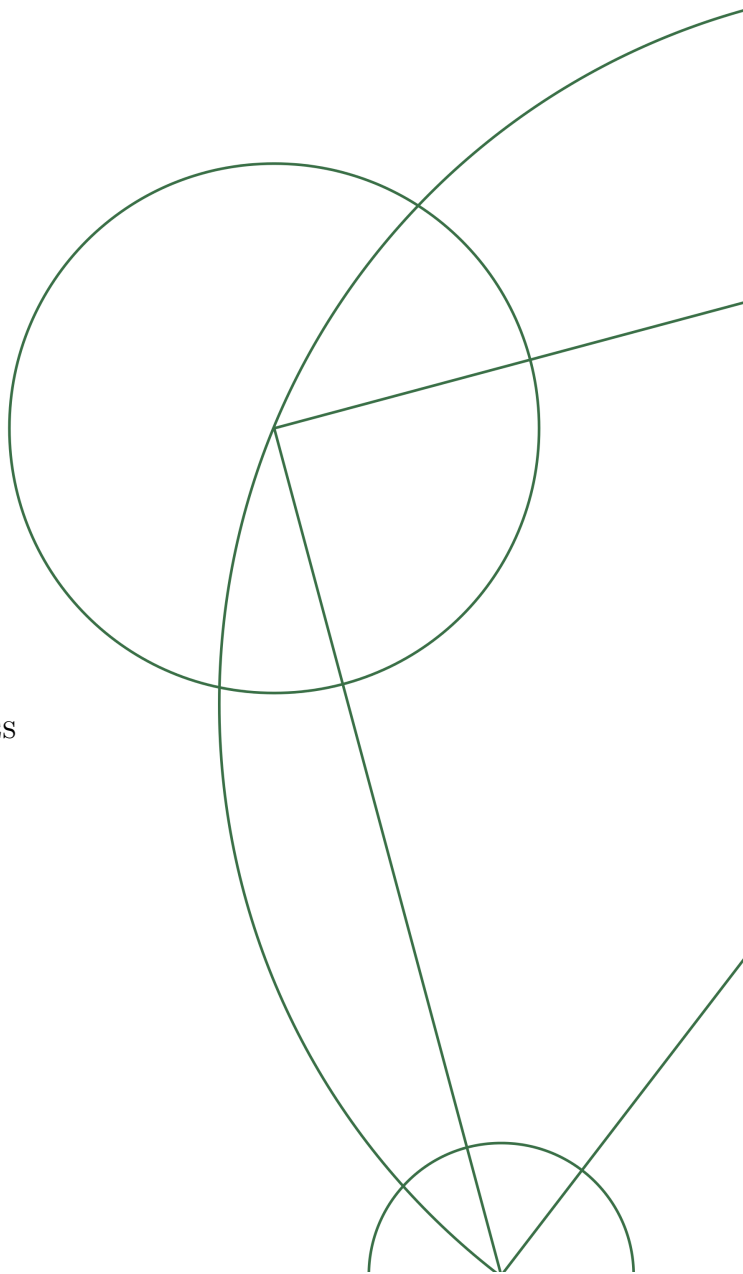
Department of Mathematical Sciences
UNIVERSITY OF COPENHAGEN

Jacob Harder

Theoretical aspects of Q-learning



Master thesis in Mathematics
Department of Mathematical Sciences
University of Copenhagen

Advisors
Stefan Horst Sommer

May 7, 2020

Abstract

This paper is mainly about the part of reinforcement learning that is called Q-learning, which is a category of algorithms which can *learn* from interaction with a decision process. We present the background theory for these algorithms, 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 Finally we will present and prove a yet unpublished result [5, Fan et al. (2020?)] which uses the fitted Q-iteration algorithm to prove convergence rates of Q-learning in the case of a continuous state space Markov decision process.

Contents

1 Introduction

1.1 Foreword

I came upon the idea to write about Q-learning when I was fascinated by the performance of the algorithms implemented by [11, Mnih et al. (2015)].

Coming from a mathematical background my main purpose of this master thesis was initially to investigate what has been proven about the convergence of Q-learning algorithms and what mathematical theory is relevant to establish such proofs. In particular Q-learning algorithms using artificial neural networks.

In the course of this I discovered that the frameworks and settings in which various Q-leaning algorithms are analysed varies greatly across litterature. 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 sequence of state and action spaces $\mathcal{S}_1, \mathcal{A}_1, \mathcal{S}_2, \dots$ and rules $P_1, R_1, P_2, \dots$ specifying which states and rewards and likely to follow after some action is chosen. One can then specify rules π, called a *policy*, for how the agent should act in every situation is the environment. Given an environment and a policy one obtains stochastic process, that is, a distribution on sequences of states, actions and rewards. One can then measure the performance of the policy by looking at the expected sum of rewards called the *value function* V_π of the policy. The goal of reinforcement learning is to find an optimal policy π^*, maximizing the value function.

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

A particular kind class of environments are called Markov decision processes (MDPs), and work with the same state space $\mathcal{S}$, action space $\mathcal{A}$ and rules P, R throughout the process. In an MDP we can use a value function V_1 to obtain a policy π_1 by choosing actions leading to states with high values (according to V_1). We can then evaluate value of π_1 yielding a new value function V_2. This process can be continued indefinitely yielding a sequence of value functions and policies. Variations of this idea are called *value iteration* and *policy iteration*, and have been shown to converge to the optimal policy in many cases.

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

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

When the process dynamics are hidden designing algorithms becomes trickier. In such settings approaches to the problem fall in two categories. In the *direct* approaches one attempts to estimate the process dynamics first and then afterwards methods for the known-dynamics are applied. The *indirect* approaches basically covers *the rest*. Here the process dynamics is not directly estimated but somehow captured by the Q-function anyway. In the indirect category we find the popular *temporal difference* algorithms on which *fitted Q-iteration* and *deep Q-learning* as used in [11] is based.

A further categorization of RL-algorithms can be made into *off-policy* and *on-policy* classes. This is simply whether the algorithm learns from data (states, actions and rewards) arising from following its own policy (on-policy) or it can learn from more arbitrary data (off-policy). This *more arbitrary data* could for example be the trajectory of another algorithm when interacting with a decision process, or simply state-action-reward pairs drawn from some distribution. As an example *fitted Q-iteration* is off-policy while *deep Q-learning* is on-policy.

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 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 σ-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 σ-algebra $\Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y}$, unless otherwise specified. A map $f : \mathcal{X} \to \mathcal{Y}$ 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 Σ_Ω-$\Sigma_\mathcal{X}$ measurable map.

1.3.2 Kernels

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

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

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

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

When the σ-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 $\int f(y) \, \mathrm{d}\kappa(y \mid x)$.

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

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

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

Proof. We refer to [12] 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 κ can also be seen as a mapping from $\mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$. Also note that $\kappa \circ \delta_x = \kappa(\cdot \mid x)$.

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

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

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

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

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

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

Proof. We refer to [12] thm. 1.3.2 and 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

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

is a $\mathcal{X} \rightsquigarrow \mathcal{Z}$ probability kernel called the composition of κ and ψ. 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 ψ, κ and μ.

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 [12] lem. 4.5.4. □

Proposition 1 actually makes the class of measurable spaces into a category [8, Lawvere (1962)], with identity $\mathrm{id}_{\mathcal{X}}(\cdot \mid x) = \delta_x$. Notice that the mapping $(A \times B, x) \mapsto \delta_x(A)\kappa(B \mid x)$ defines a probability kernel $\mathcal{X} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$ which we could denote $\mathrm{id}_{\mathcal{X}} \times \kappa$. Now if $\psi : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ is a kernel then by proposition 1 the composition $(\mathrm{id}_{\mathcal{X} \times \mathcal{Y}} \times \psi) \circ (\mathrm{id}_{\mathcal{X}} \times \kappa)$ is a kernel $\mathcal{X} \to \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ which we will denote $\psi\kappa$. It inherits associativity from $\circ$ and again this associativity extends to application on measures: if μ is a measure on $\mathcal{X}$ then $\psi(\kappa\mu) = (\psi\kappa)\mu$.

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

Proof. Simple functions are measurable since κ 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 $\mathcal{X}^{\underline{n}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $\Sigma_{\mathcal{X}^{\underline{n}}} := \Sigma_{\mathcal{X}_1} \otimes \cdots \otimes \Sigma_{\mathcal{X}_n}$ and let $\kappa_n : \mathcal{X}^{\underline{n}} \rightsquigarrow \mathcal{X}_{n+1}$ be a probability kernel. Then $\kappa^{\underline{n}} := \kappa_n \dots \kappa_1$ is a kernel from $\mathcal{X}_1$ to $\mathcal{X}^{\underline{n+1}}$. So for any probability measure $\rho_1 \in \mathcal{P}(\mathcal{X}_1)$ there exists a unique probability measure ρ_n on $\mathcal{X}^{\underline{n}}$ defined by $\kappa^{\underline{n-1}}\rho_1$. This however does not imply existence of a measure on

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

where $\mathcal{X}^{\underline{\infty}} := \prod_{n\in\mathbb{N}} \mathcal{X}_n$ and $\Sigma_{\mathcal{X}^{\underline{\infty}}} := \bigotimes_{n\in\mathbb{N}} \Sigma_{\mathcal{X}_n}$. 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 $\rho \in \mathcal{P}(\mathcal{X}^{\underline{\infty}})$ such that

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

Proof. We refer to [7, Kallenberg (2002)] 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 μ_x denote the Ionescu-Tulcea measure of a sequence of probability kernels $\kappa_i : \mathcal{X}^{\underline{i}} \to \mathcal{X}_{i+1}$ with starting measure δ_x on $\mathcal{X}_1$ for any $x \in \mathcal{X}_1$. Then $\kappa(A \mid x) = \mu_x(A)$ defines a probability kernel $\kappa : \mathcal{X}_1 \to \mathcal{X}^{\underline{\infty}}$.

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

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

The cylinder algebra

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

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

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

and any $\kappa^{\underline{i-1}}$ is a kernel (proposition 1). We will show that $\mathbb{G}$ is a Dynkin class. Then by Dynkins π-λ theorem (see theorem 20)

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

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

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

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

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 $A \in \Sigma_{\mathcal{X}_{\underline{n}}}$. By the uniqueness in theorem 3 we are done. □

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, including the question of optimal policy existence.

