Contents

Chapter 1

Decision models and value functions

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

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

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

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

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

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

The following example shows how probability kernels are easily constructed.

Example 1. If $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is positive a measurable function with the property that

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

then $\kappa(B \mid x) = \int_B f(x, y) \, \mathrm{d}\mu(y)$ defines a $\mathcal{X}$-probability kernel on $\mathcal{Y}$.

A handy property of kernels is

Proposition 1. Let $\kappa : \mathcal{X} \to \mathcal{Y}$ be a probability kernel and $f : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be measurable satisfying that $f(x, \cdot)$ is $\kappa(\cdot \mid x)$-integrable for every $x \in \mathcal{X}$. Then $x \mapsto \int f \, \mathrm{d}\kappa(\cdot \mid x)$ is measurable into $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$.

Proof. Simple functions are measurable since κ is a kernel. Now extend by sums and limits. □

We can now state the definition of a decision process

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

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

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

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

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

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

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

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. $A_n(h_n) \subseteq \mathcal{A}_n$ 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 3 (Policy)**.** A (randomized) **policy** $\pi = (\pi_n)_{n\in\mathbb{N}}$ for a HDP is a sequence of probability kernels $\pi_n : \mathcal{H}_n \rightsquigarrow \mathcal{A}_n$, such that $\pi_n(A(h_i) \mid h_i) = 1$ for alle $h_i \in \mathcal{H}_i$, i.e. the policy chooses only admissable actions (with probability 1). The set of all policies we denote $R\Pi$.

With a HDP, a starting state S_1 and a policy π intuitively we should be able to obtain a history by sampling

- an action $A_1 \in A(H_1)$ from $\pi_1(\cdot \mid H_1)$ (where $H_1 = S_1$),
- a state $S_2 \in P(H_2)$ from $P(\cdot \mid H_1, A_1)$,
- a reward $R_1 \in \mathbb{R}$ from $R_1(\cdot \mid H_2)$,
- an action $A_2 \in A(H_2)$ from $\pi_2(\cdot \mid H_2)$
- and so on.

To make this precise we need some additional measure theory on probability kernels.

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

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

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

Proof. 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

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

Using this definition we see that $\lambda(\mathcal{X} \times \mathcal{Y}) = 1$ and by monotone convergence for disjoint $G_1, G_2, \ldots$

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

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 1. In light of theorem 1 we can view a probability kernel as a mapping $\kappa : \mathcal{P}(X) \rightsquigarrow \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ defined by $\mu \mapsto \kappa\mu$.

Definition 4 (Composition of kernels)**.** Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\phi : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ be probability kernels. We define the composition $\phi\kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$ by

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

.

Remark 2. Following remark 1 $\phi\kappa$ can be viewed as a mapping a from from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{X} \times \mathcal{Z})$. This is somewhat unsatisfactory. We are missing the intermediary space $\mathcal{Y}$. However writing $\phi(\kappa\mu)$ we obtain a measure on $\mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ as wanted. We will therefore use a slight abuse of notation and interpret compositions of kernels as including all intermediary spaces, when viewed as maps of measures, that is we will write $\phi\kappa\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. It is a trivial exercise to verify that composition when viewed this way is associative. That is if $\psi : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \rightsquigarrow \mathcal{W}$ is another probability kernel, then $((\psi\phi)\kappa)\mu = (\psi(\phi\kappa))\mu$.

Remark 3. When one has $\varphi : \mathcal{Y} \to \mathcal{Z}$ we can also use definition 4 since φ can be viewed as a $\mathcal{X} \times \mathcal{Y}$-kernel which does not depend on its input from $\mathcal{X}$. We write $\varphi \circ \kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$. This is often also referred to a *composition of kernels*. In fact it makes the class of measurable spaces into a category [4, Lawvere (1962)], with identity $\mathrm{id}_{\mathcal{X}}(\cdot \mid x) = \delta_x$.

