Handbook of Markov Decision Processes: Methods and Applications

Book · January 2002

DOI: 10.1007/978-1-4615-0805-2

2 authors, including:

Adam Shwartz
Technion - Israel Institute of Technology
110 PUBLICATIONS **2,597** CITATIONS

Handbook of Markov Decision Processes

Handbook of Markov Decision Processes

Methods and Applications

Edited by
Eugene A. Feinberg
State University of New York at Stony Brook
Adam Schwartz
Technion—Israel Institute of Technology

Kluwer Academic Publishers
Boston/Dordrecht/London

We dedicate this volume to

Harold J. Kushner and Alexander A. Yushkevich,
our thesis advisors and major contributors to the
field of stochastic control.

The Editors

Contents

0 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 0.2. Most chapters should be accessible by graduate or advanced undergraduate students in fields of operations research, electrical engineering, and computer science.

0.1 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 [49] 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 [46] 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 [42, 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 [32] 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 [42]). Also in the 1960s, three distinguished probabilitists, Dynkin, Krylov, and Shiryaev [47], worked on MDPs in Russia. Hinderer's book [31] 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, there were many interesting, important and surprising discoveries in MDPs that lead to an interesting 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 [40]. 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 13, by Van Roy, in this volume). However, most of the research over the last two decades has been focused on 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 problems.

0.2 DEFINITIONS AND 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 $x_0 a_0 x_1 a_1 \dots$ 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 $h_n = x_0 a_0 \dots x_{n-1} a_{n-1} x_n$. 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 $\phi_n(x_0 a_0 \dots x_{n-1} a_{n-1} x_n) \in \mathbb{A}(x_n)$. 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 $\pi_n(\mathbb{A}(x_n)|x_0 a_0 \dots x_{n-1} a_{n-1} x_n) = 1$. A policy π is called randomized Markov if $\pi_n(a_n|x_0 a_0 \dots x_{n-1} a_{n-1} x_n) = \pi_n(a_n|x_n)$. 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 $x_0 a_0 x_1 a_1 \dots$. 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

$$Q(\pi) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n(\phi) \tag{0.1}$$

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, $n \in \mathbb{N}$:

$$v_N(x,\pi,\beta,f) = \mathbb{E}_x^\pi \left[\sum_{n=0}^{N-1} \beta^n r(x_n,a_n) + \beta^N f(x_N) \right], \tag{0.2}$$

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

$$v(x,\pi) = v(x,\pi,\beta) = v_\infty(x,\pi,\beta,0)\,. \tag{0.3}$$

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

$$w(x,\pi) = \liminf_{n\to\infty} \frac{1}{N} v_N(x,\pi,1,0). \tag{0.4}$$

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

$$G(x) = \sup_{\pi\in\Pi^R} g(x,\pi). \tag{0.5}$$

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

$$V_N(x,\beta,f) \triangleq \sup_{\pi\in\Pi^R} v_N(x,\pi,\beta,f)\,, \tag{0.6}$$

$$V(x) = V(x,\beta) \triangleq \sup_{\pi\in\Pi^R} v(x,\pi,\beta)\,, \tag{0.7}$$

$$W(x) \triangleq \sup_{\pi\in\Pi^R} w(x,\pi)\,. \tag{0.8}$$

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:

$$P^a g(x) \triangleq \mathbb{E}[g(x_1) \mid x_0 = x, a_0 = a]\,, \tag{0.9}$$

$$T_\beta^a g(x) \triangleq r(x,a) + \beta P^a g(x) \tag{0.10}$$

and the optimality operators:

$$Pg(x) \triangleq \sup_{a \in \mathbb{A}(x)} P^a g(x)\,, \tag{0.11}$$

$$T_\beta g(x) \triangleq \sup_{a \in \mathbb{A}(x)} T_\beta^a g(x)\,. \tag{0.12}$$

The finite horizon Optimality Equation is

$$V_{N+1}(x) = T_\beta V_N(x), \qquad x \in \mathbb{X},\ N = 0, 1, \dots\,, \tag{0.13}$$

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

The discounted reward Optimality Equation is

$$V(x) = T_\beta V(x) \qquad x \in \mathbb{X}. \tag{0.14}$$

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,

$$V(x) = TV(x), \qquad x \in \mathbb{X}, \tag{0.15}$$

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

$$W(x) = PW(x), \qquad x \in \mathbb{X}, \tag{0.16}$$

$$W(x) + h(x) = \sup_{a \in \mathbb{A}'(x)} T^a h(x), \qquad x \in \mathbb{X}, \tag{0.17}$$

where

$$\mathbb{A}'(x) = \{a \in \mathbb{A}(x) : P^a W(x) = PW(x)\}\,, \quad x \in \mathbb{X}. \tag{0.18}$$

Equation (0.16) is called the First Optimality Equation and equation (0.17) 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

$$W + h(x) = Th(x), \qquad x \in \mathbb{X}, \tag{0.19}$$

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 $\mathbb{P}_\mu^\pi$, $\mathbb{E}_\mu^\pi$, $v(\mu,\pi)$, $V(\mu)$, $w(\mu,\pi)$, 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 $r_k(x,a)$, $k=0,\dots,K$. For finite horizon problems, terminal rewards may also depend on k. In this case, we index by $k=0,\dots,K$ 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

$$g_k(x,\pi) \geq C_k, \qquad k=1,\dots,K\,. \tag{0.20}$$

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

$$g_0(x,\pi) \geq g_0(x,\sigma) \quad \text{for any feasible policy } \sigma. \tag{0.21}$$

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 [38, 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

$$\nu(f) \triangleq \int f(\alpha)\, d\nu(\alpha) \triangleq f(\nu)\,. \tag{0.22}$$

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 Feinberg, chapter 5. 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 [34].

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 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 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 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. 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 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.

0.3 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 discounting 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 linear perturbation when 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 so-call small parameters. 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 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 the gambling theory over the last three decades (see Maitra and Sudderth [37] for references and many beautiful results on gambling and games), establishes some links between gambling and MDPs. Though gambling theory and MDPs are close relatives, 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 [43].

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. The 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 [44, 29], 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 addi-

tion to a policy. This approach is described in chapter 11 by Hernández-Lema 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 [29] 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 Kallenberg, chapter 1, or original contributions [14, 19]. Attempts to expand this result to broader state and action spaces, undertaken between 1960s–1980s, identified significant difficulties. For finite state MDPs with compact actions 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 [45, 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 5), for any $\epsilon > 0$ there exists an ϵ-optimal randomized Markov policy. If lim inf is replaced with lim sup in (0.4), 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 13, 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 14, considers MDP applications to finance. Eitan Altman, chapter 15, surveys MDP applications to telecommunications. Bernard Lamond and Abdeslem Boukhtouta, chapter 16, describe water reservoir applications.

We hope that this book covers most of major directions in MDPs. It covers all major criteria except risk-sensitive ones; see [50]. It covers only discrete time models with complete information. It does not cover continuous time problems and models with incomplete information. 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 many details can be found in books by Altman [1], Borkar [16], Kallenberg [33], and Piunovskiy [39]. 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's books [42, 9, 10]. Here we just mention the fundamental importance of MDPs for economic dynamics methods [48], transportation science [36], control of queues [35, 44], and production, inventory and supply chain management [23, 30, 41, 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-Gaucherand, 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 and J.B. Lasserre, *Further Topics in Discrete-Time Markov Control Processes,* Springer, New York, 1999.

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

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

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

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

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

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

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

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

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

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

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

[41] 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.

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

[43] 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.

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

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

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

[47] 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.

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

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

[50] 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

I Finite State and Action Models

1 FINITE STATE AND ACTION MDPS

Lodewijk Kallenberg

Abstract: In this chapter we study Markov decision processes (MDPs) with finite state and action spaces. This is the classical theory developed since the end of the fifties. We consider finite and infinite horizon models. For the finite horizon model the utility function of the total expected reward is commonly used. For the infinite horizon the utility function is less obvious. We consider several criteria: total discounted expected reward, average expected reward and more sensitive optimality criteria including the Blackwell optimality criterion. We end with a variety of other subjects.
The emphasis is on computational methods to compute optimal policies for these criteria. These methods are based on concepts like value iteration, policy iteration and linear programming. This survey covers about three hundred papers. Although the subject of finite state and action MDPs is classical, there are still open problems. We also mention some of them.

1.1 INTRODUCTION

1.1.1 Origin

Bellman's book [13], can be considered as the starting point of Markov decision processes (MDPs). However, already in 1953, Shapley's paper [221] on stochastic games includes as a special case the value iteration method for MDPs, but this was recognized only later on. About 1960 the basics for the other computational methods (policy iteration and linear programming) were developed in publications like Howard [121], De Ghellinck [42], D'Epenoux [55], Manne [164] and Blackwell [27]. Since the early sixties, many results on MDPs are published in numerous journals, monographs, books and proceedings. Thousands of papers were published in scientific journals. There are about fifty books on MDPs. Around 1970 a first series of books was published. These books (e.g. Derman [58], Hinderer [107], Kushner [148], Mine and Osaki [167] and Ross [198]) contain the fundamentals of the theory of finite MDPs. Since that time nearly

every year one or more MDP-books appeared. These books cover special topics (e.g. Van Nunen [250], Van der Wal [246], Kallenberg [134], Federgruen [69], Vrieze [260], Hernández-Lerma [102], Altman [2] and Sennott [218]) or they deal with the basic and advanced theory of MDPs (e.g. Bertsekas [15], Whittle [289], [290], Ross [200], Dietz and Nollau [63], Bertekas [17], Denardo [50], Heyman and Sobel [106], White [285], Puterman [186], Bertsekas [18], [19], Hernández-Lerma and Lasserre [103], [104], and Filar and Vrieze [79].

1.1.2 The model

We will restrict ourselves to discrete, finite Markovian decision problems, i.e. the *state space* $\mathbb{X}$ and the *action sets* $\mathbb{A}(i), i \in \mathbb{X}$, are finite, and the decision time points t are equidistant, say $t = 1, 2, \ldots$. If, at time point t, the system is in state i and action $a \in \mathbb{A}(i)$ is chosen, then the following happens independently of the history of the process:
(1) a *reward* $r(i, a)$ is earned immediately;
(2) the process moves to state $j \in \mathbb{X}$ with *transition probability* $p(j|i, a)$, where $p(j|i, a)) \geq 0$ and $\sum_j p(j|i, a) = 1$ for all i, j and a.

The objective is to determine a policy, i.e. a rule at each decision time point, which optimizes the performance of the system. This performance is expressed as a certain *utility function*. Such utility function may be the expected total (discounted) reward over the planning horizon or the average expected reward per unit time. The decision maker has to find the optimal balance between immediate reward and future reward: a high immediate reward may bring the process in a bad situation for later rewards.

In Chapter 0 several classes of *policies* are introduced: general policies, Markov policies and stationary policies. There are randomized and nonrandomized (pure) policies. Denote the set of pure stationary policies by F and a particular policy of that set by f. Let $\mathbb{X} \times \mathbb{A} = \{(i, a) \mid i \in \mathbb{X},\ a \in \mathbb{A}(i)\}$, let the random variables X_t and Y_t denote the state and action at time t and let $\mathbb{P}_{\beta,\pi}[X_t = j, Y_t = a]$ be the notation for the probability that at time t the state is j and the action is a, given that policy π is used and β is the initial distribution. The next theorem shows that for any initial distribution β, any sequence of policies $\pi_1, \pi_2, \ldots$ and any convex combination of the marginal distributions of $\mathbb{P}_{\beta,\pi_k}, k \in \mathbb{N}$, there exists a Markov policy with the same marginal distribution.

Theorem 1.1 *Given any initial distribution β, any sequence of policies π_1, $\pi_2, \ldots$ and any sequence of nonnegative real numbers $p_1, p_2 \ldots$ with $\sum_k p_k = 1$, there exists a Markov policy π_* such that for every $(j, a) \in \mathbb{X} \times \mathbb{A}$*

$$\mathbb{P}_{\beta,\pi_*}[X_t = j, Y_t = a] = \sum\nolimits_k p_k \cdot \mathbb{P}_{\beta,\pi_k}[X_t = j, Y_t = a], \quad t \in \mathbb{N}. \tag{1.1}$$

Corollary 1.1 *For any starting state i and any policy π, there exists a Markov policy π_* such that*

$$\mathbb{P}_{i,\pi_*}[X_t = j, Y_t = a] = \mathbb{P}_{i,\pi}[X_t = j, Y_t = a], \quad t \in \mathbb{N}, (j, a) \in \mathbb{X} \times \mathbb{A}. \tag{1.2}$$

The results of Theorem 1.1 and Corollary 1.1 imply the sufficiency of Markov policies for performance measures which only depend on the marginal distributions. Corollary 1.1 is due to Derman and Strauch [61] and the extension to Theorem 1.1 was given by Strauch and Veinott [237]. The result is further generalized to more general state and actions spaces by Hordijk [112] and Van Hee [247].

1.1.3 Optimality criteria

Let $v(i,\pi)$ be the utility function if policy π is used and state i is the starting state, $i \in \mathbb{X}$. The *value vector* v of this utility function is defined by

$$v(i) := \sup_{\pi} v(i,\pi), \quad i \in \mathbb{X}. \tag{1.3}$$

A policy π is an *optimal policy* if $v(i,\pi) = v(i), i \in \mathbb{X}$. In Markov decision theory the existence and the computation of optimal policies is studied. For this purpose a so-called *optimality equation* is derived, i.e. a functional equation for the value vector. Then a solution of this equation is constructed which produces both the value vector and an optimal policy. There are three standard methods to perform this: value iteration, policy iteration and linear programming.

In *value iteration* the optimality equation is solved by successive approximation. Starting with some v^0, v^{t+1} is computed from $v^t, t = 0, 1, \ldots$. The sequence $v^0, v^1, \ldots$ converges to the solution of the optimality equation. In *policy iteration* a sequence of improving policies $f_0, f_1, \ldots$ is determined, i.e. $v(f_{t+1}) \geq v(f_t)$ for all t, until an optimal policy is reached. The *linear programming* method can be used because the value vector is the smallest solution of a set of linear inequalities; an optimal policy can be obtained from its dual program.

In this survey we consider the following utility functions:
(1) total expected reward over a finite horizon;
(2) total expected discounted reward over an infinite horizon;
(3) average expected reward over an infinite horizon;
(4) more sensitive optimality criteria for the infinite horizon.

Suppose that the system has to be controlled over a finite planning horizon of T periods. As performance measure we use the *total expected reward* over the planning horizon, i.e. for policy π we will consider for starting state i

$$v^T(i,\pi) := \sum_{t=1}^{T} \mathbb{E}_{i,\pi}[r(X_t, Y_t)] = \sum_{t=1}^{T} \sum_{j,a} \mathbb{P}_{i,\pi}[X_t = j, Y_t = a] \cdot r(j,a). \tag{1.4}$$

A matrix $P = (p_{ij})$ is called a *transition matrix* if $p_{ij} \geq 0$ for all (i,j) and $\sum_j p_{ij} = 1$ for all i. Markov policies, and consequently also stationary policies, induce transition matrices. For the randomized Markov policy $\pi = (\pi^1, \pi^2, \ldots)$ we define, for every $t \in \mathbb{N}$, the transition matrix $P(\pi^t)$ by

$$[P(\pi^t)]_{ij} := \sum\nolimits_a p(j|i,a)\pi^t(i,a) \text{ for all } i,j \in \mathbb{X}, \tag{1.5}$$

and the reward vector $r(\pi^t)$ by

$$r_i(\pi^t) := \sum\nolimits_a \pi^t(i,a) r(i,a) \text{ for all } i \in \mathbb{X} \tag{1.6}$$

Hence the total expected reward for the Markov policy π can be written in vector notation as

$$v^T(R) = \sum\nolimits_{t=1}^{T} P(\pi^1) P(\pi^2) \cdots P(\pi^{t-1}) r(\pi^t). \tag{1.7}$$

It can be shown that an optimal Markov policy $\pi_* = (f_*^1, f_*^2, \ldots, f_*^T)$ exists, where f_*^t is a pure decision rule $1 \leq t \leq$ T. The nonstationarity is due to the finiteness of the planning horizon.

Next, we consider an infinite planning horizon. In that case there is no unique optimality criterion. Different optimality criteria are meaningful: discounted reward, total reward, average reward or more sensitive criteria.

The *total expected α-discounted reward*, given *discount factor* $\alpha \in [0,1)$, initial state i and policy π, is denoted by $v^\alpha(i,\pi)$ and defined by

$$\begin{aligned} v^\alpha(i,\pi) &:= \sum\nolimits_{t=1}^{\infty} \mathbb{E}_{i,\pi}[\alpha^{t-1} r(X_t, (Y_t)] \\ &= \sum\nolimits_{t=1}^{\infty} \alpha^{t-1} \sum\nolimits_{j,a} \mathbb{P}_{i,\pi}[X_t = j, Y_t = a] r(j,a). \end{aligned} \tag{1.8}$$

In section 1.3.1 it will be shown that there exists an optimal policy $f \in F$ and that any stationary policy π satisfies

$$v^\alpha(\pi) = \sum\nolimits_{t=1}^{\infty} \alpha^{t-1} P(\pi)^{t-1} r(\pi) = [I - \alpha P(\pi)]^{-1} r(\pi). \tag{1.9}$$

When there is no discounting, i.e. the discount factor α equals 1, then—for instance—we may consider the total expected reward and the average expected reward criterion. In the total expected reward criterion the utility function is $\lim_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}[r(X_t, Y_t)]$. Without further assumptions, this limit can be infinite or the limsup can be unequal to the liminf. When the average reward criterion is used, the limiting behavior of the expectation of $\frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)$ is considered. Since $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[r(X_t, Y_t)]$ or $\mathbb{E}[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)]$ does not exist, in general, and interchanging limit and expectation may not be allowed, there are four different evaluation measures, which can be considered for a given policy:
(a) the lower limit of the average expected reward:

$$\phi(i,\pi) := \liminf_{T\to\infty} \frac{1}{T} \sum\nolimits_{t=1}^{T} \mathbb{E}_{i,\pi}[r(X_t, Y_t)], i \in \mathbb{X}; \tag{1.10}$$

(b) the upper limit of the average expected reward:

$$\Phi(i,\pi) := \limsup_{T\to\infty} \frac{1}{T} \sum\nolimits_{t=1}^{T} \mathbb{E}_{i,\pi}[r(X_t, Y_t)], i \in \mathbb{X}; \tag{1.11}$$

(c) the expectation of the lower limit of the average reward:

$$\psi(i,\pi) := \mathbb{E}_{i,\pi}[\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)], i \in \mathbb{X}; \tag{1.12}$$

(d) the expectation of the upper limit of the average reward:

$$\Psi(i,\pi) := \mathbb{E}_{i,\pi}[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)], i \in \mathbb{X}. \tag{1.13}$$

Lemma 1.1
(i) $\psi(\pi) \leq \phi(\pi) \leq \Phi(\pi) \leq \Psi(\pi)$ *for every policy* π*;*
(ii) $\psi(\pi) = \phi(\pi) = \Phi(\pi) = \Psi(\pi)$ *for every stationary policy* π*.*

Remark
In Bierth [26] it is shown that the four criteria are equivalent in the sense that the value vectors can be attained for one and the same deterministic policy. Examples can be constructed in which for some policy π the inequalities of Lemma 1.1 part (i) are strict.

The long-run average reward criterion has the disadvantage that it does not consider rewards earned in a finite number of periods. Hence, there may be a preference for more selective criteria. There are several ways to be more selective. One way is to consider discounting for discount factors that tend to 1. Another way is to use more subtle kinds of averaging. We will present some criteria and results. For all criteria it can be shown that optimal policies in class F exist and that these policies are (at least) average optimal.

A policy π_* is called *n-discount optimal* for some integer $n \geq -1$, if $\liminf_{\alpha\uparrow 1}(1-\alpha)^{-n}[v^\alpha(\pi_*) - v^\alpha(\pi)] \geq 0$ for all policies π. 0-discount optimality is also called *bias-optimality*. There is also the concept of *n-average optimality*. For any policy π, any $t \in \mathbb{N}$ and for $n = -1, 0, 1, \ldots$, let the vector $v^{n,t}(\pi)$ be defined by

$$v^{n,t}(\pi) := \begin{cases} v^t(\pi) & for \quad n = -1 \\ \sum_{s=1}^{t} v^{n-1,s}(\pi) & for \quad n = 0, 1 \ldots \end{cases} \tag{1.14}$$

π_* is said to be n-average optimal if $\liminf_{T\to\infty} \frac{1}{T}[v^{n,T}(\pi_*) - v^{n,T}(\pi)] \geq 0$ for all policies π.

A policy π_* is said to be *Blackwell optimal* if π_* is α-discounted optimal for all discount factors $\alpha \in [\alpha_0, 1)$ for some $0 \leq \alpha_0 < 1$. In a fundamental paper Blackwell [27] presented a mathematically rigorous proof for the policy iteration method to compute an α-discounted optimal policy. He also introduced the concept of bias-optimality (Blackwell called it *nearly optimality)* and established the existence of a discounted optimal policy for all discount factors sufficiently close to 1. In honor of Blackwell, such policy is called a Blackwell optimal policy.
It can be shown that n-discount optimality is equivalent to n-average optimality, that (-1)-discount optimality is equivalent to average optimality, and

that Blackwell optimality is n-discount optimality for all $n \geq N-1$, where $N = \#\mathbb{X}$ (in this chapter we will always use the notation N for the number of states).
The n-discount optimality criterion and the policy iteration method for finding an n-discount optimal policy, were proposed by Veinott [257]. He also showed that Blackwell optimality is the same as n-discount optimality for $n \geq N-1$. Sladky [223] has introduced the concept of n-average optimality; furthermore, he also showed the equivalence between this criterion and n-discount optimality. More details on bias optimality and Blackwell optimality can be found in Chapter 2 and Chapter 7.

1.1.4 Applications

White has published three papers on 'real applications' of Markov decision theory (White [280], [281] and [284]). Many stochastic optimization problems can be formulated as MDPs. In this section we shortly introduce the following examples: routing problems, stopping and target problems, replacement problems, maintenance and repair problems, inventory problems, the optimal control of queues, stochastic scheduling and multi-armed bandit problems. In this book there are also chapters on applications in finance (Chapter 14) and in telecommunication (Chapter 15). We also mention the contribution Chapter 16 on water reservoir applications.

Routing problems
In routing problems the problem is to find an optimal route through a network. Well known is the shortest path problem. A shortest path problem in a layered network can be formulated as an MDP over a finite horizon. Another application of this kind is the maximum reliability problem. In this network the connections are unreliable: let p_{ij} be the probability of reaching node j when the arc from node i to node j is chosen. The objective is to maximize the probability of reaching a terminal node n when the process is started in some node, say node 1. Results for a stochastic version of the shortest path problem can for instance be found in Bertsekas and Tsitsiklis [23]. The maximum reliability problem is discussed in Roosta [194].

Optimal stopping problems
In an optimal stopping problem there are two actions in each state. The first action is the stopping action and the second action corresponds to continue. If we continue in state i, a cost c_i is incurred and the probability of being in state j at the next time point is p_{ij}. If the stopping action is chosen in state i, then a final reward r_i is earned and the process terminates. In an optimal stopping problem, in each state one has to determine which action is chosen with respect to the total expected reward criterion. This kind of problem often has an optimal policy that is a so-called *control limit policy*.
The original analysis of optimal stopping problems appeared in Derman and Sacks [60], and Chow and Robbins [36]. A dynamic programming approach can be found in Breiman [28] who showed the optimality of a control limit policy.

Target problems
In a target problem one wants to reach a distinguished state (or a set of states) in some optimal way, where in this context optimal means, for instance, at minimum cost or with maximum probability. The target states are absorbing, i.e. there are no transitions to other states and the process can be assumed to terminate in the target states. These target problems can be modeled as MDPs with the total expected reward as optimality criterion. To the class of target problems we may count the so-called *first passage problem*. In this problem there is one target state and the objective is to reach this state (for the first time) at minimum cost. A second class of target problems are *gambling problems* (the gambler's goal is to reach a certain fortune N and the problem is to determine a policy which maximizes the probability to reach this goal). For more information about MDPs and gambling problems we refer to Chapter 12. The first passage problem was introduced by Eaton and Zadeh [67] under the name "pursuit problem". The dynamic programming approach was introduced in Derman [56]. A standard reference on gambling is Dubins and Savage [64]. Dynamic programming approaches are given in Ross [199] and Dynkin [66].

Replacement problems
Consider an item which is in a certain state. The state of the item describes its condition. Suppose that in each period, given the state of the item, the decision has to be made whether or not to replace the item by a new one. When an item of state i is replaced by a new one, the old item is sold at price s_i, a new item is bought at price c, and the transition to the new state is instantaneous. In case of nonreplacement, let p_{ij} be the probability that an item of state i is at the beginning of the next period in state j, and suppose that c_i is the maintenance cost—during one period—for an item of state i. This problem can be modeled as an MDP. It turns out that for the computation of an optimal policy an efficient algorithm, with complexity $\mathcal{O}(N^3)$, exists (see Gal [83]).
Next, we mention the model of deterioration with failure. In this model the states are interpreted as 'ages'. In state i there is a failure probability p_i and, when failure occurs, there is an extra cost f_i and the item has to be replaced by a new one. If there is no failure the next state is state $i+1$. It can be shown that, under natural assumptions about the failure probabilities and the costs, a control limit policy is optimal, i.e. there is an age i_* and the item is replaced by a new one if its age exceeds i_*. This property holds for the discounted reward criterion as well as for the average reward criterion.
There are a lot of references on replacement models. The early survey of Sherif and Smith [222] contained already over 500 references. Results on the optimality of control limit policies for replacement problems can be found in Derman [57, 58], Kolesar [146], Ross [198] and Kao [138].

Maintenance and repair problems
In maintenance and repair problems there is a system which is subject to deterioration and failure. Usually, the state is a characterization of the condition of the system. When the state is observed, an action has to be chosen, e.g. to keep the system unchanged, to execute some maintenance or repair, or to replace one or more components by new ones. Each action has corresponding

costs. The objective is to minimize the total costs, the discounted costs or the average costs. These problems can easily be modeled as an MDP.
A one-component problem is described in Klein [145]. The two-component maintenance problem was introduced by Vergin and Scriabin [259]. Other contributions in this area are e.g. Oezekici [173], and Van der Duyn Schouten and Vanneste [244]. An n-component series system is discussed in Katehakis and Derman [139]. Asymptotic results for highly reliable systems can be found in Smith [226], Katehakis and Derman [140], and Frostig [81].

Inventory problems
In inventory problems an optimal balance between inventory costs and ordering costs has to be determined. We assume that the probability distribution of the demand is known. There are different variants of the inventory problem. They differ, for instance, in the following aspects:
- stationary or nonstationary costs and demands;
- a finite planning horizon or an infinite planning horizon;
- backlogging or no backlogging.

For all these variants different performance measures may be considered.
In many inventory models the optimal policy is of (s, S)-type, i.e. when the inventory is smaller than or equal to s, then replenish the stock to level S. The existence of optimal (s, S)-policies in finite horizon models with fixed cost K is based on the so-called K-convexity, introduced by Scarf [202]. The existence of an optimal (s, S)-policy in the infinite horizon model is shown by Iglehart [126]. Another related paper is Veinott [255]. For the relation between discounted and average costs we refer to Hordijk and Tijms [119]. For the computation of the values s and S we refer to papers like Federgruen and Zipkin [76], and Zheng and Federgruen [292].

Optimal control of queues
Consider a queueing system where customers arrive according to a Poisson process and where the service time of a customer is exponentially distributed. Suppose that the arrival and service rates can be controlled by a finite number of actions. When the system is in state i, i.e. there are i customers in the system, action a means that the arrival or the service rates are $\lambda_i(a)$ or $\mu_i(a)$, respectively. The arrival and service processes are continuous-time processes. However, by the memoryless property of the exponential distribution, we can find an embedded discrete-time Markov chain which is appropriate for our analysis. This technique is called uniformization (see e.g. Tijms [241]).
A queue, or a network of queues, is a useful model for many applications, e.g. manufacturing, computer, telecommunication and traffic systems. See the survey of MDPs in telecommunication, Chapter 15. Control models can optimize certain performance measures by varying the control parameters of the system. We distinguish between *admission control* and *service rate control*.
In a service rate model, the service rate can be chosen from an interval $[0, \bar{\mu}]$. If rate μ is chosen, there are service costs $c(\mu)$ per period; we also assume that there are holding costs $h(i)$ per period when there are i customers in the system. Under natural conditions it can be shown that a *bang-bang policy* is optimal, i.e. $\mu = 0$ or $\mu = \bar{\mu}$. For details see Weber and Stidham [268]. Surveys of optimal

control of (networks of) queues can be found in the book by Walrand [265] and the papers by Stidham [234] and Stidham and Weber [235].

Stochastic scheduling
In a scheduling problem, jobs are processed on machines. Each machine can process only one job at a time. A job has a given processing time on the machines. In stochastic scheduling, these processing times are random variables. At certain time points decisions have to be made, e.g. which job is assigned to which machine. There are two types of models: the *customer assignment* models, in which each arriving customer has to be assigned to one of the queues (each queue with its own server) and *server assignment* models, where the server has to be assigned to one of the queues (each queue has its own customers).
Also in queueing models optimal policies often have a nice structure. Examples of this structure are:
- *μc-rule* : this rule assigns the server to queue k, with k the queue with $\mu_k c_k = \max_i\{\mu_i c_i \mid \text{queue } i \text{ is nonempty}\}$, where c_i is the cost which is charged per unit of time that the customer is in queue i and the service times in queue i are geometrically distributed with rate μ_i;
- *shortest queue policy (SQP)*: an arriving customer is assigned to the shortest queue;
- *longest expected processing time (LEPT)*: the jobs are allocated to the machines in decreasing order of their expected processing times;
- *shortest expected processing time (SEPT)*: the jobs are allocated to the machines in increasing order of their expected processing times.
The optimality of the μc-rule is established in Baras, Ma and Makowsky [9]. Ephremides, Varayia and Walrand [68] have shown the optimality of the shortest queue policy. The results for the optimality of the LEPT and SEPT policies are due to Bruno, Downey and Frederickson [30]. Related results are obtained by Weber [266] and by Chang, Hordijk, Righter and Weiss [33]. For reviews on stochastic scheduling we refer to Weiss [269], Walrand [265] (chapter 8) and Righter [193].

Multi-armed bandit problem
The multi-armed bandit problem is a model for dynamic allocation of a resource to one of n independent alternative projects. Any project may be in one of a finite number of states. At each period the decision maker has the option of working on exactly one of the projects. When a project is chosen, the immediate reward and the transition probabilities only depend on the active project and the states of the remaining projects are frozen. Applications of this model appear in machine scheduling, in the control of queueing systems and in the selection of decision trials in medicine. It can be shown that an optimal policy is the policy that selects the project which has the largest so-called *Gittins-index*. Fortunately, these indices can be computed for each project separately. As a consequence, the multi-armed bandit problem can be solved by a sequence of n one-armed bandit problems. This is a decomposition result by which the dimensionality of the problem is reduced considerably. Efficient algorithms for the computation of the Gittins indices exist. The most fundamental contribution on multi-armed bandit problems was made by Gittins (cf. Gittins and

Jones [86], and Gittins [85]). In Whittle [288] an elegant proof is presented. Other proofs are given by Ross [200], Varaiya, Walrand and Buyoccoc [254], Weber [267] and Tsitsiklis [243]. Several methods are developed for the computation of the Gittins indices: Varaiya, Walrand and Buyukkoc [254], Chen and Katehakis [35], Kallenberg [135], Katehakis and Veinott [141], Ben-Israel and S.D.Flåm [14], and Liu and Liu [155].

1.2 FINITE HORIZON

Consider an MDP with a finite horizon of T periods. In fact, we can analyze with the same effort a nonstationary MDP, i.e. with rewards and transition probabilities which may depend on the time t ($1 \leq t \leq T$). These nonstationary rewards and transition probabilities are denotated by $r^t(i,a)$ and $p^t(j|i,a)$. By the *principle of optimality*, an optimal policy can be determined by *backward induction* as the next theorem shows. The proof can be given by induction on the length T of the horizon. The use of the principle of optimality and the technique of dynamic programming for sequential optimization was provided by Bellman [13].

Theorem 1.2 *Let $x_i^{T+1} = 0, i \in \mathbb{X}$. Determine for $t = T, T-1, ..., 1$ a pure decision rule f^t such that*

$$[r^t(f^t)]_i + [P(f^t)x^{t+1}]_i = \max_{a \in \mathbb{A}(i)} \{r^t(i,a) + \sum\nolimits_j p^t(j|i,a) \cdot x_j^{t+1}\}, i \in \mathbb{X},$$

and let $x^t = r^t(f^t) + P^t(f^t)x^{t+1}$. Then, $R_* = (f^1, f^2, ..., f^T)$ is an optimal policy and x^1 is the value vector.

If $[r^t(f^t)]_i + [P^t(f^t)x^{t+1}]_i = \max_{a \in A(i)} \{r^t(i,a) + \sum_j p^t(j|i,a) \cdot x_j^{t+1}\}, i \in \mathbb{X}$, then we denote $r^t(f^t) + P^t(f^t)x = \max_{\mathbb{X} \times \mathbb{A}} \{r^t + P^t x\}$ and $f^t \in \underset{\mathbb{X} \times \mathbb{A}}{\operatorname{argmax}} \{r^t + P^t x\}$.

Algorithm I (finite horizon)

1. $x := 0$.
2. Determine for $t = T, T-1, \ldots, 1$:
   $f^t \in \underset{\mathbb{X} \times \mathbb{A}}{\operatorname{argmax}} \{r^t + P^t x\}$ and $x := r^t(f^t) + P^t(f^t)x$.
3. $R_* := (f^1, f^2, \ldots, f^T)$ is an optimal policy and x is the value vector.

Remarks

1. It is also possible to include in this algorithm *elimination of suboptimal actions* . Suboptimal actions are actions that will not occur in an optimal policy. References are Hastings and Van Nunen [99] and Hübner [124].
2. A finite horizon nonstationary MDP can be transformed in an equivalent stationary infinite horizon model. In such an infinite horizon model other options, as the treatment of *side constraints*, also called *additional constraints*, are applicable. These results can be found in Derman and Klein [59] and in Kallenberg [131], [132].

1.3 DISCOUNTED REWARD CRITERION

1.3.1 Introduction

In order to find an optimal policy and the value vector v^α, the so-called optimality equation

$$v_i^\alpha = \max_a \{r(i,a) + \alpha \sum\nolimits_j p(j|i,a) v_j^\alpha\}, i \in \mathbb{X} \tag{3.1}$$

plays a central role. Consider the mapping $T : \mathbb{R}^N \to \mathbb{R}^N$, defined by

$$[Tz]_i := \max_a \{r(i,a) + \alpha \sum\nolimits_j p(j|i,a) z_j\}, z \in \mathbb{R}^N, i \in \mathrm{X}. \tag{3.2}$$

It turns out that T is a *monotone contraction mapping* with as fixed point the value vector v^α. We also introduce, for any stationary policy π, the mapping $T_\pi : \mathbb{R}^N \to \mathbb{R}^N$, defined by (in vector notation)

$$T_\pi z := r(\pi) + \alpha P(\pi) z. \tag{3.3}$$

Let $f_z(i) \in \underset{\mathbb{A}(i)}{\operatorname{argmax}} \{r(i,a) + \alpha \sum_j p(j|i,a) z_j\}$, then

$$T_{f_z} z = Tz = \max_\pi T_\pi z. \tag{3.4}$$

First, we summarize some well-known results on discounted MDPs. The proofs can be found in the standard MDP textbooks. They are based on the theory of monotone contraction mappings. For this theory we refer to the book written by Stoer and Bulirsch [236].

Theorem 1.3 *With respect to the norm $\|\cdot\|_\infty, T_\pi$ and T are monotone contraction mappings in $\mathbb{R}^N$ with contraction factor α.*

Theorem 1.4 *For any stationary policy $\pi, v^\alpha(\pi)$ is the unique solution of the functional equation $T_\pi z = \mathrm{z}$.*

Corollary 1.2 *$v^\alpha(\pi) = \lim_{n\to\infty} T_\pi^n z$ for any $z \in \mathbb{R}^N$.*

Theorem 1.5 *v^α is the unique solution of the equation $Tz = z$.*

Because, $v^\alpha = Tv^\alpha = T_{f_{v^\alpha}} v^\alpha$, the last equality by (3.4), it follows from Theorem 1.5 that $v^\alpha = v^\alpha(f_{v^\alpha})$, i.e. f_{v^α} is an optimal policy. If f satisfies $r(i,f(i)) + \alpha \sum_j p(j|i,f(i)) v_j^\alpha = \max_a \{r(i,a) + \alpha \sum_j p(j|i,a) v_j^\alpha\}, i \in \mathbb{X}$, then f is called a *conserving* policy. f_{v^α} is a conserving policy and conserving policies are optimal. Therefore, the equation $Tz = z$ is called the optimality equation.

Corollary 1.3
(i) There exists a pure stationary optimal policy.
(ii) $v^\alpha = \lim_{n\to\infty} T^n z$ *for any* $z \in \mathbb{R}^N$.
(iii) Any conserving policy is optimal.

We use the following result from the theory of contracting mappings.

Lemma 1.2 *Let* B *be a monotone contraction in* $\mathbb{R}^N$ *with respect to* $\|\cdot\|_\infty$ *, with contraction factor* β*, fixed-point* z^* *and with the property that* $B(z+c\cdot e) = Bz + \beta c \cdot e$ *for every* $z \in \mathbb{R}^N$ *and scalar* c*. Suppose that for some scalars* a *and* b *and for some* $z \in \mathbb{R}^N$*,* $a \cdot e \le Bz - z \le b \cdot e$*. Then,* $z + (1-\beta)^{-1} a \cdot e \le Bz + \beta(1-\beta)^{-1} a \cdot e \le z^* \le Bz + \beta(1-\beta)^{-1} b \cdot e \le z + (1-\beta)^{-1} b \cdot e$.

Since $T_f(z + c\cdot e) = T_f z + \alpha c \cdot e$ and $T(z + c\cdot e) = Tz + \alpha c \cdot e$ for any $z \in \mathbb{R}^N$ and any scalar c, we can apply Lemma 1.2 to obtain bounds for the fixed points $v^\alpha(f)$ and v^α of the operators T_f and T, respectively.

Lemma 1.3 *For any* $z \in \mathbb{R}^N$ *we have*
(i) $z + (1-\alpha)^{-1} \min_i (Tz - z)_i \cdot e \le Tz + \alpha(1-\alpha)^{-1} \min_i (Tz-z)_i \cdot e \le v^\alpha(f_z) \le v^\alpha \le Tz + \alpha(1-\alpha)^{-1} \max_i (Tz-z)_i \cdot e \le z + (1-\alpha)^{-1} \max_i (Tz-z)_i \cdot e$.
(ii) $\| v^\alpha - v^\alpha(f_z) \|_\infty \le \alpha(1-\alpha)^{-1} \mathrm{span}(Tz - z)$*, where* $\mathrm{span}(z)$ *is defined by* $\mathrm{span}(z) := \max_i z_i - \min_i z_i$.

An action $a \in \mathbb{A}(i)$ is called *suboptimal* if there does not exist an optimal policy f with $f(i) = a$. Because f is optimal if and only if $v^\alpha(f) = v^\alpha$, and because $v^\alpha = Tv^\alpha$, an action $a \in \mathbb{A}(i)$ is suboptimal if and only if

$$v_i^\alpha > r(i,a) + \alpha \sum\nolimits_j p(j|i,a) v_j^\alpha. \tag{3.5}$$

Suboptimal actions can be excluded. Not directly by (3.5), because v^α is unknown, but by using the bounds on v^α as given by Lemma 1.3. Then, by the monotonicity of T, the next result is obtained.

Theorem 1.6
(i) Suppose that $x \le v^\alpha \le y$*. If* $r(i,a) + \alpha \sum_j p(j|i,a) y_j < (Tx)_i$*, then action* $a \in \mathbb{A}(i)$ *is suboptimal.*
(ii) Suppose that for some scalars b *and* c*,* $x + b\cdot e \le v^\alpha \le x + c \cdot e$*. If* $r(i,a) + \alpha \sum_j p(j|i,a) x_j < (Tx)_i - \alpha(c-b)$*, then action* $a \in \mathbb{A}(i)$ *is suboptimal.*

Using the bounds for v^α from Lemma 1.3, we obtain suboptimality for an action $a \in \mathbb{A}(i)$ if

$$r(i,a) + \alpha \sum\nolimits_j p(j|i,a) z_j < (Tz)_i - \alpha(1-\alpha)^{-1} \mathrm{span}(Tz - z) \tag{3.6}$$

or

$$r(i,a) + \alpha \sum\nolimits_j p(j|i,a) (Tz)_j < (T^2 z)_i - \alpha^2(1-\alpha)^{-1} \mathrm{span}(Tz - z) \tag{3.7}$$

Remark
If we relax the property that $\sum_j p(j|i,a) = 1$ to $\sum_j p(j|i,a) \leq 1$ for all (i,a) and require that the model is *transient*, i.e. the matrix $\sum_{t=1}^{\infty} [P(f)]^t$ has finite elements for every policy f, then the total expected reward criterion, i.e. the discounting case with discount factor $\alpha = 1$, is well-defined. For this criterion similar results can be obtained as in the discounted model. The investigation whether an MDP is transient can be done efficiently (cf. Veinott [257] and Kallenberg [134]). Other references on this topic are van Hee, Hordijk and van der Wal [248], Denardo and Rothblum [53], and Hordijk and Kallenberg [115].

Already in 1953, Shapley [221] analyzed contraction properties for stochastic games. In the special case of a one-player game a stochastic game becomes an MDP. A comprehensive treatment of the theory of contraction mappings for discounted Markov decision processes was given by Denardo [45]. The details of the proof of Theorem 1.5 can be found in Ross [198]. An alternative proof that T has a fixed point, based on Brouwer's theorem, was given in Shapiro [220]. The concepts 'conserving' and 'span' were introduced by Dubins and Savage [64] and by Bather [10]. Concerning the bounds of Lemma 1.3, the weakest bounds were proposed by MacQueen [162] and the strongest by Porteus [179]. Related papers are Porteus [180] and Bertsekas [16]. The notion that suboptimal actions can be excluded if bounds on the value vector are available can be found in MacQueen [163], which paper includes the test (3.6). Test (3.7) is proposed in Porteus [179]. Other suboptimality tests can be found in Hastings and Mello [97], White [278] and Thomas [239].

1.3.2 Policy iteration

For $x, y \in \mathbb{R}^N$, $x > y$ means that $x_i \geq y_i$ for every i and $x_i > y_i$ for at least one i. In the method of policy iteration a sequence of pure stationary policies $f_1, f_2, \ldots$ is constructed such that

$$v^{\alpha}(f_{k+1}) > v^{\alpha}(f_k) \text{ for } k = 1, 2, \ldots \tag{3.8}$$

Because there are finitely many pure stationary policies, the method of policy iteration is finite. Furthermore, it can be shown that the method terminates with an α-discounted optimal policy. We first remark that the following lemma is a consequence of Theorem 1.3.

Lemma 1.4
(i) If $T_f z \leq z$, then $v^{\alpha}(f) = \lim_{n\to\infty} T_f^n z \leq T_f z \leq z$.
(ii) If $T_f z > z$, then $v^{\alpha}(f) = \lim_{n\to\infty} T_f^n z \geq T_f z > z$.

For every $i \in \mathbb{X}$ and every $f \in F$, the set $A(i,f)$ is defined by

$$A(i,f) := \{a \in \mathbb{A}(i) \mid r(i,a) + \alpha \sum\nolimits_j p(j|i,a) v_j^{\alpha}(f) > v_i^{\alpha}(f)\}. \tag{3.9}$$

The intuitive idea of policy iteration is that if action $f(i)$ is replaced by an action $a \in A(i,f)$ the resulting policy improves the α-discounted rewards.

Therefore, the actions of $A(i,f)$ are called *improving actions*. The correctness of this idea is established by the following theorem.

Theorem 1.7
(i) If $A(i,f) = \emptyset$ for every $i \in \mathbb{X}$, then f is α-discounted optimal.
(ii) If $A(i,f) \neq \emptyset$ for some $i \in \mathbb{X}$, then $v^\alpha(g) > v^\alpha(f)$ for any $g \in F$ with $g \neq f$ and $g(i) \in A(i,f)$ if $g(i) \neq f(i)$.

Let

$$s_{ia}(f) := r(i,a) + \alpha \sum\nolimits_j p(j|i,a) v_j^\alpha(f) - v_i^\alpha(f), \ a \in \mathbb{A}(i) \text{ and } i \in \mathbb{X}. \quad (3.10)$$

Algorithm II (policy iteration; discounted rewards)

1. Start with any $f \in F$.
2. Compute $v^\alpha(f)$ as unique solution of the linear system $T_f z = \mathrm{z}$.
3. $A(i,f) := \{a \in \mathbb{A}(i) \mid s_{ia}(f) > 0\}$ for every $i \in \mathbb{X}$.
4. If $A(i,f) = \emptyset$ for every $i \in \mathbb{X}$: go to step 6.
   Otherwise: take any $g \neq f$ such that, if $g(i) \neq f(i), g(i) \in A(i,f)$.
5. $f := g$ and go to step 2.
6. f is an α-discounted optimal policy.

The idea to use policy iteration to determine an optimal policy appeared in Howard [121]. Blackwell [27] has provided a strong mathematical treatment of this method. In Porteus [183] and in Hartley, Lavercombe and Thomas [92] efficient ways are analyzed in order to calculate $v^\alpha(f)$ as solution of the linear system $T_f z = z$.

Remarks
1. There is some freedom in the choice of policy g in step 4. A usual choice is to take g such that $s_{ig(i)}(f) = \max_a s_{ia}(f)$, i.e. $g(i) \in \operatorname{argmax}_a \ s_{ia}(f)$.
2. It can be shown (see Puterman and Brumelle [187]) that the policy iteration method, with the above choice for g, is equivalent to solving the optimality equation $Tz = z$ by Newton's method.
3. Furthermore, we can derive a result on the convergence rate. It can be shown that $x^n = v^\alpha(f_n), n = 1, 2, \ldots$, where x^n are the iterates of the Newton method and f_n the policies of the policy iteration method. Since it can be shown that $\| v^\alpha - v^\alpha(f_{n+1}) \|_\infty \leq 2\alpha(1-\alpha)^{-1} \| v^\alpha - v^\alpha(f_n) \|_\infty$, there is geometric convergence. Already in Pollatschek and Avi-Itzhak [178], in the context of stochastic games, the equivalence between the policy iteration method and Newton's method was noticed. A related paper is Schweitzer [211]. Puterman and Brumelle [187] were the first who derived results for the rate of convergence.
4. One can also exclude suboptimal actions. E.g. by using test (3.6) with $z = v^\alpha(f)$. Since, for $z = v^\alpha(f)$,

$(Tz - z)_i = \max_a \{r(i,a) + \sum_j p(j|i,a)v_j^\alpha(f) - v_i^\alpha(f) = \max_a s_{ia}(f), i \in \mathbb{X}$, we have $\text{span}(Tz - z) = \max_i[\max_a \ s_{ia}(f)] - \min_i[\max_a \ s_{ia}(f)]$. Hence (3.6) becomes:

if $s_{ib}(f) < \max_a \ s_{ia}(f) - \alpha(1-\alpha)^{-1}[\max_i \max_a s_{ia}(f) - \min_i \max_a \ s_{ia}(f)]$, then action $b \in \mathbb{A}(i)$ is suboptimal.

Grinold [91] pointed out that suboptimality tests can be implemented in policy iteration. The above test is stronger than Grinold's test.

5. The following modification, which was shown to be correct by Hastings [94], often gives faster convergence. Instead of the steps 3 and 4, the steps 3' and 4' are used, where:

Step 3':

For i = 1 to N do

a. $d_{ia}(f) := r(i,a) + \alpha\sum_{j=1}^{i-1} p(j|i,a)z_j + \alpha\sum_{j=i}^{N} p(j|i,a)v_j^\alpha(f), a \in \mathbb{A}(i)$;

b. if $d_{ia}(f) \le v_i^\alpha(f)$ for every $a \in \mathbb{A}(i)$:

$z_i := v_i^\alpha(f)$ and $g(i) := f(i)$;

c. if $d_{ia}(f) > v_i^\alpha(f)$ for some $a \in \mathbb{A}(i)$:

$z_i := \max_a \ d_{ia}(f)$ and take $g(i) \in \text{argmax} \ d_{ia}(f)$.

Step 4':

If $g(i) = f(i)$ for every $i \in \mathbb{X}$, then go to step 6.

6. In Schmitz [204] the question is raised: "Does there exist a polynomial bound for the number of iterations in the policy iteration?". Meister and Holzbaur [165] have shown that this method is polynomial in time. In Ng [171] is shown that the complexity of one iteration is $\mathcal{O}(mN^2)$, where m is the number of states i for which $g(i) \neq f(i)$.

1.3.3 Linear programming

A vector $v \in \mathbb{R}^N$ is said to be *α-superharmonic* if

$$v_i \ge r(i,a) + \alpha\sum\nolimits_j p(j|i,a)v_j \text{ for every } (i,a) \in \mathbb{X} \times \mathbb{A}. \tag{3.11}$$

Theorem 1.8 *v^α is the (componentwise) smallest α-superharmonic vector.*

Corollary 1.4 *v^α is the unique optimal solution of the LP-problem*

$$\min\left\{\sum\nolimits_j \beta_j v_j \mid \sum\nolimits_j [\delta_{ij} - \alpha p(j|i,a)]v_j \ge r(i,a), (i,a) \in \mathbb{X} \times \mathbb{A}\right\} \tag{3.12}$$

where $\beta_j > 0$ for every $j \in \mathbb{X}$.

By Corollary 1.4, the value vector v^α can be found as the optimal solution of the linear program (3.12). This program does not give an optimal policy. However, an optimal policy can be obtained from the solution of the dual program:

$$\max\left\{\sum\nolimits_i\sum\nolimits_a r(i,a)x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a[\delta_{ij} - \alpha p(j|i,a)] & x_{ia} = \beta_j, j \in \mathbb{X} \\ & x_{ia} \ge 0, (i,a) \in \mathbb{X} \times \mathbb{A} \end{array}\right\} \tag{3.13}$$

Theorem 1.9 *Let x^* be an optimal solution of (3.13). Then, a policy f with $x^*_{jf(j)} > 0$ for every $j \in \mathbb{X}$ exists and is an optimal policy.*

There is a one-to-one correspondence between the set of feasible solutions of (3.13) and the set of stationary policies, given by the following relations. For a stationary policy π the feasible solution $x(\pi)$ satisfies

$$x_{ia}(\pi) := [\beta^T (I - \alpha P(\pi))^{-1}]_i \cdot \pi_{ia}, (i,a) \in \mathbb{X} \times \mathbb{A}. \tag{3.14}$$

Conversely, for a feasible solution x of (3.13), define $\pi(x)$ by

$$\pi_{ia}(x) := x_{ia} / \sum\nolimits_a x_{ia}, (i,a) \in \mathbb{X} \times \mathbb{A}. \tag{3.15}$$

Theorem 1.10 *The mapping (3.14) is a one-to-one mapping of the set of stationary policies onto the set of feasible solutions of the dual program (3.13) with (3.15) as the inverse mapping; furthermore, the set of extreme feasible solutions of (3.13) corresponds to the set F of pure stationary policies.*

Algorithm III (linear programming; discounted rewards)

1. Take any $\beta \in \mathbb{R}^N$ with $\beta_j > 0, j \in \mathbb{X}$.
2. Compute optimal solutions v^* and x^* of the dual pair LP-problems (3.12) and (3.13).
3. Take any f_* such that $x^*_{if_*(i)} > 0$, $i \in \mathbb{X}$.
4. v^* is the value vector and f_* is an α-discounted optimal policy.

It turns out that the linear programming method is, in some sense, equivalent to policy iteration. This is formulated in the next theorem, in which the term *block-pivoting* simplex algorithm is used. A simplex LP-algorithm, which in one iteration more than one pivot step may use, is called a block-pivoting simplex algorithm (cf. Dantzig [40]).

Theorem 1.11
(i) Any policy iteration algorithm is equivalent to a block-pivoting simplex algorithm.
(ii) Any simplex algorithm is equivalent to a particular policy iteration algorithm.

Remarks
1. Since the LP-method and policy iteration are equivalent, exclusion of suboptimal actions can also be implemented in the LP-method. The relevant data $s_{ia}(f)$ for this test (see (3.10)) are available in the simplex tableaus as the so-called reduced costs.
2. The variables $x_{ia}(\pi)$, defined in (3.14) can be interpreted as *discounted state-action frequencies*, i.e. if policy π is used, then $x_{ia}(\pi)$ is equal to the total

expected discounted number of times that state i is visited and then also action a is chosen, given that the starting state is state j with probability $\beta_j, j \in \mathbb{X}$.
3. The linear programming method is the only method which can easily handle additional constraints. Constrained optimization arises in many MDP applications, e.g. in inventory and queueing models. For examples we refer to Derman [58], chapter 7, and to Puterman [186], section 8.9. The constraints have to be expressed in terms of the state-action frequencies and added to the dual program (3.13). Constrained problems have a stationary, but not necessarily pure optimal policy. For the details we refer to Kallenberg [134], and to Hordijk and Kallenberg [115]. In Altman and Shwartz [4] the sensitivity of constrained MDPs is investigated. Altman, Hordijk and Kallenberg [3] have analyzed the behavior of the value function in constrained MDP.

The idea to use linear programming to compute an optimal policy originated with D'Epenoux [55]. The one-to-one correspondence between the feasible solutions of the dual program and the set of stationary policies can be found in De Ghellinck and Eppen [43]. The equivalence between block-pivoting and policy iteration was mentioned in De Ghellinck [42]. The implementation of the suboptimality tests was proposed by Grinold [91] and by Hordijk and Kallenberg [115]. In Sun [238] an implementation of the LP-method is described, based on the revised simplex method. Stein [233] has investigated the computational aspects of the linear programming method in comparison with other methods as policy iteration, modified policy iteration and value iteration. It turns out that the LP-method is preferable if the discount factor is close to unity and the state space is not too large. Chapter 11 deals with the linear programming approach for MDPs with infinite state and action spaces.

1.3.4 Value iteration

In the value iteration method the value vector v^α is approximated by a sequence $\{v^n\}_{n=1}^\infty$, which converges to v^α and in this way a nearly optimal policy is obtained. For $\epsilon > 0$, a vector $v \in \mathbb{R}^N$ is an *ϵ-approximation* of v^α if $\| v^\alpha - v \|_\infty \leq \epsilon$; a policy R is an *ϵ-optimal policy* if $\| v^\alpha - v^\alpha(R) \|_\infty \leq \epsilon$, i.e. $v^\alpha(R)$ is an ϵ-approximation of v^α. From Corollary 1.3 (ii) it follows that $v^\alpha = \lim_{n\to\infty} T^n x$ for every $x \in \mathbb{R}^N$.

Define the sequence $\{v^n\}_{n=1}^\infty$ by

$$\begin{cases} v^1 \in \mathbb{R}^N & \text{arbitrarily chosen} \\ v^{n+1} := Tv^n, & n = 1, 2, \ldots \end{cases} \tag{3.16}$$

with a corresponding sequence $f_1, f_2, \ldots$ of policies where $f_n = f_{v^n}$ for every $n \in \mathbb{N}$, i.e.

$$v^{n+1} = Tv^n = T_{f_n} v^n = r(f_n) + \alpha P(f_n) v^n, n \in \mathbb{N}. \tag{3.17}$$

The next lemma shows that f_n is an ϵ-optimal policy for n sufficiently large. The proof is based on contraction properties.

Lemma 1.5 $\| v^\alpha(f_n) - v^\alpha \|_\infty \leq 2\alpha^n(1-\alpha)^{-1} \cdot \| v^2 - v^1 \|_\infty, n \in \mathbb{N}.$

Algorithm IV (value iteration; discounted rewards)

1. Choose $\epsilon > 0$ and $x \in \mathbb{R}^N$ arbitrarily.
2. Compute $y = Tx$ and take $f = f_x$.
3. If $\| y - x \|_\infty \leq (1-\alpha)\alpha^{-1}\epsilon$, then f is a 2ϵ-optimal policy and y is an ϵ-approximation of v^α (Stop);

   Otherwise: $x := y$ and go to step 2.

The correctness of algorithm IV is a consequence of the next theorem, which also follows from the contraction properties.

Theorem 1.12
(i) $\| v^\alpha(f_x) - v^\alpha \|_\infty \leq 2\alpha(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty$;
(ii) $\| Tx - v^\alpha \|_\infty \leq \alpha(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty$.

In the next theorem we summarize some suboptimality tests.

Theorem 1.13 *An action $a \in \mathbb{A}(i)$ is suboptimal if one of the following tests is satisfied:*

$$r(i,a) + \alpha \sum\nolimits_j p(j|i,a)x_j < (Tx)_i - 2\alpha(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty \tag{3.18}$$

$$r(i,a) + \alpha \sum\nolimits_j p(j|i,a)(Tx)_j < (T^2x)_i - 2\alpha^2(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty \tag{3.19}$$

$$r(i,a) + \alpha \sum\nolimits_j p(j|i,a)x_j < (Tx)_i - \alpha(1-\alpha)^{-1}\mathrm{span}(Tx - x) \tag{3.20}$$

$$\begin{aligned} r(i,a) + \alpha \textstyle\sum_j p(j|i,a)(Tx)_j < \quad & (Tx)_i + \alpha(1-\alpha)^{-1} \min_i (Tx-x)_i \\ & -\alpha^2(1-\alpha)^{-1} \max_i (Tx-x)_i \end{aligned} \tag{3.21}$$

Remarks
1. In the usual computation scheme of the value iteration algorithm, test (3.20) is the best available test.
2. We also mention two variants of the standard algorithm. In the *Pre-Gauss-Seidel* variant we use for the computation of y_i the components y_j which are already computed, i.e.

$$y_i = \max_a \{ r(i,a) + \alpha \sum\nolimits_{j=1}^{i-1} p(j|i,a)y_j + \alpha \sum\nolimits_{j=i}^{N} p(j|i,a)x_j \}, \quad i = 1, 2, \ldots, N \tag{3.22}$$

In the *Gauss-Seidel* variant also the i-th component x_i is replaced by y_i, which gives

$$\begin{aligned} y_i = \max_a [1 - \alpha p(i|i,a)]^{-1} \cdot \{ & r(i,a) + \alpha \sum\nolimits_{j=1}^{i-1} p(j|i,a)y_j \\ & + \alpha \sum\nolimits_{j=i+1}^{N} p(j|i,a)x_j \}, \ i = 1, 2, \ldots, N \end{aligned} \tag{3.23}$$

For both variants it can be shown that the corresponding operators are contraction mappings with fixed point v^α and with contraction factor at most α. Hence, they may be considered as an acceleration of the basic algorithm. Suboptimaly tests for the exclusion of actions can also be included in the value iteration method.

Value iteration goes back to the seminal paper of Shapley [221] on stochastic games. For a survey of the basic properties of value iteration we refer to Federgruen and Schweitzer [70]. The idea to accelerate the convergence by the pre-Gauss-Seidel method was proposed in Hastings [94]. The Gauss-Seidel method can be found in Kushner and Kleinman [149]. An overview of these variants is presented in Porteus [182]. Other techniques, based on successive overrelaxation and stopping times, in order to accelerate the convergence can be found in Reetz [190] and [191], Schellhaas [203], Wessels [272], Van Nunen [249], Van Nunen and Wessels [252], [253], Porteus and Totten [184], Porteus [181], Herzberg and Yechiali [105], and Bertsekas [20]. Holzbaur [110] has presented a theoretically polynomial bound for the number of steps in the value iteration method.

1.3.5 Modified policy iteration

In section 1.3.2 the policy iteration method was discussed. This method, with the usual choice for the improving actions, can be considered as Newton's method for the solution of the optimality equation. A new iterand y is obtained from x by the formula

$$y = x + A(Tx - x), \text{ where } A = [I - \alpha P(g)]^{-1} \text{ with } g \text{ such that } T_g x = Tx \tag{3.24}$$

The determination of $[I-\alpha P(g)]^{-1}$, which is equal to $\sum_{i=0}^{\infty} \alpha^i [P(g)]^i$, requires in general a lot of work. In the *modified policy iteration method* the matrix A is truncated by

$$A^{(k)} = \sum_{i=0}^{k-1} \alpha^i [P(g)]^i \text{ for some } 1 \leq k \leq \infty\,. \tag{3.25}$$

For $k = 1$, $A^{(k)} = I$ and the value iteration method is obtained; for $k = \infty, A^{(k)} = A$, and we have policy iteration. For $1 < k < \infty$, the modified policy iteration method can be considered as a combination of policy iteration and value iteration, or as an inexact Newton method for the solution of the optimality equation.

We may allow that in each iteration another value of k is chosen, and we denote $k(n)$ for the value in iteration n. Hence, we obtain the following iteration scheme, where f_n is the policy in iteration n, i.e. $T_{f_n} x^n = Tx^n$.

$$\begin{aligned} x^{n+1} &= x^n + A^{(k(n))}(Tx^n - x^n) \\ &= x^n + \textstyle\sum_{i=0}^{k(n)-1} \alpha^i P^i(f_n)[r(f_n) + \alpha P(f_n)x^n - x^n] \\ &= r(f_n) + \alpha P(f_n) r(f) + \cdots + [\alpha P(f_n)]^{k(n)-1} r(f_n) + [\alpha P(f_n)]^{k(n)} x^n \\ &= T_{f_n}^{k(n)} x^n. \end{aligned}$$

Algorithm V (modified policy iteration; discounted rewards)

1. Choose $x \in \mathbb{R}^N$, $\epsilon > 0$ and $f \in F$.

2. a. Choose k with $1 \leq k \leq \infty$;

   b. Determine g such that $T_g x = Tx$, where $g(i) = f(i)$ if possible.

3. If $\| Tx - x \|_\infty \leq (1-\alpha)\epsilon$: g is an 2ϵ-optimal policy and Tx is an $\alpha\epsilon$-approximation of v^α (Stop);

   Otherwise: $x := T_g^k x$, $f := g$ and go to step 2.

Remarks
1. Since $x^{n+1} = T_{f_n}^{k(n)} x^n$, the iteration operator depends on n, and it is not obvious that this operator is monotone and/or contracting. Indeed, in general, this operator is neither a contraction nor monotone. Nevertheless, it can be shown that $v^\alpha = \lim_{n\to\infty} T_{f_n}^{k(n)} x^n$ for any starting vector x^1.
2. Also in this method, it is possible to implement tests for the exclusion of suboptimal actions.

Puterman and Shin [188] and independently Van Nunen [249], [250] and [251] have developed the modified policy iteration method. The first authors have shown the convergence under the assumption that the starting vector x satisfies $Tx \geq x$. The convergence of the method for an arbitrary starting vector was proved by Rothblum [201]. In Van Nunen [249] an example is given which shows that the operator of the modified policy iteration method can be neither contracting nor monotonic. The observation that the modified policy iteration method can be viewed as an inexact Newton method was made by Dembo and Haviv [44]. The exclusion of suboptimal actions for this method was developed by Puterman and Shin [189]. Puterman [185] reviews computational results for the modified policy iteration method.

1.4 AVERAGE REWARD CRITERION

1.4.1 Introduction

We start this section with some properties of a transition matrix P. The *stationary matrix* P^* of P is defined by the Cesaro-limit of P^n, i.e.

$$P^* := \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P^{k-1} \tag{4.1}$$

The next theorem summarizes some properties of the stationary matrix.

Theorem 1.14
(i) $P^*P = PP^* = P^*P^* = P^*$.
(ii) $[P - P^*]^n = P^n - P^*$, $n \geq 1$.
(iii) $\lim_{\alpha\uparrow 1}(1-\alpha)\sum_{n=0}^{\infty} \alpha^n(P^n - P^*) = 0$.
(iv) $[I - P + P^*]^{-1} = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k}[P - P^*]^{i-1}$.
(v) For every stationary policy π, *the average reward* $\phi(\pi)$ *satisfies* $\phi(\pi) = P^*(\pi)r(\pi)$, *where* $P^*(\pi)$ *is the stationary matrix of the transition matrix* $P(\pi)$.

(vi) $\phi(\pi) = \lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(\pi)$.

$[I - P + P^*]^{-1}$ is denoted by Z and is called the *fundamental matrix*. The *deviation matrix* D is defined by

$$D := Z - P^* \tag{4.2}$$

Theorem 1.15
(i) $D = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k}[P^{i-1} - P^*]$
(ii) $P^*D = DP^* = [I - P]D + P^* - I = D[I - P] + P^* - I = 0$.

For the proofs of the Theorems 1.14 and 1.15 we refer to books on Markov chains (e.g. Kemeny and Snell [144]). A treatment, related to MDPs, of the stationary, the fundamental and the deviation matrix can be found in Veinott [258]. For a stationary policy π, the deviation matrix of the transition matrix $P(\pi)$ is denoted by $D(\pi)$.

We continue this section with a classification of MDPs based on the ergodic structure. We distinguish between multichain, unichain and irreducible MDPs. The reason for this distinction is that MDPs can be analyzed easier in case they are unichain or irreducible, which may lead to simplified algorithms for solving these MDPs. We assume the reader is familiar with concepts as recurrent state, transient state, recurrent class, irreducibility, unichain and multichain. The determination whether an MDP is irreducible is an easy, i.e. polynomially solvable, problem (the number of steps is bounded by a polynomial function of the problem's data, see Kallenberg [137]).

Open problem
Does there exist a polynomial algorithm to determine whether an MDP is unichain or multichain?

Next, we will formulate a theorem on the existence of a Blackwell optimal policy f_0, i.e. f_0 is α-discounted optimal for all discount factors $\alpha \in [\alpha_0, 1)$ for some $0 \le \alpha_0 < 1$. The next theorem shows even more, namely that the interval $[0, 1)$ can be partitioned in a finite number of subintervals such that in each subinterval there exists a policy which is discounted optimal over the whole subinterval. Since a proof of this result cannot be found in the textbooks on MDPs, we include an outline of this proof.

Theorem 1.16 *There are numbers* $\alpha_m, \alpha_{m-1}, ..., \alpha_0, \alpha_{-1}$ *and policies* f_m, $f_{m-1}, ..., f_0$ *such that* $0 = \alpha_m < \alpha_{m-1} < \cdots < \alpha_0 < \alpha_{-1} = 1$ *and* $v^\alpha(f_j) = v^\alpha$ *for all* $\alpha \in [\alpha_j, \alpha_{j-1}), j = m, m-1, ..., 0$.

Proof (outline). Since $v^\alpha(f)$ is the solution of the system $[I - \alpha P(f)]x = r(f)$, each component $v_i^\alpha(f)$ is a rational function in α. Suppose that a Blackwell optimal policy does not exist. Since for any fixed α a deterministic α-discounted optimal policy exists, this implies that there are series $\{\alpha_k \mid k = 1, 2, \dots\}$ and $\{f_k \mid k = 1, 2, \dots\}$ such that $\alpha_1 \le \alpha_2 \le \cdots$ with $\lim_{k\to\infty} \alpha_k = 1$ and $v^\alpha = v^\alpha(f_k) > v^\alpha(f_{k-1})$ for $\alpha = \alpha_k, k = 2, 3, \dots$. Because F is finite, there

are two pure policies, say f and g, that both are in turn optimal for an infinite number of increasing α's with limit $\alpha = 1$. Let $h(\alpha) = v^\alpha(f) - v^\alpha(g)$, then for any $i \in \mathbb{X}$, $h_i(\alpha)$ is a continuous rational function in α on $[0,1)$, which has an infinite number of zeros. This contradicts the rationality of $h_i(\alpha)$. Hence, there exists a Blackwell optimal policy. With similar arguments, it can be shown that for each fixed $\alpha \in (0,1]$ there is an interval around α and a policy which is optimal in that interval. These intervals are a covering of the closed bounded set $[0,1]$. Hence, by the Heine-Borel-Lebesque theorem, it follows that there is a covering by a finite number of intervals. ∎

We close this section with the *Laurent expansion* of a stationary policy π.

Theorem 1.17 *Let $u^k(\pi), k = -1, 0, \ldots$ be defined by $u^{-1}(\pi) = P^*(\pi)r(\pi)$, $u^0(\pi) = D(\pi)r(\pi)$ and $u^{k+1}(\pi) = -D(\pi)u^k(\pi)$, $k \geq 0$. Then, for $\alpha_0(\pi) < \alpha < 1$, we have $v^\alpha(\pi) = \alpha^{-1}\sum_{k=-1}^{\infty}[(1-\alpha)/\alpha]^k \cdot u^k(\pi)$, where $\alpha_0(\pi) = \| D(\pi) \| / [\| D(\pi) \| + 1]$.*

Corollary 1.5
(i) $\phi(\pi) = \lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(\pi)$.
(ii) $v^\alpha(\pi) = \frac{\phi(\pi)}{1-\alpha} + u^0(\pi) + \epsilon(\alpha)$, *where* $\lim_{\alpha\uparrow 1}\epsilon(\alpha) = 0$.

The first part of the Laurent expansion as presented in Corollary 1.5 (ii) was derived by Blackwell [27]. The complete Laurent expansion of Theorem 1.17 was proposed by Miller and Veinott [166]. The vector $u^0(\pi)$ is called the *bias vector* of policy π.

1.4.2 The optimality equation

A. The multichain case.
Before we introduce the optimality equation, we first give some prerequisites.

Lemma 1.6 $\lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(R) \geq \phi(R)$ *for any policy R.*

The proof of this theorem is based on Tauberian arguments which can be found in Derman [58] or Hordijk [111].

Corollary 1.6 *Any stationary Blackwell optimal policy is average optimal.*

In the discounted case, the value vector is the unique solution of an optimality equation. A similar result holds for the average reward criterion, but the derivation is more complex.

Theorem 1.18 *Consider the system*

$$\begin{cases} x_i & = \max_{a\in\mathbb{A}(i)} \sum_j p(j|i,a)x_j, & i \in \mathbb{X} \\ x_i + y_i & = \max_{a\in\mathbb{A}(i,x)} \{r(i,a) + \sum_j p(j|i,a)y_j\}, & i \in \mathbb{X} \end{cases} \tag{4.3}$$

where $\mathbb{A}(i,x) := \{a \in \mathbb{A}(i) \mid x_i = \sum_j p(j|i,a)x_j\}$, $i \in \mathbb{X}$.

This system has the following properties
(i) With f_0 any Blackwell optimal policy, $x = u^{-1}(f_0)$ and $y = u^0(f_0)$ satisfy (4.3).
(ii) If (x, y) is a solution of (4.3), then x equals the value vector ϕ.

B. The unichain case.
In the unichain case, for every policy f, the stationary matrix $P^*(f)$ has identical components. Hence, the value vector ϕ is a constant vector. We will denote this constant vector by $\phi \cdot e$ (ϕ is a scalar). The first part of the optimality equation is always satisfied and the following result can be derived.

Theorem 1.19 *Consider the system $x+y_i = \max_a\{r(i,a)+\sum_j p(j|i,a)y_j\}, i \in \mathbb{X}$. This system has the following properties:*
(i) With f_0 any Blackwell optimal policy, $x \cdot e = u^{-1}(f_0)$ and $y = u^0(f_0)$, satisfy this system.
(ii) If (x, y) is a solution of the system, then $x = \phi$ and $y = u^0(f_0) + c \cdot e$ for some constant c.

The functional equation (4.3) is extensively investigated in Schweitzer and Federgruen [216]. Another proof for the solution of the optimality equation can also be provided by applying Brouwer's fixed point theorem (see Federgruen and Schweitzer [72], and Schweitzer [212]). In the unichain case the solution of the optimality equation can be exhibited as the fixed point of an N-step contraction (cf. Federgruen, Schweitzer and Tijms [74]).

1.4.3 Policy iteration

In the policy iteration method a sequence of policies $f_1, f_2, \ldots$ is constructed such that $\phi(f_{k+1}) \geq \phi(f_k)$ and $v^\alpha(f_{k+1}) > v^\alpha(f_k)$ for all $\alpha \in (\alpha_k, 1)$. Since F is finite and all policies f_k are different, this method has finite termination with an optimal policy.

A. The multichain case.

Theorem 1.20 *Consider the following system of linear equations*

$$\left\{\begin{array}{llll} [I-P(f)] & x & & = 0 \\ & x+[I-P(f)] & y & = r(f) \\ & & y+[I-P(f)]z & = 0. \end{array}\right. \tag{4.4}$$

Then, (4.4) has a solution $(x(f), y(f), z(f))$, where $x(f)$ and $y(f)$ are unique with $x(f) = u^{-1}(f)$ and $y(f) = u^0(f)$.

For every $i \in \mathbb{X}$ and every policy f, we define the action subset $B(i,f)$ by

$$B(i,f) := \left\{ a \in \mathbb{A}(i) \;\middle|\; \begin{array}{l} \sum_j p(j|i,a)\phi_j(f) > \phi_i(f) \text{ or} \\ \sum_j p(j|i,a)\phi_j(f) = \phi_i(f) \text{ and} \\ r(i,a) + \sum_j p(j|i,a)u_j^0(f) > \phi_i(f) + u_i^0(f) \end{array} \right\} \tag{4.5}$$

Theorem 1.21
(i) If $B(i,f) = \emptyset$ for every $i \in \mathbb{X}$, then *f is an average optimal policy.*
(ii) If $B(i,f) \neq \emptyset$ for at least one i and the policy $g \neq f$ satisfies for each state i: $g(i) \in B(i,f)$ if $g(i) \neq f(i)$, then $\phi(g) \geq \phi(f)$ and $v^\alpha(g) > v^\alpha(f)$ for α sufficiently close to 1.

Algorithm VI (policy iteration; average reward, multichain case)

1. Start with any $f \in F$.
2. Determine $\phi(f)$ and $u^0(f)$ as the unique (x,y)-part in a solution of the linear system (4.4).
3. For every $i \in \mathbb{X}$: determine $B(i,f)$ as defined in (4.5).
4. If $B(i,f) = \emptyset$ for every $i \in \mathbb{X}$: go to step 6.
   Otherwise: take any $g \neq f$ such that $g(i) \in B(i,f)$ if $g(i) \neq f(i)$.
5. $f := g$ and go to step 2.
6. f is an average optimal policy.

B. The unichain case.
In the unichain case, since the average reward vectors are constant, the set $B(i,f)$ can be simplified to

$$B(i,f) := \{a \in \mathbb{A}(i) \mid r(i,a) + \sum\nolimits_j p(j|i,a)u_j^0 > \phi(f) + u_i^0(f)\} \qquad (4.6)$$

The following result holds.

Theorem 1.22 *The linear system $x \cdot e + [I - P(f)]y = r(f)$ with $y_1 = 0$, has a unique solution $x = \phi(f)$ and $y = u^0(f) - u_1^0(f)$.*

Algorithm VII (policy iteration; average reward, unichain case)

1. Start with any $f \in F$.
2. Determine $\phi(f)$ and $u^0(f)$ as the unique solution of the linear system $x \cdot e + [I - P(f)]y = r(f)$ with $y_1 = 0$.
3. For every $i \in \mathbb{X}$: determine $B(i,f)$ as defined in (4.6).
4. If $B(i,f) = \emptyset$ for every $i \in \mathbb{X}$: go to step 6.
   Otherwise: take any $g \neq f$ such that $g(i) \in B(i,f)$ if $g(i) \neq f(i)$.
5. $f := g$ and go to step 2.
6. f is an average optimal policy.

The concept of policy iteration is originated by Howard [121] who considered the first two parts of system (4.4). However, in that case, the convergence is

not always guaranteed and cycling can occur. Blackwell [27] has given a convergent version by imposing the constraint $P^*(f)y = 0$; the formulation with system (4.4) was proposed by Miller and Veinott [166]. In Blackwell's version, in order to compute $P^*(f)$, the chain structure of the transition matrix $P(f)$ has to be analyzed. Other anti-cycling rules, which avoid the analysis of the chain structure, are introduced in Schweitzer and Federgruen [215], Federgruen and Spreen [75], and Spreen [232]. Various treatments of the policy iteration method in the unichain case (or other special cases) can be found in Schweitzer [208], Denardo [49], Haviv and Puterman [100], and Lasserre [151].

1.4.4 Linear programming

A vector $v \in \mathbb{R}^N$ is said to be *average-superharmonic* if there exists a vector $u \in \mathbb{R}^N$ such that the pair (u, v) satisfies

$$\left\{ \begin{array}{lll} v_i & \geq \sum_j p(j|i,a)v_j & \text{for every} (i,a) \in \mathbb{X} \times \mathbb{A} \\ v_i + u_i & \geq r(i,a) + \sum_j p(j|i,a)u_j & \text{for every} (i,a) \in \mathbb{X} \times \mathbb{A} \end{array} \right. \tag{4.7}$$

Theorem 1.23 *The value vector ϕ is the (componentwise) smallest average-superharmonic vector.*

Proof (outline). Let f_0 be a Blackwell optimal policy. From Theorem 1.18 it follows that $\phi_i \geq \sum_j p(j|i,a)\phi_j, (i,a) \in \mathbb{X} \times \mathbb{A}$, and $\phi_i + u_i^0(f_0) \geq r(i,a) + \sum_j p(j|i,a)u_j^0(f_0)$ for every $i \in \mathbb{X}$ and $a \in \mathbb{A}(i,\phi)$. Then, it can be shown that (ϕ, u) is average-superharmonic, where $u = u^0(f_0) + M \cdot \phi$ with M sufficiently large. Suppose that y is also average-superharmonic with corresponding x. Then, $y \geq P(f_0)y$, implying that $y \geq P^*(f_0)y \geq P^*(f_0)\{r(f_0)+[P(f_0)-I]x\} = P^*(f_0)r(f_0) = \phi(f_0) = \phi$, i.e. ϕ is the smallest average-superharmonic vector.■

A. The multichain case.

Corollary 1.7 *Let (u, v) be an optimal solution of the linear program*

$$\min \left\{ \sum\nolimits_j \beta_j v_j \;\middle|\; \begin{array}{ll} \sum_j [\delta_{ij} - p(j|i,a)]v_j \geq 0 & , (i,a) \in \mathbb{X} \times \mathbb{A} \\ v_i + \sum_j [\delta_{ij} - p(j|i,a)]u_j \geq r(i,a) & , (i,a) \in \mathbb{X} \times \mathbb{A} \end{array} \right\} \tag{4.8}$$

where $\beta_j > 0$, $j \in \mathbb{X}$,is arbitrarily chosen, then $u = \phi$.

The dual program of (4.8) is

$$\max \left\{ \sum\nolimits_{i,a} r(i,a)x_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} [\delta_{ij} - p(j|i,a)]x_{ia} = 0 & , j \in \mathbb{X} \\ \sum_a x_{ja} + \sum_{i,a} [\delta_{ij} - p(j|i,a)]y_{ia} = \beta_j & , j \in \mathbb{X} \\ x_{ia}, y_{ia} \geq 0, (i,a) \in \mathbb{X} \times \mathbb{A} & \end{array} \right\} \tag{4.9}$$

For any feasible solution (x, y) of (4.9) we denote by $\mathbb{X}_x$

$$\mathbb{X}_x := \{ j \in \mathbb{X} \mid \sum\nolimits_a x_{ja} > 0 \} \tag{4.10}$$

Theorem 1.24 *Let (x, y) be an extreme optimal solution of (4.9). Then, any policy f such that* $\begin{cases} x_{if(i)} > 0 & if \quad i \in \mathbb{X}_x \\ y_{if(i)} > 0 & if \quad i \notin \mathbb{X}_x \end{cases}$ *is an average optimal policy.*

Proof (outline). Because for every $j \in \mathbb{X}$, $\sum_a x_{ja} + \sum_a y_{ja} = \sum_{i,a} p(j|i,a) y_{ia} + \beta_j > 0$, policy f is well defined. Since $x_{if(i)} > 0$, $i \in \mathbb{X}_x$ and $y_{if(i)} > 0$, $i \notin \mathbb{X}_x$, it follows from the complementary slackness of linear programming that $\phi_i + \sum_j [\delta_{ij} - p(j|i,(f(i))] u_j = r(i, f(i))$, $i \in \mathbb{X}_x$ and $\sum_j [\delta_{ij} - p(j|i, f(i))] \phi_j = 0$, $i \notin \mathbb{X}_x$. Program (4.8) implies that $\sum_j [\delta_{ij} - p(j|i,a)] \phi_j \geq 0$, $i \in \mathbb{X}$, $a \in \mathbb{A}$. Suppose that $\sum_j [\delta_{kj} - p(j|k, f(k))] \phi_j > 0$ for some $k \in \mathbb{X}_x$. Since $x_{kf(k)} > 0$, $\sum_j [\delta_{kj} - p(j|k, f(k))] \phi_j \cdot x_{kf(k)} > 0$. Furthermore, $\sum_j [\delta_{ij} - p(j|i,a)] \phi_j \cdot x_{ia} \geq 0$, $(i,a) \in \mathbb{X} \times \mathbb{A}$. Hence, $\sum_{i,a} \sum_j [\delta_{ij} - p(j|i,a)] \phi_j \cdot x_{ia} > 0$.
On the other hand, this result is contradictory to the constraints of program (4.9) because $\sum_{i,a} \sum_j [\delta_{ij} - p(j|i,a)] \phi_j \cdot x_{ia} = \sum_j \{\sum_{i,a} [\delta_{ij} - p(j|i,a) x_{ia}]\} \phi_j = 0$. Therefore, we have shown that $\sum_j [\delta_{ij} - p(j|i, f(i))] \phi_j = 0$ for every $i \in \mathbb{X}$, i.e. $\phi = P(f)\phi$, and consequently $\phi = P^*(f)\phi$. Next, it can easily be shown that $\mathbb{X}_x$ is closed under $P(f)$. Then, we can prove that the states of $\mathbb{X} \backslash \mathbb{X}_x$ are transient in the Markov chain induced by $P(f)$. This implies that the columns of $\mathbb{X} \backslash \mathbb{X}_x$ in $P^*(f)$ are zero. Now, we can finish the proof as follows. For every $k \in \mathbb{X}$, we can write,
$\phi_k(f) = [P^*(f) r(f)]_k = \sum_i [P^*(f)]_{ki} r(i, f(i)) = \sum_{i \in \mathbb{X}_x} [P^*(f)]_{ki} r(i, f(i)) = \sum_{i \in \mathbb{X}_x} [P^*(f)]_{ki} \{\phi_i + \sum_j [\delta_{ij} - p(j|i, f(i))] u_j\} = [P^*(f)\{\phi + \{(I - P(f))u\}]_k = \phi_k$.
Hence, f is an average optimal policy. ∎

Algorithm VIII (linear programming; average reward, multichain case)

1. Take any β with $\beta_j > 0, j \in \mathbb{X}$, and compute extreme optimal solutions (u, v) and (x, y), e.g. by the simplex method, of the dual pair linear programs (4.8) and (4.9) respectively.

2. Choose f such that $x_{if(i)} > 0$ if $i \in \mathbb{X}_x$ and $y_{if(i)} > 0$ if $i \notin \mathbb{X}_x$. Then, f is an average optimal policy and v is the value vector ϕ.

In the average reward case there is no one-to-one correspondence between the feasible solutions of the dual program (4.9) and the stationary policies. However, there are interesting relations. For a feasible solution (x, y) of (4.9) we define a stationary policy $\pi(x, y)$ by

$$\pi_{ia}(x, y) := \begin{cases} x_{ia} / \sum_a x_{ia} & a \in \mathbb{A}(i), \ i \in \mathbb{X}_x \\ y_{ia} / \sum_a y_{ia} & a \in \mathbb{A}(i), \ i \notin \mathbb{X}_x \end{cases} \tag{4.11}$$

Conversely, consider a stationary policy π, and define $(x(\pi), y(\pi))$ by

$$\begin{cases} x_{ia}(\pi) & := \{\sum_k \beta_k [P^*(\pi)]_{ki}\} \cdot \pi_{ia} & a \in \mathbb{A}(i), \quad i \in \mathbb{X} \\ y_{ia}(\pi) & := \{\sum_k \beta_k [D(\pi)]_{ki} + \sum_k \gamma_k [P^*(\pi)]_{ki}\} \cdot \pi_{ia} & a \in \mathbb{A}(i), \quad i \in \mathbb{X} \end{cases} \tag{4.12}$$

with $\gamma_k := \max_{i \in \mathbb{X}_j} \{-\sum_k \beta_k [D(\pi)]_{ki} / \sum_k [P^*(\pi)]_{ki}\}$, $k \in \mathbb{X}_j$, where $\mathbb{X}_j$ is the j-th ergodic set of the transition matrix $P(\pi)$, and $\gamma_k := 0$ for k a transient state. Then, the following results can be derived (see Kallenberg [134]).

Theorem 1.25
(i) For any stationary policy π, $(x(\pi), y(\pi))$ is feasible for (4.9).
(ii) For any pure stationary policy f, $(x(f), y(f))$ is an extreme point of (4.9).
(iii) If π is an average optimal policy, then $(x(\pi), y(\pi))$ is an optimal solution of (4.9) and vice-versa.

Remarks
1. As mentioned before, for MDPs with constraints the linear programming approach is appropriate. For multichain constrained MDPs, there is no optimal stationary policy, in general. The variables x_{ia} of program (4.9) can, analogously to the discounted case, be interpreted as average state-action frequencies, but the analysis is much more complex. For the unichain case, this analysis can be found in Derman [58]; the multichain case is treated by Hordijk and Kallenberg [116]. An interpretation of the second type of variables, the variables y_{ia}, is not obvious. They are related to the deviation matrix (Kallenberg [134]) and can be interpreted as biased deviation measures (Altman and Spieksma [8]). Other contributions in this area, based on a sample path approach, are Ross [195] and Ross and Varadarajan [196]. Beutler and Ross [25] discuss the constrained MDP by a Lagrangean approach. In Altman and Shwartz [4] the sensitivity of constrained MDPs is investigated.
2. MDPs with multi-objectives can be treated as constrained MDPs. For this topic we refer to Hordijk and Kallenberg [116], and to Durinovics, Lee, Katehakis and Filar [65].
3. For a decision maker it can be unsatisfactory to consider only the expectation of the rewards. It may be preferable to consider also the variability. Papers on this subject are Sobel [228], [229], Kawai and Katoh [143], White [282], [283] and [286], Filar, Kallenberg and Lee [78], Chung [37], [38] and [39], Bayal-Gursoy and Ross [12], and Huang and Kallenberg [123].

Open problem
For MDPs with constraints, an interesting question is find the best policy in the class of stationary policies or in the class F of pure stationary policies. In the multichain case, no satisfactory algorithm is known for these problems. For the problem to find the best policy within the class of stationary policies, the natural candidate $\pi(x, y)$, with (x, y) the optimal solution of (4.9) with additional constraints, does not satisfy (see Kallenberg [134]).

B. The unichain case.
Since in the unichain case ϕ is a vector with identical components, the property average-superharmonic is equivalent to the existence of a scalar v and a vector u such that $v + u_i \geq r(i,a) + \sum_j p(j|i,a)u_j$ for every $(i,a) \in \mathbb{X} \times \mathbb{A}$. Hence, the LP-problem for the smallest average-superharmonic vector becomes

$$\min \left\{ v \mid v + \sum\nolimits_j [\delta_{ij} - p(j|i,a)]u_j \geq r(i,a) \text{ for every } (i,a) \in \mathbb{X} \times \mathbb{A} \right\} \quad (4.13)$$

with dual program

$$\max\left\{\sum_{i,a} r(i,a)x_{ia} \;\middle|\; \begin{array}{lll} \sum_{i,a}[\delta_{ij}-p(j|i,a)]\,x_{ia} & = 0, & j\in\mathbb{X} \\ \sum_{i,a} x_{ia} & = 1 & \\ x_{ia} & \geq 0, & (i,a)\in\mathbb{X}\times\mathbb{A} \end{array}\right\} \tag{4.14}$$

Algorithm IX (linear programming; average reward, unichain case)

1. Compute extreme optimal solutions (u,v) and x of the dual pair LPs (4.13) and (4.14) respectively.
2. Choose f such that $x_{if(i)} > 0$ if $i \in \mathbb{X}_x$ and $f(i)$ arbitrary if $i \notin X_x$. Then, f is an average optimal policy and $v \cdot e$ is the value vector ϕ.

Remarks

1. In the irreducible case any feasible solution of (4.13) satisfies $\sum_a x_{ia} > 0$, $i \in \mathbb{X}$. Furthermore, the mapping $x_{ia} \to \pi(x)$ with $\pi_{ia}(x) = x_{ia}/\sum_a x_{ia}$ is a one-to-one mapping of the feasible solutions of (4.13) onto the stationary policies with as inverse mapping $\pi \to x_{ia}(\pi)$, where $x_{ia}(\pi) = [p^*(\pi)]_i \cdot \pi_{ia}$ with $p^*(\pi)$ the equilibrium distribution. The set F of pure stationary policies corresponds to the set of extreme solutions of (4.13). In this case, similar to the discounted reward criterion, it can be shown that the linear programming method is equivalent to policy iteration. For the relation between the discounted linear program and the undiscounted linear program in the irreducible case, we refer also to Nazareth and Kulkarni [170].
2. In the unichain case, also a suboptimality test can be implemented, in the policy iteration method as well as in the linear programming method (cf. Hastings [96] and Lasserre [152]). Furthermore, in the unichain case, problems with constraints have a solution in the set of stationary policies: if (x,y) is the optimal solution of the LP-problem with constraints, then $\pi(x,y)$ with $\pi_{ia}(x,y) = x_{ia}/\sum_a x_{ia}$, $a \in \mathbb{A}(i)$, $i \in \mathbb{X}_x$ (and arbitrary decisions in $\mathbb{X}\backslash\mathbb{X}_x$) is a stationary optimal policy (see Derman [58]).

The pioneering work in solving MDPs by linear programming was made by Manne [164] and De Ghellinck [42], who considered the irreducible case. The first analysis in the general multichain case was described in Denardo and Fox [51] and in Denardo [47], who proposed a sequential procedure. Hordijk and Kallenberg [114] have shown that also in the multichain case one linear program suffices. Many results about the linear programming method can be found in the monograph Kallenberg [134].

1.4.5 Value iteration

It seems natural to investigate for value iteration the formula of the discounted rewards with discount factor $\alpha = 1$, i.e.

$$\begin{cases} v_i^{n+1} := \max_a\{r(i,a) + \sum_j p(j|i,a)v_j^n\}, & i\in\mathbb{X}, n\geq 0 \\ v_i^0 \text{ arbitrary}, & i\in\mathbb{X} \end{cases} \tag{4.15}$$

with corresponding policies $f_0, f_1, \ldots$ such that $v^{n+1} = r(f_n) + P(f_n)v^n$, $n \geq 0$.

This approach, however, causes difficulties: in general, there is no convergence of the sequence $\{v^n \mid n \geq 0\}$ nor of the sequence $\{v^n - v^{n-1} \mid n \geq 1\}$. Since v^n corresponds to the total reward during n periods, the sequence $\{v^n \mid n \geq 0\}$ is in general unbounded and grows linearly in n. Therefore, it is plausible to consider the sequence $\{v^n - n \cdot \phi \mid n \geq 0\}$. The next lemma, which appeared in Brown [29], shows that this sequence is bounded. The behavior of this sequence is also studied by Lanery [150].

Lemma 1.7 The sequence $\{v^n - n \cdot \phi \mid n \geq 0\}$ is bounded.

Corollary 1.8 $\phi = \lim_{n\to\infty} \frac{1}{n} v^n = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (v^k - v^{k-1})$.

Although Corollary 1.8 shows that ϕ can be approximated by the sequence $\{\frac{1}{n} v^n \mid n \geq 1\}$, this result does not provide sufficient information for the computation of an ϵ-optimal policy or an ϵ-approximation of ϕ. Therefore, we need stronger results, e.g. the convergence of the sequence $\{e^n\}_{n=0}^{\infty}$, where e^n is defined by $e^n := v^n - n \cdot \phi$. In general, however, this sequence may fail to converge if some of the transition matrices $P(f)$ are periodic. Fortunately, periodicity can be avoided by the following *data transformation*, proposed by Schweitzer [209]. Schweitzer and Federgruen [214] have given necessary and sufficient conditions which guarantee the convergence of the sequence $\{e^n\}_{n=0}^{\infty}$.

Consider for an arbitrary $\lambda \in (0, 1)$ the modified transition probabilities

$$p^{\lambda}(j|i, a) = \lambda \delta_{ij} + (1 - \lambda) p(j|i, a), i, j \in \mathbb{X} \text{ and } a \in \mathbb{A}(i) \tag{4.16}$$

Since $p^{\lambda}(i|i, f(i)) \geq \lambda > 0$, the transition matrix $P^{\lambda}(f)$ is aperiodic. Let $\phi^{\lambda}(f)$ be the average reward of policy f with respect to the transitions (4.16), then the next lemma shows that $\phi^{\lambda}(f) = \phi(f)$. Hence, we may assume that for every f the Markov chain with transition matrix $P(f)$ is aperiodic, in which case $P^*(f) = \lim_{n\to\infty} P^n(f)$.

Lemma 1.8 $\phi^{\lambda}(f) = \phi(f)$ for every $f \in F$.

To show that, under the aperiodicity assumption, the sequence $\{e^n\}_{n=0}^{\infty}$ is convergent, we need the following theorem.

Theorem 1.26 *Let* $b(i, a) = r(i, a) - \phi_i + \sum_j p(j|i, a) u_j - u_i, i \in \mathbb{X}, a \in \mathbb{A}$; $m_i = \liminf_{n\to\infty} e_i^n$, $i \in \mathbb{X}$; $M_i = \limsup_{n\to\infty} e_i^n$, $i \in \mathbb{X}$, *and* $\mathbb{A}_*(i) = \{a \in \mathbb{A}(i) | \phi_i = \sum_j p(j|i, a) \phi_j\}$, $i \in \mathbb{X}$. *Then,* $\max_{a\in\mathbb{A}_*(i)} \{b(i, a) + \sum_j p(j|i, a) m_j\} \leq m_i \leq M_i \leq \max_{a\in\mathbb{A}_*(i)} \{b(i, a) + \sum_j p(j|i, a) M_j\}$.

Theorem 1.27 *Assume that the aperiodicity assumption holds. Then, the sequence* $\{e^n \mid n \geq 0\}$ *is convergent.*

Lemma 1.9 *Assume that the sequence* $\{e^n \mid n \geq 0\}$ *converges. Then,*

(i) f_n is average optimal for n sufficiently large.
(ii) $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$.

A. The multichain case

Since, for large n, $\phi \approx v^{n+1} - v^n$, $\| v^{n+1} - v^n \|$ and span $(v^{n+1} - v^n)$ do not provide a valid stopping criterion. If ϕ is not a constant vector, no stopping criteria are available. Therefore, another approach is necessary. Schweitzer [210] employs a hierarchical decomposition of the MDP into a set of communicating MDPs. This decomposition was proposed by Bather [11]. Schweitzer and Federgruen [216] have shown that this decomposition is unique.