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

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

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

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

with associated product σ-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 litterature but many of them generalize to the above. Some authors also use the name *dynamic progamming model*.

Notice the slight irregularity in the beginning of the history spaces: We are missing a reward state after $\mathcal{S}_1$. This could have been avoided by introducing some start reward but 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)**.** $\mathcal{S}_1 = \mathcal{S}_2 = \ldots := \mathcal{S}$, $\mathcal{A}_1 = \mathcal{A}_2 = \ldots := \mathcal{A}$

We will do without this for the rest of this section in order to present some results in the generality they deserve. 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 [1] chapter 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 π is called **semi Markov** if each π_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 π is called **deterministic** if all π_i are degenerate, i.e. for all i we have $\pi_i(\{a_i\} \mid h_i) = 1$ for some $a_i \in \mathcal{A}_i$. Under assumption 2 it makes sense to make a (Markov) policy $(\pi, \pi, \ldots)$, where π 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

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

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

Proof. This is the Ionescu-Tulcea kernel generated by $\ldots R_2P_2\pi_2R_1P_1\pi_1$. □

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 $\kappa_\pi(\cdot \mid s)$ and we shall occasionally write this $\kappa_\pi s$ and integration with respect to it $\mathbb{E}_s^\pi$.

Across litterature 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)**.** $\int_{[0,\infty]} x \, \mathrm{d}R_i(x \mid h) < \infty$ for all $h \in \mathcal{H}_{i+1}$ and $i \in \mathbb{N}$

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

The letter F comes from [1]. 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 $\underline{R}_i : \mathcal{H}_\infty \to \overline{\mathbb{R}}$ be the projection onto the ith reward. Define

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

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

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

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

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

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

Again the letters P, N and D are adopted from [1].

Definition 5. We define the infinite horizon value function by

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

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

Remark 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 be marking it accordingly with a (F^+).

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

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 $r_i : \mathcal{H}_{i+1} \to \overline{\mathbb{R}} = h \mapsto \int r \, \mathrm{d}R_i(r \mid h)$ (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. $\lim_{n\to\infty} V_{n,\pi} = V_\pi$ for all $\pi \in R\Pi$.

Proof. By monotone or dominated convergence. □

Proposition 7. Under (D) for any $\pi \in R\Pi$ we have
$\left|V_{n,\pi}\right|, |V_\pi| \leqslant R_{\max}(1-\gamma) < \infty$.

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

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

This also covers $V_{n,\pi}$. □

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)**.**

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

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

Proposition 8. (D)
$\left|V_k^*\right|, |V^*| \leqslant V_{\max}$.

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

Remark 3. It is known that the optimal value function might not be Borel measurable (see ex. 2 p. 233 [1]). Perhaps this is not suprising since we are taking a supremum over sets of policies which have cardinality of at least the continuum. However it is often 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 π^* exist?
2. Does V_n^* converge to V^*?
3. When can optimal policies be chosen to be Markov, deterministic, etc.?
4. Can an algorithm be designed to efficiently find V^* and π^*?

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 Schäls theorem

In a quite general setting, questions 1 and 2 is investigated in [14, Schäl (1975)]. 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 S_n.

3. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})$ is similarly assumed to be standard Borel.

4. $\mathcal{A}_n$ is compact.

5. $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N \geqslant n} \sup_{\pi \in R\Pi} \sum_{t=n+1}^{N} \mathbb{E}_s^\pi r_n \to 0$ as $n \to \infty$.

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

Condition S. 1. The function

$$(a_1, a_2, \ldots, a_n) \mapsto P_n(\cdot \mid s_1, a_1, s_2, a_2, \ldots, s_n, a_n)$$

is set-wise continuous (hence the name **S**) for all $s_1, \ldots, s_n \in \mathcal{S}^{\underline{n}}$.

2. r_n is upper semi-continuous.

Condition W. 1. The function

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

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

2. r_n is continuous.

Theorem 4 (Schäl)**.** When either (S) or (W) hold then

1. There exist an optimal policy $\pi^* \in R\Pi$.

2. $V_n^* \to V^*$ as $n \to \infty$.

Proof. We refer to [14]. □

Schäls 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 in 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. $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow \overline{\underline{\mathbb{R}}}$ a reward kernel.

5. An optional disount factor $\gamma \in [0, 1]$ (when not discounting put $\gamma = 1$).

This is a special case of the history dependent decision process (definition 2) with

- Assumption 2 is satisfied i.e. $\mathcal{S}_1 = \mathcal{S}_2 = \cdots = \mathcal{S}, \quad \mathcal{A}_1 = \mathcal{A}_2 = \cdots = \mathcal{A}$.
- P_n depends only on s_n and a_n and does not differ with n, i.e. $P_n(\cdot \mid s_1, \dots, s_n, a_n) = P(\cdot \mid s_n, a_n)$ 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 π and measurable $V : \mathcal{S} \to \overline{\underline{\mathbb{R}}}$ with $V \geqslant 0$, $V \leqslant 0$ or $|V| < \infty$ we define the operators

$$P_\pi V := s \mapsto \int V(s') \, \mathrm{d}P\pi(s' \mid s)$$

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

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

where $T_a = T_{\delta_a}$.

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

1. The operators P_π, T_π 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. $V_\pi = \lim_{k\to\infty} T_{\pi_1} \dots T_{\pi_k} V_0$
4. If π is stationary $T_\pi V_\pi = V_\pi$.
5. (D) T and T_π are γ-contractive on $\mathcal{L}_\infty(\mathcal{S})$.
6. (D) V_π is the unique bounded fixed point of T_π in $\mathcal{L}_\infty(\mathcal{S})$

Proof.

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

   Now use this inductively.

3. This is by 2. and proposition 6.

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

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

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

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

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

□

2.3 Q-functions

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

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

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

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

The idea of Q-functions (and the letter Q) originates to [16, Watkins (1989)]. Upon the definition he notes

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

A clear advantage of working with Q-function $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ rather than a value function $V : \mathcal{S} \to \overline{\mathbb{R}}$, is that finding the optimal action in state s requires only a maximization over the Q-function itself: $a = \operatorname{argmax}_{a\in\mathcal{A}} Q(s,a)$. This should be compared to finding a best action according to a value function V: $a = \operatorname{argmax}_{a\in\mathcal{A}} r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V$. 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. (D)
$\lim_{k\to\infty} Q_{k,\pi} = Q_\pi$. Furthermore it holds that $\left|Q_{k,\pi}\right|, |Q_\pi|, \left|Q_k^*\right|, |Q^*| \leqslant V_{\max}$.

Proof. By dominated convergence or monotone convergence and proposition 7. □

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 $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ with $Q \geqslant 0, Q \leqslant 0$ or $|Q| < \infty$ we define

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

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

where $T_a = T_{\delta_a}$.

Proposition 11 (Properties of T-operators for Q-functions)**.** Let $\pi = (\pi_1, \pi_2, \dots) \in M\Pi$ be a Markov policy and $\tau \in S\Pi$ stationary.

1. $T_\tau Q_{k,\pi} = r + \gamma \mathbb{E} T_\tau V_{k,\pi}$
2. $Q_{k,\pi} = T_{\pi_1} \dots T_{\pi_k} Q_0$.
3. $T_\tau Q_\tau = Q_\tau$.
4. (D) T_τ is γ-contractive on $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ and Q_τ is the unique fixed point of T_τ 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
$$T_\mu Q_{k,\pi} = r + \gamma \int r + \gamma V_{k,\pi} \, \mathrm{d}P \, \mathrm{d}\mu P = r + \gamma \int r + \gamma V_{k,\pi} \, \mathrm{d}P \mu \, \mathrm{d}P = r + \gamma \int T_\mu V_{k,\pi} \, \mathrm{d}P$$
2. Use 1. iteratively starting with $\mu = \pi_1, \pi = (\pi_2, \pi_3, \dots)$.
3. By 2. $T_\pi Q_\pi = T_\pi(r + \gamma \mathbb{E} \lim_{k\to\infty} T_\pi^k V_0) = \lim_{k\to\infty} T_\pi(r + \gamma \mathbb{E} T_\pi^k V_0) = \lim_{k\to\infty}(r + \gamma \mathbb{E} T_\pi^{k+1} V_0) = r + \gamma \mathbb{E} \lim_{k\to\infty} T_\pi^{k+1} V_0 = r + \gamma \mathbb{E} V_\pi = Q_\pi$.
4. The contrativeness of T_π follows from the same argument as for value functions. 2. and Banach fixed point theorem does the rest.

□

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 π is said to be **greedy** with respect to Q and is denoted π_Q.

Proposition 12. For any integrable $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ if π_Q is greedy with respect to Q then $T_{\pi_Q} Q = TQ$.

Proof.

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

□

2.4 Bertsekas-Shreve framework

The theory described here is largely based on the text book *Stochastic Optimal Control: Discrete-time Case* by [1, Bertsekas and Shreve (2007)]. Their framework is cost-based as opposed to the this paper reward-based outset. This means that (P) and (N), upper and lower semicontinuity, suprema and infima, ect. are opposite to the source.

Setting 2 (BS)**.**

1. We consider an MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (see definition 7).
2. $\mathcal{S}$ and $\mathcal{A}$ are Borel spaces.
3. $\mathcal{A}$ is compact.
4. $P(S \mid \cdot)$ is continuous for any $S \in \Sigma_{\mathcal{S}}$.
5. $r(s,a) = \gamma^{1-i} \int x \, \mathrm{d}R(x \mid s,a)$ is upper semicontinuous and uniformly bounded from above (least upper bound denoted $0 < R_{\max} < \infty$).
6. The policies must consist of universally measurable probability kernels.

