Definitions and notation

Eugene Feinberg
SUNY Stony Brook
Stony Brook etc.

Adam Shwartz
Electrical Engineering.
Technion Etc.

May 29, 2000

Chapter 1

Introduction

ch:intro

Abstract

In this chapter we describe the basic structure of a Markov Decision Process. We introduce the notation and then give a quick tour of this volume. At this point, we use this as information for authors, with examples.

1.1 Introduction

In this section we shall introduce the volume, its purpose, etc. including a heuristic description of what an MDP is (in contrast to the formal presentation of section C1sec:notation 1.2).

1.2 Notation

C1sec:notation

Let $\mathbb{N} = \{0, 1, \dots\}$ and let $\mathbb{R}^n$ be an n-dimensional Euclidean space, $\mathbb{R} = \mathbb{R}^1$. A Markov Decision Process (MDP) is defined through the following objects:

a state space $\mathbb{X}$;

an action space $\mathbb{A}$;

sets $\mathbb{A}(x)$ of available actions at states $x \in \mathbb{X}$;

transition probabilities, denoted by $p(Y|x, a)$;

reward functions $r(x, a)$ denoting the one-step reward using action a in state x.

The above objects have the following meaning. There is a stochastic system with a state space $\mathbb{X}$. When the system is at state $x \in \mathbb{X}$, a decision-maker selects an action a from the set of actions $\mathbb{A}(x)$ available at state x. After an action a is selected, the system moves to the next state according to the probability distribution $p(\cdot|x, a)$ and the decision-maker collects a one-step reward $r(x, a)$. The selection of an action a may depend on the current state of the system, the current time, and the available information about the history of the system. At each step, the decision maker may select a particular action or, in a more general way, a probability distribution on the set of available actions $\mathbb{A}(x)$. Decisions of the first type are called nonrandomized and decisions of the second type are called randomized.

Discrete MDPs. An MDP is called finite if the state and action sets are finite. We say that a set is discrete if it is finite or countable. An MDP is called discrete if the state and action sets are discrete.

A significant part of research and applications related to MDPs deals with discrete MDPs. For discrete MDPs, we do not need additional measurability assumptions on the major objects introduced above. Readers who are not familiar with measure theory can still read the papers of this volume, since most of the papers deal with discrete MDPs: for the other papers, the results may be restricted to discrete state and action sets.

For a discrete state space $\mathbb{X}$ we denote the transition probabilities by $p(y|x, a)$ or $p_{xy}(a)$, and use (in addition to x, y) also the letters i, j, k etc. to denote states. Unless mentioned otherwise, we always assume that $p(\mathbb{X}|x,a) = 1$.

The time parameter is t, s or $n \in \mathbb{N}$ and a trajectory is a sequence $x_0a_0x_1a_1\dots$. The set of all trajectories is $H_\infty = (\mathbb{X} \times \mathbb{A})^\infty$. A trajectory of length n is called a history, and denoted by $h_n = x_0a_0\dots x_{n-1}a_{n-1}x_n$. Let $H_n = \mathbb{X} \times (\mathbb{A} \times \mathbb{X})^n$ be the space of histories up to epoch $n \in \mathbb{N}$. A nonrandomized policy is a sequence of mappings ϕ_n, $n \in \mathbb{N}$, from H_n to $\mathbb{A}$ such that $\phi_n(x_0a_0\dots x_{n-1}a_{n-1}x_n) \in \mathbb{A}(x_n)$. If for each n this mapping depends only on x_n, then the policy ϕ is called Markov. In other words, a Markov policy ϕ is defined by mappings $\phi_n : \mathbb{X} \to \mathbb{A}$ such that $\phi_n(x) \in \mathbb{A}(x)$ for all $x \in \mathbb{X}$, $n = 0, 1, \dots$. A Markov policy ϕ is called stationary if the ϕ_n do not depend on n. A stationary policy is therefore defined by a single mapping $\phi : \mathbb{X} \to \mathbb{A}$ such that $\phi(x) \in \mathbb{A}(x)$ for all $x \in \mathbb{X}$. We denote by Π, Π^M, and Π^S the sets of all nonrandomized, Markov, and stationary policies respectively. We observe that $\Pi^S \subseteq \Pi^M \subseteq \Pi$.