Open problem
Formulate a value iteration algorithm (without a hierarchical decomposition of the MDP and without chain analysis) for multichain undiscounted MDPs.

Fundamental research of value iteration for undiscounted multichain MDPs was made by Schweitzer and Federgruen. In Schweitzer and Federgruen [217] it is shown, without any assumptions about the periodicity or the chain structure, that if the sequence $\{v^n - n \cdot \phi\}_{n=0}^{\infty}$ is convergent, the convergence rate is geometric. This is surprising because the operator of the mapping (4.15) is, in general, not a contraction nor a J-step contraction with respect to any norm or the seminorm *span*. Conditions, other than aperiodicity, for the convergence of $\{v^n - n \cdot \phi\}_{n=0}^{\infty}$ are given by Schweitzer [207], Denardo [49] and Bather [10]. Surveys on value iteration for undiscounted multichain MDPs can be found in Schweitzer and Federgruen [214] and in Federgruen and Schweitzer [70] and [71].

B. The 'constant value vector' case

Assume that the value vector is constant, i.e. $\phi = \phi_0 \cdot e$, where $\phi_0 \in \mathbb{R}$. This assumption is more general than the unichain assumption. Furthermore, we assume aperiodicity, which implies (cf. Lemma 1.9) that $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$. We will formulate an algorithm to compute an ϵ-optimal policy.

Theorem 1.28 *Let $l_n = \min_i(v_i^n - v_i^{n-1})$ and $u_n = \max_i(v_i^n - v_i^{n-1})$, $n \in \mathbb{N}$. Then,*
(i) $l_n \uparrow \phi_0$ and $u_n \downarrow \phi_0$.
(ii) $l_n \cdot e \le \phi(f_{n-1}) \le \phi_0 \cdot e \le u_n \cdot e$, $n \ge 1$.

By the results of Theorem 1.28 an algorithm can be formulated. Since v^n grows linearly in n, a direct application of (4.14) may cause numerical difficulties. Therefore, we use the following transformation, which yields the so-called *relative value algorithm*.
Let $w_i^n := v_i^n - v_N^n$, $i \in \mathbb{X}$, $n \ge 0$; $g^n := v_N^n - v_N^{n-1}$, $n \ge 1$. Then, one can easily show that $\{w^n\}_{n=0}^{\infty}$ and $\{g^n\}_{n=0}^{\infty}$ are bounded sequences. Furthermore, the next iterands can be computed by the formulae $g^{n+1} = \max_{a\in\mathbb{A}(N)}\{r(N,a) + \sum_j p(j|N,a)w_j^n\}$, and $w_i^{n+1} = \max_{a\in\mathbb{A}(i)}\{r(i,a) + \sum_j p(j|i,a)w_j^n\} - g^{n+1}$, $i \in \mathbb{X}$. For the bounds l_n and u_n, we have $l_n = \min_i(w_i^n - w_i^{n-1}) + g^n$ and $u_n = \max_i(w_i^n - w_i^{n-1}) + g^n$.

Algorithm X (relative value iteration; average reward; aperiodic; constant value vector)

1. Choose $\epsilon > 0$, and take $v \in \mathbb{R}^N$ arbitrarily.
2. Compute:
   a. $s(i,a) := r(i,a) + \sum_j p(j|i,a)v_j$, $(i,a) \in \mathbb{X} \times \mathbb{A}$;
   b. $g := \max_{a \in \mathbb{A}(N)} s(N,a)$;
   c. $w_i := \max_{a \in \mathbb{A}(i)} s(i,a) - g$, $i \in \mathbb{X}$ and take f such that $w = r(f) + P(f)v - g$;
   d. $u := \max_i(w_i - v_i)$; $l := \min_i(w_i - v_i)$.
3. If $u - l \leq \epsilon$: f is an ϵ-optimal policy and $(u+l)/2$ is a $\frac{1}{2}\epsilon$-approximation of the value ϕ_0 (Stop);

   Otherwise: $v := w$ and go to step 2.

One may ask whether exclusion of suboptimal actions can be implemented for the average reward criterion. Similar to the discounted rewards, it can be shown that an action $a \in \mathbb{A}(i)$ is *suboptimal* if

$$\phi + u_i > r(i,a) + \sum_j p(j|i,a)u_j, \tag{4.17}$$

where (ϕ, u) is a solution of the optimality equation of Theorem 1.19. Since such a solution is unknown in advance, in order to apply (4.17) in an algorithm, we need bounds for ϕ and u. Theorem 1.28 provides bounds for ϕ; however, bounds for u are unknown. One may well apply a suboptimality test in one iteration of formula (4.15). In fact, v^n is the total reward over a horizon of n stages. Hence, suboptimality tests for finite horizon models can be used (see Hastings [93], Hastings and Van Nunen [99], and Hübner [124]).

Bounds on the value vector as formulated in Theorem 1.28 can be found in Hastings [95], Odoni [172], Hordijk and Tijms [118], and Platzman [177]. Hordijk and Tijms [120] have proposed an approximation method with a sequence of discounted value iterations with discount factors tending to 1. Algorithm X is established by White [277]. Recently, a new value iteration algorithm was proposed by Bertsekas [21], under the assumption that all policies are unichain and that there exists a state that is recurrent under all policies. This method is inspired by a relation with an associated stochastic shortest path problem.

1.4.6 Modified policy iteration

As in the discounted reward case, modified policy iteration can be applied. However, we need an assumption: we assume that the value vector ϕ is a constant vector: $\phi = \phi_0 \cdot e$. Furthermore, we assume the *strong aperiodicity assumption*, i.e. for some $0 < \lambda < 1$, $p(i|i,a) \geq \lambda > 0$ for all $i \in \mathbb{X}$, $a \in \mathbb{A}(i)$. As shown in the previous section by Schweitzer's aperiodicity transformation (4.16), any MDP can be transformed to an equivalent MDP which has the strong aperiodicity property.

Let T and T_f be the operators as defined in (3.2) and (3.3), respectively, and with $\alpha = 1$.

Lemma 1.10 *Let $l_n := \min_i (Tx^n - x^n)_i$, $n \in \mathbb{N}$. Then, the sequence $\{l_n\}_{n=1}^{\infty}$ is monotonically nondecreasing.*

Remark
Let $u_n := \max_i (Tx^n - x^n)_i$, $n \in \mathbb{N}$. Then the sequence $\{u_n \mid n \in \mathbb{N}\}$ is in general not monotonically nonincreasing.

Theorem 1.29
(i) Both sequences $\{l_n \mid n \in \mathbb{N}\}$ and $\{u_n \mid n \in \mathbb{N}\}$ converge to the value ϕ_0.
(ii) The convergence of span $(Ux^n - x^n)$ to zero is geometrically fast.
(iii) Algorithm XI (see below) terminates with an ϵ-optimal policy f and $\frac{1}{2}[u+l]$ is an $\frac{1}{2}\epsilon$-approximation of ϕ_0.

Algorithm XI (modified policy iteration; average reward; aperiodic; constant value vector)

1. Choose $x \in \mathbb{R}^N$ and $\epsilon > 0$ arbitrarily.
2. a. Choose k with $1 \leq k \leq \infty$;
   b. Determine f such that $T_f x = Tx$;
   c. Let $l := \min_i (Tx - x)_i$ and $u := \max_i (Tx - x)_i$.
3. If $u - l \leq \epsilon$: f is an ϵ-optimal policy and $(u+l)/2$ is a $\frac{1}{2}\epsilon$-approximation of the value ϕ_0 (Stop);
   otherwise: $x := T_f^k x$ and go to step 2.

Remark
If $k = 1$ the method becomes the standard value iteration method (without White's relative values). We will also argue that, in the unichain case, policy iteration corresponds to $k = \infty$. By Theorem 1.15, we have $\phi(f) + [I - P(f)]u(f) = r(f)$, $T_{f_n}^k x^n = T_{f_n}^k [u(f_n)] + P^k(f_n)[x^n - u(f_n)] = u(f_n) + k \cdot \phi(f_n) + P^k(f_n)[x^n - u(f_{n+1})]$. If k tends to infinity, $P^k(f_n)$ converges to $P^*(f_n)$, a matrix with equal rows, i.e. $P^k(f_n)[x^n - u(f_n)]$ converges to a constant vector. Since, $\phi(f_n)$ is also a constant vector, the difference between $T_{f_n}^k x^n$ and $u(f_n)$ converges to a constant vector. In the policy iteration algorithm VII with best improving actions, a new policy corresponds to maximization of $r(i,a) + \sum_j p(j|i,a)u_j(f_n)$, which is the same as $T[L_{f_n}^k x^n]$. Hence, both methods are very similar.

The modified policy iteration method was first mentioned by Morton [169]. Van der Wal [245] and [246] has analyzed this method extensively under various chain structure assumptions (irreducible case, unichain case, communicating case and simply connected case).

1.5 MORE SENSITIVE OPTIMALITY CRITERIA

1.5.1 Introduction

In section 1.1.3 the concepts of n-discount optimality, n-average optimality and Blackwell optimality were introduced. For all these criteria optimal policies, which are pure and stationary, exist.

Theorem 1.30 *For $n = -1, 0, 1, \ldots$ the criteria n-discount optimal and n-average optimal are equivalent.*

Remarks
1. The criterion (-1)-discount optimality, and hence also (-1)-average optimality, is equivalent to average optimality.
2. The criteria 0-discount optimality and 0-average optimality are also called bias optimality. For more details about bias optimality we refer to the Chapter 2 of this book.

In Blackwell [27] the concept of bias optimality was introduced. Veinott [256] presented a policy iteration algorithm for finding a bias optimal policy. In Veinott [256] is also shown that an *average overtaking* pure and stationary policy is bias optimal, and conjectured that the reverse statement is also true. A policy R_* is average overtaking optimal if $\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}[v^t(R_*) - v^t(R)] \geq 0$ for every policy R. In the class of stationary policies, the conjecture was proved by Denardo and Miller [52]. Lippman [154] showed the equivalence for general (possibly nonstationary) policies. Other contributions to the computation of a bias optimal policy are Denardo [47] and [49], Fox [80] and Kallenberg [133]. Denardo and Rothblum [54] have studied the stronger criterion of *overtaking optimality*. A policy R_* is overtaking optimal if $\liminf_{T\to\infty} \sum_{t=1}^{T}[v^t(R_*) - v^t(R)] \geq 0$ for every policy R. For this criterion the existence of an optimal policy is not guaranteed, in general, as already was shown in Brown [29]. Denardo and Rothblum [54] provided conditions under which a stationary overtaking optimal policy exists. The n-discount optimality criterion was proposed in Veinott [257]. Sladky [223] has introduced the concept of n-average optimality; furthermore, he showed the equivalence between this criterion and the n-discount optimality.

1.5.2 n-Discount optimality and policy iteration

In this section we present a policy iteration algorithm to compute a policy that lexicographically maximizes the vector $(u^{-1}(f), u^0(f), \ldots, u^n(f))$ over the set F. For $n = -1$ an average optimal policy and for $n = 0$ a bias optimal policy is obtained. Furthermore, for all $n \geq N - 1$ an n-discount optimal policy is Blackwell optimal.

Theorem 1.31 *The linear system*

$$\left\{ \begin{array}{ll} [I - P(f)]x^{-1} & = 0 \\ x^{-1} + [I - P(f)]x^0 & = r(f) \\ x^{k-1} + [I - P(f)]x^k & = 0 \ 1 \leq k \leq n+1; \ P^*(f)x^{n+1} = 0 \end{array} \right\}$$

has the unique solution $(u^{-1}(f), u^0(f), \ldots, u^{n+1}(f))$.

Algorithm XII (policy iteration; n-discount optimality)

1. Take an arbitrary policy f.
2. Determine $(u^{-1}(f), u^0(f), \ldots, u^{n+1}(f))$ as unique solution of the linear system

$$\left\{ \begin{array}{ll} [I - P(f)]x^{-1} & = 0 \\ x^{-1} + [I - P(f)]x^0 & = r(f) \\ x^{k-1} + [I - P(f)]x^k & = 0 \; 1 \leq k \leq n+1; \; P^*(f)x^{n+1} = 0 \end{array} \right\}$$

3.

   a. If $\max_{\mathbb{X} \times \mathbb{A}}[Pu^{-1}(f) - u^{-1}(f)] > 0$, then $\mathbb{A}^{(-1)} = \underset{\mathbb{X} \times \mathbb{A}}{\operatorname{argmax}} [Pu^{-1}(f) - u^{-1}(f)]$, choose g from $\mathbb{A}^{(-1)}$ and go to step 5.

   b. If $\max_{\mathbb{X} \times \mathbb{A}^{(-1)}} [r + Pu^0(f) - u^0(f) - u^{-1}(f)] > 0$, then $\mathbb{A}^{(0)} = \underset{\mathbb{X} \times \mathbb{A}^{(-1)}}{\operatorname{argmax}}[r + Pu^0(f) - u^0(f) - u^{-1}(f)]$, choose g from $\mathbb{A}^{(0)}$ and go to step 5.

   c. For $k = 0$ until n do:
   If $\max_{\mathbb{X} \times \mathbb{A}^{(k)}} [Pu^{k+1}(f) - u^{k+1}(f) - u^k(f)] > 0$, then $\mathbb{A}^{(k+1)} = \underset{\mathbb{X} \times \mathbb{A}^{(k)}}{\operatorname{argmax}}[Pu^{k+1}(f) - u^{k+1}(f) - u^k(f)]$, choose g from $\mathbb{A}^{(k+1)}$ and go to step 5.

4. f is n-discount optimal (Stop).
5. $f(i) := g(i)$, $i \in \mathbb{X}$, and go to step 2.

Remarks
1. Instead of $P^*(f)x^{n+1} = 0$, we can also consider $x^{n+1} + [I - P(f)]x^{n+2} = 0$, since multiplication with $P^*(f)$ gives $P^*(f)x^{n+1} = 0$.
2. For $n = -1$ the algorithm is equivalent to algorithm VI.

Theorem 1.32 *Let f and g be subsequent policies in algorithm XII, then $v^\alpha(g) > v^\alpha(f)$ for α sufficiently close to 1.*

Theorem 1.33 *Algorithm XII terminates in a finite number of iterations with an n-discount optimal policy.*

Finally, we mention that an n-discount optimal policy is a Blackwell optimal policy if $n \geq N - 1$.

Theorem 1.34 *If algorithm XII is used to determine an $n \geq (N-1)$-discount optimal policy f, then f is also a Blackwell optimal policy.*

The policy iteration method of this section was proposed in Veinott [257] and in Miller and Veinott [166]. They have also shown that Blackwell optimality is the same as n-discount optimality for $n \geq N-1$. In Veinott [258] refined results are given. In Federgruen and Schweitzer [73] a value iteration method is suggested for solving nested functional equations. These equations arise e.g. when more sensitive discount optimal policies are found. In particular, a method is given to find the optimal bias vector and a bias-optimal policy.

1.5.3 Blackwell optimality and linear programming

In this section we show how linear programming in the space of the rational functions can be developed to compute optimal policies over the entire range of the discount factor. Especially, a procedure is presented for the computation of a Blackwell optimal policy.

Let $\mathbb{R}$ be the ordered field of the real numbers with the usual ordering denoted by $>$. By $P(\mathbb{R})$ we denote the set of all polynomials with real coefficients:

$$\mathrm{P}(\mathbb{R}) = \{p(x) \mid p(x) = a_0 + a_1 x + \cdots + a_n x^n, a_i \in \mathbb{R}, 1 \leq i \leq n\}. \qquad (5.1)$$

By p_0 and p_1 we denote the polynomials $p_0(x) = 0$ and $p_1(x) = 1$ for every x. The field $F(\mathbb{R})$ of rational functions with real coefficients consists of the elements $\frac{p(x)}{q(x)}$, where p and q are from $P(\mathbb{R})$ and $q \neq p_0$. The polynomial $p(x)$ is considered as identical to the rational function $\frac{p(x)}{p_1(x)}$. Two rational functions $\frac{p}{q}$ and $\frac{r}{s}$ are considered identical, denoted $\frac{p}{q} =_\ell \frac{r}{s}$, if $p(x)s(x) = q(x)r(x)$. The operations $+$ and $\cdot$ are the natural addition and multiplication, i.e.
$\frac{p(x)}{q(x)} + \frac{r(x)}{s(x)} =_\ell \frac{p(x)s(x)+r(x)q(x)}{q(x)s(x)}$ and $\frac{p(x)}{q(x)} \cdot \frac{r(x)}{s(x)} =_\ell \frac{p(x)r(x)}{q(x)s(x)}$.
The polynomials p_0 and p_1 are the identities with respect to the operations addition and multiplication. A complete ordering in $F(\mathbb{R})$ is obtained by $\frac{p}{q} >_\ell p_0$ if and only if $d(p)d(q) > 0$, where $d(p)$ is the first nonzero coefficient of $p(x)$. If $\frac{p}{q} >_\ell p_0$, then the rational function $\frac{p}{q}$ is called positive. $\frac{p}{q} \geq_\ell p_0$ means that either $p =_\ell p_0$ or $\frac{p}{q} >_\ell p_0$. $F(\mathbb{R})$ is a non-Archimedean ordered field. The continuity of polynomials implies that the rational function $\frac{p}{q}$ is positive if and only if $\frac{p(x)}{q(x)} > 0$ for all x sufficiently close to 0. Hence, we obtain the following result.

Theorem 1.35 *The rational function $\frac{p}{q}$ is positive if and only if there exists an $x_0 > 0$ such that $\frac{p(x)}{q(x)} > 0$ for every $x \in (0, x_0]$.*

Instead of the discount factor α we can also use the interest rate ρ, where the relation between α and ρ is given by $\rho = (1-\alpha)/\alpha$. Notice that the total expected discounted reward $v^\rho(f)$ is the unique solution of the linear system $[(1+\rho)I - P(f)]x = (1+\rho)r(f)$. Solving this equation by Cramer's rule shows that $v_i^\rho(f)$, $i \in \mathbb{X}$, is an element of $F(\mathbb{R})$, say $\frac{p}{q}$, where the degree of the polynomials p and q is at most N. It is well known (Theorem 1.16) that the interval [0,1) of the discount factor can be divided into a finite number of intervals, say $[0 = \alpha_m, \alpha_{m-1}), \ldots, [\alpha_0, \alpha_{-1} = 1)$, in such a way that there exist

policies f_i, $0 \leq i \leq m$, where f_i is α-optimal for all $\alpha \in [\alpha_i, \alpha_{i-1})$. Hence, on any of these intervals the components of the value vector v^ρ are elements of $F(\mathbb{R})$.

Furthermore, the optimality equation (3.2) implies that $(1+\rho)v_i^\rho \geq (1+\rho)r(i,a) + \sum_j p(j|i,a)v_j^\rho$, $(i,a) \in \mathbb{X} \times \mathbb{A}$, $\rho > 0$. Therefore, in the ordered field $F(\mathbb{R})$, we have $(1+\rho)v_i^\rho \geq_\ell (1+\rho)r(i,a) + \sum_j p(j|i,a)v_j^\rho$, $(i,a) \in \mathbb{X} \times \mathbb{A}$. In general, v_i^ρ is not an element of $F(\mathbb{R})$, but there are elements of $F(\mathbb{R})$ which coincide piecewise with v_i^ρ.

An N-vector $w(\rho)$ with components in $F(\mathbb{R})$ is called superharmonic if $(1+\rho)w_i(\rho) \geq_\ell (1+\rho)r(i,a) + \sum_j p(j|i,a)w_j(\rho)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$. Hence, v^ρ is superharmonic. The concept of superharmonicity is useful to derive linear programs for MDPs.

Lemma 1.11 *v^ρ is the smallest superharmonic vector with components in $F(\mathbb{R})$, i.e. for any superharmonic vector $w(\rho)$, $w_i(\rho) \geq_\ell v_i^\rho$, $i \in \mathbb{X}$.*

Lemma 1.11 implies that the value vector v^ρ on the interval $(0, \rho_0]$ can be found as optimal solution of the following linear program in $F(\mathbb{R})$:

$$\min\{\sum\nolimits_j w_j(\rho) \mid \sum\nolimits_j [(1+\rho)\delta_{ij} - p(j|i,a)]w_j(\rho) \geq_\ell (1+\rho)r(i,a), (i,a) \in \mathbb{X}\times\mathbb{A}\}. \tag{5.2}$$

Consider also the following linear program in $F(\mathbb{R})$, called the *dual program*:

$$\max\left\{ \sum\nolimits_{i,a} (1+\rho)r(i,a) \cdot x_{ia}(\rho) \;\middle|\; \begin{array}{ll} \sum_{i,a}[(1+\rho)\ \delta_{ij} - p(j|i,a)]\cdot & \\ \qquad\qquad x_{ia}(\rho) =_\ell p_1, j \in \mathbb{X} & \\ x_{ia} \geq_\ell p_0, \quad (i,a) \in \mathbb{X} \times \mathbb{A} & \end{array} \right\} \tag{5.3}$$

For a fixed real value of ρ, the linear programs (5.2) and (5.3) are the linear programs (3.12) and (3.13) respectively. Also from section 1.3.3 it is known that there is a one-to-one correspondence between the extreme points of (5.3) and the set F. As in the simplex method, we will rewrite the equalities of (5.3) such that in each iteration there is precisely one positive $x(\rho)$ component in each state. The main difference with the usual simplex method for a fixed value of ρ is that, instead of real numbers, the elements are rational functions. During any iteration, the set of constraints is written in the special form

$$x_B = B^{-1}e - B^{-1}Ax_N \tag{5.4}$$

where e is the vector with the right-hand-side of (5.3) as components, i.e. p_1; x_B and x_N are the basis and nonbasis variables, B is the basic matrix and A consists of the remaining (nonbasis) columns of (5.3).

We solve the dual program (5.3) in such a way that the optimality of some basic solution, or equivalently some policy f, is shown on a certain interval for the value of ρ. This is possible, because for every fixed ρ in that interval the corresponding simplex tableau is an optimal one. At any iteration of the simplex tableau there is a feasible solution $x(\rho)$ of (5.3) and a corresponding

"trial solution" $w(\rho)$ of (5.2), i.e. $w(\rho)$ satisfies the complementary slackness conditions

$$x_{ia}(\rho) \cdot \Big\{ \sum\nolimits_j [(1+\rho)\delta_{ij} - p(j|i,a)] w_j(\rho) - (1+\rho) r(i,a) \Big\} = 0, (i,a) \in \mathbb{X} \times \mathbb{A} \tag{5.5}$$

for all ρ in the interval that is considered. Since any basic solution corresponds to a policy f, in each state i there is exactly one action, namely $f(i)$, such that $x_{if(i)}(\rho) > 0$ for all ρ in the actual interval. Hence, by (5.5),

$$[(1+\rho)I - P(f)]w(\rho) = (1+\rho)r(f), \text{ i.e. } w(\rho) = v^{\rho}(f). \tag{5.6}$$

The organization of the special simplex method with elements that are rational functions is based on the following theorem.

Theorem 1.36
(i) The elements of the simplex tableau can be written as rational functions with a common denominator, which is the product of all previous pivot elements.
(ii) The numerator and denominator of the rational functions are polynomials with degree N at most, except for the reduced costs where the numerator may have degree $N+1$.
(iii) For ρ sufficiently large, the optimal solution $x(\rho)$ is given by the basic variables $x_{if(i)}(\rho)$, where $f(i)$ is such that $r(i,(f(i)) = \max_a r(i,a)$, $i \in \mathbb{X}$.
(iv) The pivot operations in the simplex tableau are as follows ($n(\rho)$ is the common denominator): (a) the numerator of the pivot becomes the next common denominator, and the last common denominator becomes the new numerator of the pivot; (b) the numerators of the other elements in the pivot row are unchanged and the numerators of the other elements in the pivot column are multiplied by -1; (c) for the other elements, say numerator $p(\rho)$, we replace $p(\rho)$ by $\frac{p(\rho)t(\rho) - r(\rho)s(\rho)}{n(\rho)}$, which is a polynomial where $t(\rho)$ is the numerator of the last pivot and $r(\rho)$ is the numerator of the pivot row which is in the same column as $p(\rho)$, and $s(\rho)$ is the numerator in the pivot column which is in the same row as $p(\rho)$.

Starting with the artificial variables $z_j(\rho)$, $j \in \mathbb{X}$ as basic variables, we can compute the optimal simplex tableau for $\rho = \infty$ by exchanging $x_{1f(1)}$ with z_1, $x_{2f(2)}$ with $z_2, \ldots, x_{Nf(N)}$ with z_N, where $f(i)$ is such that $r(i,(f(i)) = \max_a r(i,a)$, $1 \leq i \leq N$. This tableau is optimal for $\rho \geq \rho_1$, where ρ_1 is the smallest value such that the reduced costs are nonnegative. To compute ρ_1 we have to determine the zeroes of some polynomials. The column that determines ρ_1 becomes the next pivot column. After a pivot transformation the next tableau is optimal for $[\rho_2, \rho_1)$ for some ρ_2. In this way we continue until the last interval $[\rho_m = 0, \rho_{m-1})$.

If we are only interested in computing a Blackwell optimal policy, and not in the computation of the intervals with corresponding optimal policies, the method can be described as follows:

1. Start with any policy f and compute the corresponding tableau.

2. If every reduced cost is nonnegative with respect to the ordering in $F(\mathbb{R})$, i.e. the dominating coefficient of the numerator of any reduced cost is nonnegative, then the corresponding policy is Blackwell optimal.

   Otherwise: take any column with a negative reduced cost as pivot column and execute a pivot transformation.

3. Go to step 2.

Remarks
1. Since in any transformation the value of the objective function increases, none of the basis can return and therefore the method is finite.
2. The complexity of one pivot transformation is of order $N^3[\sum_{i=1}^{N} \#\mathbb{A}(i)]$.

Hordijk, Dekker and Kallenberg [113] have developed the simplex method for rational functions for the computation of discounted optimal policies over the whole range of the discount factors, including the computation of a Blackwell optimal policy. Related works are Smallwood [224], Jeroslow [128] and Holzbaur [108], [109].

1.6 A VARIETY OF OTHER TOPICS

1.6.1 Mean - variability trade-off

The standard criteria for MDPs based on average or (sensitive) discounted rewards are not always satisfactory. In this section we consider another approach. This approach is especially suitable for a decision maker who prefers to use a criterion which also considers the *variability* induced by a given policy. How do we measure this variability? We want to have a variability measure which is sensible, mathematically tractable, and for which an optimality concept can be used. It turns out that optimality for all starting states simultaneously is too strong a requirement. Therefore, we consider the criterion for a fixed initial distribution β. Then, as mean of the rewards for a given policy R, we use

$$\phi(\beta, R) := \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\beta,R}[r(X_t, Y_t)] \tag{6.1}$$

If a policy R satisfies $\phi(\beta, R) = \sum_i \beta_i \phi_i$, where ϕ is the value vector, then R is called a β-average-optimal policy. There are several ways to define the variability $v(\beta, R)$, where β is the initial distribution and R the policy. We use the definition

$$v(\beta, R) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\beta,R}[r(X_t, Y_t) - \phi(\beta, R)]^2 \tag{6.2}$$

The quantities $\phi(\beta, R)$ and $v(\beta, R)$ can be expressed in the so-called *state-action frequencies*. For any policy R, any $T \in \mathbb{N}$, and any initial distribution β, we denote the *expected state-action frequencies* in the first T periods by the $x^T(R)$, i.e.

$$x_{ja}^T(R) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\beta,R}[X_t = j,\ Y_t = a], (i, a) \in \mathbb{X} \times \mathbb{A} \tag{6.3}$$

By $X(R)$ we denote the limit points of the vectors $\{x^T(R),\ T = 1, 2, \dots\}$, and by L, $L(M)$, $L(S)$ and $L(D)$ the elements of $X(R)$ corresponding to general, Markov, stationary and deterministic policies, respectively. For policies R with $\#X(R) = 1$, e.g. stationary policies, we denote the unique element of $X(R)$ by $x(R)$. For such policies the variability satisfies

$$v(\beta, R) = \sum\nolimits_{j,a} x_{ja}(R)[r(j,a)]^2 - \left[\sum\nolimits_{j,a} x_{ja}(R) r(j,a)\right]^2 \tag{6.4}$$

Let X be the projection on the x-space of the feasible solutions (x, y) of the linear program (4.9) , i.e.

$$X = \left\{ x \;\middle|\; \begin{array}{lll} \sum_{i,a}[\delta_{ij} - p(j|i,a)]x_{ia} & = 0, & j \in \mathbb{X} \\ \sum_a x_{ja} + \sum_{i,a}[\delta_{ij} - p(j|i,a)]y_{ia} & = \beta_j, & j \in \mathbb{X} \\ x_{ia}, y_{ia} \geq 0, (i,a) \in \mathbb{X} \times \mathbb{A} & & \end{array} \right\} \tag{6.5}$$

Theorem 1.37 $\overline{L(D)} = \overline{L(S)} = L(M) = L = X$, *where* $\overline{S}$ *is the closed convex hull of a set* S.

There are several sensible formulations for the mean-variability problem. We consider the following three formulations.

(1) Maximal mean-variability ratio with lower bound on the mean

$$\max\left\{ \frac{[\phi(\beta, R]^2}{v(\beta, R)} \;\middle|\; \phi(\beta, R) \geq l \right\} \tag{6.6}$$

(2) Minimal variability with lower bound on the mean

$$\min\{v(\beta, R) \mid \phi(\beta, R) \geq l\} \tag{6.7}$$

(3) Variability-penalized formulation

$$\max\{\phi(\beta, R) - \lambda \cdot v(\beta, R)\} \text{ for some penalty } \lambda > 0 \tag{6.8}$$

Using the state-action frequencies, the problems (6.6), (6.7) and (6.8) can be formulated as mathematical programs. These programs are special cases of the following unifying program

$$\max\left\{ \frac{\sum_{j,a} B_{ja} x_{ja}}{D(\sum_{j,a} R_{ja} x_{ja})} + C(\sum\nolimits_{j,a} R_{ja} x_{ja}) \;\middle|\; \begin{array}{l} x \in X \\ l \leq \sum_{j,a} R_{ja} x_{ja} \leq u \end{array} \right\} \tag{6.9}$$

with (a) C is a convex function; (b) if D is not a constant, then: (i) D is positive, convex and nondecreasing; (ii) C is nondecreasing; (iii) $\sum_{j,a} B_{ja} x_{ja} \leq 0$ for every $x \in X$.

In order to solve (6.9), we consider a parametric version of (4.9) with $B_{ia} + \vartheta R_{ia}$ instead of $r(i,a)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$, i.e.

$$\max\left\{\sum_{i,a} x_{ia}[B_{ia}+\vartheta R_{ia}] \;\middle|\; \begin{array}{ll} \sum_{i,a}[\delta_{ij}-p(j|i,a)]x_{ia} & =0, \quad j\in\mathbb{X} \\ \sum_{a} x_{ja}+\sum_{i,a}[\delta_{ij}-p(j|i,a)]y_{ia} & =\beta_j, \; j\in\mathbb{X} \\ x_{ia}, y_{ia}\geq 0, (i,a)\in\mathbb{X}\times\mathbb{A} & \end{array}\right\} \tag{6.10}$$

with $\vartheta \in (-\infty, +\infty)$ as the parameter. The optimal solution $x(\vartheta)$ is a piecewise constant function of ϑ with values being extreme points of X, and the optimal value is a piecewise linear, convex function of ϑ. Thus, there exist $\vartheta_0 \equiv -\infty < \vartheta_1 < \cdots < \vartheta_{m-1} < \vartheta_m \equiv +\infty$ such that $x(\vartheta) = x^n$ for $\vartheta \in [\vartheta_{n-1}, \vartheta_n]$, $1 \leq n \leq m$, with x^n an extreme point of X.

Let $k+1$ and $j+1$ be respectively the smallest integers among $0, 1, \ldots, m$ such that $\sum_{i,a} R_{ia} x_{ia}^{k+1} > u$ and $\sum_{i,a} R_{ia} x_{ia}^{j+1} \geq l$. Furthermore, let $\alpha \in (0,1]$ and $\beta \in [0,1)$ be such that $x^u = \alpha x^k + (1-\alpha)x^{k+1}$ and $x^l = \beta x^j + (1-\beta)x^{j+1}$ satisfy $\sum_{i,a} R_{ia} x_{ia}^u = u$ and $\sum_{i,a} R_{ia} x_{ia}^l = l$.
Let $G(x) = \sum_{j,a} B_{ja} x_{ja}$, $g(x) = \sum_{j,a} R_{ja} x_{ja}$ and $V(x) = \frac{G(x)}{D(g(x))} + C(g(x))$ for $x \in X$, and let $G_n = G(x^n)$, $g_n = g(x^n)$ and $V^n = V(x^n)$, $1 \leq n \leq m$. Furthermore, define $V_{\text{opt}} = \max\{\max_{j+1\leq n\leq k} V^n, V(x^l), V(x^u)\}$.

Theorem 1.38
(i) Program (6.9) is feasible if and only if $g(x^m) \geq l$ and $g(x^1) \leq u$.
(ii) If program (6.9) is feasible, then V_{opt} is the optimal value of (6.9), and the maximizing x is the optimal solution x_{opt}.

If $l = -\infty$ and $u = +\infty$, then $V_{\text{opt}} = V(x^n)$ for some extreme point x^n of X. Theorem 1.38 provides a way to find an optimal solution for program (6.9), but it does not provide a procedure to construct an optimal policy. The next two theorems show how an optimal policy can be obtained.

Theorem 1.39 *If (x, y) is an extreme optimal solution for (6.10) for all ϑ in an open interval, then any $f \in F$ with $x_{if(i)} > 0$ if $i \in \mathbb{X}_x$ and $y_{if(i)} > 0$ if $i \notin \mathbb{X}_x$ has a state-action frequency vector $x(f)$ satisfying $\sum_{i,a} B_{ia} x_{ia}(f) = \sum_{i,a} B_{ia} x_{ia}$, $\sum_{i,a} R_{ia} x_{ia}(f) = \sum_{i,a} R_{ia} x_{ia}$ and $V(x(f)) = V(x)$.*

Theorem 1.40 *If program (6.9) is feasible, then either $x_{\text{opt}} = x^n$ for some $j+1 \leq n \leq k$ and there exists an optimal deterministic policy, or $x_{\text{opt}} = x^l$ (or x^u) and an initial randomization of two deterministic policies is optimal. These policies can be determined analogously to the policy in Theorem 1.39.*

Corollary 1.9 *For an unconstrained problem, i.e. $l = -\infty$ and $u = +\infty$, there exists an optimal policy $f \in F$.*

Remarks
1. The discounted and the average-unichain case can be treated in the same way as above. In fact, these cases are more simple.
2. The optimal policy R_* is also *Pareto-optimal* with respect to the pair $(\phi(R), -V(R))$.

State-action frequencies for the unichain case are discussed in Derman [58]. The multichain case is analyzed in Kallenberg [134] and Hordijk en Kallenberg [116], who have shown Theorem 1.37. State-action frequencies play also an important role in multiple-objective MDP and for MDPs with additional constraints. Contributions in this area are made by Derman and Veinott [62], Thomas [240], Ross [195], Ross and Varadarajan [196], Altman and Shwartz [7] and by Liu and Ohno [156]. The formulations (6.6), (6.7) and (6.8) were proposed by Sobel [228], by Kawai [142] and by Filar, Kallenberg and Lee [78], respectively. Other contributions in this area are Kawai and Katoh [143], White [282], [283] and [286], Chung [37], [38] and [39], Bayal-Gursoy and Ross [12], and Sobel [229]. The unifying framework is proposed by Huang and Kallenberg [123], who also have shown the Theorems 1.38, 1.39 and 1.40. Another model with a different criterion is an MDP in which a (weighted) sum of a number of discounting rewards, each with a different discount factor, has to be maximized. This model is studied in Feinberg and Shwartz [77]: see Chapter 6.

1.6.2 Optimal stopping

Optimal stopping problems were introduced in section 1.1.4. In an optimal stopping problem there are two actions in each state. The first action is the stopping action and the second action corresponds to continue. If the stopping action is chosen in state i, then a final reward r_i is earned and the process terminates. If the second action is chosen in state i, then a cost c_i is incurred and the probability of being in state j at the next time point is p_{ij}. Therefore the MDP model is:

$\mathbb{X} = \{1, 2, \ldots, N\}$; $\mathbb{A}(i) = \{1, 2\}, i \in \mathbb{X}; r(i,1) = r_i, i \in \mathbb{X}$;
$r(i,2) = -c_i, i \in \mathbb{X}; p(j|i,1) = 0, i, j \in \mathbb{X}; p(j|i,2) = p_{ij}, i, j \in \mathbb{X}$.

We are interested in finding an optimal stopping rule, i.e. we consider only transient policies. A policy R is called *transient* if $\sum_{t=1}^{\infty} \mathbb{P}_{i,R}[X_t \in \mathbb{X}] < \infty$ for all $i \in \mathbb{X}$, i.e. for any starting state i the process terminates in a finite time with probability 1. As optimality criterion the total expected reward is considered i.e.

$$v_i(R) := \sum\nolimits_{t=1}^{\infty} \sum\nolimits_{j,a} \mathbb{P}_{i,R}[X_t = j, Y = a] \cdot r(j,a) \tag{6.11}$$

For the computation of an optimal transient policy, the usual properties of discounted MDPs hold (cf. Kallenberg [134] chapter 3). Let v be the value vector, i.e. $v = \sup\{v(R) \mid R \text{ is transient}\}$. Then, similar to the discounted reward criterion, it can be shown that v is the smallest superharmonic vector, i.e. the smallest vector that satisfies

$$\begin{cases} v_i \geq r_i & , i \in \mathbb{X} \\ v_i \geq -c_i + \sum_j p_{ij} v_j & , i \in \mathbb{X}. \end{cases} \tag{6.12}$$

Hence, the value vector is the unique solution of the linear program

$$\min\left\{\sum\nolimits_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i & ,i \in \mathbb{X} \\ v_i \geq -c_i + \sum_j p_{ij} v_j & ,i \in \mathbb{X} \end{array}\right\} \tag{6.13}$$

As in the discounted case, an optimal policy can be obtained by the dual program. Therefore, the following algorithm can be used.

Algorithm XIII (optimal stopping; linear programming)

1. Determine an optimal solution (x, y) of the dual program

$$\max\left\{\sum\nolimits_i r_i x_i - \sum\nolimits_i c_i y_i \;\middle|\; \begin{array}{lll} x_j + y_j - \sum_i p_{ij} y_i & = 1, & j \in \mathbb{X} \\ x_j, y_j & \geq 0, & j \in \mathbb{X} \end{array}\right\}$$

2. Choose f such that $f(i) = \begin{cases} 1 \text{ if } x_j > 0 \\ 2 \text{ if } x_j = 0. \end{cases}$

Let $S := \{i \in \mathbb{X} \mid r_i \geq -c_i + \sum_j p_{ij} r_j\}$, i.e. S is the set of states in which immediate stopping is not worse than continuing for one period and than choose the stopping action. An optimal stopping problem is *monotone* if $p_{ij} = 0$ for all $i \in S$, $j \notin S$, i.e. S is closed under P.

Theorem 1.41 *In a monotone optimal stopping problem the policy f, where $f(i) = 1$ if and only if $i \in S$, is optimal.*

For monotone stopping problems it is sufficient to determine the set S. The determination of S has complexity of order $\mathcal{O}(N)$.

A classical paper on optimal stopping problems is Breiman [28]. Other papers in this area are Chen [34], Ross [197], Yasuda [291] and Sonin [231].

1.6.3 Multi-armed bandit problems

We have introduced this model in section 1.1.4. At each decision time point the decision maker has the option to work on exactly one project. Any project may be in a finite number of states, say project j in the set X_j, $1 \leq j \leq n$. Hence, the state space is the Cartesian product: $\mathbb{X} = X_1 \times X_2 \times \cdots \times X_n$. Each state has the same action set $\mathbb{A} = \{1, 2, \ldots, n\}$, where action a means that project a is chosen, $1 \leq a \leq n$. When project a is chosen, i.e. project a is the active project, the immediate reward and the transition probabilities only depend on project a and the state $i \in X_a$. Let $r(i, a)$ and $p(j|i, a)$, $j \in X_a$, denote these quantities. The states of the inactive projects are frozen. As utility function the discounted reward is used.

The one-armed bandit stopping problem
Consider the one-armed bandit stopping problem, i.e. in each state there are two actions: action 1 is the stopping action where we earn a final reward M and by action 2 the process continues with immediate reward r_i and transition

probabilities p_{ij}. Let $v^\alpha(M)$ be the value vector of this optimal stopping problem. In the previous section it was discussed how this vector $v^\alpha(M)$ and an optimal policy can be computed by the linear programming programs

$$\min \left\{ \sum\nolimits_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i + \alpha \sum_j p_{ij} v_j, & i \in \mathbb{X} \\ v_i \geq M, & i \in \mathbb{X} \end{array} \right\} \tag{6.14}$$

and its dual

$$\max \left\{ \sum\nolimits_i r_i x_i + M \cdot \sum\nolimits_i y_i \;\middle|\; \begin{array}{lll} \sum_i (\delta_{ij} - \alpha p_{ij})\ x_i + y_j & = 1, & j \in \mathbb{X} \\ x_i, y_i & \geq 0, & i \in \mathbb{X} \end{array} \right\} \tag{6.15}$$

Lemma 1.12 *For all $i \in \mathbb{X}$, $v_i^\alpha(M) - M$ is a nonnegative continuous nonincreasing function in M.*

Let $M_i^\alpha = \min\{M \mid v_i^\alpha(M) = M\}$, $i \in \mathbb{X}$, called the *Gittins indices.*

Theorem 1.42 *The policy f, which chooses the stopping action in state i if and only if $M_i^\alpha \leq M$, is optimal.*

For $M = M_i^\alpha$ both actions (stop or continue) are optimal in state i. Hence, an interpretation of the Gittins index M_i^α is the value of M where both actions are simultaneously optimal, and therefore M_i^α is also called the *indifference value.*

Multi-armed bandits
Next, we assume that there are in each state $n+1$ actions: action a, $1 \leq a \leq n$, means continue with project a, and action 0 stops the process with a terminal reward M. Let $v^\alpha(M)$ be the value vector, f_M the optimal policy and $T(M)$ the stopping time, i.e. the expected time before the process terminates with the final reward M. Let $C = (1-\alpha)^{-1} \cdot \max_{i,a} |r(i,a)|$, then C is an upper bound of the total discounted rewards (without the terminal rewards). Hence, if $M \geq C$, then immediate stopping is optimal in all states. The following result is in some sense obvious: $v^\alpha(M)$ is nondecreasing in M and a small change in M will change the value (per unit change) with the discounted (unit) terminal reward $\alpha^{T(M)}$.

Lemma 1.13
(i) $v_i^\alpha(M)$ is a nondecreasing, convex function in M, for all $i \in \mathbb{X}$.
(ii) $\frac{\partial}{\partial M} v_i^\alpha(M) = \mathbb{E}_{i,f_M}[\alpha^{T(M)}]$, for all $i \in \mathbb{X}$.

The next theorem is the key theorem for the multi-armed bandit problem. It says that an optimal action in a state $i = (i_1, i_2, \ldots, i_n)$ is to choose that project which has, for the given state of the project, the smallest Gittins index. This is an interesting result. It is surprising that these indices depend only on the individual project and not on the other projects. Hence, they can be computed independently for each project. By this property, the dimensionality of the problem is considerably reduced.

Theorem 1.43 *In state $i = (i_1, i_2, \ldots, i_n)$ the optimal policy chooses action a, where a is such that $M^\alpha_{i_a} = \max_j M^\alpha_{i_j}$.*

Alternative interpretation of the Gittins index
Consider the one-armed bandit process with initial state i. If $M = M^\alpha_i$ the optimal policy is indifferent between stopping and continuing, so that for any stopping time T, $M^\alpha_i \geq \mathbb{E}[\text{discounted reward before } T] + M^\alpha_i \cdot \mathbb{E}[\alpha^T]$, with equality for the optimal policy. Hence,

$(1-\alpha)M^\alpha_i = \max_{T\geq 1} \mathbb{E}[\text{discounted reward before } T]/[\{1-\mathbb{E}(\alpha^T)\}/(1-\alpha)]$
$= \max_{T\geq 1} \mathbb{E}[\text{discounted reward before } T]/\mathbb{E}[\text{discounted time before } T]$,

where the expectations are conditional on the initial state i. Thus, another way to describe the optimal policy in the multi-armed bandit problem is as follows. For each individual project look for the stopping time T whose ratio of expected discounted reward and expected discounted time prior to T is maximal. Then work on the project with the largest ratio. In the case there also is the extra option of stopping, one should stop if all ratios are smaller than $(1-\alpha)$M.

Computation of the Gittins indices by parametric linear programming
We have already seen that for one project the Gittins index is related to the linear programs (6.14) and (6.15). For M big enough, an optimal solution (x, y) of (6.15) will satisfy $y_i > 0$, $i \in \mathbb{X}$. Decreasing M will give that some y_i becomes 0 for a certain value of M. For this M there is indifference between stopping and continuing, i.e. this M is the Gittins index in state i. By further decreasing M one can compute the next Gittins index, and so on. Hence, by parametric linear programming with parameter M which goes from $+\infty$ to $-\infty$, all Gittins indices can be computed for one project. The complexity of this approach is $\mathcal{O}(N^3)$.

Interpretation as restart-in-k problem
There is also another interpretation for the Gittins index M^α_k in a fixed state k. For any terminal value M, we have

$$v^\alpha_i(M) = \max\{M, r_i + \alpha \sum\nolimits_j p_{ij} v^\alpha_j(M)\}, i \in \mathbb{X} \tag{6.16}$$

and in state k, for $M = M^\alpha_k$,

$$v^\alpha_k(M) = M^\alpha_k = r_k + \alpha \sum\nolimits_j p_{kj} v^\alpha_j(M^\alpha_k) \tag{6.17}$$

Substituting (6.17) in (6.18) gives

$$v^\alpha_i(M^\alpha_k) = \max\left\{ r_k + \alpha \sum\nolimits_j p_{kj} v^\alpha_j(M^\alpha_k),\ r_i + \alpha \sum\nolimits_j p_{ij} v^\alpha_j(M^\alpha_k) \right\}, i \in \mathbb{X} \tag{6.18}$$

Hence, M^α_k is the k-th component of the value vector of the MDP where there are in each state two actions. By the first action the process is restarted in state k, and the second action continues the process. Since M^α_k can be found as the k-th component of the value vector of the restart-in-k problem, it can be computed by the following linear program

$$\max\left\{ \sum\nolimits_j v_j \;\middle|\; \begin{array}{l} \sum_j (\delta_{ij} - \alpha p_{ij}) v_j \geq r_i, \quad i \neq k \\ \sum_j (\delta_{ij} - \alpha p_{kj}) v_j \geq r_k, \ i \in \mathbb{X} \end{array} \right\} \tag{6.19}$$

For this restart-in-k problem, one can also characterize the states where it is optimal to choose action 'continue'.

Theorem 1.44 *Let $C_k = \{i \mid$ for the restart-in-k problem it is optimal to continue in state $i\}$. Then, $C_k = \{i \in \mathbb{X} \mid M_i^\alpha \geq M_k^\alpha\}$.*

Largest remaining index
In the largest remaining index approach the indices can be computed in a sequence, as in parametric linear programming, starting with the largest index.

Theorem 1.45 *Suppose that, for some k, $M_1^\alpha \geq M_2^\alpha \geq \cdots \geq M_k^\alpha$, and $M_k^\alpha \geq M_i^\alpha$ for all $i > k$. Let l_k be such that $M_{l_k}^\alpha = \max_{i>k} M_i^\alpha$ (the largest remaining index). Then, we have,*
$(1-\alpha)M_{l_k}^\alpha = \max_{i>k} \frac{[(I-\alpha P^k)^{-1} r]_i}{[(I-\alpha P^k)^{-1} e]_i}$, where $[P^k]_{ij} = \begin{cases} p_{ij}, & j \leq k \\ 0, & j > k. \end{cases}$

In order to find $M_{l_k}^\alpha$, we have to invert $[I - \alpha P^k]$. Since successive P^k matrices are similar, this can be done efficiently in a recursive way. The computations can be done in $\mathcal{O}(k^2)$. Hence, the overall complexity is $\sum_{k=1}^{N} \mathcal{O}(k^2) = \mathcal{O}(N^3)$.

The basic results on the multi-armed bandit problem are originated by Gittins (Gittins and Jones [86] and Gittins [85]). Other proofs of the optimality of the index rule can be found in Whittle [288] and [289], Ross [200], Tsitsiklis [242] and [243], Katehakis and Veinott [141], Weber [267] and Ishikida and Varaiya [127]. In honor of Gittins, Whittle has introduced the term Gittins indices. A first linear programming method of $\mathcal{O}(N^4)$ is proposed by Chen and Kathehakis [35]. Kallenberg [135] has improved this method to $\mathcal{O}(N^3)$. The interpretation as restart-in-k problem is made by Kathehakis and Veinott [141]. The method of the largest remaining index rule is due to Varaiya, Walrand and Buyokkoc [254]. A method based on bisection was proposed in Ben-Israel and Flåm [14]. Extension are made in various directions. Branching bandits were studied, e.g. by Weiss [270]; generalized bandits, e.g. in Glazebrook and Owen [89], and in Glazebrook and Greatrix [88]. Bertsimas and Niño-Mora [24] have proposed a new approach by generalizing the theory of extended polymatroids. Other papers based on this new approach are Glazebrook and Garbe [87], and Garbe and Glazebrook [84].

1.6.4 Separable Markov decision problems

Separable MDPs have the property that for certain pairs (i, a) of a state i and an action a: (i) the immediate reward is the sum of terms due to the current state and action, i.e. $r(i, a) = s(i) + t(a)$, (ii) the transition probabilities depend only on the action and not on the state from which the transition occurs, i.e. $p(j|i, a) = p(j|a)$. For separable problems an LP formulation can be given, which involves a smaller number of variables than in the general LP formulation. In this section we consider the multichain undiscounted case. For the discounted case and the unichain undiscounted case we refer to De Ghellinck and Eppen [43] and to Denardo [46], respectively.

A *separable Markov decision problem* has the following structure:
(1) In some states, say the states of $\mathbb{X}_1 = \{1, 2, \ldots, m)$, there are subsets of the action sets, say subset $\mathbb{A}_1(i)$ in state $i \in \mathbb{X}_1$, such that:
(i) $r((i,a) = s(i) + t(a)$, $i \in \mathbb{X}_1$, $a \in \mathbb{A}_1(i)$;
(ii) $p(j|i,a)$ is independent of $i : p(j|i,a) = p(j|a)$, $i \in \mathbb{X}_1$, $a \in \mathbb{A}_1(i)$, $j \in \mathbb{X}$.
(2) The action subsets are nested: $\mathbb{A}_1(1) \supseteq \mathbb{A}_1(2) \supseteq \cdots \supseteq \mathbb{A}_1(m) \neq \emptyset$.
Let $\mathbb{X}_2 := \mathbb{X} \backslash \mathbb{X}_1$, $\mathbb{A}_2(i) := \mathbb{A}(i) \backslash \mathbb{A}_1(i)$, $1 \leq i \leq m$, $\mathbb{A}_2(i) := \mathbb{A}(i)$, $m+1 \leq i \leq N$, and $B(i) := \mathbb{A}_1(i) - \mathbb{A}_1(i+1)$, $1 \leq i \leq m-1$, $B(m) := \mathbb{A}_1(m)$. Then $\mathbb{A}_1(i) = \cup_{j=i}^{m} B(j)$, and the sets $B(j)$ are disjoint. $\mathbb{X}_1$, $\mathbb{X}_2$, $\mathbb{A}_2(i)$ or $B(i)$ may be empty.

If the system is observed in a state $i \in \mathbb{X}_1$, and the decision maker will choose an action from $\mathbb{A}_1(i)$, the decision process can be considered as follows. First a reward $s(i)$ is earned and the system makes a zero-time transition to an additional state $N+i$. In this state there are two options: either to take an action $a \in B(i)$ or to take an action from $\mathbb{A}_1(i+1)$. In the first case reward $t(a)$ is earned and the process moves to state j with probability $p(j|a)$, $j \in \mathbb{X}$; in the second case we have the same situation as in state $N+i+1$, i.e. a zero-time transition is made from state $N+i$ to state $N+i+1$. This formulation can be interpreted as a semi-Markov decision process (see section 1.6.5). It can be shown that a linear program, which directly provides an average optimal policy, can be formulated. This linear program is based on linear programming for semi-Markov decision problems (cf. Kallenberg [134], chapter 7).

Consider the linear program

$$\text{minimize} \sum_{j=1}^{N} g_j + \sum_{j=1}^{m} h_j \tag{6.20}$$

$$\sum_{j=1}^{N} [\delta_{ij} - p(j|i,a)] g_j \geq 0, 1 \leq i \leq N, a \in \mathbb{A}_2(i); g_i - h_i \geq 0, 1 \leq i \leq m$$

$$-\sum_{j=1}^{N} p(j|a) g_j + h_i \geq 0, 1 \leq i \leq m, a \in B(i); h_i - h_{i+1} \geq 0, 1 \leq i \leq m-1$$

$$g_i + \sum_{j=1}^{N} [\delta_{ij} - p(j|i,a)] u_j \geq r(i,a), 1 \leq i \leq N, a \in \mathbb{A}_2(i); u_i - v_i \geq s_i, 1 \leq i \leq m$$

$$h_i - \sum_{j=1}^{N} p(j|a) u_j + v_i \geq t(a), 1 \leq i \leq m, a \in B(i); v_i - v_{i+1} \geq 0, 1 \leq i \leq m-1$$

The corresponding dual program is (the dual variables corresponding to the constraints of (6.20) are y_{ia}, μ_i, z_{ia}, σ_i, x_{ia}, λ_i, w_{ia} and ρ_i, respectively:)

$$\text{maximize} \sum_{i=1}^{N} \sum_{a \in \mathbb{A}(i)} r(i,a) x_{ia} + \sum_{i=1}^{m} s_i \lambda_i + \sum_{i=1}^{m} \sum_{a \in B(i)} t(a) w_{ia} \tag{6.21}$$

$$\sum_{i=1}^{N}\sum_{a\in\mathbb{A}(i)}[\delta_{ij}-p(j|i,a)]y_{ia}+\sum_{i=1}^{m}\delta_{ij}\mu_i\sum_{i=1}^{m}\sum_{a\in B(i)}p(j|a)z_{ia}$$
$$+\sum_{a\in\mathbb{A}(j)}x_{ja}=1,\quad 1\le j\le N$$
$$\sigma_j-\sigma_{j-1}+\sum_{a\in B(j)}w_{ja}-\mu_j+\sum_{a\in B(j)}z_{ja}=1,\quad 1\le j\le m$$
$$\sum_{i=1}^{N}\sum_{a\in\mathbb{A}_2(i)}[\delta_{ij}-p(j|i,a)]x_{ia}+\sum_{i=1}^{m}\delta_{ij}\lambda_i$$
$$-\sum_{i=1}^{m}\sum_{a\in B(i)}p(j|a)w_{ia}=0,\quad 1\le j\le N$$
$$\rho_j-\rho_{j-1}+\sum_{a\in B(j)}w_{ja}-\lambda_j=0, 1\le j\le m;\quad \rho_0=\rho_m=\sigma_0=\sigma_m=0;$$
$$x_{ia},y_{ia},z_{ia},w_{ia},\lambda_i,\mu_i,\rho_i,\sigma_i\ge 0 \text{ for all } i \text{ and } a.$$

A proof for the next result can be found in Kallenberg [136].

Theorem 1.46
(i) The linear programs (6.20) and (6.21) have finite optimal solutions.
(ii) If (g,h,u,v) is an optimal solution of program (6.20), then g is the value vector.
(iii) Let $(y,\mu,z,\sigma,x,\lambda,w,\rho)$ be an extreme optimal solution of program (6.21). Define m_i and n_i by $m_i=\min\{j\ge i \mid \sum_a w_{ja}>0\}$ and $n_i=\min\{j\ge i \mid \sum_a(w_{ja}+z_{ja})>0\}$, $i\in\mathbb{X}$. Take a policy f such that in state $i: x_{if(i)}>0$ if $\sum_a x_{ia}>0$; $w_{m_if(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i>0$; $y_{if(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i=0\wedge\sum_a y_{ia}>0$; $w_{n_if(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i=0\wedge\sum_a y_{ia}=0\wedge\sum_a w_{n_ia}>0$; $z_{n_if(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i=0\wedge\sum_a y_{ia}=0\wedge\sum_a w_{n_ia}=0$. Then, f is well defined and an average optimal policy.

There are many applications which can be formulated as separable MDPs. We mention some of them.

Replacement problem (cf. Howard's [121] automobile problem).
The decision maker has two options in each state i: either to continue or to replace the item by another of a certain state $j\in\{1,2,\dots,N\}$. The linear program to solve this problem as 'normal' Markov decision problem contains $2N(N+1)$ variables and $2N$ constraints. The reduced linear programming formulation has only $6N$ variables and $2N+1$ constraints.

Inventory problem
Consider the following inventory model. At the end of each period, the amount i of inventory is observed, where $0\le i\le N$. The possible actions are: either to order nothing or to order $a-i$ items, where $i+1\le a\le N$, with fixed ordering costs K and cost c for each ordered item. We assume that the delivery is instantaneous and that there is no backlogging. The linear program to solve this problem as 'normal' Markov decision problem has $(N+1)(N+2)$ variables and $2(N+1)$ constraints. In the reduced formulation as separable problem, we have $8N-2$ variables and $2(2N+1)$ constraints. In the case that the optimal policy is an (s,S)-policy the underlying Markov chain is unichained. Then a linear program with $3(N-1)$ variables and $N+2$ constraints suffices.

Totally separable problem
Suppose that the Markov decision problem has the following structure:
$\mathbb{X} = \{1, 2, \ldots, N\}$; $\mathbb{A}(i) = \{1, 2, \ldots, M\}$, $i \in \mathbb{X}$; $r(i,a) = s(i) + t(a)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$; $p(j|i,a) = p(j|a)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$ and $j \in \mathbb{X}$.
Examples of this model can be found in Sobel [227]. Without exploiting the structure, the linear program has $2NM$ variables and $2N$ constraints. It can be shown that an optimal myopic solution exists, i.e. the action a_* is optimal in state i, where a_* is determined by:

$$t(a_*) + \sum_{j=1}^{N} p(j|a) s_j = \max_{1 \le a \le M} \left\{ t(a) + \sum_{j=1}^{N} p(j|a) s_j \right\} \tag{6.22}$$

This result is a special case of the stochastic game studied in Sobel [227] and Parthasarathy, Tijs and Vrieze [176].

1.6.5 Further subjects

In this chapter some of the main topics of finite MDPs are discussed. In this section we shortly mention some other aspects of MDPs without going into detail.

Semi-Markov decision models
In many applications the times between consecutive decision time points are not identical but random. Such processes are called semi-Markov decision processes if the time until the next decision depends only on the present state i and the action a chosen in state i. We assume that the distribution function $F_{ij}^{a}(t)$ for the random variable $\tau_{ij}(a)$, which is the sojourn time until the next decision point if decision a is chosen when the system is in state i and the transition is into state j, is known for all $i, j \in E$ and $a \in A(i)$.

Semi-Markov decision models are also called *Markov renewal programs*. The essential results of MDPs can be generalized to semi-MDPs. The semi-MDP model was introduced by Jewell [129], [130], Howard [122], De Cani [41] and Schweitzer [206]. Contributions for discounted rewards are e.g. Denardo [45], De Ghellinck and Eppen [43], Kallenberg [134], Wessels and Van Nunen [271], Ohno [174] and Schweitzer [213].

In the average reward case, there is a very elegant data transformation, proposed by Schweitzer [209], which converts a semi-MDP into an equivalent MDP. Let $\tau_i(a)$ be the expected time until the next decision epoch if action a is chosen when the system is in state i. For $0 < \tau \le \min_{i,a} \tau(i,a)$, let

$$\begin{cases} \overline{r}(i,a) = r(i,a)/\tau(i,a) & ,i \in \mathbb{X}, a \in \mathbb{A}(i) \\ \overline{p}(j|i,a) = \delta_{ij} - [\delta_{ij} - p(j|i,a)] \cdot \tau/\tau(i,a) & ,i,j \in \mathbb{X}, a \in \mathbb{A}(i) \end{cases} \tag{6.23}$$

Then, $\phi(\pi) = \overline{\phi}(\pi)$, where $\phi(\pi)$ is the average reward per unit time of the semi-MDP and $\overline{\phi}(\pi)$ the average reward of the discrete-time MDP with rewards $\overline{r}(i,a)$ and transition probabilities $\overline{p}(j|i,a)$ as defined in (6.23).

Other papers on average reward MDPs are Schweitzer and Federgruen [216], Federgruen and Spreen [75], Denardo and Fox [51], Osaki and Mine [175],

Kallenberg [134], Schweitzer and Federgruen [217], Schweitzer [210] and Denardo [48].

MDPs with partial information, partial observation and adaptive control
In an MDP with *partial information* the exact state of the process cannot be observed at decision epochs. The only information available about the state is a subset of the state space to which the state belongs. Formally, an MDP has partial information if the state space is partitioned into subsets $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_m$ such that at each decision epoch the only available information is the subset $\mathbb{X}_k$ to which the state belongs. In the partial information case not all decision rules are feasible: in all states of a subset $\mathbb{X}_k$ $(1 \leq k \leq m)$ the same decision has to be chosen. Such decision rule is called an *admissible* decision rule. The objective is to find an optimal admissible policy for some optimality criterion with respect to a given initial distribution. Papers on MDPs with partial information are e.g. Smallwood and Sondik [225], Hastings and Sadjani [98], Hordijk and Loeve [117], and Loeve [158].

A related model is an MDP with *partial observation*. In this model there is probabilistic information about the state. Using Bayes' rules this model can be translated in a model with full information but with a continuous state space, which incorporates the complete history of the process. Papers in this area are Sondik [230], Albright [1], Monahan [168], Altman and Shwartz [5] and [6], Lovejoy [159], [160] and [161], Rieder [192], Sernik and Markus [219], White III [273] and [274], White III and Scherer [275] and [276], and White [287].

In *adaptive control* models the transition probabilities $p(j|i,a)$ and the rewards $r(i,a)$ depend on an unknown parameter ϑ from a parameter space Θ. About these parameters increasing information is obtained when observing the ongoing process. At each decision epoch the decision maker must estimate the true parameter and then adapt the policy to the estimated value. Further literature about this topic is e.g. Kurano [147], Hübner [125], Hernandez-Lerma [102], Cavazos-Cadena [32] and Burnetas and Katehakis [31].

Vector-valued MDPs
In vector-valued MDPs, when the system is in state i and action a is chosen, there is not a single reward $r(i,a)$, but a vector $r^k(i,a)$, $1 \leq k \leq m$, of rewards. For this model, the concept of optimality is not unambiguous. Given an initial distribution β, a policy R and a utility function u (e.g. discounted or average expected reward), there is an m-vector $u(\beta,R)$ of returns, where the k-th component $u_k(\beta,R)$ corresponds to the rewards $r^k(i,a)$. Optimality is defined with respect to a cone $C \subseteq \mathbb{R}^m$. Such cone defines a partial ordering in $\mathbb{R}^m : x \geq_C y$ iff $x - y \in C$. A policy R^* is optimal if $u(\beta,R^*) \geq_C u(\beta,R)$ for all policies R. In general, there does not exist an optimal policy. Therefore, we use the concept of an *efficient policy*. A policy R^* is efficient if there is no 'better' policy, i.e. there is no policy R with $u(\beta,R) > u(\beta,R^*)$. If the cone $C = \mathbb{R}^m_+$, then efficient policies are also called Pareto-optimal policies. For vector-valued MDPs also the term *multi-objective MDPs* is used.

Papers about vector-valued MDPs are e.g. Furakawa [82], White [279], Henig [101], Kallenberg [134], Durinovic, Lee, Katehakis and Filar [65], Ghosh [90], Liu, Ohno and Nakayama [157], and Wakuta [261], [262], [263] and [264].

Acknowledgment

I am grateful to Arie Hordijk for introducing me in the interesting subject of MDPs as well as for the cooperation during a long period.

References

[1] S.C. Albright, [1979]: "Structural results for partially observable Markov decision processes", Operations Research *27*, 1041–1053.

[2] E. Altman, [1999]: "Constrained Markov decision processes", Chapman & Hall/CRC, Boca Raton, Florida.

[3] E. Altman, A. Hordijk and L.C.M. Kallenberg [1996]: "On the value function in constrained control of Markov chains", Mathematical Methods of Operations Research *44*, 387–399.

[4] E. Altman and A. Shwartz [1991a]: "Sensitivity of constrained Markov decision processes", Annals of Operations Research *33*, 1–22.

[5] E. Altman and A. Shwartz [1991b]: "Adaptive control of constrained Markov chains", IEEE-Transactions on Automatic Control *36*, 454–462.

[6] E. Altman and A. Shwartz [1991c]: "Adaptive control of constrained Markov decision chains: criteria and policies", Annals of Operations Research *28*, 101–134.

[7] E. Altman and A. Shwartz [1991]: "Sensitivity of constrained Markov decision processes", Annals of Operations Research *33*, 1–22.

[8] E. Altman and F.M. Spieksma [1995]: "The linear program approach in Markov decision processes", Mathematical Methods of Operations Research *42*, 169–188.

[9] J.S. Baras, D.J. Ma and A.M. Makowsky [1985]: "K competing queues with linear costs and geometric service requirements: the μc-rule is always optimal" Systems Control Letters *6*, 173–180.

[10] J. Bather [1973a]: "Optimal decision procedures for finite Markov chains. Part II: Communicating systems", Advances in Applied Probability *5*, 521–540.

[11] J. Bather [1973b]: "Optimal decision procedures for finite Markov chains. Part III: General convex systems", Advances in Applied Probability *5*, 541–553.

[12] M. Bayal-Gursoy and K.W. Ross [1992]: "Variability-sensitivity Markov decision processes", Mathematics of Operations Research *17*, 558–571.

[13] R. Bellman [1957]: "Dynamic programming", Princeton University Press, Princeton.

[14] A. Ben-Israel and S.D. Flam [1990]: "A bisection/successive approximation method for computing Gittins indices", Zeitschrift für Operations Research *34*, 411–422.

[15] D.P. Bertsekas [1976]: "Dynamic programming and stochastic control", Academic Press, New York.

[16] D.P. Bertsekas [1976b]: "On error bounds for successive approximation methods", IEEE Transactions on Automatic Control *21*, 394–396.

[17] D.P. Bertsekas [1987]: "Dynamic programming: deterministic and stochastic models", Prentice-Hall, Englewood Cliff.

[18] D.P. Bertsekas [1995]: "Dynamic programming and optimal control I", Athena Scientific, Belmont, Massachusetts.

[19] D.P. Bertsekas [1995]: "Dynamic programming and optimal control II", Athena Scientific, Belmont, Massachusetts+.

[20] D.P. Bertsekas [1995c]: "Generic rank-one corrections for value iteration in Markovian decision problems", OR Letters *17*, 111–119.

[21] D.P. Bertsekas [1998]: "A new value iteration method for the average cost dynamic programming problem", SIAM Journal on Control and Optimization *36*, 742–759.

[22] D.P. Bertsekas and S.E. Shreve [1978] "Stochastic Optimal Control", Academic Press, New York.

[23] D.P. Bertsekas and J.N. Tsitsiklis [1991]: "An analysis of stochastic shortest path problems", Mathematics of Operations Research *16*, 580–595.

[24] D. Bertsimas and J. Niño-Mora [1996]: "Conservations laws, extended polymatroids and multi-armed bandit problems; a polyhedral approach to indexable systems", Mathematics of Operations Research *21*, 257–306.

[25] F.J. Beutler and K.W. Ross [1985]: "Optimal policies for controlled Markov chains with a constraint", Journal of Mathematical Analysis and Applications *112*, 236–252.

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

[27] D. Blackwell [1962]: "Discrete dynamic programming", Annals of Mathematical Statistics, 719–726.

[28] L. Breiman [1964]: "Stopping-rule problems", in: E.F. Beckenbach (ed.), Applied Combinatorial Mathematics", Wiley, New York, 284–319.

[29] B.W. Brown [1965]: "On the iterative method of dynamic programming on a finite space discrete time Markov process", Annals of Mathematical Statistics *36*, 1279–1285.

[30] J. Bruno, P. Downey and G.N. Frederickson [1981]: "Sequencing tasks with exponential service times to minimize the expected flowtime or makespan", Journal of the Association for Computing Machinery *28*, 100–113.

[31] A.N. Burnetas, and M.N. Katehakis [1997]: "Optimal adaptive policies for Markov decision processes", Mathematics of Op. Research *22*, 222–255.

[32] R. Cavazos-Cadena [1991]: "Nonparametric estimation and adaptive control in a class of finite Markov decision chains", Annals of Operations Research *28*, 169–184.

[33] C.-S. Chang, A. Hordijk, R. Righter and G. Weiss [1994]: "The stochastic optimality of SEPT in parallel machine scheduling", Probability in the Engineering and Information Sciences *8*, 179–188.

[34] M.C. Chen, Jr. [1973]: "Optimal stopping in a discrete search problem", Operations Research *21*, 741–747.

[35] Y.-R. Chen and M.N. Katehakis [1986] : "Linear programming for finite state bandit problems", Mathematics of Operations Research *11*, 180–183.

[36] Y.S. Chow and H. Robbins [1961]: "A martingale system theorem and applications" in: J. Neyman (ed), "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability", Vol.1, University of Berkeley Press, Berkeley, 93–104.

[37] K.-J. Chung [1989]: "A note on maximal mean/standard deviation ratio in an undiscounted MDP", OR Letters *8*, 201–204.

[38] K.-J. Chung [1992]: "Remarks on maximal mean/standard deviation ratio in an undiscounted MDPs", Opimization *26*, 385–392.

[39] K.-J. Chung [1994]: "Mean-variance trade-offs in an undiscounted MDP: the unichain case", Operations Research *42*, 184–188.

[40] G.B. Dantzig [1963]: "Linear programming and extensions", Princeton University Press, Princeton, New Jersey.

[41] J.S. De Cani [1964]: "A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity", Management Science *10*, 716–733.

[42] G.T. De Ghellinck [1960]: "Les problèmes de décisions séquentielles", Cahiers du Centre de Recherche Opérationelle, 161–179.

[43] G.T. De Ghellinck and G.D. Eppen [1967]: "Linear programming solutions for separable Markovian decision problems", Management Science *13*, 371–394.

[44] R.S. Dembo and M. Haviv [1984]: "Truncated policy iteration methods", OR Letters *3*, 243–246.

[45] E.V. Denardo [1967]: "Contraction mappings in the theory underlying dynamic programming", SIAM Review *9*, 165–167.

[46] E.V. Denardo [1968]: "Separable Markovian decision problems", Management Science *14*, 451–462.

[47] E.V. Denardo [1970]: "Computing a bias-optimal policy in a discrete-time Markov decision problem", Operations Research *18*, 279–289.

[48] E.V. Denardo [1971]: "Markov renewal programs with small interest rates", Annals of Mathematical Statistics *42*, 477–496.

[49] E.V. Denardo [1973]: "A Markov decision problem", in: T.C. Hu and S.M. Robinson (eds.), "Mathematical Programming", Academic Press, 33–68.

[50] E.V. Denardo [1982]: "Dynamic programming: models and applications", Prentice-Hall, Englewood Cliff.

[51] E.V. Denardo and B.L. Fox [1968]: "Multichain Markov renewal programs", SIAM Journal on Applied Mathematics *16*, 468–487.

[52] E.V. Denardo and B.L. Miller [1968]: "An optimality condition for discrete dynamic programming with no discounting", Annals of Mathematical Statistics *39*, 1220–1227.

[53] E.V. Denardo and U.G. Rothblum [1979a]: "Optimal stopping, exponential utility and linear programming", Mathematical Programming *16*, 228–244.

[54] E.V. Denardo and U.G. Rothblum [1979b]: "Overtaking optimality for Markov decision chains", Mathematics of Operations Research *4*, 144–152.

[55] F. D'Epenoux [1960]: "Sur un problème de production et de stockage dans l'aléatoire", Revue Française de Recherche Opérationelle, 3–16.

[56] C. Derman [1962]: "On sequential decisions and Markov chains", Management Science *9*, 16–24.

[57] C. Derman [1963]: "Optimal replacement rules when changes of states are Markovian", in: R. Bellman (ed.), "Mathematical optimization techniques", The Rand Corporation, R-396-PR, 201–212.

[58] C. Derman [1970]: "Finite state Markovian decision processes", Academic Press, New York.

[59] C. Derman and M. Klein [1965]: "Some remarks on finite horizon Markovian decision models", Operations Research *13*, 272–278.

[60] C. Derman and J. Sacks [1960]: "Replacement of periodically inspected equipment (an optimal stopping rule)", Naval Research Logistics Quarterly *7*, 597–607.

[61] C. Derman and R. Strauch [1966]: "A note on memoryless rules for controlling sequential control problems", Annals of Mathematical Statistics *37*, 276–278.

[62] C. Derman and A.F. Veinott, Jr. [1972]: "Constrained Markov decision chains", Management Science *19*, 389–390.

[63] H.M. Dietz and V. Nollau [1983]: "Markov decision problems with countable state space", Akademie-Verlag, Berlin.

[64] L. Dubins and L.J. Savage [1965]: "How to gamble if you must", McGraw-Hill, New York.

[65] S. Durinovics, H.M. Lee, M.N. Katehakis and J.A. Filar [1986]: "Multiobjective Markov decision processes with average reward criterion", Large Scale Systems *10*, 215–226.

[66] E.B. Dynkin [1979]: "Controlled Markov process", Springer-Verlag, New York.

[67] J.H. Eaton and L.A. Zadeh [1962]: "Optimal pursuit strategies in discrete state probabilistic systems", Transactions ASME Series D, Journal of Basic Engineering *84*, 23–29.

[68] A. Ephremides, P. Varaiya and J. Walrand [1980]: "A simple dynamic routing problem", IEEE Transactions on Automatic Control AC-25, 690–693.

[69] A. Federgruen [1984]: "Markovian control problems: functional equations and algorithms", Mathematical Centre Tract 97, Mathematical Centre, Amsterdam.

[70] A. Federgruen and P.J. Schweitzer [1978]: "Discounted and undiscounted value iteration in Markov decision problems: a survey", in: M.L. Puterman (ed), "Dynamic programming and its applications", Academic Press, New York, 23–52.

[71] A. Federgruen and P.J. Schweitzer [1980]: "A survey of asymptotic value-iteration for undiscounted Markovian decision processes", in: R. Hartley, L.C. Thomas and D.J. White (eds.), "Recent development in Markov decision processes", Academic Press, New York, 73–109.

[72] A. Federgruen and P.J. Schweitzer [1984a]: "A fixed-point approach to undiscounted Markov renewal programs", SIAM Journal on Algebraic Discrete Methods *5*, 539–550.

[73] A. Federgruen and P.J. Schweitzer [1984b]: "Successive approximation methods for solving nested functional equations in Markov decision problems", Mathematics of Operations Research *9*, 319–344.

[74] A. Federgruen, P.J. Schweitzer and H.C. Tijms [1978]: "Contraction mappings underlying undiscounted Markov decision problems", Journal of Mathematical Analysis and Applications *65*, 711–730.

[75] A. Federgruen and D. Spreen [1980]: "A new specification of the multichain policy iteration algorithm in undiscounted Markov renewal programs", Management Science *26*, 1211–1217.

[76] A. Federgruen and P. Zipkin [1984]: "An efficient algorithm for computing optimal (s, S) policies", Operations Research *34*, 1268–1285.

[77] E.A. Feinberg and A. Shwartz [1994]: "Markov decision models with weighted discounted criteria", Mathematics of Operations Research *19*, 152–168.

[78] J.A. Filar, L.C.M. Kallenberg and H.M. Lee [1989]: "Variance-penalized Markov decision processes", Mathematics of Operations Research *14*, 147–161.

[79] J.A. Filar and O. J. Vrieze [1997]: "Competitive Markov decision processes", Springer-Verlag, New York.

[80] B.L. Fox [1968]: "(g, w)-optima in Markov renewal programs", Management Science *15*, 210–212.

[81] E. Frostig [1993]: "Optimal policies for machine repairmen problems", Journal of Applied Probability *30*, 703–715.

[82] N. Furakawa [1980]: "Characterization of optimal policies in vector-valued Markovian decision processes", Mathematics of Operations Research *5*, 271–279.

[83] S. Gal [1984]: "An $\mathcal{O}(N^3)$ algorithm for optimal replacement problems", SIAM Journal on Control and Optimization *22*, 902–910.

[84] R. Garbe and K.D. Glazebrook [1998]: "On a new approach to the analysis of complex multi-armed bandit problems", Mathematical Methods of Operations Research *48*, 419–442.

[85] J.C. Gittins [1979]: "Bandit processes and dynamic allocation indices", Journal of the Royal Statistic Society Series B *14*, 148–177.

[86] J.C. Gittins and D.M. Jones [1974]: "A dynamic allocation index for the sequential design of experiments", in J.Gani (ed.) "Progress in Statistics", North Holland, Amsterdam, 241–266.

[87] K.D. Glazebrook and R. Garbe [1996]: "Reflections on a new approach to Gittins indexation", Journal of the Operational Research Society *47*, 1301–1309.

[88] K.D. Glazebrook and S. Greatrix [1995]: "On transforming an index for generalized bandit problems", J. of App. Prob. *32*, 168–182.

[89] K.D. Glazebrook and R.W. Owen [1991]: "New results for generalized bandit problems", International Journal of System Sciences *22*, 479–494.

[90] M.K. Ghosh [1990]: "Markov decision processes with multiple costs", OR Letters *9*, 257–260.

[91] R. Grinold [1973]: "Elimination of suboptimal actions in Markov decision problems", Operations Research *21*, 848–851.

[92] R. Hartley, A.C. Lavercombe and L.C. Thomas [1986]: "Computational comparison of policy iteration algorithms for discounted Markov decision processes", Computers and Operations Research *13*, 411–420.

[93] N.A.J. Hastings [1968]: "Some notes on dynamic programming and replacement", Operational Research Quarterly *19*, 453–464.

[94] N.A.J. Hastings [1969]: "Optimization of discounted Markov decision problems", Operations Research Quarterly *20*, 499–500.

[95] N.A.J. Hastings [1971]: "Bounds on the gain of a Markov decision process", Operations Research *19*, 240–243.

[96] N.A.J. Hastings [1976]: "A test for nonoptimal actions in undiscounted finite Markov decision chains", Management Science *23*, 87–92.

[97] N.A.J. Hastings and J.M.C.Mello [1973]: "Tests for nonoptimal actions in discounted Markov decision problems", Management Science *19*, 1019–1022.

[98] N.A.J. Hastings and D.Sadjani [1979]: "Markov programming with policy constraints", European Journal of Operations Research *3*, 253–255.

[99] N.A.J. Hastings and J.A.E.E. Van Nunen [1977]: "The action elimination algorithm for Markov decision processes", in H.C. Tijms and J. Wessels (eds), "Markov decision theory", Mathematical Centre Tract 100, 161–170, Mathematical Centre, Amsterdam.

[100] M. Haviv and M.L. Puterman [1991]: "An improved algorithm for solving communicating average reward Markov decision processes", Annals of Operations Research *28*, 229–242.

[101] M.I. Henig [1983]: "Vector-valued dynamic programming", SIAM Journal on Control and Optimization *21*, 490–499.

[102] O. Hernández-Lerma [1987]: "Adaptive Markov control processes", Springer-Verlag, New York.

[103] O. Hernández-Lerma and J. B. Lasserre [1996]: "Discrete-time Markov control processes: Basic optimality criteria", Springer-Verlag, New York.

[104] O. Hernández-Lerma and J. B. Lasserre [1999]: "Further topics on discrete-time Markov control processes", Springer-Verlag, New York.

[105] M. Herzberg and U. Yechiali [1994]: "Accelerating procedures of the value iteration algorithm for discounted Markov decision processes, based on a one-step look-ahead analysis", Operations Research *42*, 940–946.

[106] D.P. Heyman and M. J. Sobel [1984]: "Stochastic models in Operations Research, Volume II, MacGraw-Hill, New York.

[107] K. Hinderer [1970]: "Foundations of non-stationary dynamic programming with discrete time parameter", Springer-Verlag, New York.

[108] U.D. Holzbaur [1986a]: "Entscheidungsmodelle über angeordneten Körpern", Optimization *17*, 515–524.

[109] U.D. Holzbaur [1986b]: "Sensitivitätsanalysen in Entscheidungsmodellen", Optimization *17*, 525–533.

[110] U.D. Holzbaur [1994]: "Bounds for the quality and the number of steps in Bellman's value iteration algorithm", OR Spektrum *15*, 231–234.

[111] A. Hordijk [1971]: "A sufficient condition for the existence of an optimal policy with respect to the average cost criterion in Markovian decision processes", Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Academia, Prague, 263–274.

[112] A. Hordijk [1974]: "Dynamic programming and Markov potential theory", Mathematical Centre Tract 51, Amsterdam.

[113] A. Hordijk, R. Dekker and L.C.M. Kallenberg [1985]: "Sensitivity-analysis in discounted Markovian decision problems", OR Spektrum *7*, 143–151.

[114] A. Hordijk and L.C.M. Kallenberg [1979]: "Linear programming and Markov decision chains", Management Science *25*, 352–362.

[115] A. Hordijk and L.C.M. Kallenberg [1984a]: "Transient policies in discrete dynamic programming: linear programming including suboptimality tests and additional constraints", Mathematical Programming *30*, 46–70.

[116] A. Hordijk and L.C.M. Kallenberg [1984b]: "Constrained undiscounted stochastic dynamic programming", Mathematics of Operations Research *9*, 276–289.

[117] A. Hordijk and J.A. Loeve [1994]: "Undiscounted Markov decision chains with partial information; an algorithm for computing a locally optimal periodic policy", Mathematical Methods of Operations Research *40*, 163–181.

[118] A. Hordijk and H.C. Tijms [1974]: "The method of successive approximations and Markovian decision problems", Operations Research *22*, 519–521.

[119] A. Hordijk and H.C. Tijms [1975]: "A modified form of the iterative method of dynamic programming", Annals of Statictics *3*, 203–208.

[120] A. Hordijk and H.C. Tijms [1975]: "On a conjecture of Iglehart", Management Science *11*, 1342–1345.

[121] R.A. Howard [1960]: "Dynamic programming and Markov processes", MIT Press, Cambridge.

[122] R.A. Howard [1963]: "Semi-Markovian decision processes", Proceedings International Statistical Institute, Ottawa, Canada.

[123] Y. Huang and L.C.M. Kallenberg [1994]: "On finding optimal policies for Markov decision chains: a unifying framework for mean-variance tradeoffs", Mathematics of Operations Research *19*, 434–448.

[124] G. Hübner [1977]: "Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties", Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions, Reidel, Dordrecht, 257–263.

[125] G. Hübner [1988]: "A unified approach to adaptive control of average reward Markov decision processes", OR Spektrum *10*, 161–166.

[126] D. Iglehart [1963]: "Optimality of (s, S)-policies in the infinite horizon dynamic inventory problem", Management Science *9*, 259–267.

[127] T. Ishikida and P. Varaiya [1994]: "Multi-armed bandit problem revisited", Journal of Optimization Theory and Applications *83*, 113-154.

[128] R.G. Jeroslow [1972]: "An algorithm for discrete dynamic programming with interest rates near zero", Management Science Research Report no. 300, Carnegie-Mellon University, Pittsburgh.

[129] W.S. Jewell [1963a]: "Markov renewal programming. I: Formulation, finite return models", Operations Research *11*, 938–948.

[130] W.S. Jewell [1963b]: "Markov renewal programming. II: Infinite return models, example", Operations Research *11*, 949–971.

[131] L.C.M. Kallenberg [1981a]: "Finite horizon dynamic programming and linear programming", Methods of Operations Research *43*, 105–112.

[132] L.C.M. Kallenberg [1981b]: "Unconstrained and constrained dynamic programming over a finite horizon", Report, University of Leiden, The Netherlands.

[133] L.C.M. Kallenberg [1981c]: "Linear programming to compute a bias-optimal policy", in: B. Fleischmann et al. (eds.) "Operations Research Proceedings", 433–440.

[134] L.C.M. Kallenberg [1983]: "Linear programming and finite Markovian control problems", Mathematical Centre Tract 148, Mathematical Centre, Amsterdam.

[135] L.C.M. Kallenberg [1986]: "Note on M.N.Katehakis and Y.-R.Chen's computation of the Gittins index", Mathematics of Operations Research *11*, 184–186.

[136] L.C.M. Kallenberg [1992]: "Separable Markovian decision problem: the linear programming method in the multichain case", OR Spektrum *14*,43–52.

[137] L.C.M. Kallenberg [1999]: "Combinatorial problems in MDPs", Report, University of Leiden, The Netherlands (to appear in the Proceedings of the Changsha International Workshop on Markov Processes & Controlled Markov Chains).

[138] P.C. Kao [1973]: "Optimal replacement rules when the changes of states are semi-Markovian", Operations Research *21*, 1231–1249.

[139] M.N. Katehakis and C. Derman [1984]: "Optimal repair allocation in a series system", Mathematics of Operations Research *9*, 615–623.

[140] M.N. Katehakis and C. Derman [1989]: "On the maintenance of systems composed of highly reliable components", Management Science *35*, 551–560.

[141] M.N. Katehakis and A.F. Veinott, Jr. [1987]: "The multi-armed bandit problem: decomposition and computation", Mathematics of Operations Research *12*, 262–268.

[142] H. Kawai [1987]: "A variance minimization problem for a Markov decision process", European Journal of Operational Research *31*, 140–145.

[143] H. Kawai and N. Katoh [1987]: "Variance constrained Markov decision process", Journal of the Operations Research Society of Japan *30*, 88–100.

[144] J.G. Kemeny and J.L. Snell [1960]: "Finite Markov chains", Van Nostrand, Princeton.

[145] M. Klein [1962]: "Inspection-maintenance-replacement schedules under Markovian deterioration", Management Science *9*, 25–32.

[146] P. Kolesar [1966]: "Minimum-cost replacement under Markovian deterioration", Management Science *12*, 694–706.

[147] M. Kurano [1983]: "Adaptive policies in Markov decision processes with uncertain transition matrices", Journal of Information and Optimization Sciences *4*, 21–40.

[148] H. Kushner [1971]: "Introduction to stochastic control", Holt, Rineholt and Winston, New York.

[149] H. Kushner and A.J. Keinmann [1971]: "Accelerated procedures for the solution of discrete Markov control problems", IEEE Transactions on Automatic Control *16*, 147–152.

[150] E. Lanery [1967]: "Etude asymptotique des systèmes Markovien à commande", Revue d'Informatique et Recherche Operationelle *1*, 3–56.

[151] J.B. Lasserre [1994a]: "A new policy iteration scheme for Markov decision processes using Schweitzer's formula", Journal of Applied Probability *31*, 268–273.

[152] J.B. Lasserre [1994b]: "Detecting optimal and non-optimal actions in average-cost Markov decision processes", Journal of Applied Probability *31*, 979–990.

[153] W. Lin and P.R. Kumar [1984]: "Optimal control of a queueing system with two heterogeneous servers", IEEE Transactions on Automatic Control AC-29, 696–705.

[154] S.A. Lippman [1969]: "Criterion equivalence in discrete dynamic programming", Operations Research *17*, 920–923.

[155] J.Y. Liu and K. Liu [1994]: "An algorithm on the Gittins index", Systems Science and Mathematical Science *7*, 106–114.

[156] Q.-S. Liu and K. Ohno [1992]: "Multiobjective undiscounted Markov renewal program and its application to a tool replacement problem in an FMS", Information and Decision Techniques *18*, 67–77.

[157] Q.-S. Liu, K. Ohno and H. Nakayama [1992]: "Multi-objective discounted Markov processes with expectation and variance criteria", International Journal of System Science *23*, 903–914.

[158] J.A. Loeve [1995]: "Markov decision chains with partial information", PhD dissertation, University of Leiden, The Netherlands.

[159] W.S. Lovejoy [1987]: "Some monotonicity results for partially observed Markov processes", Operations Research *35*, 736–743.

[160] W.S. Lovejoy [1991a]: "Computationally feasible bounds for partially observed Markov decision processes", Operations Research *39*, 162–175.

[161] W.S. Lovejoy [1991b] "A survey of algorithmic methods for partially observed Markov decision processes", Annals of Op. Research *28*, 47–66.

[162] J. Macqueen [1966]: "A modified programming method for Markovian decision problems", Journal of Mathematical Analysis and Applications *14*, 38–43.

[163] J. Macqueen [1967]: "A test for suboptimal actions in Markov decision problems", Operations Research *15*, 559–561.

[164] A.S. Manne [1960]: "Linear programming and sequential decisions", Management Science, 259–267.

[165] U. Meister and U. Holzbaur [1986]: "A polynomial time bound for Howard's policy improvement algorithm", OR Spektrum *8*, 37–40.

[166] B.L. Miller and A.F. Veinott Jr. [1969]: "Discrete dynamic programming with a small interest rate", Annals of Mathematical Statistics *40*, 366–370.

[167] H. Mine and S. Osaki [1970]: "Markov decision processes", American Elsevier, New York.

[168] G.E. Monahan [1982]: "A survey of partially observable Markov decision processes: theory, models and algorithms", Management Science *28*, 1–16.

[169] T. Morton [1971]: "Undiscounted Markov renewal programming via modified successive approximations", Operations Research *19*, 1081–1089.

[170] J.L. Nazareth and R.B. Kulkarni [1986]: "Linear programming formulations of Markov decision processes", OR Letters *5*, 13–16.

[171] M.K. Ng [1999]: "A note on policy iteration algorithms for discounted Markov decision problems", OR Letters *25*, 195–197.

[172] A. Odoni [1969]: "On finding the maximal gain for Markov decision processes", Operations Research *17*, 857–860.

[173] S. Oezekici [1988]: "Optimal periodic replacement of multicomponent reliability systems", Operations Research *36*, 542–552.

[174] K. Ohno [1981]: "A unified approach to algorithms with a suboptimality test in discounted semi-Markov decision processes", Journal of the Operations Research Society of Japan *24*, 296–323.

[175] S. Osaki and H. Mine [1968]: "Linear programming algorithms for semi-Markovian decision processes", Journal of Mathematical Analysis and Applications *22*, 356–381.

[176] T. Parthasarathy, S.H. Tijs and O.J. Vrieze [1984], "Stochastic games with state independent transitions and reparable rewards in: G. Hammer and D. Pallaschke (eds.), Selected Topics in Operations Research and Mathematical Economics.

[177] L.K. Platzman [1977]: "Improved conditions for convergence in undiscounted Markov renewal programming", Op. Research *25*, 529–533.

[178] M.A. Pollatschek and B. Avi-Itzhak [1969]: "Algorithms for stochastic games with geometric interpretation", Management Science *15*, 399–415.

[179] E.L. Porteus [1971]: "Some bounds for discounted sequential decision processes", Management Science *18*, 7–11.

[180] E.L. Porteus [1975]: "Bounds and transformations for discounted finite Markov decision chains", Operations Research *23*, 761–784.

[181] E.L. Porteus [1980a]: "Improved iterative computation of the expected return in Markov and semi-Markov chains", Zeitschrift für Operations Research *24*, 155–170.

[182] E.L. Porteus [1980b]: "Overview of iterative methods for discounted finite Markov and semi-Markov chains", in: R. Hartley, L.C. Thomas and D.J. White (eds.), "Recent development in Markov decision processes", Academic Press, New York, 1–20.

[183] E.L. Porteus [1981]: "Computing the discounted return in Markov and semi- Markov chains", Naval Research Logistics Quarterly *28*, 567–577.

[184] E. L. Porteus and J.C. Totten [1978]: "Accelerated computation of the expected discounted return in a Markov chain", Operations Research *26*, 350–358.

[185] M.L. Puterman [1981]: "Computational methods for Markov decision methods", Proceedings of 1981 Joint Automatic Control Conference.

[186] M.L. Puterman [1994]: "Markov decision processes", Wiley, New York.

[187] M.L. Puterman and S.L. Brumelle [1979]: "On the convergence of policy iteration in stationary dynamic programming", Mathematics of Operations Research *4*, 60–69.

[188] M.L. Puterman and M.C. Shin [1978]: "Modified policy iteration algorithms for discounted Markov decision chains", Management Science *24*, 1127–1137.

[189] M.L. Puterman and M.C. Shin [1982]: "Action elimination procedures for modified policy iteration algorithms" Operations Research *30*, 301–318.

[190] D. Reetz [1973]: "Solution of a Markovian decision problem by successive overrelaxation", Zeitschrift für Operations Research *17*, 29–32.

[191] D. Reetz [1976]: "A decision exclusion algorithm for a class of Markovian decision processes" , Zeitschrift für Operations Research *20*, 125–131.

[192] U. Rieder [1991]: "Structural results for partially observed control problems", Zeitschrift für Operations Research *35*, 473–490.

[193] R. Righter [1994]: "Scheduling", in: M. Shaked and J.G. Shantikumar (eds.), "Stochastic orders and their applications", Academic Press, 381–432.

[194] M. Roosta [1982]: "Routing through a network with maximum reliability", Journal of Mathematical Analysis and Applications *88*, 341–347.

[195] K.W. Ross [1989]: "Randomized and past-dependent policies for Markov decision processes with multiple constraints", Operations Research *37*, 474–477.

[196] K.W. Ross and R. Varadarajan [1991]: "Multichain Markov decision processes with a sample path constraint: a decomposition approach", Mathematics of Operations Research *16*, 195–207.

[197] S.M. Ross [1969]: "A problem in optimal search and stop", Operations Research *17*, 984–992.

[198] S.M. Ross [1970]: "Applied probability models with optimization applications", Holden-Day, San Francisco.

[199] S.M. Ross [1974]: "Dynamic programming and gambling models", Advances in Applied Probability *6*, 593–606.

[200] S.M. Ross [1983]: "Introduction to stochastic dynamic programming", Academic Press, New York.

[201] U.G. Rothblum [1979]: "Iterated successive approximation for sequential decision processes", in J.W.B. van Overhagen and H.C. Tijms (eds.), "Stochastic control and optimization", Free University, Amsterdam, 30–32.

[202] H. Scarf [1960]: "The optimality of (s, S) polices in the dynamic inventory problem", Chapter 13 in: K.J. Arrow, S. Karlin and P. Suppes (eds.), "Mathematical methods in the social sciences", Stanford University Press, Stanford.

[203] H. Schellhaas [1974]: "Zur extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung", Zeitschrift für Operations Research *18*, 91–104.

[204] N. Schmitz [1985]: "How good is Howard's policy improvement algorithm?", Zeitschrift fur Operations Research *29*, 315–316.

[205] L. Schrage [1968]: "A proof of the optimality of the shortest remaining processing time discipline", Operations Research *16*, 687–690.

[206] P.J. Schweitzer [1965]: "Perturbation theory and Markovian decision processes", Ph.D. dissertation, M.I.T., Op. Research Center Report 15.

