{\max}^2 \log(N_\delta) + \omega(\mathcal{F}) \right) \end{aligned} \quad (79)$$

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

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

$$\left| \left( \hat{Q}(x) - TQ(x) \right)^2 - (f_{k*}(x) - TQ(x))^2 \right| = \left| \hat{Q}(x) - f_{k*}(x) \right| \cdot \left| \hat{Q}(x) + f_{k*}(x) - 2TQ(x) \right| \quad (80)$$

$$\leq 4V_{\max} \delta \quad (81)$$

Using this twice we can say

$$(\hat{Q}(\hat{X}_i) - TQ(\hat{X}_i))^2 \quad (82)$$

$$\leq (\hat{Q}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 - (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 + (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 \quad (83)$$

$$\begin{aligned} &\leq (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 + (\hat{Q}(X_i) - TQ(X_i))^2 - (\hat{Q}(X_i) - TQ(X_i))^2 \\ &\quad + (f_{k*}(X_i) - TQ(X_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 + 4V_{\max} \delta \end{aligned} \quad (84)$$

$$\leq (\hat{Q}(X_i) - TQ(X_i))^2 + (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 + 8V_{\max} \delta \quad (85)$$

Thus we get

$$\left\| \hat{Q} - TQ \right\|_\sigma^2 \quad (86)$$

$$= \mathbb{E} \frac{1}{n} \sum_{i=1}^n (\hat{Q}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 \quad (87)$$

$$\leq \mathbb{E} \frac{1}{n} \sum_{i=1}^n \left( (\hat{Q}(X_i) - TQ(X_i))^2 + (f_{k*}(\tilde{X}_i) - TQ(\tilde{X}_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 \right) + 8V_{\max} \delta \quad (88)$$

$$= \frac{1}{n} \left\| \hat{Q} - TQ \right\|^2 + \frac{1}{n} \sum_{i=1}^n h_{k*}(X_i, \tilde{X}_i) + 8V_{\max} \delta \quad (89)$$

Where we define

$$h_j(x, y) := (f_j(y) - TQ(y))^2 - (f_j(x) - TQ(x))^2 \quad (90)$$

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

$$T := \max_{j \in [N_\delta]} \left| \sum_{i=1}^n h_j(X_i, \tilde{X}_i) / \Upsilon \right| \quad (91)$$

Then we can bound the middle term in eq. (89)

$$\mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n h_{k*}(X_i, \tilde{X}_i) \right) \leq \Upsilon / n \mathbb{E} \max_{j \in [N_\delta]} \left( \left| \sum_{i=1}^n h_j(X_i, \tilde{X}_i) / \Upsilon \right| \right) \quad (92)$$

$$\leq \Upsilon / n \mathbb{E} T \quad (93)$$

We want to use Bernsteins inequality (theorem 17) with  $U_i = h_j(X_i, \tilde{X}_i)$ . Therefore notice that  $|h_j| \leq \Upsilon^2$  and

$$\text{Var} h_j(X_i, \tilde{X}_i) = 2 \text{Var} (f_j(X_i) - TQ(X_i))^2 \quad (94)$$

$$\leq 2 \mathbb{E} (f_j(X_i) - TQ(X_i))^4 \quad (95)$$

$$\leq 2 \Upsilon^4 \quad (96)$$



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

$$\mathbb{E}T = \int_0^\infty \mathbb{P}(T \geq t) dt \quad (97)$$

$$\leq u + \int_u^\infty \mathbb{P}(T \geq t) dt \quad (98)$$

$$\leq u + \int_u^\infty 2N_\delta \exp\left(\frac{-t^2}{2\Upsilon t/3 + 4n\Upsilon^2}\right) dt \quad (99)$$

$$\leq u + 2N_\delta \int_u^\infty \exp\left(\frac{-t^2}{2\Upsilon^2(t/(3\Upsilon) + 2n)}\right) dt \quad (100)$$

$$\leq u + 2N_\delta \left( \int_u^{6n\Upsilon} \exp\left(\frac{-t^2}{8n\Upsilon^2}\right) dt + \int_{6n\Upsilon}^\infty \exp\left(\frac{-t}{4/3\Upsilon}\right) dt \right) \quad (101)$$

