On Q-iteration with uncountable state spaces

Jacob Harder
University of Copenhagen

July 11, 2020

Abstract

In this paper we present

1. a framework for studying Q-learning for decision processes with in the generality of non-Markov dynamics and continuous state and action spaces
2. sufficient criteria for existence of optimal policies in general (possibly non-Markov and with continuous state and action spaces) decision processes based on [13, Schäl (1975)] and [1, Barto et al. (1983)]
3. relations between value-iteration and Q-iteration and their convergence properties in the setting of Markov decision processes with continuous state and action space
4. bounds on deviations from optimality of Q-iteration when using function approximators focusing in particular on two function classes:
   (a) Artifical neural networks
   (b) Bernstein polynomials

Contents

1 Introduction

1.1 Reinforcement Learning

RL is a broad topic and a main branch of machine learning. Because of its broadness it overlaps with other disciplines such as control theory and dynamic programming.

In Reinforcement Learning, as in dynamic programming, we are concerned with finding an optimal policy for an agent in some environment. This environment is described by a sequence of state and action spaces $\mathcal{S}_1, \mathcal{A}_1, \mathcal{S}_2, \ldots$ and rules (or dynamics) formalized as probability kernels $P_1, R_1, P_2, \ldots$ specifying which states and rewards are likely to follow after some action is chosen. One can then specify rules π, called a *policy*, for how the agent should choose actions in every situation in the environment. Given an environment and a policy one obtains a stochastic process, that is a distribution on sequences of states, actions and rewards. One can then measure the performance of the policy by looking at the expected accumulated rewards called the *policy evaluations* V_π of the policy. The goal of reinforcement learning is to find an optimal policy $\pi*$, maximizing the value function.

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

A particular class of environments which are called Markov decision processes (MDPs). In an MDP the same state space $\mathcal{S}$, action space $\mathcal{A}$ and rules P, R are used throughout the process. They are by far the most well-studied environments. With an MDP and a value function V_1 satisfying certain assumptions one can obtain a policy π_1 by choosing actions leading to the maximum average values (according to V_1). Such policies are called *greedy policies.* We can then evaluate the policy π_1 yielding a new value function $V_2 = V_{\pi_1}$. The process of evaluating policies and picking greedy policies is formalised by so called *T-operators* T_π, T. One of these (T) is called the *Bellman optimality operator* and combines policy evaluation and greedy choices. This process of applying the T operators and picking greedy policies can be continued indefinitely yielding a sequence of value functions and policies. Variations of this idea are called *value iteration* and *policy iteration*, and is derived from dynamic programming. We show that value iteration converges to the optimal value functions given mild assumptions on the MDP. Furthermore we show that the optimal value functions is a fixed point of the Bellman optimality operator: $TV^* = V^*$ This is called the *Bellman optimality equation* and is central to all problems in dynamic programming.

We have now described the roots of RL in dynamic programming. However RL usually refers to algorithms that are not merely value iterations, but instead work without directly using the transition and reward dynamics, and instead estimate value functions based only on sampling from the environment. Such algorithms are called *model-free.* We will not look at algorithms which are based on sampling, and instead focus on theoretical aspects assuming it is possible to use the transition dynamics directly.

1.2 Q-learning

A problem with value functions defined on the set of states $\mathcal{S}$ is that picking optimal actions require knowledge of the transition dynamics P. This is especially a problem for model-free algorithms. To get around this problem *Q-functions* were introduced, which evaluates the value of a state-action pair, instead of only a state.

Given a Q-function Q, picking best actions according to Q now merely require maximization over Q itself. Also it turns out that Q-functions is more convenient to work with computationally. In this 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 *indirect* approaches one attempts to estimate the process dynamics first and then afterwards methods for the known-dynamics are applied. The *direct* approaches basically covers *the rest.* In the direct category we find the popular *temporal difference* algorithms on which *fitted Q-iteration* (FQI) and the *deep Q-network* (DQN) algorithm of [11] is based. Many direct approaching such as FQI and DQN can be seen as stochastic approximations of the Bellman optimality equation.

Q-learning is the category of algorithms that iteratively updates Q-functions in the attempt to improve the derived policy. *Deep* Q-learning is then the subcategory of algorithms which uses deep neural networks as approximators for the Q-functions. We will see in this paper how Q-functions are used to find optimal policies (strategies) for decision processes and how they work as the underlying *knowledge* that drives the decisions of the agent. We will use a wide array of function classes in the attempt to approximate ideal Q-functions such as the policy evaluations and optimal Q-functions.

All this will be made precise in the next section. Before proceeding to this we include a brief introduction to the basic concept and notation we are going to use throughout the paper.

1.3 Basic concepts and notation

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

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

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

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

2 Decision models and value functions

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

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

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

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

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

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

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

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

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

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

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

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

A handy property of kernels is

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

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

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

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

We can now state the definition of a decision process

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

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

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

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

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

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

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

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

4. $(R_n)_{n\in\mathbb{N}}$ a sequence of $\mathcal{H}_{n+1} \rightsquigarrow \mathbb{R}$ probability kernels called the **reward** kernels.

5. $\mathfrak{A}_n(h_n) \subseteq \mathcal{A}_n$ a set of admissable actions for each $h_n \in \mathcal{H}_n$ and $n \in \mathbb{N}$.

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

Definition 2.6 (Policy)**.** A (randomized) **policy** $\pi = (\pi_n)_{n\in\mathbb{N}}$ for a HDP is a sequence of probability kernels $\pi_n : \mathcal{H}_n \rightsquigarrow \mathcal{A}_n$, such that $\pi_n(\mathfrak{A}(h_i) \mid h_i) = 1$ for alle $h_i \in \mathcal{H}_i$, i.e. the policy chooses only admissable actions (with probability 1). The set of all policies we denote $R\Pi$.

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

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

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

Theorem 2.7 (Integration of a kernel)**.** Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$. Then there exists a uniquely determined probability measure $\lambda \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that for all $A \in \Sigma_{\mathcal{X}},\ B \in \Sigma_{\mathcal{Y}}$

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

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

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

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

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

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

Uniqueness follows because the property

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

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

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

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

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

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

Furthermore if this is finite, i.e. $f \in \mathcal{L}_1(\kappa(\cdot, \mu))$ then $A_0 := \left\{ x \in \mathcal{X} \mid \int f \; \mathrm{d}\kappa(\cdot \mid x) < \infty \right\} \in \Sigma_{\mathcal{X}}$ 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 2.10 (Composition of kernels)**.** Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\phi : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ be probability kernels. Then there exists a unique probability kernel $\phi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ satisfying

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

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

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

Proof. See section A.2. □

Remark 2.11. A kernel $\varphi : \mathcal{Y} \rightsquigarrow \mathcal{Z}$ can also be considered a kernel $\varphi' : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ by defining $\varphi'(\cdot \mid x, y) = \kappa(\cdot \mid y)$ that is ignoring the first input. Therefore we can use proposition 2.10 on φ' to get $\varphi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$. Projection $\rho_{\mathcal{Z}}$ onto $\mathcal{Z}$ also preserves the kernel property, so $\rho_{\mathcal{Z}}(\varphi'\kappa) : \mathcal{X} \rightsquigarrow \mathcal{Z}$ is yet another probability kernel. We denote this $\varphi \circ \kappa = \rho_{\mathcal{Z}}(\varphi'\kappa)$ and this is also called *composition of kernels* by some authors. We can extend this to the kernel-measure composition (of theorem 2.7): If $\mu \in \mathcal{P}(\mathcal{X})$ is a probability measure on $\mathcal{X}$ then $\rho_{\mathcal{Y}}(\kappa\mu) \in \mathcal{P}(\mathcal{Y})$ and we denote this $\kappa \circ \mu = \rho_{\mathcal{Y}}(\kappa\mu)$. In fact $\circ$ makes the class of measurable spaces into a category (see [9, Lawvere (1962)]), with identity $\mathrm{id}_{\mathcal{X}}(\cdot \mid x) = \delta_x$ (for a proof of this last statement see proposition A.16).

Example 2.12. To get feel for how kernels combine, let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$, $\phi : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ and $\psi : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \rightsquigarrow \mathcal{W}$ be probability kernels. Then $\psi\phi(\cdot, \cdot \mid \cdot, \cdot)$ is a $\mathcal{X} \times \mathcal{Y}$-kernel on $\mathcal{Z} \times \mathcal{W}$ and so can be combined with κ to get $(\psi\phi)\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y} \times (\mathcal{Z} \times \mathcal{W})$. On the other hand $\phi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ can be combined with ψ to get $\psi(\phi\kappa) : \mathcal{X} \rightsquigarrow (\mathcal{Y} \times \mathcal{Z}) \times \mathcal{W}$. By the associativity result in proposition 2.10 (and of the cartesian product) these kernels are the same.

2.0.1 From kernels to processes

Throughout this section let $(\mathcal{X}_n)_{n\in\mathbb{N}}$ be a sequence of measurable spaces. For each $n \in \mathbb{N}$ let $\mathcal{X}^{\underline{n}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and $\kappa_n : \mathcal{X}^{\underline{n}} \rightsquigarrow \mathcal{X}_{n+1}$ be a probability kernel.