[207] P.J. Schweitzer [1968]: "Perturbation theory and finite Markov chains" Journal of Applied Probability *5*, 401–413.

[208] P.J. Schweitzer [1971a]: "Multiple policy improvements in undiscounted Markov renewal programming", Operations Research *19*. 784–793.

[209] P.J. Schweitzer [1971b]: "Iterative solution of the functional equations of undiscounted Markov renewal programming", Journal of Mathematical Analysis and Applications *34*, 495–501.

[210] P.J. Schweitzer [1984]: "A value-iteration scheme for undiscounted multichain Markov renewal programs", ZOR—Zeitschrift für Operations Research *28*, 143–152.

[211] P.J. Schweitzer [1985]: "The variational calculus and approximations in policy space for Markov decision processes", Journal of Mathematical Analysis and Applications *110*, 568–582.

[212] P.J. Schweitzer [1987]: "A Brouwer fixed-point mapping approach to communicating Markov decision processes", Journal of Mathematical Analysis and Applications *123*, 117–130.

[213] P.J. Schweitzer [1991]: "Block-scaling of value-iteration for discounted Markov renewal programming", Annals of Op. Research *29*, 603–630.

[214] P.J. Schweitzer and A. Federgruen [1977]: "The asymptotic behavior of value iteration in Markov decision problems", Mathematics of Operations Research *2*, 360–381.

[215] P.J. Schweitzer and A. Federgruen [1978a]: "Foolproof convergence in multichain policy iteration", Journal of Mathematical Analysis and Applications *64*, 360–368.

[216] P.J. Schweitzer and A. Federgruen [1978b]: "The functional equations of undiscounted Markov renewal programming", Mathematics of Operations Research *3*, 308–321.

[217] P.J. Schweitzer and A. Federgruen [1979]: "Geometric convergence of value iteration in multichain Markov decision problems", Advances of Applied Probability *11*, 188–217.

[218] L.I. Sennott [1999]: "Stochastic dynamic programming and the control of queueing systems", Wiley, New York.

[219] E.L. Sernik and S.I. Markus [1991]: "On the computation of the optimal cost function for discrete time Markov models with partial observations", Annals of Operations Research *29*, 471–512.

[220] J.F. Shapiro [1975]: "Brouwer's fixed point theorem and finite state space Markovian decision theory", Journal of Mathematical Analysis and Applications *49*, 710–712.

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

[222] Y.S. Sherif and M.L. Smith [1981]: "Optimal maintenance policies for systems subject to failure—A review", Naval Research Logistics Quarterly *28*, 47-74.

[223] K. Sladky [1974]: "On the set of optimal controls for Markov chains with rewards", Kybernatika *10*, 350–367.

[224] R.D. Smallwood [1966]: "Optimum policy regions for Markov processes with discounting", Operations Research *14*, 658–669.

[225] R.D. Smallwood and E.Sondik [1973]: "The optimal control of partially observable Markov processes over a finite horizon", Operations Research *21*, 1071–1088.

[226] D.R. Smith [1978]: "Optimal repairman allocation—asymptotic results", Management Science *24*, 665–674.

[227] M.J. Sobel [1981], "Myopic solutions of Markov decision processes and stochastic games", Operations Research *29*, 995–1009.

[228] M.J. Sobel [1985]: "Maximal mean/standard deviation ratio in an undiscounted MDP", OR Letters *4*, 157–159.

[229] M.J. Sobel [1994]: "Mean-variance trade-offs in an undiscounted MDP", Operations Research *42*, 175–183.

[230] E. Sondik [1978]: "The optimal control of partially observable Markov processes over the infinite horizon: discounted costs", Operations Research *26*, 282–304.

[231] I.M. Sonin [1999]: "The elimination algorithm for the problem of optimal stopping", Mathematical Methods of Operations Research *49*, 111–124.

[232] D. Spreen [1981]: "A further anti-cycling rule in multi-chain policy iteration for undiscounted Markov renewal programs", Zeitschrift für Operations Research *25*, 225–234.

[233] J. Stein [1988]: "On efficiency of linear programming applied to discounted Markovian decision problems", OR Spektrum *10*, 153–160.

[234] S.S. Stidham, Jr. [1985]: "Optimal control of admission to a queueing system", IEEE Transactions on Automatic Control *AC-30*, 705–713.

[235] S.S. Stidham, Jr. and R.R. Weber [1993]: "A survey of Markov decision models for control of networks of queues", Queueing Systems *13*, 291–314.

[236] J. Stoer and R. Bulirsch [1980]: "Introduction to numerical analysis", Springer-Verlag, New York.

[237] R. Strauch and A.F. Veinott, Jr. [1966]: "A property of sequential control processes", Report, Rand McNally, Chicago.

[238] M. Sun [1993]: "Revised simplex algorithm for finite Markov decision processes", Journal of Optimization Theory and Applications *79*, 405–413.

[239] L.C. Thomas [1981]: "Second order bounds for Markov decision processes", Journal of Mathematical Analysis and Applications *80*, 294–297.

[240] L.C. Thomas [1983]: "Constrained Markov decision processes as multi-objective problems", in: "Multi-objective decision making", Academic Press, 77–94.

[241] H.C. Tijms [1986]: "Stochastic modelling and analysis: a computational approach", Wiley, Chichester.

[242] J.N. Tsitsiklis [1986]: "A lemma on the multi-armed bandit problem", IEEE Transactions on Automatic Control *31*, 576–577.

[243] J.N. Tsitsiklis [1993]: "A short proof of the Gittins index theorem", Annals of Applied Probability *4*, 194–199.

[244] F.A. Van der Duyn Schouten and S.G. Vanneste [1990]: "Analysis and computation of (n, N)-strategies for maintenance of a two-component system", European Journal of Operations Research *48*, 260–274.

[245] J. Van der Wal [1980]: "The method of value oriented successive approximations for the average reward Markov decision processes", OR Spektrum *1*, 233–242.

[246] J. Van der Wal [1981]: "Stochastic dynamic programming", Mathematical Centre Tract 139, Mathematical Centre, Amsterdam.

[247] K.M. Van Hee [1978]: "Markov strategies in dynamic programming", Mathematics of Operations Research *3*, 191–201.

[248] K.M. Van Hee, A. Hordijk and J. Van der Wal [1977]: "Successive approximations for convergent dynamic programming", in: H.C. Tijms and J. Wessels (eds.), "Markov decision theory", Mathematical Centre Tract no. 93, Mathematical Centre, Amsterdam, 183–211.

[249] J.A.E.E. Van Nunen [1976a]: "A set of successive approximation method for discounted Markovian decision problems", Zeitschrift für Operations Research *20*, 203–208.

[250] J.A.E.E. Van Nunen [1976b]: "Contracting Markov decision processes", Mathematical Centre Tract 71, Mathematical Centre, Amsterdam.

[251] J.A.E.E. Van Nunen [1976c]: "Improved successive approximation methods for discounted Markovian decision processes", in: A. Prekopa (ed.), "Progress in Operations Research", North Holland, Amsterdam, 667–682.

[252] J.A.E.E. Van Nunen and J. Wessels [1976]: "A principle for generating optimization procedures for discounted Markov decision processes", Colloquia Mathematica Societatis Bolyai Janos, Vol. 12, North Holland, Amsterdam, 683–695.

[253] J.A.E.E. Van Nunen and J. Wessels [1977]: "The generation of successive approximations for Markov decision processes using stopping times", in: "Markov decision theory", H. Tijms and J. Wessels (eds.), Mathematical Centre Tract 93, Mathematical Centre, Amsterdam, 25–37.

[254] P.P. Varaiya, J.C. Walrand and C. Buyukkoc [1985]: "Extensions of the multi-armed bandit problem: the discounted case", IEEE Transactions on Automatic Control *30*, 426–439.

[255] A.F. Veinott, Jr. [1966a]: "On the optimality of (s, S) inventory policies: new condition and a new proof", SIAM Journal on Applied Mathematics *14*, 1067–1083.

[256] A.F. Veinott, Jr. [1966b]: "On finding optimal policies in discrete dynamic programming with no discounting", Annals of Math. Stats. *37*, 1284–1294.

[257] A.F. Veinott, Jr. [1969]: "Discrete dynamic programming with sensitive discount optimality criteria", Annals of Math. Stats. *40*, 1635–1660.

[258] A.F. Veinott, Jr. [1974]: "Markov decision chains", in: G.B. Dantzig and B.C. Eaves (eds.), "Studies in Optimization", Studies in Mathematics, Volume 10, The Mathematical Association of America, 124–159.

[259] R.C. Vergin and M. Scribian [1977]: "Maintenance scheduling for multi-component equipment", AIIE Transactions *9*, 297–305.

[260] O.J. Vrieze, [1987]: "Stochastic games with finite state and action spaces", CWI Tract 33, Centre for Mathematics and Computer Science, Amsterdam.

[261] K. Wakuta [1992]: "Optimal stationary policies in the vector-valued Markov decision process", Stochastic Processes and its Applications *42*, 149–156.

[262] K. Wakuta [1995]: "Vector-valued Markov decision processes and the systems of linear inequalities", Stochastic Processes and its Applications *56*, 159–169.

[263] K. Wakuta [1996]: "A new class of policies in vector-valued Markov decision processes", Journal of Mathematical Analysis and Applications *202*, 623–628.

[264] K. Wakuta [1999]: "A note on the structure of value spaces in vector-valued Markov decision processes", Mathematical Methods of Operations Research *49*, 77–86.

[265] J. Walrand [1988]: "An introduction to queueing networks", Prentice-Hall, Englewood Cliffs, New Jersey.

[266] R.R. Weber [1982]: "Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime.

[267] R.R. Weber [1992]: "On the Gittins index for multi-armed bandits", Annals of Applied Probability *2*, 1024–1033.

[268] R.R. Weber and S.S. Stidham, Jr. [1987]: "Optimal control of services rates in networks of queues", Advances in Applied Probability *19*, 202–218.

[269] G. Weiss [1982]: "Multiserver stochastic scheduling", in: M.A.H. Dempster, J.K. Lenstra and A.H.G. Rinnooy Kan (eds.), "Deterministic and stochastic scheduling", Reidel, Dordrecht, Holland, 157–179.

[270] G. Weiss [1988]: "Branching bandit processes", Probability in the Engineering and Information Sciences *2*, 269–278.

[271] J. Wessels and J.A.E.E. Van Nunen [1975]: "Discounted semi-Markov decision processes: linear programming and policy iteration", Statistical Neerlandica *29*, 1–7.

[272] J. Wessels [1977]: "Stopping times on Markov programming", in: Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Academia, Prague, pp. 575–585.

[273] C.C. White, III [1976]: "Procedures for the solution of a finite-horizon, partially observed, semi-Markov optimization problem", Operations Research *24*, 348–358.

[274] C.C. White, III [1991]: "A survey of solution techniques for the partially observed Markov decision process", Annals of Operations Research *33*, 215–230.

[275] C.C. White, III and W.T. Scherer [1989]: "Solution procedures for partially observed Markov decision processes", Op. Research *37*, 791–797.

[276] C.C. White, III and W.T. Scherer [1994]: "Finite-memory suboptimal design for partially observed Markov decision processes", Op. Research *42*, 439–455.

[277] D.J. White [1963]: "Dynamic programming, Markov chains, and the method of successive approximations", Journal of Mathematical Analysis and Applications *6*, 373–376.

[278] D.J. White [1978]: "Elimination of non-optimal actions in Markov decision processes", in: M.L. Puterman (ed.) Dynamic programming and its applications, Academic Press, New York, 131–160.

[279] D.J. White [1982]: "Multi-objective infinite-horizon discounted Markov decision processes", Journal of Mathematical Analysis and Applications *89*, 639–647.

[280] D.J. White [1985]: "Real applications of Markov decision theory", Interfaces *15:6*, 73–83.

[281] D.J. White [1988]: "Further real applications of Markov decision theory", Interfaces *18:5*, 55–61.

[282] D.J. White [1988]: "Mean, variance and probabilistic criteria in finite Markov decision processes: a review", Journal of Optimization Theory and Applications *56*, 1–30.

[283] D.J. White [1992]: "Computational approaches to variance-penalized Markov decision processes", OR Spektrum *14*, 79–83.

[284] D.J. White [1993]: "A survey of applications of Markov decision processes", Journal of the Operational Research Society *44*, 1073–1096.

[285] D.J. White [1993]: "Markov decision processes", Wiley, Chichester.

[286] D.J. White [1994]: "A mathematical programming approach to a problem in variance penalised Markov decision processes", OR Spektrum *15*, 225–230.

[287] D.J. White [1995]: "A superharmonic approach to solving infinite horizon partially observable Markov decision problems", Mathematical Methods of Operations Research *41*, 71–88.

[288] P. Whittle [1980]: "Multi-armed bandits and the Gittins index", Journal of the Royal Statistical Society, Series B *42*, 143–149.

[289] P. Whittle [1982]: "Optimization over time; dynamic programming and stochastic control", Volume I, Wiley, New York.

[290] P. Whittle [1982]: "Optimization over time; dynamic programming and stochastic control", Volume II, Wiley, New York.

[291] M. Yasuda [1988]: "The optimal value of Markov stopping problems with one-step look ahead policy", Journal of Applied Probability *25*, 544–552.

[292] Y.-S. Zheng and A. Federgruen [1991]: "Finding optimal (s, S)-policies is about as simple as evaluating a single policy", Op. Research *39*, 654–665.

Lodewijk Kallenberg
Mathematical Institute
University of Leiden
2300 RA Leiden, The Netherlands
kallenberg@math.leidenuniv.nl

2 BIAS OPTIMALITY

Mark E. Lewis
Martin L. Puterman

Abstract: The use of the long-run average reward or the *gain* as an optimality criterion has received considerable attention in the literature. However, for many practical models the gain has the undesirable property of being *underselective*, that is, there may be several gain optimal policies. After finding the set of policies that achieve the primary objective of maximizing the long-run average reward one might search for that which maximizes the "short-run" reward. This reward, called the *bias* aids in distinguishing among multiple gain optimal policies. This chapter focuses on the usefulness of the bias in distinguishing multiple gain optimal policies, its computation, and the implicit discounting captured by bias on recurrent states.

2.1 INTRODUCTION

The use of the long-run average reward or the *gain* as an optimality criterion has received considerable attention in the literature. However, for many practical models the gain has the undesirable property of being *underselective*, that is, there may be several gain optimal policies. Since gain optimality is only concerned with the long-run behavior of the system there is the possibility of many gain optimal policies. Often, this leads decision-makers to seek more sensitive optimality criteria that take into account short-term system behavior. We consider a special case of the sensitive optimality criteria which are considered in Chapter 7 of this volume.

Suppose the manager of a warehouse has decided through market studies and a bit of analysis that when long-run average cost is the optimality criterion an "(s, S)" ordering policy is optimal. That is to say that past demand patterns suggest it is optimal to reorder when the inventory falls below the level s and that it should be increased to S units when orders are made. Furthermore, suppose that there are many such limits that achieve long-run average optimality.

With this in mind, the manager has arbitrarily chosen the long-run average optimal policy (s', S'). In fact, in this example the manager could choose any ordering policy for any (finite) amount of time, and then start using any one of the optimal average cost policies and still achieve the optimal average cost. However, the decision-maker should be able to discern which of the optimal average cost policies is best from a management perspective and use that policy for all time. The use of the *bias* can assist in making such decisions.

In essence, after finding the set of policies that achieve the primary objective of maximizing the long-run average reward we search for that which maximizes the bias. The bias is the obvious next step among optimality criterion since it appears as the second term of the Laurent series expansion of the discount reward function. In very simple models with a single absorbing state and multiple policies to choose from on transient states the concept of bias optimality is easy to understand. In these models all policies are average optimal and the bias optimal policy is the one which maximizes the expected total reward before reaching the absorbing state. Consider the following simple example:

Example 2.1 Let $\mathbb{X} = \{1, 2\}$, $A_1 = \{a, b\}$, and $A_2 = \{c\}$. Furthermore, let $p(2|1, a) = p(2|1, b) = p(2|2, c) = 1$ and $r(1, a) = 100$, $r(1, b) = 1$, and $r(2, c) = 1$. It is easy to see that an average reward maximizing decision-maker would be indifferent which action is chosen in state 1, but any rational decision-maker would clearly prefer action 1. The analysis in this chapter will show, among other things, that using bias will resolve this limitation of the average reward criterion.

Unfortunately, this example gives an oversimplified perspective of the meaning of bias. In models in which all states are recurrent or models in which different policies have different recurrent classes, the meaning of bias optimality is not as transparent. It is one of our main objectives in this chapter to provide some insight on this point by developing a "transient" analysis for recurrent models based on relative value functions. We present an algorithmic and a probabilistic analysis of bias optimality and motivate the criterion with numerous examples. The reader of this chapter should keep the following questions in mind:

- How is bias related to average, total, and discounted rewards?
- How do we compute the bias?
- How are bias and gain computation related?
- In a particular problem, what intuition is available to identify bias optimal policies?
- Can we use sample path arguments to identify bias optimal policies?
- How is bias related to the timing of rewards?
- What does bias really mean in recurrent models?

2.2 HISTORICAL REFERENCES

Most Markov Decision Process (MDP) research has regarded bias as a theoretical concept. It was viewed as one of many optimality criteria that is more sensitive than long-run average optimality, but its application has received little attention. In many applications when there are multiple gain optimal policies there is only one bias optimal policy. Hence, the *bias-based* decision-maker need not look any further to decide between a group of gain optimal policies. Recently, Haviv and Puterman [5] showed in a queueing admission control model, with one server and a holding cost, that one can distinguish between two average optimal solutions by appealing to their bias. Their work was extended by Lewis, et al. [10] to a finite capacity, multi-class system with the possibility of multiple gain optimal policies. Further, Lewis and Puterman [12] showed that in the Haviv-Puterman model, the timing of rewards impacts bias optimality. Whereas the Haviv-Puterman paper showed that when rewards are received upon admitting a customer and there are two consecutive gain optimal control limits, say L and $L+1$, only $L+1$ is bias optimal, the Lewis-Puterman paper showed that if the rewards are received upon departure, only control limit L is bias optimal. This suggests that bias may implicitly discount rewards received later. Lewis and Puterman [11] present a new approach to compute the bias directly from the average optimality equations. This leads to sample path arguments that provide alternative derivations of the above mentioned results. In addition to the previously mentioned papers ([5], [10], [12]), the use of bias to distinguish between gain optimal policies has only been discussed in a short section of an expository chapter by Veinott [21]. Methods of computing optimal bias were considered for the finite state and action space case by Denardo [4] and Veinott [19] and on countable state and compact action spaces by Mann [14]. The extension of bias to general state spaces has not received much attention with the exception of section 10.3 of Hernandez-Lerma and Lasserre [8] where under certain assumptions it is shown to be equivalent to other sensitive optimality criteria.

Discount and average optimality have been considered extensively in the literature, therefore we will not provide a complete review here. For a comprehensive review refer to the survey paper of Arapostathis et al. [1] or Chapters 8 and 9 of Puterman [16]. Howard [9] introduced a policy iteration algorithm to solve the average reward model in the finite state space case. This has been considerably extended. For example, see the recent work of Hernández-Lerma and Lasserre [7] or Meyn [15]. Blackwell's [3] classic paper showed the existence of stationary optimal policies in the discounted finite state case and introduced a more sensitive optimality criterion now called *Blackwell* optimality. In the same paper, Blackwell introduced the concept of *nearly optimal* which is **equivalent** to bias optimality. In essence, Blackwell optimal policies are discount optimal for all discount rates close to 1. It turns out that Blackwell optimality implies bias optimality, so that we have the existence of bias optimal policies in the finite state and action space case as well. There is also a vast literature on sensitive optimality that indirectly addresses bias optimality (cf. Chapter 7 of this book or Veinott [20]). However, none of these works give an intuitive explanation for *which* policy the bias based decision-maker prefers and why.

2.3 DEFINITIONS

Assume that both the state space, $\mathbb{X}$, and the action space, $\mathbb{A}$, are finite. We offer several definitions of the bias of a stationary policy. Recall the definition of the long-run average reward or the **gain** of a policy π from equation 0.4, Chapter 0. A policy, π^*, that is optimal for the average reward per unit time criterion will be called **gain optimal**. Denote the set of stationary, deterministic (nonrandomized) policies by Π^S and a particular element of that set by ϕ. We now formalize the definition of bias.

Definition 2.1 *Suppose the Markov chain generated by a stationary, deterministic policy ϕ is aperiodic. The* **bias** *of ϕ given that the system started in state x, denoted $h(x,\phi)$, is defined to be*

$$h(x,\phi) \quad = \quad \sum_{n=0}^{\infty} \mathbb{E}_x^{\phi}[r(x_n,\phi(x_n)) - w(x_n,\phi)]. \tag{2.1}$$

Similarly, if the Markov chain generated by ϕ is periodic, we define the bias to be

$$h(x,\phi) \quad = \quad \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}_x^{\phi} \sum_{t=0}^{n} [r(x_t,\phi(x_t)) - w(x_t,\phi)]. \tag{2.2}$$

We say that a policy, (ϕ^*) is **bias optimal** *if it is* **gain optimal**, *and in addition*

$$h(x,(\phi^*)) \geq h(x,\phi) \text{ for all } x \in \mathbb{X}, \text{for all gain optimal } \phi \in \Pi^S.$$

Note that although we have only defined the bias for stationary policies, when the state and action spaces are finite, this class of policies is large enough to ensure that it contains a policy that maximizes bias among all policies. This will be discussed further momentarily. Furthermore, since we have assumed that the state space is finite and we will only be interested in stationary policies, the lim inf in equation 0.4, Chapter 0 can be replaced by a limit. As we will soon see, equation 0.4, Chapter 0 allows for an interpretation of bias as the total reward for a slightly modified process. This requires another definition.

Definition 2.2 *For a particular stationary policy ϕ and $x \in \mathbb{X}$ let*

$$e(x,\phi) \quad = \quad r(x,\phi(x)) - w(x,\phi) \tag{2.3}$$

be called the **excess reward** *of ϕ.*

We assume throughout the rest of this chapter that the Markov chain generated by a policy ϕ is aperiodic. We will adopt the convention of using subscripts or superscripts when writing vectors or matrices corresponding to particular policies. Let v_{ϕ}^N denote the vector of total expected rewards over the first N periods when using the policy ϕ so that

$$v_{\phi}^N \quad = \quad \sum_{t=1}^{N} P_{\phi}^{t-1} r_{\phi}. \tag{2.4}$$

From (2.1),

$$h_\phi = \sum_{t=1}^{N} P_\phi^{t-1} r_\phi - N w_\phi + \sum_{t=N+1}^{\infty} (P_\phi^{t-1} - P_\phi^*) r_\phi, \tag{2.5}$$

where $P_\phi^* = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P_\phi^i$ is the *limiting matrix* of the Markov chain generated by ϕ. Note $w_\phi = P_\phi^* r_\phi$. Since h_ϕ is finite, the third term in (2.5) approaches zero as $N \to \infty$. Hence, we may write,

$$v_\phi^N = N w_\phi + h_\phi + o(1) \tag{2.6}$$

where $o(1)$ denotes a vector with components that approach zero as $N \to \infty$. In component notation, we write $v_\phi^N(x) = v^N(x, \phi)$. As N becomes large, $v^N(x, \phi)$ approaches a line with slope $w(x, \phi)$ and intercept $h(x, \phi)$. Thus, for the process generated by the stationary policy ϕ we have an interpretation of the gain as the asymptotic rate of increase relative to the horizon length of the total reward and the bias as the intercept or initial level.

An alternative interpretation can be realized from (2.1) if one defines a new system in which for each stationary policy, the reward function is replaced with the excess reward function. The bias is then the expected (finite) total reward in the modified system. Alternatively, the bias represents the expected difference in total reward under policy ϕ between two different initial conditions; when the process begins in state s and when the process begins with the state selected according to the probability distribution defined by the s^{th} row of P_ϕ^*. If we assume that the process under ϕ is unichain, this initial distribution is the stationary distribution of the chain. When the process is multichain, the distributions specified by the rows of P_ϕ^* may vary with the initial state. Under either scenario, it is well-known that the convergence to steady state occurs exponentially fast.

The interpretation of the bias as the total reward while correct, also has its limitations and can be misleading. In economic applications financial rewards received earlier, rather than later, are more valuable. In fact, earlier rewards translate into decision-making flexibility regardless of the volatility of the industry. Consider the discounted reward function, $v(x, \pi, \beta)$ where β is the discount rate,

Definition 2.3 *We say that a policy π^* is* **n-discount optimal** *for some integer $n \geq -1$ if*

$$\liminf_{\beta\uparrow 1} (1-\beta)^{-n} [v(x, \pi^*, \beta) - v(x, \pi, \beta)] \geq 0 \tag{2.7}$$

for all $x \in \mathbb{X}$ and $\pi \in \Pi$.

Since the state and action space are finite, we know that there exists stationary, deterministic $n-$discount optimal policies for all n (see Theorem 10.1.5 of Puterman [16]).

On the surface it might appear overly restrictive to not define bias for nonstationary policies and say that a policy is bias optimal when it maximizes

bias only among gain optimal *stationary policies*. On the other hand, the notion of 0−discount optimality is more general since a policy is 0−discount optimal when (2.7) holds for $n = 0$ in the class of *all* policies. Theorem 10.1.6 of Puterman [16] resolves this possible point of confusion. It asserts that a stationary policy maximizes the bias among all gain optimal stationary policies, if and only if it is 0−discount optimal. Consequently by restricting attention to bias optimality within the class of stationary gain optimal policies, we are also finding a 0−discount optimal policy in the class of all policies. With this close relationship to discounting, it stands to reason that the bias retains some of the attributes of discounting. We will elaborate more on this subject later in the chapter.

The next theorem that we present does not add intuition to the meaning of the bias. However, it does hint at the similarities between the computation of the gain and the bias. Let $H_P \equiv (I - P + P^*)^{-1}(I - P^*)$ be the deviation matrix of P.

Theorem 2.1 *The bias of a stationary, policy ϕ satisfies*

$$h(x, \phi) \quad = \quad (H_{P_\phi} r_\phi)(x) \tag{2.8}$$

If we expand $(I - P_\phi + P_\phi^*)^{-1}$ in a power series the equivalence of (2.1) and (2.8) becomes clear. Recall that if P_ϕ^* is the stationary distribution of the Markov chain generated by the policy ϕ the gain may be computed

$$w_\phi \quad = \quad P_\phi^* r_\phi. \tag{2.9}$$

Hence, the deviation matrix replaces the stationary distribution when computing the bias. This begs the question, can the bias be computed using some of the methods available in the vast literature on the long-run average reward problem? This is the subject of the next section.

A final interpretation of bias is available as the second term in the Laurent series expansion of the discount value function. The following appears in Puterman [16], Theorem 8.2.3 and is treated thoroughly in Chapter 7 of this book. Let $\rho = (1 - \beta)/\beta$ be the interest rate.

Theorem 2.2 *Let ν denote the nonzero eigenvalue of $I - P_\phi$ with the smallest modulus. Then, for $0 < \rho < |\nu|$,*

$$v(\phi, \beta) \quad = \quad (1 + \rho) \left[\sum_{n=-1}^{\infty} \rho^n y_n^\phi \right] \tag{2.10}$$

where $y_{-1}^\phi = w_\phi$, $y_0^\phi = h_\phi$, and $y_n^\phi = (-1)^n H_{P_\phi}^{n+1} r_\phi$.

As was alluded to earlier, the bias appears as the second term in the Laurent series expansion of the total expected discounted reward function. The above observations lead to interpretations of the bias of a stationary policy as:

1. the intercept of the asymptotic total reward process (in the aperiodic case),

2. the total difference between the process beginning in a particular state and that which begins in stationarity and,

3. the second term of the Laurent series expansion of the expected discounted reward.

We now turn to methods for evaluating the bias.

2.4 COMPUTING THE BIAS FROM THE EVALUATION EQUATIONS

In most practical examples, computation of the bias directly from the above definitions is not feasible. We discuss some practical methods for the computation of the bias of a fixed stationary policy ϕ. These methods also lead to an intuitive understanding of bias. The gain and the bias of ϕ may be computed by solving the following system of linear equations:

$$w = P_\phi w, \tag{2.11}$$

$$h = r_\phi - w + P_\phi h, \tag{2.12}$$

and

$$k = -h + P_\phi k \tag{2.13}$$

for vectors w, h, and k. Specifically, the gain and the bias of ϕ satisfy (2.11) and (2.12) and there exists some vector k which together with the bias satisfies (2.13). Moreover, the gain and the bias are the unique vectors with these properties. Furthermore, to compute the bias one may replace (2.13) with the condition that $P_\phi^* h = 0$. However, this expression is not useful in computation except in simple illustrative examples such as Example 2.1. Recall that bias optimality is equivalent to 0−discount optimality. It should then come as no surprise that (2.11)-(2.13) are a particular case of the $n+3$ equations for $n-$discount optimality discussed in Chapter 7. To see that the gain and bias (along with some vector k) satisfy (2.11) and (2.12), note that if we multiply (2.12) by P_ϕ^* we have

$$P_\phi^* h = P_\phi^* r_\phi - P_\phi^* w + P_\phi^* h,$$

thus,

$$P_\phi^* w = P_\phi^* r_\phi. \tag{2.14}$$

However, by repeated application of (2.11)

$$w = \frac{1}{N}[w + P_\phi w + \cdots + P_\phi^{N-1} w].$$

Taking limits as $N \to \infty$ we have $w = P_\phi^* w$. Combining this with (2.14) yields $w = P_\phi^* r_\phi$. To show that h_ϕ satisfies (2.12), and (2.13) with $w = w_\phi$,

multiply (2.13) by P_ϕ^* to get $P_\phi^* h = 0$. Hence, from (2.12)

$$\begin{aligned} h - P_\phi h + P_\phi^* h &= r_\phi - w_\phi \\ &= r_\phi - P_\phi^* r_\phi. \end{aligned}$$

Since $I - P_\phi + P_\phi^*$ is invertible we get $h = (I - P_\phi + P_\phi^*)^{-1}(I - P_\phi^*)r_\phi = H_{P_\phi} r_\phi$ as desired. We refer to (2.11) and (2.12) as the *average evaluation equations* (AEE) and to (2.11), (2.12), and (2.13) as the *bias evaluation equations* (BEE). If we restrict attention to (2.11) and (2.12) only, the gain is uniquely determined and the bias is determined up to m additive constants, where m is the number of closed, recurrent classes for the Markov chain generated by ϕ. Furthermore, each of the equations can be written in the form

$$(I - P_\phi)u = v. \tag{2.15}$$

It would be nice if $I - P_\phi$ was invertible. Of course it is not. However, H_{P_ϕ} is often called the *Drazin inverse* of $I - P_\phi$ (denoted $(I - P_\phi)^\#$) and exhibits many desirable properties of matrix inverses. Namely,

$$H_{P_\phi}^\# H_{P_\phi} H_{P_\phi}^\# = H_{P_\phi}^\#, \quad H_{P_\phi} H_{P_\phi}^\# = H_{P_\phi}^\# H_{P_\phi}, \quad \text{and} \quad H_{P_\phi} H_{P_\phi}^\# H_{P_\phi} = H_{P_\phi} \tag{2.16}$$

The Drazin inverse is used to derive the chain structure of a Markov chain. For more information on the Drazin inverse see Appendix A of Puterman [16]. In the following example we compute the bias using the BEE and the Drazin inverse of $I - P_\phi$.

Example 2.2 Let $\mathbb{X} = \{x_1, x_2, x_3\}$ and suppose the Markov chain generated by the policy ϕ has transition structure

$$P_\phi = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}. \tag{2.17}$$

It is not hard to show that

$$P_\phi^* = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \tag{2.18}$$

and

$$H_{P_\phi} = \begin{bmatrix} 1/3 & 0 & -1/3 \\ -1/3 & 1/3 & 0 \\ 0 & -1/3 & 1/3 \end{bmatrix} \tag{2.19}$$

If $r'_\phi = \{1, -1, 0\}$ we get,

$$w_\phi = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$h_\phi = \begin{bmatrix} 1/3 \\ -2/3 \\ 1/3 \end{bmatrix}.$$

Furthermore, the BEE are satisfied since

$$\begin{aligned} 1/3 &= 1 - 0 - 2/3 \\ -2/3 &= -1 - 0 + 1/3 \\ 1/3 &= 0 - 0 + 1/3. \end{aligned}$$

and

$$\begin{aligned} k(x_1) &= -1/3 + k(x_2) \\ k(x_2) &= 2/3 + k(x_3) \\ k(x_3) &= -1/3 + k(x_1) \end{aligned}$$

has the solution, $k(x_1) = 0, k(x_2) = 1/3$, and $k(x_3) = -1/3$. Similar to the bias and the AEE, the vector k is not the unique vector that satisfies (2.13). Since the model consists of one recurrent class, $(k + c1, h_\phi)$ is also a solution to (2.13) for any constant c, where 1 denotes a vector with all components equal to 1. As we will see, a similar result can be used to simplify the computation of the bias of Markov decision processes.

2.4.1 Bias and total reward

In this section, we describe some models in which the bias is equivalent to the expected total reward. This is important because the expected total reward criteria has received considerable attention recently in the reinforcement learning literature (cf. Barto and Sutton [2]). This also gives us insight for alternative methods for computing the bias sample pathwise. Let $a^+ = \max\{a, 0\}$ be the positive part and $a^- = \max\{-a, 0\}$ be the negative part of a real number a. Suppose we define

$$v_+(x, \pi) = \mathbb{E}_x^\pi \left\{ \sum_{n=0}^{\infty} r^+(x_n, y_n) \right\}, \tag{2.20}$$

and

$$v_-(x, \pi) = \mathbb{E}_x^\pi \left\{ \sum_{n=0}^{\infty} r^-(x_n, y_n) \right\}. \tag{2.21}$$

If either (2.20) or (2.21) is finite for all $x \in \mathbb{X}$, then $\lim_{N\to\infty} v_\pi^N$ exists. If both are finite then so is $\lim_{N\to\infty} v_\pi^N$. Furthermore, if (2.20) and (2.21) are finite for **all** $\pi \in \Pi$, then $w_\pi = 0$ for all $\pi \in \Pi$ (see Proposition 10.4.1, part (c) of Puterman [16]). Using Definition 2.1, it stands to reason that if the total reward is finite, the bias and total reward should coincide. This is precisely the case.

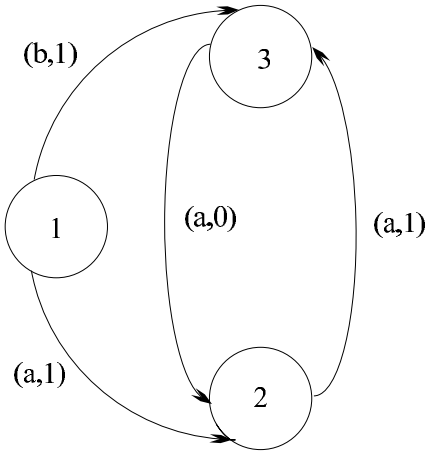

Figure 2.1 A deterministic example with finite gain, where the bias differs from the total reward until absorption into a recurrent class.

Proposition 2.1 *Let $\phi \in \Pi^S$ and suppose $v_+(x,\phi)$ and $v_-(x,\phi)$ are finite for all $x \in \mathbb{X}$, then $\lim_{N\to\infty} v_\phi^N = v_\phi = h_\phi$.*

The importance of this result is that in models with expected total reward criterion whenever v_+ and v_- are finite, methods developed for determining bias optimal policies apply to compute optimal policies. This avoids many of the complexities that have developed in the theory of models with expected total reward criterion; especially the need to distinguish positive and negative models. See Puterman [16], especially Chapter 7, for more on this issue.

2.4.2 Unichain Markov decision processes

Notice that in Example 2.1 computing the bias is straightforward. For the policy that uses action a in state 1, the bias is 99 and for the policy that uses action b the bias is zero. Given this example and the previous discussion, one might conjecture that if there is but one recurrent class, we could replace the reward with the excess reward function and compute the bias on the transient states as the total reward until reaching the recurrent class. After all, this has the effect of treating the recurrent class as a single state with zero reward and confining the transient analysis to the transient states. The following example shows that this is not the case.

Example 2.3

Let $\mathbb{X} = \{1,2,3\}$. Suppose $A_1 = \{a,b\}$, $A_2 = \{a\}$ and $A_3 = \{a\}$, such that $p(2|1,a) = p(3|1,b) = 1$ and $p(3|2,a) = p(2|3,a) = 1$. Furthermore, assume that $r(1,a) = r(1,b) = r(2,a) = 1$ while $r(3,a) = 0$. See Figure 2.3. If γ chooses action a in state 1 and δ chooses action b, it is not hard to show that $h_\gamma = \{3/4, 1/4, -1/4\}$ and $h_\delta = \{1/4, 1/4, -1/4\}$. Hence, despite the fact that the total rewards until reaching the recurrent class for both policies are

the same, the biases differ. This hints at a point that we make later in our discussion; the bias-based decision-maker distinguishes when during a recurrent cycle rewards are received. Except where noted otherwise, assume for now that the process generated by any stationary policy is unichain; that is, the process has a single ergodic class and perhaps some transient states. Hence, the following hold

1. w is constant which we express as $w1$.
2. (2.11) is redundant and (2.12) becomes
$$h = r_\phi - w1 + P_\phi h. \tag{2.22}$$
3. If (w, h) satisfies (2.22), $w = w_\phi$ and h is unique up to a constant.
4. If $(w, h_\phi^{rv(\alpha)})$ satisfies (2.22), and $h_\phi^{rv(\alpha)}(\alpha) = 0$, $h_\phi^{rv(\alpha)}$ is unique and is called the *relative value function* of ϕ at α.
5. (w_ϕ, h_ϕ) is the unique solution of (2.22) and the additional condition $P^* h = 0$.

With these observations in mind, we have the following definition which was originally introduced in [11].

Definition 2.4 *Let $\phi \in \Pi^S$ be a fixed stationary policy for which P_ϕ is unichain. For each solution to the average evaluation equations (w_ϕ, h), the constant difference between h and the bias of ϕ, h_ϕ, denoted $c_\phi(h)$, is called the* **bias constant** *associated with h.*

We now show how to use an arbitrary solution (w, h) of the AEE to compute the bias constant and therefore the bias. Suppose α is a recurrent state for the Markov chain generated by ϕ. Denote the first time the process enters the state α by τ_α. That is,

$$\tau_\alpha \quad = \quad \min\{n > 0 | x_n = \alpha\}. \tag{2.23}$$

Let

$$h_\alpha^\phi(s) \quad = \quad \mathbb{E}_s^\phi \left(\sum_{n=0}^{\tau_\alpha - 1} [r(x_n, \phi(x_n)) - w_\phi] \right). \tag{2.24}$$

Note

$$\begin{aligned} h_\alpha^\phi(s) &= r_\phi(s, \phi(s)) - w_\phi + \mathbb{E}^\phi \left(\sum_{n=1}^{\tau_\alpha - 1} [r(x_n, \phi(x_n)) - w_\phi] \middle| x_0 = s \right) \\ &= r_\phi(s, \phi(s)) - w_\phi + (P_\phi h_\alpha^\phi)(s). \end{aligned}$$

Hence, (w_ϕ, h_α^ϕ) satisfies (2.22) for policy ϕ. Furthermore, $h_\alpha^\phi(\alpha) = 0$. From point 4 above, h_α^ϕ is the relative value function of the policy ϕ at the reference state α; $h_\alpha^\phi = h_\phi^{rv(\alpha)}$. In addition, from (2.24) we interpret the relative

value function as *the expected total excess reward until the system reaches the recurrent state* α.

Since for a fixed policy ϕ the relative value functions and the bias satisfy the AEE they must differ by a constant. Choose positive recurrent state α and let $c_\phi(h_\phi^{rv(\alpha)})$ be the bias constant associated with the relative value function. Then

$$h_\phi = h_\phi^{rv(\alpha)} + c_\phi(h_\phi^{rv(\alpha)})1, \tag{2.25}$$

and

$$P_\phi^* h_\phi = P_\phi^* h_\phi^{rv(\alpha)} + P_\phi^* c_\phi(h_\phi^{rv(\alpha)})1. \tag{2.26}$$

However, since $P_\phi^* h_\phi = 0$,

$$P_\phi^* h_\phi^{rv(\alpha)} = -P_\phi^* c_\phi(h_\phi^{rv(\alpha)})1 \tag{2.27}$$

$$= -c_\phi(h_\phi^{rv(\alpha)})1, \tag{2.28}$$

where the last equality follows from the fact that the unichain assumption implies that the rows of P_ϕ^* are equal. Making the appropriate substitution into (2.25) yields the following proposition.

Proposition 2.2 *Suppose a finite state and action space Markov decision process is unichain. Let* ϕ *be a stationary policy. Denote the relative value function of* d *at a recurrent state* α *by* $h_\phi^{rv(\alpha)}$. *Let* $c_\phi = (P_\phi^* h_\phi^{rv(\alpha)})(s)$ *for any state* $s \in \mathbb{X}$. *Then the bias of* ϕ *is given by* $h_\phi = h_\phi^{rv(\alpha)} - c_\phi 1$.

Remark 2.1 *The above result holds in the countable state case provided* α *is positive recurrent.*

Hence, while in Definition 2.1 we replace the stationary distribution in the computation of the gain by the deviation matrix to compute the bias, we can instead replace the reward function in $w = P_\phi^* r_\phi$, by the relative value function to compute the bias, by computing $P_\phi^* h_\phi^{rv}$. Furthermore, applying a classic result in renewal theory we have (cf. Chapter 3.6 of Ross [17]),

$$h_\phi(s) = h_\phi^{rv(\alpha)}(s) - \frac{\mathbb{E}_\alpha^\phi \sum_{n=0}^{\tau_\alpha - 1} h_\phi^{rv(\alpha)}(x_n)}{\mathbb{E}_\alpha^\phi \tau_\alpha}. \tag{2.29}$$

This expression allows us to compute the bias of a stationary policy sample pathwise. The following simple example illustrates each of these methods for computing bias.

Example 2.4 Suppose $\mathbb{X} = \{0,1,2,3\}$, $A_0 = \{a,b,c\}$, $A_1 = \{a\}$, $A_2 = \{b\}$, and $A_3 = \{c\}$, $r(0,a) = r(2,b) = 1$, $r(0,b) = r(1,a) = -1$, $r(0,c) = r(3,c) = 0$ and $p(1|0,a) = p(3|0,c) = p(2|0,b) = p(3|1,a) = p(3|2,b) = p(0|3,c) = 1$. Let δ be the decision rule that chooses action a in state zero, γ be the decision rule that selects action c and ψ be that which chooses action b. Clearly, this model is unichain and $w_\delta = w_\gamma = w_\psi = 0$. Suppose we arbitrarily choose $\{0\}$

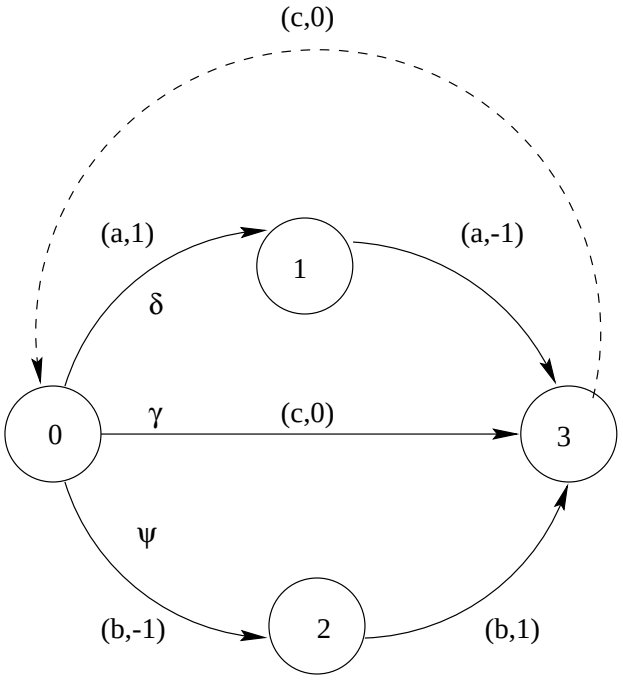

Figure 2.2 A deterministic example with average reward 0.

as the reference state (so $h_\phi^{rv(\alpha)}(0) = 0$ for each ϕ). Since state zero is the only state that requires a decision, the relative value function at α is the same for all policies. Hence, we suppress the dependence on ϕ. By examination of Figure 2.2 we have $h^{rv(\alpha)}(1) = -1$, $h^{rv(\alpha)}(2) = 1$, and $h^{rv(\alpha)}(3) = 0$. The stationary distributions, β_ϕ^* say, are $\beta_\gamma^* = \{1/2, 0, 0, 1/2\}$, $\beta_\delta^* = \{1/3, 1/3, 0, 1/3\}$, and $\beta_\psi^* = \{1/3, 0, 1/3, 1/3\}$. The bias constants are $-\beta_\gamma^* h^{rv(\alpha)} = 0$, $-\beta_\delta^* h^{rv(\alpha)} = 1/3$, and $-\beta_\psi^* h^{rv(\alpha)} = -1/3$. Hence, we have

$$h_\gamma = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \quad h_\delta = \begin{bmatrix} 1/3 \\ -2/3 \\ 4/3 \\ 1/3 \end{bmatrix}, \text{ and } \quad h_\psi = \begin{bmatrix} -1/3 \\ -4/3 \\ 2/3 \\ -1/3 \end{bmatrix}.$$

Note that if we neglect state 2, the transition structure and rewards of the Markov reward process generated by δ are identical to that discussed in Example 2.2 and thus the computed bias vectors (without state 2) are the same. We will return to this example throughout the rest of this section.

2.4.3 The Average Optimality Equation (Unichain Case)

The definition of bias optimality requires that one know the set of gain optimal policies *a priori*. This is not usually the case, except in the total reward models

of Chapter 2.4.1. In this section we review conditions for gain optimality which lead to algorithms for computing bias optimal policies in a similar way to computing gain optimal policies.

Since the state and action spaces are finite, computation of average optimal policies reduces to solving the *average optimality equations* (AOE)

$$h = \max_{\phi \in \Pi^S} \{r_\phi - w1 + P_\phi h\} \tag{2.30}$$

for w and h. Let $G(h)$ be the set of policies that achieve the maximum in (2.30). That is,

$$\delta \in argmax_{\phi \in \Pi^S} \{r_\phi + P_\phi h\} \equiv G(h). \tag{2.31}$$

We refer to (2.31) as the *average optimality selection equations* (AOSE). To begin our analysis of the average optimality equations, we consider a special case of a result of Schweitzer and Federgruen [18]. In essence, the result states that solutions of the AOE must differ by a constant just as they do for the AEE.

Proposition 2.3 *Suppose all stationary policies are unichain and let* (w_1, h_1) *and* (w_2, h_2) *be solutions to the AOE. Then* $w_1 = w_2$ *and*

$$h_1 = h_2 + c1 \tag{2.32}$$

for some constant c. In particular, if $h_1 = h^*$ *is the optimal bias, then*

$$h^* = h_2 + c^*(h_2)1. \tag{2.33}$$

We refer to $c^*(h_2)$ as the **optimal bias constant** associated with h_2. Note that if for δ^*, $h_{\delta^*} = h^*$, the optimal bias constant is the bias constant for h_{δ^*} and h^*. Further note that this result does not require that $\mathbb{X}$ be finite, only that the gain is constant. A nice discussion of the average optimality equations on Borel spaces can be found in Hernandez-Lerma and Lasserre [6].

It is easy to see that all three policies considered in Example 2.4 satisfy the AOSE and that the bias of each differs by a constant as indicated by Proposition 2.3.

We now return to the question of whether there are decision rules that are gain optimal, but do not satisfy the AOSE. When all policies generate irreducible Markov chains it is known that the average optimality equations are indeed necessary and sufficient (see Lewis, et al. [10]). The following example shows that this need not be the case in unichain models.

Example 2.5 Suppose $\mathbb{X} = \{1, 2\}$, $A_1 = \{a, b\}$ and $A_2 = \{c\}$, $r(1, a) = 2$, $r(1, b) = 3$ $r(2, c) = 1$ and $p(2|1, a) = p(2|1, b) = p(2|2, c) = 1$. Let δ be the decision rule that chooses action a in state 1 and let γ be the decision rule that chooses action b in state 1. This model is unichain and $w_\delta = w_\gamma = 1$, $h_\delta^{rv(2)}(1) = 1$, $h_\gamma^{rv(2)}(1) = 2$, $h_\delta^{rv(2)}(2) = h_\gamma^{rv(2)}(2) = 0$. Since $h_\delta^{rv(2)}$ and $h_\gamma^{rv(2)}$ do *not* differ by a constant, it follows from Proposition 2.3, that $(w_\delta, h_\delta^{rv(2)})$ and $(w_\gamma, h_\gamma^{rv(2)})$ cannot both satisfy the average optimality equations, even though both policies are average optimal. ∎

In the sequel we show that our previous observations lead to simple sample path arguments as well as algorithmic solution methods for finding the optimal bias.

2.5 THE BIAS OPTIMALITY EQUATION

Suppose in addition to satisfying the AOE, there exists a vector k, such that h satisfies

$$k = \max_{\phi \in G(h)} \{-h + P_\phi k\} \tag{2.34}$$

and δ satisfies

$$\delta \in argmax_{\phi \in G(h)} \{P_\phi k\} \tag{2.35}$$

Then δ is bias optimal and h is the optimal bias. We refer to the combined set (2.34) and the AOE as the *bias optimality equations* (BOE) and to (2.35) as the *bias optimality selection equations* (BOSE).

We can take advantage of the result of Proposition 2.3; if (w, h_1) and (w, h_2) are solutions to the AOE then h_1 and h_2 differ by a constant. Upon substituting (2.33) into (2.34) for the relative value with reference state α when $(w, h^{rv(\alpha)})$ satisfies the AOE, we have the following important result.

Theorem 2.3 *Suppose $h^{rv(\alpha)}$ is a relative value function with reference state α such that $(w^*, h^{rv(\alpha)})$ is a solution to the AOE. The BOE (2.34) can be rewritten*

$$k = max_{\phi \in G(h^{rv(\alpha)})} \{-h^{rv(\alpha)} - c1 + P_\phi k\}. \tag{2.36}$$

Further, suppose (k^*, c^*) satisfies (2.36). Then k^* is unique up to a constant and c^* is the optimal bias constant associated with $h^{rv(\alpha)}$.

To see that the second part of the theorem follows directly from the first, observe that (2.36) has exactly the same form as the AOE (2.30). That is to say, setting $r_\phi = -h^{rv(\alpha)}$ and $w = -c1$ we have again the AOE. Thus, in a unichain model, the result is immediate from existing theory on the AOE. Furthermore, all solution methods and theory for the AOE apply directly in this case. In particular, (2.36) can be solved by the same value iteration or policy iteration algorithms used to find gain optimal policies in unichain Markov Decision Processes, however, now we base them on (2.36) .

Example 2.6 Consider the model of Example 2.4. Suppose we begin policy iteration with policy $\rho_0 = \gamma$. Recall $(h^{rv(0)})' = \{0, -1, 1, 0\}$. We must find (c_0, k_0) to satisfy

$$k_0 = -h^{rv(0)} - c_0 1 + P_\gamma k_0. \tag{2.37}$$

One can easily show that $k_0' = \{0, 1, -1, 0\}$ and $c_0 = 0$ is a solution to this system. To choose the next policy, we find a policy, ρ_1 such that

$$\rho_1 \in argmax\{-h^{rv(0)} + P_\gamma k_0\}. \tag{2.38}$$

Since we need only make a decision in state 0, note that

$$\begin{aligned} \rho_1(0) &= argmax\{0+1, 0, 0-1\} && (2.39)\\ &= a. && (2.40) \end{aligned}$$

Thus, $\rho_1 = \delta$. For the second iteration of the algorithm we must solve,

$$\begin{aligned} k_1(0) &= 0 - c + k_1(1),\\ k_1(1) &= 1 - c + k_1(3),\\ k_1(2) &= -1 - c + k_1(3),\\ k_1(3) &= 0 - c + k_1(0). \end{aligned}$$

One can verify that $k_1 = \{0, 1/3, -5/3, -1/3\}$ and $c = 1/3$ satisfies the above equations. Furthermore, no further improvements can be made. Notice that the bias of ρ_1 is $1/3$ higher than the relative value function. That is, c is the bias constant of ρ_1. This agrees with the solution found in Example 2.4. ∎

Alternatively, as in the AEE, if (w_ϕ, h) satisfy the AOE and

$$P_\phi^* h = 0 \qquad (2.41)$$

where P_ϕ^* is the stationary distribution of the chain generated by ϕ, then h is the optimal bias. Neglecting the trivial case $r_\phi = 0$ for all $\phi \in \Pi^S$, it is interesting to note that since P_ϕ^* is positive on the recurrent class generated by ϕ, the optimal bias must have both positive and negative elements. We will show in the examples that follow that we can take advantage of this fact. Suppose that ϕ is bias optimal. From Proposition 2.2

$$h^* = h^{rv(\alpha)} - (P_\phi^* h^{rv(\alpha)}). \qquad (2.42)$$

In essence, solving for the policy with maximum bias reduces to finding the policy that achieves the maximum bias constant, say c^*. That is,

$$c^* 1 = \max_{\phi \in G(h^{rv(\alpha)})} \{-P_\phi^* h^{rv(\alpha)}\} = -\min_{\phi \in G(h^{rv(\alpha)})} \{P_\phi^* h^{rv(\alpha)}\} \qquad (2.43)$$

where $h^{rv(\alpha)}$ is any relative value function of a gain optimal policy. Thus, under the assumption that there exists a state α that is recurrent for all decision rules in $G(h^{rv(\alpha)})$ we can alternatively compute the optimal bias by solving

$$c^* = -\min_{d \in G(h^{rv(\alpha)})} \left(\frac{\mathbb{E}_\alpha \sum_{n=0}^{\tau_\alpha - 1} h^{rv(\alpha)}(x_n, \phi(x_n))}{\mathbb{E}_\alpha^\phi \tau_\alpha} \right) \qquad (2.44)$$

where the expectation is taken with respect to the probability transition function conditioned on starting in state α. Since we are minimizing, $h^{rv(\alpha)}$ can be interpreted as a cost function. Thus, finding a bias optimal policy corresponds to a minimum average cost problem. Furthermore, one might notice that given a relative value function we can solve for the gain, by noting $w_\phi = r_\phi + P_\phi h_\phi^{rv(\alpha)} - h_\phi^{rv(\alpha)}$. That is, the relative value function can be used

to obtain both the gain and the bias. The crux of the analysis then for finding gain and bias, lies in understanding the relative value functions. We emphasize the importance of these observations in the following examples.

Since we will often be interested in the difference in the cost starting in states s and $s+1$, define for a function b on $\mathbb{N}$, $\Delta b(s) \equiv b(s+1) - b(s)$.

Example 2.7 Consider an admission controlled $M/M/1/k$ queueing system with Poisson arrival rate λ and exponential service rate μ. Assume that a holding cost is accrued at rate $f(s)$ while there are s customers in the system. If admitted the job enters the queue and the decision-maker immediately receives reward R. Rejected customers are lost. Assume that the cost is convex and increasing in s and $f(0) = 0$. Furthermore, assume that we discretize the model by applying the standard uniformization technique of Lippman [13]. Without loss of generality let the uniformization constant $\lambda + \mu = 1$. Since rejecting all customers yields $w = 0$, we assume customers are accepted in state zero. This example was previously considered in Haviv and Puterman [5] where it was shown algebraically that bias distinguishes between gain optimal policies. For this model, the average optimality equations are

$$\begin{aligned} h(s) &= \max\{\lambda R - w - f(s) + \lambda h(s+1) + \mu h((s-1)^+), \\ &\qquad -w - f(s) + \lambda h(s) + \mu h((s-1)^+)\} \end{aligned} \tag{2.45}$$

Consider the set of policies T^∞ that accept customers until the number of customers in the system reaches some control limit $L > 0$ and rejects customers for all $s \geq L$. Denote the stationary policy that uses control limit L by L. It is known that there exists a Blackwell optimal policy within this set. The following lemma asserts the intuitive idea that it is better to start with fewer customers in the system. We will use this result in the sample path arguments to follow.

Lemma 2.1 *Suppose (w^*, h) satisfy the optimality equations. For $s \in S$, $\Delta h(s) < 0$.*

Haviv and Puterman [5] show that if there are two gain optimal control limits, only the higher one is bias optimal. Let L and $L+1$ be gain optimal control limits. Notice that since $L+1$ accepts customers in state L and control limit L rejects them, the two policies have different recurrent classes. The gain optimality equations are

$$h(s) = \begin{cases} \lambda R - g + \lambda h(s+1) + \mu h(s) & s = 0 \\ \lambda R - g - f(s) + \lambda h(s+1) + \mu h(s-1) & 0 < s \leq L \\ -g - f(s) + \lambda h(s) + \mu h(s-1) & s \geq L \end{cases} \tag{2.46}$$

At $s = L$ we have equality in the optimality equations since both L and $L+1$ are gain optimal. Furthermore, note that state L is recurrent for both policies. Let c_L be the bias constant for control limit L. Similarly for c_{L+1}. Choose $\alpha = L$ as the reference state. Suppose we follow the sample paths of two processes on the same probability space both starting in state L. Process 1 uses control limit L and process 2 uses control limit $L+1$. It is easy to see if the first event

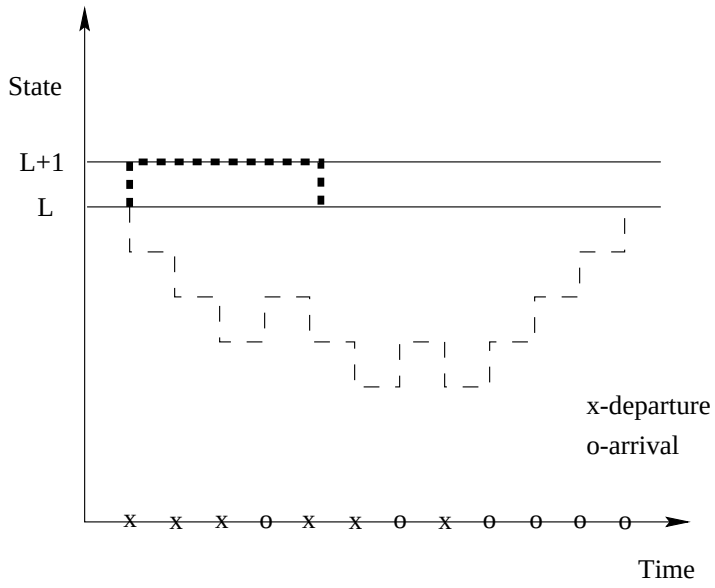

Figure 2.3 We would like to compare control limits L and $L+1$.

is a departure, both processes move to state $L-1$. Since the policies are the same on all states below L, the costs accrued (measured by $h^{rv(\alpha)}$) until the return to state L are the same. This is denoted by the lighter dashed line of Figure 2.5. Further, if the first event is an arrival, Process 1 rejects arriving customers and thus, immediately returns to L accruing cost $h^{rv(\alpha)}(L) = 0$ on the cycle. Process 2 accepts the arriving customer, accrues cost $h^{rv(\alpha)}(L)$, and moves to state $L+1$. The process then accrues cost $h^{rv(\alpha)}(L+1)$ a geometric number of times (with parameter μ) for false arrivals in the uniformization (recall $\lambda + \mu = 1$) before returning to state L. This is denoted by the bold line in Figure 2.5. Hence, while the total cost before returning to state L is the same for each policy when a departure is the first event, when an arrival occurs first, process 2 accrues $h^{rv(\alpha)}(L+1)$ for each extra decision epoch in the cycle. Let $p = \frac{\mu}{\lambda+\mu}$ be the probability that the first event is a departure, t be the total expected number of decision epochs on the cycle given that the first event is a departure, and let M be the total expected cost accrued on that

cycle. We have

$$\begin{aligned}
\frac{c_{L+1}}{c_L} &= \frac{\mathbb{E}_\alpha^L(\tau_\alpha)\left(\mathbb{E}_\alpha^{L+1}\sum_{n=0}^{\tau_\alpha-1}h^{rv(\alpha)}(x_n,\phi(x_n))\right)}{\mathbb{E}_\alpha^{L+1}(\tau_\alpha)\left(\mathbb{E}_\alpha^{L}\sum_{n=0}^{\tau_\alpha-1}h^{rv(\alpha)}(x_n,\phi(x_n))\right)} \\
&= \frac{\mathbb{E}_\alpha^L(\tau_\alpha)\left(pM+(1-p)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)\right)}{\mathbb{E}_\alpha^{L+1}(\tau_\alpha)\left(pM+(1-p)\cdot 0\right)} \\
&= \frac{(pt+(1-p))\left(pM+(1-p)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)\right)}{(pt+(1-p)(\frac{1}{\mu}+1))pM} \\
&= \frac{(pt+(1-p))\left(pM+(1-p)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)\right)}{(pt+(1-p))pM+(1-p)\frac{1}{\mu}pM}
\end{aligned}$$

Recall, that we have assumed that $\lambda+\mu=1$ so $p/\mu=1$. Thus,

$$\frac{c_{L+1}}{c_L} = \frac{\mathbb{E}_\alpha^L(\tau_\alpha)pM+(1-p)\,\mathbb{E}_\alpha^L(\tau_\alpha)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)}{\mathbb{E}_\alpha^L(\tau_\alpha)pM+(1-p)M} \tag{2.47}$$

Using (2.47) we need only compare $\mathbb{E}_\alpha^L(\tau_\alpha)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)$ and M. From Lemma 2.1, $h^{rv(\alpha)}(L+1) < h^{rv(\alpha)}(s)$ for all $s \le L$. Since M is the total expected cost given that the first event is a departure, we know M consists only of costs as measured by $h^{rv(\alpha)}(s)$ for $s \le L$. Hence, we have $\mathbb{E}_\alpha^L(\tau_\alpha)h^{rv(\alpha)}(L+1) < M$; each extra decision epoch in process 2 can only stand to **decrease** the average cost. That is to say, $c_{L+1} > c_L$, and the bias of control limit $L+1$ is larger than that of L. ∎

The previous example shows that by an astute choice of the reference state a simple sample path argument can be used to show the usefulness of bias in distinguishing between gain optimal policies. This analysis begs the question, why is the higher control limit preferred? In essence, the choice the decision-maker must make is whether to add more waiting space. If optimal gain is the primary objective, it is clear that if adding this server reduces the gain, it should not be added. On the other hand, if adding the waiting space, does not change the gain, but decreases the average cost as measured by $h^{rv(\alpha)}$ the decision-maker would prefer to add the space. The question of why the relative value functions measure cost remains open. However, we can make the observation that with L as the reference state (so $h^{rv(\alpha)}(L)=0$), Lemma 2.1 implies that $h^{rv(\alpha)}(s)<0$ for $s>L$ while $h^{rv(\alpha)}(s)>0$ for $s<L$. Thus, the average cost is decreased by time spent with more than L customers in the system. The bias-based decision-maker prefers negative relative value functions.

Suppose now we consider Example 2.7 except that rewards are received upon service completion instead of upon acceptance to the system. Using the bias optimality equation (2.34), the authors [12] showed that if there are two gain optimal control limits, it is in fact the **lower** control limit that is bias optimal. By formulating this problem as in the previous example it is not difficult to show that with L as the reference state, $h^{rv(\alpha)}(s)>0$ for $s>L$ while $h^{rv(\alpha)}(s)<0$

for $s < L$. Using precisely the same sample path argument as Example 2.7, we get that the lower control limit is bias optimal.

This leads us to two conclusions. First, the intuitive idea of the bias as the total reward is quite restrictive since in total reward models the decision-maker is indifferent to when rewards are received. And, secondly the interpretation of the bias as an average cost problem is not sufficient either, since discounting is usually lost in the long-run analysis. In fact, some of the properties of discounting are retained after the limit is taken in the definition of $0-$discount optimality (see (2.7)). This line of discussion is pursued next.

2.5.1 Bias and implicit discounting

Implicit discounting in bias allows us to explain why bias prefers control limit L or $L+1$. First note that the decision to accept or reject a customer in Example 2.7 for the long-run average reward based decision-maker is in essence based on whether or not the reward offered is higher than the cost of having the customer in the system. Thus, when the decision-maker is indifferent, the rewards and costs must balance. When rewards are received at arrivals the reward is received before the cost of having the customer in the system is accrued. On the other hand, when rewards are received upon service completion the decision-maker must accrue the cost of having a customer in the system before receiving the reward. The decision-maker only chooses to increase the amount of waiting space if the reward is received before the cost and the discounting is apparent. The following theorem explains how the bias-based decision-maker discounts future rewards.

Theorem 2.4 *Suppose that α is a positive recurrent state for a fixed policy $\phi \in \Pi^S$. Further suppose that $h_\phi^{rv(\alpha)}$ is the relative value function of ϕ with $h_\phi^{rv(\alpha)}(\alpha) = 0$. Let c_ϕ be the bias constant associated with $h_\phi^{rv(\alpha)}$. Then*

$$c_\phi = -\frac{\mathbb{E}_\alpha^\phi \sum_{n=0}^{\tau_\alpha - 1}(n+1)[r(x_n) - w_\phi]}{\mathbb{E}_\alpha^\phi \tau_\alpha} \tag{2.48}$$

So

$$h_\phi = h_\phi^{rv(\alpha)} - \frac{\mathbb{E}_\alpha^\phi \sum_{n=0}^{\tau_\alpha - 1}(n+1)[r(x_n) - w_\phi]}{\mathbb{E}_\alpha^\phi \tau_\alpha} \tag{2.49}$$

The bias-based decision-maker attempts to maximize (2.48). The factor "$n+1$" discounts the excess rewards received later in the cycle. Thus, if r exceeds w it is better if it occurs earlier in the cycle when it is multiplied by a smaller factor.

Example 2.8 Again return to Example 2.4 and compute the bias constant using (2.48).

$$\begin{aligned} c_\delta &= -\frac{\{[r(0)-w]+[r(1)-w]+[r(3)-w]\}}{3} \\ &\quad -\frac{\{[r(1)-w]+[r(3)-w]\}+\{[r(3)-w]\}}{3} \\ &= -\frac{\{[r(0)-w]+2[r(1)-w]+3[r(3)-w]\}}{3} \\ &= -\{1+2\cdot(-1)+3\cdot(0)\}/3 = 1/3. \end{aligned}$$

Similarly,

$$c_\psi = -(-1+2\cdot(1)+3\cdot(0))/3 = -1/3$$

Notice that when the excess reward is received earlier it is worth more (-1 compared to -2), and when the cost is received earlier, it is more costly than later (1 compared to $2\cdot 1$). Suppose we write "$d_1 \succ d_2$" if the bias-based decision-maker prefers d_1 to d_2. In this example, $\delta \succ \gamma$ since the decision-maker chooses to receive the immediate reward and accrue the later cost. Similarly, $\psi \succ \gamma$; the decision-maker prefers not to accrue the immediate cost, despite the fact that there is a reward to be received later. Finally, comparing ψ to δ yields the following relation $\delta \succ \psi \succ \gamma$.

Precisely the same logic can be applied to the prior queueing example. When the reward is received upon acceptance, the decision-maker is willing to accept the arriving customer and $L+1 \succ L$. On the other hand, when the reward is received upon service completion and therefore discounted, the decision-maker chooses not to accept the customer and the relation is reversed.

2.6 CONCLUSIONS AND FUTURE RESEARCH

When we began our study of bias we presented a sequence of questions. The relationship of bias to the total reward problem was considered on two levels. First, if there is but one recurrent state, on which the reward is zero, the gain is clearly zero, and the bias is equivalent to the total reward until entering the recurrent state. Suppose now that there is a single recurrent class, but there is more than one state in this class. One might immediately conjecture that we need only subtract the gain from each of the rewards on the transient states, and again compute the total reward until entering the recurrent class. While this leads to a function which satisfies the AEE, the relative value function, it does not equal the bias. The reason this is so, is that the bias includes implicit discounting. Hence, instead of simply computing total reward, we must consider when these rewards are received.

Computationally, we have shown that in the unichain case the bias can be computed by any of the methods used to compute the gain. By noticing that the form of the BOE is exactly the same as that of the AOE, we need not introduce any new methods for computation. This also leads to sample path methods which illuminate the fact that the bias based decision-maker prefers policies that spend more time in states with negative relative values on the recurrent class. The interpretation of this fact is open. However, since this also leads to implicit discounting we may shed a little light on the subject. Since the relative value function is zero on the chosen reference state, if a state entered

after leaving the reference state has a negative relative value, an equivalent, positive excess reward must have previously been received in order to balance the rewards and the gain on the cycle. On the other hand, if a state entered before returning to the reference state has positive relative value, a reward less than the gain must be earned prior to entering that state. This is the crucial point of our analysis of implicit discounting and implies that higher rewards received earlier in the cycle are preferred.

We have restricted ourselves to the unichain finite state and action space case. We feel that it is clear that there is a need to extend each of these ideas to multichain, countable, and general state space cases. It is also important to notice that the discounting bias captures, is only captured on the recurrent states. For discounting on the transient states, one would need to use the next term in the Laurent series expansion. Why this is so also remains unanswered.

References

[1] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM Journal on Control and Optimization* **31** pp. 282–344, 1993.

[2] A. G. Barto and R. S. Sutton, *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*, MIT Press, Cambridge, MA, 1998.

[3] D. Blackwell, "Discrete dynamic programming," *Annals of Mathematical Statistics* **33** pp. 719–726, 1962.

[4] E. V. Denardo, "Computing a bias optimal policy in a discrete-time Markov decision problem, *Operations Research* **18** pp. 279–289, 1970.

[5] M. Haviv and M. L. Puterman, "Bias optimality in controlled queueing systems," *Journal of Applied Probability* **35** pp. 136–150, 1998.

[6] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[7] O. Hernández-Lerma and J. B. Lasserre, "Policy iteration in average cost Markov control processes on Borel spaces," *Acta Applicandae Mathematicae* **47** pp. 125–154, 1997.

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

[9] R. A. Howard, *Dynamic Programming and Markov Processes,* John Wiley & Sons, New York, 1960.

[10] M. E. Lewis, H. Ayhan, and R. D. Foley, "Bias optimality in a queue with admission control," *Probability in the Engineering and Informational Sciences* **13** pp. 309–327, 1999.

[11] M. E. Lewis and M. L. Puterman, "A probabilistic analysis of bias optimality in unichain Markov decision processes," 2000, submitted.

[12] M. E. Lewis and M. L. Puterman, "A note on bias optimality in controlled queueing systems," *Journal of Applied Probability* **37** pp. 300–305, 2000.

[13] S. A. Lippman, "Applying a new device in the optimization of exponential queueing systems," *Operations Research* **23** pp. 687–712, 1975.

[14] E. Mann, "Optimality equations and sensitive optimality in bounded Markov decision processes," *Optimization* **16** pp. 767–781, 1985.

[15] S. Meyn, "The policy iteration algorithm for average reward Markov decision processes with general state space," *IEEE Transactions on Automatic Control* **42** pp. 1663–1680, 1997.

[16] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming,* John Wiley and Sons, New York, 1994.

[17] S. M. Ross, *Stochastic Processes*, John Wiley and Sons, New York, 1983.

[18] P. Schweitzer and A. Federgruen, "The functional equations of undiscounted Markov renewal programming," *Mathematics of Operations Research* **3** pp. 308–321, 1977.

[19] A. F. Veinott, "On finding optimal policies in discrete dynamic programming with no discounting," *Annals of Mathematical Statistics* **37** pp. 1284–1294, 1966.

[20] A. F. Veinott, Jr. "Discrete dynamic programming," *Annals of Mathematical Statistics* **40** pp. 1635–1660, 1969.

[21] A. F. Veinott, Jr. "Markov decision chains," in *Studies in Optimization*, volume 10 of *Studies in Mathematics* pp. 124–159. Mathematics Association of America, 1974.

Mark E. Lewis
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109-2117, USA
melewis@engin.umich.edu

Martin L. Puterman
Faculty of Commerce and Business Administration
University of British Columbia
Vancouver, BC Canada V6T 1Z2
marty@coe.ubc.ca

3 SINGULAR PERTURBATIONS OF MARKOV CHAINS AND DECISION PROCESSES

Konstantin E. Avrachenkov
Jerzy Filar*
Moshe Haviv

Abstract: In this survey we present a unified treatment of both singular and regular perturbations in finite Markov chains and decision processes. The treatment is based on the analysis of series expansions of various important entities such as the perturbed stationary distribution matrix, the deviation matrix, the mean-passage times matrix and others.

3.1 BACKGROUND AND MOTIVATION

Finite state Markov Chains (MC's) are among the most widely used probabilistic models of discrete event stochastic phenomena. Named after A.A. Markov, a famous Russian mathematician, they capture the essence of the existentialist "here and now" philosophy in the so-called "Markov property" which, roughly speaking, states that probability transitions to a subsequent state depend only on the current state and time. This property is less restrictive than might appear at first because there is a great deal of flexibility in the choice of what constitutes the "current state". Because of their ubiquitous nature Markov Chains are, nowadays, taught in many undergraduate and graduate courses ranging from mathematics, through engineering to business administration and finance.

*This research was supported in part by the ARC grant #A49906132.

Whereas a MC often forms a good description of some discrete event stochastic process, it is not automatically equipped with a capability to model such a process in the situation where there may be a "controller" or a "decision-maker" who—by a judicious choice of actions—can influence the trajectory of the process. This innovation was not introduced until the seminal works of Howard [44] and Blackwell [16] that are generally regarded as the starting point of the modern theory of Markov Decision Processes (or MDP's for short). Since then, MDP's have evolved rapidly to the point that there is now a fairly complete existence theory, and a number of good algorithms for computing optimal policies with respect to criteria such as maximization of limiting average expected reward, or the discounted expected reward.

The bulk of the, now vast, literature on both MCs and MDPs deals with the "perfect information" situations where all the model parameters—in particular probability transitions—are assumed to be known precisely. However, in most applications this assumption will be violated. For instance, a typical parameter, ρ, would normally be replaced by an estimate

$$\hat{\rho} = \rho + \varepsilon(n)$$

where the error term, $\varepsilon(n)$, comes from a statistical procedure used to estimate ρ and n is the number of observations used in that estimation. In most of the valid statistical procedures $|\varepsilon(n)| \downarrow 0$ as $n \uparrow \infty$, in an appropriate sense. Thus, from a perturbation analysis point of view, it is reasonable to suppress the argument n and simply concern ourselves with the effects of $\varepsilon \to 0$.

Roughly speaking, the subject of perturbation analysis of MC's and MDP's divides naturally into the study of "regular" and "singular" perturbations. Intuitively, regular perturbations are "good" in the sense that the effect of the perturbation dissipates harmlessly as $\varepsilon \to 0$, whereas singular perturbations are "bad" in the sense that small changes of ε (in a neighborhood of 0) can induce "large" effects. Mathematically, it can be shown that singular perturbations are associated with a change of the rank of a suitably selected matrix. This can be easily seen from the now classical example due to Schweitzer [66] where the perturbed probability transition matrix

$$P(\varepsilon) = \begin{pmatrix} 1-\frac{\varepsilon}{2} & \frac{\varepsilon}{2} \\ \frac{\varepsilon}{2} & 1-\frac{\varepsilon}{2} \end{pmatrix} \overset{\varepsilon \downarrow 0}{\longrightarrow} P(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

but the limit matrix for $\varepsilon > 0$

$$P^*(\varepsilon) \equiv \lim_{t\to\infty} P^t(\varepsilon) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \not\to P^*(0) \equiv \lim_{t\to\infty} P^t(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Indeed, the rank of $P^*(\varepsilon)$ is 1 for all $\varepsilon > 0$ (regardless of how close it is to zero), but it jumps to 2 at $\varepsilon = 0$; despite the fact that $P(\varepsilon) \to P(0)$. Thus, we see that singular perturbations can occur in MC's in a very natural and essential way. The latter point can be underscored by observing the behavior of $(I - \lambda P)^{-1}$ as $\lambda \to 1$, where P is a Markov (probability transition) matrix. It is well-known (e.g., see [16] or [57]) that this inverse can be expanded as a

Laurent series (with a pole of order 1) in the powers of $\varepsilon := 1-\lambda$. Indeed, much of the theory devoted to the connections between the discounted and limiting average MDP's exploits the asymptotic properties of this expansion, as $\varepsilon \downarrow 0$. Of course, the rank of $(I-\lambda P)$ changes when $\lambda = 1 (\varepsilon = 0)$.

It is not surprising, therefore, that the literature devoted to singularly perturbed MC's and MDP's has been growing steadily in recent years. In fact there have been quite a few developments since the 1995 survey by Abbad and Filar [2].

The purpose of this survey paper is to present an up to date outline of a unified treatment of both singular and regular perturbations in MC's and MDP's that is based on series expansions of various important entities such as the perturbed stationary distribution matrix, the deviation matrix, the mean-passage times matrix and the resolvent-like matrix $(I-\lambda P)^{-1}$. From this series expansion perspective, the regular perturbations are simply the cases where Laurent series reduce to power series. Consequently, the capability to characterize and/or compute the coefficients of these expansions and the order of the pole (if any, at $\varepsilon = 0$) becomes of paramount importance.

This survey covers only the results on *discrete time* MC's and MDP's. For a parallel development of the *continuous time* models we refer an interested reader to the comprehensive book of Yin and Zhang [76]. For related emerging results in game theory see Altman et al. [6].

The logical structure of the survey is as follows: in Section 2, perturbations of (uncontrolled) Markov chains are discussed from the series expansions perspective; in Section 3 the consequences of these results are discussed in the context of optimization problems arising naturally in the (controlled) MDP case; and, finally, in Section 4 applications of perturbed MDP's to the Hamiltonian Cycle Problem are outlined. This last section demonstrates that the theory of perturbed MDP's has applications outside of its own domain.

3.2 UNCONTROLLED PERTURBED MARKOV CHAINS

3.2.1 Introduction and preliminaries

Let $P \in R^{n\times n}$ be a transition stochastic matrix representing transition probabilities in a Markov chain. Suppose that the structure of the underlying Markov chain is aperiodic. Let $P^* = \lim_{t\to\infty} P^t$ which is well-known to exist for aperiodic processes. In the case when the process is also ergodic, P^* has identical rows, each of which is the stationary distribution of P, denoted by π. Let Y be the *deviation matrix* of P which is defined by $Y = (I - P + P^*)^{-1} - P^*$. It is well known (e.g., see [51]) that Y exists and it is the unique matrix satisfying

$$Y(I-P) = I - P^* = (I-P)Y \text{ and } P^*Y = 0 = Y\underline{1} \tag{3.1}$$

(where $\underline{1}$ is a matrix full of 1's) making it the group inverse of $I-P$. Finally, $Y = \lim_{T\to\infty} \Sigma_{t=0}^{T}(P^t - P^*)$. Let M_{ij} be the mean passage time from state-i into state-j. It is also known that when the corresponding random variable is proper, then M_{ij} is finite. Of course, the matrix M is well-defined if and only

if the Markov chain is ergodic. In this case, we have

$$M_{ij} = (\delta_{ij} + Y_{jj} - Y_{ij})/\pi_j \tag{3.2}$$

and, in particular, $M_{ii} = 1/\pi_i$. The above mentioned results can be found in many sources, for instance, see Meyer [56].

We consider (linear) perturbations of the matrix P and their impact on the structure of the process and on various essential matrices such as the stationary distribution matrix, the deviation matrix and the mean passage time matrix. Specifically, for a scalar ε, $0 < \varepsilon < \varepsilon_{\max}$, and for some zero rowsum matrix C, we look at the set of perturbed stochastic matrices $P(\varepsilon) = P + \varepsilon C$ which are assumed to be ergodic for any ε in the above mentioned region. Note that ergodicity is not assumed with regard to $P = P(0)$. Actually, the case where $P(0)$ contains some unrelated chains (with or without transient states) is our main focus. Our goal here is to survey the existing literature on series expansions for $\pi(\varepsilon)$, $P^*(\varepsilon)$, $Y(\varepsilon)$ and $M(\varepsilon)$, which denote the stationary distribution, the limit matrix, the deviation matrix and the mean passage time matrix, respectively, of $P(\varepsilon)$, for $0 < \varepsilon < \varepsilon_{\max}$ and consider their relationship to the corresponding entities in the unperturbed MC for $P = P(0)$.

The rest of this section is organized as follows. In the next subsection we discuss the regular case, namely the case in which the unperturbed system is ergodic. Then, in Subsection 2.3, we look at the *nearly completely decomposable* (NCD) case, namely the case in which the state space under the unperturbed process is decomposed into a number of ergodic classes. This number is assumed here to be at least two and no transient states are allowed. This assumption is removed in Subsection 2.4, where we allow the unperturbed system to have two or more ergodic classes plus a number of transient states. It will be seen that the presence of transient states induces some interesting phenomena. In Subsections 2.2 through 2.4, we assume that the perturbed process is ergodic. Subsection 2.5, we comment on some issues involving the removal of this assumption.

3.2.2 The regular case

In this subsection we assume that the unperturbed Markov chain is ergodic. This leads to the case of *regular perturbations*. The following results appear in a seminal paper by Schweitzer [65].

Theorem 3.1 *Assume that the unperturbed Markov chain is ergodic. Then,*

(i) The matrix functions $P^*(\varepsilon)$, $Y(\varepsilon)$ and $M(\varepsilon)$ are analytic in some (undeleted) neighborhood of zero. In particular, they all admit Maclaurin series expansions:

$$P^*(\varepsilon) = \sum_{m=0}^{\infty} \varepsilon^m P^{(*m)} \quad , \quad Y(\varepsilon) = \sum_{m=0}^{\infty} \varepsilon^m Y^{(m)} \quad \text{and} \quad M(\varepsilon) = \sum_{m=0}^{\infty} \varepsilon^m M^{(m)}$$

with some coefficient sequences $\{P^{(*m)}\}_{m=0}^{\infty}$, $\{Y^{(m)}\}_{m=0}^{\infty}$ and $\{M^{(m)}\}_{m=0}^{\infty}$.

(ii) The limit matrices $P^*(\varepsilon)$ and the deviation matrix of the perturbed Markov chain admit the following updating formulae

$$P^*(\varepsilon) = P^*(0)[I - \varepsilon U]^{-1} \tag{3.3}$$

and

$$\begin{aligned} Y(\varepsilon) &= [I - P^*(\varepsilon)]Y(0)[I - \varepsilon U]^{-1} \\ &= Y(0)[I - \varepsilon U]^{-1} - P^*(0)[I - \varepsilon U]^{-1}Y(0)[I - \varepsilon U]^{-1}, \end{aligned}$$

where $U := CY(0)$.

(iii) These updating formulae yield the following expressions for the power series coefficients.

$$P^{(*0)} = P^*(0), \quad Y^{(0)} = Y(0), \quad M^{(0)} = M(0)$$

$$P^{(*m)} = P^{(*0)}U^m, \quad m \geq 0$$

$$Y^{(m)} = Y(0)U^m - P^*(0)\sum_{j=1}^{m} U^j Y(0)U^{m-j}, \quad m \geq 0$$

$$M_{ij}^{(m)} = \frac{1}{\pi_j^{(0)}}(Y_{jj}^{(m)} - Y_{ij}^{(m)}) - \frac{1}{\pi_j^{(0)}}\sum_{l=1}^{m} \pi_j^{(l)} M_{ij}^{(m-l)}, \quad m \geq 0$$

(iv) The validity of any of the above series expansion holds for any ε, $0 \leq \varepsilon < \min\{\varepsilon_{\max}, \rho^{-1}(U)\}$ where $\rho(U)$ is the spectral radius of U.

The algebraic technique used to prove (3.3) contains the "flavor" of the required analysis. The following argument is based on [41]. It shows the validity of an equivalent statement: $\pi(\varepsilon) - \pi(0) = \varepsilon U(I - \varepsilon U)^{-1}$.

The latter follows from the observation that,

$$\pi(\varepsilon) - \pi(0) = \pi(\varepsilon)P(\varepsilon) - \pi(0)P(0) = \pi(\varepsilon)(P(0) + \varepsilon C) - \pi(0)P(0)$$

$$= (\pi(\varepsilon) - \pi(0))P(0) + \pi(\varepsilon)\varepsilon C$$

or

$$(\pi(\varepsilon) - \pi(0))(I - P(0)) = \pi(\varepsilon)\varepsilon C$$

Postmultiply the last equation by $Y(0)$ and use (3.1) in order to get that $(\pi(\varepsilon) - \pi(0))(I - P^*(0)) = \varepsilon\pi(\varepsilon)U$. But $(\pi(\varepsilon) - \pi(0))P^* = 0$ (as we multiply a zero sum vector by a matrix with identical rows). Hence, $\pi(\varepsilon) - \pi(0) = \varepsilon\pi(\varepsilon)U$. Replace $\pi(\varepsilon)$ in the rhs with $[\pi(\varepsilon) - \pi(0)] + \pi(0)$, move the product due to the term in brackets to the lhs and get that $(\pi(\varepsilon) - \pi(0))(I - \varepsilon U) = \varepsilon\pi(0)U$. Postmultiplication of both handsides with $(I - \varepsilon U)^{-1}$ completes the argument.

Example 1. For $0 \leq \varepsilon < .25$ let

$$P(\varepsilon) = P(0) + \varepsilon C = \begin{pmatrix} .5 & .5 \\ .5 & .5 \end{pmatrix} + \varepsilon \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix} \quad .$$

Clearly,

$$P^*(0) = P^{(*0)} = \begin{pmatrix} .5 & .5 \\ .5 & .5 \end{pmatrix}$$

Also,

$$Y(0) = Y^{(0)} = \begin{pmatrix} .5 & -.5 \\ -.5 & .5 \end{pmatrix} \quad \text{and} \quad M(0) = M^{(0)} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \quad .$$

Hence,

$$U = CY(0) = \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix} \quad .$$

It is easy to see that for $m \geq 1$, $U^m = 3^{m-1}U$ and hence for $m \geq 1$,

$$P^{(*m)} = P^*(0)U^m = 3^{m-1} \begin{pmatrix} .5 & -.5 \\ .5 & -.5 \end{pmatrix} \quad .$$

Also, for $m \geq 1$,

$$Y^{(m)} = 3^{m-1}Y(0)U - 3^{m-2}(m-1)P^*(0)UY(0)U - 3^{m-1}P^*(0)UY(0)$$

$$= 3^{m-1} \begin{pmatrix} 1.5 & -1.5 \\ -1.5 & 1.5 \end{pmatrix} - 3^{m-2}(m-1) \begin{pmatrix} 1.5 & -1.5 \\ 1.5 & -1.5 \end{pmatrix} - 3^{m-1} \begin{pmatrix} .5 & -.5 \\ .5 & -.5 \end{pmatrix} \quad .$$

Finally,

$$M_{12}(\varepsilon) = 2 + 8\varepsilon + ... \quad \text{and} \quad M_{21}(\varepsilon) = 2 + 4\varepsilon + ...$$

3.2.3 The nearly completely decomposable case

Let $P(0) \in R^{n \times n}$ be a stochastic matrix representing transition probabilities in a completely decomposable Markov chain. By the latter we mean that there exists a partition Ω of the state space into p, $p \geq 2$, subsets $\Omega = \{I_1, \ldots, I_p\}$ each of which being an ergodic class. We assume that the order of the rows and of the columns of P is compatible with Ω, i.e., for p stochastic matrices, $P_{I_1}, \ldots, P_{I_p}$,

$$P(0) = \begin{pmatrix} P_{I_1} & 0 & \cdots & 0 \\ 0 & P_{I_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & P_{I_p} \end{pmatrix}$$

Note that we assume above that none of the states is transient. Let $C \in R^{n \times n}$ be a zero rowsum matrix such that for some $\varepsilon_{\max} > 0$, the matrix $P + \varepsilon C$ is stochastic for $\varepsilon \in (0, \varepsilon_{\max})$ representing transition probabilities in an ergodic Markov chain. For small values of ε, $P(\varepsilon)$ is called *nearly completely decomposable* (NCD) or sometimes *nearly uncoupled*. Clearly, $C_{ij} \geq 0$ for any pair of states i and j belonging to different subsets.

Probably, the first motivation to study the singular perturbed Markov chains was given by Simon and Ando [70]. They demonstrated that several problems in econometrics lead to the mathematical model based on singularly perturbed Markov chains. The first rigorous theoretical developments of the singularly perturbed Markov chains have been carried out by Pervozvanski and

Smirnov [59] and Gaitsgori and Pervozvanskii [32]. In particular, they have shown that the *limiting* probability distribution $\pi^{(0)} = \lim_{\varepsilon \to 0} \pi(\varepsilon)$ can be expressed in terms of the invariant probability distributions of the ergodic classes $I_k, k = 1, ..., p$ and of the stationary distribution of an *aggregated* chain. Similar ideas were also developed in works of Courtois and his co-authors [20, 21, 22], Haviv and his co-authors [35, 36, 37, 38, 39, 40, 41] and by others [67, 69, 71, 73]. These works contain a number of iterative procedures for computing the stationary distribution, each of which having an aggregation step and a disaggregation step. In particular, in [38] it is shown that an error reduction of order of $O(\varepsilon)$ is achieved in each iteration of such procedures.

Schweitzer [66] showed that $\pi(\varepsilon)$ is analytic in some deleted neighborhood of zero. That is, $\pi(\varepsilon) = \sum_{m=0}^{\infty} \pi^{(m)} \varepsilon^m$, where $\pi^{(0)} = \lim_{\varepsilon \to 0} \pi(\varepsilon)$ and where $\pi^{(m)}$, $m \geq 1$, are zerosum vectors. Note that $\pi_i^{(0)} > 0$ for all i. Moreover, $\{\pi^{(m)}\}_{m=0}^{\infty}$ is a geometric series, that is, for some matrix $U \in R^{n \times n}$, $\pi^{(m)} = \pi^{(0)} U^m$, $0 \leq m < \infty$. See (3.7) below for an explicit expression for U. Finally, the series expansion holds for $0 < \varepsilon < \max\{\varepsilon_{\max}, \rho^{-1}(U)\}$.

For any subset $I \in \Omega$, let

$$k_I = \Sigma_{i \in I} \pi_i^{(0)} \tag{3.4}$$

Note that $k_I > 0$ for any $I \in \Omega$. Also, let γ_I be the subvector of $\pi^{(0)}$ corresponding to subset I rescaled so as its entry-sum is now one. Then, γ_I is the unique stationary distribution of P_I. Note that computing γ_I is relatively easy as only the knowledge of P_I is needed.

Next define the matrix $Q \in R^{p \times p}$ which is usually referred to as the *aggregate* transition matrix. Each row, and likewise each column in Q corresponds to a subset in Ω. Then, for subsets I and J, $I \neq J$, let

$$Q_{IJ} = \sum_{i \in I} (\gamma_I)_i \sum_{j \in J} C_{ij} \tag{3.5}$$

and let

$$Q_{II} = 1 + \sum_{i \in I} (\gamma_I)_i \sum_{j \in I} C_{ij} = 1 - \sum_{J \neq I} Q_{IJ} \quad . \tag{3.6}$$

Without loss of generality, assume that Q_{II} is non-negative for all subsets I and hence Q is easily seen to be a stochastic matrix.[1] Moreover, Q is irreducible and the vector $k \in R^p$ (see (3.4)) is easily checked to be its unique stationary distribution. Often it is convenient to express the aggregated transition matrix Q in matrix terms. Specifically, let $V \in R^{p \times n}$ be such that its i-th row is full

[1] Note that the matrix C can be divided by any constant and ε can be multiplied by this constant leading to the same $n \times n$ transition matrices. Taking this constant small enough guarantees the stochasticity of Q and hence this is assumed without loss of generality. In particular, the stationary distribution of Q is invariant with respect to the choice of this constant. Alternatively, one can define Q_{II} by $-\Sigma_{J \neq I} Q_{IJ}$ and consider Q as the generator of the aggregated process, i.e., the process among subsets (and hence no need to assume anything further with regard to the size of the entries of the matrix C.)

of zeros except for γ_{I_i} at the entries corresponding to subset I_i, and where $W \in R^{n\times p}$ is such that its j-th column is full of zeros except for 1's in the entries corresponding to subset I_j.[2] Then, we can write

$$Q = I + VCW.$$

The aggregate stochastic matrix Q represents transition probabilities between subsets which in this context are sometimes referred to as *macro-states*. However, although the original process among states is Markovian, this is not necessarily the case with the process among macro-states (and indeed typically it is not).[3] Yet, as the following indicates, much can be learned on the original process from the analysis of the aggregate matrix.

Theorem 3.2 *Let the perturbed Markov chain be nearly completely decomposable. The stationary distribution $\pi(\varepsilon)$ admits a Maclaurin series expansion in a deleted neighborhood of zero. Specifically, for some vectors $\{\pi^{(m)}\}_{m=0}^{\infty}$ with $\pi^{(0)}$ being a probability vector positive in all its entries and satisfying $\pi^{(0)} = \pi^{(0)}P(0)$, and for some zerosum vectors $\pi^{(m)}$, $m \geq 1$, $\pi(\varepsilon) = \Sigma_{m=0}^{\infty}\pi^{(m)}\varepsilon^m$. Moreover, for $I \in \Omega$, $\pi_I^{(0)} = k_I\gamma_I$.[4] Also, the series $\{\pi^{(m)}\}_{m=0}^{\infty}$ is geometric, i.e., for some square matrix U, $\pi^{(m)} = \pi^{(0)}U^m$ for any $m \geq 0$. Actually,*

$$U = CY(0)(I + CWDV), \tag{3.7}$$

where D is the deviation matrix of the aggregated transition matrix Q. Alternatively,

$$U = CY^{(0)} \tag{3.8}$$

where $Y^{(0)}$ is the first regular term of the Laurent series expansion for $Y(\varepsilon)$ (see also the next theorem). Finally, the validity of the series expansion holds for any ε, $0 \leq \varepsilon < \min\{\varepsilon_{\max}, \rho^{-1}(U)\}$ where $\rho(U)$ is the spectral radius of U.

We note that a series expression for $\pi(\varepsilon)$ appeared originally in [66]. The expression for U in (3.7) taken from [39] sheds more light on the role of the aggregate process played in the original process than its original expression given in [66] (see Equation 3-1 there). Also we note that $U = CY^{(0)}$ is an elegant generalization [68] of the regular case where $U = CY(0)$.

For ε, $0 \leq \varepsilon < \varepsilon_{\max}$, let $Y(\varepsilon)$ be the deviation matrix of $P(\varepsilon)$. This matrix is uniquely defined and the case $\varepsilon = 0$ is no exception. Yet, as we see shortly, there is no continuity of $Y(\varepsilon)$ at $\varepsilon = 0$. In particular, $Y(0)$ has the same shape

[2] Note that $VW \in R^{p\times p}$ is the identity matrix. Moreover, V and W correspond to orthonormal sets of eigenvectors of $P(0)$ belonging to the eigenvalue 1, V as left eigenvectors and W as right eigenvectors.

[3] The process among macro-states is an example of a partially observable Markov process.