As mentioned above, by selecting actions randomly, it is possible to expand the set of policies. A randomized policy π is a sequence of transition probabilities $\pi_n(a_n|h_n)$ from H_n to $\mathbb{A}$, $n \in \mathbb{N}$, such that $\pi_n(\mathbb{A}(x_n)|x_0a_0\dots x_{n-1}a_{n-1}x_n) = 1$. A policy π is called randomized Markov if $\pi_n(a_n|x_0a_0\dots x_{n-1}a_{n-1}x_n) = \pi_n(a_n|x_n)$. If $\pi_m(\cdot|x) = \pi_n(\cdot|x)$ for all $m, n \in \mathbb{N}$ then the Markov policy π is called randomized stationary. A randomized stationary policy π is thus defined by a transition probability π from $\mathbb{X}$ to $\mathbb{A}$ such that $\pi(\mathbb{A}(x)|x) = 1$ for all $x \in \mathbb{X}$. We denote by Π^R, Π^{RM}, Π^{RS} the sets of all randomized, randomized Markov, and randomized stationary policies respectively. We have that $\Pi^{RS} \subseteq \Pi^{RM} \subseteq \Pi^R$, and in addition $\Pi^S \subseteq \Pi^{RS}$, $\Pi^M \subseteq \Pi^{RM}$, and $\Pi \subseteq \Pi^R$.

Note that, while we try to be consistent with the above definitions, there is no standard terminology for policies: in particular, there is no general agreement as to whether "stationary" implies nonrandomized or, more generally, whether the "default" should be randomized (the more general case) or nonrandomized. The following additional terms are sometimes also used:

pure policy means nonrandomized;

deterministic policy means (nonrandomized) stationary.

The stochastic process evolves as follows. If at time n the process is in state x, having followed the history h_n, then an action is chosen (perhaps randomly) according to the policy π. If action a ensued, then at time $n+1$ the process will be in the state y with probability $p(y|x,a)$.

Given an initial state x and a policy π, the "evolution rule" described above defines all finite-dimensional distributions $x_0, a_0, \dots, x_n$, $n \in \mathbb{N}$. Kolmogorov's extension theorem guarantees that any initial state x and any policy π define a stochastic sequence $x_0 a_0 x_1 a_1 \ldots$. We denote by $\mathbb{P}_x^\pi$ and $\mathbb{E}_x^\pi$ respectively the probabilities and expectations related to this stochastic sequence; $\mathbb{P}_x^\pi\{x_0 = x\} = 1$.

Any stationary policy ϕ defines a Markov chain with transition probabilities $p_{xy}(\phi) = p(y|x, \phi(x))$ on the state space $\mathbb{X}$. A randomized stationary policy π also defines a Markov chain with the state space $\mathbb{X}$. In the latter case, the transition probabilities are $p_{xy}(\pi) = \sum_{a \in \mathbb{A}(x)} \pi(a) p(y|x, a)$. We denote by $P(\pi)$ the transition matrix with elements $\{p_{xy}(\pi)\}$. The limiting matrix

$$Q(\pi) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n(\phi). \tag{1.1}$$

always exists and it is stochastic if $\mathbb{X}$ is finite; Chung [2, Section 1.6]. Let f be a terminal reward function and β be a discount factor. We denote by $v_N(x, \pi, \beta, f)$ the expected total reward over the first n steps, $n \in \mathbb{N}$:

$$v_N(x, \pi, \beta, f) = \mathbb{E}_x^\pi \left[\sum_{n=0}^{N-1} \beta^n r(x_n, a_n) + \beta^N f(x_N) \right], \tag{1.2}$$

C1e:DefFinRew

whenever this expectation is well-defined.

