1 INTRODUCTION

Eugene A. Feinberg
Adam Shwartz

This volume deals with the theory of Markov Decision Processes (MDPs) and their applications. Each chapter was written by a leading expert in the respective area. The papers cover major research areas and methodologies, and discuss open questions and future research directions. The papers can be read independently, with the basic notation and concepts of Section 1.2. Most chapters should be accessible by graduate or advanced undergraduate students in fields of operations research, electrical engineering, and computer science.

1.1 AN OVERVIEW OF MARKOV DECISION PROCESSES

The theory of Markov Decision Processes—also known under several other names including sequential stochastic optimization, discrete-time stochastic control, and stochastic dynamic programming—studies sequential optimization of discrete time stochastic systems. The basic object is a discrete-time stochastic system whose transition mechanism can be controlled over time. Each control policy defines the stochastic process and values of objective functions associated with this process. The goal is to select a "good" control policy.

In real life, decisions that humans and computers make on all levels usually have two types of impacts: (i) they cost or save time, money, or other resources, or they bring revenues, as well as (ii) they have an impact on the future, by influencing the dynamics. In many situations, decisions with the largest immediate profit may not be good in view of future events. MDPs model this paradigm and provide results on the structure and existence of good policies and on methods for their calculation.

MDPs have attracted the attention of many researchers because they are important both from the practical and the intellectual points of view. MDPs provide tools for the solution of important real-life problems. In particular, many business and engineering applications use MDP models. Analysis of various problems arising in MDPs leads to a large variety of interesting mathematical and computational problems. Accordingly, this volume is split into two major categories: theory (Parts I and II) and applications (Part III).

The concept of dynamic programming, which is very important for MDPs, was systematically studied by Bellman in many papers and in the book [6]. This concept is natural and several authors used dynamic programming methods in 1940s–early 1950s or probably earlier to approach various problems. Examples include the work on statistical sequential analysis by Wald [50] and by Arrow, Blackwell, and Girshick[3], the work by Arrow, Harris, and Marschack [4] and by Dvoretsky, Kiefer, and Wolfowitz [21] on inventory control, and the work by Bellman and Blackwell [7] and Bellman and LaSalle [8] on games.

Shapley's [47] seminal work on stochastic games introduced important definitions and results. The relationship between MDPs and stochastic games is similar to the relationship between a usual game and an optimization problem: a stochastic game with one player is an MDP. Therefore, many experts consider this Shapley's paper as the first study of MDPs. In addition to the mentioned individuals, Puterman [43, p. 16] refers to Isaacs, Karlin, Massé, and Robbins as major contributors to early breaking work in 1940s and early 50s. The book by Dubins and Savage [20] on gambling theory played an important role. Howard [33] introduced policy iteration algorithms and that book started the systematic study of MDPs. Several seminal contributions were done by Blackwell, Denardo, Derman, Ross, and Veinott in the 1960s (references to their work, and to work of other individuals mentioned by names in this paragraph, can be found in Puterman [43]). Also in the 1960s, three distinguished probabilitists, Dynkin, Krylov, and Shiryaev [48], worked on MDPs in Russia. Hinderer's book [32] was an important contribution. Over the following thirty years, there were many fundamental and exciting developments in MDPs and their applications. Most are either described in this volume or associated with the names of its contributors.

Since their introduction in the 1950s, MDPs have become an important research area with a rich and deep theory and various applications. In fact, MDPs became basic tools for the analysis of many problems in operations research, electrical engineering and computer science. Algorithms for inventory control and telecommunications protocols are two examples of such engineering applications.

During the first thirty years of the MDP theory, roughly speaking until early 1980s, most of the research was centered around optimality equations and methods for their solution, namely policy and value iteration. Value iteration algorithms and their various versions are also known under the names of successive approximation, backward induction, and dynamic programming. The dynamic programming principle in its classical form can be applied only to problems with an appropriate single objective function. For example, the dynamic programming algorithm is applicable to optimization of an expected total reward over a finite time horizon. It can usually be applied to single-criterion infinite horizon problems with a total expected reward or average reward per unit time. For some other objective functions, or when the goal is to optimize one objective function under constraints on other criteria, the problem usually cannot be solved directly by dynamic programming; for an indirect approach see Piunovskiy and Mao [41]. Convex analytic methods, including linear and convex programming in finite and infinite dimensional spaces are usually more natural in these situations. There are many exciting recent developments, especially in applications, in which the dynamic programming principle plays an important role (see e.g. [12] and chapter 14, by Van Roy, in this volume). However, most of the research over the last two decades has been focused on difficult problems to which the dynamic programming principle cannot be applied in its direct form. In particular, a significant part of current research deals with multiple criteria.

