0.1 General notes

In RL there is a categorization of algorithms into *off-policy* and *on-policy* classes. This is simply whether the algorithm learns from data (states, actions and rewards) arising from following its own policy (on-policy) or it can learn from more arbitrary data (off-policy). This *more arbitrary data* could for example be the trajectory of another algorithm when interacting with a decision process, or simply state-action-reward pairs drawn from some distribution. In this paper we exclusively consider off-policy algorithms.

0.2 Measure Theory

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

0.2.1 Kernels

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

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

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

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

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

We now state some fundamental results on probability kernels

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

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

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

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

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

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

Theorem 2 (Extended Tonelli and Fubini)**.** Let $\mu \in \mathcal{P}(\mathcal{X})$, $f \in \mathcal{M}(\Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y}, \mathbb{B})$ be a measurable function and $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a probability kernel. Then

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

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

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

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

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

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

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

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

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

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

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

0.2.2 Kernel derived processes

Let $(\mathcal{X}_n, \Sigma_{\mathcal{X}_n})_{n\in\mathbb{N}}$ be a sequence of measurable spaces. For each $n \in \mathbb{N}$ define $\mathcal{X}^{\underline{n}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $\Sigma_{\mathcal{X}^{\underline{n}}} := \Sigma_{\mathcal{X}_1} \otimes \cdots \otimes \Sigma_{\mathcal{X}_n}$ and let $\kappa_n : \mathcal{X}^{\underline{n}} \rightsquigarrow \mathcal{X}_{n+1}$ be a probability kernel. Then $\kappa^{\underline{n}} := \kappa_n \ldots \kappa_1$ is a kernel from $\mathcal{X}_1$ to $\mathcal{X}^{\underline{n}}$. So for any probability measure $\rho_1 \in \mathcal{P}(\mathcal{X}_1)$ there exists a unique probability measure ρ_n on $\mathcal{X}^{\underline{n}}$ defined by $\kappa^{\underline{n}}\rho_1$.

Let $\mathcal{X}^{\underline{\infty}} := \prod_{n\in\mathbb{N}} \mathcal{X}_n$ and $\Sigma_{\mathcal{X}^{\underline{\infty}}} := \bigotimes_{n\in\mathbb{N}} \Sigma_{\mathcal{X}_n}$. We are not equipped to establish existence of a kernel generated measure on $(\mathcal{X}^{\underline{\infty}}, \Sigma_{\mathcal{X}^{\underline{\infty}}})$ yet which we will need. This problem was solved by Cassius Ionescu-Tulcea in 1949:

Theorem 3 (Ionescu-Tulcea extension theorem)**.** For every $\mu \in \mathcal{P}(\mathcal{X}_1)$ there exists a unique probability measure $\rho \in \mathcal{P}(\mathcal{X}^{\underline{\infty}})$ such that

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

We denote this measure $\ldots \kappa_2\kappa_1\mu = \prod_{i=1}^{\infty} \kappa_i\mu := \rho$.

Proof. Todo: what about this. □

Proposition 3 (Ionescu-Tulcea kernel)**.** Let μ_x denote the Ionescu-Tulcea measure of a sequence of probability kernels $\kappa_i : \mathcal{X}^{\underline{i}} \to \mathcal{X}_{i+1}$ with starting measure δ_x on $\mathcal{X}_1$ for any $x \in \mathcal{X}_1$. Then $\kappa(A \mid x) = \mu_x(A)$ defines a probability kernel $\kappa : \mathcal{X}_1 \to \mathcal{X}^{\underline{\infty}}$.

Proof. Since we already know that μ_x is a probability measure for any $x \in \mathcal{X}_1$, we just have to show that $\kappa(A \mid x) = \mu_x(A)$ is measurable for all $A \in \bigotimes_i \Sigma_{\mathcal{X}_i}$. …todo □

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

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

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

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

Proposition 4. Let $\mathcal{X}, \mathcal{Y}$ be separable and metrizable, $\kappa : \mathcal{X} \to \mathcal{Y}$ be a continuous probability kernel and $f : \mathcal{X} \times \mathcal{Y} \to \overline{\underline{\mathbb{R}}}$ be Borel-measurable satisfying one of $f \leqslant 0, f \geqslant 0, |f| < \infty$. If f is bounded from above (below) and upper (lower) semicontinuous then

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

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

Proof. We refer to [BS SOC, prop. 7.31]. □

0.3 General decision model

In the quest to have a united framework to talk about results from several different models we define here a quite general model. One which is quite close to in generality can be found in [ref. to Schal]. In this section recall that $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ and $\overline{\underline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$.