[4] Recall that γ_I is the stationary distribution of P_I and that k in the stationary distribution of the aggregated matrix Q as defined in (3.5) and (3.6).

as P has, namely

$$Y(0) = \begin{pmatrix} Y_{I_1} & 0 & \cdots & 0 \\ 0 & Y_{I_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Y_{I_p} \end{pmatrix} \tag{3.9}$$

where Y_{I_i} is the deviation matrix of P_{I_i}, $1 \leq i \leq p$.

Theorem 3.3 *In the case of NCD Markov chains, the matrix $Y(\varepsilon)$ admits a Laurent series expansion in a deleted neighborhood of zero with the order of the pole being exactly one. Specifically, for some matrices $\{Y^{(m)}\}_{m=-1}^{\infty}$ with $Y^{(-1)} \neq 0$,*

$$Y(\varepsilon) = \frac{1}{\varepsilon} Y^{(-1)} + Y^{(0)} + \varepsilon Y^{(1)} + \varepsilon^2 Y^{(2)} + \cdots \tag{3.10}$$

for $0 < \varepsilon < \varepsilon_{\max}$. Moreover,

$$Y^{(-1)} = WDV,$$

or in a component form,

$$Y_{ij}^{(-1)} = D_{IJ}(\gamma_J)_j \quad , \quad i \in I \ , \ j \in J. \tag{3.11}$$

A series expression for $Y(\varepsilon)$ was originally developed in [66]. Yet, the expression for $Y^{(-1)}$ given above in (3.11) is taken from [39]. We find (3.11) appealing as it explicitly shows the role played by the aggregate matrix (and hence by the process among macro-states) in the original process. This point was already mentioned in [37] (see Equation 3 there).

We now focus our attention on $M(\varepsilon)$. Note that as opposed to $Y(0)$, $M(0)$ is not well-defined as the corresponding mean value (when $\varepsilon = 0$ and states i and j belong to two different subsets) does not exist. Let $E \in R^{p \times p}$ be the mean passage time matrix associated with the aggregated process. That is, for any pair of subsets I and J ($I = J$ included), E_{IJ} is the mean passage time from the macro-state I into the macro-state J when transition probabilities are governed by the stochastic matrix Q.

The following theorem will appear in [11].

Theorem 3.4 *The matrix $M(\varepsilon)$ admits a Laurent series expansion in a deleted neighborhood of zero with the order of the pole being exactly one. Specifically, for some matrices $\{M^{(m)}\}_{m=-1}^{\infty}$ with $M^{(-1)} \neq 0$,*

$$M(\varepsilon) = \frac{1}{\varepsilon} M^{(-1)} + M^{(0)} + \varepsilon M^{(1)} + \varepsilon^2 M^{(2)} + \cdots \tag{3.12}$$

for $0 < \varepsilon < \varepsilon_{\max}$. Moreover, for $i \in I$ and $j \in J$,

$$M_{ij}^{(-1)} = \begin{cases} 0 & if \ J = I \\ E_{IJ} & if J \neq I \end{cases} \tag{3.13}$$

$$M_{ij}^{(m)} = \frac{1}{\pi_j^{(0)}}(Y_{jj}^{(m)} - Y_{ij}^{(m)}) - \frac{1}{\pi_j^{(0)}} \sum_{l=1}^{m+1} \pi_j^{(l)} M_{ij}^{(m-l)} \quad , \quad m \geq -1$$

Proof. (Included since [11] has not yet appeared.) From (3.2) coupled with the fact that the Markov chain is ergodic when $0 < \varepsilon < \varepsilon_{\max}$,

$$M_{ij}(\varepsilon) = \frac{\delta_{ij} + Y_{jj}(\varepsilon) - Y_{ij}(\varepsilon)}{\pi_j(\varepsilon)} \quad , \quad 0 < \varepsilon < \varepsilon_{\max} \tag{3.14}$$

Hence, by (3.10)

$$M_{ij}^{(-1)} = \frac{Y_{jj}^{(-1)} - Y_{ij}^{(-1)}}{\pi_j^{(0)}} \quad . \tag{3.15}$$

By (3.11), $Y_{jj}^{(-1)} = Y_{ij}^{(-1)}$ whenever states i and j are in the same subset, hence $M_{ij}^{(-1)} = 0$ in this case. Using (3.11) again for the case where $J \neq I$, (3.15) is with a numerator which equals $(D_{JJ} - D_{IJ})(\gamma_J)_j$. By (3.4) and the definition of γ_J, the denominator equals $k_J(\gamma_J)_j$. Thus for this case, (3.15) equals $(D_{JJ} - D_{JI})/k_J$. Apply now (3.2) to the matrix Q and conclude that

$$M_{ij}^{(-1)} = \frac{D_{JJ} - D_{JI}}{k_J} = E_{IJ}$$

whenever $i \in I$, $j \in J$ and $J \neq I$. ∎

The following example is taken from [68].

Example 2 Let

$$P(0) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & .5 & .5 \\ 0 & .5 & .5 \end{pmatrix} \quad \text{and} \quad C = \frac{2}{7} \begin{pmatrix} -2 & 1 & 1 \\ 3 & -1 & -2 \\ 4 & -3 & -1 \end{pmatrix}.$$

The number of subset equals 2 with $\gamma_{I_1} = 1$ and $\gamma_{I_2} = (0.5, 0.5)$. First, we construct the following matrices.

$$V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix} \quad W = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad Y(0) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.5 & -0.5 \\ 0 & -0.5 & 0.5 \end{pmatrix}$$

The aggregated transition matrix is given by

$$Q = I + VCW = \begin{pmatrix} 3/7 & 4/7 \\ 1 & 0 \end{pmatrix} \quad .$$

Hence, $k = (7/11, 4/11)$ and $\pi^{(0)} = (7/11, 2/11, 2/11)$. Next, we calculate D, the deviation matrix of Q,

$$D = (I - Q + Q^*)^{-1} - Q^* = \frac{1}{121} \begin{pmatrix} 28 & -28 \\ -49 & 49 \end{pmatrix}$$

and hence, using (3.11),

$$Y^{(-1)} = WDV = \frac{1}{242}\begin{pmatrix} 56 & -28 & -28 \\ -98 & 49 & 49 \\ -98 & 49 & 49 \end{pmatrix}$$

The matrix E, which is the mean passage time matrix for the aggregated process, equals

$$E = \begin{pmatrix} 11/7 & 7/4 \\ 1 & 11/4 \end{pmatrix}$$

and hence, using (3.13),

$$M^{(-1)} = \begin{pmatrix} 0 & 7/4 & 7/4 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Finally, from (3.7) we get that

$$U = CY(0)(I + CWDV) = \frac{1}{77}\begin{pmatrix} 0 & 0 & 0 \\ -2 & 12 & -10 \\ 4 & -24 & 20 \end{pmatrix}$$

3.2.4 The general case

In this section we impose no assumptions on the ergodic structure of the unperturbed chain. That is, the unperturbed chain may now consist of several ergodic classes plus a set of transient states. Probably, Delebecque and Quadrat [23] were the first who studied the model in this general setting. However, they have analyzed only the case of the first order singularity, namely the case where the Laurent series expansion for the perturbed deviation matrix has a simple pole. This analysis is not much different from the analysis of the NCD case. Later Delebecque [24], using the reduction technique of Kato [48], removed the first order singularity assumption and showed how to compute the asymptotic expansion for $\pi(\varepsilon)$ in the general case. However, the reduction technique of [48] does not provide an efficient computational scheme. Other related papers devoted to the case where the unperturbed chain has a transient set are [18, 19, 15, 52, 63].

The following typical example of the general situation was considered extensively in the past (see [39, 34, 40, 13]):

Example 3.

$$P(\varepsilon) = P(0) + \varepsilon C = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \varepsilon \begin{pmatrix} 0 & -1 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} .$$

In this example the unperturbed chain contains two ergodic classes (states 2 and 4) and two transient states (states 1 and 3). They all are coupled to a single chain when $\varepsilon > 0$. Moreover, states 2 and 4 (i.e., the ergodic chains

in the unperturbed process) communicate under the perturbed case only via states 1 and 3 (i.e., transient states in the unperturbed case). This phenomenon makes this case more involved than the NCD case. For example, the reader is invited to check that the order of magnitude of $M_{24}(\varepsilon)$ is ε^{-2}.

Theorem 3.5

(i) The stationary distribution $\pi(\varepsilon)$ admits a Maclaurin series expansion in a deleted neighborhood of zero, that is, for some sequence $\{\pi^{(m)}\}_{m=0}^{\infty}$ with $\pi^{(0)}$ being a probability vector and $\pi^{(m)}$ for $m \geq 1$, being zero sum vectors,

$$\pi(\varepsilon) = \sum_{m=0}^{\infty} \varepsilon^m \pi^{(m)} \ .$$

Moreover, $\pi^{(0)} P(0) = \pi^{(0)}$ and for some matrix U,

$$\pi^{(m)} = \pi^{(0)} U^m \quad . \tag{3.16}$$

(ii) The deviation matrix $Y(\varepsilon)$ admits a Laurent series expansion in a deleted neighborhood of zero with a non-essential pole, that is, for some non-negative integer s and for some sequence of matrices $\{Y^{(m)}\}_{m=-s}^{\infty}$ with $Y^{(-s)} \neq 0$,

$$Y(\varepsilon) = \frac{Y^{(-s)}}{\varepsilon^s} + \frac{Y^{(-s+1)}}{\varepsilon^{s-1}} + \cdots + Y^{(0)} + Y^{(1)}\varepsilon + \cdots \ .$$

Moreover, the regular part $Y^R(\varepsilon) = Y^{(0)} + \varepsilon Y^{(1)} + \cdots$ can be expressed by the updating formula

$$Y^R(\varepsilon) = [I - P^*(\varepsilon)]Y^{(0)}[I - \varepsilon U]^{-1} - P^*(\varepsilon) \sum_{i=1}^{s} U^i Y^{(-i)} \quad . \tag{3.17}$$

The latter provides a computationally efficient recursive formula for the coefficients of the regular part:

$$Y^{(m)} = Y^{(0)} U^m - P_0^* \sum_{j=0}^{m} U^j Y^{(0)} U^{m-j} + P^{*(0)} U^m [I - \sum_{i=1}^{s} U^i Y^{(-i)}], \quad m \geq 1. \tag{3.18}$$

Furthermore, the quotient matrix U can be expressed as

$$U = CY^{(0)} \tag{3.19}$$

(iii) The mean passage time matrix $M(\varepsilon)$ admits a Laurent series expansion in a deleted neighborhood of zero with a non-essential pole, that is, for some

nonnegative integer t and for some sequence of matrices $\{M^{(m)}\}_{m=-t}^{\infty}$ with $M^{(-t)} \neq 0$,

$$M(\varepsilon) = \frac{M^{(-t)}}{\varepsilon^t} + \frac{M^{(-t+1)}}{\varepsilon^{t-1}} + \cdots + M^{(0)} + M^{(1)}\varepsilon + \cdots \quad .$$

Finally, $t \geq s$.

We would like to emphasize that once one has found the value of s and computed the terms $Y^{(-s)}, ..., Y^{(0)}$ and $\pi^{(0)}$, all the other terms in the series expansion for $Y(\varepsilon)$ and $\pi(\varepsilon)$ can be computed by (3.18). We refer below to a procedure for computing s. The term $\pi^{(0)}$ (which we believe to be the most important coefficient of the expansion) can be computed either by a reduction process [24] or via the determination of the null space of an auxiliary augmented matrix. The latter method will be described in some detail below. Finally, one can find methods for computing $Y^{(-s)}, ..., Y^{(0)}$ (and hence U) (once s and $\pi^{(0)}$ are in hand) by methods outlined in [13] and [39].

Remark 3.1 *Note that the formulae (3.16) and (3.17) are the natural generalization of the regular case. Formula (3.16) and (3.19) appeared originally in [68]. Finally, (3.17) was derived in [13].*

The next theorem shows that the order of poles of $Y(\varepsilon)$ and $M(\varepsilon)$ (denoted above by s and t, respectively), can be computed by a combinatorial algorithm.

Theorem 3.6 *For $1 \leq i, j \leq n$, let u_{ij} be the order of the pole of $M_{ij}(\varepsilon)$ at zero. Then, there exists an $O(n^4)$ algorithm for computing these orders whose input is the addresses of the nonzero entries in $P(0)$ and in C. Likewise, for $1 \leq i, j \leq n$, let v_{ij} be the order of the pole of $Y_{ij}(\varepsilon)$ at zero. Then,*

$$t = \max_{ij} u_{ij} \geq \max_{ij}(u_{ij} - u_{jj}) = \max_{ij} v_{ij} = s \quad . \tag{3.20}$$

The above theorem and a detailed algorithm are given in [34]. Finally, note that as $1/M_{ii}(\varepsilon) = \pi_i(\varepsilon)$, then u_{ii} is also the power of the first non-zero coefficient in the series expansion for $\pi_i(\varepsilon)$.

Example 3 (cont.). Running the algorithm developed in [34] for computing u_{ij}, $1 \leq i, j \leq 4$, results in

$$(u_{ij}) = \begin{pmatrix} 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 2 \\ 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 0 \end{pmatrix}$$

In particular, in this case $t = s = 2$.

For $m \geq 0$, let $B_m \in R^{n(m+1) \times n(m+1)}$ be the matrix such that each of its $n \times n$ block diagonal matrices equals $I - P(0)$ and each of its $n \times n$ matrices

above the diagonal equals $-C$. Namely,

$$B_m = \begin{pmatrix} I-P(0) & -C & 0 & \cdots & 0 & 0 \\ 0 & I-P(0) & -C & \cdots & 0 & 0 \\ 0 & 0 & I-P(0) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I-P(0) & -C \\ 0 & 0 & 0 & \cdots & 0 & I-P(0) \end{pmatrix} .$$

where $I-P(0)$ appears $m+1$ times while $-C$ appears m times. By equating coefficients of the same order in the identity $\pi(\varepsilon) = \pi(\varepsilon)(I-P(0)-\varepsilon C)$, it is easy to conclude that for any $m \geq 0$, $(\pi^{(0)}, \pi^{(1)}, \cdots, \pi^{(m)})$ is a left eigenvector of B_m belonging to the eigenvalue zero. Of course, it is not a unique eigenvector. Also, the dimension of this eigenspace of B_m is non-increasing in m. Let V_m be the subspace of vectors of length n such that the first n entires of vectors in the above mentioned eigenspace of B_m induce. Again, the dimension of V_m is non-increasing in m. Finally, these dimensions are always greater than or equal to 1. The following theorem is taken from [40].

Theorem 3.7 *(The general case: computing $\pi^{(0)}$).*

$$s+1 = min\{m;\ dim V_m = 1\} \quad .$$

The above two theorems suggest a way to compute $\pi^{(0)}$. Specifically, one can first find the value of s by the combinatorial algorithm suggested in [34] coupled with (3.20). Then the left eigenspace of B_s belonging to the eigenvalue zero can be explored. For the latter task some procedures exist. A few of them are based on *reduction* steps [48]. In [39] one reduction step is carried out while [24] and [13] suggest performing a number of reduction steps. The latter number is bounded by s as at the s-th step one gets a non-singular system. Actually, the procedure was defined in the previous subsection and takes care of NCD matrices with one (and terminal) reduction step.

We again note that once $\pi^{(0)}$ and $Y^{(0)}$ are calculated, the other coefficients of the series expansion for $\pi(\varepsilon)$ can be computed by the recursive formula $\pi^{(m)} = \pi^{(m-1)}U, m \geq 1$, where $U = CY^{(0)}$.

Example 3 (cont.). As we have already pointed out above, $s = 2$. Hence, solving for the left null space of

$$B_2 = \left(\begin{array}{c|c|c} I-P(0) & -C & 0 \\ 0 & I-P(0) & -C \\ 0 & 0 & I-P(0) \end{array}\right)$$

$$= \left(\begin{array}{cccc|cccc|cccc} 1 & -1 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

results in $(0, \alpha, 0, \alpha, *, *, *, *, *, *, *, *)$ for any α.[5] Hence, $\pi^{(0)} = (0, .5, 0, .5)$. In [39] it is shown that the same set of equations which is needed in order to solve for $\pi^{(0)}$ is also all required in order to compute U and $Y^{(-s)}$. For this numerical example, using the methods of either [39] or [13], one obtains

$$Y^{(0)} = \begin{pmatrix} .5 & -.5 & .5 & -.5 \\ 0 & 0 & 0 & 0 \\ .5 & -.5 & .5 & -.5 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and hence

$$U = CY^{(0)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ 0 & 0 & 0 & 0 \\ 0.5 & -0.5 & 0.5 & -0.5 \end{pmatrix}$$

The terms of the singular part of the Laurent expansion for the deviation matrix are:

$$Y^{(-2)} = \begin{pmatrix} 0 & .25 & 0 & -.25 \\ 0 & .25 & 0 & -.25 \\ 0 & -.25 & 0 & .25 \\ 0 & -.25 & 0 & .25 \end{pmatrix} \text{ and } Y^{(-1)} = \begin{pmatrix} .25 & -.5 & -.25 & .5 \\ .25 & 0 & -.25 & 0 \\ -.25 & .5 & .25 & -.5 \\ -.25 & 0 & .25 & 0 \end{pmatrix}$$

In particular,

$$(v_{ij}) = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}$$

[5] The sign '*' corresponds to values not computed (and not relevant for our current purposes.)

Finally,

$$M_{21}(\varepsilon) = \frac{Y_{11}(\varepsilon) - Y_{21}(\varepsilon)}{\pi_1(\varepsilon)} \tag{3.21}$$

$$= [(\frac{0.25}{\varepsilon} + 0.5 + ...) - (\frac{0.25}{\varepsilon} + 0 + ...)]/[0.5\varepsilon - 0.5\varepsilon^2 + ...]$$

$$= \frac{1}{\varepsilon} + ..., \tag{3.22}$$

$$M_{31}(\varepsilon) = \frac{1}{\varepsilon^2} + \frac{1}{\varepsilon} + ..., \quad M_{41}(\varepsilon) = \frac{1}{\varepsilon^2} + \frac{2}{\varepsilon} + ...,$$

$$M_{12}(\varepsilon) = \frac{1}{\varepsilon} + ..., \quad M_{32}(\varepsilon) = \frac{1}{\varepsilon^2} + \frac{0}{\varepsilon} + ..., \quad M_{42}(\varepsilon) = \frac{1}{\varepsilon^2} + \frac{1}{\varepsilon} + ... \ .$$

We like to point out that the expression for $Y^{(-1)}$ for this example given in [39] had a numerical error that was corrected in [13]. An open question left here is to find an expression similar to (3.13) (which suits only the NCD case) also for the general case.

The last lesson from the above example is that (3.20) cannot be refined and the conjecture that $u_{ij} - u_{jj} = v_{ij}$ is false even for the cases when $i \neq j$ (let alone when $i = j$). For example, $u_{21} - u_{11} = 0$ but $v_{21} = 1$.

3.2.5 The case of non ergodic perturbed Markov chains

In this section we briefly discuss the singularity phenomena that can appear if the perturbed Markov chain itself is not ergodic. The reader can find a more detailed exposition in [10].

In case where the perturbed Markov chain has several ergodic classes plus a non- empty set of transient states, then the perturbed transition matrix can be written in the following canonical form

$$P(\varepsilon) = \begin{bmatrix} P_1(\varepsilon) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & P_\ell(\varepsilon) & 0 \\ R_1(\varepsilon) & \cdots & R_\ell(\varepsilon) & S(\varepsilon) \end{bmatrix} \begin{matrix} \} \ \Omega_1 \\ \vdots \\ \} \ \Omega_\ell \\ \} \ \Omega_T \end{matrix} \tag{3.23}$$

where $P_i(\varepsilon) = P_i(0) + \varepsilon C_i$, $R_i(\varepsilon) = R_i(0) + \varepsilon C_{Ri}$, $S(\varepsilon) = S(0) + \varepsilon C_S$. Now note that all invariant measures $m_i(\varepsilon)$ of the perturbed Markov chain can be immediately constructed from the invariant measures of the ergodic classes associated with stochastic matrices $P_i(\varepsilon), i = 1, ..., \ell$. Namely, $m_i(\varepsilon) = [0 \cdots 0 \ \pi_i(\varepsilon) \ 0 \cdots 0]$, where $\pi_i(\varepsilon)$ is uniquely determined by the system

$$\begin{cases} \pi_i(\varepsilon) P_i(\varepsilon) = \pi_i(\varepsilon), \\ \pi_i(\varepsilon) \underline{1} = 1. \end{cases}$$

This is exactly the perturbation problem under the assumption that there are several ergodic classes plus possibly a set of transient states coupled by the perturbation into a single ergodic class. This issue was covered in the previous subsections.

The challenging task that we face in this case of multi-chain perturbed process is the computation of power series for the *right* eigenvectors of the perturbed Markov chains. The perturbed right eigenvectors contain the probabilities of being absorbed in each of the ergodic classes under the perturbed processes. These quantities exhibit the multi-chain ergodic properties of the perturbed process. For instance, nothing conceptually new would emerge had the set of transient states in the perturbed process been empty.

The right 0-eigenvectors of the perturbed chain can be written in the following form [51]

$$q_i(\varepsilon) = \begin{bmatrix} 0 \\ \underline{1} \\ 0 \\ \varphi_i(\varepsilon) \end{bmatrix} \begin{matrix} \} \ \Omega_i \\ \\ \} \ \Omega_T \end{matrix} \tag{3.24}$$

where subvector $\varphi_i(\varepsilon)$ is given by

$$\varphi_i(\varepsilon) = (I - S(\varepsilon))^{-1}(\varepsilon) R_i(\varepsilon)\underline{1}, \tag{3.25}$$

Note that if some ergodic classes become transient after the perturbation, then the matrix valued function $(I - S(\varepsilon))^{-1}$ has a singularity at $\varepsilon = 0$. To further elaborate on this phenomenon, let us consider the structure of the substochastic matrix $S(0)$.

$$S(0) = \begin{bmatrix} \tilde{P}_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \tilde{P}_m & 0 \\ \tilde{R}_1 & \cdots & \tilde{R}_m & \tilde{S} \end{bmatrix}.$$

Blocks $\tilde{P}_1, ..., \tilde{P}_m$ represent the ergodic classes of the original Markov chain which merge with the transient set after the perturbation. Of course, $m = 0$ is possible. Since each of $\tilde{P}_1, ..., \tilde{P}_m$ is a stochastic matrix, we conclude that matrix $I - S(0)$ has zero as an eigenvalue with multiplicity of at least m. Of course, $I - S(0)$ is not invertible. However, the matrix $(I - S(\varepsilon))^{-1}$ exists for small positive (but not zero) values of ε. From the results of [12], it follows that one can expand $(I - S(\varepsilon))^{-1}$ as a Laurent series at $\varepsilon = 0$

$$(I - S(\varepsilon))^{-1} = \frac{1}{\varepsilon^p} U^{(-p)} + ... + \frac{1}{\varepsilon} U^{(-1)} + U^{(0)} + \varepsilon U^{(1)} + ... \tag{3.26}$$

One can use the methods of [12] to calculate the coefficients of the above series. Substituting the expression $R_i(\varepsilon) = R_i(0) + \varepsilon C_{Ri}$ and the Laurent series (3.26) into formula (3.25), we obtain the asymptotic expansion

$$\varphi_i(\varepsilon) = \varphi_i^{(0)} + \varepsilon \varphi_i^{(1)} + \varepsilon^2 \varphi_i^{(2)} + ... \ , \tag{3.27}$$

where

$$\varphi_i^{(m)} = U^{(m)} R_i^{(0)} + U^{(m-1)} C_{Ri}, \quad m \geq 0. \tag{3.28}$$

The above formulae are valid in the most general setting. Some interesting particular cases are discussed in [10].

In some sense, the mathematical motivation behind many contributions to this topic has been the desire to explain the phenomenon that, in general

$$\lim_{\varepsilon\downarrow 0}\lim_{T\to\infty}\left[\frac{1}{T+1}\sum_{t=0}^{T}P^t(\varepsilon)\right]\neq\lim_{T\to\infty}\lim_{\varepsilon\downarrow 0}\left[\frac{1}{T+1}\sum_{t=0}^{T}P^t(\varepsilon)\right].$$

Yet, intuitively speaking, the above lack of equality stems from the fact that on the left hand side the perturbation ε is first "allowed" unlimited time to wreck havoc on the ergodic structure of $P(\varepsilon)$ before it is taken to 0. However, on the right hand side the effect of the perturbation is dissipated before the infinite tail of the series has any effect. This interpretation leads us to the following problem (see [72] for some recent partial results). *Investigate the existence and structure of the limit:*

$$\lim_{\varepsilon\downarrow 0}\left[\frac{1}{T(\varepsilon)+1}\sum_{t=0}^{T(\varepsilon)}P_\varepsilon^t(\pi)\right],$$

where $T(\varepsilon)$ is an integer valued function that tends to infinity as ε tends to 0. In particular, investigate the cases where $T(\varepsilon)=O(\frac{1}{\varepsilon^k}), k>0$.

3.3 SINGULARLY PERTURBED MARKOV DECISION PROCESSES

In this section we review some results on singular perturbations of Markov decision processes (MDPs) and introduce a unified approach to singular perturbation, Blackwell optimality and branching Markov decision processes. First we briefly introduce notation and define various optimality criteria. The reader can find a more detailed study on MDPs in the books and surveys [25, 42, 44, 49, 50, 61, 75] and the references therein.

We consider a discrete-time MDP with a finite state space $\mathbb{X}=\{1,...,N\}$ and a finite action space $\mathbb{A}(i)=\{1,...,m_i\}$ for each state $i\in\mathbb{X}$. At any time point t the system is in one of the states $i\in\mathbb{X}$ and the controller or "decision-maker" chooses an action $a\in\mathbb{A}(i)$; as a result the following occur: (a) the controller gains an immediate reward $r(i,a)$, and (b) the process moves to a state $j\in\mathbb{X}$ with probability $p(j|i,a)$, where $p(j|i,a)\geq 0$ and $\sum_{j\in\mathbb{X}}p(j|i,a)=1$.

A *decision rule* π_t at time t is a function which assigns a probability to the event that any particular action a is taken at time t. In general, π_t may depend on history $h_t=(i_0,a_0,i_1,a_1,...,a_{t-1},i_t)$ up to time t. The distribution $\pi_t(\cdot|h_t)$ defines the probability of selecting any action a at time t given the history h_t.

A *policy* (or *strategy*) is a sequence of decision rules $\pi=(\pi_0,\pi_1,...,\pi_t,...)$. A policy π is called *Markov* if $\pi_t(\cdot|h_t)=\pi_t(\cdot|i_t)$. If $\pi_t(\cdot|i)=\pi_{t'}(\cdot|i)$ for all $t,t'\in\mathbb{N}$ then the Markov policy π is called *stationary*. Furthermore, a *deterministic* policy π is a stationary policy whose single decision rule is nonrandomized. It can be defined by the function $f(i)=a, a\in\mathbb{A}(i)$.

Let Π, Π^S and Π^D denote the sets of all policies, of all stationary policies and of all deterministic policies, respectively. For the stationary policy $\pi\in\Pi^S$ define the corresponding transition matrix $P(\pi)=\{p_{ij}(\pi)\}_{i,j=1}^N$ and the reward vector $r(\pi)=\{r_i(\pi)\}_{i=1}^N$

$$p_{ij}(\pi):=\sum_{a\in\mathbb{A}(i)}p(j|i,a)\pi(a|i),\quad r_i(\pi):=\sum_{a\in\mathbb{A}(i)}r(i,a)\pi(a|i).$$

Let the *limit matrix* $P^*(\pi)$ and the *deviation matrix* (or *reduced resolvent* as it is sometimes refered to in this setting) $Y(\pi)$ associated with policy π be defined by

$$P^*(\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P^{t-1}(\pi)$$

and

$$Y(\pi) := (I - P(\pi) + P^*(\pi))^{-1} - P^*(\pi).$$

The *expected average reward* $g_i(\pi)$ and the *expected discounted reward* $v_i^\lambda(\pi)$ are defined as follows:

$$g_i(\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left[P^{t-1}(\pi) r(\pi)\right]_i$$

and

$$v_i^\lambda(\pi) := \sum_{t=1}^{\infty} \lambda^{t-1} \left[P^{t-1}(\pi) r(\pi)\right]_i = \left[(I - \lambda P(\pi))^{-1} r(\pi)\right]_i,$$

where $i \in \mathbb{X}$ is an initial state and $\lambda \in (0,1)$ is a discount factor. One can also use the interest rate $\rho = (1-\lambda)/\lambda \in (0,\infty]$ instead of the discount factor λ. Note that $\lambda \uparrow 1$ is equivalent to $\rho \downarrow 0$. Then, the expected discount reward $v^\rho(\pi) := v^{\lambda(\rho)}(\pi)$ can be expanded as a Laurent series [57, 74, 55] for sufficiently small ρ

$$v^\rho(\pi) = (1+\rho) \left[\rho^{-1} y_{-1}(\pi) + \sum_{m=0}^{\infty} \rho^m y_m(\pi)\right], \tag{3.29}$$

where $y_{-1}(\pi) = P^*(\pi) r(\pi)$ and and $y_m = (-1)^m Y(\pi)^{m+1} r(\pi)$ for $m \geq 0$. Note that the above series can be considered as a particular case of resolvent expansion [48, 64], which in turn can be viewed as a particular case of the inversion of singularly perturbed matrices. Note that in particular $y_{-1}(\pi) = g(\pi)$ and $y_0 = h(\pi)$, where $g(\pi)$ is the expected average reward vector defined above and $h(\pi)$ is the bias vector.

Definition 3.1 *The stationary policy π_* is called a discount optimal policy for the discount factor $\lambda \in (0,1)$ if $v_i^\lambda(\pi_*) \geq v_i^\lambda(\pi)$ for all $i \in \mathbb{X}$ and all $\pi \in \Pi^S$.*

Definition 3.2 *The stationary policy π_* is called a gain optimal policy if $g_i(\pi_*) \geq g_i(\pi)$ for each $i \in \mathbb{X}$ and all $\pi \in \Pi^S$.*

It is well known that there exists a deterministic discount policy and a gain optimal policy, each of which can be computed by a number of efficient algorithms, [25, 49, 61] and Chapter 1. The Laurent expansion (3.29) allows us to define more selective optimality criteria [57, 74, 75].

Definition 3.3 *A policy $\pi^* \in \Pi^S$ is called an m-discount optimal policy for some integer $m \geq 0$ if*

$$[y_m(\pi^*)]_i \geq [y_m(\pi)]_i$$

for all $i \in \mathbb{X}$ and all policies π that are $(m-1)$-discount optimal. By convention, the gain optimal policy is (-1)-discount optimal policy.

In particular, 0-discount optimal policy is called *bias optimal.*

Moreover, it is known that there exists a stationary (and even deterministic) policy that is n-discount optimal for all $n \geq -1$ [57, 74, 75]. It is referred to as the *Blackwell optimal policy* [16] and is defined equivalently[6] in the following way.

Definition 3.4 *A policy π_* is said to be Blackwell optimal if there exists some $\rho_0 > 0$ such that for all $\rho \in (0, \rho_0]$ $v^\rho(\pi^*) \geq v^\rho(\pi)$ for all $\pi \in \Pi^S$.*

In other words, a Blackwell optimal policy is a policy which is discount optimal policy for any discount factor sufficiently close to one.

The results of singular perturbation theory have several applications to Markov decision processes. The first example is the Laurent series (3.29) for the expected discount reward. Another important application is singularly perturbed MDP. For the clarity of exposition we restrict ourselves to the case of linear perturbation, namely, we assume that the transition probabilities of the perturbed MDP are given by

$$p^\varepsilon(j|i,a) = p(j|i,a) + \varepsilon d(j|i,a) \geq 0, \tag{3.30}$$

where $p(j|i,a)$ are transition probabilities of the original unperturbed chain, $\sum_j d(j|i,a) = 0$, and ε is a "small" perturbation parameter. We are especially interested in the case of *singular* perturbations, that is when the perturbation changes the ergodic structure of the underlying Markov chain.

As the following example shows, the policies that are optimal for the unperturbed MDP may not coincide with the optimal policies for the perturbed MDP.

Example 2.1 Let us consider an MDP model with $\mathbb{X} = \{1,2\}$, $\mathbb{A}(1) = \{a_1, b_1\}$, $\mathbb{A}(2) = \{a_2\}$ and

$$p^\varepsilon(1|1,a_1) = 1, \quad p^\varepsilon(2|1,a_1) = 0;$$

$$p^\varepsilon(1|1,b_1) = 1-\varepsilon, \quad p^\varepsilon(2|1,b_1) = \varepsilon;$$

$$p^\varepsilon(1|2,a_2) = \varepsilon, \quad P^\varepsilon(2|2,a_2) = 1-\varepsilon;$$

$$r(1,a_1) = 1, \quad r(1,b_1) = 1.5, \quad r(2,a_2) = 0$$

There are only two deterministic policies, $u = [a_1, a_2]$ and $v = [b_1, a_2]$. For these policies one can easily calculate the average reward vectors. Namely,

$$g^\varepsilon(u) = \begin{cases} \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \varepsilon = 0 \\ \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \varepsilon > 0 \end{cases} \quad \text{and} \quad g^\varepsilon(v) = \begin{cases} \begin{bmatrix} 1.5 \\ 0 \end{bmatrix} & \varepsilon = 0 \\ \begin{bmatrix} 0.75 \\ 0 \end{bmatrix} & \varepsilon > 0 \end{cases}$$

[6] This equivalence holds only for finite state spaces: see Chapter 1. In the general case the two definitions are not equivalent see Chapter 7.

Thus, we can see that for $\varepsilon = 0$ the average optimal policy is v, whereas for $\varepsilon > 0$ the average optimal policy is u.

Since often we do not know the exact value of the perturbation parameter ε, we are interested in finding the policy which is "close" to the optimal one for ε small but different from zero. We will call it a *suboptimal* policy or a *limit control optimal* policy. A strict definition will be given later in this section.

Most research in this topic was carried out with the assumption that the unperturbed MDP is completely decomposable and that the perturbed process is ergodic. More precisely, one can introduce the following four assumptions:

A1) $\mathbb{X} = \cup_{k=1}^{p} \mathbb{X}_k$, where $\mathbb{X}_k \cap \mathbb{X}_l = \emptyset$ if $k \neq l, n > 1$, and $\sum_{k=1}^{p} card(\mathbb{X}_k) = N$

A2) $p(j|i,a) = 0$ whenever $i \in \mathbb{X}_k, j \in \mathbb{X}_l$ and $k \neq l$.

A3) For every $i = 1, ..., p$ the unperturbed MDP associated with the subspace $\mathbb{X}_k$ is ergodic.

A4) The transition matrix $P^{\varepsilon}(\pi)$ is irreducible for any $\pi \in \Pi^S$ and any ε sufficiently small but different from zero, that is the perturbed MDP is ergodic.

The structure defined by assumptions A1)-A4) has a clear interpretation. The perturbed MDP model can be viewed as a complex system consisting of p "weakly-interacting" subsystems associated with $\mathbb{X}_k, k = 1, ..., p$. Note that perturbation $\varepsilon d(j|i,a)$, where i and j are the states of different subsystems $\mathbb{X}_k$ and $\mathbb{X}_l$ respectively, represents the probability of rare transitions between the subsystems, which are independent in the unperturbed chain.

This model of singularly perturbed MDP was first studied by Pervozvanski and Gaitsgory [58]. They proposed an *aggregation - disaggregation* algorithm for the computation of suboptimal policies for the perturbed MDP with average and discount optimality criteria. Furthermore, they have shown that in the case of discount optimality criterion, an optimal policy for the original problem is also suboptimal for the perturbed problem. The latter is not true in general if one uses the gain optimality criterion. In particular, if the perturbation is singular, the suboptimal policy of the perturbed problem can be quite different from the optimal policies of the unperturbed problem (see Example 2.1). The explanation for this phenomenon is that the interaction between weakly-coupled subsystems in the singularly perturbed process makes an effect only after sufficiently long time interval, and, if one uses discounting, the end of the trajectory has no significant contribution to the expected discount reward.

To investigate the above phenomenon, it was proposed in [23] to consider the discount optimality criterion for the case where the interest rate goes to zero (equivalently, the discount factor goes to one) as the perturbation parameter vanishes. In particular, in [23] it is assumed that $\rho = \rho(\varepsilon) = \mu\varepsilon$ for some constant μ. The latter allows one to exhibit the two time scale behavior of the singularly perturbed MDP. In addition, the model in [23] admits a set of transient states. The dynamic programming equation for the perturbed model is solved in [23] by using an approach based on lexicographical ordering. This concept was first applied to MDPs by Veinott [74] to calculate a Blackwell optimal policy. Moreover, as one can see there, singular perturbed MDPs have a close relation with Blackwell optimality [16, 57, 74, 61, 75] and Markov branching decision chains (for comprehensive study of Markov branching chains the reader is referred to the paper by Huang and Veinott [45]). Later, based on

[24], results which appeared first in [23], were extended in [62] to include the possibility of several time scales. The key point in [24] and [62] is the use of the following dependency of the interest rate on the perturbation parameter: $\rho = \rho(\varepsilon) = \mu\varepsilon^l$, where l represents the order of a time scale. In [4] Altman and Gaitsgory have generalized the results of [58] to constrained Markov decision processes.

Abbad, Bielecki and Filar [1, 3, 14, 2] formulated the *limit control principle*, which provides a formal setting for the concept of suboptimal policies. First, as elaborated in Section 2, we note that for any stationary policy $\pi \in \Pi^S$, there exists a limiting ergodic projection

$$\hat{P}^*(\pi) := \lim_{\varepsilon \to 0} P^{*,\varepsilon}(\pi).$$

The average reward optimization problem for the perturbed MDP can be written in the form

$$g_i^{opt,\varepsilon} = \max_{\pi \in \Pi^S} [P^{*,\varepsilon}(\pi) r(\pi)]_i \quad (L^\varepsilon).$$

The limit control principle says that instead of the above singular program one can consider a well-defined *limit Markov control problem*:

$$\hat{g}_i^{opt} = \max_{\pi \in \Pi^S} [\hat{P}^*(\pi) r(\pi)]_i \quad (L).$$

It is natural to expect that an optimal strategy to (L), if exists, will approximate well the optimal strategy for the perturbed problem (L^ε), in the case where the perturbation parameter is small. In [14] it was shown that this is indeed the case. Specifically, let π_* be any maximizer in (L), then

$$\lim_{\varepsilon \to 0} \max_{i \in \mathbb{X}} |g_i^\varepsilon(\pi_*) - g^{opt,\varepsilon}| = 0.$$

In [1] it is proved that there exists a deterministic policy that solves the limit control problem (L). Recently, Bielecki and Stettner [15] have generalized the limit control principle to MDPs with general Borel state spaces and compact action spaces.

Under the assumptions A1) - A4) the limit Markov control problem (L) can be solved by solving the following linear programming problem (P):

maximize $\sum_{k=1}^n \sum_{i \in \mathbb{X}_k} \sum_{a \in A(i)} r(i,a) z_{ia}^k$

subject to:

$$\sum_{i \in \mathbb{X}_k} \sum_{a \in A(i)} (\delta_{ij} - p(j|i,a)) z_{ia}^k = 0, \quad j \in \mathbb{X}_k; k = 1, \ldots, n$$

$$\sum_{k=1}^n \sum_{j \in \mathbb{X}_\ell} \sum_{i \in \mathbb{X}_k} \sum_{a \in A(i)} d(j|i,a) z_{ia}^k = 0; \quad \ell = 1, \ldots, n$$

$$\sum_{k=1}^n \sum_{i \in \mathbb{X}_k} \sum_{a \in A(i)} z_{ia}^k = 1$$

$$z_{ia}^k \geq 0 \quad k = 1, \ldots, n; \quad i \in \mathbb{X}_k, a \in A(i).$$

It can be shown (see [3]) that an optimal strategy in the limit Markov control problem (L) can be constructed as follows.

Theorem 3.8 *Let $\{z_{ia}^k | k = 1, \ldots, n;\ i \in \mathbb{X}_k;\ a \in A(i)\}$ be an optimal extreme solution to the linear program (P), then the deterministic strategy defined by*

$$f_*(i) = a, \quad i \in \mathbb{X}_k,\ k = 1, \ldots, n \Longleftrightarrow z_{ia}^k > 0$$

is optimal in the limit Markov control problem (L).

The linear program (P) is similar to one given by Gaitsgory and Pervozvanski [33]. However, these authors used techniques different from those in [3].

As a result of the solution of problem (L), one gets *limit control optimal* policy $f_* \in \Pi^D$ such that for any other deterministic policy $f \in \Pi^D$ the holds

$$\lim_{\varepsilon \to 0} (g_i^\varepsilon(f_*) - g_i^\varepsilon(f)) \geq 0, \quad i \in \mathbb{X}.$$

However, as was noted in [1], a policy that solves (L), in general, is only sub-optimal. Interestingly, there exists a policy that is optimal for all sufficiently small ε but different from zero (see [1]). This policy is called *uniform optimal* [5] and it satisfies the following inequality

$$g_i^\varepsilon(f_*) \geq g_i^\varepsilon(f), \quad i \in \mathbb{X},$$

for any deterministic policy f and all ε sufficiently small but different from zero. We like to emphasize that a uniform optimal policy is limit control optimal, but the converse need not hold, as the following example illustrates.

Example 2.2 Consider $\mathbb{X} = \{1, 2\}$, $\mathbb{A}(1) = \{a_1, b_1\}$, $\mathbb{A}(2) = \{a_2\}$; let

$$\begin{array}{ll} p^\varepsilon(1|1, a_1) = 1, & r(1, a_1) = 10 \\ p^\varepsilon(1|1, b_1) = 1 - \varepsilon, & r(1, b_1) = 10 \\ p^\varepsilon(2|1, b_1) = \varepsilon, & \\ p^\varepsilon(1|2, a_2) = 1, & r(2, a_2) = 0 \end{array}$$

Then the stationary policy $u(1) = a_1, u(2) = a_2$ is uniformly optimal with expected average reward $g_i^\varepsilon(u) = 10$. The stationary policy $v(1) = b_1, v(2) = a_2$ is limit control optimal as $\lim_{\varepsilon \to 0} g_i^\varepsilon(v) = 10$, but for every $\varepsilon > 0$,

$$g_i^\varepsilon(v) = \frac{10}{1 + \varepsilon} < g_i^\varepsilon(u).$$

A heuristic procedure to determine the uniform optimal policy is proposed in [1]. In contrast, a well-defined computational procedure was proposed in [5], which is based on asymptotic linear programming. The *asymptotic linear programming* was first introduced by Jeroslow [46, 47] and later refined by Hordijk, Dekker and Kallenberg [43]. It is designed to solve the following parametric linear program

$$\max c(\varepsilon)x \tag{3.31}$$

$$A(\varepsilon)x = b(\varepsilon), \quad x \geq 0 \tag{3.32}$$

for ε small but different from zero. The elements of $A(\varepsilon), b(\varepsilon)$ and $c(\varepsilon)$ are assumed to be polynomials of ε. The basic idea of this method is to perform the simplex algorithm over the field of rational functions instead of the field of real numbers. The latter results in a solution that is optimal for all ε sufficiently small but different from zero. We would like to note that instead of asymptotic linear programming one can use its modification, *the asymptotic simplex method* [58, 53, 54, 8], which operates over the field of Laurent series. Moreover, if the asymptotic linear programming or asymptotic simplex method is used, then one can drop Assumptions A1)-A4) and consider the most general situation.

As can be concluded from the above overview of Markov decision processes, there is an intrinsic relationship between the Blackwell optimality and singular perturbations. Actually, the contributions of Delebecque and Quadrat [23, 24, 62] clearly emphasize this connection. It turns out that the Blackwell optimality, singularly perturbed MDPs and branching Markov decision chains can all be considered in a unified framework based on asymptotic simplex method. Towards this end, let us consider the linear programming formulation for the following four MDP models.

Model I: MDP with Blackwell optimality criterion

A Blackwell optimal policy is a discount optimal policy for all discount factor sufficiently close to one, or, equivalently, the policy which is optimal for all interest rate sufficiently close to zero. A Blackwell optimal policy can be determined by using the asymptotic linear programming approach [7, 43, 53, 54]. Namely, one needs to find a solution for the following linear program which is optimal for all interest rate $\rho > 0$ which are sufficiently small:

$$\max \sum_{i,a} (1+\rho) r(i,a) x_{ia}$$

$$\sum_{i,a} [(1+\rho)\delta_{ij} - p(j|i,a)] x_{ia} = 1, \quad x_{ia} \geq 0. \tag{3.33}$$

Note that the above linear program can be immediately written in the form (3.31),(3.32) with $\varepsilon = \rho$. In [43] a refined version of Jeroslow's asymptotic linear programming method [46, 47] for solving (3.33) was developed. Later Lamond [54] applied his version of the asymptotic simplex method to the above parametric LP. Of course, our algorithm [8] can be applied to this model as well. As pointed out in [54], in this particular case, the Laurent series for the basis matrix always admits a simple pole. This in turn implies that the complexity of the basis updating is $O(n^2)$, which is comparable with the complexity of the basis updating for the ordinary simplex method.

Model II: Markov branching decision chains.

The Markov branching decision chains were introduced in [45]. These are MDPs with immediate rewards which dependent on the interest rate. Namely, it is assumed that $r(i,a) = r^{\rho}(i,a)$ is a polynomial in the interest rate ρ. To

find a policy which is optimal for all sufficiently small ρ (Strong-value policy [45]), we need to solve only slightly modified version of (3.33), that is

$$\max \sum_{i,a} (1+\rho) r^{\rho}(i,a) x_{ia}$$

$$\sum_{i,a} [(1+\rho)\delta_{ij} - p(j|i,a)] x_{ia} = 1, \quad x_{ia} \geq 0. \tag{3.34}$$

Again the asymptotic simplex method can be immediately applied. A method of finding the optimal policy proposed by Huang and Veinott [45] also employs the asymptotic programming approach. The method is based on the subsequent solution of the augmented LPs, whose objective functions and right hand sides (but not the constraint coefficient matrix) depend on the parameter.

Model III: Singularly perturbed MDP with average criterion

In Section 3.4 we applied the asymptotic linear programming [46, 47, 43], which is based on the ordering in the field of rational functions, to singularly perturbed MDPs with average criterion. Since one needs to solve the following parametric program

$$\max \sum_{i,a} r(i,a) x_{ia}$$

$$\sum_{i,a} [\delta_{ij} - p(j|i,a) - \varepsilon d(j|i,a)] x_{ia} = 0,$$

$$\sum_{a} x_{ja} + \sum_{i,a} [(\delta_{ij} - p(j|i,a) - \varepsilon d(j|i,a)] y_{ia} = \beta_j,$$

$$x_{ia} \geq 0, \qquad y_{ia} \geq 0, \tag{3.35}$$

the asymptotic simplex method, based on the Laurent series expansions, can be applied as well. Moreover, since the basis updating in the asymptotic linear programming takes $O(m^4 log(m))$ flops and the asymptotic simplex method takes $\bar{s}m^2$ flops, the latter method appears to be faster.

Model IV: Singularly perturbed MDP with killing interest rate

In [23, 62] singularly perturbed MDPs with "killing interest rate" $\rho(\varepsilon) = \mu\varepsilon^l$, where l is the order of a time scale were considered. This model exhibits the necessity of different control regimes for the different time scales. The papers [23, 62] provided a lexicographical policy improvement algorithm for the solution of the perturbed dynamic programming equation. Alternatively, the extension of our asymptotic simplex method for the polynomial perturbation can be used in this problem. Here the parametric linear program takes the following form

$$\max \sum_{i,a} (1+\mu\varepsilon^l) r(i,a) x_{ia}$$

$$\sum_{i,a}[(1+\mu\varepsilon^l)\delta_{ij} - p(j|i,a) - \varepsilon d(j|i,a)]x_{ia} = 1, \quad x_{ia} \geq 0. \tag{3.36}$$

Generalized Model:

Finally, we would like to note that the Models I,II and IV can be viewed as particular cases of the next unified scheme. Let transition probabilities $p^\varepsilon(j|i,a)$, immediate rewards $r^\varepsilon(i,a)$ and interest rate $\rho(\varepsilon)$ of an MDP model be polynomials of the parameter ε. Then a policy which is optimal for all sufficiently small values of parameter ε can be found from the next perturbed linear program.

$$\max \sum_{i,a}(1+\rho(\varepsilon))r^\varepsilon(i,a)x_{ia}$$

$$\sum_{i,a}[(1+\rho(\varepsilon))\delta_{ij} - p^\varepsilon(j|i,a)]x_{ia} = 1, \quad x_{ia} \geq 0. \tag{3.37}$$

The above perturbed LP can be efficiently solved by the generalization of the asymptotic simplex method proposed in [8]. Note that we retrieve Model I with $\rho(\varepsilon) = \varepsilon$, $r^\varepsilon(i,a) = r(i,a)$, $p^\varepsilon(j|i,a) = p(j|i,a)$; Model II with $\rho(\varepsilon) = \varepsilon$, $r^\varepsilon(i,a) = \sum_{k=0}^{p}\varepsilon^k r_k(i,a)$, $p^\varepsilon(j|i,a) = p(j|i,a)$; and Model IV with $\rho(\varepsilon) = \mu\varepsilon^l$, $r^\varepsilon(i,a) = r(i,a)$, $p^\varepsilon(j|i,a) = p(j|i,a) + \varepsilon d(j|i,a)$.

3.4 SINGULARLY PERTURBED MDP'S AND THE COMBINATORIAL HAMILTONIAN CYCLE PROBLEM

In this section we demonstrate that a cornerstone combinatorial optimization problem such as the Hamiltonian Cycle Problem (HCP) can be reduced to a singularly perturbed Markov Decision Process. We begin with a brief description of only one version of the HCP.

In graph theoretic terms, the problem is to find a simple cycle of N arcs, that is a *Hamiltonian Cycle* or a *tour*, in a directed graph G with N nodes and with arcs $(i,j) \in N \times N$, or to determine that none exists. Recall that a simple cycle is one that passes exactly once through each node comprising the cycle. It is known that HCP belongs to the NP-complete class of problems and, as such, is considered very difficult from an algorithmic perspective.

In this section we propose the following, unorthodox, perspective of the Hamiltonian cycle problem: Consider a moving object tracing out a directed path on the graph G with its movement "controlled" by a function f mapping the set of nodes $\mathbb{X} = \{1, 2, \ldots, N\}$ into the set of arcs A.

Clearly, we can think of this set of nodes as the state space of a Markov decision process Γ where for each state/node i, the action space

$$A(i) = \{a = j | (i,j) \in A\}$$

is in one to one correspondence with the set of arcs emanating from that node.

Of course, we shall ignore the trivial case $A(i) = \emptyset$, because in such a case, obviously, no Hamiltonian cycle exists. Furthermore, if we restrict the function f above in such a way that $f(i) \in A(i)$, for each $i \in \mathbb{X}$, then we see that f can be thought of as a deterministic strategy f in an MDP Γ. Designating node 1 as the "home node" in G, we shall say that f is a *Hamiltonian cycle in* G if the set of arcs $\{(1, f(1)), (2, f(2)), \ldots, (N, f(N))\}$ is a Hamiltonian cycle in G. If the above set of arcs contains cycles of length less than N, we shall say that f *has subcycles in* G.

Note that if $P(f)$ is the transition probability matrix of a Markov chain induced by f that is a Hamiltonian cycle, then $P(f)$ is irreducible and the long-run frequency of visits to any state $x_i(f) = 1/N$. Of course, if f has subcycles in G, then $P(f)$ contains multiple ergodic classes which complicates the analysis of the Markov decision process Γ, in which we have embedded our graph theoretic problem.

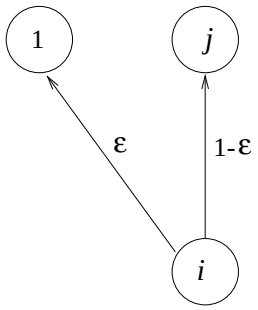

Figure 3.1

A class of limiting average Markov decision processes that retains most of the desirable properties of the irreducible processes is the so-called "unichained" class. Briefly, a Markov decision process is *unichained* if for every deterministic stationary control $\mathbf{f}$, $P(f)$ contains only a single ergodic class and possibly a nonempty set of transient states. We now perturb the transition probabilities of Γ slightly to create an ε-perturbed process $\Gamma(\varepsilon)$ (for $0 < \varepsilon < 1$) defined by:

$$p^{\varepsilon}(j|i,a) \quad = \quad \begin{cases} 1 & \text{if } i = 1 \text{ and } a = j \\ 0 & \text{if } i = 1 \text{ and } a \neq j \\ 1 & \text{if } i > 1 \text{ and } a = j = 1 \\ \varepsilon & \text{if } i > 1,\ a \neq j, \text{ and } j = 1 \\ 1 - \varepsilon & \text{if } i > 1,\ a = j, \text{ and } j > 1 \\ 0 & \text{if } i > 1,\ a \neq j, \text{ and } j > 1. \end{cases}$$

Note that with the above perturbation, for each pair of nodes i, j (not equal to 1) corresponding to a "deterministic arc" (i, j) our perturbation replaces that arc by a pair of "stochastic arcs" $(i, 1)$ and (i, j), see Figure 1, with weights ε and $(1 - \varepsilon)$ r respectively ($\varepsilon \in (0, 1)$). Note that this perturbation changes Γ to an ε-perturbed Markov decision process $\Gamma(\varepsilon)$.

Example 4.1
Consider the following complete graph G on four nodes (with no self-loops):

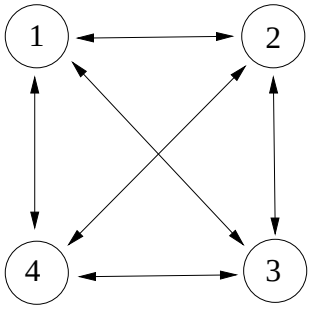

Figure 3.2

and think of the nodes as the states of an MDP, denoted by Γ, and of the arcs emanating from a given node as actions available at that state. The Hamiltonian cycle $c_1 : 1 \to 2 \to 3 \to 4 \to 1$ corresponds to the deterministic stationary strategy $f_1 : \{1,2,3,4\} \to \{2,3,4,1\}$ where $f_1(2) = 3$ corresponds to the controller choosing arc (2,3) in state 2 with probability 1. The Markov chain induced by f_1 is given by the transition matrix

$$P(f_1) \quad = \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

which is irreducible. On the other hand, the union of two sub-cycles: $1 \to 2 \to 1$ and $3 \to 4 \to 3$ corresponds to the policy $f_2 : \{1,2,3,4\} \to \{2,1,4,3\}$ which identifies the Markov chain transition matrix

$$P(f_2) \quad = \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

containing two distinct ergodic classes.

As mentioned earlier, the perturbation destroys multiple ergodic classes and induces a unichained, singularly perturbed, Markov decision process $\Gamma(\varepsilon)$. For instance, the policy f_2 now has the Markov chain matrix

$$P(f_2) \quad = \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \varepsilon & 0 & 0 & 1-\varepsilon \\ \varepsilon & 0 & 1-\varepsilon & 0 \end{pmatrix}.$$

Remark 4.1
Our perturbation has made the home node/state 1 rather special. In particular, the home state always belongs to the single ergodic class of $P(f)$ for any

$f \in \Pi^S$. Of course, some other states could be transient.

We shall undertake the analysis of the Hamiltonian cycle problem in the "frequency space" of the perturbed process $\Gamma(\varepsilon)$. Recall that with every $f \in \Pi^S$ we can associate the long-run frequency vector $\mathbf{x}(f)$. This is achieved by defining a map $M : \Pi^S \to X(\varepsilon)$ by

$$x_{ia}(f) = \pi_i(\varepsilon, f) f(a|i); \quad f \in \Pi^S$$

for each $i \in \mathbb{X}$ and $a \in A(i)$, where $\pi_i(\varepsilon, f)$ is the i-th element of the stationary distribution vector of the perturbed Markov chain transitions matrix $P(f)$, and $f(a|i)$ is the probability of choosing action a in state i.

Now consider the polyhedral set $\mathbf{X}(\varepsilon)$ defined by the constraints

(i) $$\sum_{i=1}^{N} \sum_{a \in A(i)} [\delta(i,j) - p_\varepsilon(j|i,a)]\, x_{ia} = 0;\ j \in \mathbb{X}.$$

(ii) $$\sum_{i=1}^{N} \sum_{a \in A(i)} x_{ia} = 1.$$

(iii) $x_{ia} \geq 0;\ a \in A(i),\ i \in \mathbb{X}$.

Next define a map $\hat{M} : \mathbf{X}(\varepsilon) \to \Pi^S$ by

$$f_{\mathbf{x}}(i,a) \quad = \quad \begin{cases} \frac{x_{ia}}{x_i}; & \text{if } x_i = \sum_{a \in A(i)} x_{ia} > 0 \\ 1; & \text{if } x_i = 0 \text{ and } a = a_1 \\ 0; & \text{if } x_i = 0 \text{ and } a \neq a_1, \end{cases}$$

for every $a \in A(i)$, $i \in \mathbb{X}$ where a_1 denotes the first available action in a given state according to some ordering. The following result can be found in [28] and [31].

Lemma 4.1

(i) *The set* $\mathbf{X}(\varepsilon) = \{\mathbf{x}(f) | f \in \Pi^S\}$ *and will henceforth be called the (long-run) "frequency space" of* $\Gamma(\varepsilon)$.

(ii) *For every* $\mathbf{x} \in \mathbf{X}(\varepsilon)$,

$$M\left(\hat{M}(\mathbf{x})\right) = \mathbf{x}$$

but the inverse of M *need not exist.*

(iii) *If* $\mathbf{x}$ *is an extreme point of* $\mathbf{X}(\varepsilon)$, *then*

$$f_{\mathbf{x}} = \hat{M}(\mathbf{x}) \in \Pi^D.$$

(iv) *If* $f \in \Pi^D$ *is a Hamiltonian cycle, then* $\mathbf{x}(\mathbf{f})$ *is an extreme point of* $\mathbf{X}(\varepsilon)$.

We shall now derive a useful partition of the class Π^D of deterministic strategies that is based on the graphs they "trace out" in G. In particular, note that with each $f \in \Pi^D$ we can associate a subgraph G_f of G defined by

$$\text{arc } (i,j) \in G_f \iff f(i) = j$$

We shall also denote a simple cycle of length m and beginning at 1 by a set of arcs

$$c_m^1 = \{(i_1 = 1, i_2), (i_2, i_3), \ldots, (i_m, i_{m+1} = 1)\}; \quad m = 2, 3, \ldots, N.$$

Of course, c_N^1 is a Hamiltonian cycle. If G_f contains a cycle c_m^1 we write $G_f \supset c_m^1$. For $2 \leq m \leq N$, let $C_m := \{f \in \Pi^D | G_f \supset c_m^1\}$, namely, the set of deterministic strategies that trace out a simple cycle of length m, beginning at 1, for each $m = 2, 3, \ldots, N$. Of course, C_N is the set of strategies that correspond to Hamiltonian cycles and any single C_m can be empty, depending on the structure of the original graph G. Thus a typical strategy $f \in C_3$, for example, traces out a graph G_f in G that might look like Figure 3 where the dots indicate the "immaterial" remainder of G_f that corresponds to states that are transient in $P(f)$.

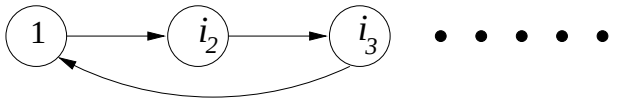

Figure 3.3

The partition of the deterministic strategies that seems to be most relevant for our purposes is

$$\Pi^D = \left[\bigcup_{m=2}^{N} C_m\right] \bigcup B, \tag{3.38}$$

where B contains[7] all the deterministic strategies that are not in any of the C_m's.

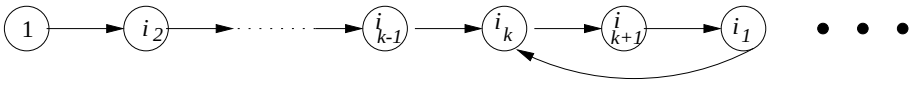

Figure 3.4

Note that a typical strategy f in B traces out a graph G_f in G that looks like Figure 4, where the dots again denote the immaterial part of G_f. However, it is important to note that for any $\varepsilon > 0$, the states $1, i_2, \ldots, i_{k-1}$ are not transient in $\Gamma(\varepsilon)$.

It is, perhaps, interesting to observe that all strategies in a given set in the partition (3.38) induce the same long-run frequency $x_1(f)$ of visits to the home

[7] It will soon be seen that the strategies in B are in a certain sense "bad" or, more precisely, difficult to analyze, thereby motivating the symbol B.

node 1. This observation is captured in the following proposition which can be found in [17] and [31]

Proposition 4.2
Let $\varepsilon \in (0,1)$, $f \in \Pi^D$, and $\mathbf{x}(f)$ be its long-run frequency vector (that is, $\mathbf{x}(f) = M(f)$). The long-run frequency of visits to the home state 1 is given by

$$x_1(f) = \sum_{a \in A(1)} x_{1a}(f) = \begin{cases} \dfrac{1}{d_m(\varepsilon)}; & \text{if } f \in C_m,\ m = 2,3,\ldots,N \\ \dfrac{\varepsilon}{1+\varepsilon}; & \text{if } f \in B, \end{cases}$$

where $d_m(\varepsilon) = 1 + \sum_{i=2}^{m} (1-\varepsilon)^{i-2}$ for $m = 2,3,\ldots,N$.

The above proposition leads the following characterizations of the Hamiltonian cycles of a directed graph.

Theorem 3.9

(i) Let $f \in \Pi^D$ be a Hamiltonian cycle in the graph G. Then $G_f = c_N^1$, $\mathbf{x}(f)$ is an extreme point of $\mathbf{X}(\varepsilon)$ and $x_1(f) = \frac{1}{d_N(\varepsilon)}$.

(ii) Conversely, suppose that $\mathbf{x}$ is an extreme point of $\mathbf{X}(\varepsilon)$ and that $x_1 = \sum_{a \in A(1)} x_{1a} = \frac{1}{d_N(\varepsilon)}$, then $f = \hat{M}(\mathbf{x})$ is a Hamiltonian cycle in G.

(iii) Hamiltonian cycles of the graph G are in $1:1$ correspondence with those points of $\mathbf{X}(\varepsilon)$ which satisfy

(a) $x_1 = \sum_{a \in A(1)} x_{1a} = \dfrac{1}{d_N(\varepsilon)}$.

(b) For every $i \in \mathbb{X}$, $x_i = \sum_{a \in A(1)} x_{ia} > 0$ and $\frac{x_{ia}}{x_i} \in \{0,1\}$ for each $a \in A(i)$, $i \in \mathbb{X}$.

Remark 4.2: It is, perhaps, significant to note that for all $\varepsilon \in (0,1)$, $m = 2,3,\ldots,N-1$

$$\frac{1}{d_m(\varepsilon)} > \frac{1}{d_{m+1}(\varepsilon)} > \frac{\varepsilon}{1+\varepsilon}.$$

Thus Theorem 9 demonstrates that the extreme points $\mathbf{x}$ of $\mathbf{X}(\varepsilon)$ can be "ranked" according to the values of the linear function $l(\mathbf{x}) = \sum_{a \in A(1)} x_{1a}$. Unfortunately, the Hamiltonian cycles (if they exist) may attain only the "second lowest" value of $l(\mathbf{x})$, namely, $\frac{1}{d_N(\varepsilon)}$. The latter problem is, partially, rectified in the following result that can be found in [29].

Theorem 3.10 *Let $f^* \in \Pi^D$ be a Hamiltonian cycle in the graph G. Then for $\varepsilon \geq 0$ and sufficiently small, f^* is a global minimizer of the following perturbed optimization problem:*

$$\min_{f \in \Pi^D} \{[I - P(f) + P^*(f)]_{11}^{-1}\},$$

where H_{11} denotes the (1,1)th entry of matrix H.

The above theorem shows that if Hamiltonian cycles exist, they are the minimizers of the top left element of the fundamental matrix, over $f \in \Pi^D$, as long as ε is sufficiently near 0. Recently it was shown (see [26]) that $\varepsilon \leq 1/N^2$ qualifies as being sufficiently small, but a sharp upper bound on ε is not known. Further, from the optimization point of view, elements of the fundamental matrix are not straightforward to analyze. This leads naturally to the *open problem*: can asymptotic expansions such as the one in Theorem 5, facilitate algorithmic approaches to the optimization problem in Theorem 10? We conjecture that a necessary first step is to establish that

$$\min_{f\in\Pi^D}\{[I - P(f) + P^*(f)]^{-1}_{11}\} = \min_{f\in\Pi^S}\{[I - P(f) + P^*(f)]^{-1}_{11}\},$$

but we do not have a proof of the above statement.

To date the best algorithmic results based on this stochastic approach to the HCP are reported in [9]. These results exploit the properties in (i)–(iii) of Theorem 9, above. In particular, it can be checked that the most awkward requirement $x_{ia}/x_i \in \{0,1\}$ for all $i \in \mathbb{X}, a \in \mathbb{A}(i)$ is equivalent $\min\{x_{ia}, x_{ib}\} = 0$ for all $i \in \mathbb{X},\ a,b \in \mathbb{A}(i)$ and $a \neq b$. This observation immediately leads to the following mixed integer programming formulation of the HCP problem:

$$\min \textstyle\sum_i \sum_a c_{ia} x_{ia}$$

$$\begin{array}{lrcll}
\text{s.t.} & x & \in & X(\varepsilon) & \\
& x_1 & = & 1/d_N(\varepsilon) & \\
& x_{ia} & \leq & My_{ia} & : \quad i \in \mathbb{X}, a \in \mathbb{A}(i) \\
& y_{ia} + y_{ib} & \leq & 1 & ; \quad i \in \mathbb{X}, a,b \in \mathbb{A}(i), a \neq b \\
& y_{ia} & \in & \{0,1\} & ; \quad i \in \mathbb{X}, a \in \mathbb{A}(i).
\end{array}$$

In the above $M \geq 1/d_N(\varepsilon)$ and c_{ia}'s can be experimented with. In preliminary numerical experiments reported in [9], randomly generated problems with up to 100 nodes and 300 arcs were solved in less than 150 cpu seconds on a Sun Workstation. Recently, Filar and Lasserre [30] proposed a non-standard branch and bound algorithm for the formulation

$$\begin{array}{ll}
& \min \displaystyle\sum_i \sum_a c_{ia} x_{ia} \\
\text{s.t.} & \text{(a)} \quad \mathrm{x} \in \mathrm{X}(\varepsilon) \\
& \text{(b)} \quad \mathrm{x_1} = 1/\mathrm{d_N}(\varepsilon) \\
& \text{(c)} \quad \mathrm{x_{ia}/x_i} \in \{0,1\};\ \mathrm{i} \in \mathbb{X},\ \mathrm{a} \in \mathbb{A}(\mathrm{i})
\end{array}$$

which exploits the structural property that ensures that if the relaxation omitting (c) is solved by the simplex method, then (c) can be violated for at most one i and at most one pair of arcs a, b.

In an interesting, related, development Feinberg [27] has recently considered the embedding of a Hamiltonian cycle problem in a discounted Markov decision process. In that paper the perturbation parameter ε is not necessary but, instead, the discount factor $\lambda \in [0,1)$ plays a crucial role. In particular

Feinberg's embedding can be obtained by setting $\varepsilon = 0$ in $p_1^\varepsilon(j|i,a)$ as defined earlier and by setting

$$r(i,a) = \begin{cases} 1 & \text{if i} = 1,\ \text{a} \in \mathbb{A}(1) \\ 0 & \text{otherwise} \end{cases}$$

For any $\pi \in \Pi^S$ the expected discounted reward $v_i^\lambda(\pi)$ is now defined as in Section 3. The analysis in [27] is based on the following observation. Let i_m denote the state/node visited at stage m, then

$$v_1^\lambda(\pi) = \sum_{m=0}^{\infty} \lambda^m P_1^\pi(i_m = 1),$$

where $P_1^\pi(\cdot)$ denotes the probability measure induced by π and the initial state $i_0 = 1$, and

$$P_1^\pi(i_m = 1) = \frac{1}{m!}\left[\frac{\partial^m}{\partial \lambda^m}(v_1^\lambda(\pi))\right]_{\lambda=0}.$$

The above lead to novel characterizations of Hamiltonian cycles that are summarized below.

Theorem 3.11 *With the embedding in Γ_λ described above the following statements are equivalent:*

(i) A policy $\pi = f$ is deterministic and a Hamiltonian cycle in G.

(ii) A policy π is stationary and a Hamiltonian cycle in G.

(iii) A policy f is deterministic and $v_1^\lambda(f) = (1-\lambda^N)^{-1}$ for at least $\lambda \in [0,1)$.

(iv) A policy π is stationary and $v_1^\lambda(\pi) = (1-\lambda^N)^{-1}$ for $2N-1$ distinct discount factors $\lambda_k \in (0,1)$; $k = 1, 2, \ldots, 2N-1$.

The above characterization naturally leads to a number of mathematical programming formulations of both HCP and TSP that are described in [27]. There is clearly a need to explore the algorithmic potential of these formulations.

References

[1] M. Abbad and J.A. Filar, "Perturbation and stability theory for Markov control problems", *IEEE Trans. Auto. Contr.*, AC-37, no. 9, pp. 1415–1420, 1992.

[2] M. Abbad and J.A. Filar, "Algorithms for singularly perturbed Markov control problems: A survey", in *Techniques in discrete-time stochastic control systems* (ed.) C.T. Leondes, Series: Control and Dynamic Systems, v. 73, Academic Press, New York, 1995.

[3] M. Abbad, J.A. Filar and T.R. Bielecki, "Algorithms for singularly perturbed limiting average Markov control problems," *IEEE Trans. Auto. Contr.* AC-37, pp. 1421–1425, 1992.

[4] E. Altman and V.G. Gaitsgory, "Stability and Singular Perturbations in Constrained Markov Decision Problems", *IEEE Trans. Auto. Control*, v. 38, pp. 971–975, 1993.

[5] E. Altman, K.E. Avrachenkov, and J.A. Filar, "Asymptotic linear programming and policy improvement for singularly perturbed Markov decision processes", *ZOR: Math. Meth. Oper. Res.*, v. 49, pp. 97–109, 1999.

[6] E. Altman, E. Feinberg, J.A. Filar, and V.A. Gaitsgory, "Perturbed Zero-sum Games with Applications to Dynamic Games," in *Proc. 8th International Symposium on Dynamic Games and Aplications*, pp. 45–51, Maastricht, The Netherlands, 1998 (to appear in the Annals of Dynamic Games; Birkhauser).

[7] E. Altman, A. Hordijk, and L.C.M. Kallenberg, "On the value function in constrained control of Markov chains", *ZOR: Math. Meth. Oper. Res.*, v. 44, pp. 387–399, 1996.

[8] J.A. Filar, E. Altman and K.E. Avrachenkov, "An asymptotic simplex method for singularly perturbed linear programs", submitted to *Operations Research Letters*, 1999.

[9] M. Andramonov, J. Filar, A. Rubinov and P. Pardalos, "Hamiltonian Cycle Problem via Markov Chains and Min-type Approaches" in *Approximation and Complexity in Numerical Optimization: Continuous and Discrete Problems*, Ed. P.M. Pardalos, Kluwer Academic Publishers, 2000.

[10] K.E. Avrachenkov, "Analytic perturbation theory and its applications", PhD Thesis, University of South Australia, 1999.

[11] K.E. Avrachenkov and M. Haviv, "The highest singular coefficients in singular perturbation of stochastic matrices," INRIA Sophia Antipolis, (in preparation).

[12] K.E. Avrachenkov, M. Haviv and P.G. Howlett, "Inversion of analytic matrix functions that are singular at the origin", *SIAM J. Matrix Anal. Appl.,* (to appear).

[13] K.E. Avrachenkov and J.B. Lasserre, "The fundamental matrix of singularly perturbed Markov chains," *Advances in Applied Probability*, v. 31, (to appear).

[14] T.R. Bielecki and J.A. Filar, "Singularly perturbed Markov control problem: Limiting average cost", *Annals of O.R.*, v. 28, pp. 153–168, 1991.

[15] T.R. Bielecki and L. Stettner, "Ergodic control of singularly perturbed Markov process in discrete time with general state and compact action spaces", preprint, Mathematics Department, Northeastern Illinois University, 1996.

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

[17] M. Chen and J.A. Filar (1992),"Hamiltonian Cycles, Quadratic Programming and Ranking of Extreme Points" in *Global Optimization,* C. Floudas and P. Pardalos, eds. Princeton University Press.

[18] M. Cordech, A.S. Willsky, S.S. Sastry and D.A. Castanon, "Hierarchical aggregation of linear systems with multiple time scales," *IEEE Trans. Autom. Contr.*, AC-28, pp. 1029–1071, 1083.

[19] M. Cordech, A.S. Willsky, S.S. Sastry, and D.A. Castanon, "Hierarchical aggregation of singularly perturbed finite state Markov processes," *Stochastics.* v. 8, pp. 259–289, 1983.

[20] P.J. Courtois, *Decomposability: queueing and computer system applications*, Academic Press, New York, 1977.

[21] P.J. Courtois and G. Louchard, "Approximation of eigencharacteristics in nearly-completely decomposable stochastic systems", *Stoch. Process. Appl.*, v. 4, pp. 283–296, 1976.

[22] P.J. Courtois and P. Semel, "Bounds for the positive eigenvectors of nonnegative matrices and their approximation by decomposition", *JACM*, v. 31, pp. 804–825, 1984.

[23] F. Delebecque and J.P. Quadrat, "Optimal control of Markov chains admitting strong and weak interactions", *Automatica*, v. 17, pp. 281–296, 1981.

[24] F. Delebecque, "A reduction process for perturbed Markov chain," SIAM J. Appl. Math., v. 43, pp. 325–350, 1983.

[25] C. Derman, *Finite state Markovian decision processes*, Academic Press, New York, 1970.

[26] V.V. Ejov, J.A. Filar and M.T. Nguyen, "Hamiltonian cycles and singularly perturbed Markov chains", School of Mathematics, University of South Australia, (in preparation).

[27] E.A. Feinberg, "Constrained discounted Markov decision processes and Hamiltonian cycles", *Math. Oper. Res.*, v. 25, pp. 130–140, 2000.

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

[29] J.A. Filar and Ke Liu, "Hamiltonian cycle problem and singularly perturbed Markov decision process", in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, IMS Lecture Notes - Monograph Series, USA, 1996.

[30] J.A. Filar and J-B Lasserre, "A Non-Standard Branch and Bound Method for the Hamiltonian Cycle Problem", *ANZIAM J. (on line)*, v. 42, 2000 (to appear).

[31] J.A. Filar and K. Vrieze, *Competitive Markov Decision Processes,* Springer-Verlag, N.Y., 1996.

[32] V.G. Gaitsgori and A.A. Pervozvanskii, "Aggregation of states in a Markov chain with weak interactions", *Cybernetics*, v. 11, pp. 441–450, 1975. (Translation of Russian original in *Kibernetika*, v. 11, pp. 91–98, 1975.)

[33] V.G. Gaitsgori and A.A. Pervozvanskii, *Theory of Suboptimal Decisions*, Kluwer Academic Publishers, 1988.

[34] R. Hassin, and M. Haviv, "Mean passage times and nearly uncoupled Markov chains," *SIAM Journal of Discrete Mathematics*, v. 5, pp. 386–397, 1992.

[35] M. Haviv, "An approximation to the stationary distribution of a nearly completely decomposable Markov chain and its error analysis," *SIAM Journal on Algebraic and Discrete Methods*, v. 7, pp. 589–594, 1986.

[36] M. Haviv, "Aggregation/disaggregation methods for computing the stationary distribution of a Markov chain," *SIAM Journal on Numerical Analysis*, v. 24, pp. 952–966, 1987.

[37] M. Haviv, "More on the Rayleigh-Ritz refinement technique for nearly uncoupled matrices," *SIAM Journal of Matrix Analysis and Application*, v. 10, pp. 287–293, 1989.

[38] M. Haviv, "An aggregation/disaggregation algorithm for computing the stationary distribution of a large Markov chain," Communications in Statistics - Stochastic Models, v. 8, pp. 565–575, 1992.

[39] M. Haviv and Y. Ritov, "Series expansions for stochastic matrices," unpublished manuscript, Department of Statistics, The Hebrew University of Jerusalem, 1989.

[40] M. Haviv and Y. Ritov, "On series expansions for stochastic matrices," *SIAM Journal on Matrix Analysis and Applications*, v. 14, pp. 670–677, 1993.

[41] M. Haviv and L. Van der Heyden, "Perturbation bounds for the stationary probabilities of a finite Markov chain," *Advances in Applied Probability*, v. 16, pp. 804–818, 1984.

[42] O. Hernandez-Lerma and J.B. Lasserre, *Discrete- time Markov control processes: basic optimality criteria*, Springer-Verlag, New York, 1996.

[43] A. Hordijk, R. Dekker, and L.C.M. Kallenberg, "Sensitivity analysis in discounted Markovian decision problems", *OR Spectrum*, v. 7, pp. 143–151, 1985.

[44] R.A. Howard, *Dynamic programming and Markov processes*, Cambridge, MA: MIT Press, 1960.

[45] Y. Huang and A.F. Veinott, Jr., "Markov branching decision chains with interest-rate-dependent rewards", *Probability in the Engineering and Information Sciences*, v. 9, pp. 99–121, 1995.

[46] R.G. Jeroslow, "Asymptotic Linear Programming", *Oper. Res.*, v. 21, pp. 1128–1141, 1973.

[47] R.G. Jeroslow, "Linear Programs Dependent on a Single Parameter", *Disc. Math.*, v. 6, pp. 119–140, 1973.

[48] T. Kato, *Perturbation theory for linear operators*, Springer-Verlag, Berlin, 1966.

[49] L. C. M. Kallenberg, *Linear programming and finite Markovian control problems*, Mathematical Centre Tracts 148, Amsterdam, 1983.

[50] L. C. M. Kallenberg, "Survey of linear programming for standard and nonstandard Markovian control problems, Part I: Theory", *ZOR - Methods and Models in Operations Research*, v. 40, pp. 1–42, 1994.

[51] J.G. Kemeny and J.L. Snell, *Finite Markov Chains,* Von Nostrand, New York, 1960.

[52] V.S. Korolyuk and A.F. Turbin, *Mathematical foundations of the state lumping of large systems*, Naukova Dumka, Kiev, 1978, (in Russian), translated by Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[53] B.F. Lamond, "A generalized inverse method for asymptotic linear programming", *Math. Programming*, v. 43, pp. 71–86, 1989.

[54] B.F. Lamond, "An efficient basis update for asymptotic linear programming", *Lin. Alg. Appl.*, v. 184, pp. 83–102, 1993.

[55] B. F. Lamond and M. L. Puterman, "Generalized inverses in discrete time Markov decision processes", *SIAM J. Matrix Anal. Appl.*, v. 10, pp. 118–134, 1989.

[56] C.D. Meyer, "The role of the group generalized inverse in the theory of finite Markov chains," *SIAM Review*, v. 17, pp. 443–464, 1975.

[57] B. L. Miller and A. F. Veinott, Jr., "Discrete dynamic programming with a small interest rate", *Ann. Math. Stat.* v. 40, pp. 366–370, 1969.

[58] A.A. Pervozvanski and V.G. Gaitsgori, *Theory of suboptimal decisions*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988, (Translation from the Russian original: *Decomposition, aggregation and approximate optimization*, Nauka, Moscow, 1979.)

[59] A.A. Pervozvanskii and I.N. Smirnov, "Stationary-state evaluation for a complex system with slowly varying couplings", *Cybernetics*, v. 10, pp. 603–611, 1974. (Translation of Russian original in *Kibernetika,* v. 10, pp. 45–51, 1974.)

[60] R.G. Phillips and P.V. Kokotovic, "A singular perturbation approach to modeling and control of Markov chains", *IEEE Trans. Auto. Contr.*, AC-26, no. 5, pp. 1087–1094, 1981.

[61] M. L. Puterman, *Markov decision processes*, John Wiley & Sons, New York, 1994.

[62] J.P. Quadrat, "Optimal control of perturbed Markov chains: the multitime scale case", in *Singular perturbations in systems and control*, ed. M.D. Ardema, CISM Courses and Lectures no. 280, Springer-Verlag, Wien - New York, 1983.

[63] J.R. Rohlicek and A.S. Willsky, "The reduction of Markov generators: An algorithm exposing the role of transient states", *JACM*, v. 35, pp. 675–696, 1988.

[64] U. Rothblum, "Resolvent expansions of matrices and applications", *Linear Algebra Appl.,* v. 38, pp. 33–49, 1981.

[65] P.J. Schweitzer, "Perturbation theory and finite Markov chains," *J. Appl. Probability*, v. 5, pp. 401–413, 1968.

[66] P.J. Schweitzer, "Perturbation series expansion of nearly completely-decomposable Markov chains," Working Paper Series No. 8122, The Graduate School of Management, The University of Rochester, 1981.

[67] P.J. Schweitzer, "Aggregation methods for large Markov chains," in *Mathematical Computer Performance and Reliability*, by G. Iazeolla, P.J. Courtios and A. Hordijk (editors), Elsevier Science Publishers B.V. (North Holland), pp. 275–286, 1984.

[68] P.J. Schweitzer, "Perturbation series expansion of nearly completely-decomposable Markov chains," in *Teletraffic Analysis and Computer Performance Evaluation,* by O.J. Boxma, J.W. Cohen and H.C. Tijms (editors), pp. 319–328, Elsevier Science Publishers B.V. (North Holland), 1986.

[69] P.J. Schweitzer, "A survey of aggregation-disaggregation in large Markov chains," in *Proceedings of the First International Workshop on Numerical Solution for Markov Chains* by W.J. Stewart (editor), pp. 63–87, 1991.

[70] H.A. Simon and A. Ando, "Aggregation of variables in dynamic systems", *Econometrica*, v. 29, pp. 111–138, 1961.

[71] U. Sumita and M. Rieders, "Numerical Comparison for the replacement process approach with the aggregation-disaggregation algorithm for row-continuous Markov chains" in *Proceedings of the First International Workshop on Numerical Solution for Markov Chains* by W.J. Stewart (editor), pp. 287–302, 1991.

[72] Z. U. Syed, "Algorithms for stochastic games and related topics", PhD Thesis in Mathematics, University of Illinois at Chicago, 1999.

[73] H. Vantilborgh, "Aggregation with an error of $O(\varepsilon^2)$", *Journal of the Association for Computing Machinery*, v. 32, pp. 161–190, 1985.

[74] A. F. Veinott, Jr., "Discrete dynamic programming with sensitive discount optimality criteria", *Ann. Math. Stat.* v. 40, pp. 1635–1660, 1969.

[75] A. F. Veinott, Jr., "Markov decision chains", in *Studies in Optimization*, eds. G. B. Dantzig and B. C. Eaves, pp. 124–159, 1974.

[76] G.G. Yin and Q. Zhang, "Continuous-time Markov chains and applications: A singular perturbation approach", Series: Applications of Mathematics, v. 37, Springer-Verlag, New York, 1998.

Konstantin E. Avrachenkov
INRIA Sophia Antipolis
2004 route des Lucioles, B.P.93, 06902, France
k.avrachenkov@sophia.inria.fr

Jerzy Filar
Department of Mathematics
The University of South Australia
The Levels, South Australia 5095, Australia
jerzy.filar@unisa.edu.au

Moshe Haviv
Department of Statistics
The Hebrew University, 91905 Jerusalem, Israel
and Department of Econometrics, The University of Sydney
Sydney NSW 2006, Australia
haviv@mscc.huji.ac.il

II Infinite State Models

4 AVERAGE REWARD OPTIMIZATION THEORY FOR DENUMERABLE STATE SPACES

Linn I. Sennott

4.1 INTRODUCTION

In this chapter we deal with certain aspects of average reward optimality. It is assumed that the state space $\mathbb{X}$ is denumerably infinite, and that for each $x \in \mathbb{X}$, the set $\mathbb{A}(x)$ of available actions is finite. It is possible to extend the theory to compact action sets, but at the expense of increased mathematical complexity. Finite action sets are sufficient for digitally implemented controls, and so we restrict our attention to this case.

For initial state x, the quantity $W(x)$ is the best possible limiting expected average reward per unit time (*average reward*, for short). This is an appropriate measure of the largest expected reward per unit time that can possibly be achieved far into the future, neglecting short-term behavior. Many interesting applications have the property that the average reward is independent of the initial state, i.e. $W(x)$ is a constant.

This chapter develops a theory to guarantee the existence of a stationary policy ϕ and finite constant W such that

$$W(x) = w(x, \phi) \equiv W, \qquad x \in \mathbb{X}. \tag{4.1}$$

Such a policy is an *average reward optimal* stationary policy. In this chapter a stationary policy means a nonrandomized (pure) stationary policy. Implementing such a policy requires the controller to know only the current state x of the system. Table look-up may then determine the fixed action $\phi(x)$ appropriate in that state.

The development takes place under the assumption that there exists a nonnegative (finite) constant R such that $r(x,a) \leq R$, for all $x \in \mathbb{X}$ and $a \in \mathbb{A}(x)$. Note that rewards may be unbounded below. In some applications, the actual reward is a random quantity. In these cases, $r(x,a)$ is to be interpreted as an expected reward.

In a typical reward maximization setting, it may be possible to incur costs as well as earn rewards. Costs can be built into the system as negative rewards. For example, to minimize over the set $\{5,2,8\}$ of costs, we may calculate $\max\{-5,-2,-8\} = -2$, and then the answer is $-(-2) = 2$. Our framework allows rewards to be unbounded below, thereby handling the common case of costs unbounded above. For example, queueing control problems may involve holding costs that are linear in the number of customers. If the buffers are unlimited (able to hold all arriving customers), then this would entail costs unbounded above. The theory does not allow the controller to earn arbitrarily large positive rewards. This is not a severe limitation in queueing control problems and other applications. For example, assume that the controller earns a unit reward each time a customer is admitted to the system. If the number of customers that can arrive in any slot is bounded, then the assumption will hold. If the distribution on customer batch sizes is unbounded, then we may allow the controller to earn a reward that is a function of the mean batch size.

We may define a new reward structure by subtracting R from the rewards in the original system. By so doing, the optimal policy will not be affected, and it will be the case that all rewards are nonpositive. Let us assume that this has already been done, so that for the rest of the chapter we make the following assumption.

Assumption 4.1 *We have $r(x,a) \leq 0$, for all $x \in \mathbb{X}$ and $a \in \mathbb{A}(x)$.*

Note that to recover the average reward in the original setting, it is only necessary to add R to W.

To motivate our approach, let us consider the situation when $\mathbb{X}$ is finite. In this case, it is well-known that there exist $\beta_0 \in (0,1)$ and a stationary policy ϕ that is discount optimal for $\beta \in (\beta_0, 1)$. Such a policy is called *Blackwell optimal*, and it must also be average optimal. These claims are proved in the chapter by Hordijk and Yushkevich in this volume; also see Sennott [37, Proposition 6.2.3]. Note that in the general case, $W(x)$ may not be constant. To motivate the assumptions to be introduced in Section 3, we give the following result. It was stated in [37, Proposition 6.4.1] for the cost minimization framework, and the proof may be recast into the reward maximization framework.

Proposition 4.1 *Let $\mathbb{X}$ be finite. The following are equivalent:*

(i) *$W(x) \equiv W$, for $x \in \mathbb{X}$.*

(ii) *There exists $z \in \mathbb{X}$ and a finite constant L such that $|V(x,\beta) - V(z,\beta)| \leq L$, for all $x \in \mathbb{X}$ and $\beta \in (0,1)$.*

(iii) *Given $y \in \mathbb{X}$, there exists a finite constant L such that $|V(x,\beta) - V(y,\beta)| \leq L$, for all $x \in \mathbb{X}$ and $\beta \in (0,1)$.*

4.2 COUNTEREXAMPLES

When $\mathbb{X}$ is denumerably infinite, the situation is quite different from that when $\mathbb{X}$ is finite. An average reward optimal stationary policy may not exist. The following example shows that, in fact, an optimal policy of any sort may not exist.

Example 4.1 *The state space has two "layers." The top layer consists of states* $\{1, 2, 3, \ldots\}$ *and the bottom layer of states* $\{1^*, 2^*, 3^*, \ldots\}$. *There is a single action in each bottom state and these are all absorbing, i.e.* $p_{x^*x^*} = 1$. *There are actions a and b in each top state, with* $1 = p_{xx+1}(a) = p_{xx^*}(b)$. *There are no rewards in any top state, and* $r(x^*) = 1 - \frac{1}{x}$.

Beginning in state x, to achieve a positive reward requires that b be eventually chosen. Assume that b is first chosen in state y, where $y \geq x$. *Then from that point on, a reward of* $1 - \frac{1}{y}$ *per unit time is earned. It is clear (and can be proved) that the rewards earned up to this point do not affect the limiting average reward, which is thus* $1 - \frac{1}{y} < 1$. *Clearly,* $W(x) = 1$. *However, no policy achieves an average reward of 1.*

The next example shows that, even if an average cost optimal policy exists, it may be nonstationary.

Example 4.2 *Let* $\mathbb{X} = \{1, 2, 3, \ldots\}$. *There are two actions in each state with* $1 = p_{xx+1}(a) = p_{xx}(b)$. *There is no reward under a and* $r(x, b) = 1 - \frac{1}{x}$. *In words, we may advance to the next higher state and earn nothing, or remain in x and earn* $1 - \frac{1}{x}$. *Let us assume that the process begins in state 1 and operates under a stationary policy. If this policy chooses b for the first time in state x, then it must continue to choose b, and the average reward under this stationary policy is* $1 - \frac{1}{x}$.

However, consider the nonstationary policy π *that operates as follows: Upon first entry into state x, it chooses b a total of x times and then chooses a. The sequence of rewards generated under* π *is*

$$0, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, 0, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}, 0, \ldots \tag{4.2}$$

It may be shown that $w(1, \pi) = 1$, *and hence* π *is average reward optimal.*

These examples appear in Ross [34]. In both examples, it is the case that there exists a stationary policy achieving an average reward that is within ϵ of optimality. If this were always the case, we would probably be satisfied to know that we could produce a stationary policy with any desired degree of closeness to the optimal value. However [34, p. 91], gives an example for which no stationary policy is within ϵ of optimality. Furthermore, Dynkin and Yushkevich [14, Sec. 7.8, Ex. 3] and Feinberg [18] give examples for which there are no Markov policies within ϵ of optimality.

4.3 ASSUMPTIONS FOR VALIDITY OF EQUATION 4.1

The examples show that some assumptions are necessary to achieve the goal of (4.1). One possible form for those assumptions is suggested by the situation

when $\mathbb{X}$ is finite. In this case, as we have observed, there exists a stationary policy that is discount optimal, for every discount factor sufficiently close to 1, and is also average optimal. We may seek to obtain a similar result when $\mathbb{X}$ is denumerable. Since we also want $W(x)$ to be constant, this suggests that the condition in Proposition 4.1 (ii) might be a suitable assumption. It is shown in Ross [34, p. 95] that this will indeed work. [This approach was initiated by Taylor [43] and generalized by Ross [33].] However, it turns out that this condition is too strong for the treatment of many important applications.

The following set of assumptions is based on a generalization of the condition in Proposition 4.1. These assumptions are the reward version of those in Sennott [37, p. 132] and are a slight modification of those in Puterman [31, Section 8.10.2]. A version of these assumptions is also discussed in Arapostathis, et al [2] and Kitaev and Rykov [28]. Keep in mind the fact that quantities that are automatically finite for $\mathbb{X}$ finite may become infinite when $\mathbb{X}$ is infinite. Expectations exist because we are assuming that rewards are nonpositive, but they may be $-\infty$.

Let z be a distinguished state.

Assumption 4.2 *The quantity* $(1-\beta)V(z,\beta)$ *is bounded, for* $\beta \in (0,1)$.

This implies that $V(z,\beta) > -\infty$ and hence we may let $h(x,\beta) \triangleq V(x,\beta) - V(z,\beta)$ without fear of introducing the indeterminate form $-\infty + \infty$, which would not be well-defined.

Assumption 4.3 *There exists a (finite) nonnegative function* M *such that* $h(x,\beta) \geq -M(x)$, *for* $x \in \mathbb{X}$ *and* $\beta \in (0,1)$.

The final assumption is:

Assumption 4.4 *There exists a (finite) nonnegative constant* L *such that* $h(x,\beta) \leq L$, *for* $x \in \mathbb{X}$ *and* $\beta \in (0,1)$.

Note that $h(z,\beta) \equiv 0$ and hence we may always take $M(z) = 0$. As we will see, Assumption 4.2 is related to the requirement that the average reward be finite. Assumptions 4.3 and 4.4 are basically the condition in Proposition 4.1 (ii), but modified to allow the lower bound to be a function, rather than a constant. Section 10 discusses a weaker set of assumptions that allows the upper bound to also be a function. It turns out that for many applications, the upper bound is constant, and this assumption simplifies the theory.

One may also wonder whether the distinguished state z plays a special role in the assumptions. The answer is no. For the cost minimization framework, it is shown in [37, Proposition 7.2.4] that if Assumptions 4.1–4.4 hold for z, then they hold when z is replaced by any other state.

Here is an important lemma.

Lemma 4.1 *Let* ϕ *be a stationary policy. Assume that there exist a (finite) constant* W *and a (finite) function* h, *that is bounded above, such that*

$$W + h(x) \leq T^{\phi}h(x), \qquad x \in \mathbb{X}, \tag{4.3}$$

where

$$T^{\phi}h(x) = r(x,\phi(x)) + \sum_{y\in\mathbb{X}} p_{xy}(\phi(x))h(y). \tag{4.4}$$

Then $w(x,\phi) \geq W$, *for* $x \in \mathbb{X}$.

Proof. Assume that the process starts in state x, operates under ϕ, and let $X_0 = x, X_1, X_2, \ldots$ be the sequence of values. Then from (4.3) it follows that

$$W + h(X_t) \leq T^{\phi}h(X_t), \qquad t \geq 0. \tag{4.5}$$

We now show that $\mathbb{E}_x^{\phi}[h(X_t)] > -\infty$. In fact, we prove by induction on t that the expectation is bounded below by $tW + h(x)$. This is clearly true for $t = 0$. Now assume that it is true for t . Then from (4.4), (4.5), and Assumption 4.1, it follows that $\mathbb{E}_x^{\phi}[h(X_{t+1})|X_t] \geq W + h(X_t)$. Taking the expectation of both sides, using a property of expectation (i.e. $\mathbb{E}(\mathbb{E}[X|Y]) = \mathbb{E}[X]$), together with the induction hypothesis, we find that $\mathbb{E}_x^{\phi}[h(X_{t+1})] \geq W + \mathbb{E}_x^{\phi}[h(X_t)] \geq W + tW + h(x) = (t+1)W + h(x)$. This completes the induction.