The original setup in [1] 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* $p : \mathcal{S} \times \mathcal{A} \to W$, instead of a transition kernel which on the other hand is a deterministic *system function* $f : \mathcal{S} \times \mathcal{A} \times W \to \mathcal{S}$ which should be Borel measurable. Moreover it allows for constrains on the action space for each state. This is made precise by a function $U : \mathcal{S} \to \Sigma_{\mathcal{A}}$ and a restriction on $R\Pi$ that all policies π 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, $\kappa : \mathcal{X} \to \mathcal{Y}$ be a continuous probability kernel and $f : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be Borel-measurable satisfying one of $f \leqslant 0, f \geqslant 0, |f| < \infty$. If f is bounded from above (below) and upper (lower) semicontinuous then

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

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

Proof. We refer to [1] prop. 7.31. □

Proposition 14 (Prop. 8.6 in BS)**.** $V_k^* = T^k V_0$ and is upper semicontinuous. Furthermore there exists a sequence of deterministic, stationary, Borel-measurable policies $\tau_1^*, \tau_2^*, \cdots \in DS\Pi$ such that $\pi_k^* = (\tau_k^*, \ldots, \tau_1^*)$ is k-optimal.

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

2.4.1 Analytic setting

For comparison, we include here an similar result in an alternative setting, also considered by [1].

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 $\forall \lambda \in \mathbb{R}, n \geqslant k, s \in \mathcal{S}$

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

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

This setting 3, was actually more widely discussed in [1]. We have put more emphasis on the semicontinuous setting, as it appears restrictive to assume the semianalytical property.

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

Proof. We refer to [1, Bertsekas and Shreve (2007)] prop. 9.17. □

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

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) $Q^* = \lim_{k\to\infty} Q_k^*$.
5. (N) (D) $Q^* = Q_{\pi^*}$.

Proof.

1. Since V_k^* is measurable due to proposition 14 we see that $Q_k^* = \sup_{\pi\in R\Pi}(r + \gamma \mathbb{E} V_{k,\pi}) \leqslant r + \gamma \mathbb{E} V_k^* = r + \gamma \mathbb{E} V_{\pi_k^*} \leqslant Q_k^*$. 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. $TQ_k^* = r + \gamma \mathbb{E} T V_k^*$ 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. $TQ^* = r + \gamma \mathbb{E} T V^*$ and $TQ^* = Q^*$.
3. (D) T is γ-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. $$\begin{aligned} TQ_k^*(s,a) &= T(r + \gamma \mathbb{E} V_k^*)(s,a) \\ &= r(s,a) + \gamma \int \sup_{a' \in \mathcal{A}} (r(s',a') + \gamma \mathbb{E}_{P(\cdot \mid s',a')} V_k^*) \, \mathrm{d}P(s' \mid s,a) \\ &= r(s,a) + \gamma \int \sup_{a' \in \mathcal{A}} \left(r(s',a') + \gamma \int V_k^*(s'') \, \mathrm{d}P(s'' \mid s',a') \right) \, \mathrm{d}P(s' \mid s,a) \\ &= r(s,a) + \gamma \int T V_k^*(s') \, \mathrm{d}P(s' \mid s,a) \end{aligned}$$

   To get $Q_k^* = T^k r$ use this inductively $Q_k^* = r + \gamma \mathbb{E} V_k^* = r + \gamma T V_{k-1}^* = TQ_{k-1}^* = \dots$. 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 $TQ^* = r + \gamma \mathbb{E} T V^* = r + \gamma \mathbb{E} V^* = Q^*$ 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 γ^k. That is

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

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

Proposition 18.

1. Let π_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 $\pi(\cdot \mid s) = \delta_{a_s}$ 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'.

2.5 Theoretical Q-iteration

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

$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : r(s,a) \leftarrow \int x \, \mathrm{d}R(x \mid s,a)$.

$\widetilde{Q}_0 \leftarrow r$

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

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

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

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

Proposition 19. (D)

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

Proof. This is by corollary 1. □

2.5.1 Finite Q-iteration

We have shown how if one knows the dynamics of a stationary decision process satisfying rather broad criteria, such as continuity and compactness, the optimal policy and state-value function can be found simply by iteration over the T-operator and picking a greedy strategy (see proposition 19). Of course this is practical computationally, only if the resulting Q functions can be represented and computed in finite space and time. An obvious situation in which such a representation and computation is possible, is the finite case.

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*

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

then the algorithm becomes

Algorithm 2: Simple finite Q-iteration

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

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

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

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

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

  Update action-value function:

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

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

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

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

Proof. See proposition 19. □

2.6 Approximation

In this section we will look at what happens if we instead use approximations of the Q-functions and T operator. This means that we are in a setting where we can somehow calculate r and TQ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, but it is hard or infeasible to represent them (or at least one of them) directly. This setting is not very well-studied in the case of a continuous state space (at least in the sources known to this writer). This is perhaps because it is considered solved by the results of theoretical Q-learning presented in the previous section. However as we have argued, this only have practical relevance when it is feasible to represent TQ. Therefore we find it relevant to consider this setting in more detail. What *is* very well-studied is a further generalized setting where T and r are assumed to be unknown, that is, one has only access to their distributions via sampling from them. We will deal with this setting in the next section. In following we present some rather simple bounding techniques which is inspired by arguments found in e.g. [5], together with some standard results from approximation theory on artificial neural networks and Bernstein polynomials. Throughout this section we assume (D) i.e. that we are discounting with some $\gamma \in [0,1)$.

Let us consider any norm $\|\cdot\|$ on $(\mathcal{F},\|\cdot\|)$ where $\mathcal{F} \subseteq \mathcal{Q}$ is a subset of the space of bounded Q-functions $\mathcal{Q} = \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$. Let $\widetilde{Q}_0$ be any Q-function which is bounded in $\|\cdot\|$. Suppose we approximate $T\widetilde{Q}_0$ by a Q-function $\widetilde{Q}_1$ to $\varepsilon_1 > 0$ precision and then approximate $T\widetilde{Q}_1$ by $\widetilde{Q}_2$ and so on. This way we get a sequence of Q-functions satisfying

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

First observe that

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

Using this iteratively we get

$$\left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \leqslant \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i := \varepsilon_{\text{approx}}(k)$$

Then we can bound

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

These terms are called respectively the *algorithmic* error and the *approximation* error.

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

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

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

Proposition 21. $\sum_{i-1}^{k} \gamma^{k-i}\varepsilon_i \to 0$ whenever $\varepsilon_k \to 0$ as $k \to \infty$.

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

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

□

2.6.1 Using artifical neural networks

Setting 4. An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0,1]^w$ and $\mathcal{A}$ finite. Assume that r is continuous and P is setwise-continuous.

Definition 12. An **ANN** (Artificial Neural Network) with structure $(d_i)_{i=0}^{L+1} \subseteq \mathbb{N}$, activation functions $\sigma_i = (\sigma_{ij})_{j=1}^{d_i}$, where $\sigma_{ij} : \mathbb{R} \to \mathbb{R}$ are real-valued functions on $\mathbb{R}$, and weights $W_i \in M^{d_i \times d_{i-1}}$, $v_i \in \mathbb{R}^{d_i}$, $i \in [L+1]$ is the function $F : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$

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

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

To clarify we have $\sigma_i(x_1, \ldots, x_{d_i}) = (\sigma_{i1}(x_1), \ldots, \sigma_{id_i}(x_{d_i}))$. $L \in \mathbb{N}_0$ is interpreted as the number of *hidden layers* and d_i is the number of neurons or nodes in layer i.

We denote the class of these networks (or functions)

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

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

Theorem 7 (Universal Approximation Theorem for ANNs)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be non-constant, bounded and continuous activation function. Let $\varepsilon > 0$ and $f \in C([0,1]^w)$. Then there exists an $N \in \mathbb{N}$ and a network $F \in \mathcal{DN}(\sigma, (w, N, 1))$ with one hidden layer and activation function σ such that

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

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

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

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

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

Proposition 22. Consider setting 4 let and $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-constant, bounded, continuous activation function. Let $\varepsilon > 0$. Then for every $k \in \mathbb{N}$ there exists a $N \in \mathbb{N}$ and a sequence of Q-networks $(\widetilde{Q}_i)_{i=1}^k \subseteq \mathcal{DN}(\sigma, \{w|\mathcal{A}|, N, 1\})$ such that

$$\left\| T\widetilde{Q}_{i-1} - \widetilde{Q}_i \right\|_\infty < \varepsilon$$

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

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

This gives us the first method of how to approximate Q^* arbitrarily closely on continuous state spaces, in the case where it is infeasible to represent TQ directly.

2.6.2 Using Bernstein polynomials

We here discuss another approach using multivariate Bernstein polynomials for approximation instead of neural networks. In this case the need a slightly stronger form of continuity, namely Lipschitz continuity, to establish the bounds.

Setting 5. An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0,1]^w$ and $\mathcal{A}$ finite. Assume that there exists a probability measure $\mu \in \mathcal{S}$, such that $P(\cdot \mid s, a)$ has density $p(\cdot \mid s, a) : \mathcal{S} \to \mathbb{R}$ with respect to μ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Furthermore assume that $r(\cdot, a)$, $p(s \mid \cdot, a)$ are Lipschitz with constants L_r, L_p respectively for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Definition 13 (Bernstein polynomial)**.** The multivariate Bernstein polynomial $B_{f,n}$ with exponents $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ approximating the function $f : [0,1]^w \to \mathbb{R}$ is defined by

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

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

Theorem 8. Let $f : [0,1]^w \to \mathbb{R}$ be Lipschitz (see definition 26) w.r.t. the standard euclidean 2-norm induced metrics on $[0,1]^w$ and $\mathbb{R}$ with constant L. Then for any $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ there exists a polynomial $B_{f,n} : [0,1]^w \to \mathbb{R}$ of degree $\leqslant \|n\|_1$ such that

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

Lemma 2. $TQ(\cdot, a)$ is Lipschitz in $\|\cdot\|_2$ with constant $L_T = (L_r + \gamma V_{\max} L_p)$ for all $a \in \mathcal{A}$ and $Q : \mathcal{S} \times \mathcal{A} \to [-V_{\max}, V_{\max}]$.

Now we can bound

Proposition 23.

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

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

Proposition 24.

