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 $\kappa(\cdot \mid \mu)$. This way $\kappa(\cdot \mid \cdot)$ can also be seen as a mapping from $\mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ (in the second entry).

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

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

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. $\forall \mu \in \mathcal{X} : (\psi \circ \kappa)(\cdot \mid \mu) = \psi(\cdot \mid \kappa(\cdot \mid \mu))$ 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 $\psi * (\kappa * \mu) = (\psi * \kappa) * \mu$.

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.

Proposition 2 (Existence and uniqueness of finite kernel processes)**.** For any probability measure $\rho_1 \in \mathcal{P}(\mathcal{X}_1)$ and every $n \in \mathbb{N}$ there exists a unique probability measure ρ_n on $\mathcal{X}^{\underline{n}}$ defined by

$$\rho_n := \kappa_{n-1} * \cdots * \kappa_1 * \rho_1$$

(with the convention that an empty $*$-product is $\mathrm{id}_{\mathcal{X}_1}$ by context)

Proof. This follows simply by induction using proposition 1. □

Let $\mathcal{X}^{\underline{\infty}} := \prod_{n\in\mathbb{N}} \mathcal{X}_n$ and $\Sigma_{\mathcal{X}^{\underline{\infty}}} := \bigotimes_{n\in\mathbb{N}} \Sigma_{\mathcal{X}_n}$. Proposition 2 is not enough to establish existence of a kernel generated measure on $(\mathcal{X}^{\underline{\infty}}, \Sigma_{\mathcal{X}^{\underline{\infty}}})$ which we will need later. 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

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

for all $n \in \mathbb{N}$. We denote this measure $\rho_\mu^{(\kappa_1,\kappa_2,\dots)}$.

Proof. Todo: what about this. □

We here include lemma about the behavior of the Ionescu-Tulcea measure for use later.

Lemma 1. The Ionescu-Tulcea measure satisfies $\rho_\mu^{(\kappa_1,\dots)} = \rho_{\kappa_1*\mu}^{(\kappa_2,\dots)}$.

Proof. Notice that by associativity $\kappa_n * \cdots * \kappa_1 * \mu = (\kappa_n * \cdots * \kappa_2) * (\kappa_1 * \mu)$. This implies that

$$\rho_\mu^{(\kappa_1,\dots)}\left(A \times \prod_{k=n+1}^{\infty} \mathcal{X}_k\right) = \rho_{\kappa_1*\mu}^{(\kappa_2,\dots)}\left(A \times \prod_{k=n+1}^{\infty} \mathcal{X}_k\right)$$

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

Definition 2 (Dynamic programming 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 $n \in \mathbb{N}$ write $\mathcal{H}_n = \mathcal{S}_1 \times \mathcal{A}_1 \times \cdots \times \mathcal{S}_n$, $\mathcal{H}_\infty = \mathcal{S}_1 \times \mathcal{A}_1 \times \dots$, with associated σ-algebras $\Sigma_{\mathcal{H}_n} := \left(\bigotimes_{n=1}^{n-1}(\Sigma_{\mathcal{S}_n} \otimes \Sigma_{\mathcal{A}_n})\right) \otimes \Sigma_{\mathcal{S}_n}$ and $\Sigma_{\mathcal{H}_\infty} := \bigotimes_{n=1}^{\infty}(\Sigma_{\mathcal{S}_n} \otimes \Sigma_{\mathcal{A}_n})$.

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 $\mathcal{H}_{n+1} \rightsquigarrow \mathbb{R}$ kernels called the **reward** kernels.

For such a 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 $R\Pi$.