Proposition 2.13. For all $n \in \mathbb{N}$ the composition $\kappa^{\underline{n}} := \kappa_n \dots \kappa_1$ yields a $\mathcal{X}_1$-probability kernel on $\mathcal{X}_2 \times \cdots \times \mathcal{X}_{n+1}$.

Proof. This is by induction using proposition 2.10. □

Proposition 2.13 allows us to make sense to finite decision processes. That is for any $n \in \mathbb{N}$, distribution $\mu \in \mathcal{P}(\mathcal{S}_1)$ of S_1 and policy $(\pi_1, \pi_2, \dots) \in R\Pi$ we can get a distribution of the nth history $H_n \in \mathcal{H}_n$ by the composition of kernels

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

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

Theorem 2.14 (Ionescu-Tulcea extension theorem)**.** For every $\mu \in \mathcal{P}(\mathcal{X}_1)$ there exists a unique probability measure $\rho \in \mathcal{P}(\mathcal{X}^{\underline{\infty}})$ such that

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

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

In particular theorem 2.14 applied to the measure Dirac-measure δ_x can be interpreted as starting the process in $x \in \mathcal{X}_1$. Later we would like to consider function defined

Proposition 2.15 (Ionescu-Tulcea kernel)**.** Let μ_x denote the Ionescu-Tulcea measure of a sequence of probability kernels $(\kappa_i)_{i\in\mathbb{N}}$ where $\kappa_i : \mathcal{X}^{\underline{i}} \to \mathcal{X}_{i+1}$ with starting measure δ_x on $\mathcal{X}_1$ for any $x \in \mathcal{X}_1$. Then $\kappa(A \mid x) = \mu_x(\mathcal{X}_1 \times A)$ defines a probability kernel $\kappa : \mathcal{X}_1 \rightsquigarrow \mathcal{X}_2 \times \mathcal{X}_3 \times \dots$.

Proof. Since we already know by theorem 2.14 that μ_x is a probability measure for any $x \in \mathcal{X}_1$, we just have to show that $\kappa(A \mid x) = \mu_x(A)$ is measurable as a function of x for all $A \in \Sigma$ where $\Sigma = \bigotimes_{i=2}^{\infty} \Sigma_{\mathcal{X}_i}$. Let $\phi_A = x \mapsto \mu_x(A)$ for all $A \in \Sigma$ and define

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

The cylinder sets

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

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

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

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

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

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

For showing that $\mathbb{G}$ is a Dynkin class, notice that clearly $\mathcal{X}_2 \times \mathcal{X}_3 \times \dots$ and $\varnothing$ are in $\mathbb{G}$. If A, $B \in \mathbb{G}$ with $A \subseteq B$ then $\phi_{B\setminus A} = \phi_B - \phi_A \in \mathbb{G}$. Finally if $(B_n)_{n\in\mathbb{N}}$ is an ($\subseteq$-) increasing sequence in $\mathbb{G}$ then $\phi_{\bigcup_{n=1}^{\infty} B_n} = \lim_{n\to\infty} \phi_{B_n}$ is again measurable as it is a limit of measurable functions, showing that $\mathbb{G}$ is a Dynkin class. □

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

Lemma 2.16. The Ionescu-Tulcea kernel satisfies $\prod_{i=1}^{\infty}\kappa_i = (\prod_{i=2}^{\infty}\kappa_i)\kappa_1$.

Proof. Let $x \in \mathcal{X}_1$. Since for the composition of finitely many kernels by associativity (proposition 2.10) it holds that $\kappa_n \dots \kappa_1 = (\kappa_n \dots \kappa_2)\kappa_1$. Therefore for any $n \in \mathbb{N}$ and $A \in \Sigma_{\mathcal{X}_2} \otimes \cdots \otimes \Sigma_{\mathcal{X}_n}$, using theorem 2.14 we have

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

for all $n \in \mathbb{N}$ and $A \in \Sigma_{\mathcal{X}^{\underline{n}}}$. By the uniqueness in theorem 2.14 we are done. □

Let $\mu \in \mathcal{P}(\mathcal{S}_1)$ be a measure on the first state space. By theorem 2.14 a HDP and a policy $\pi = (\pi_1, \pi_2, \dots) \in R\Pi$ gives rise to a kernel $\kappa_\pi : \mathcal{P}(\mathcal{S}_1) \to \mathcal{P}(\mathcal{H}_\infty)$, namely

$$\kappa_\pi\mu = \dots P_2\pi_2P_1\pi_1\mu \in \mathcal{P}(\mathcal{H}_\infty) \tag{1}$$

The measure $\kappa_\pi\mu$ can be interpreted as the stochastic process arising from sampling the first state from μ and then follow π for a countable number of steps.

Remark 2.17. We will denote expectation with respect to $\kappa_\pi\mu$ by $\mathbb{E}_\mu^\pi$. In the case where $\mu = \delta_s$ for some $s \in \mathcal{S}_1$ integration with respect to the measure $\kappa_\pi\delta_s$ and $\kappa_\pi(\cdot \mid s)$ is equivalent in the sense that

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

for some integrable $f : \mathcal{H}_\infty \to \mathbb{R}$. Both of these measures can be interpreted as the stochastic process arising from starting in state s and following policy π. We will sometimes abuse notation slightly, writing $\kappa_\pi\delta_s = \kappa_\pi s$ and $\mathbb{E}_{\delta_s}^\pi = \mathbb{E}_s^\pi$. If a measurable function f depends on only parts of the history, say $f : \mathcal{S}_3 \times \mathcal{A}_3 \to \mathbb{R}$ we will also assume that it is applied with *right* inputs when taking expectations, that is we will write

$$\mathbb{E}_\mu^\pi f = \mathbb{E}_\mu^\pi f \circ \rho_{\mathcal{S}_3 \times \mathcal{A}_3} = \int f\, \mathrm{d}\kappa_\pi\mu = \int f \circ \rho_{\mathcal{S}_3 \times \mathcal{A}_3}\, \mathrm{d}\kappa_\pi\mu = \int f(s_3, a_3)\, \mathrm{d}\kappa_\pi\mu(s_1, a_1, s_2, a_2, \dots)$$

where $\rho_{\mathcal{S}_3 \times \mathcal{A}_3} : \mathcal{H}_\infty \to \mathcal{S}_3 \times \mathcal{A}_3$ denotes projection onto $\mathcal{S}_3 \times \mathcal{A}_3$.

3 Policy evaluation and value functions

The next step is to evaluate how *good* a policy is. This is where the reward kernels $R_1, R_2, \dots$ come into play. However their stochastic properties will not be relevant for this section. For now we will only care about their expected values. Therefore we will need

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

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

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

Remark 3.1. Assumption 1 implies that all $R_n(\cdot \mid h_n)$ has moments of all orders for any $h_n \in \mathcal{H}_n$.

Definition 3.2. We define for each $n \in \mathbb{N}$

$$r_n : \mathcal{H}_{n+1} \to \mathbb{R}, \quad r(h_{n+1}) = \int x \, \mathrm{d}R_n(x \mid h_{n+1})$$

called the nth **expected reward function**.

Remark 3.3. The expected reward function r_n is measurable due to proposition 2.4. If X is a random variable with distribution $X \sim R_n(\cdot \mid h_{n+1})$ (or stated differently $X(\mathbb{P}) = R_n(\cdot \mid h_{n+1})$) then we have $\mathbb{E}X = \mathbb{E}_{\mathbb{P}}X = r_n(h_{n+1})$.

We are now ready to talk about *value functions.* A value function is any function $V : \mathcal{S}_1 \to \mathbb{R}$ which assigns a real number to a starting state (a state in $\mathcal{S}_1$). Usually the purpose of a value function is to give an estimate $V(s_1)$ for each state $s_1 \in \mathcal{S}_1$ of the accumulated rewards of a decision process starting at s_1. We will now define an important class of value functions. In these definitions, recall the notation of integration with respect to kernel-derived processes discussed in remark 2.17.

Definition 3.4 (Finite policy evaluation)**.** We define the function $V_{n,\pi} : \mathcal{S}_1 \to \mathbb{R}$ by

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

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

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

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

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

This assumption allows the following definition of the following value function

Definition 3.5. We define the (infinite) **policy evaluation** by

$$V_\pi(s) = \mathbb{E}^{\pi}_{s} \lim_{n\to\infty} \sum_{i=1}^{n} r_i = \int \sum_{i=1}^{\infty} r_i \, \mathrm{d}\kappa_\pi s \tag{2}$$

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

Proposition 3.6. Let $\pi \in R\Pi$ be a policy. Under assumption 2 (discounting) the following holds

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

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

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

Proof.

1. By remark 3.3 the expected reward functions are measurable (see proposition 2.4). Therefore sums and limits of them are aswell. Integrability follows once we show item 3.