Definition 2 (Decision model)**.** A general **decision** model is determined by

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

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

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

with associated product σ-algebras

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

Some authors refers to this as a *dynamic progamming* model. Notice the slight irregularity in the beginning of the history spaces: We are missing a reward state after $\mathcal{S}_1$. We could avoid this by introducing some start reward, but we will be careless.

The vast majority of sources considered in this paper actually specialize the DP model with the following:

Assumption 1 (One state and action space)**.** $\mathcal{S}_1 = \mathcal{S}_2 = \ldots := \mathcal{S}$ $\mathcal{A}_1 = \mathcal{A}_2 = \ldots := \mathcal{A}$

However we will do without this for the rest of this section in order to present some results in the generality they deserve. One could ask if it is possible to embed the general DP model into one with assumption 1 by setting $\mathcal{S} := \bigcup_{i \in \mathbb{N}} \mathcal{S}_i$ and $\mathcal{A} := \bigcup_{i \in \mathbb{N}} \mathcal{A}_i$ or similar. One attempt at this can be found in [BS SOC, chp. 10], but this will not be covered here.

For a DP model we can define

Definition 3 (Policy)**.** A (randomized) **policy** $\pi = (\pi_n)_{n \in \mathbb{N}}$ is a sequence of $\mathcal{H}_n \rightsquigarrow \mathcal{A}_n$ kernels. The set of all policies we denote $R\Pi$. The policy π is called **semi Markov** if each π_i only depends on the first and last state in the history and is called **Markov** if only the last. The sets are denoted $sM\Pi$ and $M\Pi$. SFurthermore π is called **deterministic** if all π_i are degenerate, i.e. are actually measurable functions from $\mathcal{H}_n$ to $\mathcal{A}_n$. Under assumption 1 it makes sense to make a (Markov) policy $(\pi, \pi, \ldots)$, where π only depends on the last state. Such a policy is called **stationary**, and the set of them denoted $S\Pi$.

We have the following inclusions

$$S\Pi \subseteq M\Pi \subseteq sM\Pi \subseteq R\Pi$$
$$DS\Pi \subseteq DM\Pi \subseteq DsM\Pi \subseteq D\Pi$$

Proposition 5. A dynamic progamming model together with a policy π defines a probability kernel $\kappa_\pi : \mathcal{S}_1 \to \mathcal{H}_\infty$.

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

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

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

Condition F^+. $R_i(\{\infty\} \mid h) = 0$ for all $h \in \mathcal{H}_{i+1}$ and $i \in \mathbb{N}$

Condition F^-. $R_i(\{-\infty\} \mid h) = 0$ for all $h \in \mathcal{H}_{i+1}$ and $i \in \mathbb{N}$

When assuming either of (F^+) or (F^-) adding rewards cannot lead to a $\infty - \infty$ situation, and the following definition makes sense

Definition 4. Let $\mathcal{R}_i : \mathcal{H}_\infty \to \overline{\mathbb{R}}$ be the projection onto the ith reward. Define

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

called the kth finite **ideal** value function. When $n = 0$ $\forall \pi : V_{0,\pi} = V_0 := 0$.

These are also called *finite horizon* value functions.

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

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

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

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

Definition 5. Assume (P), (N) or (D). We define the (infinite horizon) ideal value function by

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

Proposition 6. The ideal value functions $V_{n,\pi}, V_\pi$ are measurable into $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$.

Proof. Use proposition 2. □

Proposition 7. Under P, N or D we have $\lim_{n\to\infty} V_{n,\pi} = V_\pi$ for all $\pi \in R\Pi$.

Proof. By monotone or dominated convergence. □

0.3.1 Optimal policies

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

Assumption 2. (Reward independence) P_n, R_n and policies are only allowed to depend on the states and actions.

In all sources known to this writer assumption 2 is assumed. This is a bit of a puzzle since it is obvious that one could want to define algorithms (policies) that take into account which rewards they received in the past. We will also do this but stick to the standard and never attempt to evaluate ideal value functions of policies that depend on rewards. Thus we will assume assumption 2 henceforth with including the shrinkage of the set of general policies $R\Pi$ that it entails.