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

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

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

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

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

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

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

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

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

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

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

There are broadly two ways of dealing with these problems. In the *direct* approaches one tries to first estimate P and R by sampling. Then since P and R are now "known" we can apply the model-dependent methods. The *indirect* approaches broadly cover *the rest* of the cases, and it is mainly these were are going to look at throughout this paper. A popular indirect approach is called *temporal difference* (TD) learning

3.1 Finite case

TD learning is based on the following update scheme

$$\widetilde{Q}_{k+1}(s,a) \leftarrow (1-\alpha_k)\widetilde{Q}_k(s,a) + \alpha_k(r' + \gamma \cdot \max_{a' \in \mathcal{A}} \widetilde{Q}_k(s', a')) \tag{2}$$

Here r' and s' are the reward and next-state drawn from the reward and transition kernels, and $\alpha_k \in [0,1]$ is the so-called **learning rate** (of the kth step). The 'temporal difference' is also the name of term $\alpha_k(r' + \gamma \cdot \max_{a \in \mathcal{A}} \widetilde{Q}_k(s', a') - \widetilde{Q}_k(s,a))$ occuring from rearranging eq. (2). Usually the learning rate is fixed before running the algoritm (does not depend on the history) and is set to decay from 1 to 0 in some fashion as $k \to \infty$.

We will now look at a convergence result originally obtained in [17, Watkins and Dayan (1992)] of a TD algorithm using Q-functions. The result was extended slightly in [6, Jaakkola et al. (1994)] and is here presented more in the style of [6].

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 $(s_1, a_1, \ldots, s_K, a_K)$, learning rates $(\alpha'_1, \ldots, \alpha'_K)$, initial $\widetilde{Q}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

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

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

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

Update action-value function:

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

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

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

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

Theorem 9 (Watkins, Dayan 1992)**.** Let $s_1, a_1, s_2, a_2, \cdots \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A} \times \ldots$ be random variables, and $\alpha_1, \alpha_2, \cdots \in [0,1]$. The output $\widetilde{Q}_K$ of algorithm 4 converges to Q^* provided

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

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

In a special case of the same setup, convergence rates where established by [15, Szepesvári (1997)].

Theorem 10 (Szepesvári)**.** Let $t \in \mathbb{N}$ and $s_1, a_1, s_2, \ldots, s_t, a_t$ be sampled i.i.d. from $p \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$. Set the learning rates such that $\alpha'_k = |\{i \in [k-1] \mid (s_i, a_i) = (s_k, a_k)\}|^{-1}$, i.e. the reciprocal of the frequency of (s_k, a_k) at step k. Let $\beta = \max_{x\in\mathcal{S}\times\mathcal{A}} p(x) / \min_{x\in\mathcal{S}\times\mathcal{A}} p(x)$. Then for some $B > 0$ the following holds asymptotically almost surely

$$\left|\widetilde{Q}_t - Q^*\right| \leqslant B \frac{1}{t^{\beta(1-\gamma)}} \tag{3}$$

and

$$\left|\widetilde{Q}_t - Q^*\right| \leqslant B \sqrt{\frac{\log\log t}{t}} \tag{4}$$

Here eq. (3) is tightest when $\gamma > 1 - \beta/2$ otherwise eq. (4) is tighter.

A paper [9, Majeed and Hutter (2018)] 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 $\widetilde{Q}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

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

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

Update action-value function:

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

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

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

Theorem 11 (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

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

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

If the asynchronos algorithm (algorithm 4) with

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

and the state-action pairs $(s_1, a_1, \dots, s_K, a_K)$ 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 $\left\|\widetilde{Q}_K - Q^*\right\|_\infty < \varepsilon$.

In [9] L is called the *covering rate*.

Remark 5. An interesting side note to theorem 11 is that one can use the bounds to give hints at how to tune the learning rate by changing ω. Optimizing for different scenarios yield different learning theoretically optimal values for ω. For example if we want to optimize for the bound on K for $\gamma \to 1$ using the synchronos algorithm, we get the following rate (treating other variables as constant) $K \geqslant C'(1/(1-\gamma)^{4/\omega} + 1/(1-\gamma)^{1/(1-\omega)})$ for some $C' > 0$. Thus picking $\omega = 4/5$ is optimal. As another example: running the asynchronos algorithm and wanting to minimize for large covering rates L. We get $K \geqslant C''(L^{2+1/\omega} + L^{1/(1-\omega)})$ for some $C'' > 0$. This is optimized for $\omega \approx 0.77$. Then in [9] experiments points to $\omega = 0.85$ as being optimal using a scheme generating random finite MDPs. Other authors have since used this number as a standard value for the learning rate (see e.g. [4, Devraj and Meyn (2017)])

3.1.1 History dependent setting

Setting 6 (Finite HDP)**.**

1. A history dependent decision process (see definition 2), with a single *finite* state space, a single finite action space $(\mathcal{S}, \mathcal{A})$, and transition and reward kernels $(P_n, R_n)_{n\in\mathbb{N}}$. Define $\mathcal{H}^* := \bigcup_{i\in\mathbb{N}} \mathcal{H}_n$, the space of finite histories.

2. $(P_n)_{n\in\mathbb{N}}$ is viewed as a single kernel $P : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{S}$.

3. $(R_n)_{n\in\mathbb{N}}$ is deterministic and viewed as a single function $r : \mathcal{H}^* \to \mathbb{R}$. This is discounted by $\gamma \in [0, 1)$ in accordance to condition (D). That is $r(h_n)$ is bounded in the interval $[-\gamma^{n-2}R_{\max}, \gamma^{n-2}R_{\max}]$ for any $h_n \in \mathcal{H}_n$. Furthermore r depends only on $s_1a_1 \dots s_k r_{k-1}a_k$ when evaluated on $h_{k+1} = s_1a_1 \dots r_{k-1}a_k s_{k+1} \in \mathcal{H}_{k+1}$.

Remark 6. Note that the finite setting 6 is a special case of setting 1 considered by [Schäl, 1974], because Polishness and compactness of $\mathcal{S}, \mathcal{A}$ is readily implied by using the discrete topology in the finite state and action spaces, and the fact that (D) implies pt. 5 in setting 1. Further the conditions (S) and (W) of Schal are also both implied by the discreteness. This implies by theorem 4 the existence of an optimal $\pi^* \in R\Pi$ and that $V_n^* \to V^*$.

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

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

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

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

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

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

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

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

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

Assumption 4 (State-uniformity condition)**.** For any $h, h' \in \mathcal{H}^*$ we have

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

A process as in setting 6 together with the state-uniformity condition is by [9] called a *Q-Value uniform decision process* (QDP).

Theorem 12 (Hutter, 2016)**.** Under assumption 4 we have $q^*_{h'}(\phi(h), a) = Q^*(h, a)$ for any $h' \in \mathcal{H}^*$.

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

$$q_{t+1}(x, a) = q_t(x, a) + \alpha_t(x, a) \left(r' + \gamma \max_{a \in \mathcal{A}} q_t(x', a') - q_t(x, a) \right), \quad x = \phi(h), r' = r(h, a) \tag{5}$$

Theorem 13. Within setting 6 assume

1. State-uniformity (assumption 4).
2. Any state is reached eventually under any policy (called *state-process ergodicity* in [9]).
3. Learning rate satisfies
$$\sum_{t=0}^{\infty} \alpha_t(x, a) = \infty, \quad \sum_{t=0}^{\infty} \alpha_t(x, a)^2 < \infty$$

Then starting with any $q_0 : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ the update step eq. (5) 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. [9] answers this by an array of examples showing the following relations of the classes of decision processes:

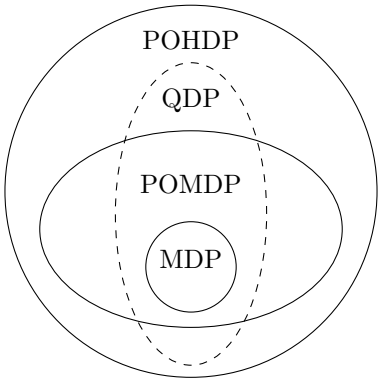


Figure 1: Classes of finite decision processes considered in [9].

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

One point remains unclear after reading [9]: Why is q^* and Q^* well defined by their recursive definition and how are they related to the optimal value function $V^*(s) = \sup_{\pi \in R\Pi} \mathbb{E}^\pi_s \sum_{i=1}^{\infty} \gamma^{i-1} r_i$ (see definition 6) of a general HDP? A sensible thing to ask would be that $Q^*(h, a) = r(h, a) + \gamma \mathbb{E}_{P(\cdot|ha)} V^*$. 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 [10, Melo and Ribeiro (2007)].

Setting 7 (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 7. Item 5 was actually not part of the assumptions in [10]. We include it here in order to ensure the existence of an optimal policy and thus measurability of V^*.

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

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

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

The gradient of Q_θ over θ is

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

This gives the idea for a temporal difference with approximation from $\mathcal{Q}$ 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 π, number of iterations K, learning rates $(\alpha_1, \dots, \alpha_K)$, initial $\theta_1 \in \mathbb{R}^M$

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

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

Update action-value parameter:

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

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

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

In order to understand the results of the analysis of algorithm 6 found in [10], we need to define some concepts from ergodic theory.

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

Definition 14 (Invariant measure)**.** A countably additive measure $\mu \in \mathcal{P}(\mathcal{X})$ is said to be **invariant** w.r.t κ if $\kappa \circ \mu = \mu$.

Definition 15 (Positivity)**.**

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

Definition 16 (Irreducibility)**.** $\mathfrak{P}$ is called φ-irreducible $\mathfrak{P}_x(\tau_A < \infty) > 0$ for all $A \in \Sigma_{\mathcal{X}}$ with $\varphi(A) > 0$ and all $x \in \mathcal{X}$.

Definition 17 (Harris recurrency)**.** $\mathfrak{P}$ is called φ-Harris recurrent if it it φ-irreducible and $\mathfrak{P}_x(\eta_A = \infty) = 1$ for all $A \in \Sigma_{\mathcal{X}}$ with $\varphi(A) > 0$ and all $x \in \mathcal{X}$.