2. Follows by dominated convergence once we show item 3.

3. Let $\pi \in R\Pi$ be any policy. Because of assumption 2 we have

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

This also covers $V_{n,\pi}$.

□

Remark 3.7. As this bound will occur again and again we denote it

$$V_{\max} := R_{\max}/(1-\gamma) \tag{3}$$

3.1 The optimal value function

Definition 3.8 (Optimal value functions).

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

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

Remark 3.9. Under assumption 2 we have $\left|V_k^*\right|, |V^*| \leqslant V_{\max}$ since by proposition 3.6 all terms in the supremum are within this bound.

Remark 3.10. It is known that the optimal value function might not be Borel measurable (see ex. 2 p. 233 [2]). Perhaps this is not suprising since we are taking a supremum over sets of policies and these sets have cardinality is at least the continuum, even if both all states and actions spaces are finite.

At this point some relevant questions can be asked.

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

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

3. Can optimal policies be chosen to be deterministic?

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

We will wait with questions 3 and 4 until the next section. In a quite general setting, questions 1 and 2 is investigated in [13, Schäl (1975)]. Here some additional structure on our process is imposed.

Definition 3.11 (Standard Borel measurable space). A measurable space $(\mathcal{X}, \Sigma_\mathcal{X})$ is called **standard Borel** if $\mathcal{X}$ is Polish space, that is a seperable completely metrizable space, and $\Sigma_\mathcal{X}$ is the Borel σ-algebra of $\mathcal{X}$, that is the σ-algebra generated by all open sets.

Setting 1 (Schäl).

1. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})$ is standard Borel for each $n \in \mathbb{N}$.

2. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})$ is standard Borel. for each $n \in \mathbb{N}$.

3. The set of admissible actions $\mathfrak{A}_n(h_n)$ is compact for any $h_n \in \mathcal{H}_n,\ n \in \mathbb{N}$.

4. $|V_\pi| < \infty$ for all policies $\pi \in R\Pi$.

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

Remark 3.12. The last two points in setting 1 is readily implied by assumption 2 because of proposition 3.6 and the fact that under assumption 2 we have $\left|\sum_{t=n+1}^{N} r_t\right| \leqslant \gamma^n V_{\max}$.

To understand the results of [13] which we will present shortly, we will need some additional concepts, namely the weak topology and semicontinuity.

Definition 3.13 (Weak topology)**.** Let $\mathcal{X}$ be a metrizable space equipped with the Borel σ-algebra. Consider the family of sets of probability measures on $\mathcal{X}$, that is a family of subsets of the space of $\mathcal{X}$-probability measures $\mathcal{P}(\mathcal{X})$

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

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

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

Proposition 3.14 (Properties of the weak topology)**.** *Let $\mathcal{X}$ be a seperable metric space with metric d. Denote by $C(\mathcal{X})$ the set of continuous real-valued functions on $\mathcal{X}$ and by $U_d(\mathcal{X})$ the set of uniformly continuous real-valued functions on $\mathcal{X}$. Let $p, p_1, p_2, \dots \in \mathcal{P}(\mathcal{X})$. Then the following is equivalent:*

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

Proof. We refer to [2] prop.7.21. □

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

Proposition 3.15. *Let $\mathcal{X}$ be a standard Borel measurable space. Then the space $\mathcal{P}(\mathcal{X})$ of probabilty measures on $\mathcal{X}$ equipped with the weak topology is standard Borel. If furthermore $\mathcal{X}$ is compact then $\mathcal{P}(\mathcal{X})$ is also compact.*

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

Definition 3.16 (Semicontinuity)**.** Let $\mathcal{X}$ be a topological space and $f : \mathcal{X} \to \overline{\mathbb{R}}$ be a extended real-valued function. Then f is **upper** semicontinuous at $x_0 \in \mathcal{X}$ if for every $y > f(x_0)$ there exists a neighborhood U of x_0 such that $f(x) < y$ for all $x \in U$. We define f to by **lower** semicontinuous if $-f$ is upper semicontinuous.

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

Proposition 3.17.

1. If $\mathcal{X}$ is a metrizable space then $f : \mathcal{X} \to \overline{\mathbb{R}}$ is upper semicontinuous if and only if for each sequence $x_1, x_2, \dots \in \mathcal{X}$ with $x_n \to x \in \mathcal{X}$ we have $\limsup f(x_n) \leqslant f(x)$ (analogously $\liminf f(x_n) \geqslant f(x)$ for lower semicontinuous f).
2. If $f,\ g : \mathcal{X} \to \overline{\mathbb{R}}$ are upper (lower) semicontinuous then $f + g$ is upper (lower) semicontinuous.
3. If furthermore g is continuous and non-negative then $f \cdot g$ is upper (lower) semicontinuous.
4. If $(f_i)_{i \in I}$ are an arbitrary collection of upper (lower) semicontinuous functions then the infimum $\inf_{i \in I} f_i$ (supremum $\sup_{i \in I} f_i$) is again upper (lower) semicontinuous.

Proof. We refer to [2] p. 147. □

We are now ready to present the results of [13]. Under setting 1 Schäl introduced two sets of criteria for the existence of an optimal policy:

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

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

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

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

Remark 3.18. In (S) set-wise continuity is to be understood as the requirement that $\psi_{S,(s_i)} : \mathcal{A}^{\underline{n}} \to \mathbb{R} = (a_1, a_2, \dots, a_n) \mapsto P_n(S \mid s_1, a_1, \dots, s_n, a_n)$ is continuous for any $S \in \Sigma_{\mathcal{S}_{n+1}}$ and $s_i \in \mathcal{S}_i,\ i \in [n]$. In (W) weak continuity means that the function $\phi : \mathcal{H}_n \times \mathcal{A}_n \to \mathcal{P}(\mathcal{S}_{n+1}) = (h_n, a_n) \mapsto P_n(\cdot \mid h_n, a_n)$ is continuous when $P_n(\cdot \mid h_n, a_n)$ is endowed with the weak topology.

Theorem 3.19 (Schäl)**.** Under setting 1 when either (S) or (W) hold then

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

Proof. We refer to [13]. □

Corollary 3.20. Under setting 1 when either (S) or (W) hold then V^* is (Borel) measurable.

Proof. Since by theorem 3.19 there exists an optimal policy π^* we have $V^* = V_{\pi^*}$ which is measurable due to proposition 3.6. □

Schäls theorem tells us that optimal policies exist and that the n-stage optimal value functions converge to the infinite optimal value function in a wide class of decision processes.

In many cases we are looking at processes in which the next state is independent of the history, that is *Markov*. In such cases it seems natural to ask if optimal policies can be chosen to be Markov aswell. This is the topic of the next section which we will turn to now.

4 Markov decision processes

Definition 4.1 (Markov decision process)**.** A **Markov decision process** (MDP) consists of

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

Remark 4.2. This is a special case of the history dependent decision process (definition 2.5) with

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

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

Remark 4.3. One could ask if it is possible to embed a HDP into an MDP by taking unions or cartesian products of the state and action spaces:

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

One attempt at this can be found in [2] chapter 10, but this will not be covered here. Note however that any properties, such as those in setting 1, one assumes regarding the spaces $\mathcal{S}_1, \mathcal{A}_1, \ldots$, one must reconsider if each such property hold in the new constructed MDP.

Remark 4.4. In an MDP the policy evaluations can be written

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

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

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

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

Definition 4.5 (Policy classes)**.** A policy $\pi = (\pi_1, \pi_2, \ldots) \in R\Pi$ is called **Markov** if it only depends on the last state is the history. That is there exist $\pi_1, \pi_2, \cdots : \mathcal{S} \rightsquigarrow \mathcal{A}$ such that $\pi_i(\cdot \mid s_1, \ldots s_i) = \pi_i(\cdot \mid s_i)$. We denote the set of (random) Markov policies by $M\Pi$. If $\pi_1 = \pi_2 = \ldots$ the Markov policy is called **stationary** and the set of them denote by $S\Pi$. Furthermore π is called **deterministic** if all π_i are degenerate, i.e. for all i we have $\pi_i(\{a_i\} \mid h_i) = 1$ for some $a_i \in \mathcal{A}_i$. We denote the deterministic version of the policy classes by the letter D.

Remark 4.6. We have the following inclusions of policy classes

$$\begin{array}{ccc} S\Pi & \subseteq M\Pi & \subseteq R\Pi \\ \rotatebox{90}{$\subseteq$} & \rotatebox{90}{$\subseteq$} & \rotatebox{90}{$\subseteq$} \\ DS\Pi & \subseteq DM\Pi & \subseteq D\Pi \end{array}$$

Note that stationary policies might not exist in HDPs, but always exist for MDPs. A policy $(\pi_1, \pi_2, \dots) \in R\Pi$ is deterministic if and only if there exist measurable functions $\varphi_n : \mathcal{H}_n \to \mathcal{A}$ such that $\pi_n(\cdot \mid h_n) = \delta_{\varphi_n(h_n)}$. Therefore we shall sometimes write $\pi_n(h_n) = \varphi_n(h_n)$, viewing π_n as a function. For convenience will view stationary policies $\tau \in S\Pi$ interchangeably as kernels $\tau : \mathcal{S} \rightsquigarrow \mathcal{A}$ and as the policy $(\tau, \tau, \dots)$.