Proposition 3 (Existence and uniqueness of policy generated processes)**.** Let $(\pi_n)_{n\in\mathbb{N}}$ be a policy and $\mu \in \mathcal{P}(\mathcal{S}_1)$ be a probability measure. Then for every $n \in \mathbb{N}$ there exists a unique probability measure $\rho_n \in \mathcal{P}(\mathcal{H}_n)$ such that $\rho_1 = \mu$ and $\rho_n = (P_n * \pi_n)(\cdot \mid \rho_{n-1})$. Futhermore there exists a unique probability measure $\rho \in \mathcal{P}(\mathcal{H}_\infty)$ satisfying

$$\rho_n(H_n) = \rho\left(H_n \times \prod_{k=n+1}^{\infty} (\mathcal{A}_k \times \mathcal{S}_k)\right)$$

We will call this the **process** measure for π and μ and denote it ρ_μ^π with a slight abuse of notation ($\rho_\mu^{(P_n*\pi_n)_{n\in\mathbb{N}}}$ would be less abusive). Expectations it we denote $\mathbb{E}_\mu^\pi$. In particular when $\mu = \delta_s$, i.e. the one-point measure of $s \in \mathcal{S}_1$ we write ρ_s^π and $\mathbb{E}_s^\pi$ for short.

Proof. This is directly from proposition 2 and theorem 3 with $\kappa_1 = P_1 * \pi_1, \kappa_2 = P_2 * \pi_2 \dots$. □

0.2.1 Optimal policies

Let $(\mathcal{S}_n, \mathcal{A}_n, P_n, R_n)_{n\in\mathbb{N}}$ be a dynamic programming model. Define the nth **expected reward function** $r_n : \mathcal{H}_{n+1} \to \mathbb{R}$ by $r_n(h) = \int r \mathrm{d}R_n(r \mid h)$ for any $h \in \mathcal{H}_{n+1}$.

In litterature the terminology varies and generally any function mapping a state space $\mathcal{S}$ to $\mathbb{R}$ can be called a (state) **value** function. Similarly any function mapping some the product of a state space and an action space $\mathcal{A}$ to $\mathbb{R}$ 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.

Assumption 1 (General assumption)**.** We assume that $\sum_{i\in\mathbb{N}} \mathbb{E}_\pi r_i^+ < \infty$ for all policies $\pi \in R\Pi$.

Then following definition makes sense

Definition 4 (Ideal and optimal value functions)**.** Let π be a policy. We define

$$V_n^\pi(s) := \mathbb{E}_s^\pi \sum_{i\in[n]} r_i \qquad\qquad V^\pi(s) := \mathbb{E}_s^\pi \sum_{i\in\mathbb{N}} r_i$$

called the **ideal** value functions for the policy π and

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

called the **optimal** value functions. A policy for which $V^{\pi^*} = V^*$ is called an **optimal** policy.

At this point many interesting questions can be asked.

1. Does an optimal policy π^* exist?

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

3. Can π^* be chosen to be deterministic?

These questions has been answered in a variety of settings. 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. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})$ 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.

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

3. $\mathcal{A}_n$ is compact.
4. $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N \geqslant n} \sup_{\pi \in R\Pi} \sum_{t=n+1}^{N} \mathbb{E}_s^\pi r_n \to 0$ 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 condition S or condition W hold then

1. There exist an optimal oplicy $\pi^* \in R\Pi$.
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 Stationary policies

We will now specialize the dynamic programming model to:

Setting 2 (Finite action decision model)**.** 1. $\mathcal{S}_1 = \mathcal{S}_2 = \ldots := \mathcal{S}$ and $\mathcal{S}$ is standard Borel.

2. $\mathcal{A}_1 = \mathcal{A}_2 = \ldots := \mathcal{A}$ and $\mathcal{A}$ is finite.
3. There exist a kernel $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ such that

$$P_n(\cdot \mid s_1, a_1, \dots, s_n, a_n) = P(\cdot \mid s_n, a_n), \qquad \forall n \in \mathbb{N}$$

4. There exist a kernel $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow (-\infty, R_{\max}]$ such that

$$R_n(\cdot \mid s_1, a_1, \dots, s_n, a_n, s_{n+1}) = \gamma^{n-1} R(\cdot \mid s_n, a_n), \qquad \forall n \in \mathbb{N}$$

where $\gamma \in [0, 1)$ is called the **discount factor**. We donte $r(s, a) := \int r' \mathrm{d}R(r' \mid s, a)$.

5. r is semi upper continuous.

Proposition 4. Setting 2 implies setting 1 and condition S.