Now take the expectation of both sides of (4.5) and rearrange, to obtain

$$\mathbb{E}_x^{\phi}[r(X_t)] \geq W + \mathbb{E}_x^{\phi}[h(X_t)] - \mathbb{E}_x^{\phi}[h(X_{t+1})], \quad t \geq 0. \tag{4.6}$$

What has just been proved assures us that we have not created the indeterminate form $-\infty + \infty$. Add the terms in (4.6), for $t = 0$ to $n-1$, and divide by n to obtain

$$\begin{aligned} \frac{v(x,\phi,n)}{n} &\geq W + \frac{h(x) - \mathbb{E}_x^{\phi}[h(X_n)]}{n} \\ &\geq W + \frac{h(x) - L}{n}. \end{aligned} \tag{4.7}$$

Here L is the upper bound on h. Taking the limit infimum of both sides of (4.7) yields the result. ∎

The proper generalization of the finite state space result deals with sequences of discount factors rather than sufficiently large discount factors. The following definition sets up the basic concepts. (It is independent of the assumptions.) We will be taking subsequences of sequences, and for notational convenience, each time this is done, the subsequence will be indexed by n.

Definition 4.1

(i) *Let* z *be a distinguished state and assume that the function* $h(x,\beta) \triangleq V(x,\beta) - V(z,\beta)$ *involves no indeterminate form. Let* β_n *be a sequence of discount factors converging to 1. (All sequences are assumed to converge*

to 1 from the left.) If there exist a subsequence δ_n and a function h such that

$$\lim_{n\to\infty} h(x,\delta_n) = h(x), \qquad x \in \mathbb{X}, \tag{4.8}$$

then h is a limit function (of the sequence $h(-,\beta_n)$).

(ii) *Let $\phi(\beta)$ be a stationary policy realizing the β discount optimality equation, and let $\beta_n \to 1$. Assume that there exist a subsequence δ_n and a stationary policy ϕ such that $\lim_{n\to\infty}\phi(\delta_n) = \phi$. This means that for a given x and sufficiently large n (dependent on x) we have $\phi(\delta_n)(x) = \phi(x)$. Then ϕ is a limit point (of $\phi(\beta_n)$).*

(iii) *Let ϕ be a limit point. The limit function h is associated with ϕ if there exists a sequence β_n such that $\lim_{n\to\infty} h(-,\beta_n) = h$ and $\lim_{n\to\infty}\phi(\beta_n) = \phi$.*

4.4 THE EXISTENCE THEOREM

The following existence theorem is our major result. It is the average reward counterpart of [37, Theorem 7.2.3].

Theorem 4.1 *Assume that Assumptions 4.1–4.4 hold. Then:*

(i) *There exists a finite constant $W \triangleq \lim_{\beta\to 1}(1-\beta)V(x,\beta)$, for $x \in \mathbb{X}$.*

(ii) *There exists a limit function. Any such function h satisfies $-M \le h \le L$ and*

$$\begin{aligned} W + h(x) &\le Th(x) \\ &\triangleq \max_{a\in A(x)}\{r(x,a) + \sum_{y\in\mathbb{X}} p_{xy}(a)h(y)\}, \qquad x \in \mathbb{X}. \end{aligned} \tag{4.9}$$

Let ψ be a stationary policy realizing the maximum in (4.9). Then ψ is average reward optimal with (constant) average reward W and

$$\lim_{n\to\infty} \frac{\mathbb{E}_x^{\psi}[h(X_n)]}{n} = 0, \qquad x \in \mathbb{X}. \tag{4.10}$$

(iii) *Any limit point ϕ is average reward optimal. There exists a limit function associated with ϕ. Any such function h satisfies*

$$W + h(x) \le T^{\phi}h(x), \qquad x \in \mathbb{X}. \tag{4.11}$$

and

$$\lim_{n\to\infty} \frac{\mathbb{E}_x^{\phi}[h(X_n)]}{n} = 0, \qquad x \in \mathbb{X}. \tag{4.12}$$

(iv) *The average reward under any optimal policy is obtained as a limit, rather than a limit infimum.*

Proof. Observe that the theorem encompasses two viewpoints. It says that an optimal stationary policy may be obtained from (4.9), which is constructed by first obtaining a limit function. Or an optimal stationary policy may be obtained as a limit of discount optimal stationary policies. In any case, the average reward under any optimal policy (stationary or not) is obtained as a limit, rather than a limit infimum.

We first prove (ii). Fix a sequence $\beta_n \to 1$. It follows from Assumptions 4.3–4.4 and [37, Proposition B.6] that there exists a limit function of the sequence $h(-, \beta_n)$, and that any such function h satisfies $-M \le h \le L$.

Now fix a limit function h as in (4.8). Using Assumption 4.1 we see that $(1-\delta_n)V(z,\delta_n)$ is a bounded sequence of real numbers. Any such sequence has a convergent subsequence. Hence there exist a subsequence ϵ_n and a (finite) number W such that

$$\lim_{n\to\infty} (1-\epsilon_n)V(z,\epsilon_n) = W. \tag{4.13}$$

Note that $(1-\beta)V(x,\beta) = (1-\beta)h(x,\beta) + (1-\beta)V(z,\beta)$. Let $\beta = \epsilon_n$ and let $n \to \infty$. The last term approaches W. It follows from (4.8) and the finiteness of h that the second term approaches 0. Hence

$$\lim_{n\to\infty} (1-\epsilon_n)V(x,\epsilon_n) = W, \qquad x \in \mathbb{X}. \tag{4.14}$$

The discount optimality equation $V = T_\beta V$ may be rewritten as

$$(1-\beta)V(z,\beta) + h(x,\beta) = T_\beta h(x,\beta), \quad x \in \mathbb{X}. \tag{4.15}$$

Now fix a state x and consider the sequence $\phi(\epsilon_n)(x)$ of discount optimal actions in x. Because the action set $\mathbb{A}(x)$ is finite, it is the case that there exist an action $a(x)$ and a subsequence γ_n (dependent on x) such that $\phi(\gamma_n)(x) \equiv a(x)$. For the fixed state x and $\beta = \gamma_n$, (4.15) becomes

$$(1-\gamma_n)V(z,\gamma_n) + h(x,\gamma_n) = T_{\gamma_n}^{a(x)} h(x,\gamma_n). \tag{4.16}$$

Take the limit supremum of both sides of (4.16) as $n \to \infty$. Use (4.8), (4.13), and the limit supremum version of Fatou's lemma [37, Proposition A.2.1] to obtain

$$\begin{aligned} W + h(x) &\le T^{a(x)} h(x) \\ &\le Th(x). * \end{aligned} \tag{4.17}$$

Because this argument may be repeated for each x, it follows that (4.9) holds.

Now let ψ be a stationary policy realizing the maximum in (4.9). Then (4.3) holds for ψ. To prove that ψ is optimal, let π be an arbitrary policy and fix an initial state x. Then

$$w(x,\psi) \geq W \geq \liminf_{\beta\to 1}(1-\beta)V(x,\beta) \geq \liminf_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \geq w(x,\pi). \tag{4.18}$$

The leftmost inequality follows from Lemma 4.1. The next inequality follows from (4.14) and the definition of the limit infimum. The next inequality follows since $V(-,\beta) \geq v(-,\pi,\beta)$, and the rightmost inequality follows from (4.33) in the Appendix. This proves that ψ is average reward optimal. Moreover by setting $\pi = \psi$ we see that $w(x,\psi) \equiv W$ and hence W is the maximum average reward.

Recall that the whole argument was carried out with respect to the sequence β_n. Given this sequence we obtained a subsequence such that (4.14) holds for the maximum average reward W. This means that given any sequence, there exists a subsequence such that (4.14) holds for the fixed value W. This implies that the limit exists and hence (i) holds.

We prove (iv) and then return to the proof of (4.10). To prove (iv) let π be an arbitrary average reward optimal policy. Note that all that is assumed is that $W \equiv w(x,\pi)$. We have

$$\begin{aligned} W = \lim_{\beta\to 1}(1-\beta)V(x,\beta) &\geq \limsup_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \\ &\geq \liminf_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \geq W. \end{aligned} \tag{4.19}$$

The leftmost equality follows from (i). The rightmost inequality follows from (4.33) and the optimality of π. Hence all the terms in (4.19) are equal to W and it follows that $\lim_{\beta\to 1}(1-\beta)v(x,\pi,\beta)$ exists. Then (iv) follows from Proposition 4.7 in the Appendix.

Let us now prove (4.10). Using the optimality of ψ and (iv) it follows that we may take the limit of both sides of (4.7) to obtain (4.10).

The proof of (iii) is similar to the proof of (ii) and we omit it. (See [37, p. 137].) ∎

No implication concerning the structure of the Markov chain induced by an optimal stationary policy can be drawn from the assumptions. To see that this is the case, consider a process with any desired transition structure whatsoever and with identically 0 rewards. Then the assumptions hold and all policies are optimal.

The cost minimization analog of the assumptions is denoted (SEN) in [37]. A related set (SCH) of assumptions was introduced by Schäl [35]. Problem 7.6 of [37] claims that (SEN) $\Leftrightarrow$ (SCH), and hence these assumptions sets may be shown to be equivalent.

Equation (4.9) is the *average reward optimality inequality* (AROI). The next section explores conditions under which it will be an equality, yielding the *average reward optimality equation* (AROE).

Criteria more sensitive than average optimality are treated in the chapters by Lewis and Puterman and Hordijk and Yushkevich in this volume.

4.5 THE AVERAGE REWARD OPTIMALITY EQUATION

It is possible for the inequality in (4.9) to be strict. Cavazos-Cadena [9] presents an example for which this is the case. However, in "normal situations", (4.9) is an equality. This section gives very weak conditions for the AROE to be valid.

These conditions are closely related to some of the assumptions made in Section 6 of Chapter 8 by Makowski and Shwartz (see also Chapter 9 by Meyn). Our notation is different from theirs. Let G be a nonempty subset of X. Then $\mathcal{R}(x,G)$ is the set of policies π satisfying the following: Beginning in x and following π, the process will enter G at some time $t \geq 1$ with probability 1, and the expected time $m_{xG}(\pi)$ of a first passage from x to G is finite. We let $\mathcal{R}^*(x,G)$ be the subset of $\mathcal{R}(x,G)$ consisting of policies π such that the expected (total) reward $r_{xG}(\pi)$ of the first passage is also finite. If $G = \{y\}$, then $\mathcal{R}(x,G)$ (respectively, $\mathcal{R}^*(x,G)$) is denoted $\mathcal{R}(x,y)$ (respectively, $\mathcal{R}^*(x,y)$).

Proposition 4.2 *Assume that Assumptions 4.1–4.4 hold and let ϕ be a stationary policy realizing the maximum in (4.9). The AROI is an equality at a fixed state x under any of the following conditions:*

(i) *There exists a finite set G such that $\phi \in \mathcal{R}(x,G)$. This also implies that $\phi \in \mathcal{R}^*(x,G)$ and $h(x) = r_{xz}(\phi) - W m_{xz}(\phi) + \mathbb{E}_x^{\phi}[h(X_T)]$, where T is the time of a first passage to G.*

(ii) *We have $\phi \in \mathcal{R}(x,z)$. This also implies that $\phi \in \mathcal{R}^*(x,z)$ and $h(x) = r_{xz}(\phi) - W m_{xz}(\phi)$.*

(iii) *The Markov chain induced by ϕ is positive recurrent at x.*

(iv) *We have $p_{xy}(\phi(x)) > 0$ for only finitely many values of y.*

Proof. We omit the proof of (i). A generalization of (i), with proof, is given in [37, Theorem 7.4.3] for the average cost framework. If we grant the truth of (i), then (ii) follows immediately by setting $G = \{z\}$. Similarly, (iii) follows by setting $G = \{x\}$, and (iv) follows by setting $G = \{y | p_{xy}(\phi(x)) > 0\}$. ∎

Assume that the process operates under an optimal stationary policy determined by (4.9). From Proposition 4.2, we see that the AROI can be strict at x only if the process does not reach any finite set in a finite expected amount of time. A system with this property is unlikely to arise in applications. The impetus for Proposition 4.2 came from Cavazos-Cadena [8].

4.6 A SUFFICIENT CONDITION FOR ASSUMPTION 4.3

This section presents a sufficient condition for Assumption 4.3 to hold and then uses this to show how the assumptions can be verified in an example.

Proposition 4.3 *Assume that Assumption 4.1 holds. Let z be the distinguished state, and assume that $V(z,\beta) > -\infty$, for $\beta \in (0,1)$. Given $x \neq z$, assume that there exists a policy $\pi_x \in \mathcal{R}^*(x,z)$. Then $h(x,\beta) \geq r_{xz}(\pi_x)$, and hence Assumption 4.3 holds for the nonnegative function $M(x) = -r_{xz}(\pi_x)$.*

Proof. If the process begins in state $x \neq z$ and follows π_x, it will reach state z at some time in the future. Let T be a random variable denoting this time. Let the policy π follow π_x until z is reached, and then follow a discount optimal policy $\phi(\beta)$. Then

$$\begin{aligned} V(x,\beta) &\geq v(x,\pi,\beta) \\ &= \mathbb{E}_x^\pi \left[\sum_{t=0}^{T-1} \beta^t r(X_t, A_t) \right] + \mathbb{E}_x^\pi [\beta^T] V(z,\beta) \\ &\geq \mathbb{E}_x^\pi \left[\sum_{t=0}^{T-1} r(X_t, A_t) \right] + V(z,\beta) \\ &= r_{xz}(\pi_x) + V(z,\beta). \end{aligned} \tag{4.20}$$

The validity of the second inequality follows from Assumption 4.1. The result that follows by subtracting $V(z,\beta)$ from both sides. ∎

We now give an example. All the examples take place in discrete time.

Example 4.3 *We have a single server queue, and the service time of a customer is geometrically distributed with success parameter $\mu \in (0,1]$. There are Bernoulli arrival processes $k = 1, 2, \ldots, K$, and process k has parameter $p_k \in (0,1]$.*

The state of the system is the number $x \geq 0$ of customers currently in the system. In each state and each time slot, the action set is $\{0, 1, \ldots, K\}$. Action $k, 1 \leq k \leq K$, allows arrival process k to operate. In this case, with probability p_k, one customer will enter the system, and with probability $1 - p_k$, no customer arrives. Let us adopt the convention that the customer arrives sometime during the slot, so that, if it arrives to an empty buffer, it will enter service at the beginning of the next slot. Choosing action 0 bars the system from new customers for that slot.

There is an increasing holding cost $H(x)$, where $H(0) = 0$. If process k is chosen, then a positive reward $R(k)$ is earned. If action 0 is chosen, then we set $R(0) = 0$. If R is the maximum possible reward, then we may set this up as a reward maximization problem by defining $r(x,k) = -H(x) + R(k) - R, 0 \leq k \leq K$.

Let us argue informally that Assumptions 4.2–4.4 hold, with distinguished state 0. Let ϕ be the (stationary) policy that always chooses 0. In this case, no new customers can enter the system and eventually the process will reach state 0 and remain there. In state 0 the reward is $-R$, and hence $v(0,\phi,\beta) = -\frac{R}{1-\beta}$. Then $0 \geq (1-\beta)V(0,\beta) \geq (1-\beta)v(0,\phi,\beta) = -R$, and Assumption 4.2 holds.

For $x \geq 1$, it is easy to see that $v(x,\phi,\beta) > -\infty$, and hence $V(x,\beta)$ is finite. Moreover, $m_{x0}(\phi) = \frac{x}{\mu}$ and $r_{x0}(\phi) = -[H(x) + H(x-1) + \ldots + H(1) + xR]/\mu$, and hence $\phi \in \mathcal{R}^*(x,0)$. It follows from Proposition 4.3 that Assumption 4.3 holds.

We may prove by induction on the finite horizon that $V(0,\beta) \geq V(x,\beta)$. This is intuitively clear since every state has the same choices and the holding costs are increasing in the state. Granted this, we see that Assumption 4.4 holds with $L = 0$.

It then follows from Theorem 4.1 that an average reward optimal stationary policy may be determined from the AROI. Since Proposition 4.2 (iv) holds, the AROE is valid. If the holding cost is unbounded, we conjecture that an optimal policy will choose 0 for sufficiently large x.

Example 4.4 *This is a modification of Example 4.3. The arrival streams are as in Example 4.3. However, each stream has its own queue and server, with stream k being served at geometric rate μ_k.*

The state of the system is the vector $\mathbf{x}$ of buffer occupancies. In each state and each time slot, an allowable action is a vector $\mathbf{a}$, such that a_k equals 1 if the kth process is activated and 0 if it is not.

There is an increasing holding cost $H_k(x_k)$ on the content of buffer k, with zero cost for an empty buffer. Let $H(\mathbf{x}) = \sum_k H_k(x_k)$. If $a_k = 1$, then a positive reward R_k is earned. Let $R = \sum_k R_k$ be the maximum possible reward, and let $R(\mathbf{a}) = \sum_k R_k a_k$ be the reward earned under action $\mathbf{a}$. We may set this up as a reward maximization problem by defining $r(\mathbf{x}, \mathbf{a}) = -H(\mathbf{x}) + R(\mathbf{a}) - R$.

Let the distinguished state be $\mathbf{0}$, and let ϕ be the (stationary) policy that always chooses $\mathbf{a} = \mathbf{0}$. In this case, no new customers can enter the system and eventually the process will enter state $\mathbf{0}$ and remain there. The argument showing that Assumptions 4.2–4.3 hold is similar to that for Example 4.3.

It is intuitively clear that the best situation for the system is to be in state $\mathbf{0}$. This is so because there are the same actions in each state, and the holding cost is minimized in $\mathbf{0}$. So we have $V(\mathbf{0}, \beta) \geq V(\mathbf{x}, \beta)$. A formal induction proof on the finite horizon may be given to justify this claim. Granted this, we see that Assumption 4.4 holds with $L = 0$.

4.7 A SUFFICIENT CONDITION FOR ASSUMPTIONS 4.2–4.3

In Example 4.3, the presence of the non-admit action aided us in verifying the assumptions. What if the controller must always choose among the active arrival streams? In this section, we give a sufficient condition for Assumptions 4.2 and 4.3 to hold, and then show how this condition is useful in the modified version of Example 4.3.

Definition 4.2 *Let z be a distinguished state. A z standard policy is a (randomized) stationary policy ϕ such that $m_{xz}(\phi)$ and $r_{xz}(\phi)$ are finite, for all $x \in \mathbb{X}$.*

The implications of Definition 4.2 are: (1) The Markov chain induced by ϕ has a single positive recurrent class S containing z; (2) The expected time and reward to reach the class from any $x \notin S$ are finite; (3) The average reward is a constant $w(\phi)$, for all initial states, and

$$w(\phi) = \sum_{y \in S} q(y, \phi) r(y, \phi(y)), \tag{4.21}$$

where $q(-, \phi)$ is the steady state distribution. If ϕ is randomized, then the last term in (4.21) is the expected reward associated with state y. For details on this concept and other results concerning Markov chains used throughout the chapter, see [37, Appendix C].

This concept provides a sufficient condition for Assumptions 4.2 and 4.3 to hold.

Proposition 4.4 *Assume that Assumption 4.1 holds, and let z be the distinguished state. If there exists a z standard policy, then Assumptions 4.2–4.3 hold.*

Proof. Let ϕ be a z standard policy. Then

$$w(\phi) = \sum_{y \in S} q(y, \phi)\, \mathbb{E}_y^{\phi}[r(X_n)], \qquad n \geq 0. \tag{4.22}$$

This is easily proved by induction on n. It holds for $n = 0$ by (4.21). For full details, see [37, p. 299].

Multiplying both sides of (4.22) by β^n and summing over n yields

$$w(\phi) = (1 - \beta) \sum_{y \in S} q(y, \phi) v(y, \phi, \beta). \tag{4.23}$$

It follows from (4.23) that $w(\phi) \leq (1 - \beta) q(z, \phi) v(z, \phi, \beta)$. Hence

$$w(\phi) q^{-1}(z, \phi) \leq (1 - \beta) v(z, \phi, \beta) \leq (1 - \beta) V(z, \beta) \leq 0. \tag{4.24}$$

and Assumption 4.2 holds. The quantity on the left of (4.24) equals $r_{zz}(\phi)$.

The validity of Assumption 4.3 now follows from Proposition 4.3. ∎

Example 4.5 *This is Example 4.3, modified to remove the 0 action, so that the controller must choose among the active streams. We assume that $p_1 < \mu < 1$, and note that $c \triangleq [(1 - \mu)p_1]/[\mu(1 - p_1)] < 1$. Assume that*

$$\sum_{x=1}^{\infty} H(x) c^x < \infty. \tag{4.25}$$

Assumptions 4.1 and 4.4 will hold as before. Let ϕ be the stationary policy that always chooses the first stream. If we can show that ϕ is 0 standard, then the validity of Assumptions 4.2–4.3 will follow from Proposition 4.4.

It is clear that ϕ induces an irreducible Markov chain on $\mathbb{X}$. If this chain is positive recurrent with finite average reward, then it will follow that ϕ is 0 standard. It is shown in [37, Proposition 8.5.1] that

$$q(0, \phi) = 1 - \frac{p_1}{\mu}, \qquad q(x, \phi) = \left(\frac{q(0, \phi)}{1 - \mu} \right) c^x, \quad x \geq 1. \tag{4.26}$$

Then $w(\phi) = R_1 - R - \left(\frac{q(0,\phi)}{1-\mu} \right) \sum H(x) c^x > -\infty$, where the finiteness follows from (4.25). Hence the assumptions hold.

If the arrival streams are not Bernoulli, various Lyapunov techniques (e.g. [37, Appendix C]) may be used to prove that ϕ is 0 standard. In particular, if the mean of the first arrival stream is less than $\mu (< 1)$, the $(n+1)$th moment of this stream is finite, and $H(x)$ is bounded by a polynomial of degree n, then the result will still hold.

4.8 VERIFYING ASSUMPTION 4.4

In Examples 4.3, 4.4, and 4.5, we verified Assumption 4.4 by finding a state that maximized the value of V. Choosing that state as the distinguished state then verified Assumption 4.4 with $L = 0$. Here is a generalization of this technique.

Proposition 4.5 *Assume that Assumption 4.1 holds, and that Assumptions 4.2–4.3 hold for distinguished state z. Assume:*

(i) *There exists a finite set G of states, containing z, such that, for $x \notin G$ and $\beta \in (0,1)$, there exists $y \in G$ with $V(y,\beta) \geq V(x,\beta)$.*

(ii) *Given $y \in G - \{z\}$, there exists a policy $\pi_y \in \mathcal{R}^*(z,y)$.*

Then Assumption 4.4 holds.

Proof. We claim that the nonnegative quantity

$$L \triangleq \max_{y \in G - \{z\}} [-r_{zy}(\pi_y)] < \infty \tag{4.27}$$

will work to verify Assumption 4.4. To see this, fix $x \neq z$ and $\beta \in (0,1)$. If $x \notin G$, let y be as in (i). Whereas, if $x \in G$, let $y = x$. Let the process start in z and follow the policy π_y until y is reached, and then follow a β discount optimal stationary policy. Using similar reasoning to that in (4.20), we have

$$\begin{aligned} V(z,\beta) &\geq r_{zy}(\pi_y) + V(y,\beta) \\ &\geq -L + V(x,\beta). \end{aligned} \tag{4.28}$$

and hence Assumption 4.4 holds. ∎

It is possible to generalize Proposition 4.5 to allow G to be infinite, if we can show that L defined in (4.27) as a supremum, is finite. Here is an example using Proposition 4.5.

Example 4.6 *This is a modification of Example 4.4. In this case, if $x_k = 0$, then we must set $a_k = 1$. That is, if a buffer is empty, then we must activate the arrival process for that buffer. Otherwise, the model remains the same. Notice that the allowable set of actions now depends on the state.*

Let ϕ be the stationary policy that chooses $a_k = 0$ when $x_k \geq 1$. When a buffer is empty, it must admit, but as soon as a customer enters the system and begins service, arrivals are rejected until that service is finished. Hence each buffer contains either 0 or 1 customer. Indeed, the positive recurrent class is $S = \{\mathbf{x} | x_k = 0,1\}$. It is readily seen that ϕ is $\mathbf{0}$ standard, and hence by Proposition 4.4, it follows that Assumptions 4.2–4.3 hold.

A bit of thought convinces us that V takes on its maximum value in the finite set S. The reason is that, for any state $\mathbf{x}$ with $x_1 > 1$, the controller is in a better position as x_1 decreases to 1, while holding the other coordinates constant. Once this has been done, the same reasoning can be given for x_2, etc. until we reach a state in S. We may set $\pi_{\mathbf{y}} = \phi$, for $\mathbf{y} \in S - \{\mathbf{0}\}$. Hence it follows from Proposition 4.5 that Assumption 4.4 holds.

4.9 A STRONGER SET OF ASSUMPTIONS

This section presents a sufficient "non-structural" set of conditions for the assumptions.

Proposition 4.6 *Assume that Assumption 4.1 holds, and let z be a distinguished state. Assume:*

(i) *There exists a z standard policy ϕ with positive recurrent class S.*

(ii) *There exists $\epsilon > 0$ such that $G = \{x | r(x,a) \geq w(\phi) - \epsilon$ for some $a\}$ is a finite set.*

(iii) *Given $y \in G - S$, there exists a policy $\pi_y \in \mathcal{R}^*(z,y)$.*

Then Assumptions 4.2–4.4 hold. Moreover:

1. *The AROE holds.*

2. *The Markov chain induced by an optimal stationary policy ψ has at least one positive recurrent state in the set $G(\psi) = \{x | r(x,\psi(x)) \geq W - \epsilon\}$. Let $S(\psi)$ be the set of positive recurrent states. The number of positive recurrent classes making up $S(\psi)$ cannot exceed $|G(\psi)|$, and there are no null recurrent classes.*

3. *If ψ realizes the maximum in the AROE, then $\psi \in \mathcal{R}^*(x, G(\psi) \cap S(\psi))$, for all x. Hence, if $S(\psi)$ consists of a single class, then ψ is y standard, for $y \in S(\psi)$.*

Proof. It follows from (i) and Proposition 4.4 that Assumptions 4.2–4.3 hold. Suppose we can show

(*): V takes on its maximum in the finite set G given in (ii).

We may then apply Proposition 4.5 to show that Assumption 4.4 holds. (Add z to the set G if necessary. For $y \in (G \cap S) - \{z\}$, set $\pi_y = \phi$.)

To show (*), let us fix $\beta \in (0,1)$ and suppress it in our notation. We first let σ be any (randomized) stationary policy and choose $x \notin G$. Let T be the time of a first passage from x to G, under σ. For notational purposes, let $\alpha = \frac{w(\phi)-\epsilon}{1-\beta}$. Then we have

$$v(x,\sigma) \leq \mathbb{E}_x^\sigma [\alpha I(T = \infty) + \{\alpha(1-\beta^T) + \beta^T v(X_T, \sigma)\} I(T < \infty)]. \quad (4.29)$$

This follows since $w(\phi) - \epsilon$ is an upper bound on the rewards outside of G.

Since the expression on the right of (4.23) is a convex combination, it follows from (4.23) that there exists $y \in S$ such that $w(\phi) \leq (1-\beta)v(y,\phi)$. We claim that

$$w(\phi) \leq (1-\beta)v(i,\phi), \text{ for some } i \in G. \quad (4.30)$$

This is proved by contradiction. Assume that (4.30) fails. Use (4.29) with $\sigma = \phi$ and $x = y$ to obtain a contradiction. (Note that $I(T = \infty) = 0$.)

Since G is finite, it follows that there exists $j \in G$ that maximizes the value of V for initial states in G. Then from (4.30) it follows that

$$\frac{w(\phi)}{1-\beta} \leq V(j). \tag{4.31}$$

Let us now begin the process in $x \notin G$ and operate under the discount optimal stationary policy $\phi(\beta)$. Applying (4.29) with $\sigma = \phi(\beta)$ and using (4.31) yields $V(x) \leq V(j)$. This shows that V takes on its maximum in G, and proves that the assumptions hold.

Proposition 4.6 is the average reward version of [37, Theorem 7.5.6]. The rather lengthy argument for the validity of (1)–(3) is given there for the average cost case, and we omit the proof. ∎

The following remarks discuss background and also some subtle ramifications of the conditions in Proposition 4.6.

Remark 4.1 *Proposition 4.6 (i)–(iii) is a version of an assumption set originally developed by Borkar [3, 4, 5, 6] and denoted (BOR) in [37]. The proof that (BOR) ⇒ (SEN) originally appeared in Cavazos-Cadena and Sennott [11]. Assume that (BOR) holds, and let ψ be an optimal stationary policy. From (2) it follows that the Markov chain induced by ψ has at least one positive recurrent state in $G(\psi)$, and no null recurrent classes. Thus all non-positive recurrent states must be transient. It is shown in Sennott [36] that the probability of reaching a positive recurrent class from any transient state is 1. However, an example is given showing that the expected time of such a first passage may be infinite. Of course, by (3) this behavior cannot occur if ψ realizes the AROE.*

Remark 4.2 *There is a slightly weaker assumption set due to Stidham and Weber [39], and denoted (WS) in [37]. The only change is that (ii) appears without the ϵ. It is the case that (BOR) ⇒ (WS) ⇒ (SEN). Under (WS) it is still possible to prove the corresponding version of (2) which has the ϵ omitted. However, interestingly enough, (3) may fail. Problem 7.7 in [37] asks the reader for such a construction, and this is available from the author.*

The following corollaries of Proposition 4.6 are due to Cavazos-Cadena [7, 8, 10]. In each case, it is easily shown that conditions (i-iii) of Proposition 4.6 hold. These results are very useful when the rewards are unbounded outside of finite sets.

Corollary 4.1 *Assume that Assumption 4.1 holds, and let z be a distinguished state. Assume:*

(i) *There exists a z standard policy ϕ with positive recurrent class S.*

(ii) *Given a positive number U, the set $G(U) = \{x | r(x,a) \geq -U \text{ for some } a\}$ is finite.*

(iii) *Given $y \in \mathbb{X} - S$, there exists a policy $\pi_y \in \mathcal{R}^*(z,y)$.*

Then the conclusions of Proposition 4.6 hold.

Corollary 4.2 *Assume that Assumption 4.1 holds. Assume that there exists a standard policy ϕ such that $S = \mathbb{X}$. If, for each positive number U, the set $G(U) = \{x | r(x,a) \geq -U \text{ for some } a\}$ is finite, then the conclusions of Proposition 4.6 hold.*

The last example shows how Corollary 4.2 may be applied.

Example 4.7 *Consider a polling system, with stations $1, 2, \ldots, K$ arranged in a ring. We will be dealing with a number of distributions, and for each one, we assume that the parameter of that distribution lies in the interval $(0, 1]$. Each station has an infinite buffer and the arrival stream to station k is Bernoulli with parameter p_k. The service time of a customer at station k follows a geometric distribution with rate μ_k. The server travels around the ring counterclockwise from station 1 to station 2, etc., and finally back to station 1. The walking time for the server to get from station $k-1$ to station k is geometrically distributed with rate ω_k. Note that station 0 is station K. The set-up time at station k (which occurs after a walk terminating at k) is geometrically distributed with rate δ_k. The arrival processes, service times, walking times, and set-up times are all independent.*

The state of the system is a vector $(\mathbf{x}, k, z)$. Here $\mathbf{x}$ is the K-dimensional vector of buffer occupancies. The quantity k indicates the number of the station currently involved, and $z = 0, 1, 2$ indicates the condition of the server. Here $z = 0$ means that the server is already set-up at k (and ready to serve customers if the buffer is nonempty); $z = 1$ means that the server is setting up at k; and $z = 2$ means that the server is walking from $k-1$ to k.

A choice of action is available only when $z = 0$ or 1. The action set is $A = \{a, b\}$, where $a =$ remain at the present station, and $b =$ initiate a walk. We are assuming that if the server is walking to a station, then the walk must be completed. However, when the station is reached and set-up is begun, it may be aborted at any time.

A holding cost of $H_k x$ is incurred on a buffer content of x at station k, where $H_k > 0$. The objective is to minimize the expected average holding cost. This may be cast as a reward maximization problem by setting the reward equal to $-\sum H_k x_k$.

We assume that

$$\sum_{k=1}^{K} \frac{p_k}{\mu_k} < 1. \tag{4.32}$$

This is the condition for stability of the polling system under a stationary policy known as exhaustive service, denoted ϕ. This policy operates as follows. If the server arrives to an empty station, it immediately initiates a walk to the next station. If it arrives to a station with customers, it sets up and serves customers at that station until the buffer empties, and it then walks to the next station. Note that when the system is empty the server will continually cycle until a customer enters the system.

It is easy to see that ϕ induces an irreducible Markov chain on $\mathbb{X}$. Moreover, it may be shown that the chain is positive recurrent with finite average

reward. It then follows immediately from Corollary 4.2 that the conclusions of Proposition 4.6 hold.

Most of the work on polling systems has been done for continuous time systems. Under the assumptions of Poisson arrivals and exponential service times, the stability of the exhaustive service policy under (4.32) was derived heuristically by Takagi [40] and rigorously by Altman et al. [1] and Georgiadis and Szpankowski [20]. See also Fricher and Jaibi [19]. Takagi [41, 42] are useful survey articles containing many references.

4.10 WEAKENING THE ASSUMPTIONS

It is possible to weaken the assumptions by allowing the constant L in Assumption 4.4 to be a function. This necessitates several additional assumptions. This development, for the cost minimization framework, is given in [37, Sec. 7.7]. The resulting set of assumptions is denoted (H). They are related to a line of development due to Hordijk [25, 26] and also to Hu [27]. Also see Spieksma [38]. Example 7.7.4 of [37] is a priority queueing system satisfying (H) but for which (SEN) may not hold.

4.11 APPENDIX

For a given policy π and initial state x, let $w^*(x,\pi)$ be the limit supremum of the expected average rewards. That is, $w^*(x,\pi)$ is defined as in (–) but with lim inf replaced by lim sup. The following result provides a crucial link between the discounted value function under π and the lim inf and lim sup expected average rewards under π.

Proposition 4.7 *For any policy π and initial state x we have*

$$w(x,\pi) \le \liminf_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \le \limsup_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \le w^*(x,\pi). \tag{4.33}$$

For any $x \in \mathbb{X}$ the following are equivalent:

(i) *All the terms in (4.33) are equal and finite.*

(ii) *$w(x,\pi) = w^*(x,\pi) > -\infty$, and hence the quantity in Chapter 0 (0.3) is obtained as a limit.*

(iii) *$\lim_{\beta\to 1}(1-\beta)v(x,\pi,\beta)$ exists and is finite.*

Proof. This result is stated for the cost minimization case as Proposition 6.1.1 in [37] and a complete proof appears in Sec. A.4 of [37]. The statement of (4.33) for the continuous case appears in Widder [45]. Granting (4.33), it is easy to see that the only non-trivial implication is (iii) $\Rightarrow$ (ii). The original proof of this is due to Karamata (see Titchmarsch [44]). This proof is adapted and explicated in [37]. ∎

4.12 BIBLIOGRAPHIC NOTES

Important early work was done by Taylor [43], Derman [12, 13], Ross [33], and Dynkin and Yushkevich [14]. Stronger assumptions than those in this chapter were developed by Federgruen and others [15, 16, 17], Lippman [29] and Wijngaard [46].

A line of development has extended some of our results to the general state space case, see Hernández-Lerma and others [21, 22, 23, 24, 30]. Also see Ritt and Sennott [32].

Detailed discussions of prior work occur in Arapostathis et al [2], Puterman [31] and Sennott [37].

References

[1] E. Altman, P. Konstantopoulos, and Z. Liu, "Stability, monotonicity and invariant quantities in general polling systems," *Queueing Sys.* **11**, 35–57, 1992.

[2] A. Arapostathis, V. Borkar, E. Fernandez-Gaucherand, M. Ghosh, and S. Marcus,"Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM J. Control Optim.* **31**, 282–344, 1993.

[3] V. Borkar, "On minimum cost per unit time control of Markov chains,"*SIAM J. Control Optim.* **22**, 965–978, 1984.

[4] V. Borkar, "Control of Markov chains with long-run average cost criterion,"in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, edited by W. Fleming and P. L. Lions, Springer-Verlag, New York, 1988.

[5] V. Borkar, "Control of Markov chains with long-run average cost criterion: the dynamic programming equations,"*SIAM J. Control Optim.* **27**, 642–657, 1989.

[6] V. Borkar, *Topics in Controlled Markov Chains*, Pitman Research Notes in Mathematics No. 240, Longman Scientific-Wiley, New York, 1991.

[7] R. Cavazos-Cadena, "Weak conditions for the existence of optimal stationary policies in average Markov decision chains with unbounded costs,"*Kybernetika* **25**, 145–156, 1989.

[8] R. Cavazos-Cadena, "Solution to the optimality equation in a class of Markov decision chains with the average cost criterion,"*Kybernetika* **27**, 23–37, 1991.

[9] R. Cavazos-Cadena, "A counterexample on the optimality equation in Markov decision chains with the average cost criterion,"*Sys. Control Letters* **16**, 387–392, 1991.

[10] R. Cavazos-Cadena, "Recent results on conditions for the existence of average optimal stationary policies,"*Ann. Op. Res.* **28**, 3–27, 1991.

[11] R. Cavazos-Cadena and L. Sennott, "Comparing recent assumptions for the existence of average optimal stationary policies,"*Op. Res. Letters* **11**, 33–37, 1992.

[12] C. Derman, "Denumerable state Markovian decision processes-average cost criterion," *Ann. Math. Stat.* **37**, 1545–1553, 1966.

[13] C. Derman,*Finite State Markovian Decision Processes*, Academic, New York, 1970.

[14] E. Dynkin and A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[15] A. Federgruen and H. Tijms, "The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms," *J. Appl. Prob.* **15**, 356–373, 1978.

[16] A. Federgruen, A. Hordijk, and H. Tijms, "Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion," *Stoc. Proc. Appl.* **9**, 223–235, 1979.

[17] A. Federgruen, P. Schweitzer, and H. Tijms, "Denumerable undiscounted semi-Markov decision processes with unbounded costs," *Math. Op. Res.* **8**, 298–313, 1983.

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

[19] C. Fricker and M. Jaibi, "Monotonicity and stability of periodic polling models," *Queueing Sys.* **15**, 211–238, 1994.

[20] L. Georgiadis and W. Szpankowski, "Stability of token passing rings," *Queueing Sys.* **11**, 7–33, 1992.

[21] O. Hernández-Lerma and J. Lasserre, "Average cost optimal policies for Markov control processes with Borel state space and unbounded costs," *Sys. Control Letters* **15**, 349–356, 1990.

[22] O. Hernández-Lerma, "Average optimality in dynamic programming on Borel spaces—unbounded costs and controls," *Sys. Control Letters* **17**, 237–242, 1991.

[23] O. Hernández-Lerma, "Existence of average optimal policies in Markov control processes with strictly unbounded costs," *Kybernetika* **29**, 1–17, 1993.

[24] O. Hernández-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1996.

[25] A. Hordijk, "Regenerative Markov decision models," *Math. Prog. Study* **6**, 49–72, 1976.

[26] A. Hordijk, *Dynamic Programming and Markov Potential Theory* Second Ed., Mathematisch Centrum Tract 51, Amsterdam, 1977.

[27] Q. Hu, "Discounted and average Markov decision processes with unbounded rewards: new conditions," *J. Math. Anal. Appl.* **171**, 111–124, 1992.

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

[29] S. Lippman, "On dynamic programming with unbounded rewards," *Man. Sci.* **21**, 1225–1233, 1975.

[30] R. Montes-de-Oca and O. Hernandez-Lerma, "Conditions for average optimality in Markov control processes with unbounded costs and controls," *J. Math. Sys. Estimation and Control* **4**, 1–19, 1994.

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

[32] R. Ritt and L. Sennott, "Optimal stationary policies in general state space Markov decision chains with finite action sets," *Math. Op. Res.* **17**, 901–909, 1992.

[33] S. Ross, "Non-discounted denumerable Markovian decision models," *Ann. Math. Stat.* **39**, 412–423, 1968.

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

[35] M. Schäl, "Average optimality in dynamic programming with general state space," *Math. Op. Res.* **18**, 163–172, 1993.

[36] L. Sennott, "The average cost optimality equation and critical number policies," *Prob. Eng. Info. Sci.* **7**, 47–67, 1993.

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

[38] F. Spieksma, *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, Leiden University, 1990.

[39] S. Stidham, Jr. and R. Weber, "Monotonic and insensitive optimal policies for control of queues with undiscounted costs," *Op. Res.* **87**, 611–625, 1989.

[40] H. Takagi, *Analysis of Polling Systems*, MIT, Cambridge, 1986.

[41] H. Takagi, "Queueing analysis of polling models: an update," in *Stochastic Analysis of Computer and Communication Shystems*, edited by H. Takagi, North Holland, New York, 1990.

[42] H. Takagi, "Queueing analysis of polling models: progress in 1990–1994," in *Frontiers in Queueing*, edited by J. Dshalalow. CRC Press, Boca Raton, 1997.

[43] H. Taylor, "Markovian seqential replacement processes," *Ann. Math. Stat.* **36**, 1677–1694, 1965.

[44] E. Titchmarsh, *Theory of Functions*, Second Ed., Oxford University Press, Oxford, 1939.

[45] D. Widder, *The Laplace Transform*, Princeton University Press, Princeton, 1941.

[46] J. Wijngaard, "Existence of average optimal strategies in Markovian decision problems with strictly unbounded costs," in *Dynamic Programming and Its Applications*, edited by M. Puterman, Academic, New York, 1978.

Linn I. Sennott
Department of Mathematics
Illinois State University
Normal, IL 61790-4520, USA
sennott@math.ilstu.edu

5 TOTAL REWARD CRITERIA

Eugene A. Feinberg

Abstract: This chapter deals with total reward criteria. We discuss the existence and structure of optimal and nearly optimal policies and the convergence of value iteration algorithms under the so-called General Convergence Condition. This condition assumes that, for any initial state and for any policy, the expected sum of positive parts of rewards is finite. Positive, negative, and discounted dynamic programming problems are special cases when the General Convergence Condition holds.

5.1 INTRODUCTION

This chapter deals with the total reward criterion. This criterion is natural for finite horizon problems and for infinite horizon problems in which the number of steps is not fixed but it is finite along most trajectories. The examples include discounted criteria, which can be interpreted as problems with geometrically distributed horizon, as well as other problems such as sequential statistical procedures, stopping, search, and optimal selection problems.

The analysis of finite horizon models is usually based on the analysis of optimality equations and optimality operators. The value function satisfies the optimality equation and a Markov policy constructed by value iteration procedures is optimal. If optimality cannot be achieved, one can sequentially construct an ε-optimal Markov policy for an N-horizon problem by backward induction. Such a policy is formed by actions for which the reward operators, when applied to the value function at the corresponding step, are (ε/N)-close to the optimal value.

For infinite horizon problems, we can distinguish two distinct approaches. The first approach deals with the analysis of optimality (also called, dynamic programming) equations and operators. The second approach studies probability distributions on the sets of trajectories. In fact, the most interesting results have been achieved by combining these two approaches.

As was observed in the sixties, certain properties of optimality operators, namely contracting and motonicity properties, imply the existence of optimal or nearly optimal policies within natural classes of policies. These properties also imply the convergence of algorithms. In general, dynamic programming operators may not possess these properties. However, if the one-step reward function is uniformly bounded and the nonnegative discount factor is less then one, then the contracting property holds. If rewards are nonnegative (nonpositive) then the value iteration algorithm, applied to the zero terminal value, forms a monotone and therefore convergent sequence. The mentioned three models are called discounted, positive, and negative respectively.

The first comprehensive results were obtained for these three models. Blackwell [8] Denardo [18], Strauch [61] studied discounted models, Blackwell [9], Ornstein [50], and Strauch [61] studied positive models, and Strauch [61] studied negative models. The results differ significantly from one model to another. To illustrate these differences, let us consider the situation when the state space is countable and there are no additional assumptions such as compactness of action sets. For discounted and positive models, for each initial state the supremum of the expected total rewards over the class of all policies is equal to the corresponding supremum over the class of all stationary policies; Blackwell [8, 9]. However, it is not true for negative models; Example 5.5. For discounted models, for any positive constant ε, there exist stationary ε-optimal policies. Such policies may not exist for positive models; Blackwell [9]. However, for positive models there exist stationary εV-optimal policies also called multiplicatively ε-optimal; Ornstein [50]. There are other significant differences between positive, negative, and discounted programming. The differences are so significant that for a long period of time it was even was not clear how to formulate unified results. As the result, all textbooks on dynamic programming and Markov decision processes, that deal with infinite-horizon models with total rewards, consider only positive, negative, and discounted models and deal with them separately; see e.g. Ross [53] and Puterman [52].

In addition to positive, negative, and discounted models, problems with arbitrary reward functions and without discounting have been considered in the literature for a long period of time. It turned out that the most comprehensive results can be proved when the so-called General Convergence Condition holds. This condition means that the positive part of the reward function satisfies the positive programming assumptions. Positive, negative, and discounted models are particular cases of models satisfying the General Convergence Condition.

Blackwell [7] and Krylov [45] proved the existence of stationary optimal policies for MDPs with finite state and action sets. Dubins and Savage [20] and Hordijk [43] described necessary and sufficient conditions for optimality, so-called conserving and equalizing conditions. Derman and Strauch [19], and Strauch [61] proved that, for a given initial state, any policy can be substituted with an equivalent randomized Markov policy. Krylov [45] and Gikhman and Skorohod [40] showed that nonrandomized policies are as good as randomized policies. Dynkin and Yushkevich [21] proved that if someone randomizes between different policies, the objective function cannot be improved. Feinberg [22, 23] proved that, for a given initial state, (nonrandomized) Markov

policies are as good as general ones; earlier van Hee [70] showed that the supremum of the expected total rewards over the class of Markov policies is equal to the supremum over the class of all policies.

Seminal papers by Blackwell [8, 9] and Strauch [61] dealt with models with Borel state and action spaces. Some of the following papers on total reward MDPs dealt just with countable state spaces. For some results, their extension from countable to uncountable models is a straightforward exercise. For other results, such extensions are either difficult or impossible. In addition, Blackwell and Strauch considered Borel-measurable policies and discovered that ε-optimal policies may not exist but for any initial measure they exist almost everywhere. In order to establish the existence of everywhere ε-optimal policies, one should expand the set of policies to universally measurable or analytically measurable policies. Such extension was introduced by Blackwell, Freedman, and Orkin [11] and it was done in a systematic and comprehensive way in the book by Bertsekas and Shreve [5]. In particular, this book expanded in a natural way almost all results on positive, negative and discounted programming in a way that ε-optimality was established instead of almost sure ε-optimality. The major exception was Ornstein's theorem [50]. It was proved by Ornstein [50] for a countable state space; see also Hordijk [43]. It's extension to Borel models in the sense of almost sure multiplicative nearly-optimality was formulated by Blackwell [9] as an open question. Frid [38] solved this problem (Schäl and Sudderth [59] found a correctable gap in Frid's proof). The natural conjecture is that under more general measurability assumptions, in the spirit of Bertsekas and Shreve [5], almost sure nearly-optimality can be replaced with nearly optimality everywhere in Ornstein's theorem. Blackwell and Ramachandran [12] constructed a counter-example to this conjecture.

For the General Convergence Condition, significantly deeper results are available for countable state models than for Borel state problems. First, three important particular results were discovered by van der Wal: (i) the supremum of the expected total rewards over all policies is equal to the supremum over stationary policies if the action sets are finite; [66, Theorem 2.22] (this result was generalized by Schäl [56] to Borel state models with compact action sets); (ii) extension of Ornstein's theorem to models in which rewards may be nonpositive and in each state, where the value function is nonpositive, a conserving action exists; [68] (this results was generalized by van Dawen and Schäl [65, 64]); (iii) existence of uniformly nearly-optimal Markov policies; [67]. The survey by van der Wal and Wessels [69] describes these and many preceding results.

Feinberg and Sonin [34] generalized Ornstein's [50] theorem to models satisfying the General Convergence Condition. For the long period of time, it had not been clear even how to formulate such results. The first clue is that in the more general formulation the value function of the class of stationary policies should be considered instead of the value function of the class of all policies. The second clue is that, in the definition of multiplicative ε-optimal policies (or in the definition of εV-optimal policies according to another terminology), the value function V should be replaced with an excessive majorant of a value function of the class of stationary policies. The proofs in Feinberg and Sonin [34] are non-trivial and differ from Ornstein's [36] proofs. Feinberg and Sonin [36]

extended these results to non-stationary policies and to more general classes of functions that approximate optimal values.

Feinberg [25, 26] described the structure of uniformly nearly-optimal policies in countable state models satisfying the General Convergence Condition. As mentioned above, stationary policies can be significantly outperformed by non-stationary policies in negative problems; see Example 5.5. It turned out that this example demonstrates the only pathological situation when the value of the class of stationary policies is less than the value of the class of all policies. Consider the set of states at which the value function equals zero and there are no conserving actions (actions at which the optimality operator applied to the value function achieves its maximum). Then there are policies which are uniformly nearly optimal and which are stationary outside of this set; see Theorems 5.20 and 5.21.

Another important feature of the papers by Feinberg and Sonin [36, 25, 26] is that they consider general classes of nonstationary policies and general methods how to deals with nonstationary policies. In particular, the information about the past plays an important role. It is possible to identify two properties of this information: (i) Non-Repeating and (ii) Transitivity Conditions. The Non-Repeating condition implies that randomized policies are as good as non-randomized ones and there exist uniformly nearly optimal policies within any class of policies that satisfies this condition. The Transitivity Condition implies that the model can be transformed into a new model in a way that the class of policies satisfying this condition in the old model becomes the class of stationary policies in the new one.

This paper is a survey of results and methods for models satisfying the General Convergence Condition. We consider countable state MDPs everywhere in this paper, except Section 5.10. We present the theory for countable state MDPs, discuss Borel state MDPs, and discuss open questions, most of which deal with uncountable state spaces. In order to illustrate major concepts and counter-examples, we start our presentation with classical discounted, positive, and negative problems.

5.2 DEFINITIONS OF DISCOUNTED, POSITIVE, NEGATIVE, AND GENERAL CONVERGENT MODELS

We say that an MDP is *discounted* if function r is bounded and there is a constant $\beta \in [0, 1[$, called the discount factor, such that

$$v(x, \pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \beta^t r(x_t, a_t). \tag{5.1}$$

An MDP is called *unbounded discounted* if (5.1) holds and the function r is bounded above, $r(x, a) \le C < \infty$ for all $x \in \mathbb{X}$, $a \in \mathbb{A}(x)$ and for some C.

Discounted and unbounded discounted MDPs can be reduced to an MDP with a discount factor equal to 1. In order to do it, we add an additional state to the state space $\mathbb{X}$. This state has only one action under which it is absorbent and all rewards are equal to zero in this state. This state sometimes is called a grave. The one-step transition probabilities between states in $\mathbb{X}$

become equal to $\beta p_{xy}(a)$ and the transition probability from any state in $\mathbb{X}$ to the new absorbent state becomes $(1-\beta)$. The expected total rewards for all initial distributions on $\mathbb{X}$ in the new MDP are equal to the expected total discounted rewards for the same initial distribution in the original system. In the new MDP, the original constant β can be interpreted as the probability that the system remains alive at the next step if it is alive at the current step. This construction is well-known and it is described in details in Altman's book [2, Section 10.1]. Thus we can and will consider discounted MDPs as a special case of total reward MDPs with the discount factor equal to 1.

An MDP is called *positive* if $r(x,a) \geq 0$ for all $x \in \mathbb{X}$ and $a \in \mathbb{A}(x)$ and $v(x,\pi) < \infty$ for all $x \in \mathbb{X}$ and for all $\pi \in \Pi^R$. An MDP is called *negative* if $r(x,a) \leq 0$ for all $x \in \mathbb{X}$ and for all $a \in \mathbb{A}(x)$.

We indicate by **D**, **UD**, **P**, and **N** when we assume that the MDP is respectively discounted, unbounded discounted, positive, and negative. For an arbitrary MDP we define

$$v_+(x,\pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} r^+(x_t,a_t), \tag{5.2}$$

$$v_-(x,\pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} r^-(x_t,a_t), \tag{5.3}$$

where $c^+ = \max\{c,0\}$ and $c^- = \min\{c,0\}$ for any number c. Let

$$V_+(x) = \sup_{\phi\in\Pi^R} v_+(x,\phi), \qquad V_-(x) = \sup_{\phi\in\Pi^R} v_-(x,\phi).$$

It is well-known that $V_+(x) = \sup_{\phi\in\Pi^S} v_+(x,\phi)$; see Blackwell [9]. The following condition holds for **D**, **UD**, **P**, and **N** MDPs.

General Convergence Condition. $V_+(x) < \infty$ for all $x \in \mathbb{X}$.

The General Convergence Condition is equivalent to the condition that $v_+(x,\pi) < \infty$ for all $x \in \mathbb{X}$ and for all $\pi \in \Pi^R$; see van der Wal [66, Theorem 2.3]. If the General Convergence Condition holds, $v(x,\pi) = v_+(x,\pi) + v_-(x,\pi)$ for all initial states and for all policies. In this paper we always assume the General Convergence Condition. The value of $v(x,\pi)$ does not change if we regroup summands $r(x_t,a_t)$ or change the order of summation. We say that an MDP is *general convergent* (**GC**) if the General Convergence Condition holds.

Let

$$s(x) = \sup_{\phi\in\Pi^S} v(x,\phi)$$

be the value of the class of stationary policies.

We remark that we consider **GC** MDPs in this paper because of two reasons: they cover broad classes of situations and there is a well-developed theory for these MDPs. The natural question is how general the General Convergence Condition is? At first glance, it looks so general that one cannot imagine any reasonable model where it is not satisfied and average rewards per unit time are not applicable because they are equal to zero for all policies. However, this conditions is not satisfied for some gambling and stopping problems; see Schäl [58]. Several approaches have been considered are

possible when the **GC** does not hold. Schäl [56] proved that $V(x) = s(x)$ if either $v_+(x,\pi) < \infty$ or $v_-(x,\pi) > -\infty$ for each initial state x and for each policy x in a Borel state MDP with compact action sets. Feinberg [22] proved that $V(x) = \sup_{\pi\in\Pi^M} v(x,\pi)$ if by definition $v(x,\pi) = -\infty$ when $v_+(x,\pi) = -v_-(x,\pi) = \infty$. Some relatively weak results are available for more interesting approaches when $v(x,\pi)$ is defined as the upper limit of $v_n(x,\pi)$ as $n \to \infty$ or as a limit of $v(x,\pi,\beta)$ as $\beta \nearrow 1$; Schäl [57, 58]. We do not know if the main results of this paper, Theorems 5.10 and 5.21, hold for the latter two definitions of the expected total rewards.

5.3 PROPERTIES OF STRATEGIC MEASURES AND OBJECTIVE FUNCTIONS

Theorem 5.1 ([61, 19, 43]) *Let $\pi^1, \pi^2, \ldots$ be an arbitrary sequence of policies and $\lambda_1, \lambda_2, \ldots$ a sequence of nonnegative numbers summing to 1. Consider a randomized Markov policy π defined by*

$$\pi_t(C \mid y) \triangleq \frac{\sum_{i=1}^{\infty} \lambda_i \mathbb{P}_x^{\pi^i}(x_t = y, a_t \in C)}{\sum_{i=1}^{\infty} \lambda_i \mathbb{P}_x^{\pi^i}(x_t = y)}, \quad t \geq 0,\ y \in \mathbb{X}, \tag{5.4}$$

whenever the denominator in (5.4) is not equal to 0. Then, for all $t \geq 0$, $y \in \mathbb{X}$ and measurable subsets C of $\mathbb{A}(y)$,

$$\mathbb{P}_x^{\pi}(x_t = y, a_t \in C) = \sum_{i=1}^{\infty} \lambda_i \mathbb{P}_x^{\pi^i}(x_t = y, a_t \in C). \tag{5.5}$$

Note that the randomized Markov policy defined in Theorem 5.1 depends on the initial state. Our first observation concerning Theorem 5.1 is that, setting $\lambda_1 = 1$ and $\lambda_i = 0$, $i > 1$, we have that for any given policy π^1 and for any initial state x, there is a randomized Markov policy π such that $\mathbb{P}_x^{\pi}(x_t, a_t) = \mathbb{P}_x^{\pi^1}(x_t, a_t)$, $t = 0, 1, \ldots$. Consequently, for any criterion depending only on such marginal distributions, Markov policies suffice. In particular, we have the following result.

Corollary 5.1 *Given an initial state x, for any policy σ consider a randomized Markov policy π defined in Theorem 5.1 for $\pi^1 = \sigma$ and $\lambda_1 = 1$. Then $v(x,\pi) = v(x,\sigma)$.*

Our second observation is that we can expand the notion of a policy by allowing the decision maker to select policies randomly at epoch 0. For example, we can say that a sequence $\gamma = \{(\lambda_i, \pi^i)\}_{i=1}^{\infty}$, where λ_i are nonnegative numbers with the sum equal to 1 and π^i are policies, is a mixed policy. Then any couple (γ, μ), where γ is a mixed policy and μ is the initial probability distribution on $\mathbb{X}$, defines a *strategic measure* $\mathbb{P}_{\mu}^{\gamma} = \sum_{i=1}^{\infty} \lambda_i \mathbb{P}_{\mu}^{\pi^i}$. Theorem 5.1 shows that mixed policies do not outperform randomized Markov policies. Given a mixed policy γ, formula (6) in Section 1.3 in Dynkin and Yushkevich [21] defines a usual randomized policy σ such that $\mathbb{P}_x^{\sigma} = \mathbb{P}_x^{\gamma}$. This is even a stronger argument that there is no need to deal with mixed policies.

Thus, Theorem 5.1 implies that

$$V(x) = \sup_{\pi \in \Pi^{RM}} v(x, \pi), \qquad x \in \mathbb{X}. \tag{5.6}$$

Our next step is to show that nonrandomized Markov policies are as good randomized Markov ones. It follows from the fact that, for a given initial state or distribution, any strategic measure for a randomized Markov policy can be presented as a convex combination of strategic measures for nonrandomized Markov policies.

We recall that a Markov policy ϕ selects action $\phi_n(x)$ when the system is at state x on epoch n. Sometimes it is convenient to write $\phi(x, n)$ instead $\phi_n(x)$. We shall use both these notations. We observe that the set of all Markov policies Π^M is the set of all functions from $\mathbb{X} \times \mathbb{N}$ to $\mathbb{A}$ such that $\phi(x, n) \in \mathbb{A}(x)$ for all x and n. For $B \in \mathcal{A}$ we denote by $\mathcal{A}(B)$ the σ-field on B which elements belong to $\mathcal{A}$.

Now we introduce a measurable structure on Π^M. We notice that $\Pi^M = \times_{(x,n) \in \mathbb{X} \times \mathbb{N}} \mathbb{A}(x, n)$, where $\mathbb{A}(x, n) = A(x)$. Then we define the σ-field $\mathcal{F}$ on Π^M,

$$\mathcal{F} = \times_{(x,n) \in \mathbb{X} \times \mathbb{N}} \mathcal{A}(\mathbb{A}(x, n)).$$

We observe that $\phi \to \mathbb{P}_x^\phi(C)$ is a measurable function on $(\Pi^M, \mathcal{F})$ for any $C \in \mathcal{H}_\infty$ and for any given initial state x. The main idea of the proof is that we interpret the problem in the following way. The decision-maker selects a Markov policy first and then uses this policy. The initial state is always x. So, we add the point ϕ before each trajectory. A trajectory $x, a_0, x_1, a_1 \dots$ transforms into $\phi, x, a_0, x_1, a_1 \dots$. The set of trajectories H_∞ transforms the set H_∞ into the set $\Pi^M \times H_\infty$. Transition probabilities from a_n to x_{n+1}, given the history $\phi, x_0, a_0, \dots, x_n, a_n$ are defined by $p(dx_{n+1}|x_n, a_n)$, $n = 0, 1, \dots$. Transition probabilities from x_n to a_n, given the history $\phi, x_0, a_0, \dots, x_n$ are defined by $\mathbf{I}\{\phi(x_n, n) = a_n\}$, $n = 0, 1, \dots$. Since the measurability conditions from the Ionesco Tulcea theorem [51, Proposition V.1.1] hold, we have that the mapping $\phi \to \mathbb{P}_x^\phi(C)$ is measurable on $(\Pi^M, \mathcal{F})$ for any $C \in \mathcal{H}_\infty$.

Consider a randomized Markov policy π. We also will use sometimes the notation $\pi(\cdot|x, n)$ instead of $\pi_n(\cdot|x)$. Consider the measure m^π on $(\Pi^M, \mathcal{F})$ defined as the product of measures $\pi(\cdot|x, n)$. This means that for any set of Markov policies $B = \{\phi \in \Pi^M |\ \phi(x, n) \in B(x, n), B(x, n) \in \mathcal{A}(\mathbb{A}(x)), (x, n) \in \mathbb{X} \times \mathbb{A}\}$,

$$m^\pi(B) = \prod_{(x,n) \in \mathbb{X} \times \mathbb{N}} \pi(B(x, n)|x, n).$$

The measure m^π exists and is unique for each Markov policy π. This follows from the general construction of product measures on products of measurable spaces; see e.g. Proposition V.1.2 in Neveu [51]. The following theorem is a particular case of Theorem 5.19. A direct proof of Theorem 5.2 is enclosed in the Appendix.

Theorem 5.2 *For any randomized Markov policy π and for any $x \in \mathbb{X}$*

$$\mathbb{P}_x^\pi(C) = \int_{\Pi^M} \mathbb{P}_x^\phi(C) m^\pi(d\phi), \qquad C \in \mathcal{H}_\infty. \tag{5.7}$$

The General Convergence Condition and Theorem 5.2 yield the following result.

Corollary 5.2 *For any randomized Markov policy π and for any initial state $x \in \mathbb{X}$*

$$v(x,\pi) = \int\limits_{\Pi^M} v(x,\phi) m^{\pi}(d\phi).$$

Corollaries 5.1 and 5.2 imply the following statement.

Corollary 5.3 *Given an initial state x, for any policy σ there exists a Markov policy ϕ such that*

$$v(x,\phi) \geq v(x,\sigma).$$

The latter corollary yields the following result.

Corollary 5.4 $V(x) = \sup\{v(x,\phi) |\ \phi \in \Pi^M\}$ *for all* $x \in \mathbb{X}$.

A nonrandomized policy ϕ is called *semi-Markov* if $\phi(h_n) = \phi(x_0, x_n)$ for all $h_n = x_0, a_0, \dots, x_n$, $n = 1, 2, \dots$. Since the state space $\mathbb{X}$ is discrete, Corollary 5.4 implies the following statement.

Corollary 5.5 *For any positive function $\varepsilon(x)$ on $\mathbb{X}$ there exists a semi-Markov policy ϕ such that $v(x,\phi) \geq V(x) - \varepsilon(x)$ for all $x \in \mathbb{X}$.*

Theorem 5.3 (Optimality Equation) $V(x) = TV(x)$ *for all* $x \in \mathbb{X}$.

Proof. For a Markov policy $\phi = \{\phi_0, \phi_1, \dots\}$ we define a shifted policy $\phi^1 = \{\phi_1, \phi_2, \dots\}$. Then $v(x,\phi) = T^{\phi_0(x)} v(x, \phi^1)$. We fix an arbitrary $x \in \mathbb{X}$. The proof consists of two simple steps.

Step 1 ($V(x) \leq TV(x)$). Consider an arbitrary constant $\varepsilon > 0$. According to Corollary 5.4 there exists a Markov policy ϕ for which $v(x,\phi) \geq V(x) - \varepsilon$. We use obvious monotonicity properties of optimality operators and get $V(x) - \varepsilon \leq v(x,\phi) = T^{\phi_0(x)} v(x,\phi^1) \leq Tv(x,\phi^1) \leq TV(x)$. Since $\varepsilon > 0$ is arbitrary, step 1 is proved.

Step 2 ($V(x) \geq TV(x)$). Again, we consider an arbitrary positive constant ε. Corollary 5.5 implies that for any $\varepsilon > 0$ there exists an ε-optimal semi-Markov policy ϕ. For any $a \in \mathbb{A}(x)$ we consider a nonrandomized policy $\sigma[a,\phi]$ such that

$$\sigma[a,\phi](x_0, a_0, x_1, \dots, x_n) = \begin{cases} a, & \text{if } n = 0 \text{ and } x_0 = x; \\ \phi_0(x_1), & \text{if } n = 1; \\ \phi_{n-1}(x_1, x_n), & \text{if } n > 1. \end{cases}$$

This policy uses action a at epoch 0 in the initial state x and than it switches to the semi-Markov policy ϕ with the initial state x_1 and starting epoch $t = 1$. We have that

$$V(x) \geq v(x, \sigma[a,\phi]) = T^a v(x,\phi) \geq T^a(V - \varepsilon)(x) = T^a V(x) - \varepsilon$$

and $V(x) \geq T^a V(x) - \varepsilon$ for all $a \in \mathbb{A}(x)$ and for all $\varepsilon > 0$. Step 2 is proved. ∎

We recall that an action $a \in \mathbb{A}(x)$ is called *conserving* in state x if $V(x) = T^a V(x)$. A policy that uses only conserving actions in all states is called conserving. Obviously, a stationary optimal policy is conserving. The following simple example shows that a conserving stationary policy may not be optimal.

Example 5.1 Let $\mathbb{X} = \mathbb{A} = \{0, 1\}$, $\mathbb{A}(0) = \mathbb{A}$, and $\mathbb{A}(1) = \{0\}$. State 1 is absorbing with one-step reward equal to 0. If action 0 is selected in state 0, the process remains in state 0 and the reward is not collected. Action 1 moves the process to state 1 and brings the one-step reward equal to 1. The formal definitions are $p(x|x, 0) = p(1|0, 1) = 1$, $r(x, 0) = 0$, and $r(0, 1) = 1$ where $x \in \mathbb{X}$. In this example there are two stationary policies ϕ_a, $a = 0, 1$. Policy ϕ_a selects action a in state 0. Both ϕ_0 and ϕ_1 are conserving. However, $V(0, \phi_0) = 0$ and $V(0, \phi_1) = 1$. ∎

A policy π is called *equalizing* at state x if $\limsup_{n\to\infty} \mathbb{E}_x^\pi V(x_n) \leq 0$. A policy is called equalizing if it is equalizing at all states x. It is easy to see that if a policy is conserving and equalizing then it is optimal. An optimal policy π is conserving and the limit of $\mathbb{E}_x^\pi V(x_n)$ exists and is equal to 0 for all states x in which $V(x) > -\infty$.

A natural question is when conserving policies exist. The Optimality Equation implies that a stationary conserving policy exists if and only if TV(x) is achieved for each $x \in \mathbb{X}$ at some action $a \in \mathbb{A}(x)$.

Consider the following conditions that are essential for the existence of conserving policies.

- (i) A is a metric space and $\mathcal{A}$ is its Borel σ-field;
- (ii) $\mathbb{A}(x)$ is compact;
- (iii) $p(y|x, a)$ is continuous in a and $r(x, a)$ is upper semi-continuous in a.

Lemma 5.1 *If conditions (i-iii) hold for a given $x \in \mathbb{X}$ and f is a bounded above function on $\mathbb{X}$ then $T^a f(x) = Tf(x)$ for some $a \in \mathbb{A}(x)$.*

See Lemma 4.2(i) in Feinberg and Shwartz [32] for the proof. In order to guarantee the existence of conserving policies, we consider the following assumption.

Compactness Assumption Condition (i) holds and Conditions (ii) and (iii) hold for all $x \in \mathbb{X}$.

Lemma 5.1 and the Optimality Equation imply the following statement.

Corollary 5.6 *If the value function $V(x)$ bounded above then the Compactness Assumption implies the existence of stationary conserving policies.*

In some applications, action sets are not compacts but can be approximated by compacts in a way that it is expensive to select actions outside of compact subsets. For example, in some inventory models, the size of an order may not be limited but large orders are expensive. To cover this situation, we consider the following condition.

- (iv) For any positive number N there exist a compact subset $B_N(x)$ of $\mathbb{A}(x)$ such that $r(x,a) \le -N$ for $a \in \mathbb{A}(x) \setminus B_N(x)$.

Lemma 5.2 *If conditions (i,iii,iv) hold for a given $x \in \mathbb{X}$ and f is a bounded above function on $\mathbb{X}$ then $T^a f(x) = Tf(x)$ for some $a \in \mathbb{A}(x)$.*

Proof. If $Tf(x) = -\infty$ then $T^a f(x) = Tf(x)$ for all $a \in \mathbb{A}(x)$. Let $T^{a^*} f(x) > -\infty$ for some $a^* \in \mathbb{A}(x)$. Let $C \ge f(z)$ for all $z \in \mathbb{X}$. We set $N = C - T^{a^*} f(x) + 1$. Lemma 5.1 applied to $B_N(x)$ implies that $T^{a'} f(x) = \sup\{T^a f(x) : a \in B_N(x)\}$ for some $a' \in B_N(x)$. For any $a \in \mathbb{A}(x) \setminus B_N(x)$ we have that $T^a f(x) \le C - N = T^{a^*} f(x) - 1 < Tf(x)$. Thus $T^{a'} f(x) = Tf(x)$. ∎

Pre-Compactness Assumption Condition (i) holds and Conditions (iv) and (iii) hold for all $x \in \mathbb{X}$.

The following statements strengthens Corollary 5.6.

Corollary 5.7 *If the value function $V(x)$ is bounded above then the Pre-Compactness Assumption implies the existence of stationary conserving policies. In particular, the Pre-Compactness Assumption implies the existence of stationary conserving policies for* **D**, **UD**, *and* **N** *MDPs.*

Example 5.2 (There are no conserving policies in positive MDPs satisfying the Compactness Assumption) Let $\mathbb{X} = \{0, 1, \dots\} \cup \{g\}$, $\mathbb{A} = \mathbb{A}(0) = \{1/2, 1/3, \dots, 0\}$, and $\mathbb{A}(x) = \{0\}$ for $x \in \mathbb{X} \setminus \{0\}$. We set $p(g|x,0) = 0$ for $x \in \mathbb{X}$, $r(g,0) = 0$, and $r(x,0) = x - 1$ for $x = 1, 2, \dots$. We also set $r(0,a) = 0$ for all $a \in \mathbb{A}$ and $p(x|0,1/x) = 1/x$, $p(g|0,1/x) = (x-1)/x$. We have that $V(g) = 0$, $V(x) = x - 1$ for $x \ge 1$, and $T^{1/x} V(0) = 1 - 1/x$. Since $T^0 V(0) = 0$, there is no conserving action at state 0. ∎

In view of Corollary 5.3, we consider the question if for any randomized Markov policy π there exists a Markov policy ϕ such that $v(x,\phi) \ge v(x,\pi)$ for all $x \in \mathbb{X}$. This natural question was asked in Dynkin and Yushkevich [21, Section 4.7]. The following example illustrates that the answer to this question is negative even when π is randomized stationary.

Example 5.3 (Feinberg and Sonin [33]) Let $\mathbb{X} = \{(0,0)\} \cup \{(i,j)\}$, $i = 1, 2, \dots$, $j = -i+1, \dots, 0\}$, $A = \{c, s\}$, where c stands for continue and s stands for stop. We also have that $\mathbb{A}(i,j) = \{c\}$ when $j < 0$, $\mathbb{A}(0,0) = \{s\}$, and $\mathbb{A}(i,0) = \{c,s\}$ when $i > 0$. The state $(0,0)$ is a "grave." This means that $p((0,0)|(0,0),s) = 1$ and $r((0,0),s) = 0$. The system always moves with zero reward from a state (i,j) to $(i,j+1)$ when $j < 0$. In other words, $p((i,j+1)|(i,j),c) = 1$ and $r((i,j),c) = 0$ when $j < 0$. For $i > 0$ we set $p((i+1,0)|(i,0),c) = p((0,0)|(i,0),s) = 1$ and $r((i,0),c) = 2^{-(i+1)}$, $r((i,0),s) = 1 - 2^{-(i+1)}$.

We consider a randomized stationary policy π that selects actions c and s with equal probabilities, $\pi(s|(i,0),c) = \pi(s|(i,0),s) = 0.5$ for $i > 0$. Expected one-step rewards in states $(i,0)$ are equal to 0.5 when $i > 0$. If the initial state is $(i,0)$ with positive i, the income stream forms a geometric progression with the first term and ratio equal 0.5. The sum of this progression is 1. From

each state (i, j) with $j < 0$, the system moves to state $(i, j + 1)$ with zero one-step rewards until it reaches state $(i, 0)$. Therefore, $v(x, \pi) = 1$ for all $x \in \mathbb{X} \setminus \{(0, 0)\}$.

Consider a Markov policy ϕ. If $\phi((i, 0), i - 1) = c$ for all $i \geq 1$ then $v((1, 0), \pi) = 0.5 < v((1, 0), \phi)$. Therefore, if $v(x, \phi) \geq v(x, \pi)$ for all $x \in \mathbb{X}$ then $\phi((i, 0), i-1) = s$ for some $i \geq 1$. However, in this case $v((i, -i+1), \phi) = 1 - 2^{-(i+1)} < v((i, -i + 1), \phi)$. ∎

5.4 NON-HOMOGENEOUS AND FINITE-HORIZON MODELS

In this section, we consider models in which action sets, transition probabilities, and rewards depend on time. For such models, called non-homogeneous, it is natural to expand the state space and consider the standard homogeneous model with the state space $\tilde{\mathbb{X}} = \mathbb{X} \times \mathbb{N}$. By doing this, we can expand all previous notations and results to non-homogeneous models. The major distinction is that in the new model the states are couples (x, n) instead of x. For example, values functions are $V(x, n)$.

A particular important example of non-homogeneous models is a finite-horizon model. For an n-horizon model, the state and action sets, reward and transition functions remain unchanged at epochs $t = 0, 1, \ldots, n - 1$. At epoch n, the one-step reward is $V_0(x)$, where V_0 is a final reward, and the system stops. The final reward is also known under several other names such as terminal reward or salvage value.

For each n-horizon model, it is natural to construct a non-homogeneous model, in which the model remains unchanged at steps $0, 1, \ldots, n-1$. At step n the reward function is V_0 and the system goes to a "grave" which is a state $\mathbf{x}$ with one available action under which this state is absorbent, i.e. $p(\mathbf{x}|\mathbf{x}, a) = 1$, and one-step rewards are equal to zero, $r(\mathbf{x}, a) = 0$.

If the original MDP contains only nonnegative rewards, V_0 is a nonnegative function, and the expected total rewards are finite for any policy and any initial state then the corresponding homogeneous model is positive. If the original model is negative and V_0 is a nonpositive function then the appropriate homogeneous model for an n-horizon model is negative. If the original model satisfies the General Convergence Condition with V_0 considered as a reward at n-th epoch then the appropriate homogeneous model for an n-horizon model satisfies the General Convergent Condition. In particular, if the original model satisfies the General Convergent Condition and $V_0 = 0$ then the homogeneous model, that corresponds to an n-horizon model, satisfies the General Convergent Condition too.

For n-horizon models, it is natural to use notations $V_i(x, V_0)$ instead of $V(x, n - i)$, $i = 0, \ldots, n$. Then Theorem 5.3 implies that

$$V_{n+1}(x, V_0) = TV_n(x, V_0), \qquad x \in \mathbb{X}. \tag{5.8}$$

We formulate a dynamic programming algorithm for N-horizon problems: first solve iteratively (5.8) for $n = 0, 1, \ldots, N - 1$ with $V_0(x, V_0) = V_0(x)$. Then for $\varepsilon > 0$ define a Markov policy $\phi = (\phi_N, \phi_{N-1}, \ldots, \phi_1)$ by $\phi_{n+1}(x) = a$ where

$n = N-1, N-2, \ldots, 0$, $x \in \mathbb{X}$, and a is an element of $\mathbb{A}(x)$ such that

$$T^a V_n(x, V_0) \geq T V_n(x, V_0) - \varepsilon/N.$$

Since any policy is equalizing in the new homogeneous model, it is easy to see that Markov policy ϕ is ε-optimal. In addition, if this policy can be defined as a conserving policy ($\varepsilon = 0$) then it is optimal. For example, if functions V_0 and r are bounded above and the Pre-Compactness Assumption holds then a conserving policy can be defined; see Corollary 5.7.

For the important case V_0 identical to 0, we write $V_n(x)$ instead of $V_n(x, V_0)$. An important question is whether the sequence $V_n(x, V_0)$ converges. We denote $V_\infty(x, V_0) = \lim_{n\to\infty} V_n(x, V_0)$ and $V_\infty(x) = V_\infty(x, 0)$ when these limits, possibly infinite, exist. Sequential computation of V_n is called *value iteration* or *successive approximation*. When the limit V_∞ exists, another important question is whether it equals V.

5.5 PARTICULAR MODELS

5.5.1 Discounted MDPs

We write the optimality operator T in the explicit form

$$Tf(x) = \sup\{r(x,a) + \beta P^a f(x) : a \in \mathbb{A}(x)\}.$$

This operator is a contracting mapping of the set of bounded functions endowed with the norm $\|f\| = \sup_x |f(x)|$ into itself; see Denardo [18] or Blackwell [8]. This implies that the optimality equation has a unique bounded solution and the limit $V_\infty(x, V_0)$ exists and is equal to this solution for any bounded V_0. Theorem 5.3 implies that this solution is V.

Theorem 5.4 (Blackwell [9], Denardo [18]) *Consider a* **D** *MDP.*

(i) For any $\varepsilon > 0$ there exists an ε-stationary optimal policy.

(ii) If a stationary policy is conserving, it is optimal.

(iii) If the Compactness Assumption holds then there exists a stationary optimal policy.

(iv) The limit V_∞ exists and equals V.

We remark that (i) and (ii) follow from the equalizing property applied either to a stationary policy ϕ with $T^{\phi(x)}V(x) \geq V(x) - (1-\beta)\varepsilon$ for all $x \in \mathbb{X}$ or to a stationary conserving policy; (iii) follows from Corollary 5.7, boundness of $V(x)$, and (iv); and (iv) follows from the fixed point theorem for contracting mappings. We also remark that, as the following simple example shows, the Optimality Equation may have additional unbounded solutions.

Example 5.4 $\mathbb{X} = \{0, 1, \ldots\}$ and $\mathbb{A}(x) = \{a\}$. All rewards are equal to 0. If the system is in state i, it moves to state $i+1$; $p(i+1|i,a) = 1$. We have that $V(x) = 0$ for all x. However, any function $u(i) = C/\beta^i$ satisfies the Optimality Equation $u = Tu$. ∎

5.5.2 Negative MDPs

Theorem 5.5 (Strauch [61]) *Consider an* **N** *MDP.*

(i) For any $\varepsilon > 0$ there exists a Markov ε-optimal policy.

(ii) If a stationary policy is conserving, it is optimal.

(iii) If the Pre-Compactness Assumption holds then there exists a stationary optimal policy.

(iv) The sequence $V_n(x)$ is nonincreasing, the limit $V_\infty(x)$ exists, and $V_\infty(x) \geq V(x)$ for all $x \in \mathbb{X}$.

(v) The value function V is the maximum nonpositive solution of the Optimality Equation.

(vi) If either $\mathbb{X}$ is finite or the Pre-Compactness Assumption holds then $V_\infty(x) = V(x)$ for all $x \in \mathbb{X}$.

In order to verify (i), one should consider a Markov policy $\phi = \{\phi_1, \phi_2, \dots\}$ such that $T^{\phi_n(x)} v(x) \geq v(x) - \varepsilon_n$ for all $x \in \mathbb{X}$ and for all $n \in \mathbb{N}$, where $\varepsilon_0 + \varepsilon_1 + \ldots \leq \varepsilon$. Then the inequality $v(x) \leq 0$ implies ε-optimality of ϕ. (ii) follows from the same reasons as (i) where ϕ is a stationary policy such that actions $\phi(x)$ are conserving at x. Since $V(x) \leq 0$, the value function V is bounded above and Corollary 5.7 implies (iii). (iv) is correct because $v(x,\pi) \leq v(x,\pi,n+1) \leq v(x,\pi,n)$ for any policy π and therefore $V(x) \leq V_{n+1}(x) \leq V_n(x)$. Statements (i), (ii), and (iii) hold for general convergent MDPs with $V(x) \leq 0$ for all $x \in \mathbb{X}$. The proof of these generalizations is the same as for **N** MDPs. It follows from the fact that any policy is equalizing. Statement (vi) was proved by Strauch [61] when either $\mathbb{X}$ is finite or all sets $\mathbb{A}(x)$ are finite. For the Pre-Compactness condition, (vi) follows from Proposition 9.17 in Bertsekas and Shreve combined with the proof of Lemma 5.2; see also Schäl [55] for the similar result when the Compactness Condition holds. When $\beta = 1$, Example 5.6 demonstrates the possibility of $V_\infty(x) > V(x)$; see also Strauch [61].

The following well-known example demonstrates that stationary ε-optimal policies may not exist for negative MDPs.

Example 5.5 $\mathbb{X} = \{1,2\}$, $\mathbb{A} = \mathbb{A}(1) = [0,1)$, and $\mathbb{A}(2) = \{0\}$. State 2 is absorbent with zero one-step rewards, $p(2|2,0) = 1$ and $r(2,0) = 0$. If action a is selected in state 1, $p(1|1,a) = a$, $p(2|1,a) = 1-a$, and $r(1,a) = (a-1)$. Then $w(1,\phi) = -1$ for any stationary policy ϕ and $V(1) = 0$. ∎

5.5.3 Unbounded discounted MDPs

Theorem 5.6 *Consider a* **UD** *MDP.*

(i) For any $\varepsilon > 0$ there exists a stationary ε-optimal policy.

(ii) If a stationary policy is conserving, it is optimal.

(iii) If the Pre-Compactness Assumption holds then there exists a stationary optimal policy.

(iv) The limit $V_\infty(x)$ exists and $V_\infty(x) \geq V(x)$ for all $x \in \mathbb{X}$.

Proof. The proof of statements (i-iii) coincides with the proof of the corresponding statements in Theorem 5.4. (iv) Let $r(x,a) \leq C$ for all x and a. If we subtract C from r then $V(x,\pi)$ will be reduced by $C/(1-\beta)$ for all x and π.

Therefore, we received an equivalent **N** MDP and the inequality follows from Theorem 5.5(iv) . ∎

The following example illustrates the possibility of $V_\infty(x) > V(x)$ in Theorems 5.5(iv) and 5.6(iv).

Example 5.6 ($V_\infty(x) > V(x)$) Let $\mathbb{X} = \{(0,0)\} \cup \{1,2,\dots\} \times \{0,1,\dots\}$, $\mathbb{A} = \mathbb{A}(0,0) = \{a^0, a^1, \dots\}$, and $\mathbb{A}(x) = \{a^0\}$ when $x \in \mathbb{X} \setminus \{(0,0)\}$. From any state (i,j) with $i > 0$, the system moves to $(i+1,j)$. This means that $p((i+1,j)|(i,j),a^0) = 1$ when $i > 0$. If action a^j is selected at state $(0,0)$, the system moves to $(1,j)$. In other words, $p((1,j)|(0,0),a^j)$=1. The rewards are

$$r((i,j),a^0) = \begin{cases} -\beta^{-i}, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

We have that $V_n(0,0) = 0$ for $n > 0$ and therefore $V_\infty(0,0) = 0$. We also have that $V(0,0) = -1$. ∎

5.5.4 Positive MDPs

The notion of an ε-optimal policies has been defined for a constant ε. It is easy to extend it when a function $g(x)$ is considered instead of constant ε. We say that a policy π is g-optimal for a given function g on $\mathbb{X}$ if $w(x,\pi) \geq v(x) - g(x)$ for all $x \in \mathbb{X}$.

Theorem 5.7 *Consider a* **P** *MDP.*

(i) (Blackwell [9]) $s(x) = V(x)$ *for all* $x \in \mathbb{X}$.

(ii) (Ornstein [50]) *For any* $\varepsilon > 0$ *there exists a stationary* εV*-optimal policy.*

(iii) (Blackwell [9]) *The sequence* $V_n(x)$ *is nondecreasing and the limit* $V_\infty(x)$ *exists and* $V_\infty(x) = V(x)$ *for all* $x \in \mathbb{X}$.

(iv) (Blackwell [9]) *The value function* V *is the minimum nonnegative solution of the Optimality Equation.*

We give the sketch of the proof of (i). We introduce a discount factor $\beta \in [0,1[$. Then $v(x,\pi,\beta)$ is non-decreasing in β and $v(x,\pi,\beta) \to v(x,\pi,1) = v(x,\pi)$ as $\beta \to 1$. We fix an arbitrary $\varepsilon > 0$. Consider an ε-optimal policy π. If ϕ is a stationary policy such that $T^{\phi(x)}V(x,\beta) \geq V(x,\beta) - (1-\beta)\varepsilon$ for all $x \in \mathbb{X}$ then $v(x,\pi,\beta) \geq V(x,\beta) - \varepsilon$ for all $x \in \mathbb{X}$. We fix $x \in \mathbb{X}$ and consider β such that $v(x,\pi,\beta) \geq v(x,\pi) - \varepsilon$. We have

$$\begin{aligned} s(x) \geq v(x,\phi) \geq v(x,\phi,\beta) \geq V(x,\beta) - \varepsilon \geq \\ v(x,\pi,\beta) - \varepsilon \geq v(x,\pi) - 2\varepsilon \geq V(x) - 3\varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, $s(x) \geq V(x)$. However, $s(x) \leq V(x)$. Thus, $s(x) = V(x)$ for all $x \in \mathbb{X}$.

The proof of (ii) is non-trivial and it was explained in greater details in Hordijk [43]. The proofs of statements (iii) and (iv) are easy. Example 5.7 shows that conserving policies may not be optimal. In addition it shows that the Compactness Assumption does not imply the existence of a stationary optimal policy even when the state space is finite.

Example 5.7 (Cavazos-Cadena, Feinberg, Montes-de-Oca [14]). Consider an MDP with state and action spaces given by $\mathbb{X} = \{0, 1\}$ and $\mathbb{A} = [0, 1]$, respectively. The sets of admissible actions are defined by $\mathbb{A}(1) = A = [0, 1]$, and $\mathbb{A}(0) = \{0\}$, whereas the transition law and the reward function are determined by

$$p(0|0, 0) = 1, \quad p(0|1, a) = a = 1 - p(1|1, a), \quad a \in [0, 1],$$

and

$$r(0, 0) = 0, \quad \text{and} \quad r(1, a) = a(1 - a), \ a \in [0, 1].$$

From these definitions, it is clear that 0 is an absorbing state under every policy and that $V(0) = 0$. Also, stationary policies are naturally indexed by the action they prescribe at state 1: $\phi_a \in \mathbb{F}$ is given by

$$\phi_a(1) = a, \quad \phi_a(0) = 0, \quad a \in [0, 1].$$

The following statements are true for this example:

(a) the expected total-reward at state 1 under policy ϕ_a is given by $v(1, \phi_a) = 1 - a$ if $a \in (0, 1]$, and $v(1, \phi_a) = 0$ if $a = 0$;
(b) $V(1) = 1$;
(c) an optimal policy does not exist.

Indeed, (a) implies $s(1) = 1$. Theorem 5.7(i) and (a) imply (b). From (a) and (b) we have that there is no stationary optimal policy. However, if an optimal policy exists in positive dynamic programming then a stationary optimal policy exists (Puterman 1994 , p. 324). Therefore, (a) and (b) yield (c).

Now we verify (a). Let $a \in (0, 1]$ be fixed. Starting at state $x_0 = 1$, under policy ϕ_a the system will arrive to state 0 at a random time T which is geometrically distributed with parameter a, that is, for each positive integer n, $\mathbb{P}_1^{\phi_a}\{T = n\} = a(1 - a)^{n-1}$ and a reward $r(1, a) = a(1 - a)$ will be earned at each integer time $t \in \{0, 1, \ldots, T - 1\}$. Since $r(0, 0) = 0$ and state 0 is absorbing, we have

$$v(1, \phi_a) = \mathbb{E}_1^{\phi_a}\left[\sum_{t=0}^{T-1} r(x_t, a_t)\right] = r(1, a)\, \mathbb{E}_1^{\phi_a}[T]$$

$$= a(1 - a) \times \mathbb{E}_1^{\phi_a}[T] = a(1 - a) \times \frac{1}{a} = (1 - a).$$

Consider now $a = 0$. In this case, state 1 is absorbing under $\phi_a = \phi_0$, and a reward $r(1, 0) = 0$ is earned forever, so that $v(1, \phi_0) = 0$. ∎

Blackwell [9] constructed an example when there is no stationary ε-optimal policy for a positive countable MDP with a finite state of actions. The following example, which is Example 2 in Feinberg and Sonin [33], shows that randomized Markov K-optimal policies may not exist for any $K > 0$; see also the relevant Example 2.26 in van der Wal [66].

Example 5.8 Let $\mathbb{X} = \{(0, 0)\} \cup \{(i, j) : i = 1, 2, \ldots, j = -i+1, \ldots, 0, 1\}$, and $\mathbb{A} = \{c, s\}$. In states $(0, 0)$ and (i, j), where $i = 1, 2, \ldots$, and $j < 0$, the action

sets, transition probabilities, and rewards are the same as in Example 5.3. We set $\mathbb{A}(i,1) = \{c\}$ and $p((0,0)|(1,1),c) = 1$, $p((i,1)|(i+1,1),c) = 1$, and $r((i,1),c) = 1$, $i = 1,2,\dots$. We also set $\mathbb{A}(i,0) = \mathbb{A}$, $p((0,0)|(i,0),c) = p((i+1,0)|(i,0),c) = 0.5$, $p((2^i - i^2 + i - 1,1)|(i,0),s) = 1$, and $r((i,0),a) = 0$, $i = 1,2,\dots$, $a \in \mathbb{A}$.

We denote $g(i) = i^2 - i + 1$. The sequence $g(i)$ is used in this example because of its three properties: (a) $\frac{g(n+i)}{2^i} \to 0$ as $i \to \infty$, (b) $g(i) > 0$ when $i \geq 1$, and (c) $\sum_{i=1}^{\infty} \frac{1}{g(i)} < \infty$.

Let i be a positive integer. If action s is selected in state $(i,0)$, the total future reward is $2^i - g(i)$. If action c is selected in this state, the system moves with fifty-fifty chances to the "grave" $(0,0)$ or to the state $(i+1,0)$. Let ϕ be a Markov policy and $m = \min\{n \geq 0 |\ \phi_n(i+n) = s,\ i = 0,1,\dots\}$. If $m < \infty$, we have that $v((i,0),\phi) = 2^{-m}(2^{i+m} - g(i+m))$. If $m = \infty$ then $v((i,0),\phi) = 0$. Therefore, $v(i,0) = 2^i$. The optimality equation implies that $v(i,j) = 2^i$ when $j < 0$.

Let ψ be a randomized Markov K-optimal policy for some constant $K > 0$. Since $v((i,j),\psi) \geq V(i,j) - K = 2^i - K$, $j \leq 0$, we have that $\pi_j(s|(i,0)) \leq K/g(i)$, $j = 0,1,\dots,i-1$. Then

$$v((i,0),\psi) = \sum_{t=0}^{\infty} P\{x_t = (i+t,0), a_t = s\}(2^{i+t} - g(i+t)) =$$

$$\pi_0(s|(i,0))(2^i - g(i)) + \sum_{n=1}^{\infty} \pi_n(s|(i+n,0))(\frac{1}{2})^n \times$$

$$[\prod_{j=0}^{n-1}(1 - \pi_j(s|(s+j,0)))](2^{n+i} - g(n+i)) \leq$$

$$2^i \sum_{n=0}^{\infty} \pi_n(s|(i+n,0)) < 2^i K \sum_{j=i}^{\infty} \frac{1}{g(j)} = 2^i - 2^i(1 - K\sum_{j=i}^{\infty} \frac{1}{g(j)}) < 2^i - K.$$

when i is large enough. The last inequality holds since $\sum_{j=i}^{\infty} \frac{1}{g(j)} \to 0$ as $i \to \infty$. Thus ψ is not K-optimal and for any $K > 0$ there is no randomized Markov K-optimal policy. ∎

Puterman [52] describes and studies a class of MDPs called positive bounded. For these MDPs the assumption that the reward function r is nonnegative is replaced with a weaker assumption that for each $x \in \mathbb{X}$ there is at least one $a \in \mathbb{A}(x)$ for which $r(x,a) \geq 0$. As explained in Puterman [52], these more general models inherits many properties of positive models. Our remark is that, though $s = V$ for positive bounded models, stationary εV-optimal may not exist for them. The following example shows that randomized Markov εV-optimal policies may not exist in positive bounded MDPs.

Example 5.9 Consider Example 5.8. Let $\tilde{v}$ and $\tilde{V}$ denote functions v and V in that example. If ψ is a randomized Markov policy such that $\tilde{v}((i,-i+1),\psi) > \tilde{V}(i,-i+1) - 1$ for all $i \geq 1$ then $\tilde{v}((i,0),\psi) < \tilde{V}(i,0) - 1$ for large i. Using the same arguments as in the previous example but with slightly more complicated

calculations, it is possible to show that $\tilde{v}((i,-1),\psi) < \tilde{V}(i,-1) - 1$ for some i. We recall that $\tilde{V}(i,j) = 2^i$.

Now we slightly modify Example 5.8 by setting $\mathbb{A}(i,-1) = \{c,s\}$ with $p((i,0)|(i,-1),c) = 1$, $r((i,-1),c) = 1 - 2^i$ and $p((0,0)|(i,-1),s) = 1$, $r((i,-1),s) = 0$. In other words, the decision maker can either pay $2^i - 1$ and move to the state $(i,0)$ or collect zero return and stop the process. Since $V(i,0) = 2^i$, the optimality equation implies that $V(i,j) = 1$ when $j < 0$. Fix $\varepsilon > 0$. Let ϕ be a randomized Markov policy and $v((i,j),\phi) \geq 1 - \varepsilon$ when $j < 0$. Consider the model from the Example 5.8. For that model consider a randomized Markov policy ψ which coincides with ϕ at all states except $(i,-1)$, $i = 1,2,\ldots$. Since in the MDP from Example 5.8 we have that $\mathbb{A}(i,-1) = \{c\}$, policy ϕ always selects action c at states $(i,-1)$. We observe that $v((i,j),\phi) \leq \tilde{v}((i,j),\psi) + 2^i - 1$. Therefore, if $v((i,j),\phi) \geq V(i,j) - \varepsilon V(i,j)$ for all states (i,j) with $j < 0$ then $\tilde{v}((i,j),\psi) \geq 2^i - \varepsilon$ for all states with $j < 0$. As explained in the previous paragraph, this is impossible. ∎

5.6 VALUE ITERATION

Since $V_n(x) \geq v_n(x,\pi)$ for any policy π, $\liminf_{n\to\infty} V_n(x) \geq V(x)$ for all $x \in \mathbb{X}$; see Schäl [55] or van der Wal [66, Lemma 3.1] for details.

The following example illustrates that, even if V_∞ exists, it is possible that $V_\infty(x) > V(x)$ for some x for MDPs with finite state and action sets.