We will prove that in MDPs under mild assumptions the optimal policy can be chosen to be deterministic, Markov, and even stationary. Before we do this we define some important tools for studying MDPs. We now need to integrate over value functions so it is necessary to talk about integrability.

Definition 4.7. Let $\mathcal{L}_\infty(\mathcal{S})$ denote the set of measurable functions on $\mathcal{S}$ which are essentially bounded with respect to all measures arising as the distribution $\rho_{\mathcal{S}_i}(\kappa_\pi \mu)$ of the ith state space given some initial distribution $\mu \in \mathcal{P}(\mathcal{S})$ and some policy $\pi \in R\Pi$. Here $\rho_{\mathcal{S}_i} : \mathcal{H}_\infty \to \mathcal{S}_i$ is projection onto the ith state space. Similarly $\mathcal{L}_p$ for $p \in [1, \infty)$ is defined with respect to all such probability measures.

Remark 4.8. Note that $\mathcal{L}_\infty(\mathcal{S})$ contains all (everywhere) bounded functions, so is non-empty. In the situation where one can prove that all measures in question are absolutely continuous w.r.t. some measure ν then it is enough to ensure boundedness ν-almost everywhere.

Definition 4.9 (The T-operators)**.** For a stationary policy $\tau \in S\Pi$ and a value function $V : \mathcal{S} \to \mathbb{R} \in \mathcal{L}_\infty(\mathcal{S})$ we define the operators

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

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

Remark 4.10. We will sometimes write $T_a = T_{\delta_a}$ for $a \in \mathfrak{A}(s)$. Using this we can express T alternatively as $TV = s \mapsto \sup_{a \in \mathfrak{A}(s)} T_a V(s)$.

The Bellman optimality operator T is harder to work with than T_τ because it envolves a supremum. Therefore we will first take a closer look at properties of T_τ.

Proposition 4.11 (Properties of the T_τ-operator)**.** Let $\pi = (\tau_1, \tau_2, \dots) \in M\Pi$ be a Markov policy, and $\tau \in S\Pi$ be a stationary policy.

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

6. V_τ is the unique bounded fixed point of T_τ in $\mathcal{L}_\infty(\mathcal{S})$.

Proof.

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

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

   Now use this inductively.

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

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

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

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

□

Proposition 4.11 gives us a way of obtaining the finite policy evaluations by iteratively using the T_τ operator (for $\tau = \tau_1, \tau_2, \dots$). We emphasize this by writing it as our first algorithm, called the *policy evaluation algorithm*. We should say that this algorithm has been considered before and is used in other algorithms like the *policy iteration algorithm* (see e.g. [2]).

Algorithm 1: Simple theoretical policy evaluation

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

The reason behind calling line 3 *theoretical* is that we have not talked about how to actually represent any of r, T_π or V in a computer. We will take up this discussion later (see e.g. example 4.27).

In many cases we are actually interested the evaluation of a stationary policy $\tau \in S\Pi$ on an infinite horizon, that is we are interested in obtaining the infinite policy evaluation V_τ. To this end line 3 can still be used as a method of approximation, since by proposition 4.11.(3,5 and 6) and Banach fixed point theorem we have that

Corollary 4.12. Let $\tau \in S\Pi$ be a stationary policy. Then the kth finite policy evaluation $V_{\pi,K}$, that is the output of line 3 with k iterations, satisfy

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

It is in fact mostly due corollary 4.12 that line 3 is used.

4.1 Greedy policies

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

Definition 4.13. Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A} \in S\Pi$ be a stationary policy and let $V : \mathcal{S} \to \mathbb{R}$ be a measurable value-function. We define

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

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

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

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

Remark 4.14. For a function $f : \mathcal{X} \to \overline{\mathbb{R}}$ and $A \subseteq \mathcal{X}$ we denote

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

In order to prove the existence of greedy policies we need some additional structure on our MDP. Recall (definition 3.11) that a measurable space is standard Borel if it is Polish and equipped with the Borel σ-algebra. Also recall that the space of probability measures $\mathcal{P}(\mathcal{X})$ over a standard Borel space $\mathcal{X}$ is also standard Borel when endowed with the weak topology (see definition 3.13 and proposition 3.15).

Definition 4.15 (Continuous kernel)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be standard Borel measurable spaces. A probability kernel $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is **continuous** if the map

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

is continuous.

Setting 2 (Greedy MDP)**.**

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

By the results of the history dependent section we can already establish import facts about MDPs under setting 2.

Proposition 4.16. Let $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be an MDP under setting 2. Then there exists an optimal policy $\pi^* \in R\Pi$, we have convergence of n-optimal value functions $V_n^* \to V^*$ and V_n^*, V^* are measurable.

Proof. As we have noted an MDP is just a special case of a HDP. Moreover setting 1 and (S).2 is directly implied by setting 2. Lastly the set-wise continuity ((S).1) follows since pr. setting 2.4, P is continuous, which by proposition 3.14 implies that $P(S \mid s_1, \cdot, \ldots, s_n, \cdot)$ is continuous for all $S \in \Sigma_{\mathcal{S}_{n+1}}$ and $(s_1, \ldots, s_n) \in \mathcal{S}^{\underline{n}}$. Therefore we can apply theorem 3.19 and corollary 3.20 and we are done. □

Remark 4.17 (Almost optimal policies)**.** For MDPs which do not fulfill the requirement of setting 2 policies which are *close to optimal* some sense may still exist. Let $\varepsilon > 0$, then a policy $\pi \in R\Pi$ is called ε-optimal if $V_\pi > V^* - \varepsilon$. One can then show that under 1. of setting 2 and the additional assumption that $\mathfrak{A}(s)$ is countable for all $s \in \mathcal{S}$, we have that deterministic ε-optimal policies always exist (see theorem 18 [6, Feinberg (2012)]). For a detailed discussion of this topic see [6].

Proposition 4.18. Let $\mathcal{X}$ and $\mathcal{Y}$ be separable metrizable and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a continuous probability kernel. Let $f : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be Borel measurable, bounded from below or above. Define

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

Then

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

Proof. We refer to [2] prop.7.31. □

Proposition 4.19. A upper (lower) semicontinuous function $f : \mathcal{X} \to \overline{\mathbb{R}}$ on a compact metrizable space $\mathcal{X}$ attains its supremum (infimum). That is there exists an $x^* \in \mathcal{X}$ ($x_* \in \mathcal{X}$) such that $f(x^*) = \sup_{x \in \mathcal{X}} f(x)$ ($f(x_*) = \inf_{x \in \mathcal{X}} f(x)$).

Proof. Let $x_1, x_2, \dots \in \mathcal{X}$ be a sequence such that $f(x_n) \to \sup_{x \in \mathcal{X}} f(x)$ then since $\mathcal{X}$ is compact this sequence has at least one accumulation point $x^* \in \mathcal{X}$. Let $x_{k_1}, x_{k_2}, \dots$ be a subsequence such that $x_{k_n} \to x^*$. Since f is upper semicontinuous

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

The statement for lower semicontinuous f is analogous. □

Proposition 4.20. Let $\mathcal{X}$ be metrizable, $\mathcal{Y}$ compact metrizable, $\Gamma \subseteq \mathcal{X} \times \mathcal{Y}$ be closed with $\rho_{\mathcal{X}}(\Gamma) = \mathcal{X}$, where $\rho_{\mathcal{X}}$ is projection onto $\mathcal{X}$ and let $f : \Gamma \to \overline{\mathbb{R}}$ be upper semicontinuous. Let

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

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

Remark 4.21. Here $\mathrm{Gr}(f) = \big\{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid y = f(x)\big\}$ denotes the graph of the function $f : \mathcal{X} \to \mathcal{Y}$.

Proof of proposition 4.20. We refer to [2] prop.7.33. □

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

Proposition 4.22 (Existence of greedy deterministic policies)**.** Under setting 2 let $V : \mathcal{S} \to \mathbb{R} \in \mathcal{L}_\infty(\mathcal{S})$ be upper semicontinuous. Then for any $s \in \mathcal{S}$ it holds that

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

Proof.

1. This is a consequence of proposition 3.17 and proposition 4.18 since r is upper semicontinuous.
2. Since by 1. $(s, a) \mapsto T_a V(s)$ is upper semicontinuous, this follows by proposition 4.19.
3. Recall that the set of admissable state-action pairs $\Gamma = \big\{(s, a) \in \mathcal{S} \times \mathcal{A} \mid a \in \mathfrak{A}(s)\big\}$ assumed to be closed in setting 2. Let $\varphi : \mathcal{S} \to \mathcal{A}$ be the Borel-measurable function obtained from proposition 4.20 with $\mathrm{Gr}(\varphi) \subseteq \Gamma$. Let $s \in \mathcal{S}$ be any state. Then

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

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

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

is greedy with respect to V. Upper semicontinuity of TV follows from proposition 4.18 since $TV = T_{\tau_V}$. If $\tau \in S\Pi$ is a stationary policy then

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

Therefore $\sup_{\tau\in S\Pi} V \leqslant TV$. On the other hand τ_V is in the supremum so $\sup_{\tau\in S\Pi} V \geqslant TV$.