Remark 4. Let $(\mathcal{X}_n, \Sigma_{\mathcal{X}_n})_{n \in \mathbb{N}}$ be a sequence of measurable spaces. For each $n \in \mathbb{N}$ define $\mathcal{X}^{\underline{n}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $\Sigma_{\mathcal{X}^{\underline{n}}} := \Sigma_{\mathcal{X}_1} \otimes \cdots \otimes \Sigma_{\mathcal{X}_n}$ and let $\kappa_n : \mathcal{X}^{\underline{n}} \rightsquigarrow \mathcal{X}_{n+1}$ be a probability kernel. Then by remark 2 $\kappa^{\underline{n}} := \kappa_n \dots \kappa_1$ defines a map from $\mathcal{P}(\mathcal{X}_1)$ to $\mathcal{P}(\mathcal{X}^{\underline{n+1}})$.

Remark 2 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

$$P_{n-1}\pi_{n-1}R_{n-2}P_{n-2}\pi_{n-2} \dots R_2P_2\pi_2R_1P_1\pi_1\mu \in \mathcal{P}(\mathcal{H}_n)$$

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

Theorem 2 (Ionescu-Tulcea extension theorem)**.** For every $\mu \in \mathcal{P}(\mathcal{X}_1)$ there exists a unique probability measure $\rho \in \mathcal{P}(\mathcal{X}^{\underline{\infty}})$ such that

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

Proof. We refer to [3, Kallenberg (2002)] thm. 5.17. □

It is even possible to view the Ionescu-Tulcea construction as a kernel

Proposition 2 (Ionescu-Tulcea kernel)**.** Let μ_x denote the Ionescu-Tulcea measure of a sequence of probability kernels $\kappa_i : \mathcal{X}^{\underline{i}} \to \mathcal{X}_{i+1}$ with starting measure δ_x on $\mathcal{X}_1$ for any $x \in \mathcal{X}_1$. Then $\kappa(A \mid x) = \mu_x(A)$ defines a probability kernel $\kappa : \mathcal{X}_1 \to \mathcal{X}^{\underline{\infty}}$.

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

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

The cylinder algebra

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

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

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

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

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

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

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

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

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

Proof. Let $x \in \mathcal{X}_1$. Notice that by associativity of the composition of finitely many kernels $\kappa_n \ldots \kappa_1 \mu = (\kappa_n \ldots \kappa_2)(\kappa_1 \mu)$. This implies that

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

for all $n \in \mathbb{N}$ and $A \in \Sigma_{\mathcal{X}^{\underline{n}}}$. By the uniqueness in theorem 2 we are done. □

Let $\mu \in \mathcal{P}(\mathcal{S}_1)$ be a measure on the first state space. By theorem 2 a HDP and a policy π gives rise to a kernel $\kappa_\pi : \mathcal{P}(\mathcal{S}_1) \to \mathcal{P}(\mathcal{H}_\infty)$, namely

$$\kappa_\pi = \ldots R_2 P_2 \pi_2 R_1 P_1 \pi_1 \mu \tag{1.1}$$

In particular $\kappa_\pi \mu$ can be interpreted as the stochastic process arising from sampling the first state from μ and then follow π for a countable number of steps. We will denote expectation with respect to $\kappa_\pi \mu$ by $\mathbb{E}^\pi_\mu$. In the case where $\kappa_\pi \delta_s$ can be interpreted as the stochastic process arise from starting in state s and following policy π. We will abuse notation slightly, writing $\kappa_\pi \delta_s = \kappa_\pi s$ and $\mathbb{E}^\pi_{\delta_s} = \mathbb{E}^\pi_s$.

1.1 Policy evaluation and value functions

The next step is to evaluate how *good* a policy is. To this end we introduce *value functions*. In order for the sum of finitely many rewards to have a meaningful expected value we will need one of the following conditions:

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

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

Now the following definition makes sense

Definition 5 (Finite horizon value function)**.** Let $\underline{R}_i : \mathcal{H}_\infty \to \overline{\mathbb{R}}$ be the projection map onto the ith reward. We define the function $V_{n,\pi} : \mathcal{S}_1 \to \overline{\mathbb{R}}$ by

$$V_{n,\pi}(s_1) = \mathbb{E}_s^\pi \sum_{i=1}^{n} \underline{R}_i$$

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

The finite horizon value function measures the expected total reward of starting in state s and then follow the policy π for n steps. This way it measures the *value* of that particular state given a policy and *horizon* (number of steps).

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

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

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

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

Remark 5. The letters F^+, F^-, P, N and D are adopted from [1].