1.2 DEFINITIONS AND NOTATIONS

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 $x_0 a_0 x_1 a_1 \dots$ 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 ϕ 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 randomized 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 $x_0a_0x_1a_1 \dots$. 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 for any initial distribution a homogeneous 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 for each initial distribution a homogeneous 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, when $\mathbb{X}$ is finite, this matrix is stochastic; Chung [18, 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, $n \in \mathbb{N}$:

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

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 the expected total undiscounted reward or simply the total reward. For infinite-horizon problems with $N = \infty$, we do not write N explicitly and the expected total rewards do not depend on the terminal reward f. Thus, we define by $v(x,\pi,\beta)$ the expected total rewards over the infinite horizon. 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

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

If the reward function r is bounded either from above or from below, the expected total rewards over the infinite horizon are well-defined when $\beta \in [0,1[$. Additional conditions are required for the expected total reward $v(x,\pi,1)$ to be well-defined. Since this sum may diverge when the discount factor is 1, it is natural to consider the expected reward per unit time

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

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

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

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

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

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

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

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.

We introduce the important notions of optimality operators and optimality equations. The conditions when optimality operators are well-defined and optimality equations hold are considered in appropriate chapters.

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

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

$$T_\beta^a g(x) \triangleq r(x,a) + \beta P^a g(x) \tag{1.10}$$

and the optimality operators:

$$Pg(x) \triangleq \sup_{a \in \mathbb{A}(x)} P^a g(x)\,, \tag{1.11}$$

$$T_\beta g(x) \triangleq \sup_{a \in \mathbb{A}(x)} T_\beta^a g(x)\,. \tag{1.12}$$

The finite horizon Optimality Equation is

$$V_{N+1}(x) = T_\beta V_N(x), \qquad x \in \mathbb{X},\ N = 0, 1, \dots\,, \tag{1.13}$$

with $V_0(x) = f(x)$ for all $x \in X$.

The discounted reward Optimality Equation is

$$V(x) = T_\beta V(x) \qquad x \in \mathbb{X}. \tag{1.14}$$

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

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

$$V(x) = TV(x), \qquad x \in \mathbb{X}, \tag{1.15}$$

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, $n = 1, \dots, N+1$ 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

$$W(x) = PW(x), \qquad x \in \mathbb{X}, \tag{1.16}$$

$$W(x) + h(x) = \sup_{a \in \mathbb{A}'(x)} T^a h(x), \qquad x \in \mathbb{X}, \tag{1.17}$$

where

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

Equation (1.16) is called the First Optimality Equation and equation (1.17) is called the Second Optimality Equation. We remark that W has a meaning of an optimal average reward per unit time and h has a meaning of a terminal reward. 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

$$W + h(x) = Th(x), \qquad x \in \mathbb{X}, \tag{1.19}$$

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 notions than the optimality and ϵ-optimality with respect to all initial states. However, in many natural cases these definitions are equivalent. For example, this 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 $g_1, \ldots, g_K$. Let C_k, $k = 1, \ldots, K$, be given numbers. We say that a policy π is feasible if

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

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

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

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 the 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 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 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 $\mathbb{X}$. Similarly, for randomized policies, π_n are transition probabilities from $(H_n, \mathcal{H}_n)$ to $(\mathbb{A}, \mathcal{A})$ and, for randomized Markov and stationary policies, they are transition probabilities from $(\mathbb{X}, \mathcal{X})$ to $(\mathbb{A}, \mathcal{A})$.

Let $\mathcal{H}_\infty = (\mathcal{X} \times \mathcal{A})^\infty$. Ionescu Tulcea theorem, Neveu [39, Section 5.1], implies that any initial probability measure μ on $\mathbb{X}$ and any policy π define a unique probability measure on $(H_\infty, \mathcal{H}_\infty)$. In particular, μ defines the initial distribution, π_n define transition probabilities from H_n to $H_n \times \mathbb{A}$, and p define transition probabilities from $H_n \times \mathbb{A}$ to H_{n+1}, $n = 0, 1, \ldots$. We denote this measure by $\mathbb{P}_\mu^\pi$. Sometimes this measure is called a "strategic" measure. We denote by $\mathbb{E}_\mu^\pi$ expectations with respect to this measure. If $\mu(x) = 1$ for some $x \in \mathbb{X}$, we write $\mathbb{P}_x^\pi$ and $\mathbb{E}_x^\pi$ instead of respectively $\mathbb{P}_\mu^\pi$ and $\mathbb{E}_\mu^\pi$. We also notice that Ionescu Tulcea theorem implies that $\mathbb{P}_x^\pi$ is a 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 the better known Kolmogorov's extension theorem primarily because the latter one requires that the process has values in a locally compact metric spaces. For MDPs this means that the state and action spaces are required to be locally compact metric spaces. Since Ionesco Tulcea theorem holds for arbitrary measurable spaces, it is more convenient to apply it to the construction of strategic measures in MDPs, rather than Kolmogorov's extension theorem.