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

□

4.2 Existence of optimal policies

In the light of proposition 4.22 (and induction) since $V_0 = 0$ is upper semicontinuous for any $k \in \mathbb{N}$ we have that $T^k V_0$ is upper semicontinuous and thus has an associated greedy policy $\tau_{T^k V_0} \in DS\Pi$ which we will denote τ_k^*.

Proposition 4.23 (Existence of n-stage optimal policies)**.** Under setting 2 we have that

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

and this is an upper semicontinuous function. Thus $(\tau_{k-1}^*, \dots, \tau_0^*)$ is a deterministic k-optimal policy where $\tau_k^* = \tau_{T^k V_0}$ is any deterministic greedy policy for $T^k V_0$ for any $k \in \mathbb{N}$.

Proof. We begin by stating some preliminary facts.

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

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

As induction basis observe that $0 = V_0 = V_0^*$ is upper semicontinuous. Assume that $T^{k-1} V_0 = V_{k-1}^*$ is upper semicontinuous. Let $s_1 \in \mathcal{S}$.

$$\begin{aligned} V_k^*(s_1) &\overset{\text{def}}{=} \sup_{\pi\in R\Pi} \int \sum_{i=1}^{k} \gamma^{i-1} r(s_i, a_i)\, \mathrm{d}\kappa_\pi(a_1, s_2, \cdots \mid s_1) && (4) \\ &\overset{2.9}{=} \sup_{\pi\in R\Pi} \int r(s_1,a_1) + \gamma \int \left(\int \sum_{i=1}^{k-1} \gamma^{i-1} r(s_{i+1}, a_{i+1})\, \mathrm{d}\kappa_{\pi^{s_1,a_1}}(a_2, s_3, \cdots \mid s_2) \right) && (5) \\ &\quad \mathrm{d}P(s_2 \mid s_1, a_1)\, \mathrm{d}\pi_1(a_1 \mid s_1) && (6) \\ &\overset{\text{hyp.}}{\leqslant} \sup_{\pi\in R\Pi} \int r(s_1,a_1) + \gamma \int V_{k-1}^*(s_2)\, \mathrm{d}P(s_2 \mid s_1, a_1)\, \mathrm{d}\pi_1(a_1 \mid s_1) && (7) \\ &= \sup_{\pi_1\in S\Pi} T_{\pi_1} V_{k-1}^*(s_1) && (8) \\ &\overset{4.22.4}{=} TV_{k-1}^*(s_1) && (9) \end{aligned}$$

The integral makes sense in the eq. (7) because of the induction hypothesis, since $V_{k-1}^* = V_{k,(\tau_{k-1}^*,\dots,\tau_0^*)}$ and this is integrable like all policy evaluations (see proposition 3.6). In eq. (8) the supremum changes because only the first step in the policy is used (this first step is a stationary policy). Equation (9) is by proposition 4.22.4. Since s_1 was arbitrary we must have $V_k^* \leqslant TV_{k-1}^*$. On the other hand by proposition 4.11 and induction hypothesis we have

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

But since $(\tau_{k-1}^*, \dots, \tau_0^*)$ occur in the supremum we must then also have $TV_{k-1}^* \leqslant V_k^*$. Note that upper semicontinuity of V_k^* follows since T preserves this property (see proposition 4.22.4). □

Proposition 4.24. Under setting 2 it holds that $V^* = \lim_{k\to\infty} T^k V_0^*$. Furthermore under the greedy policy τ_{V^*} exists and is a deterministic stationary optimal policy.

Proof. Since setting 2 implies condition (S) (see above theorem 3.19) we have by theorem 3.19 that $T^k V_0^* = V_k^* \to V^*$ (as always we mean pointwise convergence here). We know by proposition 4.23 that V_k^* is semi uppercontinuous for all $k \in \mathbb{N}$. Also we have that

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

Here ↓ denotes downwards monotone (pointwise) convergence. So by proposition 3.17 the infimum $\inf_k \widehat{V}_k = V^* - V_{\max}$ is upper semicontinuous and thus V^* is upper semicontinuous. Therefore by proposition 4.22 there exists a deterministic greedy policy τ_{V^*} which satisfies

$$T_{\tau_{V^*}} V^* = TV^* \tag{10}$$

By proposition 4.11 T and $T_{\tau_{V^*}}$ is contractive on the Banach space $B = \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A}) \ni V_0^*$. Therefore by Banach fixed point theorem (see theorem A.24) $V^* = \lim_{k\to\infty} T^k V_0^*$ is the unique fixed point of T in B. Again by Banach fixed point theorem and eq. (10) V^* is the fixed point of $T_{\tau_{V^*}}$, which by proposition 4.11 also has $V_{\tau_{V^*}}$ as fixed point. By uniqueness $V_{\tau_{V^*}} = V^*$ and thus τ_{V^*} is optimal. □

Remark 4.25. The property that $TV^* = V^*$ is often referred to as *Bellman's optimality equation.*

Corollary 4.26. Under setting 2 we have that τ_{V^*} is optimal and for any $V \in \mathcal{L}(\mathcal{S} \times \mathcal{A})$ it holds that

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

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

Proof. Since (D) implies the last point in setting 1 we can apply proposition 4.24. The bound on $|T^k V - V^*|$ is by the Banach fixed point theorem. Upper semicontinuity of the policy evaluations and optimal value functions followed from the proofs of proposition 4.22 and proposition 4.24. □

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

In [2] (prop. 8.6 and cor. 9.7.2) results very similar to proposition 4.23 and proposition 4.24 are also established with in a slightly different setup. Besides having a state and action space, [2] also considers a non-empty Borel space called the *disturbance space* W, a *disturbance kernel* $p : \mathcal{S} \times \mathcal{A} \to W$, instead of a transition kernel which on the other hand is a deterministic *system function* $f : \mathcal{S} \times \mathcal{A} \times W \to \mathcal{S}$ which should be Borel measurable. Also the rewards are interpreted as negative costs, and thus g is required to be semi *lower*continuous. In [2] are also found much theory assuming semi*analytic* functions instead of semicontinuous ones.

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

4.3 Value iteration

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

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

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

To actually use line 3 it becomes relevant if the environment can be *computed* efficiently (in a computer). An easy example such an easily computed environment is an MDP where both the state and the action space are finite i.e. $|\mathcal{S}| \cdot |\mathcal{A}| < N < \infty$ for a not too big number $N \in \mathbb{N}$. Indeed value iteration was invented for finite state and action spaces, but as we have shown, exponential convergence to the optimal infinite horizon value function is guaranteed in far more general case (an MDP under setting 2), and therefore line 3 could be applied in other cases if one has a practical way of representing the iterations $TV_0, T^2V_0, \ldots$.

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

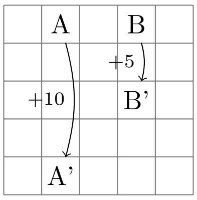


Figure 1: The simple gridworld Markov decision process.

Example 4.27 (Gridworld)**.** The *gridworld* MDP consist of 25 states $\mathcal{S} = [5]^2$ and 4 actions $\mathcal{A} = \{U, D, L, R\}$ for *up*, *down*, *left* and *right*. The transition P and reward R kernels are deterministic: the agent moves 1 square up, down, left or right according to the chosen action and receives a reward of 0, except:

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

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

Finite spaces are trivially standard Borel with the discrete topology, which also makes every map $(s, a) \mapsto \mathcal{X}$ into some topological space continuous. In particular P is continuous and r is (upper semi)continuous. The set of admissible actions $\mathfrak{A}(s)$ is equal to the full action space $\mathcal{A}$ for all $s \in \mathcal{S}$, which is trivially compact. The rewards are bounded by $R_{\max} = 10$ and therefore $V_{\max} = 10/(1 - 0.9) = 100$. Thus we can apply corollary 4.26 and for $\widetilde{V}_0 = 0$ get that

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

By proposition 4.11 for any stationary policy $\tau \in S\Pi$ we have that T_τ is also γ-contractive and we easily get the same bound on the policy evaluation

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

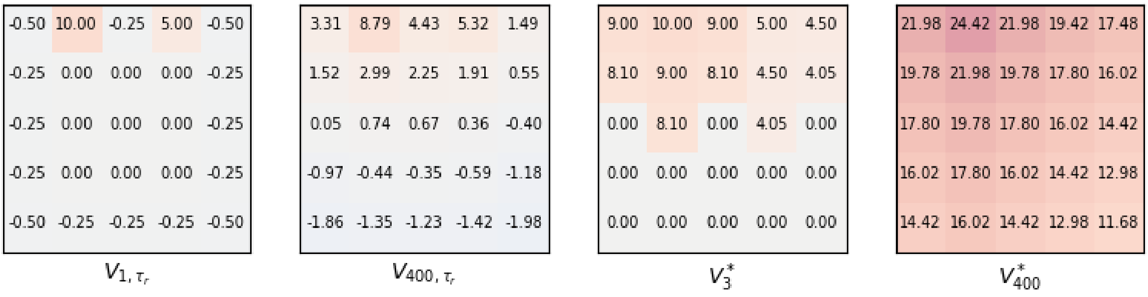
To use line 3 and line 3 on the gridworld example we need to understand how the kernels R, P, value functions, policies and operators T_τ, T are represented in a computer. Since $\mathcal{S}$ and $\mathcal{A}$ are finite and small we can simply treat value functions V as a vector $\widetilde{V} \in \mathbb{R}^{\mathcal{S}}$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

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

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

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

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



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

V_{1,τ_r}

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

V_{400,τ_r}

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

V_3^*

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

V_{400}^*

Figure 2: Value functions of the gridworld environment. Note that $V_{\max} \cdot \gamma^{400} = 100 \cdot (0.9)^{400} \approx 4.97 \cdot 10^{-17}$ so V_{400}^* and $V_{\tau_r,400}$ are very close to the true infinite horizon value functions V^* and V_{τ_r} (providing numerical errors are insignificant).