Definition 18 (Geometric ergodicity)**.** A Markov process $\mathfrak{P}$ is called **geometrically ergodic** if it is positive with invariant measure μ, φ-Harris recurrent for some $\varphi \in \mathcal{P}(\mathcal{X})$ and $\exists t > 1$ such that

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

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

Theorem 14 (Melo, Ribeiro)**.** Let $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be an MDP as of setting 7. Let $\pi \in S\Pi$ be a stationary process and $\mathfrak{P}$ the process kernel derived by $P\pi$. Assume that $\mathfrak{P}$ is geometrically ergodic with invariant measure μ and that $\pi(a \mid s) > 0$ for all $a \in \mathcal{A}$ and μ-almost all $s \in \mathcal{S}$. Assume that $\sum_{i=1}^{M} |\xi_i| \leqslant 1$. Then if algorithm 6 is run with learning rates from a sequence $\{\alpha_k\}_{k \in \mathbb{N}}$ satisfying $\alpha_k \in [0, 1]$ and

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

we have that

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

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

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

Furthermore the orthogonal projection is expressible as

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

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

This gives us our first sort of convergence garantee for Q-learning in continuous state space setting. However there is still room for improvement since theorem 14 does not tell us:

1. How fast is the convergence?
2. How far is $Q_{\theta *}$ from Q^*? Though this is obviously best handled seperately for each $\mathcal{Q}$.
3. How far is $Q_{\tilde{\pi}_K}$ from Q^*?

In a quite similar setting these questions are answered for the fitted Q-iteration algorithm in the next section (theorem 15).

4 Deep fitted Q-iteration

4.1 Introduction

4.1.1 Differences in notation

Because σ is used ambigously in theorem 15 we denote the probability distribution σ from [5] p. 20 by ν instead. I avoid the shorthand defined in [5] p. 26 bottom: $\|f\|_n^2 = 1/n \cdot \sum_{i=1}^n f(X_i)^2$. and use p-norms instead. The conversion to the notation used here becomes $\|f\|_n \rightsquigarrow \|f\|/n$. The letter r is used in [5] to denote the euclidean dimension of the state space, while here we use w.

4.1.2 The decision model

Setting 8 (Fan et al.)**.**

1. We're considereing 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 $0 < \gamma < 1$.

4.1.3 ReLU Networks

Definition 19 (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_ℓ, v_ℓ) satisfying

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

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

The idea to work with this particular subclass of neural networks come from [13], which establishes the following lemma

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

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

4.1.4 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 ν, number of iterations K, number of samples n, initial estimator $\widetilde{Q}_0$

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

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

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

Update action-value function:

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

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

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

4.2 Assumptions

4.2.1 Hölder Smoothness

Definition 20 (Hölder smoothness)**.** For $f : \mathcal{S} \to \mathbb{R}$ we define

$$\|f\|_{C_w} := \sum_{|\alpha| < \beta} \|\partial^\alpha f\|_\infty + \sum_{\|\alpha\|_1 = \lfloor \beta \rfloor} \sup_{x \neq y} \frac{|\partial^\alpha (f(x) - f(y))|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \tag{6}$$

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

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

Definition 21. Let $t_j, p_j \in \mathbb{N}$, $t_j \leqslant p_j$ and $H_j, \beta_j > 0$ for $j \in [q]$. We say that f is a **composition of Hölder smooth functions** when

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

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

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

Definition 22. Define

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

and

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

Assumption 5. $T\mathcal{F}_0 \subseteq \mathcal{G}_0$. I.e. t is assumed that $Tf \in \mathcal{G}_0$ for any $f \in \mathcal{F}_0$, so when using the Bellman optimality operator on our sparse ReLU networks, we should stay in the class of compositions of Holder smooth functions.

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

We now look at a simple example where assumption 5 holds: Seting $\mathcal{D} = [0,1]^r$, $q = 1$ and taking both the expected reward function and transition kernel to be Hölder smooth.

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

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

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

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

similarly

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

Thus

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

4.2.2 Concentration coefficients

Definition 23 (Concentration coefficients)**.** Let $\nu_1, \nu_2 \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be probability measures, absolutely continuous w.r.t. $\lambda^w \otimes \mu_{\mathcal{A}}$ (the product of the w-dimensional Lebesgue measure and the counting measure on $\mathcal{A}$). Define

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

Assumption 6. Let ν be the sampling distribution from the algorithm, and μ the distribution over which we measure the error in the main theorem, then we assume

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

4.3 Main theorem

Theorem 15 (Yang, Xie, Wang)**.** Let μ 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

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

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

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

with structure satisfying

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

such that when running algorithm 7 with $\mathcal{F}_0$ and n is sufficiently large

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

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

4.4 Proofs

The proof of theorem 15 combines two results. The first on the error propagation and the second on the error ocurring in a single step.

Theorem 16 (Error Propagation)**.** Let $\{\widetilde{Q}_i\}_{0\leqslant i\leqslant K}$ be the iterates of the fitted Q-iteration algorithm. Then

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

Where

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

Theorem 17 (One-step Approximation Error)**.** Let

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

Then

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

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 16 and theorem 17 are found below, but first we will show how to combine them to obtain theorem 15.

Proof of main theorem. Using theorem 16 we get

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} \leqslant \frac{2\phi_{\mu,\nu}\gamma}{(1-\gamma)^2}\varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2}R_{\max} \tag{7}$$

where $\varepsilon_{\max} = \max_{k\in[K]}\left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\|_{2,\nu}$. Using theorem 17 with $Q = \widetilde{Q}_{k-1}$, $\mathcal{F} = \mathcal{F}_0$, $\epsilon = 1$ and $\delta = 1/n$, we get

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

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

Step 1. We want to employ the following lemma by [13, Schmidt-Hieber (2017)] p. 22. [**?**] 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 $f(\cdot, a) = g_q \circ \cdots \circ g_1$ where the (sub-)components $(g_{jk})_{k=1}^{p_{j+1}} = g_j$ satisfy

$$g_{jk} \in C_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j), \qquad j \in [q], k \in [p_{j+1}] \tag{9}$$

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

$$\begin{aligned} h_1 =& g_1/(2H_1) + 1/2 \\ h_j(u) =& g_j(2H_{j-1}u - H_{j-1})/(2H_j) + 1/2, && j \in \{2, \ldots, q-1\} \\ h_q(u) =& g_q(2H_{q-1}u - H_{q-1}) \end{aligned} \tag{10}$$

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

$$\begin{aligned} h_{1k} &\in C_{t_1}([0,1]^{t_1}, \beta_1, 1) \\ h_{jk} &\in C_{t_j}([0,1]^{t_j}, \beta_j, (2H_{j-1})^{\beta_j}), && j \in \{2, \ldots, q-1\} \\ h_q &\in C_{t_q}([0,1]^{t_q}, \beta_q, H_q(2H_{q-1})^{\beta_q}) \end{aligned} \tag{11}$$

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

$$W := \max\left(\left\{(2H_{j-1})^{\beta_j} \mid 1 \leqslant j \leqslant q-1\right\} \cup \left\{H_q(2H_{q-1})^{\beta_q}, 1\right\}\right) \tag{12}$$

By lemma 3 there exists a ReLU network

$$\widehat{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \widetilde{d}_j p_{j+1}, \ldots, \widetilde{d}_j p_{j+1}, p_{j+1}\right\}, (\widetilde{s}_j + 4)\cdot p_{j+1}\right) \tag{13}$$

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

$$\left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty \leqslant (2W+1)6^{t_j}N2^{-m} + W3^{\beta_j}N^{-\beta_j/t_j} \leqslant 2W3^{\beta_j}N^{-\beta_j/t_j} \tag{14}$$

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

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

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

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

$$\widetilde{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \widetilde{d}_j, \ldots, \widetilde{d}_j, 1\right\}, (\widetilde{s}_j + 4)p_{j+1}\right) \tag{15}$$

and since $h_{jk}([0,1]^{t_j}) \in [0,1]$ by eq. (14)

$$\begin{aligned}
\left\|\widetilde{h}_{jk} - h_{jk}\right\|_\infty &= \left\|\tau \circ \widehat{h}_{jk} - \tau \circ h_{jk}\right\|_\infty && (16)\\
&\leqslant \left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty && (17)\\
&\leqslant 2W 3^{-\beta_j} N^{-\beta_j/t_j} && (18)
\end{aligned}$$

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

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

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

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

and for $\widetilde{s}$

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

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

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

so by induction and eq. (14)

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

and therefore

$$\begin{aligned}
\omega(\mathcal{F}_0) &= \sup_{g\in\mathcal{G}}\inf_{f\in\mathcal{F}} \\
&\leqslant C_Nn^{-2\alpha^*\min_{j\in[q]}\beta_j^*/t_j} \\
&\leqslant C_Nn^{-2\alpha^*\kappa^*} \qquad (19)
\end{aligned}$$

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

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

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

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

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

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

Lemma 5 (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 19. Let $D := \prod_{\ell=1}^{L+1}(d_\ell + 1))$. Then for any $\delta > 0$

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

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

With lemma 5 and n sufficiently large we can bound

$$\begin{aligned}
\log N_0 &= \log\left|\mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty)\right| \\
&\leqslant |\mathcal{A}| \cdot \log\left|\mathcal{N}_{1/n}\right| \\
&\leqslant |\mathcal{A}|\,(\widetilde{s}+1)\log(2(\widetilde{L}+1)\widetilde{D}^2 n) \\
&\leqslant |\mathcal{A}|\,(c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*}+1)2\log\left(2(c_{\widetilde{L}}\log(n)^{\xi^*}+1)\prod_{\ell=1}^{\widetilde{L}+1}(\widetilde{d}+1)\right) \\
&\leqslant 2|\mathcal{A}|\,(c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*}+1)\log\left(2(c_{\widetilde{L}}\log(n)^{\xi^*}+1)(6n^{\alpha^*}\log(n)^{\xi^*}+1)^{\widetilde{L}+1}\right) \\
&\leqslant 4|\mathcal{A}|\,c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*}(\widetilde{L}+1)\log\left(24 c_{\widetilde{L}}\log(n)^{\xi^*} n^{\alpha^*}\log(n)^{\xi^*}\right) \\
&\leqslant 8|\mathcal{A}|\,c_{\widetilde{s}} n^{\alpha^*}\log(n)^{\xi^*} c_{\widetilde{L}}\log(n)^{\xi^*}(\alpha^*+2)\log(n) \\
&= 8 c_{\widetilde{s}} c_{\widetilde{L}}(\alpha^*+2) n^{\alpha^*}\log(n)^{1+2\xi^*}
\end{aligned} \tag{20}$$

Finally, combining eq. (7), eq. (8), eq. (19), eq. (20) and fiddling around with constants one obtains

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

where

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

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

Now for theorem 16.