If $\beta \in [0, 1[$ then we deal with expected total discounted reward. If $\beta = 1$, we deal with expected total undiscounted reward or simply total reward. If the discount factor $\beta \in [0, 1]$ is fixed, we usually write $v(x, \pi)$ instead of $v(x, \pi, \beta)$.

The expected total reward over an infinite horizon is

$$v(x, \pi) = v(x, \pi, \beta) = v_\infty(x, \pi, \beta, 0) \,. \tag{1.3}$$

C1e:DefDiscCost

Conditions for the total reward $v(x, \pi, 1)$ to be well-defined are usually stronger than the conditions that ensure that total discounted rewards $v(x, \pi, \beta)$, $0 \le \beta < 1$, are well-defined. The expected reward per unit time is

$$w(x, \pi) = \liminf_{n \to \infty} \frac{1}{N} v_N(x, \pi, 1, 0). \tag{1.4}$$

C1e:DefAvRew

If a performance measure $g(x, \pi)$ is defined for all policies π, we denote

$$G(x) = \sup_{\pi \in \Pi^R} g(x, \pi). \tag{1.5}$$

C1e:maxValue

In terms of the performance measures defined above, this yields the values

$$V_N(x,\beta,f) \stackrel{def}{=} \sup_{\pi\in\Pi^R} v_N(x,\pi,\beta,f)\,, \tag{1.6}$$

C1e:DefFinValue

$$V(x) = V(x,\beta) \stackrel{def}{=} \sup_{\pi\in\Pi^R} v(x,\pi,\beta)\,, \tag{1.7}$$

C1e:DefDiscValue

$$W(x) \stackrel{def}{=} \sup_{\pi\in\Pi^R} w(x,\pi)\,. \tag{1.8}$$

C1e:DefAvValue

For $\epsilon \geq 0$, a policy π is called ϵ-optimal for criterion g if $g(x,\pi) \geq G(x)-\epsilon$ for all $x \in \mathbb{X}$. A 0-optimal policy is called optimal.

For a function f on $\mathbb{X}$, we consider the reward operators:

$$P^a f(x) \stackrel{def}{=} \mathbb{E}[f(x_1) \mid x_0 = x, a_0 = a]\,, \tag{1.9}$$

C1e:TransOper

$$T^a_\beta f(x) \stackrel{def}{=} r(x,a) + \beta P^a f(x) \tag{1.10}$$

C1e:DiscOper

and the optimality operators:

$$Pf(x) \stackrel{def}{=} \sup_{a\in\mathbb{A}(x)} P^a f(x)\,, \tag{1.11}$$

C1e:TransOptOper

$$T_\beta f(x) \stackrel{def}{=} \sup_{a\in\mathbb{A}(x)} T^a_\beta f(x)\,. \tag{1.12}$$

C1e:DiscOptOper

The finite horizon Optimality Equation is

$$V_{N+1}(x) = T_\beta V_N(x), \qquad N = 0,1,\ldots\,. \tag{1.13}$$

C1e:FinHorOptEq

The discounted reward Optimality Equation is

$$V(x) = T_\beta V(x)\,. \tag{1.14}$$

C1e:DiscOptEq

An action $a \in A(x)$ is called conserving at state x for the $(N+1)$-step problem if $T^a_\beta V_N(x) = T_\beta V_N(x)$. An action $a \in A(x)$ is called conserving at state x for the total discounted reward if $T^a_\beta V(x) = T_\beta V(x)$.

When $\beta = 1$ we denote $T^a = T^a_1$ and $T = T_1$. In particular,

$$V(x) = TV(x) \tag{1.15}$$

C1e:TotOptEq

is the Optimality Equation for expected total undiscounted rewards.