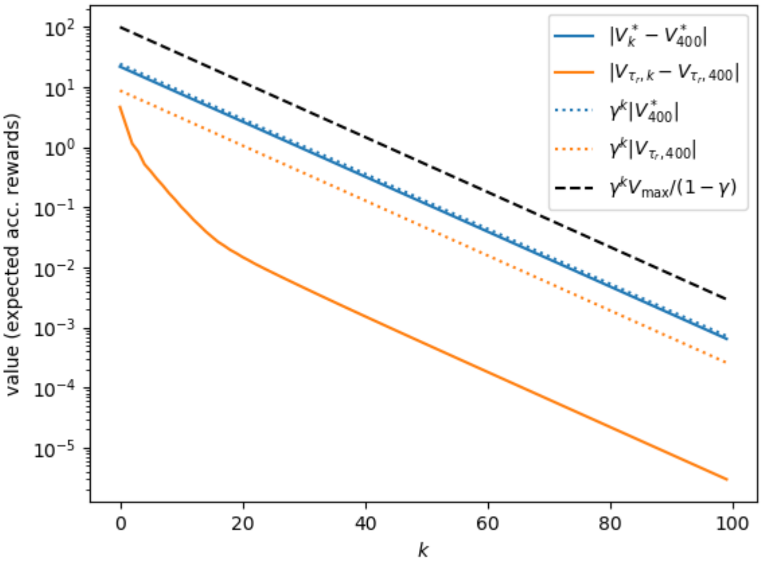


Figure 3: Convergence of gridworld value functions compared with the theoretical bounds. The black dashed line is the general theoretical bound for both T and T_τ by Banachs fixed point theorem and the maximum value $V_{\max} = R_{\max}/(1-\gamma)$. The dotted blue and orange uses $|V_k^*|$ and $|V_{\tau,k}|$ respectively, which might not be available. ($\gamma = 0.9$).

4.4 Q-functions

A **Q-function** is any function that assigns a real number to every state-action pair, that is any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Q-functions are also called *action-value* functions, to distinguish them from the *value* functions we have discussed in the previous sections. The idea of Q-functions (and the letter Q) originates to [14, Watkins (1989)]. Upon the definition he notes

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

A clear advantage of working with Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ rather than a value function $V : \mathcal{S} \to \mathbb{R}$, is that finding the optimal action $a^* \in \mathfrak{A}(s)$ at state s requires only a maximization over the Q-function itself: $a^* = \operatorname{argmax}_{a \in \mathfrak{A}(s)} Q(s,a)$. This should be compared to finding an optimal action according to a value function V: $a^* = \operatorname{argmax}_{a \in \mathfrak{A}(s)} r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V$. Besides being less simple, this requires taking an expectation with respect to both the reward and transition kernel. Later we will study settings where we can only sample from P and R when attempting to find the optimal strategy. In these situations the advantage of Q-functions is clear. For now however the transition kernel will remain known and we will in this section see how the results of state-value functions translate to Q-functions. Because of the similar role Q-functions play compared to value function, many concepts such as T-operators and the finite, infinite horizon policy evaluations and greedy policies, can be defined analogously.

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

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

Remark 4.28. We make assumption 3 to ensure that the supremum $\sup_{a \in \mathfrak{A}(s)} Q(s,a)$ is attained for any $Q \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$. One might be able discard assumption 3 and instead demand that all Q-functions be upper semicontinuous generalizing the discussion in this section. We have not pursued this generalization.

Definition 4.29 (Policy evaluation for Q-functions)**.** Let $\pi \in R\Pi$. Define

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

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

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

Definition 4.30 (Operators for Q-functions)**.** For any stationary policy $\tau \in S\Pi$ and integrable Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ we define

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

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

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

where $T_a = T_{\delta_a}$.

Remark 4.31. The next-step operator P_τ is defined for simplications in proofs, especially in the analysis of [5, Fan et al. (2020+)] in the later sections. Using P_τ we can write T_τ alternatively as $T_\tau Q(s,a) = r(s,a) + \gamma P_\tau Q(s,a)$.

Definition 4.32 (Greedy policies for Q-functions)**.** Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A}$ be a stationary policy. Define $G_Q(s) = \operatorname{argmax}_{a \in \mathfrak{A}(s)} Q(s,a)$. If there exist a measurable set $G_Q^\tau(s) \subseteq G_Q(s)$ for every $s \in \mathcal{S}$ such that

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

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

Proposition 4.33 (Relations between Q- and value functions)**.** Let $\pi = (\tau_1, \tau_2, \dots) \in M\Pi$ be a Markov policy and $\tau \in S\Pi$ stationary. Then

1. Policy evaluations are related by $\mathbb{E}_{\tau(\cdot|s)} Q_{k,\pi} = V_{k+1,(\tau,\pi)}(s)$.

2. T_τ-operators are related by $T_\tau Q_{k,\pi}(s,a) = r + \gamma \mathbb{E}_{P(\cdot|s,a)} T_\tau V_{k,\pi}$.

3. Greedy policies for policy evaluations are the same. That is

   (a) τ is greedy for $Q_{k,\pi}$ if and only if τ is greedy for $V_{k,\pi}$.

   (b) τ is greedy for Q_π if and only if τ is greedy for V_π.

4. Optimal policies are related by $\max_{a \in \mathfrak{A}(s)} Q^*(s,a) = V^*(s)$ and

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

Proposition 4.34 (Properties of Q-functions)**.** Let $\pi = (\tau_1, \tau_2, \dots) \in M\Pi$ be a Markov policy and $\tau \in S\Pi$ stationary. Then

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

Proof of proposition 4.33 and proposition 4.34. Measurability of $Q_{k,\pi}$ and Q_π follow from measurability of $V_{k,\pi}$, V_π and proposition 2.4. Upper semicontinuity of $Q_{k,\pi}$ and Q_π follows from proposition 4.18 because $V_{k,\pi}$ and V_π are upper semicontinuous (see corollary 4.26).

For proposition 4.33.1 we have

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

For proposition 4.33.2 we sketch the idea by

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

For $Q_{k,\pi} = T_{\tau_1} \dots T_{\tau_k} Q_0$ use proposition 4.33.2 iteratively starting with $\tau = \tau_1, \pi = (\tau_2, \tau_3, \dots)$.

The $\tau(Q_{k,\pi}) = \tau(V_{k,\pi})$ part of proposition 4.33.3 is by definition of the two concepts of greedy functions.

That $Q_\pi = \lim_{k\to\infty} Q_{k,\pi}$ follows from dominated convergence and proposition 3.6.3.

For proposition 4.33.4 $Q_k^* = \sup_{\pi \in R\Pi}(r + \gamma \mathbb{E} V_{k,\pi}) \leqslant r + \gamma \mathbb{E} V_k^* = r + \gamma \mathbb{E} V_{\pi_k^*} \leqslant Q_k^*$. The same argument works for the second part.

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

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

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

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

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

Therefore by proposition 4.34.1

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

since by proposition 4.33.3 τ_{k-1}^* is greedy for Q_{k-1}^*. Now use induction to get $Q_{k-1}^* = T^k Q_0^*$.

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

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

The contrativeness of T and T_π follows from the same argument as for value functions. Banach fixed point theorem now concludes proposition 4.34.3.

Since now Q^* and Q_{τ^*} are fixed points for T they must be equal, concluding the last point, namely proposition 4.34.4. □

4.4.1 Q-iteration

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

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

By proposition 4.34.3 we thus have convergence of the line 3:

Corollary 4.35. $\left|Q^*_k - Q^*\right| \leqslant \gamma^k \|Q^*\|_\infty \leqslant \gamma^k V_{\max}$.