Lemma 6. $TQ \geqslant T^\pi Q$ for any policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ and any action value function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

Proof.

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

□

Lemma 7. Let $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be an action-value function, $\tau_1, \dots, \tau_m$ be policies and $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be a probability measure. Then

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

For any measure $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ which is absolutely continuous w.r.t. $(P^{\tau_m} \dots P^{\tau_1})(\mu)$. Here κ is the concentration coefficients defined in definition 23.

Proof. Recall that

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

Thus

$$\begin{aligned}\mathbb{E}_\mu[(P^{\tau_m} \dots P^{\tau_1})(f)] &= \int (P^{\tau_m} \dots P^{\tau_1})(f)\, \mathrm{d}\mu && (21)\\ &= \int f\, \mathrm{d}(P^{\tau_m} \dots P^{\tau_1}\mu) && (22)\\ &= \int f \frac{\mathrm{d}(P^{\tau_m} \dots P^{\tau_1}\mu)}{\mathrm{d}\nu}\, \mathrm{d}\nu && (23)\\ &\leqslant \left\| \frac{\mathrm{d}(P^{\tau_m} \dots P^{\tau_1}\mu)}{\mathrm{d}\nu} \right\|_{2,\nu} \cdot \|f\|_{2,\nu} && (24)\\ &\leqslant \kappa(m, \mu, \nu)\|f\|_{2,\nu} && (25)\end{aligned}$$

Where eq. (23) is due to the Radon-Nikodym theorem and eq. (24) is Cauchy-Schwarz. □

Proof of theorem 16. First some things to keep in mind during the proof. Recall that $V_{\max} = R_{\max}/(1-\gamma)$ and that π_Q is the greedy policy w.r.t. Q. Denote

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

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

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

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

The proof consists of four steps.

Step 1 We start by relating $Q^* - Q^{\pi_K}$, the quantity of interest, to $Q^* - \widetilde{Q}_K$, which is more related to the output of the algorithm. Using lemma 6 we can make the upper bound

$$\begin{aligned}Q^* - Q^{\pi_K} &= T^{\pi^*}Q^* - T^{\pi_K}Q^{\pi_K} \\ &= T^{\pi^*}Q^* + (T^{\pi^*}\widetilde{Q}_K - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T\widetilde{Q}_K) - T^{\pi_K}Q^{\pi_K} \\ &= (T^{\pi^*}\widetilde{Q}_K - T\widetilde{Q}_K) + (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K}) \\ &\leqslant (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K}) \\ &= (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T^{\pi_K}\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K}) \\ &= \gamma P^{\pi^*}(Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K}(\widetilde{Q}_K - Q^{\pi_K}) \\ &= \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K}(Q^* - Q^{\pi_K}) \qquad (26)\end{aligned}$$

This implies

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

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

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

From this we also see that $f \geqslant f' \implies Uf \geqslant Uf'$ for any $f, f' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Therefore we can apply U on both sides of eq. (26) to obtain

$$Q^* - Q^{\pi_K} \leqslant \gamma U^{-1}(P^{\pi^*}(Q^* - \widetilde{Q}_K) - P^{\pi_K}(Q^* - \widetilde{Q}_K)) \tag{27}$$

Step 2 Using lemma 6 for any $i \in [K]$ we can get an upper bound

$$\begin{aligned}
Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi^*}\widetilde{Q}_i - T^{\pi^*}\widetilde{Q}_i) \\
&= (Q^* - T^{\pi^*}\widetilde{Q}_i) + (T\widetilde{Q}_i - \widetilde{Q}_{i+1}) + (T^{\pi^*}\widetilde{Q}_i - T\widetilde{Q}_i) \\
&= (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_i) + \varrho_{i+1} + (T^{\pi^*}\widetilde{Q}_i - T\widetilde{Q}_i) \\
&\leqslant T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_i + \varrho_{i+1} \\
&= \gamma P^{\pi^*}(Q^* - \widetilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{28}$$

and a lower bound

$$\begin{aligned}
Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi_i}Q^* - T^{\pi_i}Q^*) \\
&= (T^{\pi_i}Q^* - T^{\pi_i}\widetilde{Q}_i) + \varrho_{i+1} + (TQ^* - T^{\pi_i}Q^*) \\
&\geqslant T^{\pi_i}Q^* - T^{\pi_i}\widetilde{Q}_i + \varrho_{i+1} \\
&= \gamma P^{\pi_i}(Q^* - \widetilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{29}$$

Applying eq. (28) and eq. (29) iteratively we get

$$Q^* - \widetilde{Q}_K \leqslant \gamma^K (P^{\pi^*})^K (Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi^*})^{K-1-i} \varrho_{i+1} \tag{30}$$

and

$$Q^* - \widetilde{Q}_K \geqslant \gamma^K (P^{\pi_{K-1}} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi_{K-1}} \dots P^{\pi_{i+1}}) \varrho_{i+1} \tag{31}$$

Step 3 Combining eq. (30) and eq. (31) with eq. (27) we get

$$\begin{aligned}
Q^* - Q^{\pi_K} \leqslant U^{-1} \Bigg(& \gamma^{K+1}((P^{\pi^*})^{K+1} - P^{\pi_K} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) \\
& + \sum_{i=0}^{K-1} \gamma^{K-i}((P^*)^{K-i} - P^{\pi_K} \dots P^{\pi_{i+1}}) \varrho_{i+1} \Bigg)
\end{aligned} \tag{32}$$

For shorthand define constants

$$\alpha_i = \frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}} \quad \text{for } 0 \leqslant i \leqslant K-1 \text{ and } \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} \tag{33}$$

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

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

$$O_K = (1-\gamma)/2 U^{-1}[(P^{\pi^*})^{K+1} + (P^{\pi_K} \dots P^{\pi_0})] \tag{35}$$

Then by eq. (32)

$$\left|Q^* - Q^{\pi_K}\right| \leqslant \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{i=0}^{K-1}\alpha_i O_i|\varrho_{i+1}| + \alpha_K O_K\left|Q^* - \widetilde{Q}_0\right|\right] \tag{36}$$

So by linearity of expectation

$$\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} = \mathbb{E}_\mu\left|Q^* - Q^{\pi_K}\right| \tag{37}$$

$$\leqslant \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{i=0}^{K-1}\alpha_i \mathbb{E}_\mu(O_i|\varrho_{i+1}|) + \alpha_K \mathbb{E}_\mu(O_K\left|Q^* - \widetilde{Q}_0\right|)\right] \tag{38}$$

With the bound on rewards we (crudely) estimate

$$\mathbb{E}_\mu O_K\left|Q^* - \widetilde{Q}_0\right| \leqslant 2V_{\max} = 2R_{\max}/(1-\gamma) \tag{39}$$

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

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

$$\mathbb{E}_\mu(O_i|\varrho_{i+1}|) \tag{40}$$

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

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

$$= \frac{1-\gamma}{2}\sum_{j=0}^{\infty}\mathbb{E}_\mu\left([(P^{\pi_K})^j(P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1}P^{\pi_{K-1}}\dots P^{\pi_{i+1}}]|\varrho_{i+1}|\right) \tag{43}$$

Notice that there are $K-i+j$ P-operators on both terms in the sum. Therefore were can employ lemma 7 twice. Moreover define $\varepsilon_{\max} = \max_{i\in[K]}\|\varrho_i\|_{2,\nu}$. Then

$$\begin{aligned}\mathbb{E}_\mu(O_i|\varrho_{i+1}|) &\leqslant (1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu)\|\varrho_{i+1}\|_{2,\nu} \\ &\leqslant \varepsilon_{\max}(1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu)\end{aligned} \tag{44}$$

Using eq. (38), eq. (39) and eq. (44)

$$\begin{aligned}\left\|Q^* - Q^{\pi_K}\right\|_{1,\mu} \leqslant \frac{2\gamma(1-\gamma^{K+1})}{1-\gamma}&\left[\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\right]\varepsilon_{\max} \\ &+\frac{4\gamma(1-\gamma^{K+1})}{(1-\gamma)^3}\alpha_K R_{\max}\end{aligned} \tag{45}$$

Focusing on the first term on RHS of eq. (45), if we then we can take the norm out of the sum as

a constant. We are left with

$$\begin{aligned}
&\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\\
&=\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\frac{(1-\gamma)\gamma^{K-i+j-1}}{1-\gamma^{K+1}}\kappa(K-i+j;\mu,\nu)\\
&=\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{j=0}^{\infty}\sum_{i=0}^{K-1}\gamma^{K-i+j-1}\kappa(K-i+j;\mu,\nu)\\
&\leqslant\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{m=0}^{\infty}\gamma^{m-1}\cdot m\cdot\kappa(m;\mu,\nu)\\
&\leqslant\frac{1}{1-\gamma^{K+1}(1-\gamma)}\phi_{\mu,\nu}
\end{aligned}\tag{46}$$

Where the last inequality is due to assumption 6. Combining eq. (45) and eq. (46) we arrive at

$$\left\|Q^*-Q^{\pi_K}\right\|_{1,\mu}\leqslant\frac{2\gamma\cdot\phi_{\mu,\nu}}{(1-\gamma)^2}\cdot\varepsilon_{\max}+\frac{4\gamma^{K+1}}{(1-\gamma)^2}\cdot R_{\max}\tag{47}$$

□

Finally we show theorem 17.