For total reward criteria, value functions usually satisfy the optimality equation. In addition, the sets of conserving n-step actions, $n = 1,\ldots,N+1$ form the sets of optimal actions for $(N+1)$-step problems. Under some additional conditions, the sets of conserving actions form the sets of optimal actions for infinite horizon problems. We shall consider these results in appropriate chapters. The average reward Optimality Equations are

$$W(x) = PW(x) \qquad (1.16) \quad \boxed{\texttt{C1e:AvOptEq1}}$$

$$W(x) + h(x) = \sup_{a \in \mathbb{A}'(x)} T^a h(x)\,, \qquad (1.17) \quad \boxed{\texttt{C1e:AvOptEq2}}$$

where

$$\mathbb{A}'(x) = \{a \in \mathbb{A}(x) : P^a W(x) = PW(x)\}\ . \qquad (1.18)$$

Equation (C1e:AvOptEq1 1.16) is called the First Optimality Equation and equation (C1e:AvOptEq2 1.17) is called the Second Optimality Equation. Note that if $W(x) = W$, a constant, then the First Optimality Equation holds and $\mathbb{A}'(x) = \mathbb{A}(x)$. In this case, the Second Optimality Equations transforms into

$$W + h(x) = Th(x) \qquad (1.19) \quad \boxed{\texttt{C1e:OptEq2}}$$

which is often referred to simply as the Optimality Equation for average rewards.

We allow for the starting point x to be defined by an initial probability distribution μ. In this case, we keep the above notation and definitions but we replace the initial state x with the initial distribution μ. For example, we use $\mathbb{P}^\pi_\mu$, $\mathbb{E}^\pi_\mu$, $v(\mu, \pi)$, $V(\mu)$, $w(\mu, \pi)$, and $W(\mu)$. We remark that, generally speaking, optimality and ϵ-optimality with respect to all initial distributions are stronger than the defined above optimality and ϵ-optimality with respect to all initial states. However, in many natural cases these definitions are equivalent. For example, it is true for total reward criteria.

A more general problem arises when there are multiple objectives. Suppose there are $(K + 1)$ reward functions $r_k(x, a)$, $k = 0, \ldots, K$. For finite horizon problems, terminal rewards may also depend on k. In this case, we index by $k = 0, \ldots, K$ all functions that describe rewards. For example, we use the notation $w_k(x, \pi)$, $f_k(x)$, and $W_k(x)$.

For problems with multiple criteria, it is usually natural to fix an initial state x. It is also possible to fix an initial distribution μ, with our convention that all definitions remain the same, but we write μ instead of x. So, for simplicity, we define optimal policies when the initial state x (not a distribution) is fixed.

If the performance of a policy π is evaluated by $(K + 1)$ criteria $g_k(x, \pi)$ then one goal may be to optimize criterion g_0 subject to constraints on

$g_1, \dots, g_K$. Let C_k, $k = 1, \dots, K$, be given numbers. We say that a policy π is feasible if

$$g_k(x, \pi) \geq C_k, \qquad k = 1, \dots, K . \tag{1.20}$$ C1e:DefFeasible

A policy π is called optimal for a constrained optimization problem if it is feasible and

$$g_0(x, \pi) \geq g_0(x, \sigma) \quad \text{for any feasible policy } \sigma. \tag{1.21}$$ C1e:DefOptConstr

Nondiscrete MDPs: general constructions. When a state space $\mathbb{X}$ or an action space $\mathbb{A}$ are is not discrete, the natural assumption is that they are measurable spaces endowed with σ-fields $\mathcal{X}$ and $\mathcal{A}$ respectively. When $\mathbb{X}$ or $\mathbb{A}$ are discrete, the corresponding σ-field is the set of all subsets of the corresponding set. It is also natural to assume that the sets $\mathbb{A}(x) \in \mathcal{A}$ of feasible actions are measurable, for all states $x \in \mathbb{X}$. Of course, this assumption always holds when $\mathbb{A}$ is discrete.