A neat consequence of assumption 2 when talking about value functions is that we can reduce the reward kernels to functions $r_i : \mathcal{H}_{i+1} \to \mathbb{R} = h \to \int r \mathrm{d}R_i(r \mid h)$ which are measurable (due to proposition 2).

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

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

are called the **optimal** value functions. 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.

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

At this point many interesting questions can be asked.

1. To which extend does an optimal policy π^* exist?
2. Does V_n^* converge to V^*?
3. In case there is some sort of optimal policy in which classes of policies has a representative?

These questions has been answered in a variety of settings. We will address these question in order by strength of assumptions they require as far as this is possible.

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

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

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

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

Condition S. 1. The function

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

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

2. r_n is upper semi-continuous.

Condition W. 1. The function

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

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

2. r_n is continuous.

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

0.4 Markov Decisions Process

Setting 2 (Markov Decision Process).

- Assumption 1 i.e. there is only one state and action space $\mathcal{S}$, $\mathcal{A}$.
- P_n depends only on s_n and a_n and does not differ with n. I.e. there exists a kernel P such that $P_n(\cdot \mid s_1, \ldots, s_n, a_n) = P(\cdot \mid s_n, a_n)$ for all $n \in \mathbb{N}$. We will write P instead of P_n understanding kernel compositions as if using P_n.
- r_n depends only on s_n and a_n and does not differ with n except for a potential discount. I.e. there exists a measurable function $r : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ such that $r = r_n/\gamma^{n-1}$ for all $n \in \mathbb{N}$ (in the case where we are not discounting set $\gamma = 1$).

0.5 Bertsekas-Shreve framework

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

Setting 3 (BS).

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

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

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

Notice that setting 3 implies (F^+). Throughout this section (P), (N) or (D) are always assumed. Some results only hold for some of these conditions and we will indicate this by e.g. (D) (P) when the result only holds for the discounted and positive case. At this point it makes sense to define

Definition 7 (The T-operators)**.** For a stationary policy π and measurable $V : \mathcal{S} \to \overline{\mathbb{R}}$ with $V \geqslant 0$, $V \leqslant 0$ or $|V| < \infty$ we define the operators

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

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

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

where $T_a = T_{\delta_a}$.

Proposition 8. The operators P_π, T_π and T commutes with limits.

Proof. By monotone or dominated convergence theorems. □

Proposition 9. Let $\pi = (\pi_1, \pi_2, \dots)$ be a Markov policy. Then $V_{k,\pi} = T_{\pi_1} \dots T_{\pi_k} V_0$ and $V_\pi = \lim_{k\to\infty} T_{\pi_1} \dots T_{\pi_k} V_0$.

Proof. todo □

Proposition 10. Let π be a stationary policy then $T_\pi V_\pi = V_\pi$.

Proof. By proposition 9 $T_\pi V_\pi = T_\pi \lim_{k\to\infty} T_\pi^k V_0 = \lim_{k\to\infty} T_\pi^{k+1} V_0 = V_\pi$. □

Proposition 11. Under (D) for any $\pi \in R\Pi$ we have

$$\left|V_{n,\pi}\right|, \left|V_\pi\right|, \left|V_k^*\right|, \left|V^*\right| \leqslant V_{\max} := R_{\max}/(1-\gamma)$$

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

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

This also covers $V_{n,\pi}$. □

Proposition 12. (D)
T and T_π are γ-contractive on $\mathcal{L}_\infty(\mathcal{S})$.

Proof. Let $V, V' \in \mathcal{L}_\infty(\mathcal{S})$ and let $K = \left\|V - V'\right\|_\infty$. Then

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

For T use that 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 $f, g : X \to \overline{\mathbb{R}}$. □

Corollary 1. (D)
Let $\pi \in S\Pi$ be a stationary policy. V_π is the unique bounded fixed point of T_π in $\mathcal{L}_\infty(\mathcal{S})$.

Proof. By proposition 10 V_π is a fixed point. By proposition 12 and Banach fixed point theorem we get uniqueness. □

Proposition 13 (Prop. 8.6 in BS)**.** $V_k^* = T^k V_0$ and is upper semicontinuous. Furthermore for any $k \in \mathbb{N}$ there exists a deterministic, Markov, Borel-measurable k-optimal policy $\pi_k^* = (\pi_{k,1}^*, \pi_{k,2}^*, \dots, \pi_{k,k}^*, \dots) \in DM\Pi$. These policies satisfy for any $i < k$ $\pi_i^* = (\pi_{k,k-i}^*, \dots, \pi_{k,k}^*, \dots)$.