Lemma 8 (Rotation invariance)**.** Let $(X_i)_{i=1}^n$ be independent, centered and sub-gaussian. Then $\sum_{i=1}^n X_i$ is centered and sub-gaussian with

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

Proof. See [Vershynin 2010, p. 12]. □

Definition 24 (Sub-exponential norm)**.** For a random variable define

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

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

Lemma 9 (Sub-gaussian squared is sub-exponential)**.** A random variable X is sub-gaussian if and only if X^2 is sub-exponential and

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

Proof. See [Vershynin 2010, p. 14] □

Proposition 25. Let v be a random vector in $\mathbb{R}^n$ then

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

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

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

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

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

□

Theorem 18 (Bernstein's inequality)**.** Suppose $U_1, \dots, U_n$ are independent with $\mathbb{E}U_i = 0, |U_i| \leqslant M$ for all $i \in [n]$. Then for all $t > 0$

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

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

Proof of theorem 17. First some introductory fixing of notation and variables. Fix a minimal δ-covering of $\mathcal{F}$ with centers $f_1, \dots, f_{N_\delta}$. Define

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

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

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

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

$$\|Y\|^2 + \left\|\widehat{Q}\right\|^2 - 2Y \cdot \widehat{Q} \leqslant \|Y\|^2 + \|f\|^2 - 2Y \cdot f \tag{48}$$

$$\iff \left\|\widehat{Q}\right\|^2 + \|TQ\|^2 - 2\widehat{Q} \cdot TQ \leqslant \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \widehat{Q} - 2Y \cdot f - 2\widehat{Q} \cdot TQ + 2f \cdot TQ \tag{49}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leqslant \|f - TQ\|^2 + 2(Y - TQ) \cdot (\widehat{Q} - f) \tag{50}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leqslant \|f - TQ\|^2 + 2\xi \cdot (\widehat{Q} - f) \tag{51}$$

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

Proposition 26. $\mathbb{E}(\xi_i \mid \Sigma) = 0$ and thus $\mathbb{E}(\xi_i g(X_i)) = 0$ for any function $g : \mathbb{R} \to \mathbb{R}$.

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

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

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

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

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

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

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

By this lemma we can deduce

$$\mathbb{E}\left(\xi \cdot (\widehat{Q} - f) \right) = \mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ) \right) \tag{52}$$

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

$$\left| \mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ) \right) \right| \leqslant \left| \mathbb{E}\left(\xi \cdot (\widehat{Q} - f_{k*}) \right) \right| + \left| \mathbb{E}\left(\xi \cdot (f_{k*} - TQ) \right) \right| \tag{53}$$

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

$$\left| \mathbb{E}\xi \cdot (\widehat{Q} - f_{k*}) \right| \leqslant \mathbb{E}\left(\|\xi\| \left\| \widehat{Q} - f_{k*} \right\| \right) \leqslant \mathbb{E}(\|\xi\|) \sqrt{n} \delta \leqslant 2nV_{\max}\delta \tag{54}$$

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

$$\left\| \widehat{Q} - f_{k*} \right\|^2 = \sum_{i=1}^n (\widehat{Q}(X_i) - f_{k*}(X_i))^2 \leqslant \sum_{i=1}^n \delta^2 = n\delta^2 \tag{55}$$

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

$$\|\xi\|^2 = \sum_{i=1}^n (Y_i - TQ(X_i))^2 \leqslant \sum_{i=1}^n (2V_{\max})^2 = 4V_{\max}^2 n \tag{56}$$

To bound the second term in eq. (53) define

$$Z_j := \xi \cdot (f_j - TQ) \left\| f_j - TQ \right\|^{-1} \tag{57}$$

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

Definition 25 (Sub-gaussian norm)**.**

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

Because of proposition 26 $\xi_i(f_j(X_i) - TQ(X_i))$ is centered for any $i \in [n]$ and

$$\left\| \xi_i(f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma} \leqslant 2V_{\max} \left| f_j(X_i) - TQ(X_i) \right| \tag{58}$$

Therefore by lemma 8

$$\begin{aligned}
\left\| Z_j \right\|_{\psi_2, \Sigma}^2 &\leqslant \left\| f_j - TQ \right\|^{-2} \left\| \sum_{i=1}^n \xi_i(f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma}^2 && (59) \\
&\leqslant \left\| f_j - TQ \right\|^{-2} C_1 \sum_{i=1}^n \left\| \xi_i(f_j(X_i) - TQ(X_i)) \right\|_{\psi_2, \Sigma}^2 && (60) \\
&\leqslant \left\| f_j - TQ \right\|^{-2} C_1 \sum_{i=1}^n 4V_{\max} \left| f_j(X_i) - TQ(X_i) \right|^2 && (61) \\
&= 4V_{\max}^2 C_1 && (62)
\end{aligned}$$

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

$$\mathbb{E}\exp\left(cZ_j^2/\left\|Z_j\right\|_{\psi_2}^2\right) \leqslant e \tag{63}$$

so

$$\begin{aligned}
\mathbb{E}\max_{j\in N_\delta} Z_j^2 &= \frac{\max_{j\in[N_\delta]}\left\|Z_j\right\|_{\psi_2}^2}{c}\mathbb{E}\left(\max_{j\in[N_\delta]}\frac{cZ_j^2}{\max_{k\in[N_\delta]}\|Z_k\|_{\psi_2}}\right) && (64)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\mathbb{E}\left(\max_{j\in N_\delta}\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right) && (65)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(\mathbb{E}\max_{j\in N_\delta}\exp\left(\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right)\right) && (66)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(\sum_{j\in[N_\delta]}\mathbb{E}\exp\left(\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right)\right) && (67)\\
&\leqslant \frac{4V_{\max}^2 C_1}{c}\log\left(eN_\delta\right) && (68)\\
&\leqslant C_2^2 V_{\max}^2\log(N_\delta) && (69)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left(\xi\cdot(f_{k*}-TQ)\right) &= \mathbb{E}\left(\|f_{k*}-TQ\||Z_{k*}|\right) && (70)\\
&\leqslant \mathbb{E}\left(\left(\left\|\widehat{Q}-TQ\right\|+\left\|\widehat{Q}-f_{k*}\right\|\right)|Z_{k*}|\right) && (71)\\
&\leqslant \mathbb{E}\left(\left(\left\|\widehat{Q}-TQ\right\|+n\delta\right)|Z_{k*}|\right) && (72)\\
&\leqslant \left(\mathbb{E}\left(\left\|\widehat{Q}-TQ\right\|+n\delta\right)^2\right)^{1/2}\left(\mathbb{E}Z_{k*}^2\right)^{1/2} && (73)\\
&\leqslant \mathbb{E}\left(\left\|\widehat{Q}-TQ\right\|+n\delta\right)\left(\mathbb{E}Z_{k*}^2\right)^{1/2} && (74)\\
&\leqslant \left(\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|_2^2}+n\delta\right)\left(\mathbb{E}Z_{k*}^2\right)^{1/2} && (75)\\
&\leqslant \left(\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|_2^2}+n\delta\right)C_2^2V_{\max}^2\log(N_\delta) && (76)
\end{aligned}$$

Where eq. (70) to eq. (71) is by the triangle inequality and eq. (74) to eq. (75) is proposition 25. Combining eq. (51), eq. (53), eq. (54) and eq. (76)

$$\begin{aligned}
\mathbb{E}\left\|\widehat{Q}-TQ\right\|^2 &\leqslant \mathbb{E}\|f-TQ\|^2 + 4nV_{\max}\delta + \left(\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|^2}+\sqrt{n}\delta\right)C_2V_{\max}\sqrt{\log(N_\delta)} && (77)\\
&= C_2V_{\max}\sqrt{n\log(N_\delta)}\sqrt{\mathbb{E}\left\|\widehat{Q}-TQ\right\|^2} + nC_2^2\delta V_{\max}^2\log(N_\delta) + \mathbb{E}\|f-TQ\|^2 && (78)
\end{aligned}$$

Lemma 10. Let $a, b > 0, \kappa \in (0,1]$ then

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

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

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

□

By lemma 10 applied to eq. (78)

$$\frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 \leqslant \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa)\left(\delta C_2^2 V_{\max}^2 \log(N_\delta) + \frac{1}{n}\mathbb{E}\|f - TQ\|^2\right) \tag{79}$$

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

$$\begin{aligned} \frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 &\leqslant \frac{(1+\kappa)^2}{\kappa}\delta C_2^2 V_{\max}^2 \log(N_\delta) \\ &\quad + (1+\kappa)\left(C_2^2 V_{\max}^2 \log(N_\delta) + \sup_{g\in\mathcal{G}}\inf_{f\in\mathcal{F}}\|f - g\|_\nu^2\right) \\ &\leqslant \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2 V_{\max}^2 \log(N_\delta) + (1+\kappa)\left(C_2^2 V_{\max}^2 \log(N_\delta) + \omega(\mathcal{F})\right) \end{aligned} \tag{80}$$

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

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

$$\left|\left(\widehat{Q}(x) - TQ(x)\right)^2 - \left(f_{k*}(x) - TQ(x)\right)^2\right| = \left|\widehat{Q}(x) - f_{k*}(x)\right| \cdot \left|\widehat{Q}(x) + f_{k*}(x) - 2TQ(x)\right| \tag{81}$$

$$\leqslant 4V_{\max}\delta \tag{82}$$

Using this twice we can say

$$\begin{aligned} &(\widehat{Q}(\widehat{X}_i) - TQ(\widehat{X}_i)^2 && (83)\\ &\leqslant (\widehat{Q}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 + (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 && (84)\\ &\leqslant (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 + (\widehat{Q}(X_i) - TQ(X_i))^2 - (\widehat{Q}(X_i) - TQ(X_i))^2 \\ &\quad + (f_{k*}(X_i) - TQ(X_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 + 4V_{\max}\delta && (85)\\ &\leqslant (\widehat{Q}(X_i) - TQ(X_i))^2 + (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2 + 8V_{\max}\delta && (86) \end{aligned}$$

Thus we get

$$\left\|\widehat{Q} - TQ\right\|_\sigma^2 \tag{87}$$

$$= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(\widehat{Q}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 \tag{88}$$

$$\leqslant \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\left((\widehat{Q}(X_i) - TQ(X_i))^2 + (f_{k*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k*}(X_i) - TQ(X_i))^2\right) + 8V_{\max}\delta \tag{89}$$

$$= \frac{1}{n}\left\|\widehat{Q} - TQ\right\|^2 + \frac{1}{n}\sum_{i=1}^{n} h_{k*}(X_i, \widetilde{X}_i) + 8V_{\max}\delta \tag{90}$$

Where we define

$$h_j(x, y) := \big(f_j(y) - TQ(y)\big)^2 - \big(f_j(x) - TQ(x)\big)^2 \tag{91}$$