At the intuitive level, a randomized decision at any state is a probability measure on the set of nonrandomized decisions. In addition, in order to avoid a 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 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

$$\nu(f) \triangleq \int f(\alpha)\, d\nu(\alpha) \triangleq f(\nu)\,. \tag{1.22}$$

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 Derman-Strauch's theorem which is the first theorem in chapter 6. This can be done by constructing 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 $\mathbb{A}$ are Borel spaces. In addition, at the current state of knowledge, there is no clear need to consider MDPs with arbitrary measurable state spaces because there is no clear motivation or practical need for such objects. For example, MDPs with Borel state spaces have applications in statistics, control of models with incomplete information, and inventory management. However 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. From the definitions for general MDPs we have that 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 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 [11, Chapter 7]; see also Dynkin and Yushkevich [22] and Kechris [35].

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-field $\mathcal{B}(E)$ on a metric space E; $\mathcal{B}(E)$ is the minimal σ-field containing all open subsets of E. 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 $([0,1], \mathcal{B}([0,1]))$. Any measurable subset E' of a Polish space forms a Borel space endowed with the Borel σ-field which is the intersection of E' with Borel subsets of the original space.

The assumptions for Borel MDPs are:

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

(ii) the graph

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

is a measurable subset of $(\mathbb{X} \times \mathbb{A}, \mathcal{X} \times \mathcal{A})$ 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 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 the σ-field $\mathcal{E}_p$ as the completion of $\mathcal{E}$ with respect to the 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 E is a measurable mapping from $(E, \mathcal{U}(E))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ 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 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 be a measurable mapping of a Borel space $(E_1, \mathcal{E}_1)$ into a Borel 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 for all B in $\mathcal{E}_1$.

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) $\mathrm{Gr}(A)$ 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 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 the 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 assumption on a reward function is considered in the literature mainly because the optimality operator preserves this property.

The last important difference between Borel and universally measurable MDPs is that policies are universally measurable for the latter ones. Nonrandomized 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 $h_n = x_0 a_n \ldots x_n \in H_n$. 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 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 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 [11] for details. The important feature is that all definitions and notations, given for discrete MDPs, hold also for universally measurable MDPs.

1.3 THE SCOPE OF THIS VOLUME

The first two parts of this book deal with theoretical questions and Part III addresses some applications of MDPs. Part I deals with models with finite state and action spaces, and Part II deals with infinite state problems.

The paper by Lodewijk Kallenberg surveys the classical theory for basic criteria including total discounted expected rewards, average expected rewards per unit time, and more sensitive criteria including bias, Blackwell, and n-discount optimality criteria. The paper by Mark Lewis and Martin Puterman focuses on bias optimality. In real life, parameters of the models may be measured with finite accuracy. An important question is what happens in the case of, say a linear perturbation when the transition probabilities $p(y|x,a)$ are replaced with transition probabilities $p(y|x,a) + \epsilon d(y|x,a)$, where $\epsilon > 0$ is a small parameter. The survey by Konstantin Avrachenkov, Jerzy Filar, and Moshe Haviv describes research and applications for a nontrivial case of singular perturbation when the above transformation changes the ergodic structure of the system. One important application is an approach to the classical Hamiltonian Cycle and Traveling Salesman Problems via MDPs introduced by Filar and Krass [28].

Part II covers the following major objective criteria for infinite state models: expected total rewards (Eugene Feinberg), average rewards/costs per unit time (Linn Sennott, Armand Makowski and Adam Shwartz, Sean Meyn, and significant parts of chapters written by Vivek Borkar and by Onésimo Hernández-Lerma and Jean Lasserre), Blackwell optimality (Arie Hordijk and Alexander Yushkevich), and linear combinations of various criteria (Eugene Feinberg and Adam Shwartz). The chapter written by Vivek Borkar concentrates on convex analytic methods and the chapter written by Onésimo Hernández-Lerma and Jean Lasserre describes the infinite dimensional linear programming approach which is one of the major developments of these methods.