Definition 6. We define the infinite horizon value function by

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

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

Remark 6. Whenever we are working with the finite horizon value function we will always assume that either (F^+) or (F^-) holds without stating this explicitly. If a result only holds under e.g. (F^+) we will of course be explicit about this be marking it accordingly with a (F^+).

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

We mention some immediate properties of the finite and infinite horizon value functions

Proposition 3.

1. $V_{n,\pi}, V_\pi$ are measurable into $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$ and under (D) they are integrable with respect to $\kappa_\pi(\cdot \mid s)$ for any $\pi \in R\Pi$.

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

3. Under (D) for any $\pi \in R\Pi$

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

Proof.

1. Use proposition 1.

2. By monotone or dominated convergence.

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

□

Remark 7. As this bound will occur again and again we denote it

$$V_{\max} := R_{\max}(1-\gamma) \tag{1.2}$$

1.1.1 The optimal value function

Definition 7 (Optimal value functions)**.**

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

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

Proposition 4. Under (D) we have $\left|V_k^*\right|, |V^*| \leqslant V_{\max}$.

Proof. All terms in the suprema are within this bound. □

Remark 8. It is known that the optimal value function might not be Borel measurable (see ex. 2 p. 233 [1]). Perhaps this is not suprising since we are taking a supremum over sets of policies which might have cardinality of at least the continuum.

At this point some relevant questions can be asked.

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

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

3. When can optimal policies be chosen to be Markov, deterministic, etc.?

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

In a quite general setting, questions 1 and 2 is investigated in [5, Schäl (1975)]. Here some additional structure on our process is imposed.

Definition 8 (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. $V_\pi < \infty$ for all policies $\pi \in R\Pi$.
2. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})$ are all standard Borel.
3. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})$ are all standard Borel.
4. The set of admissible actions $A_n(h_n)$ is compact for any $h_n \in \mathcal{H}_n,\ n \in \mathbb{N}$.
5. The kernels $(P_n, R_n)_{n\in\mathbb{N}}$ are independent of rewards in the process.
6. $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N\geqslant n} \sup_{\pi\in R\Pi} \sum_{t=n+1}^{N} \mathbb{E}_s^\pi \underline{R}_n \to 0, \quad n \to \infty$

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

Condition S.

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

Condition W.

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

Theorem 3 (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 [5]. □

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

Proof. Since by theorem 3 there exists an optimal policy π^* we have $V^* = V_{\pi^*}$ which is measurable due to proposition 3. □

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

1.2 Markov decision processes

Definition 9 (Markov decision process)**.** A **Markov decision process** (MDP) consists of

1. $(\mathcal{S}, \Sigma_{\mathcal{S}})$ a measurable space of states.
2. $(\mathcal{A}, \Sigma_{\mathcal{A}})$ a measurable space of actions.
3. $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ a transition kernel.
4. $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow \overline{\underline{\mathbb{R}}}$ a reward kernel.
5. An optional disount factor $\gamma \in [0, 1]$ (when not discounting put $\gamma = 1$).
6. $A(s) \subseteq \mathcal{A}$ a set of admissable actions for each $s \in \mathcal{S}$.

This is a special case of the history dependent decision process (definition 2) 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, i.e. $P_n(\cdot \mid s_1, \ldots, s_n, a_n) = P(\cdot \mid s_n, a_n)$ for all $n \in \mathbb{N}$.
- R_n depends only on s_n and a_n and does not differ with n except for a potential discount. I.e. $R = R_n/\gamma^{n-1}$ for all $n \in \mathbb{N}$

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

Remark 9. Expectations over any reward R_i occuring in an MDP can be computed from a function $r : \mathcal{S} \times \mathcal{A} \to \overline{\underline{\mathbb{R}}}$, defined by $r(s, a) = \int r' \, \mathrm{d}R(r' \mid s, a)$. To see this note that

$$\begin{aligned}
\mathbb{E}^{\pi}_{\mu} \underline{R}_i &= \int \underline{R}_i \, \mathrm{d}\kappa_{\pi}\mu \\
&= \int r_i \, \mathrm{d}R_i P \pi_i \ldots R_1 P \pi_1 \mu(s_1, a_1, \ldots, s_{i+1}, r_i) \\
&= \int \gamma^{i-1} r(s_i, a_i) \, \mathrm{d}\pi_i \ldots R_1 P \pi_1 \mu(s_1, a_1, \ldots, s_i, a_i) \\
&= \mathbb{E}^{\pi}_{\mu} \gamma^{i-1} r(\underline{S}_i, \underline{A}_i)
\end{aligned}$$

where $\underline{S}_i, \underline{A}_i$ are projection onto the ith state and action.