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

$$T := \max_{j\in[N_\delta]}\left|\sum_{i=1}^{n} h_j(X_i, \widetilde{X}_i)/\Upsilon\right| \tag{92}$$

Then we can bound the middle term in eq. (90)

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} h_{k*}(X_i, \widetilde{X}_i)\right) \leqslant \Upsilon/n\mathbb{E}\max_{j\in[N_\delta]}\left(\left|\sum_{i=1}^{n} h_j(X_i, \widetilde{X}_i)/\Upsilon\right|\right) \tag{93}$$

$$\leqslant \Upsilon/n\mathbb{E}T \tag{94}$$

We want to use Bernsteins inequality (theorem 18) with $U_i = h_j(X_i, \widetilde{X}_i)$. Therefore notice that $\left|h_j\right| \leqslant \Upsilon^2$ and

$$\mathrm{Var}h_j(X_i, \widetilde{X}_i) = 2\mathrm{Var}\left(f_j(X_i) - TQ(X_i)\right)^2 \tag{95}$$

$$\leqslant 2\mathbb{E}\left(f_j(X_i) - TQ(X_i)\right)^4 \tag{96}$$

$$\leqslant 2\Upsilon^4 \tag{97}$$

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

$$\mathbb{E}T = \int_0^\infty \mathbb{P}(T \geqslant t) \tag{98}$$

$$\leqslant u + \int_u^\infty \mathbb{P}(T \geqslant t)\,\mathrm{d}t \tag{99}$$

$$\leqslant u + \int_u^\infty 2N_\delta \exp\left(\frac{-t^2}{2\Upsilon t/3 + 4n\Upsilon^2}\right)\,\mathrm{d}t \tag{100}$$

$$\leqslant u + 2N_\delta\int_u^\infty \exp\left(\frac{-t^2}{2\Upsilon^2(t/(3\Upsilon) + 2n)}\right)\,\mathrm{d}t \tag{101}$$

$$\leqslant u + 2N_\delta\left(\int_u^{6n\Upsilon} \exp\left(\frac{-t^2}{8n\Upsilon^2}\right)\,\mathrm{d}t + \int_{6n\Upsilon}^\infty \exp\left(\frac{-t}{4/3\Upsilon}\right)\,\mathrm{d}t\right) \tag{102}$$

$$\leqslant u + 2N_\delta\left(\frac{8n\Upsilon}{2u}\exp\left(\frac{-u^2}{8n\Upsilon}\right) + \frac{4\Upsilon}{3}\exp\left(\frac{-24n\Upsilon}{3\Upsilon}\right)\right) \tag{103}$$

where we use lemma 11 from eq. (102) to eq. (103). Now set $u = \Upsilon\sqrt{8n\log N_\delta}$ continuing from

eq. (103) we have

$$\cdots = \Upsilon\sqrt{8n\log N_\delta} + \frac{\Upsilon^2 8nN_\delta}{\Upsilon\sqrt{8n\log N_\delta}}\exp(-\log N_\delta) + 8/3N_\delta\Upsilon\exp(-9/2n) \tag{104}$$

$$= \Upsilon 2\sqrt{2n}\left(\log N_\delta + \frac{1}{\log N_\delta}\right) + 8/3N_\delta e^{-9/2n} \tag{105}$$

$$\leqslant 4\sqrt{2}\Upsilon\sqrt{n\log N_\delta} + 8/3\Upsilon \tag{106}$$

Inserting eq. (106) and eq. (80) into eq. (90)

$$\left\|\widehat{Q} - TQ\right\|_\nu^2 \leqslant \frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1} + \delta) \tag{107}$$

$$\leqslant \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2V_{\max}^2\log(N_\delta) + (1+\kappa)\left(\delta C_2^2V_{\max}^2\log(N_\delta) + \omega(\mathcal{F})\right) + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1} + \delta) \tag{108}$$

□

4.5 Critique

In [5] it is argued that the statistical aspects of DQN is preserved even though experience replay is replaced by i.i.d. sampling, however the arguments are few and vague.

5 Conclusion

In this paper we have build up the theory behind Q-learning, covering decision models, optimality of policies, value functions and their iteration methods. This gave an introduction to Q-learning and a general framework from which to understand and compare results within the field. We then turned to model-free algorithms and presented convergence results for such in a variety of settings with state space being both finite and infinite and dynamics being allowed to depend on history or not. Finally we presented and proved convergence of the fitted Q-iteration algorithm as obtained in [5]. All together this paints a picture of what Q-learning is, how it was developed, which topics it is related to, what its challenges are and what it is possible to say theoretically about its convergence to optimaliy at present. Theoretically you could say that Q-learning is solved in many situations, since, as we have established, there is convergence guarrantees for broad classes of problems. However as to how these convergence results relate to practical aspects of Q-learning we can still say little and as to the succes of the DQN of [11] we are not much further in understanding. The major reason is that the computational aspects are so important to their succes, and this part is mostly ignored in the results we have covered. Even though we establish results of the related FQI algorithm in [5], it is unclear if it captures the critical aspects of DQN, such as experience replay. In [5] convergence of FQI is guaranteed given corresponding increases in iterations, batch size and function space complexity. It is hard to interpret exactly how large these increases must be or whether it is practical.

5.1 Further directions

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

5.1.1 Suboptimality of policies

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

5.1.2 Bernstein polynomials vs. orthogonal projection

A Bernstein polynomial B_f approximating a function f are constructed by evaluating the functions at a finite number of points (see definition 13). Since we in this setting are concerned with approximation in the 2-norm, another approach would be to simply take the orthogonal projection of TQ onto the span of polynomials of degree less than n. One should keep in mind that this requires integration of $|TQ(\cdot, a)f_i|$ for every basis polynomial f_i, which is potentially hard to compute. On the other hand, as the orthogonal projection is distance minimizing, it should provide the best approximation with polynomials. The relation between the performances of the Berstein polynomial and the orthogonal projection, both in terms of accuracy and computational complexity, could be interesting analyse.

6 Appendices

6.1 Lemmas for Fan et al.

Lemma 11. For $x > 0$.

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

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

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

□

6.2 Other notes

Definition 26 (Lipschitz continuity)**.** Let $(\mathcal{X}, d_\mathcal{X})$, $(\mathcal{Y}, d_\mathcal{Y})$ be metric spaces. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to **Lipschitz** with constant $L > 0$ if

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

Definition 27 (Almost sure uniform convergence of random processes)**.** A sequence of random processes $X_n : \mathcal{X} \times \Omega \to \mathbb{R}$ is said to converge **almost surely uniformly** to $X : \mathcal{X} \times \Omega \to \mathbb{R}$ if and only if

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

Definition 28 (Uniform convergence in probability of random processes)**.** A sequence of random processes $X_n : \mathcal{X} \times \Omega \to \mathbb{R}$ is said to converge **uniformly in probability** to $X : \mathcal{X} \times \Omega \to \mathbb{R}$ if and only if

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

Proposition 27. $id_{\mathcal{P}(X)} = \mu \mapsto \kappa \circ \mu$ where $\kappa(\cdot \mid x) = \delta_x(\cdot)$. Thus κ can be seen as an identity mapping on $\mathcal{P}(X)$.

Proof.

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

□

Theorem 19 (Banach fixed point theorem)**.** Let $(\mathcal{X}, d)$ be a complete metric space and $T : \mathcal{X} \to \mathcal{X}$ be a contraction, i.e. $d(Tx, Ty) < \gamma d(x, y)$ for some $0 < \gamma < 1$ and all $x, y \in \mathcal{X}$. Then T has a unique fixed point x^* and for every $x \in \mathcal{X}$ it holds that $T^k x \to x^*$ as $k \to \infty$, with rate $d(T^k x, x^*) < \gamma^k d(x, x^*)$.

Definition 29 (Dynkin class)**.** Let D be a pavement of X, that is a collection of subsets of X. D is called a **Dynkin class** if

1. $X \in D$,

2. If $A, B \in D$ and $A \subseteq B$ then $B \backslash A \in D$,

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

Theorem 20 (Dynkins π-λ theorem)**.** Let P be a pavement of X which is stable under finite intersections (such are called π-systems) and D a Dynkin class (see definition 29). If $P \subseteq D$ then $\sigma(P) \subseteq D$ where $\sigma(P)$ is the smallest σ-algebra containing P.

6.3 Disambiguation

- $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, $\overline{\underline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$.
- $\mathrm{id}_X := x \mapsto x$ the identity function on X.
- $[\phi] := \begin{cases} 1 & \phi \\ 0 & \neg\phi \end{cases}$: 0-1 indicator for logical formulas.
- $[q] := \{1, \ldots, q\}$ for $q \in \mathbb{N}$.
- $1_A(a) := [a \in A]$: the indicator function.
- $C_{\mathbb{K}}(X) := \{f : X \to \mathbb{K} \mid f \text{ continuous}\}$, for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. $C(X) = C_{\mathbb{R}}(X)$
- $C_b(X) := \{f \in C(X) \mid \exists K : \|f\|_\infty < K\}$. I.e. the space of bounded continuous functions.
- ANN: abbreviation for artificial neural network see definition 12.
- δ_a: Dirac-measure of point a, i.e. $\delta_a(A) = [a \in A] = 1_A(a)$.
- $(\Omega, \mathcal{F}, \mathbb{P})$: background probability space, that is source space for random variables.
- $\mathbb{B}_n$: the n-dimensional Borel σ-algebra.
- λ^n: the n-dimension Lebesgue measure.
- $\mathcal{P}(A)$: the set of all probability measures on a measurable space (A, Σ_A).
- 2^A: the powerset of A.
- $\mathcal{M}(A, B)$: set of Σ_A-Σ_B measurable functions.
- $\mathcal{L}_p$: set of functions f with $\|f\|_p < \infty$ $p \in [1, \infty]$.
- $\mathbb{E}, \mathbb{E}_\mu$: expectation, that is integration w.r.t. the measure $\mathbb{P}$ or μ respectively.
- Var: variance operator.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

[14] Manfred Schäl. On dynamic programming: Compactness of the space of policies. *Stochastic Processes and their Applications*, 3(4):345 – 364, 1975.

ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(75)90031-9. URL http://www.sciencedirect.com/science/article/pii/0304414975900319.

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

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

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