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

Proposition 14. $V^* = T_{\pi^*} V^* = TV^*$
(D) V^* is the unique fixed point of T in $\mathcal{L}_\infty(\mathcal{S})$.

Proof. Since π^* is optimal $V^* = V_{\pi^*} = T_{\pi^*} V_{\pi^*}$ by proposition 10. By theorem 5 $TV^* = T \lim_{k\to\infty} T^k V_0 = \lim_{k\to\infty} T^{k+1} V_0 = V^*$. If (D) holds $V^* \in \mathcal{L}_\infty(\mathcal{S})$ so proposition 12 and Banach fixed point theorem ensures uniqueness. □

0.5.1 Analytic setting

Setting 4 (BS Analytic)**.** The same as setting 3 except: P is not necessarily continuous. r is upper semianalytic. $\mathcal{A}$ is not necessarily compact, but there exists a $k \in \mathbb{N}$ such that $\forall \lambda \in \mathbb{R}, n \geqslant k, s \in \mathcal{S}$

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

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

Theorem 6 (Prop. 9.17 BS)**.** Under setting 4 we have $V^* = \lim_{n\to\infty} V_n^*$ for all $s \in \mathcal{S}$ and there exists a optimal policy π^* which is stationary and deterministic.

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

0.6 Q-functions

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

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

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

Definition 8. Let $\pi \in R\Pi$. Define

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

We define $Q_{0,\pi} = r + \gamma \mathbb{E} V_0 = r := Q_0$.

Proposition 15. $\lim_{k \to \infty} Q_{k,\pi} = Q_\pi$

Proof. (D) By dominated convergence or (P/N) by monotone convergence theorem. □

Definition 9.

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

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

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

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

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

where $T_a = T_{\delta_a}$.

Proposition 16. Let $\pi = (\pi_1, \ldots, \pi_k, \ldots) \in M\Pi$ then

- $T_{\pi_k} Q_{k-1,\pi} = r + \gamma \mathbb{E} T_{\pi_k} V_{k-1,\pi}$
- $Q_{k,\pi} = T_{\pi_1} \ldots T_{\pi_k} r$.

Proof. The first statement is almost by definition. For the second use the first inductively. □

Proposition 17. $Q_k^* = r + \gamma \mathbb{E} V_k^*$ and $Q^* = r + \gamma \mathbb{E} V^*$.

Proof. $Q^* = \sup_\pi Q_\pi = \sup_\pi (r + \gamma \mathbb{E} V_\pi) = r + \gamma \sup_\pi \mathbb{E} V_\pi = r + \gamma \mathbb{E} V^*$ where in the fourth equality we have used that $V^* \geqslant V_\pi$ uniformly so $\sup_\pi \mathbb{E} V_\pi \leqslant \mathbb{E} V^*$ while trivially $\mathbb{E} V^* = \mathbb{E} V_{\pi^*} \leqslant \sup_\pi \mathbb{E} V_\pi$. For Q_k^* the argument is similar. □

Proposition 18. $\sup_{a \in \mathcal{A}} Q^*(s,a) = V^*(s)$

Proof. This is by definition after considering proposition 17. □

Proposition 19. $TQ_k^* = r + \gamma \mathbb{E} T V_k^*$ and if $\pi^* = (\pi_1^*, \ldots, \pi_k^*, \ldots)$ is k-optimal then $Q_k^* = T_{\pi_1^*} \ldots T_{\pi_k^*} r = T^k r$.

Proof.

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

To get $Q_k^* = T^k r$ use this inductively $Q_k^* = r + \gamma \mathbb{E} V_k^* = r + \gamma T V_{k-1}^* = TQ_{k-1}^* = \dots$. For $Q_k^* = T_{\pi_1^*} \dots T_{\pi_k^*} r$ use $Q_k^* = r + \gamma \mathbb{E} V_k^* = r + \gamma T_{\pi_1^*} \dots T_{\pi_k^*} V_0$ and first statement in proposition 16 inductively. □

The proof of proposition 19 also shows

Proposition 20. $TQ^* = r + \gamma \mathbb{E} T V^*$.

implying

Proposition 21. $TQ^* = Q^*$.