Remark 10. One could ask if it is possible to embed a HDP into an MDP by setting $\mathcal{S} := \bigcup_{i \in \mathbb{N}} \mathcal{S}_i$ and $\mathcal{A} := \bigcup_{i \in \mathbb{N}} \mathcal{A}_i$ or similar. One attempt at this can be found in [1] chapter 10, but this will not be covered here. Note however that whatever 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.

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 10 (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 $\tau_1, \tau_2, \cdots : \mathcal{S} \rightsquigarrow \mathcal{A}$ such that $\pi_i(\cdot \mid s_1, \ldots s_i) = \tau_i(\cdot \mid s_i)$. We denote the set of (random) Markov policies by $M\Pi$. If $\tau_1 = \tau_2 = \ldots$ 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 11. We have the following inclusions of policy classes

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

Note that stationary policies might not exist in HDPs, but always exist for MDPs. A policy $(\pi_1, \pi_2, \dots)$ 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.

We will prove that in MDPs under mild assumptions the optimal policy π^* can be chosen Markov, and even stationary. Before we can do this we need some tools for studying MDPs.

Definition 11 (The T-operators)**.** For a stationary policy π and measurable $V : \mathcal{S} \to \overline{\mathbb{R}}$ we define the operators

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

where $T_a = T_{\delta_a}$ for $a \in A(s)$.

Remark 12. For the integral to make sense we assume under (D) that V is bounded, under (P) that $V \geqslant 0$ and under (N) that $V \leqslant 0$.

The T operator is sometimes called the Bellman-operator. It is harder to work with than T_π because it envolves a supremum. Therefore we will first take a closer look at properties of T_π.

Proposition 5 (Properties of the T_π-operator)**.** Let $\pi = (\pi_1, \pi_2, \dots)$ be a Markov policy.

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

Proof.

1. Measurability is by proposition 1 the rest follows by monotone or dominated convergence.
2. We have

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

   Now use this inductively.

3. This is by 2. and a monotone or dominated convergence.

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

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

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

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

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

□

1.2.1 Greedy policies

Definition 12. Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A} \in S\Pi$ be a stationary policy and let $V : \mathcal{S} \to \overline{\underline{\mathbb{R}}}$ be a measurable value-function. We define

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

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

In order to talk about existence of greedy policies we need some additional structure on our MDP.

Definition 13 (Borel σ-algebra)**.** For a topological space the **Borel** σ-algebra is the smallest σ-algebra containing all open sets.

Definition 14 (Weak topology)**.** Let $\mathcal{X}$ be a metrizable space equipped with the Borel σ-algebra. Consider the family of subsets of $\mathcal{P}(\mathcal{X})$

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

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 $\mathcal{V}$.

Recall (definition 8) that a measurable space is standard Borel if it is Polish and equipped with the Borel σ-algebra.

Proposition 6. Let $\mathcal{X}$ be a standard Borel measurable space. Consider the space $\mathcal{P}(\mathcal{X})$ of probabilty measures on $\mathcal{X}$ equipped with the weak topology. Then $\mathcal{P}(\mathcal{X})$ is standard Borel. If furthermore $\mathcal{X}$ is compact then $\mathcal{P}(\mathcal{X})$ is also compact.

Proof. We refer to [1] cor.7.25.1 and prop.7.22. □

Definition 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.

Definition 16 (Semicontinuity)**.** Let $\mathcal{X}$ be a topological space and $f : \mathcal{X} \to \overline{\underline{\mathbb{R}}}$ 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 $x \in U$. If $-f$ is upper semicontinuous, then f is **lower** semicontinuous.

Proposition 7.

1. If $f,\ g : \mathcal{X} \to \overline{\underline{\mathbb{R}}}$ are upper (lower) semicontinuous then $f + g$ is upper (lower) semicontinuous.
2. If furthermore g is continuous and non-negative then fg is upper (lower) semicontinuous.
3. 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.