4.5 Why are we not done?

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

1. Gridworld (example 4.27).
2. Board games like *chess* and *go* against a fixed (possibly randomized) opponent policy can be modelled accurately as such finite MDPs (putting γ close to 1, so that winning late in the game is still considered worth the effort).
3. *Pole balancing* in a 2D physical simulation environment (see e.g. the famous *cartpole* example from [1, Barto et al. (1983)]). One may even add random effects (such as *wind* effects) to make full use of our stochastic setup.

The problem is of course that the value functions and operators which are used in line 3 are not computable in practice. For example the state space of chess is very large (roughly $|\mathcal{S}_{\text{chess}}| \geqslant 10^{43}$). This means that if we were to use line 3 naively (with finite implementation as example 4.27) then we would have to store a vector of roughly $N \cdot 10^{43}$ real numbers for each Q-function we define, where N is the average number of admissable actions at each state $\mathfrak{A}(s), s \in \mathcal{S}$ which has been estimated to around 35 for chess. This requires roughly $1.4 \cdot 10^{45}$ bytes, if each number is stored as a single precision floating point number (4 bytes). For comparison the entire digital data capacity in the world is estimated less than 10^{23} bytes as of 2020. Needless to say this is beyond any practical relevance.

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

5 Q-learning with function approximators

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

It seems to this author, that this setting is not very well-studied in the case of a continuous state space. This is perhaps because it is considered solved by the results of theoretical Q-learning presented in the previous section. However as we have argued, this only have practical relevance when it is feasible to represent TQ. Therefore we find it relevant to consider this setting in more detail.

What *is* very well-studied is a further generalized setting where T and r are assumed to be unknown, that is, one has only access to their distributions via sampling from them. Solutions for that setting are called *model-free*. We will deal with that setting in the next section.

5.1 Algorithmic and approximation errors

In the following we present some rather simple bounding techniques which is inspired by arguments found in e.g. [5], together with some standard results from approximation theory on artificial neural networks and Bernstein polynomials.

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

$$\left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\| \leqslant \varepsilon_k, \forall k \in \mathbb{N} \tag{11}$$

First observe that

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

Using this iteratively we get

$$\left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \leqslant \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i \tag{12}$$

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

$$\varepsilon_{\text{approx}}(k) := \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i \tag{13}$$

Thus we get

Theorem 5.1. Let $\|\cdot\|$ be a norm on the space of functions $\mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$. Let $\widetilde{Q}_k$ be obtained from a function class $\mathcal{F}$ such that $\left\|\widetilde{Q}_k - T\widetilde{Q}_{k-1}\right\| \leqslant \varepsilon_k$ for any $k \in \mathbb{N}$. Then

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

Proof. By the discussion above and

$$\begin{aligned}\left\|Q^* - \widetilde{Q}_k\right\| &\leqslant \left\|Q^* - T^k\widetilde{Q}_0\right\| + \left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \\ &\overset{12}{\leqslant} \gamma^k\left\|Q^* - \widetilde{Q}_0\right\| + \varepsilon_{\text{approx}}(k)\end{aligned}$$

□

The first term in theorem 5.1 is sometimes called the **algorithmic** error[1]. The algorithmic error converges exponentially, so one is usually happy with this part not spending time trying to bound this tighter. The approximation error depends on our step-wise approximations. For example if $\varepsilon_i(k) = \varepsilon$ for some $\varepsilon > 0$ we easily get the bound

$$\varepsilon_{\text{approx}}(k) = \varepsilon \frac{1-\gamma^k}{1-\gamma} \leqslant \frac{\varepsilon}{1-\gamma} \tag{14}$$

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

Proposition 5.2. $\sum_{i-1}^{k} \gamma^{k-i}\varepsilon_i \to 0$ whenever $\varepsilon_k \to 0$ as $k \to \infty$.

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

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

□

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

5.2 Using artifical neural networks

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

Definition 5.3. An **artificial neural network** (ANN) with $L \in \mathbb{N}_0$ hidden layers, structure $(d_i)_{i=0}^{L+1} \subseteq \mathbb{N}$, activation functions $(\sigma_i)_{i=1}^{L}$, weights $(W_i)_{i=1}^{L+1} \in M^{d_i \times d_{i-1}}$ and biases $(v_i)_{i=1}^{L+1} \in \mathbb{R}^{d_i}$ is the function $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$ defined by

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

where w_i is the affine function $x \mapsto W_i x + v_i$, and $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ is coordinate-wise application of components $\sigma_{ij} : \mathbb{R} \to \mathbb{R}$. We denote the class of these networks (or functions)

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

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

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

Remark 5.4. Artificial neural networks are often illustrated as $L+1$-partite graphs with d_i nodes in the ith partition. A node $n_{i,j}$ in partition i is then connected to a node $n_{i+1,k}$ if $W_{i+1}(k,j) \neq 0$. This is because they were inspired by the structure of *neurons* in nerve tissue (e.g. the brain) of living organisms, with the graph nodes corresponding to neurons and edges to *axons*. Indeed for every suitable collection of activation functions and every $L+1$-partite weighted graph G satisfying

> The ith partition is only connected to the neighboring $(i-1)$th and $(i+1)$th partition. (15)

[1] For example in [5].

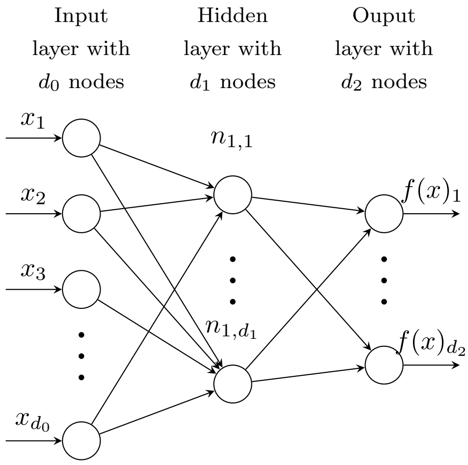


Figure 4: An ANN with one hidden layer ($L = 1$). Notice that there is no edge from $n_{0,3}$ to $n_{1,1}$ which means that $W_1(1, 3) = 0$.

Then there exists a corresponding ANN corresponding to G. From the graph view-point it easy to see that one may join neural nets to form larger ones, by either function composition or side-by-side alignment. That this if $f \in \mathcal{DN}\left((\sigma_i)_{i=1}^{L},\ (d_i)_{i=0}^{L+1}\right)$ and $g \in \mathcal{DN}\left((\sigma'_i)_{i=1}^{L'},\ (d'_i)_{i=0}^{L'+1}\right)$ are two ANNs and $d_{L+1} = d'_0$ (i.e. that the dimensions match $\mathrm{im}(f) \subseteq \mathrm{dom}(g)$) then $g \circ f \in \mathcal{DN}\left((\sigma_1, \ldots, \sigma_L, \sigma'_1, \ldots, \sigma'_{L'}), (d_1, \ldots, d_L, d'_1, \ldots, d'_{L'})\right)$. By side-by-side alignment we mean the situation where $L = L'$ and one creates the function $h : \mathbb{R}^{d_0+d'_0} \to \mathbb{R}^{d_L+d'_L}$ by defining $h(x_1, \ldots, x_{d_0+d'_0}) = (f(x_1, \ldots, x_{d_0}), g(x_1, \ldots, x_{d'_0}))$. With this way of defining h we have that h is an ANN with structure $(d_0 + d'_0, \ldots, d_{L+1} + d'_{L+1})$. Generally any way of stitching together graphs into n-partite graphs satisfying eq. (15) will gives ways of producing new ANNs.

Theorem 5.5 (Universal Approximation Theorem for ANNs)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be non-constant, bounded and continuous activation function. Let $\varepsilon > 0$ and $f \in C([0, 1]^w)$. Then there exists an $N \in \mathbb{N}$ and a network $F \in \mathcal{DN}(\sigma, (w, N, 1))$ with one hidden layer, unbiased final layer (that is $v_2 = 0$) and activation function σ such that

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

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

Discussion of proofs of theorem 5.5. The original proof in [4, Cybenko (1989)] is very short and elegant, but non-constructive, using the Riesz Representation and Hahn-Banach theorems to obtain a contractiction to the statement that $\bigcup_{N\in\mathbb{N}} \mathcal{DN}(\sigma, (w, N, 1))$ is dense in $C([0, 1]^w)$. Furthermore it considered only *sigmoidal* activations functions, meaning that σ should satisfy

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

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

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

Definition 5.6. We define the class of ReLU networks as the ANNs (see definition 5.3) with all ReLU activation functions, and write $\mathcal{RN}\left((d_i)_{i=0}^{L+1}\right) := \mathcal{DN}\left(\sigma_r, (d_i)_{i=0}^{L+1}\right)$.

Proposition 5.7. Under setting 3 let $\varepsilon > 0$. Assume that either

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

   or

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

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

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

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

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

Proof. The key points are that

a. Any ANN with continuous activation functions is continuous.