Proposition 22. For stationary $\pi \in S\Pi$ we have $T_\pi Q_\pi = Q_\pi$.

Proof. Using proposition 16 and proposition 10 $T_\pi Q_\pi = T_\pi(r + \gamma \mathbb{E} \lim_{k\to\infty} T_\pi^k V_0) = \lim_{k\to\infty} T_\pi(r + \gamma \mathbb{E} T_\pi^k V_0) = \lim_{k\to\infty}(r + \gamma \mathbb{E} T_\pi^{k+1} V_0) = r + \gamma \mathbb{E} \lim_{k\to\infty} T_\pi^{k+1} V_0 = r + \gamma \mathbb{E} V_\pi = Q_\pi$. □

Proposition 23. $Q^* = Q_{\pi*}$ and $T_{\pi*} Q^* = Q^*$.

Proof. $Q^* = r + \gamma \mathbb{E} V^* = r + \gamma \mathbb{E} V_{\pi*} = Q_{\pi*}$ for the second statement use proposition 22. □

Proposition 24. $Q^* = \lim_{k\to\infty} Q_k^*$.

Proof. By monotone or dominated convergence and theorem 5. □

Proposition 25. T and T_π is γ-contractive on $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$. If furthermore (D) holds, then $|Q_{k,\pi}|, |Q_\pi|, |Q_k^*|, |Q^*| \leqslant V_{\max}$ and Q_π, Q^* are the unique fixed points of T_π, T in $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$.

Proof. The contrativeness of T, T_π follows from the same argument as in proposition 12. If (D) holds the boundedness of the Q functions follow from an argument similar to the proof of proposition 11. Then proposition 21, proposition 22 and Banach fixed point theorem implies uniqueness. □

Proposition 26. (D)
For any $Q \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ $T^k Q$ converges to Q^* with rate γ^k.
That is $\|T^k Q - Q^*\|_\infty \leqslant \gamma^k \|Q - Q^*\|_\infty$.

Proof. By proposition 25 T γ-contracts so

$$\begin{aligned}
\left\|T^k Q - Q^*\right\|_\infty &= \left\|T^k Q - T^k Q^*\right\| \\
&\leqslant \gamma^k \|Q - Q^*\|
\end{aligned}$$

□

Proposition 27. (D)(N)
$Q^*(s,\cdot)$ is upper semicontinuous.

Proof. Since P is continuous and V^* is upper semicontinuous by theorem 5 the proposition follow by proposition 4. □

Definition 11. Let $\pi : \mathcal{S} \to \mathcal{A}$ be a stationary policy. If

$$\pi\left(\operatorname*{argmax}_{a \in \mathcal{A}} Q(s,a) \,\middle|\, s \right) = 1$$

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

Proposition 28. (D)(N)
Let $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ be measurable and upper semicontinuous in the second entry. Then there exists a deterministic greedy policy for Q.

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

Proposition 29. For any $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ if π_Q is greedy with respect to Q then $T_{\pi_Q} Q = TQ$.

Proof.

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

□

Proposition 30. Let π_i be greedy with respect to Q^*_{i-1} for $k \in \mathbb{N}$. Then $Q^*_k = T_{\pi_1} \ldots T_{\pi_k} Q_0$ for all $k \in \mathbb{N}$.

Proposition 31. (D)(N)
Any greedy policy with respect to Q^* is optimal and can be chosen to be deterministic.

Proof. By proposition 28 we can pick a greedy policy π for Q^* which can be chosen to be deterministic but let π stay general. Then by proposition 29 $T_\pi Q^* = TQ^*$ and by proposition 23 $Q_\pi = Q^*$ implying that π is optimal. □

```
Algorithm 1: Simple theoretical Q-iteration
  Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$
  $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : r(s,a) \leftarrow \int x \mathrm{d}R(x \mid s,a)$.
  $\widetilde{Q}_0 \leftarrow r$
  for $k = 0, 1, 2, \ldots, K-1$ do
      $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \widetilde{Q}_{k+1}(s,a) \leftarrow r(s,a) + \gamma \int \sup_{a' \in \mathcal{A}} \widetilde{Q}_k(s',a') \mathrm{d}P(s' \mid s,a)$
  Define $\pi_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
  Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\pi_K$
```