Proposition 8. Let $\mathcal{X}$ and $\mathcal{Y}$ be separable metrizable and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a continuous stochastic kernel. Let $f : \mathcal{X} \times \mathcal{Y} \to \overline{\underline{\mathbb{R}}}$ 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. □

Proposition 9. A upper (lower) semicontinuous function $f : \mathcal{X} \to \overline{\underline{\mathbb{R}}}$ on a compact set $\mathcal{X}$ 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. □

Proposition 10. 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 $f : \Gamma \to \overline{\underline{\mathbb{R}}}$ be upper semicontinuous.

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

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

Setting 2.

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

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

Proposition 11. Under setting 2 let $V : \mathcal{S} \to \overline{\mathbb{R}}$ be upper semicontinuous and bounded from above. Then the following holds

1. For every $s \in \mathcal{S}$ we have that $(s, a) \mapsto T_a V(s)$ is upper semicontinuous and bounded from above.
2. For every $s \in \mathcal{S}$ we have that $G_V(s)$ is non-empty.
3. There exist a deterministic greedy policy τ_V.
4. $T_{\tau_V} V = TV$ so TV is measurable.

Proof.

1. This is a consequence of proposition 7 and proposition 8 since r is upper semicontinuous.
2. Since by 1. $(s, a) \mapsto T_a V(s)$ is upper semicontinuous, this follows by proposition 9.
3. By proposition 10 there exists a Borel-measurable function $\varphi : \mathcal{S} \to \mathcal{A}$ with $\mathrm{Gr}(\varphi) \subseteq \Gamma$ such that
$$T_{\varphi(s)} V(s) = \sup_{a \in A(s)} T_a V(s)$$
thus $\varphi(s) \in \operatorname{argmax}_{a \in A(s)} T_a V(s) = G_V(s)$. Therefore the induced deterministic policy
$$\tau_V(\cdot \mid s) = \delta_{\varphi(s)}$$
is greedy with respect to V.
4. By definition $T_{\tau_V} V(s) = T_{\operatorname{argmax}_{a \in A(s)} T_a V(s)} V(s) = \sup_{a \in A(s)} T_a V(s) = TV(s)$.

□

1.2.2 Existence of optimal policies

Proposition 12. Under setting 2 we have that
$$V_k^* = T^k V_0^* = T_{\tau_{V_{k-1}^*}} \dots T_{\tau_{V_0^*}} V_0^*$$
and this is an upper semicontinuous function. Thus $(\tau_{V_k^*}, \dots, \tau_{V_0^*})$ is a deterministic k-optimal policy for any $k \in \mathbb{N}$.

Proof. 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.

$$\begin{aligned}
V_k^*(s) &= \sup_{\pi \in R\Pi} \int \sum_{i=1}^{k} \gamma^{i-1} \underline{R}_i \, \mathrm{d}\kappa_\pi(\cdot \mid s) \\
&= \sup_{\pi \in R\Pi} \int r(s,a) + \gamma \left(\sum_{i=1}^{k-1} \gamma^{i-1} \underline{R}_i \, \mathrm{d}\kappa_{(\pi_2, \pi_3, \dots)}(\cdot \mid s, a, s') \, \mathrm{d}P(s' \mid s, a) \right) \mathrm{d}\pi_1(a \mid s) \\
&\leqslant \sup_{\pi_1 \in S\Pi} \int r(s,a) + \gamma \int V_{k-1}^* \, \mathrm{d}P(s' \mid s, a) \, \mathrm{d}\pi_1(a \mid s) \\
&= \sup_{\pi_1 \in S\Pi} T_{\pi_1} V_{k-1}^*(s) = TV_{k-1}^*(s)
\end{aligned}$$

Since s was arbitrary we must have $V_k^* \leqslant TV_{k-1}^*$. On the other hand by proposition 5 and induction hypothesis we have

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

But since $(\tau_{V_{k-1}^*}, \dots, \tau_{V_0^*})$ 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_a preverses this property (see proposition 11). □

Proposition 13. Under setting 2 and the last point in setting 1, that is the condition

$$\forall s \in \mathcal{S} : \sup_{N \geqslant n} \sup_{\pi \in R\Pi} \sum_{i=n+1}^{N} \mathbb{E}_s^\pi \underline{R}_i \to 0, \quad n \to \infty$$

it holds that $V^* = \lim_{k\to\infty} T^k V_0^*$. Furthermore under (D) the greedy policy τ_{V*} exists and is a deterministic stationary optimal policy.

