0.1 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.

0.1.1 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 $A_0 := \left\{ x \in \mathcal{X} \;\middle|\; \int f \mathrm{d}\kappa(\cdot \mid x) < \infty \right\} \in \Sigma_\mathcal{X}$ with $\mu(A_0) = 1$,

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

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

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

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. □

0.1.2 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 $\kappa^{\underline{n}} := \kappa_n \dots \kappa_1$ 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 $\dots \kappa_2\kappa_1\mu = \prod_{i=1}^{\infty} \kappa_i\mu := \rho$.

Proof. Todo: what about this. □

Proposition 3. 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 measure satisfies $\prod_{i=1}^{\infty} \kappa_i = \prod_{i=2}^{\infty} \kappa_i\kappa_1$.

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

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

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

0.2 Dynamic programming

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 (DP model)**.** A general **dynamic programming** 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

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

with associated product σ-algebras

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

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

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)**.** $\mathcal{S}_1 = \mathcal{S}_2 = \dots := \mathcal{S}$ $\mathcal{A}_1 = \mathcal{A}_2 = \dots := \mathcal{A}$

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 $\mathcal{S} := \mathcal{S}^{\infty}$ and $\mathcal{A} := \mathcal{A}^{\infty}$. I was not able to find discussions about this in litterature, and chose not to pursue this end.

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 $M\Pi$. Furthermore 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 $(\pi, \pi, \dots)$, 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 4. A dynamic progamming model together with a policy π defines a probability kernel $\kappa_\pi : \mathcal{S}_1 \to \mathcal{H}_\infty$.

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

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

In litterature the terminology varies and 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 (usually) 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 (Ideal value functions)**.** Let $r_i : \mathcal{H}_\infty \to \overline{\mathbb{R}}$ be the projection onto the ith reward. Define

$$V_{n,\pi}(s) = \mathbb{E}^\pi_s \sum_{i=1}^{n} r_i, \qquad V_\pi(s) = \mathbb{E}^\pi_s \limsup_{n\to\infty} \sum_{i=1}^{n} r_i$$

called the **ideal** value functions.

Proposition 5. The ideal value functions $V_{n,\pi}, V_\pi$ are measurable into $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$.

Proof. Use proposition 2. □

Note that for pointwise convergence of $V_{n,\pi}$ to V_π we need something like the assumptions for monotone or dominated convergence. To this end we introduce

Condition P. $R_i \in [0, \infty], \forall i \in \mathbb{N}$

Condition N. $R_i \in [-\infty, 0], \forall i \in \mathbb{N}$

Condition D. There exist a bound $R_{\max} > 0$ and a $\gamma \in [0, 1)$ called the **discount** factor such that $R_i \in [-R_{\max}\gamma^i, R_{\max}\gamma^i]$ for all $i \in \mathbb{N}$.

Proposition 6. 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. □

0.2.1 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 $r_i : \mathcal{H}_{i+1} \to \underline{\mathbb{R}} = h \to \int r \mathrm{d}R_i(r \mid h)$ which are measurable (due to proposition 2).

Definition 5 (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.

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

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

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

2. r_n is upper semi-continuous.

Condition W. 1. The function

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

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

2. r_n is continuous.

Theorem 4 (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]. □

0.3 Bertsekas-Shreve framework

The theory here described 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 mirrored.

Setting 2 (BS)**.** We write the source notation in parenthesis for comparison.

- 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 function $r : \mathcal{S} \times \mathcal{A}$ such that $r = r_n/\gamma^{n-1}$ for all $n \in \mathbb{N}$ (in the case where we are not discounting set $\gamma = 1$).
- $\mathcal{S}$ and $\mathcal{A}$ are Borel spaces.
- $\mathcal{A}$ is compact.
- r is upper semicontinuous and bounded from above (least upper bound denoted $R_{\max} > 0$).
- $P(S \mid \cdot)$ is continuous for any $S \in \Sigma_{\mathcal{S}}$.

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. This is equivalent to our conditions by symmetry.

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.

With setting 2 we can define

Definition 6 (The T-operators)**.** For a stationary policy π we define the operators

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

$$T^\pi V := s \mapsto \int r(s, a) + \gamma V(s') \mathrm{d}(P\pi)(s, a, s' \mid s)$$

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

where $T^a = T^{\delta_a}$.

Proposition 7. $V_k^* = T^k 0$ and is semi uppercontinuous. Furthermore there exists a deterministic, Markov, Borel-measurable policy $\pi^* = (\pi_1^*, \pi_2^*, \dots) \in DM\Pi$ which is k-optimal for all $k \in \mathbb{N}$.

Theorem 5. Under (N) or (D) $V^* = \lim_{k \to \infty} T^k 0$ and is upper semicontinuous. Furthermore there exist a deterministic stationary, Borel-measurable policy π^*.

Setting 3 (BS Analytic)**.** The same as setting 2 except: P is not necessarily continuous. r is upper semianalytic. $\mathcal{A}$ is not necessarily compact.

Theorem 6. Under setting 3 suppose 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 $\mathcal{A}$. Then $V^*(s) = \lim_{N \to \infty} V_N^*(s)$ 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]. □

Proposition 8. Under (D) for any $\pi \in R\Pi$ we have $V_{n,\pi}, V_\pi \leqslant V_{\max} := R_{\max}/(1 - \gamma)$.

Proof.

$$\sum_{i \in \mathbb{N}} \mathbb{E}_\mu r_i^+ \leqslant \sum_{i \in \mathbb{N}} \gamma^{i-1} R_{\max} \leqslant R_{\max}/(1 - \gamma) := V_{\max} < \infty$$

□

Proposition 9. Under (D) T^π is 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

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

□

Corollary 1. Under (D) V^π is the unique bounded fixed point of T^π.

Proof. This is by the Banach fixed point theorem and proposition 9. □