0.6.1 Finite Q-iteration

Concluding on the results so far we have showed how if one knows the process dynamics of a stationary decision process satisfying rather broad criteria, such as continuity and compactness, the optimal policy and state-value function can be found simply by iteration over the T-operator and picking a greedy strategy (see proposition 26). Of course this is practical computationally, only if the resulting Q functions can be represented and computed in finite space and time. This is trivially the case when

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

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

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

then the algorithm becomes

Algorithm 2: Simple finite Q-iteration

```
Input: DP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$
Set $r \leftarrow \left(\int r \mathrm{d}R(\cdot \mid s_1, a_1), \ldots, \int r \mathrm{d}R(\cdot \mid s_k, a_\ell)\right)^T$
and $\widetilde{Q}_0 = r$.
for $k = 0, 1, 2, \ldots, K-1$ do
    Set $m(\widetilde{Q}_k) := (\max_{a \in \mathcal{A}} Q(s_1, a), \ldots, \max_{a \in \mathcal{A}} Q(s_k, a))^T$
    Update action-value function:
        $\widetilde{Q}_{k+1} \leftarrow r + \gamma P m(\widetilde{Q}_k)$
Define $\pi_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\pi_K$
```

Proposition 32. The output $\widetilde{Q}_K$ from algorithm 2 is K-optimal and $\left\|\widetilde{Q}_K - Q^*\right\|_\infty \leqslant \gamma^K \|Q^*\|_\infty$.

Proof. See proposition 26 and proposition 30. □

0.7 Approximation

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

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

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

Then we can bound

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

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

Proposition 33. $\sum_{i-1}^k \gamma^{k-i}\varepsilon_i \to 0$ 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 $m > M$

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

□

0.8 Hidden dynamics

In this section we will look at what can be done when the process dynamics are unknown. In this case we cannot calculate directly neither r, $T_\pi Q$ nor TQ because the transition and reward kernels P, R are unknown.

It is clear that algorithm 1 will not work without modification in this case. Simply because R and P are not available. To make the scheme work anyway we could simply avoid taking expectations and use the random outcomes of the kernels. Leading to

Algorithm 3: Random theoretical Q-iteration (example of thought)

```
Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$
$\forall(s,a) \in \mathcal{S} \times \mathcal{A} : \widetilde{Q}_0(s,a) \leftarrow X \sim R(\cdot \mid s,a)$.
for $k = 0, 1, 2, \dots, K-1$ do
    $\forall(s,a) \in \mathcal{S} \times \mathcal{A} : \widetilde{Q}_{k+1}(s,a) \leftarrow r' + \gamma \sup_{a' \in \mathcal{A}} \widetilde{Q}_k(s', a')$
    where $r' \sim R(\cdot \mid s,a), s' \sim P(\cdot \mid s,a)$.
Define $\pi_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\pi_K$
```

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

0.8.1 Finite case

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

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

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

We will now look at a convergence result obtained by [Jaakkola, Jordan, Singh, 1993] of a TD algorithm using Q-functions

Algorithm 4: Simple Q-learning

```
Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ such that $|\mathcal{S}||\mathcal{A}| < \infty$, number of iterations $K$, state-action
       pairs $(s_1, a_1, \dots, s_K, a_K)$, learning rates $(\alpha'_1, \dots, \alpha'_K)$, initial $\widetilde{Q}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
Put $\alpha_k = \delta_{(s_k, a_k)} \alpha'_k$.
for $k = 1, 2, \dots, K$ do
    Sample $r' \sim R(\cdot \mid s_k, a_k), s' \sim P(\cdot \mid s_k, a_k)$
    Update action-value function:
        $\widetilde{Q}_k \leftarrow \widetilde{Q}_{k-1} + \alpha_k(r' + \max_{a' \in \mathcal{A}} \widetilde{Q}_{k-1}(s', a'))$
Define $\pi_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\pi_K$
```

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

Theorem 7. (Jaakkola, Jordan, Singh) Let $s_1, a_1, s_2, a_2, \dots \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A} \times \dots$ be random variables, and $\alpha_1, \alpha_2, \dots \in [0,1]$. The output $\widetilde{Q}_K$ of algorithm 4 converges to Q^* provided

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

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

0.8.2 Perspectives

Another approach could be to estimate R and P before or while using an algorithm like algorithm 1 using the estimated kernels. I was not able to find sources that did this, however you can argue that this idea is already employed in temporal difference learning and others.

1 Appendix

Definition 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 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

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