Gambling theory, introduced by Dubins and Savage [20], is a close relative of MDPs. The chapter written by Lester Dubins, Ashok Maitra, and William Sudderth, the major contributors to gambling theory over the last three decades (see Maitra and Sudderth [38] for references and many beautiful results on gambling and games), establishes some links between gambling and MDPs. Though gambling theory and MDPs look like close fields, as far as we know, this chapter is only the third major paper that links MDPs and gambling. The other two publications are by Blackwell [15] and Schäl [44].

A significant part of this volume deals with average reward MDPs. This criterion is very important for applications. In addition, many interesting mathematical questions arise for average reward problems. A major research direction over the last fifteen years was to find ergodicity and other special conditions that hold for broad classes of applications and that ensure the existence of stationary optimal policies. The papers of this volume and many recent publications, including Sennott's and Hernández-Lerma and Lasserre's books [45, 30], demonstrate significant progress in this direction. As we mentioned, these results usually require some ergodicity or other structural assumptions which could be difficult to verify. An interesting development is a minimal pair approach, in which the controller selects an initial state in addition to a policy.

This approach is described in chapter 12 by Hernández-Lerma and Lasserre. Theorem 3 in that chapter is a beautiful result that states the existence of optimal policies for the minimal pair approach without any explicit ergodicity or other structural conditions; see also [30] and references therein.

If there are no additional structural assumptions, stationary optimal policies may not exist for the standard average reward criterion except in the case of finite state and action sets; see chapter 2, or original contributions [14, 19]. Attempts to expand this result to broader state and action spaces, undertaken between 1960 and 1980, identified significant difficulties. For finite state MDPs with compact action sets and continuous transition probabilities and reward functions, stationary optimal policies may not exist [5, 22, 7]. Stationary ϵ-optimal strategies exist for such models when continuity of reward functions is relaxed to upper-semicontinuity; see [17, 24]. For arbitrary average rewards finite state MDPs, there exist Markov ϵ-optimal policies [25, 13], ϵ-optimal policies in several other classes of nonstationary policies [27], but stationary ϵ-optimal policies may not exist [22]. If the state space is infinite, it is possible that there is no randomized Markov ϵ-optimal policy which is ϵ-optimal for two given initial states [25]. It is also possible that the supremum of average rewards over all randomized Markov policies is greater than the similar supremum over all (nonrandomized) Markov policies [22]; see also [46, p.91] for a corresponding example for stationary policies. If the initial state is fixed, in view of the Derman-Strauch theorem (the first theorem in Feinberg, chapter 6), for any $\epsilon > 0$ there exists an ϵ-optimal randomized Markov policy. If $\liminf$ is replaced with $\limsup$ in (1.4), for any give initial state and for any $\epsilon > 0$ there exists an ϵ-optimal Markov policy [26].

Part III deals with some applications of MDPs. Benjamin van Roy, chapter 14, describes recent trends and directions in neuro-dynamic programming, one of the major relatively recent developments in artificial intelligence, also known as reinforcement learning, which combines MDPs with approximation and simulations techniques. Manfred Schäl, chapter 15, considers MDP applications to finance. Eitan Altman, chapter 16, surveys MDP applications to telecommunications. Bernard Lamond and Abdeslem Boukhtouta, chapter 17, describe water reservoir applications.

This book covers most of the major directions in MDPs. It is probably impossible to cover in detail all aspects of MDPs in one book. This book focuses on discrete-time models with complete information. Models with incomplete information [29], continuous-time models, in particular Semi-Markov Decision Processes and Continuous Time MDPs (see [36, 10, 37, 45] and references therein), and risk-sensitive criteria [51] are three important topics which are out of the major scope of this book. Though this book does not have a special chapter on problems with multiple criteria and constraints, it describes the convex analytic approach which is the methodological foundation for studying such problems. Several chapters mention particular results on constrained MDPs and additional details can be found in books by Altman [1], Borkar [6], Kallenberg [37], and Piunovskiy [44]. There are numerous areas of applications of MDPs in addition to areas covered in this book. Some of them are summarized in Puterman's and Bertsekas' books [37, 9, 10]. Here we just mention the fundamental importance of MDPs for economic dynamics methods [49], transportation science [37], control of queues [36, 45], and production, inventory and supply chain management [9, 31, 42, 2].