Example 5.10 (van Hee, Hordijk, and van der Wal [71] or van der Wal [66, Example 3.2]) Let $\mathbb{X} = \{1,2,g\}$, $\mathbb{A} = \mathbb{A}(1) = \{1,2\}$, $\mathbb{A}(2) = \mathbb{A}(g) = \{1\}$. We also have that $p(1|1,1) = p(2|1,2) = p(g|2,1) = p(g|g,1)$. Thus, under action 1 selected at state 1, the process remains in this state and action 2 moves it to state 2. From state 2 the process moves to the absorbent state g. We also have $r(1,2) = 2$, $r(2,1) = -1$, and all other rewards equal zero. If $x_0 = 1$, an n-step optimal policy is to remain $(n-1)$ units of time in 1 and then to move to state 2. Thus, $V_n(1) = 2$ and $V(1) = 1$. ∎

Example 5.11 (The limit V_∞ does not exist) Consider an MDP with the same state sets, action sets, and transition probabilities as in Example 5.6. All rewards equal zero except $r(2i+2,i) = -2$, and $r(2i+1,i) = r(2i+3,i) = 1$, $i = 0,1,\ldots$. Then $\limsup_{n\to\infty} V_n(0,0) = 1$, $\liminf_{n\to\infty} V_n(0,0) = -1$, and $V(0,0) = 0$. ∎

Obviously, $V_\infty(x,V) = V(x)$, $x \in \mathbb{X}$. The following statement generalizes this fact. We recall that $V_-(x) = \sup\{v_-(x,\pi)|\ \pi \in \Pi\}$.

Theorem 5.8 (van der Wal [66, Theorem 3.7]) *If $V_-(x) \leq V_0(x) \leq V(x)$ for all $x \in \mathbb{X}$ then for each $x \in \mathbb{X}$ the limit $V_\infty(x,V_0)$ exists and equals $V(x)$.*

If either $\mathbb{X}$ is finite or all sets $\mathbb{A}(x)$ are finite, Theorems 5.5(vi) and 5.8 provide the following two-step procedure to compute V.

Step 1 Compute $V_-(x) = V_\infty^-(x)$ where "$-$" means that we deal with the MDP in which the reward function r is replaced with its negative part r^-.

Step 2 Compute $V(x) = V_\infty(x, V_-)$.

Van Dawen [64] studied the rate of convergence of value iteration for finite state **N** MDPs. Maitra and Sudderth [48] proved that the limit of V_n exists in uncountable **N** MDPs if the definition of a limit is extended by using transfinite induction. Van Hee, Hordijk, and van der Wal [71] provided a sufficient condition for convergence of the value iteration; see Theorem 3.3 in van der Wal [66]. This result generalizes Theorem 5.5(vi). Bertsekas [4] and van der Wal [66] provide modifications of value iteration algorithms such as Jacobi and Gauss-Seidel iteration algorithms. For some other results on value iteration, see Schäl [55] and Stidham [60].

5.7 THE FIRST MAIN THEOREM: UNIFORMLY NEARLY-OPTIMAL STATIONARY POLICIES

Comparison of positive and negative MDPs creates a strong impression that the results on the existence of good policies are totally different for these models; see the table on page 324 in Puterman [52]. For example, for **P** MDPs $s = V$ but it is not true for **N** MDPs. In addition, it is even not obvious how to define a unified notion of ε-optimality. The notion of εV-optimality, which is natural for **P** MDPs, is not applicable to **N** MDPs because no policy can be better than optimal. The notion of ε-optimality, which is natural for **N** MDPs, is not applicable to **P** models; see Example 5.8.

As the result, almost all books treat these models separately. The major exceptions are Hinderer [42], Dynkin and Yushkevich [21], and van der Wal [66]. The first two books were written when very little was known about **GC** MDPs. Van der Wal [66] introduced several new results including several counterexamples, the proof that $s = V$ when all actions $\mathbb{A}(x)$ are finite, and interesting results on value iteration. The following result has been known for a long period of time.

Theorem 5.9 (Blackwell [7], Krylov [45], Dynkin and Yushkevich [21], Kallenberg [44]) *If $\mathbb{X}$ and A are finite then there exists a stationary optimal policy.*

As Examples 5.5 and 5.8 demonstrate, if either $\mathbb{X}$ or one of the sets $\mathbb{A}(x)$ is infinite then optimal policies may not exist. Van der Wal [66] showed that if all states $\mathbb{A}(x)$ are finite then $s(x) = V(x)$ for all x. Schäl [56] extended this result to Borel state problems with compact action sets.

For a long time, there were no results on the existence of stationary uniformly nearly-optimal policies for **GC** MDPs. Probably the first result was van der Wal's [68] theorem that states that stationary εV_+-optimal policies exist if for each $x \in \mathbb{X}^{\leq} = \{x \in \mathbb{X} |\ V(x) \leq 0\}$ there exists a conserving action. This result generalizes Ornsten's theorem (Theorem 5.7(ii)). The weak point of this statement is that the use of function V_+ in the definition of ε-optimality does not look natural when reward functions can be negative. For example, if $V(x)$ is always nonpositive and there is a conserving action in each state, any stationary conserving policy is optimal; this fact follows from the same arguments as Theorem 5.5(ii). However, it is possible that $V_+(x)$ is unbounded from above and van der Wal's [68] result implies the existence of stationary policies which

are far from optimal. Van der Wal and Wessels [69] asked whether V_+ could be substituted with a better function.

Now we describe the result from Feinberg and Sonin [34] on the existence of uniformly nearly-optimal policies within the class of stationary policies for **GC** MDPs. This result implies Ornstein's theorem and many other specific results. First, we define the sets

$$\mathbb{X}_S = \{x \in \mathbb{X} |\ v(x,\phi) = s(x) \text{ for some } \phi \in \Pi^{RS}\};$$

$$\mathbb{X}_\Pi = \{x \in \mathbb{X} |\ v(x,\pi) = V(x) \text{ for some } \pi \in \Pi^{R}\}.$$

For any x from $\mathbb{X}_\Pi$ there is a policy π which is optimal for this initial state. Similarly, for any x from $\mathbb{X}_S$ there is the best randomized stationary policy for the initial state x.

Let Φ be the class of numerical functions g on $\mathbb{X}$ such that $Pg^+ < \infty$. We observe that functions V, s, V_+ belong to Φ. Constants also belong to Φ. For $Z \subseteq \mathbb{X}$ and for any function $g \in \Phi$, we denote by $L(g,Z)$ the set of nonnegative functions ℓ on $\mathbb{X}$ such that $\ell(x) > 0$ and $\ell(x) \geq \max\{g, P\ell\}$ when $x \in \mathbb{X} \setminus Z$. We also define $L(g) = L(g,\emptyset)$.

We remark that $L(g,Z)$ is the set of nonnegative functions which are positive excessive (or super-harmonic according to another terminology) majorants of g on $\mathbb{X} \setminus Z$. The sets $L(g,Z)$ possess the following two properties: (i) if $g(x) \geq g'(x)$ for all $x \in \mathbb{X} \setminus Z$ then $L(g',Z) \supseteq L(g,Z)$, and (ii) if $Z \supseteq Y$ then $L(g,Z) \supseteq L(g,Y)$. We also observe that if $\ell \in L(g,Z)$ then $\ell_0 \in L(g,Z)$ where $l_0(x) = l(x)$ for $x \in \mathbb{X} \setminus Z$ and $l_0(x) = 0$ for $x \in Z$.

We observe that $V_+ + 1 \in L(V)$. Therefore, $L(V,Z) \neq \emptyset$ for any Z. We also observe that $V \in L(V,\mathbb{X}_S)$ in **P** MDPs and $V_+ \in L(V,\mathbb{X}_S)$ if for each $x \in \mathbb{X}^{\leq}$ there exists a conserving action. Indeed, consider the set $\mathbb{X}^0_+ = \{x \in \mathbb{X} | V_+(x) = 0\}$. If an initial state belongs to $\mathbb{X}^0_+$ the system will never leave this set. Therefore, we get a negative MDP if we restrict $\mathbb{X}$ to $\mathbb{X}^0$. Since there are conserving actions in all sets $\mathbb{A}(x)$ when $x \in \mathbb{X}^{\leq}$ and $\mathbb{X}^0_+ \subseteq \mathbb{X}^{\leq}$, there is a stationary optimal policy in the negative MDP with the state space $\mathbb{X}^0_+$. Therefore, there is an optimal stationary policy for any initial point $x \in \mathbb{X}^0_+$. Thus, $\mathbb{X}^0_+ \subseteq \mathbb{X}_S$. And we have that $V_+ \in L(V_+, \mathbb{X}^0_+) \subseteq L(V,\mathbb{X}_S)$.

Theorem 5.10 (The first main theorem: the existence of uniformly optimal stationary policies; Feinberg and Sonin [34], Theorem 2.1.) *For any $\varepsilon > 0$ and for any $\ell \in L(s,\mathbb{X}_S)$ there exists a stationary policy ϕ such that $v(x,\phi) \geq s(x) - \varepsilon\ell$ for all $x \in \mathbb{X}$.*

The proof of Theorem 5.10 is not trivial and we do not consider it here. This theorem provides a unified way to prove the existence of uniformly $\varepsilon\ell$-optimal policies: it is sufficient to prove that $s(x) = V(x)$ for any given x and that $\ell \in L(s,\mathbb{X}_S.)$ As we saw in Theorem 5.7, the proof that $s(x) = V(x)$ is significantly easier than the direct proof that there are uniformly nearly-optimal policies. For example, for **P** MDPs, we have $s(x) = V(x)$ (Theorem 5.7(i)) and $V \in L(V,\mathbb{X}_S)$. Therefore, Theorem 5.10 implies Ornstein's theorem. If there is a conserving action for any $x \in \mathbb{X}^{\leq}$ then $s(x) = V(x)$; this and more general

results follow from our second main statement, Theorem 5.21. Therefore, Theorem 5.10 implies van der Wal's [68] result described above. We also observe that if V is bounded above by a constant K then $K \in L(V)$. Therefore, if $s = V$ and V is bounded above then there exist stationary ε-optimal policies. For example, Theorem 5.10 implies the existence of a stationary ε-optimal policy for positive bounded models from Puterman [52]. Indeed, it is easy to see that in this model $V(x) \geq 0$ for all x and if $V(x) = 0$ then there is a conserving action at x. Thus, $s(x) = V(x)$ for all x in positive bounded models.

An important corollary from Theorem 5.10 is that the value function s is the solution of the optimality equation.

Theorem 5.11 (Feinberg and Sonin [34, Theorem 2.2]) *$s(x) = Ts(x)$ for all $x \in \mathbb{X}$.*

We remark that we do not know how to prove $s = Ts$ in **GC** MDPs without using Theorem 5.10. It is easy to show that $Ts(x) \geq s(x)$ for all $x \in \mathbb{X}$. In order to prove $s(x) \geq Ts(x)$ we need to use the existence of a stationary policy ϕ such that $v(z, \phi) \geq s(z) - \varepsilon(z)$ for all $z \in \mathbb{X}$ and for some function $\varepsilon(z)$. In the proof of $V \geq TV$, we used Corollary 5.5 for the similar result. After Theorem 5.10 is established, the proof of $s = Ts$ is similar to the proof of $V = TV$; see Feinberg and Sonin [34] for details.

An important question is how to expand the class of functions $L(s, \mathbb{X}_S)$. One possible approach is to replace s in $L(s, \mathbb{X}_S)$ with a function $d \leq s$. Our particular interest is to consider d defined in a way that it is possible that s is not bounded above but d is bounded above. Van Dawen and Schäl [65], van Dawen [64], and Schäl and Sudderth [59] considered functions d of this type for particular models. These functions were related to the limiting behavior of $s(x_n)$. For countable MDPs, the broadest known class of such functions was introduced in Feinberg and Sonin [36]. Let $Q^b = \cup_{n=1}^{\infty}\{\tau |\ \tau < n\}$ be the set of all uniformly bounded stopping times. Let

$$d_S(x) = \sup_{\phi S} \inf_{\tau \in Q^b} \mathbb{E}_x^{\phi}\, s(x_\tau).$$

The following theorem generalizes Theorem 5.10 in the sense that it describes a broader class of functions ℓ.

Theorem 5.12 (Feinberg and Sonin [36], Theorem 1.) *For any $\varepsilon > 0$ and for any $\ell \in L(d, \mathbb{X}_S)$ there exists a stationary policy ϕ such that $v(x, \phi) \geq s(x) - \varepsilon\ell$ for all $x \in \mathbb{X}$.*

We say that an MDP is deterministic if all transition probabilities $p(y|x, a)$ are equal to 0 or 1. Bertsekas and Shreve [6] proved the existence of stationary ε-optimal policies for **P** MDPs.

Theorem 5.13 (Feinberg and Sonin [34, Section 5]) *Consider a deterministic MDP. Consider an arbitrary nonnegative function ℓ on $\mathbb{X}$ such that $\ell(x) \geq \ell(y)$ for $x, y \in \mathbb{X} \setminus \mathbb{X}_S$ when there is an action $a \in \mathbb{A}(x)$ such that $p(y|x, a) = 1$. Then for any $\varepsilon > 0$ there exists a stationary policy ϕ such that $v(x, \phi) \geq s(x) - \varepsilon\ell(x)$ for all $x \in \mathbb{X}$. In particular, in a **P** MDP for any $\varepsilon > 0$ there exists an $\varepsilon \min(1, V)$-optimal stationary policy.*

We see that stationary $\varepsilon\ell$-optimal policies exist in deterministic models for a broader class of functions ℓ. A natural research direction is to expand the class $L(d, \mathbb{X}_S)$ in Theorem 5.12. An open question is to get results that unify deterministic and stochastic models; see Feinberg [24] for additional details. Another intriguing question is that the function d looks like a value function in a gambling problem; see Dubins and Savage [20] and Maitra and Sudderth [49]. However, the meaning of this relationship currently is not clear.

In conclusion, we illustrate how Theorems 5.10 and 5.12 can be used to prove the existence of uniformly nearly-optimal policies in particular models.

Theorem 5.14 (cp. Cavazos-Cadena and Montes-de-Oca [15]) *Consider an* **N** *MDP. Assume that if* $V(x) > -\infty$ *for some* $x \in \mathbb{X}$ *then*

$$\lim_{n\to\infty} \inf_{\phi\in S} \mathbb{E}_x^\phi\, v(x_n, \phi) = 0. \tag{5.9}$$

If this assumption holds then for any $\varepsilon > 0$ *there exists a stationary* ε*-optimal policy.*

Proof. In view of Theorem 5.10, it is sufficient to prove that $s(x) = V(x)$ for all $x \in \mathbb{X}$. Since $s(x) \le V(x)$, we have that $s(x) = V(x)$ when $V(x) = -\infty$. We observe that it is sufficient to prove the theorem for the situation when $V(x) > -\infty$ for all $x \in \mathbb{X}$. Indeed, let $Y = \{x \in \mathbb{X} |\, V(x) = -\infty\}$. If $Y = \mathbb{X}$ then the problem is trivial. So, we consider the case, $Y \ne \mathbb{X}$. In this case, we exclude the subset Y from $\mathbb{X}$ and remove all actions a such that $p(Y|x, a) > 0$ from sets $\mathbb{A}(x)$ when $x \in \mathbb{X} \setminus Y$. We remark that $\mathbb{A}(x)$, $\mathbb{X} \in \mathbb{X} \setminus Y$ are still nonempty sets after this procedure because otherwise we would have $V(x) = -\infty$. So, we have received a new model in which condition (5.9) holds for all $x \in \mathbb{X}$.

So, it is sufficient to show that $s(x) = V(x)$ for any $x \in \mathbb{X}$ if (5.9) holds for all $x \in \mathbb{X}$. In order to do it, we fix an arbitrary $\varepsilon > 0$ and an arbitrary $x \in \mathbb{X}$. Then we select $n \ge 1$ such that

$$\inf_{\phi\in S} \mathbb{E}_x^\phi\, v(x_n, \phi) \ge -\varepsilon. \tag{5.10}$$

We also select a stationary policy ϕ such that $T^{\phi(z)}V(z) \ge V(z) - \varepsilon/n$ for all $z \in \mathbb{X}$. We have that

$$(T^\phi)^n V(x) \ge V(x) - \varepsilon. \tag{5.11}$$

We also have

$$s(x) \ge v(x, \phi) = T^\phi v(x, \phi) = (T^\phi)^n v(x, \phi) = (T^\phi)^n (v(x, \phi) - V(x) + V(x)) =$$

$$(T^\phi)^n V(x) + \mathbb{E}_x^\phi\, v(x_n, \phi) - \mathbb{E}_x^\phi\, V(x_n) \ge V(x) - 2\varepsilon.$$

The last inequality follows from (5.10, 5.11) and $V(x) \le 0$. Since $\varepsilon > 0$ is arbitrary, $s(x) \ge V(x)$. Thus, $s(x) = V(x)$. ∎

We remark that $s = V$ if $s_- = V_-$ where the minus means that the function is related to the model in which r is replaced with r^-. This statement follows from Theorem 5.21. So, if the model, in which r is replaced with r^- satisfies

conditions of Theorem 5.14 then for any $\ell \in L(d_S, \mathbb{X}_S)$ and for any $\varepsilon > 0$ there exists a stationary $\varepsilon\ell$-optimal policy. We also observe that $\mathbb{X}_\Pi = \mathbb{X}_S$ if $V = s$. The following theorem provides another condition when $V = s$.

Theorem 5.15 (Corollary 2 in Feinberg [24]) *If for any* $x \in \{x \in \mathbb{X} |\ V(x) > -\infty\}$ *and for any* $\pi \in \Pi^M$

$$\limsup_{n\to\infty} \mathbb{E}_x^\pi s(x_n) \geq 0$$

then for any $\ell \in L(d_S, \mathbb{X}_\Pi)$ *and for any* $\varepsilon > 0$ *there exists a stationary* $\varepsilon\ell$*-optimal policy.*

In conclusion, we remark that $s(x) = s_R(x)$ for all $x \in \mathbb{X}$ where $s_R(x) = \sup_{\pi\in\Pi^{RS}} v(x,\pi)$. This result was proved for the countable state MDPs by Feinberg and Sonin [36] and for Borel MDPs by Feinberg [29].

5.8 (f, I)-GENERATED POLICIES

Van der Wal [67] proved that for any $\varepsilon > 0$ there exists a Markov $\varepsilon\ell$-optimal policy with $\ell = V_+ + 1$. Theorems 5.10 and 5.12 imply similar results for broader sets of functions ℓ. Indeed, let us replace the state space $\mathbb{X}$ with the state space of couples (x, n) where $x \in \mathbb{X}$ and $n = 0, 1, \dots$. The process moves from states (x, n) to $(y, n + 1)$ with transition probabilities $p(y|x, a)$. There is one-to-one correspondence between stationary policies in the new model and Markov policies in the original one. We also observe that if there exists a policy π such that $v(x, \pi) = V(x)$ for some $x \in \mathbb{X}$ then there is a Markov policy ϕ such that $v(x, \pi) = V(x)$; see Corollary 5.3. These observations and Theorems 5.10 imply the following result.

Theorem 5.16 (Feinberg and Sonin [34]) *For any* $\varepsilon > 0$ *and for any* $\ell \in L(V, \mathbb{X}_\Pi)$ *there exists a Markov* $\varepsilon\ell$*-optimal policy.*

We remark that if one uses Theorem 5.12 instead of Theorem 5.10 then a slightly more general class of function ℓ can be considered. We shall formulate it as a part of a more general statement, Theorem 5.17. A natural question is which classes of policies contain uniformly nearly-optimal policies. In order to answer this question, we consider the following construction.

Let I be a countable or finite set and let $f : H \to I$ be a measurable mapping. A policy π is called randomized (f, I)-generated if $\pi_n(\cdot|h_n) = \sigma(\cdot|x_n, f(h_n))$ for some transition probability σ from $\mathbb{X}\times I$ to $\mathbb{A}$. We also consider nonrandomized (f, I)-generated policies. In the latter case, $\pi_n(h_n) = \sigma(x_n, f(h_n))$ where σ is a mapping from $\mathbb{X} \times I$ to $\mathbb{A}$. If we say that a policy is (f, I)-generated we mean that it is nonrandomized and (f, I)-generated.

Markov policies are an example of (f, I)-generated policies. In this case, $I = \mathbb{N}$ and $f(h_n) = n$. Stationary policies form another example. In this case, $I = \{0\}$ and $f(h_n) = 0$ for all h_n. Tracking policies introduced by Hill [41] is the third example. For tracking policies, decision depends on the current state and the number of visits to it, i.e. $I = \{1, 2, \dots\}$ and $f(h_n) = \sum_{i=0}^n \mathbf{I}\{x_i = x_n\}$. So, I is the information available about the past and f is a memory

function which indicates what information about the past is remembered. See [35, 36, 25, 26, 28, 31] for other examples of (f, I)-generated policies. We consider the following condition.

Transitivity Condition (Feinberg and Sonin [36]) If $f(h_n) = f(h'_m)$ and $x_n = x'_m$, where $h_n = x_0 a_0 x_1 \dots x_n$ and $h_m = x'_0 a'_0 x'_1 \dots x'_m$, then $f(h_n a z) = f(h'_m a z)$ for all $a \in \mathbb{A}(x_n)$ and for all $z \in \mathbb{X}$.

In other words, the current state x_n, the current information $i_n = f(h_n)$, the next action a, and the next state z completely define $i_{n+1} = f(h_n a z)$. We observe that stationary and Markov policies satisfy the Transitivity Condition and tracking policies do not satisfy this condition. We denote by I^f the set of (f, I)-generated policies. We also denote by RI^f the set of randomized (f, I)-generated policies. We define

$$V^f(x) = \sup\{v(x, \pi) : \ \pi \in I^f\}, \qquad V_R^f(x) = \sup\{v(x, \pi) : \ \pi \in RI^f\}.$$

Let

$$d^f(x) = \sup_{\phi \in I^f} \inf_{\tau \in Q^b} \mathbb{E}_x^\phi V^f(x_\tau).$$

Let also

$$\mathbb{X}^f = \{x \in \mathbb{X} | v(x, \phi) = V^f(x) \text{ for some } \phi \in I^f\}.$$

Theorem 5.17 (Feinberg and Sonin [36]) *Let I be a countable or finite set and let a measurable function $f : \ H \to I$ satisfy the Transitivity Condition. Then $V^f(x) = V_R^f(x)$ for all $x \in \mathbb{X}$ and for any $\ell \in L(d^f, \mathbb{X}^f)$ and for any $\varepsilon > 0$ there exists an (f, I)-generated policy ϕ such that $V(x, \phi) \geq V^f(x) - \varepsilon \ell(x)$ for all $x \in \mathbb{X}$.*

The proof of Theorem 5.17 is based on the following observation. Consider the MDP with the state space $\mathbb{X} \times I$. If the transitivity condition holds then there is a one-to-one correspondence between (f, I)-policies and stationary policies in the new model where states are in fact couples $(x_n, f(h_n))$. Then Theorem 5.17 follows from $s(x) = s_R(x)$ and from Theorem 5.12.

Markov and stationary policies are two examples of classes of policies satisfying the transitivity conditions. Another important example are Y-policies introduced in Feinberg [26]. For Y-policies, defined by a subset Y of $\mathbb{X}$, decisions depend on the following factors: (i) the current state, (ii) the number of visits to Y, and (iii) the time passed after the last visit to Y. For any history $h_n \in H_n$, $n = 0, 1, \dots$, we define the number of visits to Y

$$m(h_n, Y) = \sum_{i=0}^{n} \mathbf{I}\{x_i \in Y\}.$$

We also agree that by definition $x_{-1} \in Y$ and let $\xi(h_n, Y)$ be the epoch when the system visited Y for the last time,

$$\xi(h_n, Y) = \max\{i = -1, 0, 1, \dots, n| \ x_n \in Y\}$$

and $\theta(h_n, Y) = n - \xi(h_n, Y)$ be the time passed after the last visit.

For Y-embedded policies we define $I = \{0, 1, \dots\} \times \{0, 1, \dots\}$ and $f(h_n) = (m(h_n, Y), \theta(h_n, Y))$. We observe that if $f(h_n) = (m, i)$ then $f(h_n, a, z) =$

$(m, i+1)\mathbf{I}\{z \notin Y\} + (m+1, 0)\mathbf{I}\{z \in Y\}$ where $\mathbf{I}$ is an indicator function. Therefore, the Transitivity Condition holds and Theorem 5.17 can be applied to Y-embedded policies. We also remark that in two extreme cases, $Y = \mathbb{X}$ and $Y = \emptyset$, the set of Y-embedded policies coincides with the set of Markov policies.

Let τ^1 be the first epoch when the system hits Y, $\tau^1(h) = \min\{n \geq 0 |\ x_n \in Y\}$ and let τ^m be the m^{th} epoch when the process hits Y : $\tau^m = \min\{n > \tau^{m-1} |\ x_n \in Y\}$, $m = 2, 3, \ldots$. We observe that $\{\tau^m(h) = n\} = \{m(h_n, Y) = m, \theta(h_n, Y) = 0\}$. The following theorem is a direct generalization of Theorem 5.1 to Y-embedded policies. It was proved by induction in Feinberg [25] when $\lambda_1 = 1$ and $\lambda_i = 0$ for $i > 1$. However, in the case of arbitrary λ_i, the proof remains the same.

Theorem 5.18 (Theorem 4.1 in [25]) *Let $\pi^1, \pi^2, \ldots$ be an arbitrary sequence of policies and $\lambda_1, \lambda_2, \ldots$ a sequence of nonnegative numbers summing to 1. For an arbitrary $Y \subseteq \mathbb{X}$ consider a randomized Y-embedded policy π defined by*

$$\pi\left(C \mid y, m, k\right) \triangleq \frac{\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}\left(x_{\tau^m+k} = y, \tau^{m+1} > \tau^m + k, a_{\tau^m+k} \in C\right)}{\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}\left(x_{\tau^m+k} = y, \tau^{m+1} > \tau^m + k\right)}\,, \tag{5.12}$$

whenever the denominator in (5.12) is not equal to 0, where $C \in \mathcal{A}$, $y \in \mathbb{X}$, $m = 1, 2, \ldots$, and $k = 0, 1, \ldots$. Then, for all $m, k = 0, 1, \ldots$, $y \in \mathbb{X}$ and measurable subsets C of $\mathbb{A}(y)$,

$$\begin{aligned} &\mathbb{P}_x^{\pi}\left(x_{\tau^m+k} = y, \tau^{m+1} > \tau^m + k, a_{\tau^m+k} \in C\right) = \\ &\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}\left(x_{\tau^m+k} = x, \tau^{m+1} > \tau^m + k, a_{\tau^m+k} \in C\right)\,, \end{aligned} \tag{5.13}$$

and therefore

$$v(x, \pi) = \sum_{i=1}^{\infty} \lambda_i v(x, \pi^i). \tag{5.14}$$

Theorem 5.18 provides an important result in the following direction. Let Δ be some class of randomized policies. Then there is an explicit formula that indicates a policy $\pi \in \Delta$ for a given initial state x and for an arbitrary policy π such that $v(x, \sigma) = v(x, \pi)$.

The results in this direction were provided for stationary policies by Krylov [46, 47] for discounted controlled diffusion processes and by Borkar [13] for discounted MDPs. Altman [1] proved the same result for MDPs with uniformly bounded life times. Discounted MDPs are a particular case of such models. In this case, the occupation measure for the original policy is equal to the occupation measure for the corresponding randomized stationary policy. For a more general case when the expected number of visits to each state is bounded above, Altman [1] proved a weaker result that the occupation measure for the corresponding randomized stationary policy majorizes the original occupation measure. Feinberg and Sonin [37] constructed an example when the strong inequality takes place.

We remark that Theorem 5.18 has no limitation on the life time of the system. It is a direct generalization of Theorem 5.1 which follows from Theorem 5.18 when $Y = \mathbb{X}$ or $Y = \emptyset$. Unfortunately, the relationship between Theorem 5.18 and the results on the equivalent randomized stationary strategies, such as Theorem 8.2(ii,iii) in Altman [2], has not been established yet.

In addition to the Transitivity Condition, the so-called Non-Repeating Condition for (f, I)-policies plays an important role.

Non-Repeating Condition (Feinberg [25]) If a history $h_m = x_0 a_0 \dots x_m$ is a continuation of a history $h_n = x_0 a_0 \dots x_n$, i.e. $h_m = h_n a_n \dots x_m$, then $(x_n, f(h_n)) \neq (x_m, f(h_m))$.

In other words, the Non-Repeating Condition means that if $m > n$, $h_m = h_n a_n x_{n+1} \dots x_m$, and $x_n = x_m$ then $f(h_m) \neq f(h_n)$. Of course, if the Non-Repeating Condition holds, I cannot be finite. Markov, tracking, and Y-embedded policies are examples of (f, I)-generated policies that satisfy the Non-Repeating Condition. Stationary policies do not satisfy this condition.

Any randomized (f, I)-generated policy π is defined by transition probabilities $\pi(\cdot|x, i)$. Any nonrandomized (f, I)-generated policy σ is defined by a function $\sigma(x, i)$. Let I^f be the set of all functions ϕ on $\mathbb{X} \times I$ with values on $\mathbb{A}$ and such $\phi(x, i) \in \mathbb{A}(x)$ for all $x \in \mathbb{X}$. Any randomized (f, I)-generated policy π defines a measure m^π on the set I^f defined as the product of the countable set of independent measures $\pi(\cdot|x, i)$ over the set $\mathbb{X} \times I$; see Neveu [51, Proposition VI.2] for the existence and uniqueness of such products.

Theorem 5.19 (Feinberg [25], Theorem 3.1) *Let I be a countable set and let a measurable mapping $f : H \to I$ satisfy the Non-Repeating Condition. Then for any $C \in \mathcal{H}_\infty$ and for any randomized (f, I)-generated policy π*

$$\mathbb{P}_x^\pi(C) = \int_{I^f} \mathbb{P}_x^\phi(C) m^\pi(d\phi) \tag{5.15}$$

and therefore $v(x, \pi) = \int_{I^f} v(x, \phi) m(d\phi)$.

Corollary 5.8 *Let I be a countable set and let a measurable mapping $f : H \to I$ satisfy the Non-Repeating Condition. Then for any randomized (f, I)-generated policy π and for any state $x \in \mathbb{X}$ there exists a (nonrandomized) (f, I)-generated policy ϕ such that $v(x, \phi) \geq v(x, \pi)$.*

Thus, we have three groups of results that can help us to prove the existence of uniformly $\varepsilon\ell$-optimal policies in some class of nonrandomized policies. First, Theorems 5.1, 5.18, and relevant results for stationary policies related to occupation measures (see Krylov [46, 47], Borkar [13], Altman [1], and Feinberg and Sonin [37]) provide the methods to prove that for any initial state and policy there is an equivalent randomized policy in a given class of policies. Theorem 5.19, Theorem 5.17, or equality $s(x) = s_R(x)$ (see [36, 29] imply that a policy can be selected in a nonrandomized form. Theorems 5.10, 5.12, and 5.17 imply the existence of uniformly $\varepsilon\ell$-optimal policies within certain classes of policies.

5.9 THE SECOND MAIN THEOREM: UNIFORMLY NEARLY OPTIMAL LOCALLY STATIONARY POLICIES

Ornstein's theorem (Theorem 5.7(ii)) implies the existence of stationary uniformly nearly optimal policies in **P** MDPs. Example 5.5 implies that such policies do not exist in **N** MDPs. Demko and Hill [17] proved that if $r(x,a) = r(x) < 0$ all $x \in \mathbb{X}$ then for any $\varepsilon > 0$ there exists a stationary ε-optimal policy. This result can be interpreted in the following way: if rewards are negative in a strong sense then there exist stationary uniformly nearly optimal policies.

Since stationary nearly optimal policies may not exist (Example 5.5), it is natural to consider subsets of the state space on which such policies exist. We say that a nonrandomized policy ϕ is stationary on the set $Y \subseteq \mathbb{X}$ if $\phi_n(x_0 a_0 \dots x_n) = \phi(x_n)$ for some function ϕ when $x_n \in Y$. We denote by $\Pi^{S,Y}$ the set of policies stationary on Y. We observe that $\Pi^S = \Pi^{S,\mathbb{X}}$. We also denote

$$s_Y(x) = \sup_{\phi \in \Pi^{S,Y}} w(x,\phi).$$

We remark that $s_Y(x) = T s_Y(x)$ for all $x \in \mathbb{X}$ and for all $Y \subseteq \mathbb{X}$; Feinberg [27]. Let $\mathbb{A}^c(x)$ be the set of conserving actions at state $x \in \mathbb{X}$, $\mathbb{A}^c(x) = \{a \in \mathbb{A}(x) |\ T^a V(x) = V(x)\}$. We define the sets

$$\mathbb{X}^+ = \{x \in \mathbb{X} |\ V(x) > 0\}, \qquad \mathbb{X}^- = \{x \in \mathbb{X} |\ V(x) < 0\},$$
$$\mathbb{X}^c = \{x \in \mathbb{X} |\ \mathbb{A}^c(x) \neq \emptyset\}, \qquad \mathbb{X}^* = \mathbb{X}^+ \cup \mathbb{X}^- \cup \mathbb{X}^c.$$

Thus, $\mathbb{X}^+$ is the subset of states where the value function is positive, $\mathbb{X}^-$ is the subset where the value function is negative, and $\mathbb{X}^c$ is the set of states where conserving actions exist. We have that the set $\mathbb{X}^*$ contains all elements of $\mathbb{X}$ except those where the value function is equal to 0 and there is no conserving action. Chitashvili [73] showed that if $\mathbb{X}$ is finite then $s_{\mathbb{X}^*}(x) = V(x)$ for all $x \in \mathbb{X}$. Feinberg [26] proved the existence of uniformly nearly optimal policies in $\Pi^{S,\mathbb{X}^*}$ for the countable state space.

Let

$$d(x) = \sup_{\phi \in \Pi^{S,\mathbb{X}^*}} \inf_{\tau \in Q^b} \mathbb{E}_x^\phi V(x_\tau). \tag{5.16}$$

The following theorem is a particular case of Theorem 6.2 in Feinberg [26].

Theorem 5.20 *Let I be a countable set and let a measurable mapping $f : H \to I$ satisfy the Non-Repeating Condition. Consider the set I^f of (nonrandomized) (f,I)-generating policies. Then for any $\varepsilon > 0$ and for any $\ell \in L(d, \mathbb{X}_\Pi)$ there exists a policy ϕ with the following properties: (i) ϕ is stationary on $\mathbb{X}^*$, (ii) $\phi \in I^f$, and (iii) is uniformly $\varepsilon\ell$-optimal, i.e.*

$$v(x,\phi) \geq V(x) - \varepsilon \ell(x), \qquad x \in \mathbb{X}.$$

We observe that if $Y \subseteq Z$ then $s_Y(x) \geq s_Z(x)$ for all $x \in \mathbb{X}$. In view of Theorem 5.20 it is natural to find the biggest set Y for which $s_Y(x) = V(x)$ for all $x \in \mathbb{X}$. Unfortunately, even when $\mathbb{X}$ is finite, it is possible that there

are several maximum sets with this property (a subset Y is called a maximum subset with a given property if for any set Z with this property the inclusion $Z \supseteq Y$ implies $Z = Y$). The following example is similar to one provided by Chitashvili [16]. It shows that it is possible that $s_Y = V$ and $s_Z = V$ but $s_{\mathbb{X}\cup Y}(x) < V(x)$ for some x.

Example 5.12 Let $\mathbb{X} = \{0, 1, g\}$ and $\mathbb{A} = \{0\} \cup \{b^1, b^2, \dots\} \cup \{c^1, c^2, \dots\}$. The state g is absorbing, $\mathbb{A}(g) = \{0\}$, $r(g, 0) = 0$, and $p(g|g, 0) = 1$. Let $\mathbb{A}(x) = \{b^1, b^2, \dots\} \cup \{c^1, c^2, \dots\}$ for $x = 1, 2$. We also set $p(g|x, b^i) = 1/i$, $p(x|x, b^i) = 1 - 1/i$, $p(1|0, c^i) = p(0|1, c^i) = 1$, and $r(x, b^i) = r(x, c^i) = -1/i$ for $i = 1, 2, \dots$, $x = 0, 1$. It is easy to see that $V(g) = s(g) = 0$, $V(0) = V(1) = 0$, and $s(0) = s(1) = -1$. Let $Y = \{0, g\}$ and $Z = \{1, g\}$. Then $\mathbb{X} = Y \cup Z$. It is easy to see that $s_Y(x) = s_Z(x) = V(x)$, $x \in \mathbb{X}$. ∎

According to Example 6.1 in Feinberg [26], the following situation is possible. There are finite sets Z_n, $n = 1, 2, \dots$, such that $Z_n \subseteq Z_{n+1}$, $\mathbb{X} = \cup_{n=1}^{\infty} Z_n$, $s_{Z_n} = V$ but $s \neq V$. Since $s = s_{\mathbb{X}}$, this example shows that if $\mathbb{X}$ is countable, a maximal subset, for which there exist good policies stationary on it, may not exist. The following natural way to expand the set $\mathbb{X}^*$ was described in Feinberg [26].

Let $r(x, a) = r^1(x, a) + r^2(x, a)$ where r^1 and r^2 are measurable in a. We assume that r^2 is a nonnegative function and consider two MDPs with the same state space, the same action spaces, and the same transition probabilities as the original MDP but with the reward function r^1 and r^2 respectively. Let V^1 and V^2 are the value functions for these MDPs. We assume that $V^2(x) < \infty$ for all $x \in \mathbb{X}$. We also set $Z = \{x \in \mathbb{X} |\ V^1(x) < 0\}$. Of course, the set Z depends on the selection of r^2 which defines r^1 and V^1. For example, we can select $r^2(x, a) = 0$ for all x and a. In this case, $r = r + 0$ and $Z = \mathbb{X}^-$. We can select $r^2(x, a) = r^+(x, a)$. Then $r^1 = r^-$ and $Z = \{x \in \mathbb{X} |\ V_-(x) < 0.\}$ It is obvious that this set contains $\mathbb{X}^-$. We also can select $r^2 = kr^+$ where k is a constant or nonnegative bounded function of x and a.

We also remark that if a function r^2 is replaced with a function, that is greater or equal to $r^2(x, a)$ for all $x \in \mathbb{X}$ and for all $a \in \mathbb{A}(x)$, then the corresponding set Z expands. However, the function r^2 cannot be arbitrary large because of the condition $V^2(x) < \infty$ for all $x \in \mathbb{X}$.

Theorem 5.21 (The second main theorem; Feinberg [26]) *Let I be a countable set and let a measurable mapping $f : H \to I$ satisfy the Non-Repeating Condition. Consider the set I^f of (nonrandomized) (f, I)-generating policies. Let Z be a subset of $\mathbb{X}$ defined above by some nonnegative function r^2 with $V^2(x) < \infty$ for all $x \in \mathbb{X}$. Then for any $\varepsilon > 0$ and for any $\ell \in L(d, \mathbb{X}_\Pi)$ there exists a policy ϕ with the following properties: (i) ϕ is stationary on $\mathbb{X}^* \cup Z$, (ii) $\phi \in I^f$, and (iii) is uniformly $\varepsilon\ell$-optimal, i.e.*

$$v(x, \phi) \geq V(x) - \varepsilon\ell(x), \qquad x \in \mathbb{X}.$$

We remark that Theorem 5.21 implies almost all known results on the existence of uniformly nearly optimal policies. For example, for **P** MDPs, $\mathbb{X} = \mathbb{X}^*$

and Theorem 5.21 implies Ornstein's theorem. For $I^f = \Pi^M$, it implies van der Wal's [67] theorem on the existence of uniformly nearly optimal Markov policies. If I^f is selected to be the set of tracking policies, Theorem 5.21 gives the positive answer to the question asked by van der Wal and Wessels [69] on the existence of uniformly nearly optimal tracking policies. If $\mathbb{X}^0 \subseteq \mathbb{X}^c$, where $\mathbb{X}^0 = \{x \in \mathbb{X} |\ V(x) = 0\}$, then Theorem 5.21 implies the existence of stationary $\varepsilon\ell$-optimal policies and this result strengthens van der Wal's theorem on stationary policies [68]; see also van der Wal [66, Theorem 2.22] and Schäl [56] for relevant results. In addition, if $\mathbb{X}$ and $\mathbb{A}$ are finite, this result implies the existence of stationary optimal policies because the set of stationary policies if finite and the stationary $\varepsilon\ell$-optimal policy is optimal for small $\varepsilon > 0$. Theorem 5.21 also contains a statement that if there exists an optimal policy then there exists a stationary optimal policy.

There are two important open questions related to Theorem 5.21: (i) how to provide the broadest natural description of the set where stationary policies are sufficient; (ii) how to decrease the function d? With respect to the second question, it would be nice to select d in a form that it equals 0 for deterministic MDPs. Some results and discussion related to these two important questions can be found in Feinberg [24, 26]. In particular, Theorem 6.2 in Feinberg [26] contains a function d which is less than or equal to the function d defined in (5.16).

5.10 UNCOUNTABLE MDPS

Most of the results described above hold for Borel MDPs. In particular, Dynkin and Yushkevich [21] studied the sets of strategic measures $\mathbb{P}^\pi_\mu$. They showed that this is a convex Borel space. Furthermore, this set is convex in a strong sense when strategic measures are integrated with respect to any probability measure; see [21, Section 3.5]. The observation that this set is convex is important. Imagine that we expand the notion of a policy in the following way: the decision-maker selects randomly a policy at epoch 0 and then follows it. This initial randomization procedure is defined by a probability measure on the set of strategic measures with a given initial distribution or state. We call such policies mixed. The convexity of the set of strategic measures implies that for each mixed policy there is an equivalent randomized policy. Therefore, mixed policies in fact do not expand the set of policies. In view of this fact, policy π in Theorem 5.1 which was originally established by Strauch [61] for Borel MDPs and $\lambda_1 = 1$, can be constructed in two steps for a fixed initial state x: (i) for any sequence of nonnegative numbers λ_i with the sum equal to 1 and for any sequence of policies π^i, $i = 1, 2, \ldots$, there exists a policy σ with $\mathbb{P}^\sigma_x = \sum_{i=1}^{\infty} \mathbb{P}^{\sigma^i}_x$, and (ii) for any policy σ there exists a Markov policy π such that $\mathbb{P}^\pi_x\{x_n \in Y, a_n \in B\} = \mathbb{P}^\sigma_x\{x_n \in Y, a_n \in B\}$ for any Borel subsets $Y \subseteq \mathbb{X}$ and $B \subseteq \mathbb{A}$. In both cases (i) and (ii) there are simple formulae for σ and π.

Another general result for strategic measures is that strategic measures for randomized Markov policies can be presented as convex integral combinations of strategic measures for nonrandomized Markov policies; Feinberg [30]. This fact implies that for any policy and for any given initial distribution there

exists a Markov policy with equal or better performance; Feinberg [22, 23]. Schäl [54], Balder [3], and Yushkevich [72] studied topological properties of the sets on strategic measures and used them to establish sufficient conditions for the existence of optimal policies.

The optimality equation holds for Borel MDPs; see Strauch [61], Dynkin and Yushkevich [21], and Bertsekas and Shreve [5]. The major technical difficulty when a Borel state space is considered instead of a countable state space is related to so-called selection theorems. All results from Section 5.5, except Ornstein's theorem (Theorem 5.7(ii)), were originally established for classical Borel models and the existence of (p, ε)-optimal policies was proved.

Blackwell [9] recognized that the validity of Ornstein's theorem for Borel positive MDPs was a difficult question and posted it as an open problem. Frid [38] proved that Ornstein's theorem is valid for Borel MDPs if $(p, \varepsilon V)$-optimality is considered instead of εV-optimality. Schäl and Sudderth [59] found a correctable mistake in Frid's proof: one of the sets where Frid switched policies was universally measurable but not Borel measurable.

Here I would like to write few words about Efim Frid, a gifted mathematician who died at a young age as a result of an accident: he was hit by a vehicle when he was crossing a street in Odessa. Efim was a student of Nicolai Krylov. In addition to the proof of Blackwell's conjecture, Frid also wrote one of the first papers on MDPs with multiple criteria [39] and several interesting papers on stochastic games, in particular, on a sequence of nonzero sum two-person games. Although we both lived in Moscow at the same time for many years, unfortunately I never met Efim Frid.

Bertsekas and Shreve [5] studied MDPs with universally measurable policies. In particular, they proved Theorems 5.4(i) and 5.5(i) for Borel MDPs with universally measurable policies. Blackwell and Ramakrishnan [12] constructed an example of a **P** Borel MDP with a bounded function V for which there is no stationary universally measurable policy which is uniformly ε-optimal. This example implies that Ornstein's theorem as well as more general statements, Theorems 5.10, 5.12, 5.17, 5.20, and 5.21, cannot be expanded to Borel MDPs with universally measurable policies. The questions whether any of these statements or Theorem 5.16 hold for Borel MDPs when $(p, \varepsilon\ell)$-optimal policies are considered are completely open. The key issue here is Theorem 5.10. Its proof in Feinberg and Sonin [34] used explicitly that the state space is not bigger than countable. In addition to Frid [38] and Blackwell and Ramakrishnan [12], Schäl and Sudderth [59] studied extensions of Ornstein's theorem to uncountable state spaces and other relevant issues.

As observed in Feinberg [28], it is not clear whether the Non-Repeating Condition implies the result similar to Theorem 5.19 for Borel MDPs. It was shown there that the so-called Strong Non-Repeating Condition, which holds for Markov policies, implies such a result.

Feinberg [29] proved that $s_R(x) = s(x)$ for all $x \in \mathbb{X}$. Schäl [56] proved that $s(x) = V(x)$, $x \in \mathbb{X}$, for models with compact action sets. Since Theorem 5.10 is an open question for Borel MDPs when $(p, \varepsilon\ell)$-optimality is considered, we do not know if $s = Ts$ for **GC** Borel MDPS. Feinberg [30] proved that s is a universally measurable function and therefore Ts can be defined. This

fact is not trivial and the proofs in Feiberg [30] used measurability results by Sudderth [62] and Blackwell [10].

Acknowledgment. The author is grateful to Manfred Schäl for valuable comments. This research was partially supported by NSF Grant DMI-9908258.

Appendix: The proof of Theorem 5.2

Let $\mathcal{A}^*(E) = \{\emptyset, E\}$ for any set E. For $E \subseteq \mathbb{X} \times \mathbb{N}$ and for $(x,n) \in \mathbb{X} \times \mathbb{N}$ we define the σ-fields

$$\mathcal{A}_E(x,n) = \begin{cases} \mathcal{A}(A(x)), & \text{if } (x,n) \in E, \\ \mathcal{A}^*(A(x)), & \text{otherwise.} \end{cases}$$

For each set $E \subseteq \mathbb{X} \times \mathbb{N}$ we shall define a σ-field $\mathcal{F}_E$ generated by cylinder sets with the base in E,

$$\mathcal{F}_E = \times_{(x,n) \in \mathbb{X} \times N} \mathcal{A}_E(x,n).$$

We notice that $\mathcal{F}_E \subseteq \mathcal{F}_{E'}$ when $E \subseteq E'$. In particular, $\mathcal{F} = \mathcal{F}_{\mathbb{X} \times \mathbb{N}}$ and $\mathcal{F}_E \subseteq \mathcal{F}$ for all $E \subseteq \mathbb{X} \times \mathbb{N}$. By using the Ionesco Tulcea theorem [51, Proposition V.1.1], it is possible to show that for any fixed $x \in \mathbb{X}$ and $C \in \mathcal{H}_n$, the function $\phi \to \mathbb{P}_x^\phi(C)$ is $\mathcal{F}_{\mathbb{X} \times \{0,1,\ldots,n-1\}}$-measurable; see Lemmas 3.1 and 3.2 in Feinberg [25] for details.

If (5.7) holds for all $C \in \mathcal{H}_n$, $n \in \mathbb{N}$, it holds for all $C \in \mathcal{H}_\infty$ because measures P_x^σ can be continued from $\cup_{n=0}^\infty \mathcal{H}_n$ to $\mathcal{H}_\infty$. Therefore, it is sufficient to prove (5.7) for $C \in \mathcal{H}_n$, $n \in \mathbb{N}$. We shall do it by induction.

We have that $P_x^\sigma\{x_0 = y\} = \mathbf{I}\{x = y\}$ for any $y \in \mathbb{X}$ and for any policy σ. This implies that (5.7) holds for all $C \in \mathcal{H}_0$.

Let (5.7) holds for all $C \in \mathcal{H}_n$ for some $n = 0, 1 \ldots$. We take an arbitrary $C \in \mathcal{H}_n$ and an arbitrary $B \in \mathcal{A}$.

We observe that the function $\mathbb{P}_x^\phi(C')$ is $\mathcal{F}_{\mathbb{X} \times \{0,1,\ldots,n-1\}}$-measurable for all $C' \in \mathcal{H}_n$. The mapping $\phi(y,n)$ is $\mathcal{F}_{\mathbb{X} \times \{n\}}$-measurable. Since $(\mathbb{X} \times \{0,1,\ldots,n-1\}) \cap (\mathbb{X} \times \{n\}) = \emptyset$, the σ-fields $\mathcal{F}_{\mathbb{X} \times \{0,1,\ldots,n-1\}}$ and $\mathcal{F}_{\mathbb{X} \times \{n\}}$ are independent with respect to measure m^π. This implies that for any $D = \{x_0, a_0 \in B_0, x_1, \ldots, x_{n-1}, a_{n-1} \in B_{n-1}, x_n\}$

$$\int_{\Pi^M} \mathbf{I}\{\phi(x_n,n) \in B_n\} m^\pi(d\phi) \int_{\Pi^M} \mathbb{P}_x^\phi(D) m^\pi(d\phi) = \int_{\Pi^M} \mathbf{I}\{\phi(x_n,n) \in B_n\}\, \mathbb{P}_x^\phi(D) m^\pi(d\phi) \tag{5.17}$$

for all $n = 0, 1, \ldots$ and for all $x_t \in \mathbb{X}$, $B_t \in \mathcal{A}_t$, $t = 0, \ldots, n$.

We have that

$$\begin{gathered}\mathbb{P}_x^\pi\{C\times B\}=\int\limits_C \pi(B|x_n,n)\,\mathbb{P}_x^\pi(dh_n)=\\ \int\limits_C\int\limits_{\Pi^M}\mathbf{I}\{\phi(x_n,n)\in B\}m^\pi(d\phi)\int\limits_{\Pi^M}\mathbb{P}_x^\phi(dh_n)m^\pi(d\phi)=\\ \int\limits_C\int\limits_{\Pi^M}\mathbf{I}\{\phi(x_n,n)\in B\}\,\mathbb{P}_x^\phi(dh_n)m^\pi(d\phi)=\int\limits_C\mathbb{P}_x^\phi(C\times B)m^\pi(d\phi),\end{gathered}\tag{5.18}$$

where the first and the last equalities in (5.18) follow from the definition of measures P_x^σ, the second equality follows from the definition of the measure m^π and from the induction assumption, and the third equality in (5.18) follows from (5.17).

We have from (5.17) that (5.7) holds for any $C\in\mathcal{H}_n\times\mathcal{A}$. Consider arbitrary $C\in\mathcal{H}_n\times\mathcal{X}$ and $y\in\mathbb{X}$. We have

$$\begin{gathered}\mathbb{P}_x^\pi\{C\times\{y\}\}=\int\limits_C p(y|x_n,a_n)\,\mathbb{P}_x^\pi(dh_n)=\int\limits_C p(y|x_n,a_n)\int\limits_{\Pi^M}\mathbb{P}_x^\phi(dh_n)m^\pi(d\phi)\\ =\int\limits_{\Pi^M}m^\pi(d\phi)\int\limits_C p(y|x_n,a_n)\,\mathbb{P}_x^\phi(dh_n)=\int\limits_{\Pi^M}\mathbb{P}_x^\phi(C\times\{y\}\}m^\pi(d\phi),\end{gathered}\tag{5.19}$$

where the first and the last equalities in (5.19) follow from the definition of measures P_x^σ, the second one follows from the validity of (5.7) for $C\in\mathcal{H}_n\times\mathcal{A}$ established in (5.18), and the third one follows from Fubini's theorem. The theorem is proved.

References

[1] E. Altman, "Constrained Markov decision processes with total cost criteria: occupation measures and primal LP," *Math. Methods Oper. Res.* **43** pp. 45–72, 1996.

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

[3] E.J. Balder, "On compactness of the space of policies in stochastic dynamic programming," *Stoch. Proc. Appl.* **32** pp. 141–151, 1989.

[4] D.P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models,* Prentice-Hall, Englewood Cliffs, NJ, 1978

[5] D.P. Bertsekas and S. Shreve, *Stochastic Optimal Control: the Discrete Time Case,* Academic Press, New York, 1978 (reprinted by Athena Scientific, Belmont, 1996).

[6] D.P. Bertsekas and S. Shreve, "Existence of stationary optimal policies in deterministic optimal control," *J. Math. Anal. Appl.* **69**, pp. 607–620, 1979.

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

[8] D. Blackwell, "Discounted dynamic programming," *Ann. Math. Stat.* **36** pp. 226–235, 1965.

[9] D. Blackwell, "Positive dynamic programming," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* **1** pp. 415–418, University of California Press, Berkeley, 1967.

[10] D. Blackwell, "The stochastic processes of Borel gambling and dynamic programming," *Ann. Statist.* **4** pp. 370–374, 1976.

[11] D. Blackwell, D. Freedman and M. Orkin, "The optimal reward operator in dynamic programming," *Ann. Probab.* **2** pp. 926–941, 1974.

[12] D. Blackwell and S. Ramakrishnan, "Stationary plans need not be uniformly adequate for leavable, Borel gambling problems," *Proc. Am. Math. Soc.* **102** pp. 1024–1027, 1988.

[13] V.S. Borkar, "A convex analytic approach to Markov decision processes," *Prob. Theor. Relat. Fields* **78** pp. 583-602, 1988.

[14] R. Cavazos-Cadena, E.A. Feinberg and R. Montes-de-Oca, "A note on the existence of optimal policies in total reward dynamic programs with compact action sets," *Math. Oper. Res.,* to appear.

[15] R. Cavazos-Cadena and R. Montes-de-Oca, "Nearly optimal stationary policies in negative dynamic programming," *Math. Methods Oper. Res.* **49** pp. 441–456, 1999.

[16] R.Ya. Chitashvili, "On the existence of ε-optimal stationary policies for a controlled Markov chain," *Comm. Acad. Sci. Georgian SSR* **83** pp. 549–552, 1976 (in Russian).

[17] S. Demko and T.P.Hill, "Decision processes with total cost criteria," *Ann. Prob.* **9** pp. 293-301, 1981.

[18] E.V. Denardo, "Contracting mappings in the theory underlying dynamic programming," *SIAM Rev.* **9** pp. 165–177, 1967.

[19] C. Derman and R. Strauch, "A note on memoryless rules for controlling sequential control processes," *Ann. Math. Stat.* **37** pp. 276–278, 1966.

[20] L.E. Dubins and L.J. Savage, *How to Gamble If You Must: Inequalities for Stochastic Processes,* McGraw-Hill, New York; second edition: Dover, New York, 1976.

[21] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes,* Springer-Verlag, New York, 1979.

[22] E.A. Feinberg, "Nonrandomized Markov and semi-Markov strategies in dynamic programming," *SIAM Theory Prob. Appl.* **27** pp. 116–126, 1982.

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

[24] E.A. Feinberg, "The structure of persistently nearly optimal strategies in stochastic dynamic programming," *Lecture Notes in Control and Inform. Sci.* **81** pp. 22–31, 1986.

[25] E.A. Feinberg, "Sufficient classes of strategies in discrete dynamic programming. I: decomposition of randomized strategies and embedded models," *SIAM Theory Prob. Appl.* **31** pp. 658–668, 1987.

[26] E.A. Feinberg, "Sufficient classes of strategies in discrete dynamic programming. II: locally stationary strategies," *SIAM Theory Prob Appl.* **32** pp. 478–493, 1987.

[27] E.A. Feinberg, "Parametric stochastic dynamic programming," *Statistics and Control of Stochastic Processes, Steklov Seminar,* **2** (eds. A.N. Shiryaev et al.), Optimization Software, New York, pp. 103–120, 1989.

[28] E.A. Feinberg, "Nonrandomized strategies in stochastic decision processes," *Ann. Oper. Res.* **29** pp. 315–332, 1991.

[29] E.A. Feinberg, "On stationary strategies in borel dynamic programming," *Math. Oper. Res.* **17** pp. 393–397, 1992.

[30] E.A. Feinberg, "On measurability and representation of strategic measures in Markov decision processes," in *Statistics, Probability and Game Theory Papers in Honor of David Blackwell* (eds. T.S. Ferguson, L.S. Shapley and J.B. MacQueen), IMS Lecture Notes - Monograph Series, **30**, pp. 29–43, 1996.

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

[32] E.A. Feinberg and A. Shwartz, "Constrained discounted dynamic programming," *Math. Oper. Res.* **21** pp. 922–945, 1996.

[33] E.A. Feinberg and I.M. Sonin, "Markov policies in infinite horizon dynamic programming problems with bounded value functions," in *Abstracts of 4-th USSR - Japan Symposium on Probability Theory and Mathematical Statistics* **1** pp. 209–210, Tbilisi, 1982.

[34] E.A. Feinberg and I.M. Sonin, "Stationary and Markov policies in countable state dynamic programming," *Lecture Notes in Math.,* **1021** pp. 111–129, 1983.

[35] E.A. Feinberg and I.M. Sonin, "Sufficient classes of strategies in controllable Markov chains with total criterion," *Soviet Math. Dokl.* **29** pp. 308–311, 1984.

[36] E.A. Feinberg and I.M. Sonin, "Persistently nearly optimal strategies in stochastic dynamic programming," *Statistics and Control of Stochastic Processes, Steklov Seminar* (eds. N.V. Krylov, R.Sh. Liptser and A.A. Novikov), Optimization Software, New York, pp. 69–101, 1985

[37] E.A. Feinberg and I.M. Sonin, "Notes on equivalent stationary policies in Markov decision processes with total rewards," *ZOR - Math. Methods of Oper. Res.* **44** pp. 205–221, 1996.

[38] E.B. Frid, "On a problem of D. Blackwell from the theory of dynamic programming," *SIAM Theory Prob Appl.* **15** pp. 719–722, 1970.

[39] E.B. Frid, "On optimal strategies in control problems with constraints," *SIAM Theory Prob Appl.* **17** pp. 188–192, 1972.

[40] I.I. Gikhman and A.V. Skorokhod, *Controlled Random Processes,* Springer, New York, 1979.

[41] T. Hill, "On the existence of good Markov strategies," *Trans. Amer. Math. Soc.* **247** pp. 157–176, 1979.

[42] K. Hinderer, *Foundations of Non Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research **33**, Springer-Verlag, NY, 1970.

[43] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Math. Centre Tracts **51**, Math. Centrum, Amsterdam, 1974.

[44] L.C.M. Kallenberg, *Linear Programming and Finite Markovian Problem*, Math. Centre Tracts **148**, Math. Centrum, Amsterdam, 1983.

[45] N.V. Krylov, "Construction of an optimal strategy for a finite controlled chain," *SIAM Theory Prob. Appl.* **10** pp. 45–54, 1965.

[46] N.V. Krylov, "Once more about the connection between elliptic operators and Itô's stochastic equations," in *Statistics and Control of Stochastic Processes, Steklov Seminar*(eds. N.V. Krylov, R.Sh. Liptser, and A.A. Novikov), Optimization Software, New York, pp. 69–101, 1985

[47] N.V. Krylov, "An approach in the theory of controlled diffusion processes," *SIAM Theory Prob. Appl.* **31** pp. 604–626, 1987.

[48] A. Maitra and W.D. Sudderth, "The optimal return operator in negative dynamic programming," *Math. Oper. Res.* **17** pp. 921–931, 1992.

[49] A. Maitra and W.D. Sudderth, *Discrete Gambling and Stochastic Games,* Springer-Verlag, New York, 1996.

[50] D. Ornstein, "On the existence of stationary optimal strategies," *Proc. Am. Math. Soc.* **20** pp. 563–569, 1969.

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

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

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

[54] M. Schäl, "On dynamic programming: compactness of the space of policies," *Stoch. Processes Appl.* **3** pp. 345–364, 1975.

[55] M. Schäl, "Conditions for optimality in dynamic programming and for the limit of n-stage policies to be optimal," *Zeitschr. Wahrsch. Th. verw. Gebieten* **32** pp. 179–196, 1975.

[56] M. Schäl, "Stationary policies in dynamic programming models under compactness assumptions," *Math. Oper. Res.* **8** pp. 366–372, 1983.

[57] M. Schäl, "Markovian decision models with bounded finite-state rewards," *Operations Research Proceedings* 1983, Springer, Berlin, pp. 470–473, 1984.

[58] M. Schäl, "On stochastic dynamic programming: a bridge between Markov decision processes and gambling," *Markov Processes and Control Theory* (eds. H. Langer and V. Nolkau), Mathematical Research 54, Academie-Verlag, Berlin, pp. 178–216.

[59] M. Schäl and W.D. Sudderth, "Stationary policies and Markov policies in Borel dynamic programming," *Probab. Th. Rel Fields* **74** pp. 91–111, 1987.

[60] S. Stidham, "On the convergence of successive approximations in dynamic programming with non-zero terminal rewards," *Z. Operations Res.* **25** pp. 57–77, 1981.

[61] R. Strauch, "Negative dynamic programming," *Ann. Math. Stat.* **37** pp. 871–890, 1966.

[62] W.D. Sudderth, "On the existence of good stationary strategies," *Trans. Amer. Math. Soc.* **126** pp. 399–414, 1969.

[63] R. van Dawen, "Negative dynamic programming," Preprint no. 458, University of Bonn, 1985.

[64] R. van Dawen, "Pointwise and uniformly good stationary strategies in dynamic programming models," *Math. Oper. Res.* **9** pp. 521-535 (1986).

[65] R. van Dawen and M. Schäl, "On the existence of Stationary Optimal Policies in Markov Decision Processes," *Z. Angew. Math. Mech.* **63**, no 5, pp. T403-T404, 1983.

[66] J. van der Wal, *Stochastic Dynamic Programming*, Math. Centre Tracts **139**, Math. Centrum, Amsterdam, 1981.

[67] J. van der Wal, "On uniformly nearly-optimal Markov strategies," in *Operations Research Proceedings* 1982, pp. 461–467. Springer-Verlag, Berlin, 1983.

[68] J. van der Wal, "On stationary strategies in countable state total reward Markov decision processes," *Math. Oper. Res.* **9** pp. 290–300, 1984.

[69] J. van der Wal and J. Wessels, "On the use of information in Markov decision processes," *Statistics & Decisions* **2**, pp. 1–21, 1984.

[70] K.M. van Hee, "Markov strategies in dynamic programming," *Math. Oper. Res.* **3** pp. 37–41, 1978.

[71] K.M. van Hee, A. Hordijk and J. van der Wal, "Successive approximations for convergent dynamic programming," in *Markov Decision Theory* (eds. H. Tijms and J. Wessels), Math. Centre Tracts **93**, Math. Centrum, Amsterdam, pp. 183–211, 1977.

[72] A.A. Yushkevich, "The compactness of a policy space in dynamic programming via an extension theorem for Caratheodory functions," *Math. Oper. Res.* **22**, pp. 458–467, 1997.

[73] A.A. Yushkevich and R.Ya. Chitashvili, "Controlled random sequences and Markov chains," *Russian Math. Surveys* **37**, no 6, pp. 239-274, 1982.

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

6 MIXED CRITERIA

Eugene A. Feinberg
Adam Shwartz

Abstract: Mixed criteria are linear combinations of standard criteria which cannot be represented as standard criteria. Linear combinations of total discounted and average rewards as well as linear combinations of total discounted rewards are examples of mixed criteria. We discuss the structure of optimal policies and algorithms for their computation for problems with and without constraints.

6.1 INTRODUCTION

The discounted cost criterion is widely and successfully used in various application areas. When modeling economic phenomena, the discount factor models the value of money in time and is determined by return rate (or interest rate) or, in a more general context, by the "opportunity costs" which presume that a dollar now is worth more that a dollar in a year. Being invested, current funds will bring an additional return in a year. In other areas, such as the control of communications networks, the discounted criterion may reflect the imprecise but fundamental principle that future occurrences are less important than immediate ones. In reliability, discounting models systems with geometric life time distributions.

Obviously, if some part of the cost decreases (in time) at an exponential rate, then a discounted cost arises. This is the case, for example, in production processes. When a new item is manufactured, we expect some production costs to decrease as production methods are improved. Obviously there is a learning curve for all involved, along which various costs decrease. If the effect of this learning diminishes geometrically (or exponentially), then the total cost (over the infinite time-horizon) is of the discounted type. This would also be the case if the cost of obtaining a component decreases at an exponential rate. Such an

exponential decrease is evident in the computers industry (and one aspect goes by "Moore's law"): prices of various components decrease at an exponential rate, and since the processing speed increases at an exponential rate, the "unit cost" of processing power decreases exponentially fast.

However, the rates (or discount factors) for these different mechanisms are clearly unrelated. When combining several such costs (such as processing speed with economic considerations), we are naturally led to deal with several different discount factors, each applicable to a component of our optimization problem. Multiple discount factors also arise in control problems for systems with multiple parallel unreliable components. For details on applications see Feinberg and Shwartz [14, 15, 17]. Similarly, in stochastic games it is natural to consider situations where each player has its own discount factor.

In contrast to the discounted criterion, the average cost measures long-term behavior and, usually is insensitive to present and short-term conditions. Naturally, this criterion is appropriate for other applications, such as the long-term performance of systems. As before, if our criteria include system performance (measured through the average cost) as well as rewards (measured through a discounted cost), we are led to problems with mixed criteria.

In this paper we review results concerning such criteria, and point out some open questions. We shall mention specific, as well as general areas of potential applications. The main theoretical questions are:

- Existence of good (optimal, ε-optimal) policies for optimization, multi-objective optimization and in particular constrained optimization problems,
- Structure of "good policies," and
- Computational schemes for the calculation of the value and good policies.

We emphasize the mixed-discounted problem, where the criteria are all of the discounted type, but with several discount factors, since it possesses a rich structure and, in addition, much is already known. In Section 6.7 we shall review some other related results. We conclude this section with a brief survey of different mixed criterion problems.

There are several treatments of discounted models where the discounting is more general than the standard one. Hinderer [27] investigates a general model where the discounting is a function of the complete history of the model. In a number of papers, see e.g. Schäl [41], Chitashvili [7], Stidham [43], or Haurie and L'Ecuyer [22], the discount rate depends on the current state and action. This type of discounting arises when a discounted semi-Markov Decision process is converted into an MDP; see Puterman [37] or Feinberg [13] for details.

Mixed criteria and, in particular, mixed discounted criteria are linear combinations of standard criteria. The first two papers dealing with mixed criteria were published in 1982. Golabi, Kulkarni, and Way [21] considered a mixture of average reward and total discounted criteria to manage a statewide pavement system, see also [44]. Feinberg [10] proved for various standard criteria, by using a convex-analysis approach, that for any policy there exists a non-randomized Markov policy with the same or better performance. In the same paper, Feinberg [10], proved that property for mixtures of various criteria.

The 1990's saw systematic interest in criteria that mix several standard costs, which we now survey briefly. Krass, Filar and Sinha [29] studied a sum of a standard average cost and a standard discounted cost for models with finite state and action sets. They proved that for any $\varepsilon > 0$ there exists an ε-optimal randomized Markov policy which is stationary from some epoch N-onwards (so-called ultimately deterministic or (N, ∞)-stationary policies). They also provided an algorithm to compute such policies. For the special case of a *communicating MDP*, this policy is non randomized. As explained in Feinberg [11], the use of results from [10] simplifies the proofs in [29] and leads to the direct proof that for any $\varepsilon > 0$ there exists an ε-optimal (nonrandomized) Markov ultimately deterministic policy. The latter result can be derived from [29] but was formulated there only for the constant gain case.

Fernandez-Gaucherand, Ghosh and Marcus [18] considered several weighted as well as overtaking cost criteria. They treat general (Borel) state and action spaces with bounded costs, and their objective is a linear combination of several discounted costs and an average cost. They prove that, for any $\varepsilon > 0$, there exists $N(\varepsilon)$ and an ε-optimal deterministic policy which coincides with the average-optimal policy after time $N(\varepsilon)$.

Ghosh and Marcus [20] consider a similar problem in the context of a continuous time diffusion model with positive costs. There is one average component and one discounted component. It is shown that it suffices to consider Markov (randomized) policies. Under global stability conditions they show the following. There is a minimizer in the class of stationary randomized controls and randomization is not necessary. For the more general minimization problem, there exist ε-optimal policies that agree with the optimal policy for the discounted part until a fixed time, and then switch to the policy which is optimal for the average part of the cost.

Hernández-Lerma and Romera [25] noticed that the "minimal pair" approach by Lasserre and Hernández-Lerma [23, 24] can be applied to multiple criteria problems when each criterion is of the form of either average rewards per unit time or total discounted rewards where different criteria may have different discount factors. The minimal pair approach (see Chapter 11) means that a controller selects a pair (μ, π) where μ is an initial distribution and π is a policy. In classical problems, the initial distribution is either fixed or a controller optimizes with respect to all initial states. Hernandez-Lerma and Romera [25] considered an additional condition which essentially means that π is randomized stationary and μ is an invariant distribution of the Markov chain with the transition probabilities defined by π. Naturally, for a one-step reward function r, in this case expected rewards at step t are equal to $\beta^{t-1}\mu r$, where β is the discount factor. Therefore, the average rewards per unit time are equal to μr and the expected total discounted rewards are equal to $\mu r/(1-\beta)$. Therefore, the change of the original formulation by using the minimal pair approach and the additional invariant condition on the initial measure μ reduce the problem with combined criteria (average rewards per unit time and discounted rewards) to a version of a problem with only average rewards per unit time.

Filar and Vrieze [19] considered a stochastic zero-sum game with finite state and action spaces. The objective is a linear combination: either of an average cost and a discounted cost, or of two discounted costs with two different discount factors. The immediate costs are the same for all components. They prove that in both cases, the value of the stochastic zero-sum game exists. Moreover, when the objective combines two discounted costs, both players possess optimal Markov policies. When the objective combines a discounted cost with an average cost, there exist ε-optimal (behavioral) policies.

Most of the papers on mixed criteria deal either with linear combinations of total discounted rewards with different discount factors or with a linear combination of total discounted rewards and average rewards per unit time. It appears that linear combinations of discounted rewards are more natural and easier to deal with than the weighted combinations of discounted and average rewards. The latter ones model the situation when there are two goals: a short-term goal modeled by total discounted rewards and a long-term goal modeled by average rewards per unit time. In the case of two different discount factors, the weighted discounted criterion models the same situation when one of the discount factors is close to 1. When the state and action sets are finite, optimal policies exist for mixed discounted criteria, they satisfy the Optimality Equation, and can be computed. Optimal policies may not exist for mixtures of total discounted rewards and average rewards per unit time [29]. Another advantage of dealing with mixed discounting is that for this criterion there is a well-developed theory for any finite number of discount factors [14, 15], while the papers that study linear combinations of discounted and average reward criteria usually deal with linear combinations of only two criteria: discounted rewards with a fixed discount factor and average rewards per unit time.

Single discount problems are often interpreted as total reward problems with a geometric life time: however, as shown in Shwartz [42], mixed discount problems cannot, in general, be interpreted as problems with several time scales.

In this paper we concentrate on a mixed-discounted problem for Markov decision processes. The exposition is based on the detailed study of this problem, performed by Feinberg and Shwartz [14, 15, 17].

In Section 6.2 we describe the model more precisely, and show through Example 6.1 that, although the weighted criterion seems like a small variation on a standard discounted problem, it induces quite different behavior on the resulting optimal policies. Then, in Section 6.3 we show that mixed-discounted problems can be reduced to standard discounted problems if we expand the state space $\mathbb{X}$ to $\mathbb{X} \times \mathbb{N}$, and that Markov policies are sufficient for one-criterion mixed-discounted problems. Section 6.4 obtains the characterization as well as an algorithm for the computation of optimal policies for the Weighted Discount Optimization problem (**WDO**) with finite state and action sets. **WDO** problems are introduced in Definition 6.1. In Section 6.5 we treat multiple-criterion problems and in Section 6.6 we discuss finite constrained problems. In Section 6.7 we survey existing results for other relevant problems, related models of stochastic games, and discuss extensions and open problems.

6.2 THE MIXED DISCOUNTED PROBLEM

Throughout this chapter we deal with a discrete state model, as described in Chapter 0, and follow notation introduced there. We fix L discount factors which, for convenience (and without loss of generality) we order as $1 > \beta_1 > \ldots > \beta_L > 0$, and $(K+1) \times L$ one-step reward functions $r_\ell^k,\ k = 0, \ldots, K,\ \ell = 1, \ldots, L$. We assume that all reward functions are bounded above. The index k will be used only for multi-objective problems, and is omitted otherwise. Let $v_\ell^k(x, \pi, \beta_\ell)$ denote the standard total expected discounted cost corresponding to discount factor β_ℓ and to the immediate reward r_ℓ^k. We formulate the following optimization problems:

Definition 6.1 *The Weighted-Discount Optimization Problem* **WDO** *is to maximize the weighted-discount cost* $v_{\mathbb{M}}$ *over all policies* $\pi \in \Pi^R$, *where*

$$v_{\mathbb{M}}(x, \pi) \triangleq \sum_{\ell=1}^{L} v_\ell(x, \pi, \beta_\ell)\,. \tag{6.1}$$

Definition 6.2 *The Constrained Weighted-Discount Optimization Problem* **WDC** *is the constrained optimization over all policies* $\pi \in \Pi^R$

$$\textit{maximize} \qquad v_{\mathbb{M}}^0(x, \pi) \triangleq \sum_{\ell=1}^{L} v_\ell^0(x, \pi, \beta_\ell) \tag{6.2}$$

$$\textit{subject to} \qquad v_{\mathbb{M}}^k(x, \pi) \triangleq \sum_{\ell=1}^{L} v_\ell^k(x, \pi, \beta_\ell) \geq C_k, \qquad k = 1, \ldots, K. \tag{6.3}$$

A policy π *is* feasible *if the constraints* (6.3) *are satisfied.*

Given any numbers $a_1, a_2, \ldots, a_q$ we use the notation $\bar{a} \triangleq (a_1, a_2, \ldots, a_q)$. In particular we define

Definition 6.3 *The* performance vectors *associated with problems* **WDO** *and* **WDC** *respectively are*

$$\bar{v}(x, \pi) \triangleq (v_1(x, \pi, \beta_1), \ldots, v_L(x, \pi, \beta_L))\,, \tag{6.4}$$

$$\bar{v}_{\mathbb{M}}(x, \pi) \triangleq \left(v_{\mathbb{M}}^0(x, \pi), \ldots, v_{\mathbb{M}}^K(x, \pi)\right)\,. \tag{6.5}$$

We remark that the sum of the elements of the vector $\bar{v}(x, \pi)$ is the objective function for the **WDO** problem. As was discussed in Chapter 0, for unconstrained problems in general and for problem **WDO** in particular, we consider optimality with respect to all initial states. For a constrained problem, including problem **WDC**, an initial state is fixed. A **WDO** (**WDC**) problem for an MDP with finite state and action sets is called a *finite* **WDO** (**WDC**) problem.

6.2.1 An example: job versus education dilemma

The use of different discount criteria induces a time-dependence on the model since the relative impact of different immediate costs changes over time. This

implies, for example, that we cannot expect stationary policies to be optimal. This can be demonstrated with a very simple model. This negative result suggests that we search for different structural properties of optimal policies. It turns out that the structure suggested here is useful for other criteria as well, including single-discount constrained problems [16].

Example 6.1 illustrates the following dilemma. A person, say a high school or college graduate, has a choice: to accept a job offer or to continue his/her education and get a better job later. In a standard discounted model, stationary policies are optimal. Therefore, for standard models an optimal decision is either to accept a job or to continue education. For weighted discounted models, an optimal decision can suggest to accept a job for a limited period of time and then to continue education. This phenomenon cannot be modeled by standard discounted or average-reward criteria.

Example 6.1 ([14, Example 1.1]) *Consider a* **WDO** *problem. Let* $\mathbb{X} = \{x, y\}$ *with deterministic transitions: under* a *we always go to state* x*, while under* b *we always go to state* y*. Set* $r_\ell^k = r$*, where*

$$r(x,a) = 1\,, \qquad r(y,a) = r(x,b) = 0\,, \quad \textit{and} \quad r(y,b) = 2\,. \tag{6.6}$$

It is then easy to calculate that for the standard discounted cost, if $\beta \le \frac{1}{2}$*, then it is optimal to stay where you are, while for* $\beta \ge \frac{1}{2}$ *it is optimal to use only action* b*.*
For the weighted problem with $\beta_1 = \frac{3}{5}$ *and* $\beta_2 = \frac{1}{5}$*, an explicit calculation shows that the only optimal policy is to stay where you are at time* 0*, and use* b *at all later epochs.*

Another illuminating conclusion from the same example is obtained by searching for the best stationary policy. This turns out to be a randomized one! Related examples [14, Examples 1.2–1.3] show that the best stationary (non-randomized) policy may depend on the initial state. In addition, this behavior can be observed in ergodic models.

Thus, it seems that much of the basic structure of MDPs is lost when mixed discounting is used. However, it turns out that a different structure arises. In fact, that the optimal policy in Example 6.1 is Markov and becomes stationary after some initial period is a structure we shall discover in the following sections.

6.3 GENERAL PROPERTIES

Without loss of generality, we restrict our attention to the class of randomized Markov policies; see Chapter 5, Theorem 1. This general argument is well-known for average and total reward MDPs.

The main difficulty with the mixed-discounted problem is that the immediate cost changes over time and the rates of change are different for different summands in the objective function. In addition to substituting an arbitrary policy with a randomized Markov one, another technique of general applicability is the embedding of the problem into one with a larger state space. Consider an auxiliary standard discounted model with the discount factor β_1 and with

the state space $\mathbb{X} \times \mathbb{N}$. The one-step rewards in this model are equal to

$$r(x,n,a) \triangleq r_\ell(x,a) + \sum_{\ell=2}^{L} \left(\frac{\beta_\ell}{\beta_1}\right)^n r_\ell(x,a)\,. \tag{6.7}$$

If we then keep the same transition probabilities (but require a transition of one unit in the time component at each step), then we have a one-to-one correspondence between the original problem and the auxiliary problem started at $(x,0)$. There is only one immediate reward and one discount factor in the larger model. The state space for the auxiliary problem remains countable. We therefore obtain immediately the following results [14, Section II].

Theorem 6.1 ([14, Theorems 2.1–2.2]) *(i) For any $\varepsilon > 0$, the* **WDO** *problem possesses ε-optimal Markov policies. (ii) If the $\mathbb{A}(x)$ are compact subsets of a metric space, r_ℓ are upper semi-continuous and the $p(y|x,\cdot)$ are continuous in a then there are optimal Markov policies.*

Proof outline. This follows from the properties of the auxiliary problem. Note that a stationary policy for the auxiliary problem defines a Markov (but not necessarily stationary!) policy for the original problem. ∎

The above construction transforms a mixed-discounted problem with a finite or countable state spaces into a standard discounted problem with a countable state space. If the original problem has an uncountable Borel state space, so does the expanded problem.

We end this section with some insight into why problems with mixed criteria are inherently more difficult. For simplicity, let the state and action sets be finite. Define the expected occupation vectors

$$f(\beta;x,\pi;y,a) \triangleq \sum_{t=0}^{\infty} \beta^t \,\mathbb{P}_x^\pi \left(x_t = y, a_t = a \middle| x_0 = x\right)\,. \tag{6.8}$$

Then we can write the standard discounted cost as

$$v_\ell^k(x,\pi) = \sum_{y\in\mathbb{X}} \sum_{a\in\mathbb{A}(y)} f(\beta;x,\pi;y,a) r_\ell^k(y,a)\,. \tag{6.9}$$

That is, any discounted cost is a linear function of the occupation vectors $\{f(\beta;x,\pi;y,a) : y \in \mathbb{X},\ a \in \mathbb{A}(y)\}$. Moreover, these occupation vectors obey a system of linear equalities, in terms of transition probabilities. It is therefore possible to transform the optimization problem, as well as the constrained optimization problem, into a linear program. This linear program is finite if the state and action spaces are finite. This approach is described in many papers and in the books by Kallenberg [28], Borkar [6], Piunovskiy [36], Altman [1], and Hernandez-Lerma and Lasserre [24]. However, the relation between $\{f(\beta_1;x,\pi;y,a) : y \in \mathbb{X},\ a \in \mathbb{A}(y)\}$, the occupation vectors associated with discount β_1, and $\{f(\beta_2;x,\pi;y,a) : y \in \mathbb{X},\ a \in \mathbb{A}(y)\}$, the occupation vectors associated with discount β_2, is non-linear (in fact, it is even non convex!). This makes the tools of mathematical programming much more difficult to apply; see [12].

6.4 SINGLE CRITERION MODELS

Motivated by the structure we found in Example 6.1, we introduce some new notions which will be fundamental beyond this section. The lost time-homogeneity of the model is partly recovered by the notion of a funnel.

Definition 6.4 *Given measurable subsets* $\mathbb{A}_1(x) \subset \mathbb{A}(x),\ x \in \mathbb{X}$, *the* submodel $\mathbb{A}_1$ *is the Markov decision process, where the actions at* x *are restricted to* $\mathbb{A}_1(x)$.

Definition 6.5 ([15, Definition 5.4]) *Fix a positive integer* N *and subsets* $\mathbb{A}_n(x) \subset \mathbb{A}(x),\ n \geq 0,\ x \in \mathbb{X}$, *with the property that* $\mathbb{A}_n(x) = \mathbb{A}_N(x),\ n \geq N,\ x \in \mathbb{X}$. *The* funnel *associated with these data is the set of all randomized Markov policies* π *such that* $\pi_n(\mathbb{A}_n(x)|x) = 1$ *for all* $n \geq 0$ *and* $x \in \mathbb{X}$.

A funnel is thus defined by the number $N \in \mathbb{N}$ and sets $\mathbb{A}_n(x),\ n = 0, 1, \ldots, N,\ x \in \mathbb{X}$.

Definition 6.6 *Given a positive integer* N, *a Markov policy* π *is called* (N, ∞)-*stationary if there exists a stationary policy* ϕ *so that*

$$\pi_n(x) = \phi(x) \quad \textit{for all} \quad x \quad \textit{and} \quad n \geq N\ . \tag{6.10}$$

This generalizes the notion of stationarity, since obviously a $(0, \infty)$-stationary policy is stationary. As we shall see in Corollary 6.1, there is an optimal (N, ∞)-stationary policy for the **WDO** problem.

Let each set $\mathbb{A}(x)$ be finite and all functions $r_\ell(x, a)$ be bounded on $\mathbb{X} \times \mathbb{A}$, $\ell = 1, \ldots, L$. Define recursively, for $\ell = 1, \ldots, L$,

$$\Gamma_0(x) = \mathbb{A}(x)\,, \quad N_0(x) = N_0 = 0, \tag{6.11}$$

$$d_\ell \triangleq \frac{\sup_{x \in \mathbb{X}, a \in \Gamma_{\ell-1}(x)} r_\ell(x, a) - \inf_{x \in \mathbb{X}, a \in \Gamma_{\ell-1}(x)} r_\ell(x, a)}{1 - \beta_\ell}\,, \tag{6.12}$$

$$V_\ell(x) = \sup\{v_\ell(x, \pi, \beta_\ell) \,|\, \pi \text{ in submodel } \Gamma_{\ell-1}\}\ , \tag{6.13}$$

$$\Gamma_\ell(x) = \{a \in \Gamma_{\ell-1}(x) : V_\ell(x) = r_\ell(x, a) + \beta_\ell P^a V_\ell(x)\}\ . \tag{6.14}$$

If $\Gamma_\ell(x) = \Gamma_{\ell-1}(x)$ set $N_\ell(x) = N_{\ell-1}(x)$. Otherwise define

$$\varepsilon_\ell(x) = V_\ell(x) - \sup\{r_\ell(x, a) + \beta_\ell P^a V_\ell(x) : a \in \Gamma_{\ell-1}(x) \setminus \Gamma_\ell(x)\}\ , \tag{6.15}$$

$$N_\ell(x) = \min\left\{ t \geq N_{\ell-1}(x) \,\Big|\, \sum_{j=\ell+1}^{L} \left(\frac{\beta_j}{\beta_\ell}\right)^t d_j < \varepsilon_\ell(x) \right\}, \quad \ell < L\,, \tag{6.16}$$

$$N_\ell = \sup_x N_\ell(x), \tag{6.17}$$

and set $N_L(x) \triangleq N_{L-1}(x)$, $N = N_L = N_{L-1}$. If the state space is finite, $N < \infty$.

The set $\Gamma_\ell(x)$ is the set of *conserving* actions for the discounted criterion $v_\ell(x, \pi, \beta_\ell)$ in submodel $\Gamma_{\ell-1}$. The basic structure of optimal policies derives

from the following statement which follows from equalizing and thrifty properties; see Chapter 5.

Lemma 6.1 *A policy π is optimal for the criterion $v_\ell(x,\pi,\beta_\ell)$ in submodel Γ_ℓ, namely $v_\ell(x,\pi,\beta_\ell) = V_\ell(x)$ for all $x \in \mathbb{X}$, if and only if $a_t \in \Gamma_\ell(x_t)$ $\mathbb{P}_x^\pi$-a.s. for all $t = 0, 1, \ldots$ and for all $x \in \mathbb{X}$.*

In other words, all policies in submodel Γ_ℓ have the same v_ℓ-cost, and the value $V_\ell(x)$ is the optimal value in the $\Gamma_{\ell-1}$ model.

Theorem 6.2 *Suppose that all action sets $\mathbb{A}(x)$ are finite and all reward functions r_ℓ are bounded. Consider problem* **WDO**. *If π is an optimal policy, then for any $\ell = 1, \ldots, L$, any $x \in \mathbb{X}$ and any $t \geq N_\ell(x_t)$,*

$$\pi_t(x_t) \in \Gamma_\ell(x_t) \quad \mathbb{P}_x^\pi - a.s. \text{ for any } t \geq N_\ell(x_t), \text{ for all } \ell, \text{ and for all } x \in \mathbb{X}. \tag{6.18}$$

Proof outline. Write

$$v_{\mathbb{M}}(x,\pi) = v_1(x,\pi,\beta_1) + \sum_{\ell=2}^{L} v_\ell(x,\pi,\beta_\ell)\,. \tag{6.19}$$

Suppose at time t we are at state y. By Lemma 6.1, if we choose an action outside $\Gamma_1(y)$, then our v_1 reward will be smaller by at least $\beta_1^t \varepsilon_1(y)$. On the other hand, in view of (6.7), the second summand in (6.19) can be made larger by at most

$$\sum_{j=2}^{L} \beta_j^t d_j\,. \tag{6.20}$$

The result for $\ell = 1$ follows from the definitions (6.15)–(6.16) since $\beta_j < \beta_1$ for $j > 1$. Therefore we know that, after time N_1, we must restrict to submodel Γ_1. By Lemma 6.1, the v_1 cost is the same for all policies in this model, so that this component of the cost may be ignored. Repeating the same argument establishes the result for $\ell = 2, \ldots, L$. ∎

Theorem 6.2 is formulated for non-finite state spaces. However, in the finite case N is finite and it leads directly to existence as well as to an algorithm. Define the time-dependent immediate reward

$$r(t,x,a) \triangleq \sum_{\ell=1}^{L} \beta_\ell^t r_\ell(x,a)\,, \tag{6.21}$$

and the "tail reward"

$$v_\ell^{>}(x,N,y,\pi) \triangleq \mathbb{E}_x^\pi \left[\sum_{t=N}^{\infty} \beta_\ell^t r_\ell(x_t,a_t) \,\middle|\, x_N = y \right]. \tag{6.22}$$

If π is a Markov policy, then $v_\ell^{>}(x,N,y,\pi)$ does not depend on x.

Theorem 6.3 *Consider a finite* **WDO** *problem. Let* Θ *be the funnel defined by* $\mathbb{A}_t(x) = \Gamma_\ell(x)$ *for* $N_{\ell-1}(x) \le t < N_\ell(x)$, *and* $\mathbb{A}_t(x) = \Gamma_L(x)$ *for* $t \ge N$. *Let* ϕ *be any stationary policy in submodel* Γ_L. *Then (i) any optimal policy must satisfy* $v_\ell^>(x, N, y, \pi) = v_\ell^>(x, N, y, \phi)$ *for all* x *and all* y *such that* $\mathbb{P}_x^\pi(x_N = y) > 0$; *(ii) an optimal policy to* **WDO** *can be constructed as follows. Let* $\pi^N = \{\pi_0, \pi_1, \dots, \pi_{N-1}\}$ *solve the finite-horizon total cost problem with horizon* N, *immediate rewards* $r(t, x, a)$, *and terminal reward*

$$R(y) \triangleq \sum_{\ell=1}^{L} v_\ell^>(x, N, y, \phi)\,. \tag{6.23}$$

Then $\pi^* = (\pi_0, \pi_1, \dots, \pi_{N-1}, \phi, \phi, \dots)$ *is optimal.*

Proof outline. Part (i) follows from Theorem 6.2. This determines the "tail reward," and it remains to optimize over the finite horizon. ∎

The "tail" policy ϕ can be chosen stationary, and from the algorithm it follows that it does not depend on the initial state. Since these properties are also shared by solutions to finite-horizon problems, we obtain the following.

Corollary 6.1 *For a finite* **WDO** *problem there is an optimal* (N, ∞)*-stationary policy (which does not depend on the initial state).*

Formulas (6.11)–(6.17) provide an algorithm that computes an integer N and sets $\Gamma_L(x)$, $x \in \mathbb{X}$, described in Theorem 6.3. In view of Theorem 6.3, the "tail" stationary policy ϕ can be selected as any stationary policy from submodel Γ_L. Theorem 6.3 also implies that, at steps $0, 1, \cdots, N-1$, an optimal policy can be constructed by a finite-horizon dynamic programming algorithm. Thus, formulas (6.11)–(6.17) and Theorem 6.3 provide an algorithm that computes an optimal (N, ∞)-stationary policy.

This algorithm requires the computation of at most L standard discounted problems, and then the solution of a finite horizon problem. The computational complexity is obviously influenced by the size of N. This in turn in determined by the data of our problem (and in particular the ratios $\beta_\ell/\beta_{\ell-1}$), as well as by the choice of bound (e.g. (6.12)). A more complex algorithm leading to a smaller value of N is in [14].

In order to compute R in (6.23) and to find the "tail" policy ϕ, one should solve L standard discounted MDPs. By selecting a large planning horizon, it is possible to compute ε-optimal policies without computing R. To do this, we can select N directly without using the procedures described after Definition 6.6. Indeed, given $\varepsilon > 0$ select an integer N which is large enough to satisfy $\sum_{\ell=1}^{L} \beta_\ell^N d_\ell \le \varepsilon$. Then the policy $\tilde{\pi} = \{\pi_0, \pi_1, \dots, \pi_{N-1}, \phi, \phi, \dots\}$ is ε-optimal, where $\{\pi_0, \pi_1, \dots, \pi_{N-1}\}$ is an optimal Markov policy for the finite N-problem with the terminal reward equal to 0 (see Chapter 5 for details) and ϕ is an arbitrary stationary policy.

For any criterion $v(x,\pi)$ and set of policies Δ we denote

$$v(x,\Delta) = \{v(x,\pi) : \pi \in \Delta\} , \tag{6.24}$$

$$V(x,\Delta) = \sup\{v(x,\pi) : \pi \in \Delta\} , \tag{6.25}$$

$$\Delta_v^*(x) = \{\pi \in \Delta : v(x,\pi) = V(x,\Delta)\} . \tag{6.26}$$

Let Θ be a funnel. The embedding technique of Section 6.3 allows us to construct a finite model, where time becomes part of the new state space, but only until time N. A funnel in this new model corresponds to a funnel in the original MDP. Therefore Theorem 6.3 implies the following result.

Corollary 6.2 ([15, Lemma 5.5]) *Consider an MDP with finite state and action sets. Fix x and let $\Theta \subset \Pi^R$ be a non-empty funnel. Then there exists a funnel Θ' so that $v_{\mathbb{M}}(x,\pi) = V_{\mathbb{M}}(x,\Theta)$ for all $\pi \in \Theta'$, and moreover $\bar{v}(x,\Theta') = \bar{v}(x,\Delta^*_{v_{\mathbb{M}}}(x))$.*

Corollary 6.2 means that, given $x \in \mathbb{X}$, for any funnel Θ there is a funnel Θ', which is a non-empty subset of Θ and (i) $v_{\mathbb{M}}(x,\pi) \geq v_{\mathbb{M}}(x,\sigma)$ for all $\pi \in \Theta'$ and for all $\sigma \in \Theta$, (ii) if $\sigma \in \Theta$ and $v_{\mathbb{M}}(x,\sigma) \geq v_{\mathbb{M}}(x,\gamma)$ for all $\gamma \in \Theta$ then there exists policy $\pi \in \Theta'$ such that $\bar{v}(x,\pi) = \bar{v}(x,\sigma)$.

6.5 MULTIPLE CRITERION OPTIMIZATION

To describe some notions of optimality in the multiple criterion setting we need some definitions and notation. Recall the definition 6.3 of the *performance vectors* associated with problems **WDO** and **WDC** respectively.

Definition 6.7 *The* performance spaces *are, respectively,*

$$U_o(x) \triangleq \left\{\bar{v}(x,\pi) : \pi \in \Pi^R\right\} , \tag{6.27}$$

$$U_c(x) \triangleq \left\{\bar{v}_{\mathbb{M}}(x,\pi) : \pi \in \Pi^R\right\} . \tag{6.28}$$

For a vector $\bar{a}$ in $\mathbb{R}^q$ we write $\bar{a} \geq 0$ if and only if $a_i \geq 0$ for all i.

Definition 6.8 *A point $\bar{u}$* dominates *a point $\bar{v}$ if $\bar{u} - \bar{v} \geq 0$. A point $\bar{u}$ is* Pareto optimal *in a set U if there is no other $\bar{v} \in U$ which dominates $\bar{u}$. We write $(u_1, u_2, \ldots, u_q) = \bar{u} >_\ell 0$ if $u_i = 0$ for $i = 1, \ldots, j-1$ and $u_j > 0$ for some $1 \leq j \leq q$ ($j = 1$ implies $u_1 > 0$). Say $\bar{u}$ is* lexicographically larger *than $\bar{v}$ if $\bar{u} - \bar{v} >_\ell 0$.*

Note that $\bar{u} - \bar{v} \geq 0$ and $\bar{u} \neq \bar{v}$ implies $\bar{u} - \bar{v} >_\ell 0$, but the converse need not hold. We extend these notions from vectors to policies in the obvious way:

Definition 6.9 *For a fixed initial state x, a policy π is called* Pareto optimal *if the corresponding performance vector is Pareto optimal. A policy π is called* lexicographically optimal *if the corresponding performance vector is lexicographically optimal.*

With these definitions, we have the following obvious statement.

Lemma 6.2 *Any optimal policy for* **WDO** *is Pareto optimal for $\bar{v}(x,\pi)$.*

Theorem 6.3 immediately implies

Corollary 6.3 *Any policy in submodel Γ_L, and in particular ϕ, is lexicographically optimal for $\bar{v}(x,\pi)$.*

The performance space has the following convenient structure.