Proof. Since setting 2 and the last point in setting 1 implies the rest of setting 1 and condition (S) we have by theorem 3 that $T^k V_0^* = V_k^* \to V^*$. We know by proposition 12 that V_k^* is semi uppercontinuous for all $k \in \mathbb{N}$. Under (D) we have that

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

So by proposition 7 the infimum $\inf_k \widehat{V}_k = V^* - V_{\max}$ is upper semicontinuous and thus V^* is upper semicontinuous. Therefore by proposition 11 there exists a deterministic greedy policy τ_{V*} which satisfies

$$T_{\tau_{V*}} V^* = TV^* \tag{1.3}$$

By proposition 5 (under (D)) 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 **??**) $V^* = \lim_{k\to\infty} T^k V_0^*$ is the unique fixed point of T in B. Again by Banach fixed point theorem and eq. (1.3) V^* is the fixed point of $T_{\tau_{V*}}$, which by proposition 5 also has $V_{\tau_{V*}}$ as fixed point. By uniqueness $V_{\tau_{V*}} = V^*$ and thus τ_{V*} is optimal. □

Remark 13. The property that $TV^* = V^*$ is often referred to as *Bellman's optimality equation.*

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

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

Proof. Since (D) implies the last point in setting 1 we can apply proposition 13. The last part is by the Banach fixed point theorem. □

1.2.3 Value iteration

Value iteration is a broad notion that can refer to many algorithms in dynamic programming, that somehow updates value functions. We here present perhaps to most basic of such algorithms, which is simply an iterative application of the T operator.

Algorithm 1: Simple value iteration

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations K, initial value function $\widetilde{V}_0$

$r \leftarrow \int x \, \mathrm{d}R(x \mid \cdot)$.

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

$\quad \widetilde{V}_{k+1} \leftarrow r + \gamma \int \sup_{a' \in \mathcal{A}} \widetilde{V}_k(s', a') \, \mathrm{d}P(s' \mid \cdot)$

Output: An estimator $\widetilde{V}_K$ of V^*

The results of the previous section, in particular corollary 2 is the theoretical foundation for *value iteration*.

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 (setting 2 and (D)), and therefore algorithm 1 could be applied in other cases if one has a practical way of representing the iterations $TV_0, T^2V_0, \ldots$. We here include an example in the finite case.

1.2.4 Q-functions

Throughout this section we assume setting 2, (D) and furthermore that $A(s)$ is finite for every $s \in \mathcal{S}$.

A **Q-function** is any function that assigns a (extended) real number to every state-action pair, that is any function $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$. Q-function are also called *action-value* functions, to distinguish them from the *value* functions we have discussed in the previous sections. 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.

Definition 17 (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 18 (Operators for Q-functions)**.** For any stationary policy $\tau \in S\Pi$ and measurable $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ with $Q \geqslant 0, Q \leqslant 0$ or $|Q| < \infty$ we define

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

$$T_\tau Q = r + \gamma P_\tau Q$$

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

where $T_a = T_{\delta_a}$.

Remark 14. The P_τ operator is defined for simplications in proofs, especially in the analysis of [2, Fan et al. (2020?)] in the later sections.

Definition 19 (Greedy policies for Q-functions)**.** Let $\tau : \mathcal{S} \rightsquigarrow \mathcal{A}$ be a stationary policy. Define $G_Q(s) = \operatorname{argmax}_{a \in A(s)} Q(s,a)$. If there exist a measurable set $G_Q^\tau(s) \subseteq G_Q(s)$ for every $s \in \mathcal{S}$ such that

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

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

The idea of Q-functions (and the letter Q) originates to [6, Watkins (1989)]. Upon the definition he notes

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

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

Proposition 14 (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. $\mathbb{E}_{\tau(\cdot|s)} Q_{k,\pi} = V_{k+1,(\tau,\pi)}$
2. $T_\tau Q_{k,\pi} = r + \gamma \mathbb{E} T_\tau V_{k,\pi}$
3. $\tau(V_{k,\pi}) = \tau(Q_{k,\pi})$ and $\tau(V_\pi) = \tau(Q_\pi)$, in particular τ_k^* and τ^* are greedy for Q_k^* and Q^*.
4. $\max_{a \in A(s)} Q^*(s,a) = V^*(s)$.