All papers of this volume have been refereed. We would like to thank our colleagues for providing the editors and the authors with their comments and suggestions. In addition to the authors of this volume, most of whom served as referees, we would like to thank Igor Evstigneev, Emmanuel Fernandez-Gaucher and, Michael Katehakis, Victor Pestien, Ulrich Rieder, and Chelsea C. White, III for their valuable help. Dimitri P. Bertsekas and Matthew J. Sobel provided us with valuable comments on some literature sources. We are especially grateful to Martin L. Puterman who inspired us on this project. Last, but definitely not least, we would like to thank Ms. Lesley Price, who served as the de-facto technical editor of this volume. From re-typing a paper to correcting author's errors, Lesley withstood endless interchanges with authors efficiently and patiently. Her contribution to both form and substance of this volume is much appreciated.

Research of the first editor during his work on this volume was partially supported by NSF grant DMI-9908258. Research of the second editor was supported in part by the fund for promotion of research at the Technion, and in part by the fund for promotion of sponsored research at the Technion.

References

[1] E. Altman, *Constrained Markov Decision Processes,* Chapman & Hall/CRC, Boca Raton, 1999.

[2] R. Anupindi and Y. Bassok, “Supply contracts with quantity commitments and stochastic demand,” in *Quantitative Models for Supply Chain Management* (S. Tayur, R. Ganeshan, M. Magazine, eds.), pp. 197–232, Kluwer, Boston, 1999.

[3] K.J. Arrow, D. Blackwell, and M.A. Girshick, “Bayes and minimax solutions of sequential decision processes,” *Econometrica* **17**, pp. 213–244, 1949.

[4] K.J. Arrow, T. Harris, and J. Marschak, “Optimal inventory policies,” *Econometrica* **19**, pp. 250–272, 1951.

[5] J. Bather, “Optimal decision procedures for finite Markov chains I,” *Adv. Appl. Prob.* **5**, pp. 328-339, 1973.

[6] R.E. Bellman, *Dynamic Programming,* Princeton University Press, Princeton, 1957.

[7] R.E. Bellman and D. Blackwell, “On a particular non-zero sum game,” RM-250, RAND Corp., Santa Monica, 1949.

[8] R.E. Bellman and J.P. LaSalle, “On non-zero sum games and stochastic processes,” RM-212, RAND Corp., Santa Monica, 1949.

[9] D.P. Bertsekas, *Dynamic Programming and Optimal Control: Volume I,* Athena Scientific, Bellmont, MA, 2000 (second edition).

[10] D.P. Bertsekas, *Dynamic Programming and Optimal Control: Volume II,* Athena Scientific, Bellmont, MA, 1995.

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

[12] D.P. Bertsekas and J.N. Tsitsiklis, *Neuro-Dynamic Programming,* Athena Scientific, Bellmont, MA, 1996.

[13] K.-J. Bierth, "An expected average reward criterion", *Stochastic Processes and Applications* **26**, pp. 133–140, 1987.

[14] D. Blackwell, "Discrete dynamic programming," *Ann. Math. Stat.* **33**, pp. 719–726, 1962.

[15] D. Blackwell, "The stochastic processes of Borel gambling and dynamic programming," *Annals of Statistics* **4**, pp. 370–374, 1976.

[16] V.S. Borkar, *Topics in Controlled Markov Chains,* Pitman research Notes in Math., **240**, Longman Scientific and Technical, Harlow, 1991.

[17] R.Ya. Chitashvili, "A controlled finite Markov chain with an arbitrary set of decisions," *SIAM Theory Prob. Appl.* **20**, pp. 839–846, 1975.

[18] K.L. Chung, *Markov Chains with Stationary Transition Probabilities,* Springer-Verlag, Berlin, 1960.

[19] C. Derman, "On sequential decisions and Markov chains," *Man. Sci.* **9**, pp. 16–24, 1962.

[20] L.E. Dubins and L.J. Savage, *How to Gamble if You Must: Inequalities for Stochastic Processes,* McGraw-Hill, New York, 1965.

[21] A. Dvoretsky, J. Kiefer, and J. Wolfowitz, "The inventory problem: I. Case of known distribution of demand," *Econometrica* **20**, pp. 187–222, 1952.