Proof. Pt. 1 is by definition. We naturally endow $\mathcal{A}$ with the discrete topology and the powerset σ-algebra, making it standard Borel and compact. The last point in setting 1 is implied by assumption 1 and the discounting in setting 2 pt. 4. Pt. 1 in condition S is trivial since all functions are continuous from a discrete space and pt. 2 is by definition. □

Definition 5 (Markov and stationary policies)**.** A policy $(\pi_n)_{n \in \mathbb{N}}$ called **Markov** if π_n only depends on the last item in the history, that is $\pi_n(\cdot \mid s_1, a_1, \dots a_{n-1}, s_n) = \pi_n'(\cdot \mid s_n)$ for some kernel $\pi_n' : \mathcal{S} \rightsquigarrow \mathcal{A}$ for all $n \in \mathbb{N}$. The set of Markov policies we denote $M\Pi$.

A Markov policy is called **stationary** if there exist a kernel $\pi : \mathcal{S} \rightsquigarrow \mathcal{A}$ such that $\pi_n' = \pi$ for all $n \in \mathbb{N}$. The space of such policies we denote Π.

We remark that

$$\Pi \subseteq M\Pi \subseteq R\Pi$$

Proposition 5. There is an optimal policy which is Markov.

Proof. We first show that there exist optimal finite-horizon Markov policy for each $n \in \mathbb{N}$. Let $\tau_1 : \mathcal{S} \rightsquigarrow \mathcal{A}$ be a policy such that $\tau_1(\operatorname{argmax}_{a\in\mathcal{A}} r(s,a) \mid s) = 1$. Then clearly $V_1^{\tau_1}(s) = \max_{a\in\mathcal{A}} r(s,a) = V_1^*(s)$ and τ_1 is Markov as any one-step policy. For $n \in \mathbb{N}$ assume $(\tau_n, \ldots, \tau_1)$ is an optimal finite-horizon Markov policy. Let

$$\tau_{n+1}\left(\operatorname*{argmax}_{a\in\mathcal{A}} r(s,a) + \gamma \int V_n^{(\tau_n,\ldots,\tau_1)}(s')\mathrm{d}P(s' \mid s,a) \,\middle|\, s\right)$$

Then $(\tau_{n+1}, \ldots, \tau_1)$ is an $n+1$-optimal Markov policy. . . . $\square$

Definition 6 (The T-operators)**.** For a stationary policy π we define the operator T^π on $\mathcal{L}_\infty(\mathcal{S})$ by

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

When $\pi = \delta_a$ for some $a \in \mathcal{A}$ we simply write T^a.

Proposition 6. 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

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

$\square$

Let $\pi = (\pi_n)_{n\in\mathbb{N}}$ be a Markov policy.

Proposition 7. $V^\pi = T^{\pi_1}V^{(\pi_2,\ldots)}$. In particular if π is stationary then V^π is the unique bounded fixed point for T^π.

Proof. Fix $s \in \mathcal{S}$ and let ρ_s^π be the process measure (see proposition 3) of π.

$$\begin{aligned}
T^{\pi_1}V^{(\pi_2,\ldots)} &= \int V^{(\pi_2,\ldots)}(s_2)\mathrm{d}(P * \pi_1)(s_1,a_1,s_2 \mid \mu) \\
&= \int\int \sum_{n=1}^{\infty} \gamma^n r(s'_{n+1}, a'_{n+1})\mathrm{d}\rho_{s_2}^{(\pi_2,\ldots)}(s'_1, a'_1, \ldots)\mathrm{d}(P * \pi_1)(s_1,a_1,s_2|\mu) \\
&= \int\int \sum_{n=1}^{\infty} \gamma^n r(s'_{n+1}, a'_{n+1})\mathrm{d}\rho_{(P\circ\pi_1)(\cdot|\mu)}^{(\pi_2,\ldots)}(s'_1, a'_1, \ldots)
\end{aligned}$$

$\square$

Proposition 8. There exists a stationary policy τ such that $V^\tau \geqslant V^\pi$.

Proof. $\square$