Proof.

1. This is essentially due to properties of the kernels. The idea is sketched here
$$T_\mu Q_{k,\pi} = r + \gamma \int r + \gamma V_{k,\pi} \, \mathrm{d}P \, \mathrm{d}\mu P = r + \gamma \int r + \gamma V_{k,\pi} \, \mathrm{d}P \mu \, \mathrm{d}P = r + \gamma \int T_\mu V_{k,\pi} \, \mathrm{d}P$$

□

Proposition 15 (Derived 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. $\lim_{k\to\infty} Q_{k,\pi} = Q_\pi$ and $\left|Q_{k,\pi}\right|, |Q_\pi|, \left|Q_k^*\right|, |Q^*| \leqslant V_{\max}$.
2. $Q_{k,\pi} = T_{\tau_1} \dots T_{\tau_k} Q_0$.
3. T, T_τ is γ-contractive on $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ and Q^*, Q_τ are the unique fixed points of T, T_τ in $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$.
4. $Q_k^* = r + \gamma \int V_k^*(s) \, \mathrm{d}P(s \mid \cdot)$ and $Q^* = r + \gamma \int V^*(s) \, \mathrm{d}P(s \mid \cdot)$. Furthermore Q^* and Q_k^* are upper semicontinuous.
5. $Q^* = Q_{\tau^*} = \lim_{k\to\infty} Q_k^*$.

Proof.

1. We see that $Q_k^* = \sup_{\pi \in R\Pi}(r + \gamma \mathbb{E} V_{k,\pi}) \leqslant r + \gamma \mathbb{E} V_k^* = r + \gamma \mathbb{E} V_{\pi_k^*} \leqslant Q_k^*$. Upper semicontinuity follows from proposition 8.

2. Follow the argument for 1.

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

4. By dominated convergence, 1., 2. and proposition 13 $\lim_{k\to\infty} Q_k^* = r + \gamma \lim_{k\to\infty} \int V_k^* \, \mathrm{d}P = \lim_{k\to\infty} V^* = Q^*$.

5. By 2. and the definition of $Q_{\pi *}$ we have $Q^* = r + \gamma \mathbb{E} V^* = r + \gamma \mathbb{E} V_{\pi *} = Q_{\pi *}$.

6. Use 1. iteratively starting with $\mu = \pi_1, \pi = (\pi_2, \pi_3, \dots)$.

7. By 2. $T_\pi Q_\pi = T_\pi(r + \gamma \mathbb{E} \lim_{k\to\infty} T_\pi^k V_0) = \lim_{k\to\infty} T_\pi (r + \gamma \mathbb{E} T_\pi^k V_0) = \lim_{k\to\infty} (r + \gamma \mathbb{E} T_\pi^{k+1} V_0) = r + \gamma \mathbb{E} \lim_{k\to\infty} T_\pi^{k+1} V_0 = r + \gamma \mathbb{E} V_\pi = Q_\pi$.

8. The contrativeness of T_π follows from the same argument as for value functions. 2. and Banach fixed point theorem does the rest.

9.

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

   To get $Q_k^* = T^k r$ use this inductively $Q_k^* = r + \gamma \mathbb{E} V_k^* = r + \gamma T V_{k-1}^* = T Q_{k-1}^* = \dots$. The statement $Q_k^* = T_{\pi_1^*} \dots T_{\pi_k^*} r$ is from proposition 14.

10. The argument from 1. also implies this first statement in 2. Now $TQ^* = r + \gamma \mathbb{E} TV^* = r + \gamma \mathbb{E} V^* = Q^*$ by **??**.