Unless we specify otherwise, we always consider a Borel σ-field $\mathcal{B}(\mathbb{R})$ on $\mathbb{R}$: this is the minimal σ field containing all intervals. For non-discrete MDPs, we also assume that r is a measurable function on $(\mathbb{X} \times \mathbb{A}, \mathcal{X} \times \mathcal{A})$ and $p(Y|x, a)$ is a regular transition probability from $(\mathbb{X} \times \mathbb{A}, \mathcal{X} \times \mathcal{A})$ to $(\mathbb{X}, \mathcal{X})$. Recall that given two measurable spaces $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$, we call p a regular transition probability from E_1 to E_2 if the following two conditions hold: (i) $p(\cdot|e_2)$ is a probability measure on $(E_1, \mathcal{E}_1)$ for any $e_2 \in E_2$, and (ii) the function $p(B|\cdot)$ is measurable on E_2 for any $B \in \mathcal{E}_1$.

In order to define policies in the general situation, we consider σ-fields $\mathcal{H}_n = \mathcal{X} \times (\mathcal{A} \times \mathcal{X})^n$ on the sets of histories $H_n = \mathbb{X} \times (\mathbb{A} \times \mathbb{X})^n$. Nonrandomized and randomized strategies are defined in a way similar to discrete MDPs, with standard and natural additional measurability conditions: (a) nonrandomized policies π are defined by mappings π_n which are measurable on $(H_n, \mathcal{H}_n)$, and (b) stationary and Markov policies are defined by mappings which are measurable on $\mathbb{X} \times \mathbb{A}$. Similarly, for randomized policies, π_n are regular transition probabilities from $(H_n, \mathcal{H}_n)$ to $(\mathbb{A}, \mathcal{A})$ and, for randomized Markov and stationary policies, they are regular transition probabilities from $(\mathbb{X} \times \mathbb{A}, \mathcal{X} \times \mathcal{A})$ to $(\mathbb{A}, \mathcal{A})$.

Let $\mathcal{H}_\infty = (\mathcal{X} \times \mathcal{A})^\infty$. Ionescu Tulcea theorem, Neveu [5, Section 5.1], implies that any initial state x and policy π define a unique probability measure on $(H_\infty, \mathcal{H}_\infty)$. We denote this measure by $\mathbb{P}_x^\pi$. Sometimes it is called the "strategic" measure. We denote by $\mathbb{E}_x^\pi$ expectations with respect to this measure. We also notice that Ionescu Tulcea theorem implies that $\mathbb{P}_x^\pi$

is a regular transition probability from $(\mathbb{X}, \mathcal{X})$ to $(H_\infty, \mathcal{H}_\infty)$ and this implies that the functions $v_n(x,\pi,\beta,f)$ and $v(x,\pi,\beta)$ are measurable in x for any policy π (the terminal function f is also assumed to be measurable).

We remark that we use Ionescu Tulcea theorem instead of better known Kolmogorov's extension theorem primarily because the latter requires additional assumptions about the structure of the state space (it has to be Borel) and the first one has no such structural assumptions.

At the intuitive level, randomized decisions are more general than nonrandomized ones; this means that any nonrandomized policy belongs to the class of randomized policies. In addition, in order to avoid trivial situation, an MDP has to have at least one policy. In order to guarantee these two intuitive properties, we always assume the following two mild conditions: (i) all one-point sets $\{a\}$ are elements of $\mathcal{A}$, $a \in \mathbb{A}$; (ii) there is at least one measurable function ϕ from $\mathbb{X}$ to $\mathbb{A}$ such that $\phi(x) \in \mathbb{A}(x)$ for all $x \in \mathbb{X}$ The first assumption always holds for models with discrete state and action spaces. The second assumption always holds for models with discrete state spaces.

For a measure ν and a measurable function f we use the equivalent notations

$$\nu(f) \stackrel{def}{=} \int f(\alpha)\, d\nu(\alpha) \stackrel{def}{=} f(\nu)\,. \tag{1.22}$$

If we denote $\pi_x(\cdot) = \pi(\cdot|x)$ for a randomized stationary policy π then, similarly to discrete MDPs, this policy defines a Markov chain with transition probabilities $p(dy|x,\pi_x)$. If $\mathbb{X}$ is discrete, this chain has transition matrix $P(\pi)$ with elements $p_{xy}(\pi_x)$.