Theorem 6.4 *(i) The sets $U_c(x)$ are convex, $x \in \mathbb{X}$. (ii) If the $\mathbb{A}(x)$ are compact, all reward and transition functions are continuous in a and the rewards are bounded, then $U_c(x)$ are compact. (iii) If Θ is a funnel then $\bar{v}_{\mathbb{M}}(x,\Theta)$ are convex and if, in addition, the conditions of (ii) hold then $\bar{v}_{\mathbb{M}}(x,\Theta)$ are compact. (iv) The same conclusions hold for $U_o(x)$.*

Proof outline. Convexity of $U_c(x)$ follows from Chapter 5, Theorem 1. Compactness follows from compactness of $\left\{\mathbb{P}_x^\pi : \pi \in \Pi^R\right\}$ and continuity; see [15, Lemma 3.5], which is based on [41]. The extension to $\bar{v}_{\mathbb{M}}(x,\Theta)$ follows by the argument preceding Corollary 6.2. Finally, note that $U_o(x)$ is a special case of $U_c(x)$. ∎

6.5.1 Classes of policies

Multiple criterion problems often require randomization in order to achieve optimality. In order to quantify the amount of randomization, and to tie this with the notion of (N,∞)-stationarity, we introduce the following classes of policies.

We say that a randomized Markov policy π is *discrete* if each of the probabilities $\pi_t(\cdot|x)$, $t \in \mathbb{N}$, $x \in \mathbb{X}$ has a finite or countable support. We recall that any randomized stationary policy is randomized Markov.

Definition 6.10 *A randomized stationary policy ϕ is called* M-randomized stationary *if it is discrete and*

$$\sum_{x\in\mathbb{X}}\left[\sum_{a\in\mathbb{A}(x)}\mathbf{1}[\phi(a|x)>0]-1\right]\le M\,. \tag{6.29}$$

A randomized Markov policy π is called randomized Markov of order M *if it is discrete and*

$$\sum_{t=0}^{\infty}\sum_{x\in\mathbb{X}}\left[\sum_{a\in\mathbb{A}(x)}\mathbf{1}[\pi_t(a|x)>0]-1\right]\le M\,. \tag{6.30}$$

Note that the terms in square brackets are always non-negative. An M-randomized stationary policy randomizes every time when the process reaches a state where the support of $\phi(x)$ contains more than one point. By contrast, a randomized Markov policy of order M makes at most M randomizations over the entire time-horizon.

Definition 6.11 *A Markov policy π is called an* (m,N)-policy *if it is randomized Markov of order m and, in addition, it is (N,∞)-stationary, that is,*

it agrees with some stationary policy ϕ after time N. An (m,N)-policy π is called a strong (m,N)*-policy if, in addition, there is an m-randomized stationary policy ψ such that*

$$\pi_t(a|x) > 0 \quad \textit{implies} \quad \psi(a|x) > 0\,. \tag{6.31}$$

Thus (m,N)-policies have the simple structure of at most m randomizations over the entire time horizon, as well as stationarity beyond time N. For a strong (m,N)-policy, the total number of actions is further restricted in that the total number of actions beyond those of a stationary (non-randomized!) policy does not exceed m.

6.6 FINITE MODELS: CONSTRAINED OPTIMIZATION

6.6.1 Finite horizon models

As we saw, construction of an optimal policy in the weighted-discount problem goes through the computation of finite-horizon problems. We note in passing that Theorem 6.4 applies to finite-horizon problems with arbitrary time-dependence of the reward functions.

To define the finite horizon constrained optimization problem, we use Definition 6.2. However, we let the immediate rewards depend on time by setting

$$r_\ell^k(t,x,a) = \begin{cases} r_\ell^k(x,a) & \text{if } t < N, \\ f_\ell^k(x) & \text{if } t = N, \\ 0 & \text{if } t > N, \end{cases} \tag{6.32}$$

where $f_\ell^k(x)$ is a terminal reward for the reward function r_ℓ^k and discount factor β_ℓ. We usually set $f_\ell^k(x) = v_\ell^k(x,\phi,\beta_\ell)$, where ϕ is a stationary policy. For the finite horizon problem, it is possible to use a Linear Programming approach. Here, the mixed criteria is not a hindrance: if fact, this approach applies for a time-dependent cost structure, by an embedding technique as in Section 6.3.

Theorem 6.5 ([15, Theorem 4.1]) *A finite horizon, finite state and action constrained optimization problem is feasible if and only if the associated LP [15, (4.1)–(4.5)] is feasible. If it is feasible then it has an optimal randomized Markov policy of order K.*

6.6.2 Infinite horizon models

The proof of the existence of optimal (K,N)-policies requires some convex analysis. We shall relate special subsets of performance spaces to funnels. We need the following definition.

Definition 6.12 *Let W be a convex subset of a convex set E. Call W* extreme *if the relation $u_3 = \lambda u_1 + (1-\lambda)u_2$, where $0 < \lambda < 1$, $u_1, u_2 \in E$, and $u_3 \in W$ implies that necessarily $u_1, u_2 \in W$. Call W* exposed *if there is a supporting hyperplane H of E so that $W = H \cap E$.*

An exposed set is extreme, but the converse may not hold: take

$$E = \{(x,y) : -1 \le x \le 0, |y| \le 1\} \cup \{(x,y) : x \ge 0, x^2 + y^2 \le 1\}$$

and $W = \{(0,1)\}$. Then W is obviously extreme, but the only supporting hyperplane containing W satisfies $H \cap E = \{(x,y) : -1 \le x \le 0, y = 1\} \neq W$, so that W is not exposed. Note that $(0,1)$ is a Pareto-optimal point, and is a solution of the constrained optimization problem of maximizing y subject to $x \ge 0$. However, it is not isolated by any exposed subset of E.

Our plan is to show that Pareto optimal points of $U_c(x)$ are achieved by (K,N)-policies. We first show that boundary points of $U_c(x)$ are achieved by points of the set $\bar{v}_{\mathbb{M}}(x,\Theta)$ for some funnel Θ. We then show that any boundary point is achieved by a convex combination of performances of (N,∞)-stationary policies that utilize the same stationary policy from epoch N onwards. The properties of the finite-horizon problems of Theorem 6.5 are used to conclude optimality of (K,N)-policies.

Theorem 6.6 *Let Θ be a funnel and W an exposed subset of $\bar{v}_{\mathbb{M}}(x,\Theta)$. Then there exists a funnel Θ' such that $W = \bar{v}_{\mathbb{M}}(x,\Theta')$. If $E \neq \bar{v}_{\mathbb{M}}(x,\Theta)$ is an extreme subset, then then there exists a funnel Θ' such that $E = \bar{v}_{\mathbb{M}}(x,\Theta')$. In particular, these statements hold for $\Theta = \Pi^R$ and $U_c(x) = \bar{v}_{\mathbb{M}}(x,\Theta)$.*

Proof outline. A supporting hyperplane H and exposed subset W are defined by some $b, b_0, \ldots, b_K$ so that

$$H = \left\{ \bar{u} : \sum_{k=0}^{K} b_i u_i = b \right\},$$

$$W = \left\{ \bar{u} \in \bar{v}_{\mathbb{M}}(x,\Theta) : \sum_{k=0}^{K} b_i u_i = b \right\},$$

and

$$\sum_{k=0}^{K} b_i u_i \le b \quad \text{for all} \quad \bar{u} \in \bar{v}_{\mathbb{M}}(x,\Theta)\,.$$

Apply Corollary 6.2 to conclude the first result. For the extreme subset we use the fact that for a proper extreme subset E of a compact convex set $\tilde{W}$ in an Euclidean space there is a finite sequence of sets $W_0, \ldots, W_j$ such that $W_0 = \tilde{W}$, $W_j = E$, and W_{i+1} is an exposed subset of W_i; see the proof of Lemma 6.3 in [15]. The first result, applied repeatedly to the sets $\tilde{W} = \bar{v}_{\mathbb{M}}(x,\Theta), W_1, \ldots, W_{j-1}$, leads to the second statement of the theorem. ∎

When $\tilde{W} = U_c(x)$ or, in a more general situation, $\tilde{W} = \bar{v}_{\mathbb{M}}(x,\Theta)$, where Θ is a funnel, then Theorem 6.6 implies that for any proper extreme subset E of $\tilde{W}$ there is a funnel Θ' such that $E = \bar{v}_{\mathbb{M}}(x,\Theta')$. If $\bar{u}$ is an extreme point of $\bar{v}_{\mathbb{M}}(x,\Theta)$ for some funnel Θ (that is, the singleton $\{\bar{u}\}$ is an extreme subset), then its performance is achieved by a funnel, and in particular we can choose any policy π from this funnel and have $\bar{v}_{\mathbb{M}}(x,\pi) = \bar{u}$. For N large enough, we select π being (N,∞)-stationary.

If $\bar{u}$ is a Pareto optimal point of $U_c(x)$ then it belongs to the boundary of the closed convex set $U_c(x)$. Therefore, it belongs to an exposed subset E of $U_c(x)$

and, according to Theorem 6.6, E can be represented as a performance set of a funnel; $E = V(x, \Theta')$. If $\bar{u}$ is an extreme point of E, $\bar{u} = \bar{v}_{\mathbb{M}}(x, \pi)$ for some (N, ∞)-stationary policy π. If $\bar{u}$ is a relatively inner point of E, Caratheodory theorem implies that it can be represented as a convex combination of at most K extreme points of E.

Such representation holds if each extreme point is approximated by an element of E close to it. By selecting N large enough, we can approximate (N, ∞)-stationary policies, whose performance vectors are extreme points of E, with (N, ∞)-stationary policies coinciding with the same stationary policy ϕ from epoch N onwards. Thus, if $\bar{u}$ is a Pareto optimal element of $U_c(x)$, it can be represented as a convex combination of performance vectors of (N, ∞)-stationary policies with the same "tail," i.e. these policies act as the same stationary policy from some epoch N onwards; see Figure 6.1, where a, b, and c are extreme points of $E = V(x, \Theta')$ and a', b', and c' are their approximations which are performance vectors of (N, ∞)-stationary policies with the same "tail."

Theorem 6.7 ([15, Theorems 6.6–6.8]) *(i) If $\bar{u}$ is a Pareto optimal point of $U_c(x)$ then for some $N < \infty$ there exists a (K, N)-policy π such that $\bar{v}_{\mathbb{M}}(x, \pi) = \bar{u}$. (ii) If the* **WDC** *is feasible then for some $N < \infty$ there exists an optimal (K, N)-policy.*

Proof outline. (i) Fix N and the "tail" stationary policy ϕ described in the paragraph preceding the theorem. Set $\pi_t(y) = \phi(y)$ for all $y \in \mathbb{X}$ and for all $t \geq N$. In order to determine the policy π at steps $t = 0, \ldots, N$, one has to solve a constrained finite-horizon problem with $C_k = u^k$, $k = 1, \ldots, K$, where $\bar{u} = (u^0, \ldots, u^K)$. By Theorem 6.5, π is a (K, N)-policy.

(ii) Any solution of the **WDC** defines either a Pareto optimal point of $U_c(x)$ or it is dominated by a solution with this property. Therefore, (i) implies (ii). ∎

Theorem 6.8 *Fix x and consider performance vectors $\bar{v}_{\mathbb{M}}(x, \pi)$ and performance space $U_c(x)$. If $\bar{u}$ belongs to the boundary of $U_c(x)$ then there is a (K, N)-policy π with $\bar{v}_{\mathbb{M}}(x, \pi) = \bar{u}$. If $\bar{v}$ is any point in $U_c(x)$ then there is a $(K + 1, N)$-policy σ with $\bar{v}_{\mathbb{M}}(x, \sigma) = \bar{v}$.*

Proof outline. Since U is convex and compact, any point on the boundary of $U_c(x)$ can be represented as the unique solution to a constrained optimization problem, with K constraints, so Theorem 6.7 implies the result. Any point in the interior of $U_c(x)$ can be represented by a similar constrained problem, but with $K + 1$ constraints, and similar arguments apply. ∎

In general, it is not possible to achieve a given performance with (K, N)-policies, so that the result above is sharp.

6.6.3 Calculation of optimal policies

The computation of optimal policies for the constrained problem is, in general, an open problem. It is easy to compute approximate policies, provided that by

"approximate" we mean that we allow the constraints to be "slightly violated." To do this, given ε we fix a large N so that

$$\frac{K\beta_1^N \max_{a,x} |r_\ell^k(x,a)|}{1-\beta_1} < \varepsilon \tag{6.33}$$

and solve the finite horizon problem, ignoring all costs after N. This would put costs and constraints within ε of the desired values; see [15] for details.

A relaxation technique can be used to decrease the error, either in the constraints, or the value, or both. However, such algorithms are iterative, and it is difficult to obtain information about their accuracy.

Consider the case $K = 1$, and where

$$v_{\mathbb{M}}^0(x,\pi) = v_1(x,\pi,\beta_1)\,, \tag{6.34}$$

$$v_{\mathbb{M}}^1(x,\pi) = v_2(x,\pi,\beta_2). \tag{6.35}$$

That is, each criterion is a simple discounted one, but the discounts are different, $\beta_1 \neq \beta_2$. For the next result we do not assume that $\beta_1 > \beta_2$. Let the problem be feasible, that is

$$\max_\pi v_{\mathbb{M}}^1(x,\pi) \geq C\,. \tag{6.36}$$

We say a policy π is (a,b)-lexicographic optimal if it is lexicographic optimal for the vector $(v_{\mathbb{M}}^a(x,\pi), v_{\mathbb{M}}^b(x,\pi))$.

Theorem 6.9 ([17]) *(i) If* $\max_\pi v_{\mathbb{M}}^1(x,\pi) = C$ *then the* $(1,0)$*-lexicographic optimal policy solves the constrained optimization problem. (ii) Let* σ *be the*

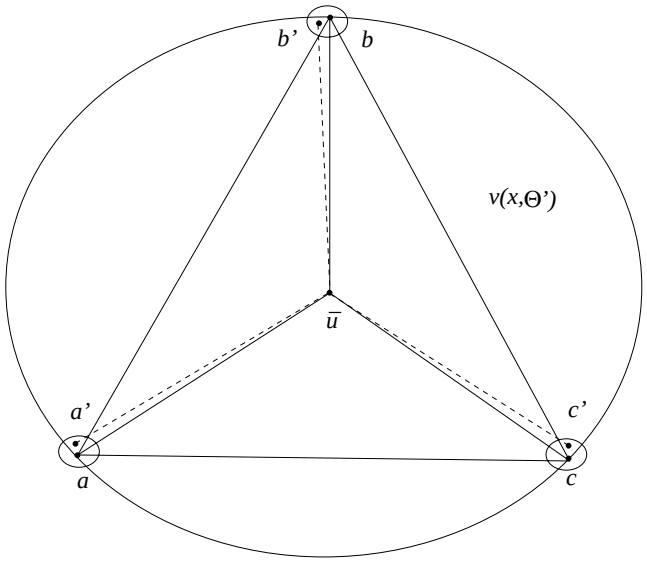

Figure 6.1 Representation of a Pareto optimal point $\bar{u}$ as a convex combinations of performance vectors a', b', and c' of (N,∞)-stationary policies with the same "tail."

$(0,1)$-lexicographic optimal policy, and suppose $v^1_{\mathbb{M}}(x,\sigma) \geq C$. Then σ is an optimal solution. (iii) If neither conditions hold and $\beta_1 > \beta_2$, then there is a finite algorithm for the computation of the optimal policy. The complexity of the algorithm is similar to the solution of the **WDO** *problem. (iv) If on the other hand $\beta_1 < \beta_2$ then there is an iterative algorithm, that terminates in a finite number of steps, for the computation of optimal policies.*

We note that even for this simple problem, in case (iv) we have no prior estimate of the complexity of this calculation.

6.6.4 Single-discount constrained optimization

The problem of constrained optimization with the standard discounted criterion has been extensively studied; see the books by Kallenberg [28], Borkar [6], Piunovskiy [36], and Altman [1]. However, the non-stationary policies introduced in this chapter give this problem a different perspective. Indeed, if $\mathbb{X}$ and $\mathbb{A}$ are finite then Theorem 6.7 states the existence of optimal (K,N)-policies. This is a new result for problems with a single discount factor! However, for problems with a single discount factor, this result can be strengthened. If $\mathbb{X}$ and $\mathbb{A}$ are finite and a problem is feasible, then there exist an optimal randomized stationary policy [28]. Standard linear programming arguments [39] imply that, if the problem is feasible, then there exists a K-randomized stationary optimal policy. Combined with Theorem 6.7, this result implies the existence of strong (K,N)-policies for some $N < \infty$ for models with finite state and action sets.

If the state space is infinite, optimal (N,∞)-stationary policies may not exist for unconstrained mixed discounted problems [14]. Therefore, optimal (K,N)-policies may not exist for constrained mixed discounted models with infinite state spaces. However, as was proved in [16], optimal strong (K,N)-policies exist for constrained discounted problems with countable state spaces if these models satisfy standard continuity assumptions. We give here a brief survey of the results of [16].

We treat the countable state model of this chapter, and make the continuity assumptions of Theorem 6.1(ii). We consider a constrained problem with K constraints. We consider the constrained problem (6.2)–(6.3) when, instead of a vector $\bar{v}_{\mathbb{M}}(x,\pi)$, the performance of a policy π is evaluated by a vector $\bar{v}(x,\pi) = (v^0(x,\pi),\dots,v^K(x,\pi))$, where $v^k(x,\pi)$ are expected discounted total rewards for reward functions $r^k(x,a)$ and the common discount factor $\beta \in [0,1[, k = 0,\dots,K$. Note that our definitions of (N,∞)-stationary, K-randomized, randomized Markov of order K, and (K,N)-policies are all well-posed. We note that, in this generality, the set $\bar{v}\left(x,\Pi^R\right)$ may not be compact because it may be unbounded However, out assumptions suffice for the following.

Lemma 6.3 *If $\bar{u}$ belongs to the closure of $\bar{v}\left(x,\Pi^R\right)$ then there exists $\bar{u}' \in \bar{v}\left(x,\Pi^R\right)$ that dominates $\bar{u}$. Consequently, there exists a policy whose performance dominates $\bar{u}$.*

Theorem 6.10 *If π is Pareto optimal then (i) there exists a K-randomized stationary policy with the same performance, and (ii) there exists a strong (K, N)-policy with the same performance.*

Theorem 6.11 *If problem* **WDC** *is feasible, then (i) there exists an optimal K-randomized stationary policy, and (ii) there exists an optimal strong (K, N)-policy.*

The strengthening of the conclusions from (K, N)-policies is worth a comment. Using conclusion (i) of Theorem 6.11, we obtain a K-randomized optimal policy σ. Consider now the submodel $\mathbb{A}'$ where

$$\mathbb{A}'(x) = \{a \in \mathbb{A}(x) : \sigma(a|x) > 0\}. \tag{6.37}$$

In this submodel, all but at most K of the $\mathbb{A}'(x)$ are singletons. We now obtain an optimal (K, N)-policy in the new model. By definition, this policy is a strong (K, N)-policy.

6.7 RELATED PROBLEMS AND CRITERIA

In this section we survey some related MDP problems and some extensions to stochastic games. Other related models are mentioned in the introduction.

6.7.1 MDP models

Average cost criteria are usually of the expected type. However, it is well known that for ergodic models the value can be achieved with probability one. This is also the case for constrained MDPs; see Borkar [6] and Altman and Shwartz [4]. Ross and Varadarajan [40] consider a finite MDP and maximize the expected average cost subject to a constraint that another average cost does not exceed a given bound with probability one. For a general multichain MDP they establish that if the problem is feasible, then there is an ε-optimal stationary policy. An algorithm for its computation is provided.

Reiman and Shwartz [38] consider a mixed-criteria problem that arises in telecommunications. Arriving users may be rejected. If accepted, they generate communication packets according to an independent random process until they leave. There are two average per unit time optimization criteria determining the Quality of Service. The percentage of lost packets by a user that is accepted at a given state (given number of users) should be below or equal to a given bound. The probability of blocking (rejecting an arriving user) should be minimized. Due to the nature of the model, only stationary policies are relevant, and the fact that users leave after a geometric session time implies that the first criterion is actually of the discounted type. Since the bound must hold for every initial state, we have a mixed criterion problem with average cost optimization and a countable number of discounted constraints. The authors provide an algorithm for the computation of optimal policies and derive a relation to a mathematical program.

6.7.2 *Stochastic games with mixed criteria*

Filar and Vrieze [19], Altman, Feinberg, and Shwartz [3], and Altman, Feinberg, Filar, and Gaitsgory [2] investigated stochastic games with weighted criteria. Filar and Vrieze [19] considered zero-sum games with finite state and action sets. They considered two criteria: a mixture of two discounted criteria and a mixture of a discounted and average reward criteria. In both cases, they proved the existence of the value and the existence of randomized (N, ∞)-stationary ε-optimal policies for $\varepsilon > 0$. Altman, Feinberg, and Shwartz [3] provided an example when optimal (N, ∞)-stationary policies do not exist in a mixed-discounted problem and proved the existence of such policies in models with perfect information.

Altman, Feinberg, and Shwartz [3] also introduced lexicographic games where the players play the game with the payoff function r_1 and discount factor β_1 first. We denote this game by Γ_1. Then the players play the game with the discount factor β_2 and reward functions r_2 on the set of optimal policies of the game Γ_1. We denote this game by $\Gamma_{1,2}$. This construction can be repeated L times and, as the result, the players play the game $\Gamma_{1,\dots,L}$. In games with perfect information, when players play optimal (N, ∞)-stationary policies for mixed-discounted games, from epoch N onwards they play any optimal policy for the game $\Gamma_{1,\dots,L}$.

We recall that a mixed-discounted game with finite or countable state space can be reduced to a standard discounted countable state game with essentially the same action spaces ([14]). This construction is briefly described in Section 6.3. Therefore, if the action sets are finite, the set of optimal actions at each step for each player exists at each state. This set is a polytope which is a subset of the set of all probability distributions on the sets of all actions $A(x)$ for player one and $B(x)$ for player two. We denote theses sets of optimal actions by $\mathbf{A}_n(x)$ and $\mathbf{B}_n(x)$ respectively. Altman, Feinberg, Filar, and Gaitsgory [2] proved for repeated mixed-discounted games that the sequence of sets $\mathbf{A}_n$ (there is only one state in repeated games and therefore $\mathbf{A}_n(x)$ do not depend on x) converges to a subset of the set of optimal policies of the game $\Gamma_{1,2}$. Whether this result holds for stochastic games with finite state and action sets is an open question. We also remark that the examples in [2] show that the limit may not be equal to the set of optimal policies in game $\Gamma_{1,2}$ and that this limit may not be a subset of the sets of optimal policies for the game $\Gamma_{1,2,3}$.

In general, the existence of values for zero-sum games and equilibrium values for non-zero sum games are nontrivial questions. For mixed criteria, we are aware of two general methods to prove the existence of such values. The first method is to represent a mixed criterion as a limsup criterion and use the results by Maitra and Sudderth [31, 32, 33]. The second method, which can be applied directly to mixed-discounted criteria, is to consider an expanded model described in Section 6.3 and then to apply the results for standard discounted criteria [30, 35, 9]. The existence of Nash equilibria is often established using fixed-point methods. Altman and Shwartz [5] consider the following stochastic game. We have L players. Player ℓ has a discount factor β_ℓ (where $\beta = 1$ means that the average cost is used) and immediate costs r_ℓ^k, $0 \le k \le B_\ell$. A

policy π is called *feasible* if

$$v_\ell^k(x,\pi,\beta_\ell) \le V_\ell^k \quad \text{for } 1 \le \ell \le L,\ 1 \le k \le B_\ell, \tag{6.38}$$

where V_ℓ^k are given numbers. It is established in [5] that, if this problem is feasible then (under some regularity conditions) there exists a Nash equilibrium. An ergodicity condition is required if the average cost is used by some player. The proof uses fixed point methods.

6.7.3 Extensions and open problems

The authors are currently considering the extension of the mixed-discounted problem to the semi-Markov setting.

Except for unconstrained problems, the algorithmic aspects of mixed-discounted criteria are still open: we do not have an algorithm for the computation of optimal, or even feasible ε-optimal policies for finite **WDC** problems. Finding optimal policies within classes of stationary and randomized stationary policies are *NP*-hard problems; see [12].

Convergence of solutions for zero-sum games to the subsets of solutions of lexicographic games, established in [2] for repeated games, is an open question for stochastic games with finite state and action sets.

Acknowledgment

Research of the first author was partially supported by NSF Grant DMI-9908258. Research of the second author was partially supported by the funds for promotion of research and the promotion of sponsored research at the Technion.

References

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

[2] E. Altman, E. Feinberg, J.A. Filar, and V.A. Gaitsgory, "Perturbed zero-sum games with applications to dynamic games," *Annals of the International Society of Dynamic Games* **6** pp. 165–181, 2001.

[3] E. Altman, E.A. Feinberg, and A. Shwartz, "Weighted discounted stochastic games with perfect information," *Annals of the International Society of Dynamic Games* **5** pp. 303–323, 2000.

[4] E. Altman and A. Shwartz, "Sensitivity of constrained Markov decision processes," *Ann. Operations Research* **32** pp. 1–22, 1994.

[5] E. Altman and A. Shwartz, "Constrained Markov games: Nash equilibria," *Annals of the International Society of Dynamic Games* **5** pp. 213–221, 2000.

[6] V.S. Borkar, *Topics in Controlled Markov Chains*, Longman Scientific & Technical, Harlow, 1991.

[7] R. Ya. Chitashvili, "A finite controlled Markov chain with small break probability," *SIAM Theory Probability Appl.* **21** pp. 157–163, 1976.

[8] C. Derman and R.E. Strauch, "A note on memoryless rules for controlling sequential processes," *Ann. Math. Stat.* **37** pp. 272–278, 1966.

[9] A. Federgruen, "On N-person stochastic games with denumerable state spaces," *Ad. Appl. Prob* **10** pp. 452–471, 1978.

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

[11] E.A. Feinberg, "Letter to the Editor," *Oper. Res.* **44** p. 526, 1996.

[12] E.A. Feinberg, "Constrained discounted Markov decision processes and Hamiltonian cycles," *Math. of Operations Research,* **25** pp. 130–140, 2000.

[13] E.A. Feinberg, "Continuous Time Discounted Jump Markov Decision Processes: Discrete-Event Approach," State University of New York at Stony Brook, Preprint, 1998.

[14] E.A. Feinberg and A. Shwartz, "Markov decision models with weighted discounted criteria," *Math. of Operations Research* **19** pp. 152–168, 1994.

[15] E.A. Feinberg and A. Shwartz, "Constrained Markov decision models with weighted discounted rewards," *Math. of Operations Research* **20** pp. 302–320, 1995.

[16] E.A. Feinberg and A. Shwartz, "Constrained discounted dynamic programming," *Math. of Operations Research* **21** pp. 922–945, 1996.

[17] E.A. Feinberg and A. Shwartz, "Constrained dynamic programming with two discount factors: applications and an algorithm," *IEEE Transactions on Automatic Control* **TAC-44** pp. 628–630, 1999.

[18] E. Fernandez-Gaucherand, M.K. Ghosh and S.I. Marcus, "Controlled Markov processes on the infinite planning horizon: weighted and overtaking cost criteria," *ZOR—Methods and Models of Operations Research* **39** pp. 131–155, 1994.

[19] J. Filar and O. Vrieze, "Weighted reward criteria in competitive Markov decision programming problems," *ZOR—Methods and Models of Operations Research* **36** pp. 343–358, 1992.

[20] M.K. Ghosh and S.I. Marcus, "Infinite horizon controlled diffusion problems with some nonstandard criteria," J. Math. Systems, Estimation and Control **1** pp. 45–69, 1991.

[21] K. Golabi, Ram B. Kulkarni and G.B. Way, "A statewide pavement management system," *Interfaces* **12** pp. 5–21, 1982.

[22] A. Haurie and P. L'Ecuyer, "Approximation and bounds in discrete event dynamic programming," *IEEE Transactions on Automatic Control* **AC-31** pp. 227–235, 1986.

[23] O. Hernandez-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes,* Springer, New York, 1996.

[24] O. Hernandez-Lerma and J. Lasserre, *Future Topics on Discrete-Time Markov Control Processes,* Springer, New York, 1999.

[25] O. Hernandez-Lerma and R. Romera, *Pareto Optimality in Multiobjective Markov Control Processes,* Preprint, 2000.

[26] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Math. Centre Tracts **51**, Math. Centrum, Amsterdam, 1974.

[27] K. Hinderer, *Foundations of Non Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research **33**, Springer-Verlag, NY, 1970.

[28] L.C.M. Kallenberg, *Linear Programming and Finite Markovian Problem*, Math. Centre Tracts **148**, Math. Centrum, Amsterdam, 1983.

[29] D. Krass, J. Filar and S.S. Sinha, "A weighted Markov decision process," *Oper. Res.* **40** pp. 1180–1187, 1992.

[30] A.P. Maitra and T. Parthasarathy, "On stochastic games," *Journal of Optimization Theory and Applications* **5** pp. 289–300, 1970.

[31] A.P. Maitra and W.D. Sudderth, "An operator solution of stochastic games," *Israel Journal of Mathematics* **78** pp. 33–49, 1992.

[32] A.P. Maitra and W.D. Sudderth, "Borel stochastic games with limsup payoff," *Annals of Probability* **21** pp. 861–885, 1993

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

[34] A.S. Nowak, "Universally measurable strategies in zero-sum stochastic games," *Annals of Probability* **13** pp. 269–287, 1985.

[35] T. Parthasarathy and E.S. Raghavan, *Some Topics in Two-Person Games,* Elsevier, New York, 1967.

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

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

[38] M.I. Reiman and A. Shwartz, "Call Admission: A new Approach to Quality of Service," to appear, QUESTA.

[39] K.W. Ross, "Randomized and past dependent policies for Markov decision processes with finite action sets," *Oper. Res.* **37** pp. 474–477, 1989.

[40] K.W. Ross and R. Varadarajan, "Multichain Markov decision processes with a sample path constraint: a decomposition approach," *Math. Operations Research* **16** pp. 195–207, 1991.

[41] M. Schäl, "Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal," *Z. Wahr. verw. Gebiete* **32** pp. 179–196, 1975.

[42] A. Shwartz, "Death and Discounting," to appear, *IEEE Trans. on Auto. Control*, 2001.

[43] S. Stidham, "On the convergence of successive approximations in dynamic programming with non-zero terminal rewards," *Z. Operations Res.* **25** pp. 57–77, 1981.

[44] K.C.P. Want and J.P. Zaniewski, "20/30 hindsight: the new pavement optimization," *Interfaces* **26** pp. 77–87, 1996.

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

7 BLACKWELL OPTIMALITY

Arie Hordijk
Alexander A. Yushkevich

7.1 FINITE MODELS

In this introductory section we consider Blackwell optimality in Controlled Markov Processes (CMPs) with finite state and action spaces; for brevity, we call them finite models. We introduce the basic definitions, the Laurent-expansion technique, the lexicographical policy improvement, and the Blackwell optimality equation, which were developed at the early stage of the study of sensitive criteria in CMPs. We also mention some extensions and generalizations obtained afterwards for the case of a finite state space. In Chapter 1 the algorithmic approach to Blackwell optimality for finite models is given. We refer to that chapter for computational methods. Especially for the linear programming method, which we do not introduce.

7.1.1 Definition and existence of Blackwell optimal policies

We consider an infinite horizon CMP with a finite state space $\mathbb{X}$, a finite action space $\mathbb{A}$, action sets $\mathbb{A}(x) = A_x$, transition probabilities $p_{xy}(a) = p(y|x,a)$, and reward function $r(x,a)$ $(x \in X, a \in A_x, y \in X)$. Let m be the number of states in $\mathbb{X}$.

We refer to Chapter 0 for definitions of various policies, of probability distributions and expectations corresponding to them, and notations. We also use the notation

$$\mathbb{K} = \{(x,a) : \ a \in \mathbb{A}(x), \ x \in \mathbb{X}\}, \tag{1}$$

so that, in particular,

$$P^a f(x) = \sum_{y \in \mathbb{X}} p_{xy}(a) f(y), \qquad (x,a) \in \mathbb{K}. \tag{2}$$

For every discount factor $\beta \in (0,1)$ the expected total reward

$$v(x,\pi,\beta) = v_\beta(x,\pi) := \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \beta^t r(x_t, a_t) \right] \tag{3}$$

converges absolutely and uniformly in the initial state x and policy π, so that the value function

$$V(x,\beta) = V_\beta(x) := \sup_{\pi \in \Pi} v_\beta(x,\pi), \qquad x \in X$$

is well defined and finite. Following Blackwell [3], in this chapter we say that a policy π is β-*optimal* if $v_\beta(x,\pi) = V_\beta(x)$ for all $x \in X$ (not to confuse with ϵ-optimal policies, for which $v_\beta(\pi) \geq V_\beta - \epsilon$; in this chapter we do not use them).

In the case of a stationary policy $\varphi \in \Pi^s$ it is convenient to write (3) in matrix notations. In that case we have an $m \times m$ transition matrix $P(\varphi) = P^\varphi$ with entries $p_{xy}(\varphi(x)) = p_{xy}^\varphi$, and (3) can be written in the form

$$v_\beta(\varphi) = v_\beta^\varphi = \sum_{t=0}^{\infty} (\beta P^\varphi)^t r^\varphi = (I - \beta P^\varphi)^{-1} r^\varphi \tag{4}$$

where r^φ is a vector with entries $r(x,\varphi(x))$, $x \in \mathbb{X}$ (formula (4) makes sense also for complex β with $|\beta| < 1$), (in the notation (2) $P^\varphi f(x) = P^{\varphi(x)} f(x)$). For every $\beta \in (0,1)$ there exists a β-optimal policy $\varphi_\beta \in \Pi^s$; namely, one may set

$$\varphi_\beta(x) = \operatorname*{argmax}_{a \in A_x} \Big[r(x,a) + \beta \sum_{y \in X} P^a v_\beta(x) \Big], \qquad x \in \mathbb{X}.$$

In the important case of undiscounted rewards, when $\beta = 1$, the total expected reward in general diverges, and the simplest performance measure is the average expected reward $w(x,\pi) = w^\pi(x)$ (see Chapter 0). For a stationary policy φ

$$w^\varphi = \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} (P^\varphi)^t r^\varphi = Q^\varphi r^\varphi = \lim_{\beta \uparrow 1} (1-\beta) v_\beta^\varphi, \tag{5}$$

where $Q^\varphi = Q(\varphi)$ is the *stationary* (or *limiting*) *matrix*

$$Q^\varphi = \lim_{N \to \infty} \frac{1}{N+1} \sum_{t=0}^{N} (P^\varphi)^t, \quad \text{and} \quad Q^\varphi P^\varphi = P^\varphi Q^\varphi = Q^\varphi. \tag{6}$$

The last expression for w^φ in (5) follows from (3) and the fact that Cesaro summability of a divergent series implies its Abel summability to the same limit. Howard [27] proved the existence of average optimal policies in finite CMPs with the class Π^s of admissible policies, and developed a policy improvement algorithm to find them, related with his name. Almost at the same time Wagner [46] determined that such policies are average optimal in the class Π too.

However, the average reward criterion is insensitive, under selective since it is entirely determined by the arbitrarily far tail of the rewards; in accordance with

this criterion, two policies providing rewards $100+0+0+\cdots$ and $0+0+0+\cdots$ are equally good (or bad). Blackwell [3] in his study of finite CMPs introduced a much more sensitive concept of optimality, that bears now his name, and proved the existence of stationary policies optimal in this new sense.

Definition 7.1 *A policy π is said to be Blackwell optimal, if π is β-optimal for all values of β in an interval $\beta_0 < \beta < 1$.*

A stationary Blackwell optimal policy φ is average optimal. Indeed, there exists a stationary average optimal policy ψ, and by (5)

$$w^{\psi} = \lim_{\beta\uparrow 1}(1-\beta)v_{\beta}^{\psi} \leq \lim_{\beta\uparrow 1}(1-\beta)v_{\beta}^{\varphi} = w^{\varphi},$$

so that $w^{\varphi} = w^{\psi}$. Since the last limit is the same for all Blackwell optimal policies, stationary or not (as follows from Definition 7.1), and since by Theorem 7.1 below there is a stationary Blackwell optimal policy, *every Blackwell optimal policy $\pi \in \Pi$ is average optimal.*

Theorem 7.1 *In finite CMP there exists a stationary Blackwell optimal policy.*

Proof. Since for every positive $\beta < 1$ there exists a β-optimal policy $\varphi_{\beta} \in \Pi^s$, and because the set Π^s of stationary policies is finite together with $\mathbb{X}$ and $\mathbb{A}$, there exists a stationary policy φ which is β-optimal for all $\beta = \beta_n$ where $\beta_n \uparrow 1$. We claim that φ is Blackwell optimal.

Suppose the contrary. Then, because $\mathbb{X}$ and Π^s are finite sets, there are a state x_0, a policy $\psi \in \Pi^s$, and a sequence $\gamma_n \uparrow 1$ such that

$$v_{\beta}^{\varphi}(x_0) < v_{\beta}^{\psi}(x_0) \quad \text{for} \quad \beta = \gamma_n, \quad \gamma_n \uparrow 1.$$

On the other hand, by the selection of φ

$$v_{\beta}^{\varphi}(x_0) \geq v_{\beta}^{\psi}(x_0) \quad \text{for} \quad \beta = \beta_n \uparrow 1.$$

It follows that the function

$$f(\beta) = v_{\beta}^{\varphi}(x_0) - v_{\beta}^{\psi}(x_0)$$

defined for all complex β with $|\beta| < 1$ takes on the value 0 at an infinite sequence of different points $z_n \uparrow 1$, and takes on nonzero values at the points $\gamma_n \uparrow 1$.

By using Cramer's rule to compute the inverse matrix, we find that each entry of $(I - \beta P^{\varphi})^{-1}$ is a rational function of β, and the same is true with ψ in place of φ. Therefore and by (4), $f(\beta)$ is a rational function of the complex variable β in the circle $|\beta| < 1$ (and hence on the whole complex plane). A rational function cannot have infinitely many different zeros z_n if it is not an identical zero. The obtained contradiction proves that φ is Blackwell optimal. ∎

The above proof is a purely existence argument, without any indication how to find a Blackwell optimal policy φ. Blackwell's original proof also did

not provide a complete algorithm to obtain φ, but it contained some essential elements in this direction. Blackwell used, besides the limiting matrix Q^φ, the *deviation matrix* D^φ corresponding to $\varphi \in \Pi^s$. If the Markov chain with the transition matrix P^φ is aperiodic, then

$$D^\varphi = \sum_{t=0}^{\infty} [(P^\varphi)^t - Q^\varphi], \tag{7}$$

and the above series converges geometrically fast; in general

$$D^\varphi = \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} \sum_{t=0}^{n} [(P^\varphi)^t - Q^\varphi]. \tag{7'}$$

An important property of this matrix is that D^φ is uniquely determined by the equations

$$D^\varphi Q^\varphi = Q^\varphi D^\varphi = 0, \tag{8}$$

$$D^\varphi (I - P^\varphi) = (I - P^\varphi) D^\varphi = I - Q^\varphi \tag{9}$$

(see, for instance, Kemeny and Snell [29]). Blackwell derived and utilized the expansion

$$v_\beta^\varphi = \frac{h_{-1}^\varphi}{1-\beta} + h_0^\varphi + o(1) \qquad \text{as} \quad \beta \uparrow 1, \tag{10}$$

where

$$h_{-1}^\varphi = Q^\varphi r^\varphi, \qquad h_0^\varphi = D^\varphi r^\varphi, \tag{11}$$

and introduced the notion of a *nearly optimal* policy $\pi \in \Pi$. For such a policy $V_\beta - v_\beta^\pi = o(1)$ as $\beta \uparrow 1$.

The existence of a Blackwell optimal policy $\varphi \in \Pi^s$ implies a similar expansion for the value function

$$V_\beta(x) = \frac{h_{-1}}{1-\beta} + h_0 + o(1) \qquad \text{as} \quad \beta \uparrow 1. \tag{12}$$

It is easy to see using (12), that a policy π is average optimal iff $v_\beta^\pi = h_{-1}/\alpha + o(1/\alpha)$, where $\alpha = 1-\beta$, and that π is nearly optimal iff $v_\beta^\pi = h_{-1}/\alpha + h_0 + o(1)$.

7.1.2 Laurent series expansions and n-discount optimality

Average optimal and nearly optimal policies, as well as relations (10)-(12), are at the start of a chain of notions and equations developed by Miller and Veinott [33] and Veinott [42], which lead to a deeper insight into Blackwell optimal policies and to an algorithm to find them. We present their main ideas in a slightly modified form.

The approach is based on the Laurent series expansion of the resolvent

$$R_\beta = (I - \beta P)^{-1} = I + \beta P + \beta^2 P^2 + \cdots \qquad (|\beta| < 1) \tag{13}$$

of a Markov chain with the transition kernel P in the neighborhood of the point $\beta = 1$. This expansion is a general fact known in functional analysis (see, for

instance, [48]). In the particular case of an aperiodic Markov chain, it follows immediately from the geometric convergence of P^t to the limiting matrix Q. Indeed, the difference

$$R_\beta - \frac{1}{1-\beta}Q = (I-Q) + \beta(P-Q) + \beta^2(P^2-Q) + \cdots,$$

in which $\|P^n - Q\| \le C\gamma^n$ for some $\gamma < 1$, is an analytic function of the complex variable β in the circle $|\beta| < 1/\gamma$ (we use the norm in the space of $(m \times m)$-matrices generated by the supremum norm in the space of m-vectors). The point $\beta = 1$ is inside this circle, thus R_β has the same singularity at the point $\beta = 1$ as $\frac{1}{1-\beta}Q$, i.e. has a single pole. Therefore in some ring $0 < |\beta - 1| < \alpha_0$ a Laurent expansion

$$R_\beta = \frac{R_{-1}}{\alpha} + R_0 + R_1\alpha + R_2\alpha^2 + \cdots, \qquad \alpha = 1-\beta \tag{14}$$

holds. If the Markov chain is periodic, consider the least common multiple d of the periods of all its ergodic classes. The chain with a kernel P^d is then aperiodic, so that P^{nd} converges geometrically fast to a stochastic matrix $\widetilde{Q}$ as $n \to \infty$. Similar to the preceding argument, it follows that the infinite sum

$$\widetilde{R}_\beta = I + \beta^d P + \beta^{2d}P^2 + \cdots$$

is analytic in a circle $|\beta|^d < 1/\gamma$ of a radius greater than 1, and thus has a simple pole at $\beta = 1$. Then the same is true for

$$R_\beta = (I + \beta P + \cdots + \beta^{d-1}P^{d-1})\widetilde{R}_\beta.$$

Instead of the Laurent series (14), one may write a similar series in powers of another small parameter ρ equivalent to α, which has the meaning of an interest rate:

$$\rho = \frac{1-\beta}{\beta} = \frac{\alpha}{1-\alpha}, \qquad \beta = \frac{1}{1+\rho} = 1-\alpha. \tag{15}$$

Veinott [42] and most of the subsequent authors used series in ρ. Chitashvili [6, 7, 56] and following him Yushkevich [49]–[55] used series in α. We present both versions.

Theorem 7.2 *In a finite CMP there exists a number $\beta_0 \in (0,1)$ such that for every policy $\varphi \in \Pi^s$*

$$v_\beta^\varphi = (1+\rho)\sum_{n=-1}^{\infty} h_n^\varphi \rho^n = \sum_{n=-1}^{\infty} k_n^\varphi \alpha^n, \qquad \beta_0 < \beta < 1 \tag{16}$$

where

$$h_{-1}^\varphi = k_{-1}^\varphi = Q^\varphi r^\varphi = w^\varphi, \qquad h_0^\varphi = k_0^\varphi = D^\varphi r^\varphi \tag{17}$$

(cf. (10) and (11)), and where for $n \ge 1$

$$h_n^\varphi = (-D^\varphi)^n h_0^\varphi, \qquad k_n^\varphi = (I - D^\varphi)^n k_0^\varphi. \tag{18}$$

A similar expansion is valid for the value function

$$V_\beta = (1+\rho)\sum_{n=-1}^{\infty} h_n \rho^n = \sum_{n=-1}^{\infty} k_n \alpha^n, \qquad \beta_0 < \beta < 1. \tag{19}$$

Proof. The existence and convergence of Laurent expansions (16) follow from expansions in powers of ρ or α of βR_β^φ, respectively R_β^φ, and from the formula $v_\beta^\varphi = R_\beta^\varphi r^\varphi$ equivalent to (4). To get the coefficients (17)-(18), observe that by (4) $v_\beta^\varphi = r^\varphi + \beta P^\varphi v_\beta^\varphi$, so that by (15) and (16)

$$(1+\rho)\sum_{-1}^{\infty} h_n^\varphi \rho^n = r^\varphi + P^\varphi \sum_{-1}^{\infty} h_n^\varphi \rho^n.$$

By the uniqueness of the coefficients of power series, this results in equations (to simplify writing, we temporarily skip the superscript φ):

$$h_{-1} = Ph_{-1}, \tag{20}$$

$$h_0 + h_{-1} = r + Ph_0, \tag{21}$$

$$h_n + h_{n-1} = Ph_n \qquad (n \geq 1). \tag{22}$$

From (6) and (20) by iteration and taking a limit, we find $h_{-1} = Qh_{-1}$. For the stationary matrix $Q = QP = PQ$, and a multiplication of (21) by Q gives $Qh_{-1} = Qr$, so that $h_{-1} = Qr$ as in (17). A multiplication of (22) by Q provides $Qh_n = 0$ $(n \geq 0)$. Using this, the relation $h_{-1} = Qh_{-1}$ and (8)–(9), we get after a multiplication of (21) by $D = D^\varphi$, that $D(I-P)h_0 + DQh_{-1} = Dr$, or $(I-Q)h_0 = Dr$, or finally $h_0 = Dr$ as in (17). Multiplying (22) by D, in a similar way we get $D(I-P)h_n + Dh_{n-1} = 0$, or $h_n - Qh_n = -Dh_{n-1}$, or $h_n = -Dh_{n-1}$ $(n \geq 1)$, and this proves that $h_n = (-D)^n h_0$ as in (18). Formulas (17)–(18) for k_n^φ follow absolutely similarly from equations $k_{-1} = Pk_{-1}$, $k_0 + Pk_{-1} = r + Pk_0$ and $k_n + Pk_{n-1} = Pk_n$ instead of (20)–(22).

Since the set Π^s is finite, we have the expansions (16) simultaneously for all $\varphi \in \Pi^s$ in some interval $(\beta_0, 1)$. Formula (19) follows now from Theorem 7.1. ∎

Formulas of Theorem 7.2 are a generalization of (10) and (11). They stimulate a similar generalization of the average optimality and nearly optimality criteria. The following definition is due to Veinott [43].

Definition 7.2 *For $n \geq 1$, a policy $\pi^* \in \Pi$ is said to be n-discount optimal, if for every $\pi \in \Pi$*

$$\underline{\lim}_{\beta \uparrow 1} \rho^{-n}[v_\beta(\pi^*) - v_\beta(\pi)] \geq 0 \tag{23}$$

(with α in place of ρ we have an equivalent condition).

By substituting in (23) a Blackwell optimal policy π, for which $v_\beta(\pi) = V_\beta$ and $v_\beta(\pi^*) - v_\beta(\pi) \leq 0$, one may see that in finite CMPs condition (23) is equivalent to a simpler (and formally stronger) condition

$$\lim_{\beta \uparrow 1} \rho^{-n}[V_\beta - v_\beta(\pi^*)] = 0. \tag{24}$$

However, condition (23) appeared to be more suitable for an extension of sensitive criteria to denumerable and Borelian CMPs. To avoid confusion, mention that in literature 0-discount optimal policies are sometimes called *bias-optimal*

or 1-*optimal*; the latter name originates from Veinott [42]. Also, as seen from a comparison of (16) and (19), a stationary policy is Blackwell optimal iff it is n-discount optimal for every natural n, or, briefly speaking, is ∞-discount optimal. For bias optimality in models with finite state and action spaces see Chapter 2.

A convenient description of n-discount optimal policies can be made in terms of sequences of coefficients of series (16) and (19) and a lexicographical ordering in spaces of them. Define

$$H^\varphi = \{h^\varphi_{-1}, h^\varphi_0, \dots\}, \quad K^\varphi = \{k^\varphi_{-1}, k^\varphi_0, \dots\}, \tag{25}$$

let H^φ_n and K^φ_n be the initial segments of H^φ and K^φ up to the n-th term, and let H, K, H_n and K_n have the same meaning for the series (19) (each h^φ_n etc. is an m-vector). For those sequences and segments we introduce a natural lexicographical ordering denoted by symbols $\succ$, $\succeq$, $\preceq$, $\prec$. So, $H^\varphi \prec H^\psi$ means that $H^\varphi \neq H^\psi$, and that there exists a number $N < \infty$ and a state $x_0 \in \mathbb{X}$, such that $H^\varphi_{N-1} = H^\psi_{N-1}$ (if $N \geq 0$), and $h^\varphi_N(x_0) < h^\psi_N(x_0)$ while $h^\varphi_N(x) \leq h^\psi_N(x)$ for all other $x \in \mathbb{X}$. The relation $H^\varphi \preceq H^\psi$ means that either $H^\varphi = H^\psi$ or $H^\varphi \prec H^\psi$. The relations $H^\psi \succ H^\varphi$ and $H^\psi \succeq H^\varphi$ are equivalent to $H^\varphi \prec H^\psi$ and $H^\varphi \preceq H^\psi$.

With this notation we have $H^\varphi \preceq H$ and $K^\varphi \preceq K$ for every $\varphi \in \Pi^s$, and the policy φ is n-discount optimal (or Blackwell optimal) iff $H^\varphi_n = H_n$ or $K^\varphi_n = K_n$ (respectively, if $H^\varphi = H$ or $K^\varphi = K$).

The following theorem due to Veinott [43] shows that in finite CMPs the n-th discount optimality of a stationary policy for large values of n coincides with its Blackwell optimality. Let Φ_n be the subset of Π^s consisting of all stationary n-discount optimal policies ($n \geq -1$), and let Φ_∞ be the set of all Blackwell optimal policies in Π^s. Evidently,

$$\Phi_{-1} \supset \Phi_0 \supset \Phi_1 \supset \cdots, \qquad \Phi_\infty = \bigcap_n \Phi_n.$$

Theorem 7.3 *In finite CMPs with $m \geq 2$ states*

$$\Phi_{m-1} = \Phi_m = \ldots = \Phi_\infty.$$

Proof. It is sufficient to show that $\Phi_{m-1} = \Phi_\infty$. Consider any policy $\varphi \in \Phi_{m-1}$. We have $H^\varphi_{m-1} = H_{m-1}$, or in more detail

$$h^\varphi_n = h_n, \qquad n = -1, 0, 1, \dots, m-1. \tag{26}$$

Since $m \geq 2$, both h_0 and h_1 are present in (26). We claim that m column m-vectors $h_0, h_1, \dots, h_{m-1}$ are linearly dependent. It is sufficient to show that m row vectors of the corresponding square matrix are linearly dependent; these rows are $\{h_0(x), \dots, h_{m-1}(x)\} = \{h^\varphi_0(x), \dots, h^\varphi_{m-1}(x)\}$, $x \in \mathbb{X}$. In fact even the infinite sequences

$$\{h^\varphi_0(x), h^\varphi_1(x), \dots, h^\varphi_t(x), \dots\}, \qquad x \in \mathbb{X} \tag{27}$$

are linearly dependent. Indeed, in the finite Markov chain generated by P^φ there exists a stationary distribution $\{\mu(x),\ x \in \mathbb{X}\}$. The total discounted

expected reward corresponding to the initial distribution μ and policy φ is equal to

$$v_\beta^\varphi(\mu) := \sum_{x\in\mathbb{X}} \mu(x) v_\beta^\varphi(x) = \sum_{x\in\mathbb{X}} \mu(x)\, \mathbb{E}_x^\varphi \sum_{t=0}^{\infty} \beta^t r(x_t, \varphi(x_t)) =$$

$$= \sum_{t=0}^{\infty} \beta^t \sum_{x\in\mathbb{X}} \mu(x)\, \mathbb{E}_x^\varphi\, r(x_t, \varphi(x_t)) = \sum_{t=0}^{\infty} \beta^t\, \mathbb{E}_\mu^\varphi\, r(x_t, \varphi(x_t)). \quad (28)$$

Here the $\mathbb{P}_\mu^\varphi$-distribution of x_t does not depend on t because μ is a stationary distribution, and hence the factor at β^t in (28) is some constant C. Thus

$$v_\beta^\varphi(\mu) = C \sum_{t=0}^{\infty} \beta^t = \frac{C}{1-\beta} = C\frac{1+\rho}{\rho} = (1+\rho)\left[\frac{C}{\rho} + \sum_{n=0}^{\infty} 0\cdot\rho^n\right] \quad (29)$$

(cf. (15)). On the other hand, by (28) and (16),

$$v_\beta^\varphi(\mu) = (1+\rho) \sum_{n=-1}^{\infty} \rho^n \sum_{x\in\mathbb{X}} \mu(x) h_n^\varphi(x).$$

A comparison with (29) together with the uniqueness of the Laurent coefficients show that $\sum_x \mu(x) h_n^\varphi(x) = 0$ for all $n \geq 0$, so that the sequences (27) are linearly dependent.

Now, by (18)

$$h_{n+1}^\varphi = -D^\varphi h_n^\varphi, \qquad h_{n+1} = -D^\psi h_n \qquad (n = 0, 1, 2, \dots) \quad (30)$$

where ψ is a Blackwell optimal policy. Let t be the maximal integer such that the vectors $h_0 = h_0^\varphi, \dots, h_t = h_t^\varphi$ in (26) are linearly independent; such $t \geq 0$ exists if only $h_0 \neq 0$, and as just proved, $t < m-1$. If $h_0 = 0$, then by (26) also $h_0^\varphi = 0$, and by (30) $h_n^\varphi = 0 = h_n$ for all $h \geq 0$, so that $H^\varphi = H$ and $\varphi \in \Phi_\infty$. If there is the required t, then $h_{t+1} = h_{t+1}^\varphi$ is a linear combination of $h_0 = h_0^\varphi, \dots, h_t = h_t^\varphi$:

$$h_{t+1} = \sum_{i=0}^{t} C_i h_i, \qquad h_{t+1}^\varphi = \sum_{i=0}^{t} C_i h_i^\varphi. \quad (31)$$

Due to (30) multiplying the first identity by $-D^\psi$ and the second by $-D^\varphi$, we only increase every subscript in (31) by 1, and since $h_i^\varphi = h_i$ for $0 \leq i \leq t+1$, we get $h_{t+2}^\varphi = h_{t+2}$. Repeating this, by induction we get $h_n^\varphi = h_n$ for every $n \geq 0$, so that $\varphi \in \Phi_\infty$. ∎

The sets Φ_{m-2} and Φ_{m-1} are in general different. The following example, taken from [43], confirms this statement. To make it more visual, we present it for $m = 5$.

Example 7.1 *There are $m = 5$ states $1, 2, \dots, 5$ with mandatory transitions $2 \to 3 \to 4 \to 5$, the state 5 is absorbing. In state 1 there is a choice between*

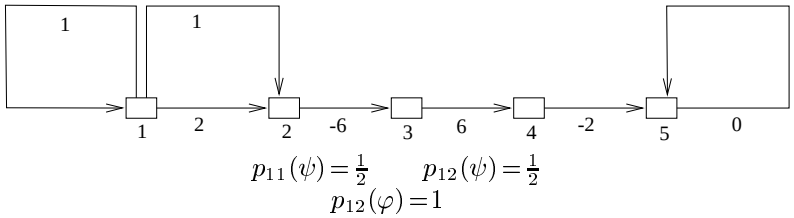

Figure 7.1

two actions, which determine two stationary policies φ and ψ (see Figure 7.1). Under φ we have a mandatory transition $1 \to 2$, under ψ transitions $1 \to 1$ and $1 \to 2$ are equally likely. The numbers under arrows indicate the rewards $r(x,a)$. Mention that the rewards $2, -6, 6, -2$ are the binomial coefficients of $(A-B)^{m-2} = (A-B)^3$ multiplied by 2.

The expected rewards v_β^φ and v_β^ψ differ only at the initial state 1. For φ we have

$$v_\beta^\varphi(1) = 2 - 6\beta + 6\beta^2 - 2\beta^3 = 2(1-\beta)^3 = 2\alpha^3.$$

For ψ, by the formula $v_\beta^\psi = r^\psi + \beta P^\psi v_\beta^\psi$ (cf. (4)) we have the equation

$$v_\beta^\psi = 1 + \beta[\frac{1}{2}v_\beta^\psi(1) + \frac{1}{2}v_\beta^\psi(2)].$$

Thus

$$(2-\beta)v_\beta^\psi(1) = 2 + \beta v_\beta^\psi(2),$$

where

$$\beta v_\beta^\psi(2) = \beta(-6 + 6\beta - 2\beta^2) = v_\beta^\varphi(1) - 2.$$

Hence

$$v_\beta^\psi(1) = \frac{v_\beta^\varphi(1)}{2-\beta} = \frac{2\alpha^3}{1+\alpha} = 2\alpha^3 - 2\alpha^4 + 2\alpha^5 - \cdots.$$

This means that $V_\beta(1) = 2\alpha^3$, that φ is Blackwell optimal, and that ψ is 3-discount optimal, but not 4-discount optimal. Thus $\Phi_3 \neq \Phi_4 = \Phi_\infty$.

7.1.3 Lexicographical policy improvement and Blackwell optimality equation

Policy improvement is both a practical method to approach an optimal policy in CMPs and an important tool in their theory. Its essence is that if φ and ψ are two stationary policies, if $\pi = \psi\varphi^\infty$ is a Markov policy coinciding with ψ at the first step of the control and coinciding with φ afterwards, and if π is better than φ, then ψ is also better than φ. This method, almost trivial for the discounted reward criterion with a fixed $\beta < 1$, was developed by Howard [27] for the average reward criterion. Howard used, besides the average reward $w^\varphi (= h_{-1}^\varphi)$, a second function, in fact equal to the term h_0^φ in the expansions (10) and (16) up to a constant term on each recurrence class of the Markov chain generated by φ. Blackwell [3] provided a rigorous proof that a slightly

different version of Howard's policy improvement method does converge. Miller and Veinott [33] have extended policy improvement to the case of Blackwell optimality, and Veinott [43] refined it using the classes Φ_n. We expose this topic in a modernized form, using an operator approach developed in Dekker and Hordijk [8] in the framework of CMPs with a countable state space $\mathbb{X}$. To avoid additional formulas, we do all calculations in terms of ρ; in terms of α formulas are slightly different.

From the structure of π and (13) we have

$$v_\beta^\pi = r^\psi + \beta P^\psi v_\beta^\varphi = r^\psi + \frac{1}{1+\rho} P^\psi v_\beta^\varphi = r^\psi + P^\psi \sum_{n=-1}^{\infty} h_n^\varphi \rho^n,$$

while

$$v_\beta^\varphi = h_{-1}^\varphi \rho^{-1} + \sum_{n=0}^{\infty} (h_n^\varphi + h_{n-1}^\varphi)\rho^n.$$

Subtracting, we get

$$v_\beta^\pi - v_\beta^\varphi = (P^\psi h_{-1} - h_{-1})\rho^{-1} + (r^\psi + P^\psi h_0 - h_0 - h_{-1}) + \sum_{n=1}^{\infty} (P^\psi h_n - h_n - h_{n-1})\rho^n \tag{32}$$

where it is understood that $h_n = h_n^\varphi$. By (18), the supremum norm $\|h_n^\varphi\|$ is growing no more than geometrically fast with n.

It is convenient to introduce the space $\mathfrak{H}$ of all sequences $H = \{h_n, n \geq -1\}$ of m-vectors satisfying this growth condition, and to treat the sequences of Laurent coefficients of the series (16), (32) etc. as elements of $\mathfrak{H}$. In particular $H^\varphi \in \mathfrak{H}$ (see (25)), and in $\mathfrak{H}$ we consider the same lexicographical ordering as we have introduced in connection with H^φ. Also, it is convenient to define the spaces $\mathfrak{H}_n$ of finite collections $H_n = \{h_t, -1 \leq t \leq n\}$ of m-vectors.

The right side of (32) defines an *operator* L^ψ in the spaces $\mathfrak{H}$ and $\mathfrak{H}_n$. Since the matrix P^ψ has entries $p_{xy}(a)$ with $a = \psi(x)$, we express L^ψ through the corresponding operators L^a transforming functions (vectors) on $\mathbb{X}$ into functions of pairs (x, a) on the state-action space $\mathbb{K}$ defined in (1). We have

$$(L^\psi H)(x) = L^{\psi(x)} H(x), \qquad x \in \mathbb{X}, \tag{33}$$

$$L^a H(x) = \{\ell h_{-1}^a(x), \ell h_0^a(x), \ell h_1^a(x), \dots\}, \qquad (x, a) \in \mathbb{K}, \tag{34}$$

where according to (32)

$$\begin{aligned} \ell h_{-1}^a(x) &= P^a h_{-1}(x) - h_{-1}(x), \\ \ell h_0^a(x) &= r(x,a) + P^a h_0(x) - h_0(x) - h_{-1}(x), \\ \ell h_n^a(x) &= P^a h_n(x) - h_n(x) - h_{n-1}(x) \qquad (n \geq 1). \end{aligned} \tag{35}$$

The same formulas define L^a and L^ψ, as operators on $\mathfrak{H}_n$.

Lemma 7.1 *Let $\varphi, \psi \in \Pi^s$. If $(L^\psi H^\varphi)_{n+1} \succeq 0$ for some $n \geq -1$, then $H_n^\psi \succeq H_n^\varphi$. Moreover, if in addition $(L^\psi H^\varphi)_{n+1}(x_0) \succ 0$ at some $x_0 \in \mathbb{X}$, then $H_n^\psi(x_0) \succ H_n^\varphi(x_0)$. The same is true with the reverse inequality signs.*

In particular, if $L^\psi H^\varphi = 0$, then $H^\psi = H^\varphi$.

Proof. The condition $(L^\psi H^\varphi)_n \succeq 0$ means that

$$v_\beta^\pi = v_\beta^\varphi + Q_n(\rho) + O(\rho^{n+1}) \tag{36}$$

where $Q_n(\rho)$ is a vector consisting of polynomials of degree $\le n$ with lexicographically nonnegative coefficients, and where $O(\rho^{n+1})$ is uniform in $x \in \mathbb{X}$ since $\mathbb{X}$ and $\mathbb{A}$ are finite sets (compare (32) with (33)-(35)). Consider policies $\pi_t = \psi^t \varphi^\infty$, and let $v(t) = v_\beta(\pi_t)$, so that, in particular, $v(0) = v_\beta^\varphi$, $v(1) = v_\beta^\pi$. We have

$$v(t+1) = r^\psi + \beta P^\psi v(t), \qquad t = 0, 1, 2, \dots \tag{37}$$

and by (36)

$$v(1) = v(0) + Q_n + R, \tag{38}$$

where the remainder R is of order ρ^{n+1}. From (37) and (38) by induction we get

$$v(t) = v(0) + (I + \beta P^\psi + (\beta^2 P^\psi)^2 + \cdots + (\beta P^\psi)^{t-1})(Q_n + R), \qquad t \ge 1.$$

(we use that $r^\psi + \beta P^\psi v(0) = v(0) + Q_n + R$ according to (37) and (38)). Since $\beta < 1$, in the limit $v(t)$ becomes $v(\infty) = v_\beta^\psi$, so that

$$v_\beta^\psi = v_\beta^\varphi + \sum_{t=0}^{\infty} (\beta P^\psi)^t (Q_n(\rho) + R) = v_\beta^\varphi + R_\beta^\psi (Q_n + R).$$

Here $Q_n \succeq 0$ for small $\rho > 0$, R is of order $O(\rho^{n+1})$, and the resolvent R_β^ψ is of order $O(1 + \beta + \beta^2 + \cdots) = O(\rho^{-1})$. This proves that $H_{n-1}^\psi \succeq H_{n-1}^\varphi$ if $L^\psi H_n^\varphi \succeq 0$. Other assertions are proved in a similar way. ∎

To proceed further, we need the lexicographical *Bellman operator* L in the spaces $\mathfrak{H}$ and $\mathfrak{H}_n$:

$$LH(x) = \max_{a \in \mathbb{A}_x} L^a H(x), \qquad H \in \mathfrak{H}, \quad x \in \mathbb{X}, \tag{39}$$

where the maximum is understood in the lexicographical sense $\succeq$; the same formula holds for $H_n \in \mathfrak{H}_n$. This maximum always exists because the sets $\mathbb{A}_x$ are finite. Since one may use all combinations of actions in stationary policies, formula

$$LH = \max_{\psi \in \Pi^s} L^\psi H \tag{40}$$

defines the same operator L.

If in (32) $\psi = \varphi$ then $\pi = \psi\varphi^\infty$ coincides with φ, and the left side of (32) is zero. Hence all the coefficients at the right side vanish, and this means that $L^\varphi H^\varphi = 0$ for every $\varphi \in \Pi^s$. Therefore $LH^\varphi \ge 0$ for every $\varphi \in \Pi^s$. If $LH^\varphi = 0$, we say that φ is *unimprovable*; if $LH_n^\varphi = 0$, then φ is *unimprovable of order* n. The equation

$$LH = 0 \qquad H \in \mathfrak{H} \tag{41}$$

is called the *Blackwell optimality equation* in honor of Blackwell; the similar equation $LH_n = 0$ for $H_n \in \mathfrak{H}_n$ is the *n-order optimality equation*. Let $H =$

$\{h_n\}$ be the element of $\mathfrak{H}$ corresponding to the value function V_β (see (19)), and let H_n be the initial segments of H. We say that a stationary policy φ is *conserving* (or n-order *conserving*) if $L^\varphi H = 0$ (respectively, $L^\varphi H_n = 0$).

Theorem 7.4 *A. The Blackwell optimality equation has a unique solution* $H^* = \max_{\varphi\in\Pi^s} H^\varphi$. *A policy* $\varphi \in \Pi^s$ *is Blackwell optimal iff* $H^\varphi = H^*$, *and iff* φ *is a conserving policy.*
B. For every $n \geq -1$, H_n^* *is uniquely determined by the equation* $LH_{n+1} = 0$. *A policy* $\varphi \in \Pi^s$ *is* n*-discount optimal iff* $H_n^\varphi = H_n^*$, *and is* n*-discount optimal if* φ *is* $(n+1)$*-order conserving.*

Proof. By Theorem 7.1 there exists a Blackwell optimal policy $\varphi \in \Pi^s$. Evidently, $H^\varphi \succeq H^\psi$, $\psi \in \Pi^s$ and φ is unimprovable, so that $H^\varphi = H^* := \max_{\psi\in\Pi^s} H^\psi$, and $LH^* = LH^\varphi = 0$. Since $L^\varphi H^\varphi = 0$, also $L^\varphi H^* = 0$, and φ is conserving. In part A it remains to prove that the solution of (41) is unique, and that a conserving stationary policy is Blackwell optimal. If ψ is conserving, then $L^\psi H^* = 0$, hence $L^\psi H^\varphi = 0$ for a Blackwell optimal φ, therefore by Lemma 7.1 (applied to every n) $H^\psi = H^\varphi = H^*$, so that ψ is Blackwell optimal too. Finally, suppose that $\widetilde{H}$ is a solution to (41). By taking for each $x \in \mathbb{X}$ a lexicographical maximizer $a \in \mathbb{A}_x$ of $L^a\widetilde{H}(x)$, we obtain a stationary policy ψ for which $L^\psi\widetilde{H} = \widetilde{H}$. One may check (we omit the proof) that Lemma 7.1 is true for any $H \in \mathfrak{H}$ in place of H^φ, in particular, for $\widetilde{H}$. It follows that $H^\psi = \widetilde{H}$, and since $LH^\psi = L\widetilde{H} = 0$, the policy ψ is unimprovable. Hence ψ is Blackwell optimal, so that $\widetilde{H} = H^\psi = H^*$.

The proof of part B is similar, with a reference to Lemma 7.1. ∎

Policy improvement is a basis for an algorithm to compute a Blackwell optimal policy in a finite CMP. Theoretically, one may proceed in the following way. Start with some $\varphi \in \Pi^s$ and compute H_m^φ using formulas of Theorem 7.2 (here m is the number of states in $\mathbb{X}$). Check the values of $\ell_{-1}^a h(x), (x,a) \in \mathbb{K}$. For $a = \varphi(x)$ those values are zeros, and if $\ell^{a^*} h_{-1}(x^*) > 0$ for some pair (x^*, a^*), then the policy

$$\psi(x) = \begin{cases} a^* & \text{if } x = x^*, \\ \varphi(x) & \text{otherwise} \end{cases}$$

improves φ. If there are no such pairs (x^*, a^*), repeat the same procedure with $\ell^a h_0$ and the shrinked sets $\mathbb{A}_0(x) = \{a \in \mathbb{A}_{-1}(x),\ \ell^a h_{-1}(x) = 0\}$, $\mathbb{K}_0 = \{(x,a) :\ a \in \mathbb{A}_0(x), x \in \mathbb{X}\}$ (where $\mathbb{A}_{-1}(x) = \mathbb{A}(x)$). A policy ψ as above with $\ell^{a^*} h_0(x^*) > 0$, $(x^*, a^*) \in \mathbb{K}$ improves φ. If there are no such pairs (x^*, a^*), repeat the procedure with all subscripts increased by 1, etc., until either you get a better policy ψ, or reach the set $\mathbb{K}_m$. In the latter case φ is $(m-1)$-order discount optimal, and therefore Blackwell optimal by Theorem 7.3. Otherwise, proceed in the same way with the obtained policy ψ. Since the set Π^s is finite, this algorithm leads to a Blackwell optimal policy in a finite number of steps. In practice, one may improve φ simultaneously at several states x^*; see Policy Iteration in chapter 10 of [34].

On the other hand, the lexicographical policy improvement approach opens a new way to prove the existence of Blackwell optimal policies via a maximiza-

tion of H^φ over all stationary policies φ and the related Blackwell optimality equation(41). The latter idea can be used in CMPs with an infinite state space $\mathbb{X}$, in which the proof of Theorem 7.1, based on the fact that the set Π^s is finite, is inapplicable.

7.1.4 Extensions and generalizations

In [45] Veinott simplified and updated results of [42, 43]. In particular, he refined Theorem 7.3 as follows: $\Phi_\infty = \Phi_{m-r}$ where r is the number of recurrent classes in X under a Blackwell optimal stationary policy.

Veinott [43] introduced also the notion of n-average optimality in addition to the n-discount optimality. Let

$$v_T^{(1)}(x,\pi) = \mathbb{E}_x^\pi \left[\sum_{t=0}^{T-1} r(x_t, a_t) \right]$$

and define recursively for $n \geq 1$

$$v_T^{(n+1)}(x,\pi) = \sum_{t=1}^{T} v_t^{(n)}(x,\pi).$$

Then π^* is n-average optimal if for every policy π

$$\varliminf_{T\to\infty} \frac{1}{T} \left[v_T^{(n+2)}(\pi^*) - v_T^{(n+2)}(\pi) \right] \geq 0.$$

Veinott [43, 44] and Sladky [38] showed that in a finite CMP a policy is n-discount optimal iff it is n-average optimal.

Chitashvili [6, 7, 56] extended results of Theorem 7.4 to more general models with a finite state space. In [6] he treated CMPs with arbitrary (indeed, compactified) action sets. He considered also what can be called (n,ϵ)-discount optimal policies; in their definition one should replace 0 by $-\epsilon$ in formula (23). In [7] he studied n-discount optimality in finite models with discount factors depending on the state x and action a: $\beta(x,a) = c_1\beta + c_2\beta^2 + \cdots + c_k\beta^k$ where k and c_i are functions of (x,a). In this case the reward functions were of some specific average form. In [56] Theorem 7.4 is generalized to a finite model with two reward functions $r(x,a)$ and $c(x,a)$. More precisely,

$$v_\beta^\varphi(x) = \mathbb{E}_x^\varphi \left[\sum_{t=0}^{\infty} \beta^t (r(x_t,a_t) + (1-\beta)c(x_t,a_t)) \right]$$

(in [56] Chitashvili considered only stationary policies). This expected discounted reward corresponds to an undiscounted reward

$$\sum_{t=0}^{\infty} r(x_t,a_t) + \varliminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c(x_t,a_t).$$

In that case all formulas related to Theorem 7.4 remain valid, with one exception: in equations (35) defining L^a, the term ℓh_1 should be changed to

$$\ell h_1^a(x) = c(x,a) + P^a h_1(x) - h_1(x) - h_0(x)$$

(similar to the term ℓh_0).

As explained in the proof of Theorem 7.1, in finite CMPs the expected discounted reward $v_\beta^\varphi(x)$ is a rational function of β. Hordijk e.a. [19] introduced a non-Archimedian ordered field of rational functions, used a simplex method in this field, and developed a linear programming method for the computation of β-optimal policies over the entire range (0,1) of the discount factor. In particular, their method allows to compute a Blackwell optimal policy. More precisely, for some m one may find numbers $\beta_0 := 0 < \beta_1 < \cdots < \beta_{m-1} < \beta_m := 1$ and stationary policies $\varphi_1, \varphi_2, \ldots, \varphi_m$ such that φ_j is β-optimal for all $\beta \in [\beta_{j-1}, \beta_j]$, $1 \leq j \leq m-1$, and φ_m is β-optimal in the interval $[\beta_{m-1}, \beta_m)$ (which means that φ_m is Blackwell optimal). In section 5.4 of Chapter 1 this algorithm is studied. In Chapter 3 an asymptotic simplex method based on Laurent series expansions for the computation of a Blackwell optimal policy, is used.

In CMPs with constraints the controller wants to maximize expected (discounted) rewards while keeping other expected (discounted) costs in some given bounds. For such CMPs Altman e.a. [1] gave a constructive proof for the following (weaker) version of the result obtained in [19]. There exist numbers m and β_j as above such that for every $j = 1, \ldots, m$ either the constrained problem is not feasible in the open interval $(\beta_{i-1}\beta_j)$ or the value function is a rational function of β in the closed interval $[\beta_{j-1}, \beta_j]$, $j \leq m-1$ and $[\beta_{m-1}, 1)$. Consequently, if the constrained problem is feasible in the neighborhood of $\beta = 1$, then v_β has a Laurent series expansion at $\beta = 1$.

As shown in the proof of Theorem 7.1, the limits of β-optimal policies, for β tending to 1, are Blackwell optimal. A counterexample in Hordijk and Spieksma [22] shows that in general this is not true in unichain CMPs with a finite state space and compact action sets. This disproves a conjecture in Cavazos-Cadena and Lasserre [4].

More recently, Huang and Veinott [28] extended many results concerning Blackwell and n-discount optimality in finite models to the case when (i) the reward $r(x, a, \rho)$ at the state x under the action a depends also on the interest rate ρ, namely is an analytic function of ρ in a neighborhood of $\rho = 0$, (ii) the nonnegative transition coefficients $p_{xy}(a)$ in general do not sum to 1, only for every Markov policy the t-step transition matrices are of order $O(1)$ as $t \to \infty$. They proved the existence of stationary Blackwell optimal policies, extended to their case the lexicographical policy improvement, showed that if $r(x, a, \rho)$ is a polynomial or a rational function of ρ then Blackwell optimality of a stationary policy is equivalent to its n-discount optimality for some n depending on the degrees of involved polynomials. In the latter case the policy iteration algorithm provides a Blackwell optimal policy in a finite number of steps. Another constructive rule is given for finding n-discount optimal policies using a linear programming approach.

7.2 DENUMERABLE STATE MODELS

In this section we consider CMPs for which the state space $\mathbb{X}$ is denumerable. There are many applications of controlled Markov chains for which it is natural to take an infinite number of states. An important class of models is that of

open stochastic networks, used for the modeling of controlled communication systems (see Spieksma [39]).

Even in the case when the action space $\mathbb{A} = \{0, 1\}$ consists of only two elements, but the state space $\mathbb{X}$ is denumerable, the situation with Blackwell optimal policies is much more complicated than in finite models. It turns out that a Blackwell optimal policy may be not average optimal. A corresponding counterexample, based on the fact that the Cesaro lower limit of a sequence of numbers can be different from the Abel lower limit of the same sequence, was constructed by Flynn [15]. Also, there can be no Blackwell optimal policy not only in the sense of Definition 7.1, but also in the sense of a weaker Definition 7.3 given below, in which the interval $(\beta_0, 1)$ may depend on the state x and the nonoptimal policy π. Maitra [31] presented such a counterexample, and in connection with it formulated this weaker definition.

The analysis of sensitive and Blackwell optimality for denumerable state models is mostly done under the following assumption.

Assumption 7.1 *(a) Action sets $\mathbb{A}(x)$, $x \in \mathbb{X}$, are compact metric sets. (b) Transition probabilities $p_{xy}(a)$ and rewards $r(x, a)$ are continuous functions of $a \in A(x)$ for all $x, y \in \mathbb{X}$.*

7.2.1 n-discount optimality

The following Lyapunov function condition introduced in Hordijk [18] implies, together with Assumption 7.1, the existence of a stationary n-discount optimal policy.

Assumption 7.2 *There exist a state, say state 0, and nonnegative functions $g_0, g_1, \dots,$ g_{n+1} on $\mathbb{X}$ such that*

(a) $\max_{a \in \mathbb{A}_x} |r(x, a)| \le g_0(x), \qquad x \in \mathbb{X},$

(b) $\inf_{x \in \mathbb{X}} g_0(x) > 0,$
$g_m(x) + \sum_{y \ne 0} p_{xy}(a) g_{m+1}(y) \le g_{m+1}(x)$ *for all* $a \in \mathbb{A}$, $x \in \mathbb{X}$,
$m = 0, 1, \dots, n,$

(c) $P^a g_{n+1}(x) = \sum_{y \in \mathbb{X}} p_{xy}(a) g_{n+1}(y)$ *is a continuous function of* $a \in \mathbb{A}_x$,
$x \in \mathbb{X}$

It is easily seen that $g_m(x) \le g_{m+1}(x)$, $m = 0, 1, \dots, n$, $x \in \mathbb{X}$. Hence, by the dominated convergence theorem, it follows from (c) that $P^a g_m(x)$ is a continuous function of $a \in \mathbb{A}_x$, $x \in \mathbb{X}$ also for $m = 0, \dots, n$.

In the case of a finite model, this assumption requires the accessibility of the state 0 under each stationary policy from each state. In the denumerable models it requires a strong version of recurrence to the state 0. More precisely, it assumes the finiteness under any policy of the n-th absolute moment of the total cost until the state 0 is reached, with immediate "cost" $c(x, a)$ equal to $|r(x, a)| \vee 1$, $(x, a) \in \mathbb{K}$. For simplicity, in [18] Assumption 7.2 is supposed

to hold for all n. However, the proofs there remain true and provide sharper results if this assumption holds for a fixed n.

As shown in [18], Assumptions 7.1 and 7.2 imply for every $\varphi \in \Pi^s$ a partial Laurent expansion of the form

$$v_\beta^\varphi = (1+\rho) \sum_{k=-1}^{n} h_k^\varphi \rho^k + O(\rho^n) \tag{42}$$

where $O(\rho^n)$ is uniform in φ. Moreover, the coefficients h_k^φ are continuous in $\varphi \in \Pi^s$ in the following topology. The space Π^s in the case of a denumerable $\mathbb{X}$ is the Cartesian (direct) product $\Pi_{x\in\mathbb{X}} \mathbb{A}(x)$, and we take in each $\mathbb{A}(x)$ the Borel topology of a metric space, and in Π^s the product topology.

Using this continuity and the compactness of $A(x)$, we can find for each $x \in \mathbb{X}$ a lexicographically maximal element $H_n^*(x) = \{h_k^*(x),\ -1 \le k \le n\}$ of $H_n^\varphi(x)$ over all $\varphi \in \Pi^s$, and the corresponding maximizer φ_x (see Subsection 1.2 for the lexicographical ordering and notations). Clearly, then

$$\varliminf_{\beta\uparrow 1} \rho^{-(n-1)} [v_\beta(x, \varphi_x) - v_\beta(x, \varphi)] \ge 0, \qquad x \in \mathbb{X}$$

for any other $\varphi \in \Pi^s$, so that φ_x can be called $(n-1)$-discount optimal in the class Π^s for the initial state x. However, in the proper n-discount optimality the same policy should fit for all states x.

The following theorem derived in Hordijk [18] leads to this goal (in Hordijk and Sladky [21] it can be found with a different proof).

Theorem 7.5 *Suppose Assumptions 7.1 and 7.2. Then there exist functions $u_{-1}, u_0, \dots, u_n$ from $\mathbb{X}$ to $\mathbb{R}$ (u_{-1} is a constant) satisfying bounds*

$$|u_m(x)| \le c_m g_m(x), \qquad m = 0, 1, \dots, n$$

for some constants c_m, and nonempty compact (in the product topology) sets

$$\mathcal{P}_{-1} = \Pi^s \supset \mathcal{P}_0 \supset \cdots \supset \mathcal{P}_n$$

of stationary policies of the form $\mathcal{P}_m = \Pi_{x\in\mathbb{X}} \mathbb{A}_m(x)$ $(\mathbb{A}_{-1}(x) = \mathbb{A}(x))$ with the following property. Let $U_m = \{u_{-1}, \dots, u_m\}$. We have $L^\varphi U_m \le 0$ for every $\varphi \in \Pi^s$ and $m = -1, \dots, n$, and $\varphi \in \mathcal{P}_m$ if and only if $L^\varphi U_m = 0$.

It is shown in [21] that $\varphi \in \mathcal{P}_{m+1}$ $(m = -1, \dots, n-1)$ *if and only if φ is m-discount optimal in the class* Π^s (this means that (23) holds for $n = m$ for $\pi^* = \varphi$ and any $\pi \in \Pi^s$). If $\varphi \in \mathcal{P}_{m+1}$, then the coefficients h_k^φ, $-1 \le k \le m$ in (42) coincide with u_k, so that in Theorem 7.5 we have an analogue of the conserving property.

Also, we can conclude from the above results that the value function V_β has a partial Laurent expansion with the coefficients equal to the functions u_m. Hence

$$V_\beta(x) = (1+\rho) \left[u_{-1}\rho^{-1} + u_0(x) + \sum_{k=1}^{m-1} u_k(x)\rho^k \right] + O(\rho^m) \qquad x \in \mathbb{X}, \tag{43}$$

where $u_k = h_k^{\varphi}$, $k = 1, \dots, m-1$ and $\varphi \in \mathcal{P}_{m+1}$. For $m = 1$ this result is established in Cavazos-Cadena and Lasserre [4] under more restrictive recurrence conditions.

Hordijk and Sladky [21] also proved that m-discount optimality is equivalent to m-average optimality for $m = 0, 1, \dots, n-1$ in denumerable CMPs satisfying Assumptions 7.1 and 7.2.

Let us now assume that Assumption 7.2 holds for all $n \in \mathbb{N}$. Then we have nonempty compact sets $\mathcal{P}_n$ for every n, and their intersection

$$\mathcal{P}_\infty = \bigcap_{n=1}^{\infty} \mathcal{P}_n$$

is also a nonempty compact set in Π^s. A policy $\varphi \in \mathcal{P}_\infty$ is n-discount optimal (in the class Π^s) for every $n \in \mathbb{N}$, and it is tempting to conjecture that φ is Blackwell optimal. However, in general this is not true. We return to this question in Section 2.2.

7.2.2 On Blackwell optimality in infinite state models

The original Blackwell definition (Definition 7.1) is too strong for the denumerable state CMPs, as we will see in the counterexample below. In the following definition it is weakened, and a policy satisfying Blackwell's version is renamed into a *strong Blackwell optimal policy*. It is easy to see that in finite models both definitions coincide. Note that Veinott's Definition 4 of n-discount optimality is stated in weak terms, applicable to general models.

Definition 7.3 *For any set $\Pi' \subset \Pi$ a policy $\pi^* \in \Pi'$ is said to be Blackwell optimal within the class Π', if for every $x \in \mathbb{X}$ and $\pi \in \Pi$ there exists a number $\beta_0(x, \pi) < 1$ such that*

$$v_\beta(x, \pi^*) \geq v_\beta(x, \pi) \quad \textit{for all } \beta \in (\beta_0(x, \pi), 1).$$

In the case $\Pi' = \Pi$, π^* is called Blackwell optimal.

Counterexample 1 The state space is

$$\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1 \cup \mathbb{X}_2 \cup \cdots$$

with

$$\mathbb{X}_0 = \{(0, 0)\},$$
$$\mathbb{X}_n = \{(n, 0), (n, 1, 1), \dots, (n, n, 1), (n, 1, 2), \dots, (n, n, 2)\}.$$

The action sets are

$$A((0, 0)) = \left\{0, 1, \frac{1}{2}, \frac{1}{3}, \dots\right\},$$
$$A(n, 0)) = \{1, 2\}, \quad A((n, i, j)) = \{1\} \quad 1 \leq i \leq n, \quad n \geq 1 \quad j = 1, 2.$$

The transition probabilities are

$$p((1,0)|(0,0),\frac{1}{n}) = 1 - p((n,0)|(0,0),\frac{1}{n}) = 1 - 2^{-n}, \quad n = 1,2,\ldots$$
$$p((1,0)|(0,0),0) = 1,$$

and for $n = 1,2,\ldots$

$$p((n,1,1)|(n,0),1) = p((n,1,2)|(n,0),2) = 1,$$
$$p((n,i+1,1)|(n,i,1),1) = p((n,i+1,2)|(n,i,2),1) = 1, \quad 1 \le i \le n-1,$$
$$p((n,0)|(n,n,1),1) = p((n,0)|(n,n,2),1) = 1.$$

The immediate rewards are

$$r((0,0),a) = 1 \quad \forall a \in A(0,0),$$

and for $n = 1,2,\ldots$

$$r((n,0),2) = n, \quad r((n,0),1) = 1,$$
$$r((n,i,1),1) = 1, \quad r((n,i,2),1) = 0, \qquad 1 \le i \le n.$$

Note that this CMP satisfies Assumption 7.1.

Define $\varphi_k, \quad k = 1,2,\ldots$ as follows:

$$\varphi_k(0,0) = \frac{1}{k}$$
$$\varphi_1(n,0) = 1 \text{ and } \varphi_k(n,0) = 2 \text{ for } k = 2,3,\ldots, n = 1,2,3,\ldots$$

It is easy to calculate that

$$v((n,0),\varphi_1,\beta) = (1-\beta)^{-1},$$

and

$$v((n,0),\varphi_k,\beta) = n(1-\beta^{n+1})^{-1} \qquad \text{for} \quad n,k \ge 2.$$

Hence

$$v((n,0),\varphi_1,\beta) \ge v((n,0),\varphi_k,\beta)$$

if and only if $\beta \ge \beta_n$ with β_n being the unique solution of the equation $1 + \beta + \cdots + \beta^n = n$ in the interval $0 \le \beta \le 1$. Since β_n is monotone increasing to 1 as $n \to \infty$, there is no $\beta_0 < 1$ such that for $k \ge 2$,

$$v(x,\varphi_1,\beta) \ge v(x,\varphi_k,\beta) \quad \text{for all} \quad \beta \in [\beta_0,1) \quad \text{and all} \quad x \in \mathbb{X}.$$

Hence φ_1 is not a *strongly* Blackwell optimal policy. Clearly, for fixed initial state $(n,0)$, $n \ge 1$ there is an β_n such that φ_1 is discounted optimal for $\beta \in [\beta_n,1)$.

This is not true for the state $(0,0)$. Indeed, for $k \ge 2$

$$v((0,0),\varphi_1,\beta) - v((0,0),\varphi_k,\beta) =$$
$$\beta[(1-\beta)^{-1} - (1-2^{-n})(1-\beta)^{-1} - 2^{-n}n(1-\beta^{n+1})^{-1}] =$$
$$\beta 2^{-n}[(1-\beta)^{-1} - n(1-\beta^{n+1})^{-1}],$$

which is nonnegative if and only if $\beta \geq \beta_k$. Thus φ_1 is (*weakly*) *Blackwell optimal* in the class Π^s.

We next show that φ_1 is Blackwell optimal in the class of randomized stationary policies. Since the set of states $\{(n,0),(n,1,1),\ldots,(n,n,1),(n,1,2),\ldots,(n,n,2)\}$ is a closed set under any policy, it follows from the results for finite models that φ_1 is Blackwell optimal on this set in the class of all policies. This holds for all $n \geq 1$. Hence it is sufficient to consider a policy φ which only randomizes in state $(0,0)$, say with probability p_k it takes action $\frac{1}{k}$. If φ takes action 1 in $(n,0)$ then $v((n,0),\varphi,\beta) = (1-\beta)^{-1} = v((n,0),\varphi_1,\beta)$, and for computing the difference between $v((0,0),\varphi_1,\beta) - v((0,0),\varphi,\beta)$ we may as well set $p_n = 0$ in this case. So without loss of generality suppose $\varphi(n,0) = 2$ if $p_n > 0$. Then

$$\begin{aligned} v((0,0),\varphi,\beta) &= 1 + \beta \sum_{k=2}^{\infty} p_k \left(2^{-k} \cdot \frac{k}{1-\beta^{k+1}} + (1-2^{-k}) \frac{1}{1-\beta} \right) \\ &= 1 + \frac{\beta}{1-\beta} \sum_{k=2}^{\infty} p_k \left(2^{-k} \frac{k}{1+\beta+\cdots+\beta^k} + (1-2^{-k}) \right). \end{aligned}$$

On the other hand

$$v((0,0),\varphi_1,\beta) = 1 + \frac{\beta}{1-\beta}.$$

Hence,

$$\begin{aligned} f(\beta) :=& \frac{1-\beta}{\beta} (v((0,0),\varphi_1,\beta) - v((0,0),\varphi,\beta)) = \\ & \sum_{k=2}^{\infty} p_k \left(2^{-k} \left(1 - \frac{k}{1+\beta+\cdots+\beta^k} \right) \right). \end{aligned}$$

By dominated convergence

$$\lim_{\beta \uparrow 1} f(\beta) = \sum_{k=2}^{\infty} p_k 2^{-k} \cdot \frac{1}{k+1}.$$

Consequently,

$$\lim_{\beta \uparrow 1} (v((0,0),\varphi_1,\beta) - v((0,0),\varphi,\beta)) = \begin{cases} 0 & \text{if } p_k = 0,\ k \geq 2 \\ \infty & \text{otherwise,} \end{cases}$$

and φ_1 dominates φ for β sufficiently close to 1. Hence φ_1 is Blackwell optimal in the class of randomized stationary policies.

With similar arguments it can be shown that it is Blackwell optimal in the class of all policies. However, φ_1 is not a strong Blackwell optimal policy, since there does not exist a $\beta_0 < 1$ such that $v((n,0),\varphi_1,\beta) = v((n,0),\beta)$ for all $n \geq 1$ and $\beta_0 < \beta < 1$.

We now return to the question whether a policy $\varphi \subset \mathcal{P}_\infty$ is Blackwell optimal (if Assumption 7.2 holds for all n). Take any policy $\psi \in \Pi^s$ and initial state $x \in \mathbb{X}$ and consider the infinite sequences $H^\varphi(x) = \{h_k^\varphi(x)\} = \{u_k(x)\}$ and

$H^{\psi}(x) = \{h_k^{\psi}(x)\}$ (see (41) and Theorem 7.5). By this theorem, $H_m^{\varphi}(x) \succeq H_m^{\psi}(x)$ for every $m \in \mathbb{N}$, hence the same is true for the infinite sequences: $H^{\varphi}(x) \succeq H^{\psi}(x)$. Therefore either $H^{\varphi}(x) = H^{\psi}(x)$, or there is an integer m such that

$$h_k^{\varphi}(x) = h_k^{\psi}(x) \quad \text{for} \quad -1 \leq k \leq m-1,$$
$$h_m^{\varphi}(x) > h_m^{\psi}(x).$$

In the second case it follows from expansions (42) for φ and ψ with $n = m+1$, that $v_{\beta}^{\varphi}(x) > v_{\beta}^{\psi}(x)$ for all β in some interval $(\beta_0(x,\psi), 1)$. The difficulty arises in the first case: there is no guarantee that v_{β}^{φ} and v_{β}^{ψ} have complete Laurent expansions of the form $v_{\beta}^{\varphi} = (1+\rho)\sum_{-1}^{\infty} h_k \rho^k$. Indeed, they may not, as one may see from the following example.

Consider the one-server queue with a controllable Poisson arrival process as studied in Hordijk [18, Section 2.2]. It is shown there that Assumptions 7.1 and 7.2 are satisfied for a given n, if $\mathbb{E}\, S^{n+1} < \infty$, where S is the service time of one customer. Hence if $\mathbb{E}\, S^k < \infty$ for all $k \in \mathbb{N}$ and the rewards $r(x,a)$ are bounded by a polynomial in x, the set $\mathcal{P}_{\infty}$ is nonempty. However, if the Laplace-Stieltjes transform of the service time

$$f(z) = \int_0^{\infty} e^{-zx} dF_S(x)$$

is not analytic at $z = 0$, then the complete Laurent expansion of discounted rewards does not exist for $r(x,a) = \mathbf{1}\{x = 0\}$ and the Poisson arrival process with a positive parameter λ. Indeed, in this case we have a homogeneous random walk as described in Hordijk et al. [20, Section 2].

7.2.3 *Operator theoretical approach to Blackwell optimality*

A satisfactory theory of denumerable state model should contain as a special case the finite model. Assumption 7.2 is not suitable for this purpose. It presupposes a unichain CMP, while the theory of finite CMPs covers the multi chain case too. So it is too restrictive. On the other hand, it does not guarantee the existence of complete Laurent expansions of discounted rewards under stationary policies, and henceforth is too weak to obtain the Blackwell optimality.

In Dekker and Hordijk [8] a theory of denumerable CMPs has been developed free of these inadequacies. The operator theoretical approach to Blackwell optimality is introduced there. In this approach a *bounding function* μ is used satisfying the condition $\mu(x) \geq 1$, $x \in \mathbb{X}$. We relate with μ the Banach space V_{μ} of all real-valued functions f on $\mathbb{X}$ with the finite μ-norm

$$\|f\|_{\mu} = \sup_{x \in \mathbb{X}} \frac{|f(x)|}{\mu(x)}.$$

The associated operator norm is

$$\|T\|_{\mu} = \sup_{f:\|f\|_{\mu} \leq 1} \|Tf\|_{\mu}$$

for any operator $T : V_\mu \to V_\mu$. In the following "bounding assumption" the notations (1), (2), and also the notation

$$\hat{f}(x) = \sup_{a \in \mathbb{A}(x)} |f(x,a)|, \qquad x \in \mathbb{X} \tag{44}$$

for any function f on K, are used.

Assumption 7.3 *For some constant $C > 0$*

(a) $\|\hat{r}\|_\mu \leq C$,

(b) $P^a\mu(x) \leq C\mu(x), \qquad (x,a) \in \mathbb{K}$.

In [8] the part (b) of the compactness-continuity Assumption 7.1 concerning the transition probabilities is strengthened to the following form.