$$\leq u + 2N_\delta \left( \frac{8n\Upsilon}{2u} \exp\left(\frac{-u^2}{8n\Upsilon}\right) + \frac{4\Upsilon}{3} \exp\left(\frac{-24n\Upsilon}{3\Upsilon}\right) \right) \quad (102)$$

where we use lemma 10 from eq. (101) to eq. (102). Now set  $u = \Upsilon\sqrt{8n \log N_\delta}$  continuing from eq. (102) we have

$$\dots = \Upsilon\sqrt{8n \log N_\delta} + \frac{\Upsilon^2 8n N_\delta}{\Upsilon\sqrt{8n \log N_\delta}} \exp(-\log N_\delta) + 8/3 N_\delta \Upsilon \exp(-9/2n) \quad (103)$$

$$= \Upsilon 2\sqrt{2n} \left( \log N_\delta + \frac{1}{\log N_\delta} \right) + 8/3 N_\delta e^{-9/2n} \quad (104)$$

$$\leq 4\sqrt{2}\Upsilon\sqrt{n \log N_\delta} + 8/3\Upsilon \quad (105)$$

Inserting eq. (105) and eq. (79) into eq. (89)

$$\left\| \hat{Q} - TQ \right\|_\nu^2 \leq \frac{1}{n} \mathbb{E} \left\| \hat{Q} - TQ \right\|^2 + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1} + \delta) \quad (106)$$

$$\begin{aligned} &\leq \frac{(1+\kappa)^2}{\kappa} \frac{1}{n} C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa) \left( \delta C_2^2 V_{\max}^2 \log(N_\delta) + \omega(\mathcal{F}) \right) \\ &\quad + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1} + \delta) \end{aligned} \quad (107)$$

□

## 5 Further directions

### 5.0.1 Suboptimality of policies

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  $\tilde{p}^i$  with respect to  $\tilde{Q}$ . We then measure our deviation from optimality in terms of the distance  $\left\| Q^* - \tilde{Q} \right\|_\infty$ . However we do not estimate the deviation of  $Q_{\tilde{\pi}}$  from  $Q^*$  which is actually a better measure of the sub-optimality of  $\tilde{\pi}$  when comparing to  $\pi^*$ . To this end it could be interesting to some how establish relations between  $\left\| Q^* - Q_{\tilde{\pi}} \right\|_\infty$  and  $\left\| Q^* - \tilde{Q} \right\|_\infty$ .

## 6 Appendices

### 6.1 Disambiguation

- $\text{id} := x \mapsto x$  the identity function.

- $[\phi] := 1$  when  $\phi$  is true/holds and 0 otherwise, for a logical formula  $\phi$ .
- $[q] := \{1, \dots, q\}$  for  $q \in \mathbb{N}$ .
- $C_{\mathbb{K}}(X) := \{f : X \rightarrow \mathbb{K} \mid f \text{ continuous}\}$ ,  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ .  $C(X) = C_{\mathbb{R}}(X)$
- $C_b(X) := \{f \in C(X) \mid \exists K : \|f\|_{\infty} < K\}$ . I.e. the space of bounded continuous functions.
- ANN: abrv. artificial neural network see definition 17.
- $\delta_a$ : Dirac-measure of point  $a$ . I.e.  $\delta_a(A) = [a \in A]$ .
- $(\Omega, \mathcal{F}, \mathbb{P})$ : the underlying measure space of all random variables and processes when not otherwise specified.
- $\mathbb{B}_n$  the  $n$ -dimensional Borel  $\sigma$ -algebra.
- $\lambda^n$  the  $n$ -dimension Lebesgue measure.

## 6.2 Lemmas for Fan et al.

**Lemma 10.** 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

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

□

**Theorem 18** (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  $d(T^k x, x^*) < \gamma^k d(x, x^*)$ .

## 6.3 Other notes

**Definition 25** (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 26** (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

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

## References

- [1] Nobody Jr. My article, 2006.