Thus, an MDP, strategies, and objective functions can be defined under very general conditions. However, very little can be done if one tries to analyze MDPs with arbitrary measurable state spaces. The first complication is that the value functions V may not be measurable even for one-step models. The second complication is that an important step in the analysis of MDPs is to construct an equivalent randomized Markov policy for an arbitrary policy; see Chapter ???. This can be done by constructing regular transition probabilities $\mathbb{P}_x^\pi(da_n|x_n)$ which may not exist for general state and action spaces. These two complications do not exist if the state space is countable. These two complications can be resolved if $\mathbb{X}$ and $\mathbb{A}$ Borel spaces. In addition, at the current state of our knowledge, there is no clear need to consider MDPs with arbitrary state measurable spaces because there is no clear motivation or practical needs for such objects. For example, MDPs with Borel state spaces have applications to statistics, control of models with incomplete information, and to inventory management. However, for example, we are not aware of possible applications of MDPs with state spaces having higher cardinality than continuum.

Discrete state MDPs. In this case, the state space $\mathbb{X}$ is discrete and the action space is a measurable space $(\mathbb{A}, \mathcal{A})$ such that all one-point sets are measurable. The sets of feasible actions $\mathbb{A}(x)$ are also elements of $\mathcal{A}$. Reward functions $r(x, a)$ and transition probabilities $p(y|x, a)$ are automatically measurable in a. All constructions described for discrete and general MDPs go through with $\mathcal{X}$ being the σ-field of all subsets of $\mathbb{X}$.

Classical Borel MDPs. Though we do not follow any particular text, all definitions, constructions, and statements, related to Borel spaces we mention in this chapter can be found in Bertsekas and Shreve [1, Chapter 7]; see also Dynkin and Yushkevich [3] and Kechris [4].

Two measurable spaces $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$ are called isomorphic if there is a one-to-one measurable mapping f of $(E_1, \mathcal{E}_1)$ onto $(E_2, \mathcal{E}_2)$ such that f^{-1} is measurable. A Polish space is a complete separable metric space. Unless we specify otherwise, we always consider a Borel σ-fields $\mathcal{B}(E)$ on a metric space E; $\mathcal{B}(E)$ is the minimal σ-field containing all open subsets of E. Of course, any measurable subset E' of a Polish space forms a Polish space endowed with the Borel σ-field which is the intersection of E' with Borel subsets of the original space. A measurable space $(E, \mathcal{E})$ is called Borel if it is isomorphic to a Polish space. All Borel spaces are either finite or countable or continuum, and two Borel spaces with the same cardinality are isomorphic. Therefore, uncountable Borel spaces are continuum. They are also isomorphic to each other and to the sets $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $([0, 1], \mathcal{B}([0, 1]))$.

The assumptions for Borel MDPs are:

(i) $\mathbb{X}$ and $\mathbb{A}$ are Borel spaces and $\mathcal{X}$ and $\mathcal{A}$ are corresponding Borel σ-fields;

(ii) the graph

$$\mathrm{Gr}\ \mathbb{A}(x) = \{(x, a)|\ x \in \mathbb{X}, a \in \mathbb{A}(x)\}$$

is a measurable subset of $\mathbb{X} \times \mathbb{A}$ and there exists at least one measurable mapping ϕ of $\mathbb{X}$ into $\mathbb{A}$ such that $\phi(x) \in \mathbb{A}(x)$ for all $x \in \mathbb{A}(x)$;

(iii) the reward functions $r(x, a)$ are measurable on $\mathbb{X} \times \mathbb{A}$ and the transition probabilities $p(\cdot|x, a)$ are regular transition probabilities from $\mathbb{X} \times \mathbb{A}$ to $\mathbb{X}$.