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

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

Using these facts we can use the universal approximation theorem (theorem 5.5) to get a series of networks

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

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

$$\left|f_{a,k} - T\widetilde{Q}_{k-1}(\cdot, a)\right| < \varepsilon \tag{16}$$

Here $\widetilde{Q}_0 = r$ and $\widetilde{Q}_k$ is obtained recursively by joining for each a the components $f_{a,k}$ into a single network $\widetilde{Q}_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $\widetilde{Q}_k(s, a) = f_{a,k}(s)$. By eq. (16) on each of its components approximates $T\widetilde{Q}_{k-1}$ to ε precision. $\widetilde{Q}_0$ is continuous by setting 3 and by a. and b. $T\widetilde{Q}_k$ is as well.

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

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

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

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

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

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

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

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

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

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

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

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

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

□

This gives us the first method of how to approximate Q^* arbitrarily closely on continuous state spaces, in the case where it is infeasible to represent TQ directly. However it is still not clear if this method is feasible computationally. To investigate this and indeed for any chance to implement the method in practice one would need to go through the construction in [3]. We will not go further into this, and instead focus on another approximation method using *Bernstein polynomials.*

5.3 Using Bernstein polynomials

In this case we need a stronger form of continuity, namely Lipschitz continuity (see definition A.17), to establish the bounds.

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

Definition 5.8 (Bernstein polynomial)**.** The (multivariate) Bernstein polynomial $B_{f,n}$ of degree $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ approximating the function $f : [0,1]^w \to \mathbb{R}$ is defined by

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

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

Theorem 5.9 (Approximation with Bernstein polynomials)**.** Let $f : [0,1]^w \to \mathbb{R}$ be Lipschitz w.r.t. the standard euclidean 2-norm induced metrics on $[0,1]^w$ and $\mathbb{R}$ with constant L. Let $n = (n_1, \dots, n_w) \in \mathbb{N}^w$. The Bernstein polynomial $B_{f,n} : [0,1]^w \to \mathbb{R}$ satisfies

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

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

Lemma 5.10. Under setting 4 $TQ(\cdot, a)$ is Lipschitz in $\|\cdot\|_2$ with constant $L_T = L_r + \gamma V_{\max} L_p$ for all $a \in \mathcal{A}$ and measurable $Q : \mathcal{S} \times \mathcal{A} \to [-V_{\max}, V_{\max}]$.

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

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

□

Using this we can make a bound on the approximation error:

Proposition 5.11. Given an MDP satisfying setting 4 and using $\|\cdot\|_\infty$ we can bound

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

Proof. Following the procedure from leading to eq. (12), starting with $\widetilde{Q}_0 = 0$ and using the $n = (n_1, \dots, n_w)$ degree Bernstein polynomium $\widetilde{Q}_k = B_{T\widetilde{Q}_{k-1}, n}$ as approximation for $T\widetilde{Q}_{k-1}$ we know by induction and the results lemma 5.10 and theorem 5.9.2 that $T\widetilde{Q}_k$ is L_T-Lipschitz for any $k \in \mathbb{N}$. Now by choosing the euclidean norm $\|\cdot\| = \|\cdot\|_2$ we have by theorem 5.9.1 that

$$\varepsilon_i = \left\|\widetilde{Q}_k - T\widetilde{Q}_{k-1}\right\|_\infty \leqslant \frac{L_T}{2} \sqrt{\sum_{j=1}^{w} \frac{1}{n_j}} = \varepsilon \tag{17}$$

where ε is the one-step error defined in eq. (11). Now by eq. (14) we have that

$$\varepsilon_{\text{approx}} \leqslant \frac{\varepsilon}{1-\gamma} \tag{18}$$

Combining eq. (17) and eq. (18) and noting that $L_T = L_r + \gamma V_{\max} L_p$ finishes the proof. □

To make more clear what are the implications of proposition 5.11 we give a corollary where we put $n_j = m$ for all $j \in [w]$:

Corollary 5.12. Under setting 4 and using Bernstein polynomials of degree $n = (m, \ldots, m) \in \mathbb{N}^w$ for $m \in \mathbb{N}$ we have the following bound

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

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

Proof. Use proposition 5.11. □

This gives a very concrete way of constructing an arbitrarily good approximation to Q^* using polynomials. A major drawback is the restriction on the transition dynamics P. For example we cannot use corollary 5.12 on deterministic decision processes, since if P is deterministic then there are no measure $\mu \in \mathcal{P}(\mathcal{S})$ which allows for a density $p(\cdot \mid s, a)$ (i.e. $p \cdot \mu = P(\cdot \mid s, a)$), unless $P(\cdot \mid s, a) = \delta_{s'}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, which would lead to a quite boring environment. Generally the processes with fast convergence bounds according to corollary 5.12 must be very stochastic.

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

6 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 success 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 success, 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. However it is also possible that DQN can be seen largely as an instance of FQI and that the results we have covered from [5] does explain part of its success.

6.1 Further directions

6.1.1 Examples

In this paper we talk little about concrete decision processes, only applying our theory on the rather trivial gridworld environment in example 4.27. Examples are important drivers much of both theoretical and more empirical research in RL. Applying the theory in this paper on concrete cases would be a interesting next step.

6.1.2 Suboptimality of policies

This is relating to decision processes and value functions. Through out this 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 most others is satisfied with a bound on $\left\|Q^* - \widetilde{Q}\right\|_\infty$. To this end it could be interesting to establish relations between $\|Q^* - Q_{\widetilde{\pi}}\|_\infty$ and $\left\|Q^* - \widetilde{Q}\right\|_\infty$.

6.1.3 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 5.8). 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 to analyse.

6.2 Notes on references

The proofs on basic measure theory are inspired by ones found in [12, Rønn-Nielsen and Hansen (2014)] and [8, Kallenberg (2002)]. A good survey on results on optimal policy existence in the special case of Markov decision processes can be found in [6, Feinberg (2012)], however proofs in this source is either missing or sketched (as one must expect in a survey). A standard reference for optimal policies in MDPs is [2], which we also rely heavily upon in this paper.

6.3 Credits

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

A Appendices

A.1 Basic definitions and results

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. $X \in D$,

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

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

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

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

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

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

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

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

1. $\varnothing \in \mathcal{X}$

2. $S,\ T \in \mathcal{S} \implies S \cap T \in \mathcal{X}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof.

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

□

Definition A.17 (Lipschitz continuity)**.** Let $(\mathcal{X}, d_{\mathcal{X}})$, $(\mathcal{Y}, d_{\mathcal{Y}})$ be metric spaces. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to **Lipschitz** with constant $L > 0$ if

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

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

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

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

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

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

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

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

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

Theorem A.24 (Banach fixed point theorem)**.** Let $(\mathcal{X}, d)$ be a complete metric space and $T : \mathcal{X} \to \mathcal{X}$ be a contraction, i.e. $d(Tx, Ty) < \gamma d(x, y)$ for some $0 < \gamma < 1$ and all $x, y \in \mathcal{X}$. Then T has a unique fixed point x^* and for every $x \in \mathcal{X}$ it holds that $T^k x \to x^*$ as $k \to \infty$, with rate $d(T^k x, x^*) < \gamma^k d(x, x^*)$.

A.2 Proofs of auxiliary results

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

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

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

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

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

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

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

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

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

□

A.3 Disambiguation

- $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ set of extended real numbers.
- $\mathrm{id}_X := x \mapsto x$ the identity function on X.
- $[\phi] := \begin{cases} 1 & \phi \\ 0 & \neg\phi \end{cases}$: 0-1 indicator for logical formulas.
- $[q] := \{1, \dots, q\}$ for $q \in \mathbb{N}$.
- $1_A(a) := [a \in A]$: the indicator function.
- $C(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{K} \mid f \text{ continuous}\}$ set of continuous real-valued functions on $\mathcal{X}$.
- ANN: abbreviation for artificial neural network see definition 5.3.
- δ_a: Dirac-measure of point a, i.e. $\delta_a(A) = [a \in A] = 1_A(a)$.
- $(\Omega, \Sigma_\Omega, \mathbb{P})$: background probability space, that is source space for random variables.
- $\mathbb{B}_n$: the n-dimensional Borel σ-algebra.
- λ^n: the n-dimension Lebesgue measure.
- $\mathcal{P}(\mathcal{X})$: the set of all probability measures on a measurable space $(\mathcal{X}, \Sigma_\mathcal{X})$.
- $2^\mathcal{X}$: the powerset of the set $\mathcal{X}$.
- $\mathcal{M}(\mathcal{X}, \mathcal{Y})$: set of $\Sigma_\mathcal{X}$-$\Sigma_\mathcal{Y}$ measurable functions.
- $\mathcal{L}_p(\mathcal{X}) = \left\{f : \mathcal{X} \to \mathbb{R} \mid \int |f|^p \, \mathrm{d}\mu < \infty,\ \forall \mu \in A\right\}$ the set of functions which are (p, μ)-integrable for all measures μ in a certain set of measures A, see definition 4.7.
- $\mathbb{E}, \mathbb{E}_\mu$: expectation, that is integration w.r.t. the measure $\mathbb{P}$ or μ respectively.
- $\mathbb{V}$: variance operator.

References

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

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

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

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

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

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

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

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

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

[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] 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`.

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