11. The argument is similar to **??** pt. 5.

□

Corollary 3. For any $Q \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ $T^k Q$ converges to Q^* with rate γ^k. That is

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

Proof. This is directly from □

1.2.5 Finite Q-iteration

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

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

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

then the algorithm becomes

Algorithm 2: Simple finite Q-iteration

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

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

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

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

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

Update action-value function:

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

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

Proposition 16. The output $\widetilde{Q}_K$ from algorithm 2 is K-optimal and $\left\|\widetilde{Q}_K - Q^*\right\|_\infty \leqslant \gamma^K \|Q^*\|_\infty$.

Proof. See **??**. □

1.2.6 Approximation

In this section we will look at what happens if we instead use approximations of the Q-functions and T operator. This means that we are in a setting where we can somehow calculate r and TQ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, but it is hard or infeasible to represent them (or at least one of them) directly. This setting is not very well-studied in the case of a continuous state space (at least in the sources known to this writer). This is perhaps because it is considered solved by the results of theoretical Q-learning presented in the previous section. However as we have argued, this only have practical relevance when it is feasible to represent TQ. Therefore we find it relevant to consider this setting in more detail. What *is* very well-studied is a further generalized setting where T and r are assumed to be unknown, that is, one has only access to their distributions via sampling from them. We will deal with this setting in the next section. In following we present some rather simple bounding techniques which is inspired by arguments found in e.g. [2], together with some standard results from approximation theory on artificial neural networks and Bernstein polynomials. Throughout this section we assume (D) i.e. that we are discounting with some $\gamma \in [0, 1)$.

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

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

First observe that

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

Using this iteratively we get

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

Then we can bound

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

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

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

$$\varepsilon_{\text{approx}}(k) = \varepsilon\frac{1-\gamma^k}{1-\gamma} \leqslant \frac{\varepsilon}{1-\gamma} \tag{1.4}$$

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 17. $\sum_{i-1}^{k} \gamma^{k-i}\varepsilon_i \to 0$ 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 $m > M$

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

□

Using artifical neural networks

Setting 3. An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0,1]^w$ and $\mathcal{A}$ finite. Assume that r is continuous and P is setwise-continuous.

Definition 20. An **ANN** (Artificial Neural Network) with structure $(d_i)_{i=0}^{L+1} \subseteq \mathbb{N}$, activation functions $\sigma_i = (\sigma_{ij})_{j=1}^{d_i}$, where $\sigma_{ij} : \mathbb{R} \to \mathbb{R}$ are real-valued functions on $\mathbb{R}$, and weights $W_i \in M^{d_i \times d_{i-1}}$, $v_i \in \mathbb{R}^{d_i}$, $i \in [L+1]$ is the function $F : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$

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

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

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

We denote the class of these networks (or functions)

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

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

Theorem 4 (Universal Approximation Theorem for ANNs)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be non-constant, bounded and continuous activation function. Let $\varepsilon > 0$ and $f \in C([0,1]^w)$. Then there exists an $N \in \mathbb{N}$ and a network $F \in \mathcal{DN}(\sigma, (w, N, 1))$ with one hidden layer and activation function σ such that

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

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

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

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

This was extended in [**?** , **?** (**?**)] to the statement as presented above and their proof is constructive. □

Proposition 18. Consider setting 3 let and $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-constant, bounded, continuous activation function. Let $\varepsilon > 0$. Then for every $k \in \mathbb{N}$ there exists a $N \in \mathbb{N}$ and a sequence of Q-networks $(\widetilde{Q}_i)_{i=1}^k \subseteq \mathcal{DN}(\sigma, \{w|\mathcal{A}|, N, 1\})$ such that

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

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

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

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

Using Bernstein polynomials

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

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

Definition 21 (Bernstein polynomial)**.** The multivariate Bernstein polynomial $B_{f,n}$ with exponents $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ approximating the function $f : [0,1]^w \to \mathbb{R}$ is defined by

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

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

Theorem 5. Let $f : [0,1]^w \to \mathbb{R}$ be Lipschitz (see **??**) w.r.t. the standard euclidean 2-norm induced metrics on $[0,1]^w$ and $\mathbb{R}$ with constant L. Then for any $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ there exists a polynomial $B_{f,n} : [0,1]^w \to \mathbb{R}$ of degree $\leqslant \|n\|_1$ such that

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

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

Now we can bound

Proposition 19.

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

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

Proposition 20.

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

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

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

Bibliography

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

[2] 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`.

[3] 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`.

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

[5] 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`.

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