Conditions (i) and (iii) are similar to the corresponding assumptions for general models. The measurability of the graph in (ii) implies that the sets $\mathbb{A}(x)$ are measurable. The existence of a measurable mapping (often called a "selector") implies that $\mathbb{A}(x) \neq \emptyset$ for all x. We remark that it is possible that the graph is Borel and all images are non-empty but the graph does not contain a Borel mapping. Therefore, the second assumption in (ii) is essential for the existence of at least one policy.

As was discussed above, the first real complication is that even for one-step problems, the values V may not be Borel measurable functions on $\mathbb{X}$. However, conditions (i)-(iii) imply that these functions are universally measurable for finite and infinite-horizon problems and therefore optimality operators can be defined.

Here we explain the concepts of universally measurable sets and functions. Let $(E, \mathcal{E})$ be a Borel space. For a given probability measure p on $(E, \mathcal{E})$, define a σ-field $\mathcal{E}_p$ which is a completion of $\mathcal{E}$ with respect to measure p. That is, $\mathcal{E}_p$ is the minimal σ-field that contains $\mathcal{E}$ and all subsets F of E such that $F \subset F'$ for some $F' \in \mathcal{E}$, and $p(F') = 0$. For example, if $(E, \mathcal{E}) = ([0,1], \mathcal{B}([0,1]))$ then we can consider the Lebesgue measure m defined by $m([a,b]) = |b-a|$. Then $\mathcal{E}_m$ is the so-called Lebesgue σ-field. Let $\mathbf{P}(E)$ be the set of all probability measures on E. Then the intersection of all σ-fields $\mathcal{E}_p$, $\mathcal{U}(E) = \cap_{\{p \in \mathbf{P}(E)\}} \mathcal{E}_p$, is a σ-field and it is called the universal σ-field. This σ-field is also called the σ-field of universally measurable sets and its elements are called universally measurable subsets of E. A universally measurable function on X is a measurable mapping from $(X, \mathcal{U}(X))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{U}(X)$ is a universal σ-field on X. Of course, any Borel set and any Borel function are universally measurable.

Thus, optimality equations can be defined for Borel MDPs. However, there is another complication for Borel models, which is annoying mostly for aesthetic reasons: ϵ-optimal policies may not exist for small positive ϵ, even for one-step Borel MDPs with bounded reward functions. The example constructed by David Blackwell is based on the observation that the value function is universally measurable but it may not be Borel. However, for any policy, the expected one-step reward is a Borel function of the initial step. Moreover, it is possible to show that for the Borel MDP described above, for any initial measure p on $\mathbb{X}$, and for any $\epsilon > 0$ there exists a policy which is $p-$a.s. ϵ-optimal. Such policies are called (p, ϵ)-optimal.

Universally measurable Borel MDPs. If we expand the set of policies and consider universally measurable policies, ϵ-optimal policies exist and the concept of (p, ϵ) optimality is not needed. However, if we expand the set of policies, the results and their proofs hold for assumptions which are broader than (ii) and (iii).

Before we give formal definitions, we explain the concept of analytic sets. Let f is a measurable mapping of a Borel space $(E_1, \mathcal{E}_1)$ into Borel x space $(E, \mathcal{E})$. If $F \in \mathcal{E}$ then by definition $f^{-1}(F) \in \mathcal{E}_1$. However, it is possible that $f(E) \notin \mathcal{E}$ for some Borel set $F \in \mathcal{E}_1$. A subset F of a Borel space $(E, \mathcal{E})$ is called analytic if there exists a Borel space $(E_1, \mathcal{E}_1)$ and a measurable mapping of E_1 to E such that $F = f(F_1)$ for some $F_1 \in \mathcal{E}_1$.

Since one can select $E_1 = E$ and $f(e) = e$, every Borel set is analytic. It is also possible to show that any analytic set is universally measurable. It is also possible to consider the σ-field of analytically measurable sets which is the smallest σ-field containing all analytic subsets of an analytic set. We remark that Borel and universally measurable σ-fields consist respectively of Borel and universally measurable sets. The situation is different for analytic sets and σ-fields of analytically measurable sets. The complement of an analytic set may not be analytic. Therefore, the σ-field of analytically measurable sets contains sets other than analytic. We remark that there are many equivalent definitions of analytic sets. For example, for Polish spaces they can be defined as continuous images or even as projections of Borel sets.