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

[23] A. Federgruen, "Centralized planning models for multi-echelon inventory systems under inventory," in *Logistics of Production and Inventory,* (S.C. Graves, A.H.G. Rinnooy Kan, and P.H. Zipkin, eds), Handbooks in Operations Research and Management Science, **4**, pp. 133–173, North-Holland, Amsterdam, 1993.

[24] E.A. Feinberg, "The existence of a stationary ϵ-optimal policy for a finite Markov chain," *SIAM Theory Prob. Appl.* **23**, pp. 297–313, 1978.

[25] E.A. Feinberg, "An ϵ-optimal control of a finite Markov chain with an average reward criterion," *SIAM Theory Prob. Appl.* **25**, pp. 70–81, 1980.

[26] E.A. Feinberg, "Controlled Markov processes with arbitrary numerical criteria," *SIAM Theory Prob. Appl.* **27** pp. 486–503, 1982.

[27] E.A. Feinberg and H. Park, "Finite state Markov decision models with average reward criteria," *Stoch. Processes Appl.,* **31** pp. 159–177, 1994.

[28] J.A. Filar and D. Krass, "Hamiltonian cycles and Markov chains," *Math. Oper. Res.* **19**, pp. 223-237, 1994.

[29] O. Hernández-Lerma, *Adaptive Markov Control Processes,* Springer, New York, 1989.

[30] O. Hernández-Lerma and J.B. Lasserre, *Further Topics in Discrete-Time Markov Control Processes,* Springer, New York, 1999.

[31] D.P. Heyman and M.J. Sobel, *Stochastic Methods in Operations Research. Volume II: Stochastic Optimization,* McGraw-Hill, New York, 1984.

[32] K. Hinderer, *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*, Springer-Verlag, New York, 1970.

[33] R.A. Howard *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, 1960.

[34] L.C.M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tract 148, Mathematical Centre, Amsterdam, 1983.

[35] A.S. Kechris, *Classical Descriptive Set Theory,* Springer-Verlag, New York, 1995.

[36] M.Yu. Kitaev and V.V. Rykov, *Controlled Queueing Systems,* CRC Press, Boca Raton, 1995.

[37] A.J. Kleywegt and J.D. Papastavrou, "Acceptance and dispatching policies for a distribution problem", *Transportation Science,* **32,** pp. 127-141, 1998.

[38] A.P. Maitra and W.D. Sudderth, *Discrete Gambling and Stochastic Games,* Springer, New York, 1996.

[39] J. Neveu, *Mathematical Foundations of the Calculus of Probability,* Holden-Day, San Francisco, 1965.

[40] A.B. Piunovskiy, *Optimal Control of Random Sequences in Problems with Constraints,* Kluwer, Dordrecht, 1997.

[41] A.B. Piunovskiy and X. Mao, "Constrained Markovian decision processes: the dynamic programming approach," *Operations Research Letters* **27**, pp. 119-126, 2000.

[42] E.L. Porteus, "Stochastic inventory theory," in *Stochastic Models,* (D.P. Heyman and M.J. Sobel, eds), Handbooks in Operations Research and Management Science, **2**, pp. 605–652, North-Holland, Amsterdam, 1990.

[43] M.L. Puterman, *Markov Decision Processes,* Wiley, New York, 1994.

[44] M. Schäl, "On stochastic dynamic programming: a bridge between Markov decision processes and gambling." *Markov processes and control theory*, pp. 178–216, *Math. Res.* **54**, Akademie-Verlag, Berlin, 1989.

[45] L. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems,* Wiley, New York, 1999.

[46] S. Ross, *Introduction to Stochastic Dynamic Programming,* Academic Press, New York, 1983.

[47] L.S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences*, pp. 1095–1100, 1953.

[48] A.N. Shiryaev, "On the theory of decision functions and control by an observation process with incomplete data," *Selected Translations in Math. Statistics and Probability* **6**, pp.162–188, 1966.

[49] N.L. Stokey and R.E. Lucas, Jr. *Recursive Methods in Economic Dynamics,* Harvard University Press, Cambridge, 1989.

[50] A. Wald, *Sequential Analysis,* Wiley, New York, 1947.

[51] P. Whittle, *Risk-Sensitive Optimal Control,* Wiley, NY, 1990

Eugene A. Feinberg
Department of Applied Mathematics and Statistics
SUNY at Stony Brook
Stony Brook, 11794-3600, NY, USA
Eugene.Feinberg@sunysb.edu

Adam Shwartz
Department of Electrical Engineering
Technion—Israel Institute of Technology
Haifa 32000, Israel
adam@ee.technion.ac.il