If $(E, \mathcal{E})$ and $(E_1, \mathcal{E}_1)$ are two Borel spaces (Borel sets with Borel σ-fields) then the mapping $f : E \to E_1$ is called universally (analytically) measurable if $f^{-1}(B)$ belongs to the σ-field of universally (analytically) measurable subsets of E.

The assumptions for universally measurable MDPs are:

(a) The state and action spaces $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{A}, \mathcal{A})$ are Borel spaces;

(b) Gr $\mathbb{A}(x)$ is an analytic subset of $\mathbb{X} \times \mathbb{A}$ and all sets $\mathbb{A}(x)$ are not empty;

(c) The reward function $r(x, a)$ is an upper analytic function on $\mathbb{X} \times \mathbb{A}$, that is, for any real number c, the set $\{r \geq c\}$ is an analytic subset of $\mathbb{X} \times \mathbb{A}$;

(d) The transition function $p(\cdot|x, a)$ is a regular transition probability from $(\mathbb{X} \times \mathbb{A}, \mathcal{X} \times \mathcal{A})$to $(\mathbb{X}, \mathcal{X})$.

Assumptions (a) and (d) coincide with similar assumptions for Borel MDPs. According to Jankov–von Neumann theorem, assumption (b) implies that there is an analytically measurable mapping ϕ from $\mathbb{X}$ to $\mathbb{A}$ such that $\phi(x) \in \mathbb{A}(x)$ for all $x \in \mathbb{X}$. Of course, any analytically measurable mapping is universally measurable. Assumption (c) is more general than the assumption that $r(x, a)$ is Borel. This generality is unimportant. It is kept in the literature just because the same proofs holds for upper analytic and Borel reward functions.

The last important difference between Borel and universally measurable MDPs is that policies are universally measurable for the latter ones. Non-randomized policies are universally measurable mappings ϕ_n of H_n to $\mathbb{A}$ such that $\phi(h_n) \in \mathbb{A}(x_n)$ for any $h_n = x_0 a_n \dots x_n \in H_n$. Markov (and stationary) policies are defined by universally measurable mappings ϕ_n of $\mathbb{X}$ to $\mathbb{A}$ such that $\phi_n(x) \in \mathbb{A}(x)$ $(\phi(x) \in \mathbb{A}(x))$ for all $x \in \mathbb{X}$. Randomized, randomized Markov, and randomized stationary policies are regular transition probabilities defined in the same way as for general models but the sets H_n and $\mathbb{X}$ are endowed with σ-fields of universally measurable subsets that play the role of σ-field $\mathcal{E}_1$ in the definition of regular transition probabilities given above. Condition (b) implies that there exists at least one policy.

There are other versions of universally measurable MDPs. For example, one can consider analytically measurable policies; see Bertsekas and Shreve [1] for details. The important feature is that all definitions and notations, given for discrete MDPs, hold also for universally measurable MDPs.

1.3 What's in this volume

To be written.

Bibliography

bs [1] [BertShr] D. P. Bertsekas and S. E. Shreve,"Stochastic Optimal Control: The Discrete-Time Case," Academic Press, New York, 1978 (republished by Athena Scientific, 1997).

chung [2] [Chung] K. L. Chung, "Markov Chains with Stationary Transition Probabilities," Springer-Verlag, Berlin, 1960.

dy [3] [DynYsh] E. B. Dynkin and A. A. Yushkevich, "Controlled Markov Processes," Springer-Verlag, New York, 1979 (translation from 1975 Russian edition).

ke [4] [Kechris] A. S. Kechris, "Classical Descriptive Set Theory," Springer-Verlag, New York, 1995.

ne [5] [Ne] J. Neveu, "Mathematical Foundations of the Calculus of Probability," Holden-Day, San Francisco, 1965.

Index