Assumption 7.4 *$P^a f(x)$ is continuous in $a \in \mathbb{A}(x)$ for every $f \in V_\mu$ and $x \in \mathbb{X}$.*

Besides the bounding and compactness-continuity assumptions, we need also a condition to guarantee the Laurent series expansion for v_β^φ, $\varphi \in \Pi^s$ (it should be pointed out that Assumption 7.2 implies Assumptions 7.3 and 7.4 with $\mu = \text{const} \cdot g_n$, but not the complete Laurent expansions). Dekker and Hordijk [8] introduced an ergodicity condition, renamed in Hordijk and Spieksma [23] into μ-geometric ergodicity in the case of a Markov chain, and into uniform μ-geometric ergodicity in the case of a CMP. Note that a CMP can be seen as a compact product set of Markov chains (see Hordijk [17]), and that therefore any ergodicity property of a CMP becomes a corresponding property of a Markov chain if CMP consists of a single chain. Since its introduction in [8], the μ-geometric ergodicity became also a new notion in the Markov processes literature, and has been intensively studied (see Meyn and Tweedie [32]). The *uniform μ-geometric ergodicity condition* is

Assumption 7.5 *For every $\varphi \in \Pi^s$, the t-th convolution $P^t(\varphi)$ of the operator $P(\varphi)$ converges to a limiting stochastic operator $Q(\varphi)$ geometrically fast in the μ-norm, i.e. for some constants $C < \infty$ and $\gamma < 1$*

$$\|P^t(\varphi) - Q(\varphi)\|_\mu \leq C\gamma^t, \qquad \varphi \in \Pi^s, t \in \mathbb{N}. \tag{45}$$

Note that (45) requires the aperiodicity of all Markov chains with kernels P^φ. As shown in Hordijk and Yushkevich [25], the following weaker version of (45) suffices for the study of Blackwell optimality, which covers the case of periodic chains.

Assumption 7.6 *For some constants T, $C < \infty$ and $\gamma < 1$*

$$\left\| \frac{1}{T} \sum_{k=1}^{T} P^{k+t}(\varphi) - Q(\varphi) \right\|_\mu \leq C\gamma^t, \quad \varphi \in \Pi^s, \quad t \in N. \tag{46}$$

Note that condition (46) (even with T independent of φ) is fulfilled in any finite CMP (because it is true for a finite Markov chain, and the number of Markov chains corresponding to stationary policies is finite). In a denumerable Markov chain, this condition implies the existence of the deviation operator analogous to the deviation matrix considered in Section 1.1. The following lemma (see [25]) is indeed a general fact on operators in Banach spaces with a convergent resolvent. We state it in the context of the μ-norm and operators $P(\varphi)$.

Lemma 7.2 *Assumption 7.6 implies the existence of the stationary operator $Q(\varphi) = Q^\varphi$ and the deviation operator $D(\varphi) = D^\varphi$ as defined in (6) and (7') (the limits are understood in the μ-norm). Moreover, the equations in (6) for Q^φ and in (8)–(9) for D^φ are satisfied.*

We have seen in Section 1.2 that in the case of a finite state space the resolvent

$$R_\beta(P) = R(\rho, P) = \sum_{t=0}^{\infty} \left(\frac{P}{1+\rho}\right)^t = \sum_{t=0}^{\infty} (\beta P)^t = (I - \beta P)^{-1} \tag{47}$$

of the transition operator P has a Laurent series expansion in the neighborhood of $\rho = 0$, reflected in formulas (13) and (16)-(18) together with $v_\beta^\varphi = R_\beta(P^\varphi) r^\varphi$, implied by equations (6) and (8)-(9). The same expansion remains true in the case of a Markov chain on a general state space $\mathbb{X}$, if there is a geometric convergence (46).

Lemma 7.3 *If Assumption 7.6 holds, then there exists a number $\rho_0 > 0$ such that for all complex values of ρ in the ring $0 < |\rho| < \rho_0$*

$$R(\rho, P^\varphi) = (1+\rho)\left[\frac{Q^\varphi}{\rho} + \sum_{n=0}^{\infty} (-\rho)^n (D^\varphi)^{n+1}\right], \tag{48}$$

where Q^φ and D^φ are the same bounded (in the μ-norm) operators as in Lemma 7.2.

Proof. Essentially, it is the same algebra based on equations (8)-(9) as in the proof of Theorem 7.2. However, there we knew beforehand that the resolvent has a simple pole at $\rho = 0$. To avoid this, one may check by direct algebra that the series (48) (which converges in some ring $0 < |\rho| < \rho_0$, because the operator D^φ is bounded) defines an operator inverse to $I - \beta P$ (see (47)). Indeed, we

have according to (8)–(9) and (6) (and omitting the index φ)

$$
\begin{aligned}
(1+\rho)\left[\frac{Q}{\rho}+\sum_0^\infty(-\rho^n)(D)^{n+1}\right]\left(I-\frac{P}{1+\rho}\right) &= \\
=\left[\frac{Q}{\rho}+\sum_0^\infty(-\rho)^n D^{n+1}\right][(I-P)+\rho I] &= \\
= Q+\sum_0^\infty(-\rho)^n D^{n+1}(I-P)+\sum_0^\infty(-1)^n(\rho D)^{n+1} &= \\
= Q+\sum_0^\infty(-\rho)^n D^n(I-Q)-\sum_1^\infty(-\rho)^n D^n &= \\
= Q+I-\sum_0^\infty(-\rho)^n D^n Q = Q+I-Q = I, &
\end{aligned}
$$

and the same holds if we multiply in the reverse order. ∎

To proceed further, we need to extend the space $\mathfrak{H}$ and operators in it, introduced in Section 1.3, to the case of a countable state space $\mathbb{X}$ and the Banach space V_μ of functions on $\mathbb{X}$. Clearly, instead of sequences of Laurent coefficients we may consider the Laurent series themselves.

Definition 7.4 *(a) The linear space $\mathfrak{H}_\mu$ consists of all Laurent series of the form*

$$h := h(x) = h(x,\rho) = \sum_{n=-1}^{\infty} h_n(x)\rho^n, \qquad h_n \in V_\mu, \quad x \in \mathbb{X}$$

in the complex variable ρ with coefficients satisfying the geometric growth condition

$$\overline{\lim_{n\to\infty}} \|h_n\|_\mu^{\frac{1}{n}} < \infty.$$

(b) For every $\psi \in \Pi^s$, operators L^ψ and U^ψ in the space $\mathfrak{H}_\mu$ are given by the formulas

$$
\begin{aligned}
L^\psi h &= r^\psi + P^\psi h - (1+\rho)h \qquad h \in \mathfrak{H}_\mu, \\
U^\psi &= \frac{\rho}{1+\rho} R(\rho, P^\psi). \qquad (49)
\end{aligned}
$$

We use in $\mathfrak{H}_\mu$ the lexicographical ordering defined in Section 1.2 for the sequences of their coefficients $H = \{h_{-1}, h_{0,\dots}\}$. Clearly, if $H' \preceq H''$, or equivalently $h' \preceq h''$, then $h'(x,\rho) \preceq h''(x,\rho)$ for all positive sufficiently small ρ, etc. Note that the definition (49) of L^ψ is consistent with formulas (33)–(35).

Lemma 7.4 *Suppose Assumption 7.6. Then for every $\varphi \in \Pi^S$ the formulas (16)–(18) of Theorem 7.2 are valid, with coefficients h_n^φ (or k_n^φ) $\in V_\mu$. If Assumption 7.5 holds (or if T in Assumption 7.6 it is the same for all φ), then*

$$\overline{\lim_{n\to\infty}} \left[\sup_{\varphi\in\Pi^s} \|h_n^\varphi\|\right]^{\frac{1}{n}} < \infty.$$

This lemma is a direct consequence of the formula $v_\beta^\varphi = R(\rho, P^\varphi) r^\varphi$ and Lemma 7.3. The corresponding element of $\mathfrak{H}_\mu$ we denote $h^\varphi = \sum_{-1}^{\infty} h_n^\varphi \rho^n$, so that

$$v_\beta^\varphi = (1+\rho) h^\varphi.$$

The following comparison lemma is an extension to denumerable models of a result derived by Veinott [43] for finite models.

Lemma 7.5 (Comparison lemma) *Suppose Assumptions 7.1, 7.3, 7.4 and 7.6. Then for every $h \in \mathfrak{H}_\mu$ and $\psi \in \Pi^s$ there exists $\rho_0 > 0$ such that*

$$h^\psi - h = \frac{1}{\rho} U^\psi L^\psi h, \qquad 0 < |\rho| < \rho_0.$$

Central in the analysis of [8] is the following lemma which is the key lemma in the operator theoretical approach.

Lemma 7.6 (Key lemma) *Under assumptions of Lemma 7.5 for every $\psi \in \Pi^s$ the operator U^ψ is a positive operator: if $h \in \mathfrak{H}_\mu$, $h \succeq 0$ then $U^\psi h \succeq 0$ (if $h \preceq 0$ then $U^\psi h \preceq 0$). Moreover, if $h \succeq 0$ and $h(x_0) \succ 0$ for some $x_0 \in \mathbb{X}$, then $Uh(x_0) \succ 0$.*

The key lemma together with the comparison lemma yield the lexicographical policy improvement approach for the denumerable models (cf. Lemma 7.1 for a finite models). For the average criterion in finite models policy improvement was developed by Howard [27] and Blackwell [3] and for the sensitive criteria in those models by Veinott [43]. In their honor we call it Howard-Blackwell-Veinott policy improvement.

In finite models policy improvement is a constructive way to find a Blackwell optimal policy, as sketched in Section 1.3. In infinite state space models this algorithm is not sufficient, since the equations are infinite and the improvements may not terminate in a finite number of steps. Lemmas 7.5–7.6 provide an approach to Blackwell optimality which is different in that it uses positive operators on the linear space of Laurent series with the norm induced by a weighted supremum norm. It overcomes the infinite state space problems, and it generalizes the theory developed by Blackwell, Miller and Veinott.

The following theorem gives several equivalent formulations of Blackwell optimality in the class Π^s of stationary policies. Its proof is rather a direct consequence of the comparison and key lemmas. We refer to formulas (39)–(41) and Theorem 7.4 for notations, terminology and a comparison with the case of a finite model.

Theorem 7.6 *Suppose Assumptions 7.1, 7.3, 7.4 and 7.6. Then the following statements concerning a policy $\varphi \in \Pi^s$ are equivalent:*

(a) φ is Blackwell optimal within the class Π^s;

(b) $h^\varphi \succeq h^\psi$ for every $\psi \in \Pi^s$;

(c) h^φ is a solution of the Blackwell optimality equation $Lh = 0$;

(d) $L^{\varphi}h = 0$ *for a solution* h *of the equation* $Lh = 0$ *(i.e.* φ *is a conserving policy).*

Moreover, the solution of the equation $Lh = 0$ *(if any) is unique.*

It is shown in [8] that under continuity and uniform μ-geometric ergodicity assumptions, $P^{\varphi}\mu$, $Q^{\varphi}\mu$ and $D^{\varphi}\mu$ are continuous functions of φ on the compact Π^s (in the product topology). This, together with the bounding assumption, implies the continuity in φ of the Laurent coefficients $h_n^{\varphi}(x)$ given by formulas (17) and (18). This, together with a diagonal process on the countable set X, allows to get a maximizer $h^{\varphi} = \max_{\psi} h^{\psi}$, $\psi \in \Pi^s$ as in part (b) of Theorem 7.6. By this theorem, φ is Blackwell optimal in the class Π^s, and the Blackwell optimality equation has a solution. A technical proof given in [8] shows that φ is Blackwell optimal also in the class Π of all policies (versions of this proof for Borel models can be found in [49, 55] and [26]). Thus, the following result holds.

Theorem 7.7 *In a denumerable state space model satisfying Assumptions 7.1 and 7.3–7.5 there exists a Blackwell optimal policy.*

A related question is whether a limit φ of β-optimal policies φ_β (if the limit exists as $\beta \uparrow 1$) is Blackwell optimal. Under weaker assumptions than above, it is shown in Hordijk [18], that such φ is 0-discount optimal. Under another set of assumptions, which can be shown to be more restrictive than Hordijk's conditions, the same result is obtained by Cavazos-Cadena and Lasserre [4]. On the other hand, Hordijk and Spieksma [22] have constructed an example in which the limiting policy φ is not 2-discount optimal, so a fortiori not Blackwell optimal.

Lasserre [30], starting from the ideas developed in [8], obtained the existence of a Blackwell optimal policy within Π^s (and hence, according to [8], in the class Π too) without the policy improvement, making use of more results in the spectral theory of bounded linear operators.

Yushkevich [52] has shown how one may get the existence of Blackwell optimal policies in denumerable models with periodic chains by perturbing them into aperiodic models. In this work, besides Assumption 7.1, the boundedness of the reward function r was assumed, as well as the following condition taken from Tijms [41]: there are a number $\epsilon > 0$ and an integer T such that for every $\varphi \in \Pi^s$ there exists a state $y_T = y(\varphi)$ such that

$$\sum_{t=1}^{T} \mathbb{P}_x^{\varphi}\{x_t = y\} \geq \epsilon, \qquad x \in X.$$

In fact, we get the most general results if we use Assumptions 7.5 (and 7.6) in a nonuniform way; this is true for the above results except the continuity of $Q(\varphi)$ and $D(\varphi)$. So, for Theorem 7.6 the following weak versions of Assumptions 7.5 and 7.6 are sufficient.

Assumption 7.7 *For every* $\varphi \in \Pi^s$*, there exist constants* $C(\varphi) < \infty$ *and* $\gamma(\varphi) < 1$ *such that*

$$\|P^t(\varphi) - Q(\varphi)\|_{\mu} \leq C(\varphi)\gamma^t(\varphi).$$

Assumption 7.8 *For every $\varphi \in \Pi^s$, there exist $T(\varphi)$, $C(\varphi)$ and $\gamma(\varphi) < 1$ such that*

$$\left\| \frac{1}{T(\varphi)} \sum_{k=1}^{T(\varphi)} P^{k+t}(\varphi) - Q(\varphi) \right\|_\mu \leq C(\varphi)\gamma^t(\varphi).$$

For continuity of $Q(\varphi)$ and $D(\varphi)$ in φ and therefore for Theorem 7.7 we need the uniform Assumption 7.5 (or 7.6).

7.2.4 Recurrence conditions for Blackwell optimality

In [10] recurrence conditions are introduced which imply the existence of complete Laurent series and of Blackwell optimal policies. Starting in Ross [35], recurrence conditions have been extensively used and studied in undiscounted nonfinite CMPs.

The first analysis based on the notion of '*simultaneous Doeblin condition*', can be found in Hordijk [17]. This condition states that for some finite set M, integer k and constant c, for every $\varphi \in \Pi^s$ and $x \in \mathbb{X}$

$$\mathbb{P}_x^\varphi\{x_k \in M\} \geq c > 0.$$

Many equivalent formulations of this condition have been derived. We refer to [17], Federgruen, Hordijk and Tijms [13, 14], Thomas [40], Hernández-Lerma, Montes-de-Oca, Cavazos-Cadena [16], Hordijk and Spieksma [23], Dekker, Hordijk and Spieksma [11]. We present here some of the results which appeared in the last of these papers.

The taboo transition matrix ${}_MP$ with taboo set $M \subset X$ is defined by

$${}_Mp_{xy} = \begin{cases} p_{xy} & y \notin M \\ 0 & y \in M, \end{cases}$$

with the convention that ${}_MP^t$ is the t-fold matrix-product of ${}_MP$, and ${}_MP^0 = I$, with I the identity-matrix. The *uniform μ-geometric recurrence condition* (in the weak form) is

Assumption 7.9 *There is a finite set M and constants $c < \infty$ and $\gamma < 1$ such that*

$$\| {}_MP^t(\varphi)\|_\mu < c\gamma^t, \qquad t = 0, 1, \ldots, \quad \varphi \in \Pi^S.$$

For the special case $\mu = e$, where e the function with $e(x) = 1$ for all $x \in X$, uniform μ-geometric recurrence is equivalent to the simultaneous Doeblin condition (see Hordijk [18, Theorem 11.3] and especially relation (11.3.2)). The generalization from e to a general (mostly unbounded) bounding function μ is important. It gives not only results for unbounded reward functions (see the bounding Assumption 7.3), but also covers the class of CMP satisfying the uniform μ-geometric recurrence condition for a suitable chosen μ which is essentially larger than that of satisfying the simultaneous Doeblin condition. Indeed, let τ be the recurrence time to the set M, then

$$\{\tau > t\} = \{x_k \notin M, \quad 1 \leq k \leq t\}.$$

Hence,

$$\mathbb{P}_x^{\varphi}\{\tau > t\} = \left({}_M P^t(\varphi)e\right)(x)$$

and under Assumption 7.9,

$$\mathbb{E}_x^{\varphi}\,\tau = \sum_{t=0}^{\infty} \mathbb{P}_x\{\tau > t\} = \sum_{t=0}^{\infty} \left({}_M P^t(\varphi)e\right)(x) \leq \frac{c}{1-\gamma} \cdot \mu(x).$$

Consequently, if μ is bounded then the expected recurrence time is bounded in the starting state. Clearly, this does not hold for most queueing models. See Spieksma [39] for CMPs, especially controlled queues, which do satisfy the uniform μ-geometric recurrence condition.

Let $\nu(\varphi)$ denote the number of closed classes in the Markov chain with transition probabilities $P(\varphi)$. A set $B(\varphi) \subset X$ is called a set of '*reference states*' if it contains precisely one state from each closed class and no other states.

An apparently stronger version of Assumption 7.9 is

Assumption 7.10 *There is a finite set M and constants $c < \infty$ and $\gamma < 1$ such that for every $\varphi \in \Pi^s$ there exists a reference set $B(\varphi) \subset M$, and moreover*

$$\| {}_{B(\varphi)} P^t(\varphi) \|_{\mu} < c\gamma^t, \quad t = 0, 1, \ldots .$$

In [11] it is shown that Assumption 7.9, together with the continuity of $\nu(\varphi)$ as function of φ, is equivalent to Assumption 7.10. Dekker and Hordijk [10] analyzed and proved the existence of Laurent expansions and Blackwell optimality under Assumption 7.10.

It was Hordijk's conjecture that (uniform) μ-geometric ergodicity is equivalent to (uniform) μ-geometric recurrence. This was proved in Hordijk and Spieksma [23] for one Markov chain and has been generalized for the unichain case to general Borel state space by Meyn and Tweedie [32]. For CMPs the equivalence is more complicated , it can be found in Dekker, Hordijk and Spieksma [11].

Note that for the finite model $\nu(\varphi)$ is automatically continuous since Π^S is a finite set. Moreover, in Assumption 7.9 we may take $M = \mathbb{X}$; then ${}_M P$ is the zero matrix. Therefore Assumption 7.9, and hence also Assumption 7.10, are always fulfilled in finite models, also in the multichain case.

One might ask whether Assumption 7.10 can be weakened. Using the existence and continuity of the Laurent expansion of the discounted rewards, one may show that for the operator theoretical approach of Dekker and Hordijk the Assumption 7.10 is also necessary (see Lasserre [30], Spieksma [39]).

Let us conclude this section with pointing out the relation between the Assumptions 7.2 and 7.9.

First, Assumption 7.2 is more restrictive, since it assumes the existence of one state which is accessible from all other states under each policy, i.e. the unichain case, whereas Assumption 7.9 allows a finite number of closed sets. Let us assume the unichain case (note that in the unichain case $\nu(\varphi) \equiv 1$ and so it is continuous), then Assumption 7.9 implies Assumption 7.10 and with $B(\varphi) = \{0\}$ we have

$$\| {}_0 P^t(\varphi) \|_{\mu} < c\gamma^t, \qquad t = 0, 1, \ldots$$

with $c < \infty$ and $\gamma < 1$.

Define

$$\widetilde{\mu} = \sup_{\varphi} \sum_{t=0}^{\infty} {}_0P^t(\varphi)\mu.$$

Then,

$$\mu +_0 P(\varphi)\widetilde{\mu} \leq \widetilde{\mu} \qquad \forall \varphi \in \Pi^S$$

and

$${}_0P(\varphi)\widetilde{\mu} \leq \widetilde{\mu} - \mu \leq \left\{1 - \frac{1-\gamma}{c}\right\} \widetilde{\mu}$$

and

$$\widetilde{\mu} +_0 P(\varphi)\tilde{\tilde{\mu}} \leq \tilde{\tilde{\mu}}. \tag{50}$$

with $\tilde{\tilde{\mu}} = \frac{c}{1-\gamma}\widetilde{\mu}$. Since $\mu \leq \widetilde{\mu} \leq \frac{c}{1-\gamma}\mu$, Assumption 7.3 for μ implies the same assumption for $\widetilde{\mu}$. By using (50) recursively, it is easily seen that Assumption 7.2 is satisfied for

$$g_n = \left(\frac{c}{1-\gamma}\right)^{n+1} \widetilde{\mu}.$$

Hence for the unichain case Assumption 7.9 implies Assumption 7.2.

With a slightly more involved argument one may show that Assumption 7.9 is equivalent to

$$s(\varphi) := \sum_{t=0}^{\infty} {}_MP^t(\varphi)\mu \leq c_1\mu, \tag{51}$$

for some constant c_1 and all $\varphi \in \Pi^S$. Indeed (cf. [10]), (51) implies that

$${}_MP(\varphi)s(\varphi) = s(\varphi) - \mu \leq \left(1 - \frac{1}{c_1}\right) s(\varphi).$$

Choose $\gamma_1 < 1$ and let t_0 be such that $\left(1 - \frac{1}{c_1}\right)^{t_0} c_1 < \gamma_1$; then

$${}_MP^{t_0}(\varphi)\mu \leq {}_MP^{t_0}(\varphi)s(\varphi) \leq \left(1 - \frac{1}{c_1}\right)^{t_0} s(\varphi) \leq \gamma_1\mu.$$

Let $c_1 = \sup_\varphi {}_MP(\varphi)\mu$, $c = (c_1 \vee 1)^{t_0}\gamma_1^{-1}$ and $\gamma = \gamma_1^{1/t_0}$. Then $\gamma < 1$, and for $kt_0 \leq t < (k+1)t_0$, $k \geq 0$ we have

$$\begin{aligned} \| {}_MP^t(\varphi)\|_\mu &\leq \| {}_MP^{kt_0}(\varphi)\|_\mu \| {}_MP^{t-kt_0}(\varphi)\|_\mu \\ &\leq \gamma_1^k c_1^{t-kt_0} \leq c\gamma^t. \end{aligned}$$

Hence Assumption 7.9 is satisfied with $c < \infty$ and $\gamma < 1$.

Mention also that Cavazos–Cadena and Lasserre [5] suggested a different approach to Blackwell optimality in countable models with bounded rewards, satisfying Assumption 7.1 and the following uniform recurrence assumption which is equivalent to the simultaneous Doeblin condition (see Hordijk [18]). Let τ_y be the first entrance time of the trajectory $x_1, x_2, \ldots$ into the state

$y \in \mathbb{X}$. There exists a constant $K < \infty$ such that for every stationary policy φ one may find a state $y(\varphi)$ so that

$$\mathbb{E}_x^{\varphi} \tau_{y(\varphi)} \le K, \quad x \in \mathbb{X}.$$

Under those assumptions, they reduce the construction of a Blackwell optimal policy to a suitably defined sequence of CMP's for which the average optimality equations should be simultaneously solved.

Remark. The Laurent series expansion of the discounted rewards and the existence of strong Blackwell optimal policies for *semi*-Markov decision chains with a finite number of states and actions has been established in Denardo [12]. Similar results under related recurrence conditions have been obtained for the denumerable state model in Dekker and Hordijk [9].

7.3 BOREL STATE MODELS

In this section we consider Blackwell optimality in CMPs with a Borel state space. We formulate the existence results, describe distinctive features of the approach to Borel models, state recurrence conditions which imply less verifiable uniform ergodicity and integrability assumptions.

7.3.1 Existence of Blackwell optimal policies

The study of Blackwell optimality in Borelian models was started by Yushkevich [49, 53] and continued by Hordijk and Yushkevich [25, 26]. An extended summary of results obtained in [53] can be found in [55]. A related paper is Yushkevich [54], where the compactness of the policy space is treated.

An advance in the direction of Borel models appeared possible in the case when the transition probabilities are given by transition densities. This is a common case in models with a continuous state space. We also need the corresponding versions of the compactness-continuity Assumption 7.1 and either of the uniform geometric ergodicity Assumption 7.5 or of recurrence conditions implying Assumption 7.5. Models with a bounded reward function and a strong minorant or simultaneous Doeblin-Doob condition were treated in [53]; the particular case of finite action sets $\mathbb{A}(x)$ was studied before that in [49]. In models with unbounded rewards considered in [25, 26], one needs a stronger version of the bounding Assumption 7.3, and the ergodicity or recurrence conditions should be stated in the terms of μ-norms; also a technical uniform integrability condition is needed in the absence of recurrence conditions.

To avoid repetitions, we first state the more general results obtained in [25, 26]. Before stating the whole set of conditions, we introduce notations related to transition densities and randomized stationary policies; the formulas will become meaningful under subsequent assumptions. There is a *reference measure* $m(dx)$ on the space $\mathbb{X}$, and we often write dx instead of $m(dx)$. The transition probabilities $p(Y \mid x, a)$ are determined by *transition densities* $p(x, a, y)$ so that (for measurable $Y \subset \mathbb{X}$)

$$p(Y \mid x, a) = \int_Y p(x, a, y) m(dy), \qquad (x, a) \in \mathbb{K}.$$

Similar to (44), we denote (because the maximum exists)

$$\hat{p}(x,y) = \max_{a \in \mathbb{A}(x)} p(x,a,y), \qquad x,y \in \mathbb{X}.$$

Formula (2) takes on the form

$$P^a f(x) = \int_{\mathbb{X}} p(x,a,y) f(y) m(dy).$$

For uniformity with other notations of this section, we denote by $\sigma(x,da)$ the probability measure $\sigma(\cdot \mid x)$ on $\mathbb{A}(x) \subset \mathbb{A}$ defined by a *randomized stationary policy* $\sigma \in \Pi^{RS}$. The transition density corresponding to $\sigma \in \Pi^{RS}$ is

$$p^\sigma(x,y) = \int_{\mathbb{A}} p(x,a,y)\sigma(x,da), \qquad x,y \in \mathbb{X},$$

the corresponding transition operator is P^σ:

$$P^\sigma f(x) = \int_{\mathbb{X}} p^\sigma(x,y) f(y) m(dy).$$

Finally, we need multistep transition densities corresponding to a *randomized Markov policy* $\pi = \{\sigma_1, \sigma_2, \dots\} \in \Pi^{RM}$ where $\sigma_t \in \Pi^{RS}$. They are defined recursively by the formulas

$$p_1^\pi(x,y) = p^{\sigma_1}(x,y), \qquad p_{t+1}^\pi(x,y) = \int_{\mathbb{X}} p_t^\pi(x,z) p^{\sigma_{t+1}}(z,y) m(dz).$$

We also have a bounding function μ on X and the corresponding μ-norms (see Section 2.3). In the definition of the space V_μ it is understood that $f \in V_\mu$ is measurable (throughout this section measurability means Borel measurability).

Assumption 7.11 *(a) $\mathbb{X}$ is a standard Borel space with a σ-finite measure m in it, $\mathbb{A}$ is a Borel set in a Polish (= complete separable metric) space, the set $\mathbb{K}$ (see (1)) is measurable in $\mathbb{X} \times \mathbb{A}$, transition densities $p(x,a,y) \geq 0$ and rewards $r(x,a)$ are measurable functions on $\mathbb{K} \times \mathbb{X}$ and $\mathbb{K}$ respectively, $\mu(x) \geq 1$ is a measurable function on $\mathbb{X}$.*

(b) $\mathbb{A}(x)$, $x \in \mathbb{X}$ are nonempty compact sets, functions $p(x,a,y)$ and $r(x,a)$ are continuous in $a \in \mathbb{A}(x)$ for every $x,y \in \mathbb{X}$.

(c) $\|\hat{r}\|_\mu < \infty$ (cf. (44)), and for some constant $C > 0$

$$\int_{\mathbb{X}} \hat{p}(x,y)\mu(y) m(dy) \leq C\mu(x), \qquad x \in \mathbb{X}.$$

(d) Operators $(P^\sigma)^t, \sigma \in \Pi^{RS}$ converge in the μ-norm to limiting operators Q^σ geometrically fast and uniformly in σ as $t \to \infty$: there exist positive constants $C < \infty$ and $\gamma < 1$ such that

$$\|(P^\sigma)^t - Q^\sigma\|_\mu \leq C\gamma^t, \qquad \sigma \in \Pi^{RS}, t = 0,1,2,\dots$$

The following result is proved in [25].

Theorem 7.8 *If Assumption 7.11 holds, then there exists a stationary policy $\varphi \in \Pi^s$ Blackwell optimal in the class Π^{RS} of randomized stationary policies. Also, all assertions of Theorem 7.6 hold (for policies $\varphi, \psi \in \Pi^S$ or Π^{RS}).*

Some partial results are true under milder assumptions. For example, Laurent series expansion of $v_\beta(\sigma)$ for $\sigma \in \Pi^{RS}$ and the analogue of Theorem 7.6 are valid under Assumption 7.8 or 7.7, and also the Laurent series expansion of $v_\beta(\sigma)$ is valid under an analogue of Assumption 7.6 in place of Assumption 7.11(d). For Blackwell optimality in the class Π of all policies in general we need the following uniform integrability assumption.

Assumption 7.12 *For every $x \in \mathbb{X}$, randomized Markov policy $\pi \in \Pi^{RM}$, and $\epsilon > 0$, there exist a set $Y \subset \mathbb{X}$ with $m(Y) < \infty$ and a constant $L > 0$ such that*

$$\int_{\mathbb{X}\setminus Y} p_t^\pi(x,y)\mu(y)m(dy) < \epsilon, \qquad t = 1,2,3,\dots,$$
$$p_t^\pi(x,y)\mu(y) \le L, \qquad y \in Y, \quad t = 1,2,3,\dots.$$

The following result is proved in [26].

Theorem 7.9 *Under Assumptions 7.11 and 7.12, every policy $\varphi \in \Pi^s$ Blackwell optimal in the class Π^{RS}, is Blackwell optimal in the class Π as well.*

In the earlier work [53], results of Theorems 7.8 and 7.9 were obtained in the case of *bounded transition densities* $p(x,a,y)$, *bounded rewards* $r(x,a)$ (so that one may take $\mu \equiv 1$), a *finite measure* m, and the following *minorant condition*: there exist a set Y with $m(Y) > 0$ and a number $\delta > 0$ such that

$$p(x,a,y) \ge \delta, \qquad (x,a) \in \mathbb{K}, \quad y \in Y.$$

In that case Assumptions 7.11(c) and 7.12 hold trivially, while the geometric convergence as in 7.11(d) is shown to be true even for the densities $p_t^\sigma(x,y)$. Of course, one has to suppose Assumptions 7.11(a,b) (with $\mu = 1$).

In related papers [50, 51] Yushkevich proved a partial expansion

$$V_\beta(x) = (1+\rho)\left(h_{\frac{-1}{\rho}} + h_0\right) + o(1)$$

(cf. (43)) for Borel models satisfying assumptions of the preceding paragraph.

7.3.2 Specific features of Borel models

In the study of Blackwell optimality in Borel state models there are several features which make it different from that of finite and denumerable CMPs. They are: (i) utilization of the class Π^{RS} instead of Π^S in the Laurent series expansions and related topics; (ii) introduction of the weak-strong topology in the space Π^{RS} based on Carathéodory functions, (iii) lexicographical maximization of expected discounted rewards not pointwise at every state but for

some absolutely continuous initial distribution; (iv) utilization of the policy improvement to get Blackwell optimal policy $\varphi \in \Pi^s$ from a maximizing policy $\sigma \in \Pi^{RS}$.

We have to work with the class Π^{RS} instead of Π^S because the latter is not a compact space in a reasonable sense. It should be clear from the following simple example. Let $\mathbb{X} = [0,1)$ and $\mathbb{A} = \mathbb{A}(x) = \{1,2\}$. For every $m = 1,2,\ldots$ let the stationary policy φ_m be defined by the rule: if $x \in [(k-1)2^{-m}, k2^{-m})$ then $\varphi_m(x) = 1$ for odd values of k and $\varphi_m(x) = 2$ for even values of k. Every $\varphi_m \in \Pi^s$, but the only reasonable limit of the sequence $\varphi_1, \varphi_2, \ldots$ is the randomized policy $\sigma \in \Pi^{RS}$ with the distribution $\sigma(1 \mid x) = \sigma(2 \mid x) = 1/2$.

Under assumptions of Section 3.1, Laurent series expansions for $v_\beta(\sigma)$ with $h^\sigma \in \mathfrak{H}_\mu$ as in Lemma 7.4 are valid for $\sigma \in \Pi^{RS}$. Along the same way as in denumerable models, with only technical differences, one justifies the lexicographical policy improvement, and this leads to an analogue of Theorem 7.6, (a) to (d), but for policies σ in the whole space $\Pi^{RS} \supset \Pi^S$.

In denumerable models we used the product topology in the space Π^S defined in Section 2.1. The appropriate topology in Π^{RS} is the so-called *weak-strong* or *ws-topology*. In this topology $\sigma_k \to \sigma$ iff

$$\lim_{k\to\infty} \int_{\mathbb{K}} f(x,a)\sigma_k(x,da)m(dx) = \int_{\mathbb{K}} f(x,a)\sigma(x,da)m(dx)$$

for all Carathéodory functions f (i.e. functions continuous in a, measurable in x) satisfying some bounding condition in terms of $\hat{f}(x)$ and measure m. In this topology, Π^{RS} is a compact space with all needed properties. In the deterministic control theory essentially the same fact was used by Warga [47] in connection with relaxed controls. In the nonstationary stochastic dynamic programming the related compactness of the set of all measures corresponding to a given initial distribution in the ws-topology was proved by Schäl [36, 37] and Balder [2]. Another proof, especially for the space Π^{RS}, is given in Yushkevich [54]. Compactness proved in the above references covers the case of a finite reference measure m. For the σ-finite measure m it is proved in [25].

Assumptions of Section 3.1 imply the continuity in $\sigma \in \Pi^{RS}$ of the operators P^σ and Q^σ, and after that, through formulas for h_n^σ as in Theorem 7.2 and the power series in P^σ for $(D^\sigma)^n$ obtained from (7), the continuity of the coefficients $h_n^\sigma(x)$. This implies the continuity of $h_n^\sigma(\ell) = \int_{\mathbb{X}} h_n^\sigma(x)\ell(x)m(dx)$ for any initial density ℓ. Taking a strictly positive density ℓ on $\mathbb{X}$, we lexicographically maximize $h^\sigma(\ell) = \{h_n^\sigma(\ell), n \geq -1\}$ over Π^{RS}, and get a "best" policy σ^* for the initial distribution ℓ.

By applying to σ^* one step of the lexicographical policy improvement simultaneously at all states of the space X, we get a policy $\varphi \in \Pi^s$. With the help of the already proven part of Theorem 7.6, it is now not difficult to show that φ is Blackwell optimal in the class Π^{RS}, and that the Blackwell optimality equation has a unique solution in $\mathfrak{H}_\mu$.

The proof that a policy Blackwell optimal in Π^{RS} is Blackwell optimal in the whole space Π is even more technical than in the denumerable case. It utilizes the main idea of the proof in [8], Assumption 7.12, and an additional property

of the ws-topology; see [26], or for the special case of bounded rewards, [49] or [53].

7.3.3 Recurrence conditions for Blackwell optimality

The uniform geometric μ-ergodicity condition and the uniform integrability condition (Assumptions 7.11(d) and 7.12) are difficult to verify in CMPs with a noncompact state space and an unbounded reward function. In Hordijk and Yushkevich [26] simpler recurrence and drift conditions are given, which imply those assumptions. This approach is based on ideas developed in Hordijk and Spieksma [23] and Hordijk et al. [24], with an additional use of the weak-strong topology. Consider the following set of conditions.

Assumption 7.13 *(a) (Uniform minorant condition) There exist sets $D, Y \subset \mathbb{X}$ with $m(D) > 0$, $m(Y) > 0$ and a number $\delta > 0$ such that*

$$p(x,a,y) \geq \delta, \qquad x \in D, \quad a \in \mathbb{A}(x), \quad y \in Y.$$

(b) (Uniform drift condition) There exist a set $D \subset \mathbb{X}$ with $m(D) > 0$ and numbers $b > 0$, $0 < \gamma < 1$ such that

$$\sup_{x \in D} \mu(x) < \infty$$

and

$$\int_{\mathbb{X}} p(x,a,y)\mu(y)m(dy) \leq \gamma\mu(x) + b\mathbf{1}_D(x), \qquad (x,a) \in \mathbb{K}.$$

(c) (Uniform accessibility condition) There exists a set $D \subset \mathbb{X}$, and for every sublevel set

$$M_c = \{x : \mu(x) \leq c\}$$

there exist a number $\eta > 0$ and an integer N such that

$$\int_D P_N^\sigma(x,y)m(dy) \geq \eta, \qquad x \in M_c, \quad \sigma \in \Pi^{RS}.$$

(d) (Dominance integrability condition) There exist a set $D \subset \mathbb{X}$ with $m(D) > 0$ and a measurable function $\ell \geq 0$ and $\mathbb{X}$ such that

$$\int_{\mathbb{X}} \ell(x)\mu(x)m(dx) < \infty \quad \text{and} \quad \hat{p}(x,y) \leq \ell(y), \quad x \in D, \quad y \in \mathbb{X},$$

and also $m(M_c) < \infty$ for every sublevel set M_c with $c \geq 1$.

Omitting some details, we summarize those relations between conditions which provide the existence of Blackwell optimal policies. It follows from [24], that Assumptions 7.13 (a,b,c) with the same set D, together with 7.11(a) and the condition $P^a\mu(x) \leq C\mu(x)$ imply the uniform integrability Assumption 7.11(d). Also, if the density $p(x,a,y)$ is bounded, Assumptions 7.13(b,d)

together with 7.11(a,b,c) imply Assumption 7.12 (with a possible change of the function μ, which does not affect the made assumptions). The proof of the last result essentially follows the proof of a similar result for denumerable models in Dekker et al. [11], with the use of ws-topology. As a consequence, we have the following theorem.

Theorem 7.10 *In CMP satisfying Assumptions 7.11(a,b,c) and 7.13(a,b,c), there exists a stationary policy $\varphi \in \Pi^s$ Blackwell optimal in the class Π^{RS}, and all assertions of Theorem 7.6 hold. If in addition the transition density $p(x,a,y)$ is bounded and Assumption 7.13(d) holds, then φ is Blackwell optimal in the class Π as well.*

References

[1] E. Altman, A. Hordijk and L.C.M. Kallenberg, "On the value function in constrained control of Markov chains", *Mathematical Methods of Operations Research* **44**, 387–399, 1996.

[2] E.I. Balder, "On compactness of the space of policies in stochastic dynamic programming", *Stochastic Processes and Applications* **32**, 141–150, 1989.

[3] D. Blackwell, "Discrete dynamic programming", *Annals of Mathematical Statistics* **33**, 719–726, 1962.

[4] R. Cavazos-Cadena and J.B. Lasserre, "Strong 1-optimal stationary policies in denumerable Markov decision processes", *Systems and Control Letters* **11**, 65–71, 1988.

[5] R. Cavazos-Cadena and J.B. Lasserre, "A direct approach to Blackwell optimality", MORFISMOS **3**, no. 1, 9–33, 1999.

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

[7] R.Ya. Chitashvili, "A finite controlled Markov chain with small termination probability", *Theory Prob. Appl.* **21**, 158–163, 1976.

[8] R. Dekker and A. Hordijk, "Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards", *Mathematics of Operations Research* **13**, 395–421, 1988.

[9] R. Dekker and A. Hordijk. "Denumerable semi-Markov decision chains with small interest rates", *Annals Oper. Res.* **28**, 185–212, 1991.

[10] R. Dekker and A. Hordijk, "Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains", *Mathematics of Operations Research* **17**, 271–289, 1992.

[11] R. Dekker, A. Hordijk and F.M. Spieksma, "On the relation between recurrence and ergodicity properties in denumerable Markov decision chains", *Mathematics of Operations Research* **19**, 539–559, 1994.

[12] E.V. Denardo, "Markov renewal programming with small interest rates", *Annals Math. Stat.* **42**, 477–496, 1971.

[13] A. Federgruen, A. Hordijk and H.C. Tijms, "A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices", *Journal of Applied Probability* **15**, 842–847, 1978.

[14] A. Federgruen, A. Hordijk and H.C. Tijms, "Recurrence conditions in denumerable state Markov decision processes", in *Dynamic Programming and Its Applications*, ed. by M.L. Puterman, 3–22, Academic Press, 1978.

[15] J. Flynn, "Averaging vs. discounting in dynamic programming: a counterexample", *Annals of Statistics* **2**, 411–413, 1974.

[16] O. Hernández-Lerma, R. Montes-de-Oca and R. Cavazos-Cadena, "Recurrence conditions for Markov decision processes with Borel state space: a survey", *Annals of Operations Research* **28**, 29–46, 1991.

[17] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tract **51**, Mathematisch Centrum, 1974.

[18] A. Hordijk, "Regenerative Markov decision models", in *Mathematical Programming Study*, **6**, ed. by R.J.B. Wets, North Holland, 1976, 49–72.

[19] A. Hordijk, R. Dekker and L.C.M. Kallenberg, "Sensitivity-analysis in discounted Markovian decision problems", *Operations Research Spektrum* **7**, 143–151, 1985.

[20] A. Hordijk, O. Passchier and F.M. Spieksma, "On the existence of the Puisseux expansion of the discounted rewards: a counterexample", *Probability in the Engineering and Informational Sciences* **13**, 229–235, 1999.

[21] A. Hordijk and K. Sladký, "Sensitive optimality criteria in countable state dynamic programming", *Math. of Oper. Res.* **2**, 1–14, 1977.

[22] A. Hordijk and F.M. Spieksma, "Are limits of α-discounted optimal policies Blackwell optimal? A counterexample", *Systems and Control Letters*, **13**, 31–41, 1989.

[23] A. Hordijk and F.M. Spieksma, "On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network", *Advances in Applied Probability* **24**, 343–376, 1992.

[24] A. Hordijk, F.M. Spieksma and R.L. Tweedie, "Uniform stability conditions for general space Markov decision processes", Technical report, Leiden University and Colorado State University, 1995.

[25] A. Hordijk and A.A. Yushkevich, "Blackwell optimality in the class of stationary policies in Markov decision chains with a Borel state space and unbounded rewards", *Math. Methods Oper. Res.* **49**, 1–39, 1999.

[26] A. Hordijk and A.A. Yushkevich, "Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards", *Mathematical Methods of Operations Research* **50**, 421–428, 1999.

[27] R.A. Howard, *Dynamic Programming and Markov Processes*, Wiley, 1960.

[28] Y. Huang and A.F. Veinott Jr., "Markov branching decision chains with interest-rate-dependent rewards, *Probability in the Engineering and Informational Sciences* **9**, 99–121, 1995.

[29] J.G. Kemeny and J.L. Snell, *Finite Markov chains*, Van Nostrand-Reinhold, 1960.

[30] J.B. Lasserre, "Conditions for existence of average and Blackwell optimal stationary policies in denumerable Markov decision processes", *Journal of Mathematical Analysis and Applications* **136**, 479–490, 1988.

[31] A. Maitra, "Dynamic programming for countable state systems", *Sankhya Ser. A* **27**, 241–248, 1965.

[32] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, 1993.

[33] B.L. Miller and A.F. Veinott, "Discrete dynamic programming with a small interest rate", *Annals of Mathematical Statistics* **40**, 366–370, 1969.

[34] M.L. Puterman, *Markov Decision Processes*, Wiley, 1994.

[35] S.M. Ross, "Non-discounted denumerable Markovian decision models", *Annals of Mathematical Statistics* **39**, 412–423, 1968.

[36] M. Schäl, "On dynamic programming: Compactness of the space of policies", *Stochastic Processes and Applications* **3**, 345–354, 1975.

[37] M. Schäl, "On dynamic programming and statistical decision theory", *Annals of Statistics* **7**, 432–445, 1979.

[38] K. Sladký, "On the set of optimal controls for Markov chains with rewards", *Kybernetika* (Prague) **10**, 350–367, 1974.

[39] F.M. Spieksma, "Geometrically ergodic Markov chains and the optimal control of queues", Ph.D. Thesis, University of Leiden, 1990.

[40] L.C. Thomas, "Connectedness conditions for denumerable state Markov decision processes", in *Recent Developments in Markov Decision Processes*, ed. by R. Hartley, L. Thomas, D. White, Academic Press, 1980, 181–204.

[41] H.C. Tijms, "Average reward optimality equation in Markov decision processes with a general state space", in *Probability, Statistics and Optimization: a Tribute to Peter Whittle*, ed. by F.P. Kelly, Wiley, 1994, 485–495.

[42] A.F. Veinott Jr., "On finding optimal policies in discrete dynamic programming with no discounting", *Annals of Mathematical Statistics* **37**, 1284–1294, 1966.

[43] A.F. Veinott Jr., "Discrete dynamic programming with sensitive optimality criteria", *Annals of Mathematical Statistics* **40**, 1635–1660, 1969.

[44] A.F. Veinott Jr., *Dynamic Programming and Stochastic Control*, Unpublished class notes.

[45] A.F. Veinott Jr., "Markov decision chains", *Studies in Optimization*, G.B. Dantzig and B.C. Eaves editors, American Mathematical Association, Providence RI 1974, 124–159.

[46] H.M. Wagner, "On optimality of pure strategies", *Management Science* **6**, 268–269, 1960.

[47] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, 1972.

[48] K. Yosida, *Functional Analysis*, Springer, 1980.

[49] A.A. Yushkevich, "Blackwell optimal policies in a Markov decision process with a Borel state space", *Mathematical Methods of Operations Research* **40**, 253–288, 1994.

[50] A.A. Yushkevich, "Strong 0-discount optimal policies in a Markov decision process with a Borel state space", *Mathematical Methods of Operations Research* **42**, 93–108, 1995.

[51] A.A. Yushkevich, "A note on asymptotics of discounted value function and strong 0-discount optimality", *Mathematical Methods of Operations Research* **44**, 223–231, 1996.

[52] A.A. Yushkevich, "Blackwell optimal policies in countable dynamic programming without aperiodicity assumptions", in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, ed. by T.S. Ferguson, L.S. Shapley and J.B. MacQueen, Inst. of Math. Stat., 1996, 401–407.

[53] A.A. Yushkevich, "Blackwell optimality in Markov decision processes with a Borel state space", *Proceedings of 36th IEEE Conference on Decision and Control* **3**, 2827–2830, 1997.

[54] A.A. Yushkevich, "The compactness of a policy space in dynamic programming via an extension theorem for Carathéodory functions", *Mathematics of Operations Research* **22**, 458–467, 1997.

[55] A.A. Yushkevich, "Blackwell optimality in Borelian continuous-in-action Markov decision processes", *SIAM Journal on Control and Optimization* **35**, 2157–2182, 1997.

[56] A.A. Yushkevich and R.Ya. Chitashvili, "Controlled random sequences and Markov chains", *Russian Mathematical Surveys* **37**, 239–274, 1982.

Arie Hordijk
University of Leiden
Mathematical Institute
P.O. Box 9512
2300 RA Leiden, The Netherlands
hordijk@math.leidenuniv.nl

Alexander A. Yushkevich
University of North Carolina at Charlotte
Department of Mathematics
Charlotte, NC 28223, USA
aayushke@email.uncc.edu

8 THE POISSON EQUATION FOR COUNTABLE MARKOV CHAINS: PROBABILISTIC METHODS AND INTERPRETATIONS

Armand M. Makowski
Adam Shwartz

Abstract: This paper considers the Poisson equation associated with time-homogeneous Markov chains on a countable state space. The discussion emphasizes probabilistic arguments and focuses on three separate issues, namely (i) the existence and uniqueness of solutions to the Poisson equation, (ii) growth estimates and bounds on these solutions and (iii) their parametric dependence. Answers to these questions are obtained under a variety of recurrence conditions.

Motivating applications can be found in the theory of Markov decision processes in both its adaptive and non-adaptive formulations, and in the theory of Stochastic Approximations. The results complement available results from Potential Theory for Markov chains, and are therefore of independent interest.

8.1 INTRODUCTION

Let $P \equiv (p_{xy})$ be the one-step transition matrix for a time-homogeneous Markov chain $\{X_t,\ t = 0, 1, \dots\}$ taking values in some countable space $\mathbb{X}$. This paper is devoted to the corresponding *Poisson* equation with forcing function $r : \mathbb{X} \to \mathbb{R}$, namely

$$h(x) + w = r(x) + \sum\nolimits_y p_{xy} h(y), \quad x \in \mathbb{X} \tag{8.1}$$

for scalar w and mapping $h : \mathbb{X} \to \mathbb{R}$. This equation arises naturally in a variety of problems associated with Markov chains as the following examples indicate.

1. As shown in Section 8.3, solving the Poisson equation provides a means to evaluate the long-run average cost w associated with the cost function r [36]: If (8.1) has a solution (h, w) and some mild growth conditions are satisfied, then Lemma 8.2 states that

$$w = \lim_t \mathbb{E}_\mu \left[\frac{1}{t+1} \sum_{s=0}^{t} r(X_s) \right] \tag{8.2}$$

where μ is the initial distribution and $\mathbb{E}_\mu$ is the corresponding expectation operator. The function h measures the sensitivity of the cost to the initial state, and represents a second-order effect captured through the "deviation matrix" [13]. This function h can also serve as a "Lyapunov function" in establishing ergodicity [26], and plays a key role in proving the convergence of the policy improvement algorithm [26]. Approximate solutions can be used for simulations—see Chapter 9.

2. During the last decade there has been widespread interest in stochastic approximation algorithms as a means to solve increasingly complex engineering problems [1, 5, 16, 17]. As a result, focus has shifted from the original Robbins-Monro algorithm to (projected) stochastic approximations driven by Markovian "noise" or "state" processes. Properties of solutions to an appropriate Poisson equation play an essential role when establishing the a.s. convergence of such adaptive algorithms [1, 18, 22, 24, 25, 39].

3. In the context of Markov decision processes (MDPs), the need for adaptive policies can arise in response to both modeling uncertainties and computational limitations [40]. Several adaptive policies have been proposed as "implementations" to a Markov stationary policy, and shown to yield the same cost performance [3, 18, 19, 23, 40]. Here too, the analysis requires precise information on the solution to the Poisson equation associated with the non-adaptive policy [40].

In many of these applications, it is natural to view the forcing function r and the transition matrix P as parameterized, say by some parameter θ (which may be loosely interpreted as a control variable). The requisite analysis then typically exploits smoothness properties (in θ) of the solution h together with various growth estimates (in x) for h. In addition, estimates on the moments of $\{h(X_t),\ t = 0, 1, 2, \dots\}$ are required, with the added difficulty that the resulting process $\{X_t,\ t = 0, 1, 2, \dots\}$ is not necessarily Markovian (say, under the given stochastic approximation scheme or adaptive policy).

Our main objective is to develop methods for addressing the concerns above in a systematic fashion. Whenever possible, we emphasize a probabilistic viewpoint as we focus mostly on the following three issues:

1. Existence and uniqueness of solutions to the Poisson equation (8.1);

2. Growth estimates and bounds on these solutions; and

3. Conditions for smoothness in the parameter of these solutions when dealing with the parametric case, as would arise when establishing the a.s.

convergence of stochastic approximations and the self-tuning property of adaptive policies.

Answers to these questions are given under a variety of recurrence conditions. As we try to keep the exposition relatively self-contained, we have included some standard material on the Poisson equation. In addition to its tutorial merit, the discussion given here provides a unified treatment to many of the issues associated with the Poisson equation, e.g. existence, uniqueness and representation of solutions. This is achieved by manipulating a single *martingale* naturally induced by the Poisson equation.

Questions of existence and uniqueness of solutions to (8.1) have obvious and natural points of contact with the Potential Theory for Markov chains [15, 29]. Unfortunately many situations of interest in applications, say in the context of MDPs, are not readily covered by classical Potential Theory. Indeed, the classical theory treats the purely transient and recurrent cases separately, with drastically different results for each situation. This approach is thus of limited use in the above-mentioned situations, where the recurrence structure of the Markov chain is typically far more complex in that it combines both transient and recurrent states. Here, in contrast with the analytical approach of classical Potential Theory, emphasis has been put on giving an explicit representation of the solution to (8.1) with a clear probabilistic interpretation.

This probabilistic approach allows for a relatively elementary treatment of questions of existence and uniqueness, under a rather general recurrence structure. We accomplish this by focusing on the discrete space case, and by keeping the assumptions as transparent as possible. The intuition developed here applies to the general state-space case, under mild conditions on the existence of petite sets—see Chapter 9 and [10, 26, 27]. Results are obtained in various degrees of completeness for both finite and countably infinite state spaces; recurrence structures include multiple positive recurrent classes, and transient classes. A representation for h is derived in detail in the case of a single positive recurrent class under integrability conditions involving the forcing function r. The derivation uses elementary methods, and provides intuition into more general situations. This representation is shown to also hold in countable case with multiple classes, and readily lends itself to establishing natural bounds on the growth rate of h (as a function of the state), and to investigating smoothness properties in the parameterized problem.

Similar results are given in [10] for the ergodic case on general state spaces. In addition, when the forcing function r is positive and "increasing" (i.e. when its sub-level sets are compact), there is an elegant theory that relates geometric ergodicity to the Poisson equation; details and references can be found in Chapter 9. As evidenced by the references section, there is a very large literature on the Poisson equation; of particular note is the monograph by Nummelin [28]. In the context of MDPs, bounded solutions are discussed by Ross [34, 35], Gubenko and Shtatland [11] and Yushkevich [42]. One of the first treatments of unbounded solutions is available in Robinson [31, 32] (with details in Chapter 4).

The paper is organized as follows: The set-up is given in Section 8.2 together with the basic martingale associated with (8.1). Various uniqueness results on

the the solution (w, h) are discussed in Section 8.3. We give two decomposition results in Section 8.4; the first is based on the decomposition of the state space $\mathbb{X}$ into its recurrent and transient classes, while the second is an analog of the standard Green decomposition and relies on an expansion of the forcing function in terms of more "elementary" functions. To set the stage for the countably infinite case, we briefly recall an algebraic treatment of the finite-state case in Section 8.5. In Section 8.6, under a single positive recurrent class assumption, an explicit representation for the solution is developed in terms of the recurrence time to some distinguished state. An example is developed in Section 8.7 to illustrate the material of previous sections. Bounds and extensions to unbounded forcing functions and multiple recurrent classes are given in Section 8.8. Equipped with this probabilistic representation of solutions, we can now investigate the smoothness properties of solutions to the parameterized problem; methods for proving continuity and Lipschitz continuity are developed in Sections 8.9 and 8.10, respectively.

To close, we note that most of the ideas which are discussed here in the context of countable Markov chains have extensions to fairly general state spaces. This is achieved by means of the so-called *splitting technique* [10, 26, 27, 28] which in essence guarantees the existence of an atom on an enlarged state space; details can be found in Chapter 9.

8.2 THE POISSON EQUATION AND ITS ASSOCIATED MARTINGALE

First, a few words on the notation used throughout the paper: The set of all real numbers is denoted by $\mathbb{R}$ and $\mathbf{1}[A]$ stands for the indicator function of a set A. Unless otherwise stated, $\lim_t$, $\underline{\lim}_t$ and $\overline{\lim}_t$ are taken with t going to infinity. Moreover, the infimum over an empty set is taken to be ∞ by convention. The Kronecker mapping $\delta : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is defined by $\delta(x, y) = 1$ if $x = y$, and $\delta(x, y) = 0$ otherwise. Finally, the notation $\sum_{x \in \mathbb{X}}$ is often abbreviated as $\sum_x$.

8.2.1 The set-up

The notion of a Markov chain we adopt in this paper is more general than the elementary one used in most applications. We do so with the view of broadening the applicability of the material developed here, especially to problems of adaptive control for Markov chains [18, 19, 22, 23, 39, 40].

The state space is a countable, and we assume the existence of a measurable space $(\Omega, \mathcal{F})$ large enough to carry all the probabilistic elements considered in this paper. In particular, let $\{\mathcal{F}_t,\ t = 0, 1, \dots\}$ denote a filtration of $\mathcal{F}$, i.e. a monotone increasing sequence of σ-fields contained in $\mathcal{F}$ such that $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ for all $t = 0, 1, \dots$, and let $\{X_t,\ t = 0, 1, \dots\}$ be a sequence of $\mathbb{X}$-valued rvs which are $\mathcal{F}_t$-adapted, i.e. the rv X_t is $\mathcal{F}_t$-measurable for all $t = 0, 1, \dots$.

The Markovian structure of interest is defined by postulating the existence of a family $\{\mathbb{P}_x,\ x \in \mathbb{X}\}$ of probability measures on $\mathcal{F}$ such that for all x and

y in $\mathbb{X}$, we have

$$\mathbb{P}_x[X_0 = y] = \delta(x, y) \tag{8.3}$$

and

$$\mathbb{P}_x[X_{t+1} = y \mid \mathbb{F}_t] = p_{X_t y} \quad \mathbb{P}_x - a.s.\ t = 0, 1, \ldots \tag{8.4}$$

This set up is more general than the standard one as we allow $\sigma\{X_s : 0 \leq s \leq t\}$ to be a strict subset of $\mathbb{F}_t$. However, in many instances, we do take $\mathcal{F}_t$ to be the σ-field generated by the rvs $X_0, \ldots, X_t$ for all $t = 0, 1, \ldots$, in which case the definition above coincides with the elementary definition of a Markov chain.

With any probability distribution μ on $\mathbb{X}$, we associate a probability measure $\mathbb{P}_\mu$ on $\mathcal{F}$ by setting

$$\mathbb{P}_\mu[A] := \sum\nolimits_x \mu(x)\mathbb{P}_x[A], \quad A \in \mathcal{F}. \tag{8.5}$$

Obviously, when μ is the Dirac measure δ_x concentrated at some x in $\mathbb{X}$, then $\mathbb{P}_\mu$ reduces to $\mathbb{P}_x$. Using (8.3)–(8.5) we easily see that

$$\mathbb{P}_\mu[X_0 = x] \quad = \quad \mu(x), \quad x \in \mathbb{X} \tag{8.6}$$

and

$$\mathbb{P}_\mu[X_{t+1} = y \mid \mathbb{F}_t] \quad = \quad p_{X_t y}, \quad y \in \mathbb{X} \quad \mathbb{P}_\mu - a.s.\ t = 0, 1, \ldots \tag{8.7}$$

Hence, under the probability measure $\mathbb{P}_\mu$ the rvs $\{X_t,\ t = 0, 1, \ldots\}$ have the Markov property with respect to the filtration $\{\mathcal{F}_t,\ t = 0, 1, \ldots\}$, and are said to form a time-homogeneous $\mathcal{F}_t$-Markov chain with one-step transition matrix P and initial probability distribution μ.

From (8.3)–(8.5) we readily conclude that

$$\mathbb{P}_\mu[A | X_0 = x] = \mathbb{P}_x[A], \quad A \in \mathcal{F} \tag{8.8}$$

whenever $\mu(\{x\}) > 0$, and $\mathbb{P}_x$ has the useful interpretation of conditional probability (under $\mathbb{P}_\mu$ for any initial distribution measure μ).

Throughout it will be convenient to denote by $\mathbb{E}_\mu$ and $\mathbb{E}_x$ the expectation operator associated with $\mathbb{P}_\mu$ and $\mathbb{P}_x$, respectively.

8.2.2 The Poisson equation

Let r be a given mapping $\mathbb{X} \to \mathbb{R}$. Throughout, it is understood that a constant w and a mapping $h : \mathbb{X} \to \mathbb{R}$ constitute a solution pair to the Poisson equation (8.1) with forcing function r whenever h satisfies the integrability conditions

$$\sum\nolimits_y p_{xy}|h(y)| \ < \ \infty, \quad x \in \mathbb{X} \tag{8.9}$$

and the relations

$$h(x) + w = r(x) + \sum\nolimits_y p_{xy} h(y), \quad x \in \mathbb{X} \tag{8.10}$$

hold. The Poisson equation is termed *homogeneous* if $r \equiv 0$.

For any initial distribution μ, we introduce several classes of $\mathbb{R}$-valued mappings defined on $\mathbb{X}$. The mapping $f : \mathbb{X} \to \mathbb{R}$ is said to be an element of

$\mathcal{I}_\mu$ if $\mathbb{E}_\mu[|f(X_t)|] < \infty$ for all $t = 0, 1, \ldots$;

$\mathcal{B}_\mu$ if $\sup_t \mathbb{E}_\mu[|f(X_t)|] < \infty$;

$\mathcal{S}_\mu$ if f belongs to $\mathcal{I}_\mu$ with $\lim_t \frac{1}{t}\mathbb{E}_\mu[f(X_t)] = 0$; and

$\mathcal{U}_\mu$ if the rvs $\{f(X_t),\ t = 0, 1, \ldots\}$ are uniformly integrable under $\mathbb{P}_\mu$.

When μ is the Dirac measure δ_x for some x in $\mathbb{X}$, we substitute the simpler notation $\mathcal{I}_x$, $\mathcal{B}_x$, $\mathcal{S}_x$ and $\mathcal{U}_x$ to $\mathcal{I}_{\delta_x}$, $\mathcal{B}_{\delta_x}$, $\mathcal{S}_{\delta_x}$ and $\mathcal{U}_{\delta_x}$, respectively.

For any initial distribution μ, it holds that

$$\mathcal{U}_\mu \subset \mathcal{B}_\mu \subset \mathcal{S}_\mu \subset \mathcal{I}_\mu, \tag{8.11}$$

and for any x in $\mathbb{X}$ such that $\mu(x) > 0$, we have $\mathcal{I}_\mu \subset \mathcal{I}_x$, $\mathcal{B}_\mu \subset \mathcal{B}_x$ and $\mathcal{U}_\mu \subset \mathcal{U}_x$.

Since any mapping $f : \mathbb{X} \to \mathbb{R}$ can be viewed as a *column* vector $(f(x),\ x \in \mathbb{X})$, the Poisson equation (8.1) can be written in matrix notation as

$$h + we = r + Ph \tag{8.12}$$

where e denotes the column vector with all its entries equal to one, i.e. $e(x) = 1$ for all x in $\mathbb{X}$. For any vector $f = (f(x),\ x \in \mathbb{X})$ and any subset E of $\mathbb{X}$, denote by f_E the restriction $(f(x),\ x \in E)$ of f to E and similarly define P_E as the restriction $(p_{xy},\ x, y \in E)$ of P to E. The identity matrix on $\mathbb{X}$ is denoted by I.

8.2.3 A martingale property

Many of the general results on solutions to the Poisson equation can be traced back to the following observation.

Lemma 8.1 *Let the pair (h, w) be a solution to the Poisson equation (8.9)–(8.10) with forcing function r. If the mapping h belongs to $\mathcal{I}_\mu$ for some probability measure μ on $\mathbb{X}$, then the following statements hold:*

1. *The forcing function r is necessarily an element of $\mathcal{I}_\mu$; and*

2. *The rvs $\{M_t,\ t = 0, 1, \ldots\}$ defined by $M_0 := h(X_0)$ and*

$$M_{t+1} := h(X_{t+1}) + \sum_{s=0}^{t} r(X_s) - (t+1)w, \quad t = 0, 1, \ldots \tag{8.13}$$

form an integrable $(\mathbb{P}_\mu, \mathcal{F}_t)$-martingale sequence.

Proof. Invoking the Markov property, we can reformulate the Poisson equation (8.9)–(8.10) as

$$h(X_t) + w = r(X_t) + \mathbb{E}_\mu[h(X_{t+1})|\mathcal{F}_t], \quad t = 0, 1, \dots \tag{8.14}$$

and the $\mathbb{P}_\mu$-integrability of the rvs $\{r(X_t),\ t = 0, 1, \dots\}$ follows from the assumption on h. This proves Claim 1.

To establish Claim 2, we first conclude from the first part of the proof that the rvs $\{M_t,\ t = 0, 1, \dots\}$ are well defined and indeed $\mathbb{P}_\mu$-integrable. From (8.13), we then get

$$\mathbb{E}_\mu[M_{t+1}|\mathcal{F}_t] = \mathbb{E}_\mu[h(X_{t+1})|\mathcal{F}_t] + \sum_{s=0}^{t} r(X_s) - (t+1)w, \quad t = 0, 1, \dots$$

because the rvs $X_0, \dots, X_t$ are all $\mathcal{F}_t$-measurable, and the martingale property follows from (8.14). ∎

8.3 UNIQUENESS RESULTS

In this section, we have collected several uniqueness results for the Poisson equation (8.9)–(8.10). In that respect, we first note that if the pair (h, w) is a solution to the Poisson equation, so is the pair $(h + \alpha e, w)$ for any constant α. In other words, uniqueness can only be obtained up to an additive constant. We also observe that for r in $\mathcal{I}_\mu$, the definition

$$w(\mu) := \overline{\lim}_t\ \mathbb{E}_\mu \left[\frac{1}{t+1} \sum_{s=0}^{t} r(X_s)\right] \tag{8.15}$$

is well posed. The next lemma is a version of a standard result from the theory of MDPs under a long-run average criterion [12, 36], [40, Lemma 3.1].

Lemma 8.2 *Let the pair (h, w) be a solution to the Poisson equation (8.9)–(8.10) with forcing function r. If the mapping h belongs to $\mathcal{S}_\mu$ for some probability measure μ on $\mathbb{X}$, then*

$$w = w(\mu) = \lim_t\ \mathbb{E}_\mu \left[\frac{1}{t+1} \sum_{s=0}^{t} r(X_s)\right] . \tag{8.16}$$

Proof. Since h is an element of $\mathcal{S}_\mu$, it is also an element of $\mathcal{I}_\mu$ by virtue of (8.11). By Claim 2 of Lemma 8.1 we readily obtain the equalities $\mathbb{E}_\mu[M_0] = \mathbb{E}_\mu[M_{t+1}]$ for all $t = 0, 1, \dots$ or, equivalently, in expanded form,

$$\mathbb{E}_\mu[h(X_0)] = \mathbb{E}_\mu[h(X_{t+1})] + \mathbb{E}_\mu \left[\sum_{s=0}^{t} r(X_s)\right] - (t+1)w. \tag{8.17}$$

Simple rearrangements yield

$$\mathbb{E}_\mu \left[\frac{1}{t+1} \sum_{s=0}^{t} r(X_s)\right] = w - \frac{1}{t+1} \left\{\mathbb{E}_\mu[h(X_{t+1})] - \mathbb{E}_\mu[h(X_0)]\right\}, \tag{8.18}$$

and the result (8.16) is now immediate upon letting $t \uparrow \infty$ in (8.18) since h is an element of $\mathcal{S}_\mu$. ∎

If the Poisson equation (8.9)–(8.10) admits a solution (h, w) with h *bounded*, then r is necessarily bounded, so that both r and h belong to $\mathcal{U}_\mu$ (thus $\mathcal{S}_\mu$) for *any* initial distribution μ. It then follows from Lemma 8.2 that $w(\mu)$ is obtained as a limit which does *not* depend on the initial distribution μ.

The uniqueness of solutions to the Poisson equation is now briefly studied in the class of "uniformly L_1-bounded" solutions, i.e. solutions in $\mathcal{B}_\mu$ for some initial state distribution μ. If the state space contains a set I of isolated states which are not reachable from $\mathbb{X}\backslash I$ and if $\mu(I) = 0$, then clearly the chain never visits the states in I. To simplify the exposition we find it convenient to reformulate the problem on the reduced state space $\mathbb{X}\backslash I$.

The next lemma is preparatory in nature and will greatly simplify the presentation: For (h_1, w_1) and (h_2, w_2) solution pairs to the Poisson equation (8.9)–(8.10), we define

$$\Delta w := w_1 - w_2 \quad \text{and} \quad \Delta h(x) := h_1(x) - h_2(x), \quad x \in \mathbb{X}. \tag{8.19}$$

Lemma 8.3 *Let (h_1, w_1) and (h_2, w_2) be two solutions of the Poisson equation (8.9)–(8.10). If Δh belongs to $\mathcal{I}_\mu$ for some probability measure μ on $\mathbb{X}$, then the rvs $\{\Delta h(X_t) - t \cdot \Delta w,\ t = 0, 1, \dots\}$ form a $(\mathbb{P}_\mu, \mathcal{F}_t)$-martingale sequence with*

$$\Delta w = \frac{1}{s}\Big(\mathbb{E}_\mu[\Delta h(X_{t+s})] - \mathbb{E}_\mu[\Delta h(X_t)]\Big), \quad t = 0, 1, \dots ; s = 1, 2, \dots \tag{8.20}$$

Proof. Denoting by $\{M_t^i,\ t = 0, 1, \dots\}$ the rvs (8.13) associated with the solution pair (h_i, w_i), $i = 1, 2$, we define the rvs $\{\Delta M_t,\ t = 0, 1, \dots\}$ by

$$\Delta M_t := M_t^1 - M_t^2 = \Delta h(X_t) - t \cdot \Delta w, \quad t = 0, 1, \dots$$

It is plain that $(\Delta h, \Delta w)$ is a solution to the homogeneous Poisson equation $\Delta h + \Delta w e = P\Delta h$. Applying Lemma 8.1 to this Poisson equation, we conclude that the rvs $\{\Delta M_t,\ t = 0, 1, \dots\}$ indeed form an integrable $(\mathbb{P}_\mu, \mathcal{F}_t)$-martingale sequence, whence $\mathbb{E}_\mu[\Delta M_{t+s}] = \mathbb{E}_\mu[\Delta M_t]$ for all $s, t = 0, 1, \dots$. In expanded form, these equalities become

$$\mathbb{E}_\mu[\Delta h(X_{t+s})] - (t+s)\Delta w = \mathbb{E}_\mu[\Delta h(X_t)] - t\Delta w, \quad t = 0, 1, \dots$$

and we obtain (8.20) after simple rearrangements. ∎

The basic uniqueness result can now be developed.

Theorem 8.1 *Let (h_1, w_1) and (h_2, w_2) be two solutions of the Poisson equation (8.9)–(8.10).*

1. *If Δh belongs to $\mathcal{S}_\mu$ for some probability measure μ on $\mathbb{X}$, then $w_1 = w_2$;*
2. *If in addition Δh is an element of $\mathcal{B}_\mu$, then Δh is constant on each recurrent class of the Markov chain $\mathbb{P}_\mu$.*

Proof. If Δh belongs to $\mathcal{S}_\mu$, then its is also an element of $\mathcal{I}_\mu$, and Claim 1 follows by letting $s \uparrow \infty$ in (8.20) and using the fact that Δh belongs to $\mathcal{S}_\mu$.

The proof of Claim 2 starts with the observation (8.11) made earlier that since Δh is an element of $\mathcal{B}_\mu$, it it is also an element of $\mathcal{S}_\mu$. Therefore, $w_1 = w_2$ by Claim 1 and the rvs $\{\Delta h(X_t),\ t = 0, 1, \dots\}$ form a $(\mathbb{P}_\mu, \mathcal{F}_t)$-martingale sequence with $\sup_t \mathbb{E}_\mu[|\Delta h(X_t)|] < \infty$. By a standard martingale convergence theorem [7, 14], the martingale sequence $\{\Delta h(X_t),\ t = 0, 1, \dots\}$ converges $\mathbb{P}_\mu$-a.s. to a proper rv.

If $\mathbb{X}$ constitutes a single recurrent class under P, then any two states in $\mathbb{X}$, say x and y, are visited infinitely often $\mathbb{P}_\mu$-a.s. It is now plain that $\Delta h(x) = \Delta h(y)$ by virtue of the $\mathbb{P}_\mu$-a.s. convergence of the martingale $\{\Delta h(X_t),\ t = 0, 1, \dots\}$, and we conclude that Δh is constant on $\mathbb{X}$.

More generally, let R be a recurrence class under P, i.e. a closed irreducible set of recurrent states. Since $p_{xy} = 0$ for all x in R and y not in R, (8.9)–(8.10) implies

$$h_{i,R} + we_R = r_R + P_R h_{i,R}, \quad i = 1, 2. \tag{8.21}$$

The matrix P_R can be interpreted as the matrix of one-step transition probabilities for an irreducible Markov chain on R with all its states recurrent, and the problem is now reduced to the previously considered case. Therefore, $h_{1,R} - h_{2,R}$ is constant on R and the proof is complete. ∎

When h_1 and h_2 belong to $\mathcal{U}_\mu$, the Ergodic Theorem can be used to derive this uniqueness result along the lines of [10, Proposition 1.1]. Under conditions weaker than the ones assumed in Theorem 8.1 we can obtain the following refinement of Claim 1 of Theorem 8.1.

Corollary 8.1 *Let (h_1, w_1) and (h_2, w_2) be two solutions to the Poisson equation (8.9)–(8.10). If for some probability measure μ on $\mathbb{X}$, h_1 belongs to $\mathcal{S}_\mu$ and h_2 belongs to $\mathcal{I}_\mu$, then*

$$\lim_t \frac{1}{t} \mathbb{E}_\mu [h_2(X_t)] = w_2 - w_1 \ . \tag{8.22}$$

Proof. First we note that if h_1 is an element of $\mathcal{S}_\mu$ and if h_2 belongs to $\mathcal{I}_\mu$, then Δh belongs to $\mathcal{I}_\mu$. By Lemma 8.3 we get

$$\Delta w = \frac{1}{t} \{\mathbb{E}_\mu[\Delta h(X_t)] - \mathbb{E}_\mu[\Delta h(X_0)]\}, \quad t = 1, 2, \dots \tag{8.23}$$

and (8.22) follows upon letting $t \uparrow \infty$ in (8.23) and using the fact that h_1 is an element of $\mathcal{S}_\mu$. The existence of the limit is a consequence of the equalities (8.23) ∎

It is very easy to demonstrate the non-uniqueness of solutions to the Poisson equation: Consider the Markov chain on the non-negative integers $\mathbb{N}$ with $p_{x,x+1} = 1$, $x = 0, 1, \dots$, and let $r \equiv 0$. Then $(h_2, w_2) \equiv (0, 0)$ is obviously a solution to the homogeneous Poisson equation with $h_1(0) = 0$. However, the pair $(h_2, w_2) \equiv (x, 1)$ is also a solution to this Poisson equation with $h_2(0) = 0$. For all $t = 0, 1, \dots$, we have $X_t = X_0 + t$ $\mathbb{P}_\mu$-a.s, whence $\mathbb{E}_\mu[h_2(X_t)] =$

$\mathbb{E}_\mu[X_0] + t$, and under the condition $\mathbb{E}_\mu[X_0] < \infty$, h_2 is an element of $\mathcal{I}_\mu$, but not of $\mathcal{S}_\mu$. In fact, (8.22) holds as $\lim_t \frac{1}{t}\mathbb{E}_\mu[h_2(X_t)] = 1 \neq 0$. In Section 8.7 we discuss the non-uniqueness issue for a more elaborate example of a positive recurrent system.

Although in practice it might be hard to verify the L_1-boundedness conditions of Theorem 8.1, a simple characterization of the set $\mathcal{B}_\mu$ is available in a special yet important case. Recall that a probability measure γ on $\mathcal{B}(\mathbb{X})$ is an *invariant* measure for the one-step transition matrix P if

$$\gamma(x) = \sum\nolimits_y \gamma(y) p_{yx}, \quad x \in \mathbb{X}. \tag{8.24}$$

Under $\mathbb{P}_\gamma$ the Markov chain $\{X_t,\ t = 0, 1, \dots\}$ forms a strictly stationary sequence with one-dimensional marginal distribution γ, so that the following characterization is immediate.

Lemma 8.4 *If γ is an invariant probability measure for the one-step transition matrix P, then $\mathcal{I}_\gamma = \mathcal{B}_\gamma = \mathcal{U}_\gamma = L_1(\mathbb{X}, \mathcal{B}(\mathbb{X}), \gamma)$.*

Derman and Veinott [8] consider the uniqueness issue for Markov chains with a single positive recurrent class (in which case the invariant measure γ is unique). They show uniqueness in the class DV of mappings $f : \mathbb{X} \to \mathbb{R}$ characterized by

$$\mathbb{E}_x\left[\sum_{t=0}^{T-1} |f(X_t)|\right] < \infty, \quad x \in \mathbb{X} \tag{8.25}$$

where $T := \inf\{t > 0 : X_t = z\}$ for some distinguished recurrent state z. Under (8.25) we conclude by standard results on Markov chains [6, 27] that

$$\mathbb{E}_\gamma[|f(X_t)|] = \frac{\mathbb{E}_z\left[\sum_{t=0}^{T-1} |f(X_t)|\right]}{\mathbb{E}_z[T]}, \quad t = 0, 1, \dots \tag{8.26}$$

and DV is seen to coincide with $\mathcal{B}_\gamma$. Note that when (8.25) holds for some mapping $f : \mathbb{X} \to \mathbb{R}$, the Markov chain is said to be f-regular [27].

8.4 DECOMPOSITION RESULTS

8.4.1 A state decomposition result

With the uniqueness result of Theorem 8.1 in mind, we consider the decomposition of the countable set $\mathbb{X}$ induced by the recurrence structure of P: Let Tr denote the (possibly empty) set of transient states, and let $\{R_\alpha,\ \alpha \in A\}$, for some countable index set A, denote the recurrent components. The sets $\{Tr, R_\alpha,\ \alpha \in \mathrm{A}\}$ form a partition of $\mathbb{X}$. Moreover, for all α in A, $p_{xy} = 0$ for x in R_α and y not in R_α, and the restriction P_α of P to the recurrent class R_α is irreducible and recurrent on it. With the vector notation of Section 8.2, the Poisson equation (8.9)–(8.10) can now be partitioned as

$$h_{R_\alpha} + we_{R_\alpha} \quad = \quad r_{R_\alpha} + P_\alpha h_{R\alpha}, \quad \alpha \in \mathrm{A} \tag{8.27}$$

and

$$h_{Tr} + we_{Tr} = r_{Tr} + \sum\nolimits_{\alpha\in\mathrm{A}} T_\alpha h_{R_\alpha} + P_{Tr}h_{Tr} \tag{8.28}$$

where the matrices $\{T_\alpha,\ \alpha \in \mathrm{A}\}$ and P_{Tr} are determined from the decomposition of P associated with the sets $\{Tr, R_\alpha,\ \alpha \in A\}$.

The decomposition (8.27)–(8.28) of the Poisson equation motivates introducing the following family of Poisson equations

$$h_\alpha + w_\alpha e_{R_\alpha} = r_{R_\alpha} + P_\alpha h_\alpha, \quad \alpha \in \mathrm{A} \tag{8.29}$$

and

$$\tilde{h} + \tilde{w}e_{Tr} = \tilde{r} + P_{Tr}\tilde{h} \tag{8.30}$$

where for each α in A, r_{R_α} and h_α are mappings $R_\alpha \to \mathbb{R}$, while $\tilde{r}$ and $\tilde{h}$ are mappings $Tr \to \mathbb{R}$, with

$$\tilde{r} = r_{Tr} + \sum\nolimits_{\alpha\in\mathrm{A}} T_\alpha h_\alpha. \tag{8.31}$$

The next result shows in what sense the solutions to the projected Poisson equations (8.29)–(8.30) determine the solution to the original equation (8.9)–(8.10). The proof is a simple consequence of (8.27)–(8.28) and of (8.29)–(8.30), and is omitted in the interest of brevity.

Theorem 8.2 *The Poisson equation (8.9)–(8.10) has a solution if and only if the following two conditions hold:*

1. *For each α in A, the Poisson equation (8.29) on R_α has a solution (h_α, w_α) such that $w_\alpha = w$ for some scalar w independent of α and*

$$\sum\nolimits_{\alpha\in\mathrm{A}} T_\alpha |h_\alpha| < \infty; \tag{8.32}$$

2. *The Poisson equation (8.30), with forcing function $\tilde{r}$ given by (8.31) has a solution $(\tilde{h}, \tilde{w})$ such that $\tilde{w} = w$.*

A solution pair to (8.9)–(8.10) is necessarily of the form (h, w) with h determined by $h_{R_\alpha} = h_\alpha$ for all α in A and $h_{Tr} = \tilde{h}$.

Condition (8.32), which is automatically satisfied when $\mathbb{X}$ is finite, guarantees that $\tilde{r}$ (and therefore (8.30)) is well defined.

8.4.2 A Green-like decomposition

Let (h_1, w_1) and (h_2, w_2) be two solutions of the Poisson equation (8.9)–(8.10) with forcing functions r_1 and r_2, respectively. Then for any β in $\mathbb{R}$, $(h, w) := (\beta h_1 + h_2, \beta w_1 + w_2)$ is a solution to the Poisson equation (8.9)–(8.10) with forcing function $\beta r_1 + r_2$. Indeed, by definition, for all x in $\mathbb{X}$, we have

$$\begin{aligned} h(x) + w &= \beta(h_1(x) + w_1) + (h_2(x) + w_2) \\ &= \beta\left(r_1(x) + \sum\nolimits_y p_{xy}h_1(y)\right) + \left(r_2(x) + \sum\nolimits_y p_{xy}h_2(y)\right) \\ &= (\beta r_1(x) + r_2(x)) + \sum\nolimits_y p_{xy}(\beta h_1(y) + h_2(y)), \end{aligned} \tag{8.33}$$

where the last sum is well defined owing to the definition (8.10) of a solution.

This simple fact can be used as follows: For each v in $\mathbb{X}$, define the function $r_v : \mathbb{X} \to \mathbb{R}$ by $r_v(x) := \delta(v; x)$ for all x in $\mathbb{X}$, and let (h_v, w_v) denote a solution to the Poisson equation with forcing function r_v. The obvious decomposition

$$r(x) = \sum\nolimits_v r(v) r_v(x), \quad x \in \mathbb{X} \tag{8.34}$$

then leads naturally to the formal representation

$$w = \sum\nolimits_v r(v) w_v \quad \text{and} \quad h(x) = \sum\nolimits_v r(v) h_v(x), \quad x \in \mathbb{X}. \tag{8.35}$$

It remains then to check that (8.35) indeed defines a legitimate solution. In view of (8.33), this is the case whenever r is constant except at a finite number of points. In the more general case, this check can be done through the constructive arguments of Corollary 8.2, or through the verification result of Theorem 8.5. Such a calculation is performed directly in Section 8.7.

8.5 FINITE STATE SPACES

A complete picture of the solution to the Poisson equation (8.9)–(8.10) is available when $\mathbb{X}$ is a *finite* set, and can be found in [2, 9, 41]. In the finite space case any solution necessarily belongs to $\mathcal{U}_\mu$ for every initial probability distribution μ. Let $P^\star$ denote the stochastic matrix defined by

$$P^\star := \lim_t \frac{1}{t+1} \sum_{s=0}^{t} P^s; \tag{8.36}$$

its existence is guaranteed by classical results from the theory of Markov chains [2, 41]. The matrix $I - P + P^\star$ being invertible, the definition

$$h := (I - P + P^\star)^{-1} (I - P^\star) r \tag{8.37}$$

is well posed, and the easy identities $P^\star P = P P^\star = P^\star P^\star = P^\star$ lead after some simple algebra to the relation

$$h + P^\star r = r + P h. \tag{8.38}$$

A simple comparison of (8.38) with (8.9)–(8.10) suggests that h defined by (8.37) will solve the Poisson equation (8.9)–(8.10) whenever the vector $P^\star r$ is *proportional* to e, i.e. all the components of the vector $P^\star r$ are identical.

To investigate the matter further, we introduce the canonical decomposition of $\mathbb{X}$ into the recurrent and transient components induced by P, as already done in Section 8.4. Here, it can be assumed that P induces m recurrent classes, say $R_1, \ldots, R_m$, as well as a (possibly empty) set Tr of transient states, with the sets $\{R_1, \ldots, R_m, Tr\}$ forming a partition of $\mathbb{X}$. For any vector f, let f_k denote the restriction of f to R_k, $k = 1, \ldots, m$.

Recall that $p_{xy} = 0$ for x in R_k and y not in R_k, and the restriction P_k of P to the recurrent class R_k is irreducible and positive recurrent on it. Possibly upon

rearranging P into a block lower triangular form, we see that the restriction $(P^\star)_k$ of $P^\star$ to R_k coincides with $(P_k)^\star$ given by

$$(P_k)^\star := \lim_t \frac{1}{t+1} \sum_{s=0}^{t} P_k^s, \quad k = 1, \ldots, m \tag{8.39}$$

with all its rows being identical to the long-run probability distribution associated with the irreducible chain P_k. Consequently, $(P^\star r)_k = w_k e_k$ where the scalar w_k depends on the class R_k. Therefore (8.38) can be decomposed as

$$h_k + w_k e_k = r_k + P_k h_k, \quad k = 1, \ldots, m \tag{8.40}$$

and

$$h_{Tr} + (P^\star r)_{Tr} = r_{Tr} + \sum_{k=1}^{m} T_k h_k + T r h_{Tr} \tag{8.41}$$

where the matrices $T_1, \ldots, T_m$ and Tr are chosen appropriately from the decomposition of P associated with the sets $\{R_1, \ldots, R_m, Tr\}$.

Theorem 8.3 *The pair (h, w) is a solution to the Poisson equation (8.9)–(8.10) if and only if the conditions*

$$w_1 = \ldots = w_m = w \quad \text{and} \quad (P^\star r)_{Tr} = w e_{Tr} \tag{8.42}$$

hold, in which case h is given—uniquely up to an additive constant on each recurrent class—by (8.37), and w is the constant appearing in (8.42). The conditions (8.42) always hold when the Markov chain P has a single recurrent class.

Proof. The first part is immediate from the discussion given earlier since $P^\star r = we$ under (8.42). The uniqueness follows from Theorem 8.1 and from the fact that $I - Tr$ is invertible. To conclude the last part, it suffices to observe that under the assumption of a single recurrent class R_1 for the Markov chain P, the rows of $P^\star$ are all identical and of the form $(\nu, 0_{Tr})$ where ν coincides with the long-run probability distribution vector associated with the irreducible chain P_1 ∎

8.6 A PROBABILISTIC FORMULA FOR SOLUTIONS

Consider now the situation where the state space $\mathbb{X}$ is *countably infinite*. The matrix $P^\star$ is still well defined, but in general the invertibility of $I - P + P^\star$ cannot be guaranteed anymore owing to the intricate nature of the recurrence structures for Markov chains over countably infinite state spaces. As a result, the algebraic discussion of Section 8.5 cannot be carried through.

However, in some situations, *probabilistic* arguments can be used to prove the existence of a solution pair to the Poisson equation. Such a situation arises when there exists a distinguished state in $\mathbb{X}$, say z, which is *positive recurrent* in a sense made precise below. In this more restricted set-up, a possible approach would mimic the arguments of [36, Section 6.7], and would yield the solution

as the limit of the discounted cost associated with r, when the discount factor tends to 1. This line of arguments was developed in [38] and does yield a probabilistic representation of the solution already obtained by Derman and Veinott [8] through algebraic means.

Here, we take a different route for deriving this probabilistic representation of solutions to the Poisson equation. We do so in several steps by exploiting the martingale property of Lemma 8.1. To precisely state the conditions, we define the *first passage time* to the state z as the $\mathbb{F}_t$-stopping time T given by

$$T := \inf\{t > 0 : X_t = z\}. \tag{8.43}$$

The *recurrence* condition **(R)** enforced thereafter is the *finite mean* condition

$$(\mathbf{R}) \quad T(x) := \mathbb{E}_x[T] < \infty, \quad x \in \mathbb{X}. \tag{8.44}$$

The condition **(R)** is automatically satisfied when the set $\mathbb{X}$ is finite and the Markov chain P admits a single (positive) recurrent class decomposition $\mathbb{X} = R \cup Tr$ into a set R of positive recurrent states and a (possibly empty) set Tr of transient states. However, when the set $\mathbb{X}$ is not finite, the condition **(R)** is far more stringent. Indeed, not only does it imply the single class decomposition $\mathbb{X} = R \cup Tr$, but it also prohibits the chain from wandering too long or exclusively amongst the transient states. We relax the first restriction in Section 8.8.

We also find it convenient to consider the following *integrability* condition **(I)**, where

$$(\mathbf{I}) \quad R_\star(x) := \mathbb{E}_x\left[\sum_{t=0}^{T-1} |r(X_t)|\right] < \infty, \quad x \in \mathbb{X}. \tag{8.45}$$

Under **(I)** the quantities

$$R(x) := \mathbb{E}_x\left[\sum_{t=0}^{T-1} r(X_t)\right], \quad x \in \mathbb{X} \tag{8.46}$$

are well defined. Under the recurrence condition **(R)**, any bounded mapping r will satisfy the integrability condition **(I)**; in fact the conditions **(R)** and **(I)** coincide for $r(x) = 1$ for all x in $\mathbb{X}$.

The next result is a consequence of the martingale property given in Lemma 8.1.

Theorem 8.4 *Assume the recurrence condition* **(R)** *to hold and let* (h, w) *be a solution pair to the Poisson equation (8.9)–(8.10). If* h *is an element of* $\mathcal{I}_x$ *for some* x *in* $\mathbb{X}$*, then*

$$\lim_n \left\{ \mathbb{E}_x[\mathbf{1}[n < T]h(X_n)] + \mathbb{E}_x\left[\sum_{t=0}^{T\wedge n-1} r(X_t)\right] \right\} = wT(x) + h(x) - h(z). \tag{8.47}$$

Proof. By Lemma 8.1, the rvs $\{M_t,\ t = 0, 1, \dots\}$ given by (8.13) form a $(\mathbb{P}_x, \mathcal{F}_t)$-martingale. By Doob's Optional Sampling Theorem [7, 14], the stopped process $\{M_{T\wedge n},\ n = 0, 1, \dots\}$ is also a $(\mathbb{P}_x, \mathcal{F}_{T\wedge n})$-martingale, so that

$$\mathbb{E}_x[M_{T\wedge n}] = \mathbb{E}_x[M_0] = h(x), \quad n = 0, 1, \dots \tag{8.48}$$

By Lemma 8.1 we see that r is an element of $\mathcal{I}_x$ because h belongs to $\mathcal{I}_x$, and therefore, for all $n = 0, 1, \dots$, the three rvs $h(X_{T\wedge n})$, $T \wedge n$ and $\sum_{t=0}^{T\wedge n-1} r(X_t)$ are integrable under $\mathbb{P}_x$. From the definition of $M_{T\wedge n}$ we conclude by direct inspection of (8.48) that

$$\begin{aligned} h(x) &= \mathbb{E}_x\left[h(X_{T\wedge n}) - (T\wedge n)w + \sum_{t=0}^{T\wedge n-1} r(X_t)\right] \\ &= h(z)\mathbb{P}_x[T \le n] + \mathbb{E}_x\left[\mathbf{1}[T > n]h(X_n)\right] \\ &\quad - w\mathbb{E}_x[T\wedge n] + \mathbb{E}_x\left[\sum_{t=0}^{T\wedge n-1} r(X_t)\right]. \end{aligned} \tag{8.49}$$

Under **(R)**, we have $\lim_n\ \mathbb{P}_x[T \le n] = 1$, whereas $\lim_n\ \mathbb{E}_x[T\wedge n] = T(x)$ by monotone convergence, and the result (8.47) follows upon letting $n \uparrow \infty$ in (8.49). ∎

As we impose additional conditions, we see the form of the probabilistic representation emerge from the relation (8.47).

Corollary 8.2 *Assume the recurrence condition* **(R)** *to hold, and let (h, w) be a solution to the Poisson equation (8.9)–(8.10). If h belongs to $\mathcal{U}_x$ for some x in $\mathbb{X}$, then the relation*

$$h(x) = \lim_n\ \mathbb{E}_x\left[\sum_{t=0}^{T\wedge n-1} r(X_t)\right] - wT(x) + h(z) \tag{8.50}$$

holds. If in addition, the integrability condition **(I)** *holds, then*

$$h(x) = R(x) - T(x)w + h(z). \tag{8.51}$$

Proof. Under **(R)**, we have $\lim_n\ \mathbb{P}_x[T > n] = 0$. The uniform integrability under $\mathbb{P}_x$ of the rvs $\{h(X_t),\ t = 0, 1, \dots\}$ yields $\lim_n\ \mathbb{E}_x\left[\mathbf{1}[T > n]h(X_n)\right] = 0$, so that (8.50) follows from (8.47). Under **(I)** we get

$$\lim_n\ \mathbb{E}_x\left[\sum_{t=0}^{T\wedge n-1} r(X_t)\right] = \mathbb{E}_x\left[\sum_{t=0}^{T-1} r(X_t)\right] = R(x) \tag{8.52}$$

by dominated convergence, and (8.51) is an immediate consequence of (8.50) and (8.52). ∎

By carefully inspecting this last proof, we can extract additional information on the interaction between the uniform integrability of solutions and the integrability condition (8.45): We define the positive and negative parts of

the forcing function r by $r_\pm(x) := \max\{0, \pm r(x)\}$ for all x in $\mathbb{X}$, so that $r(x) = r_+(x) - r_-(x)$ and $|r(x)| = r_+(x) + r_-(x)$. In analogy with (8.46), we introduce the quantities

$$R_\pm(x) := \mathbb{E}_x \left[\sum_{t=0}^{T-1} r_\pm(X_t) \right], \quad x \in \mathbb{X} \tag{8.53}$$

which are both well defined, although possibly infinite. The relation $R(x) = R_+(x) - R_-(x)$ holds provided at least one of the quantities $R_+(x)$ and $R_-(x)$ is finite, while the equality $R_\star(x) = R_+(x) + R_-(x)$ is always valid.

Corollary 8.3 *Assume the recurrence condition* **(R)** *to hold, and let (h, w) be a solution to the Poisson equation (8.9)–(8.10). If h belongs to $\mathcal{U}_x$ for some x in $\mathbb{X}$, then the relation*

$$h(x) + R_-(x) = R_+(x) - wT(x) + h(z) \tag{8.54}$$

holds. If in addition, r is either bounded above or below, then $R_\star(x)$ is finite and the relation (8.51) holds.

Proof. The fact that h is an element of $\mathcal{U}_x$ (thus of $\mathcal{I}_x$) implies that r belongs to $\mathcal{I}_x$, and membership of $r_\pm$ in $\mathcal{I}_x$ follows. Therefore, for each $n = 0, 1 \ldots$, the rvs $\sum_{t=0}^{T \wedge n-1} r_\pm(X_t)$ are integrable under $\mathbb{P}_x$. The relation (8.49), derived in the proof of Theorem 8.4, still holds and can be rewritten as

$$\begin{aligned} h(x) + \mathbb{E}_x \left[\sum_{t=0}^{T \wedge n-1} r_-(X_t) \right] &= h(z)\mathbb{P}_x[T \le n] + \mathbb{E}_x \left[\mathbf{1}[T > n] h(X_n) \right] \\ &\quad - w\mathbb{E}_x[T \wedge n] + \mathbb{E}_x \left[\sum_{t=0}^{T \wedge n-1} r_+(X_t) \right]. \end{aligned} \tag{8.55}$$

Under **(R)**, we have $\lim_n \mathbb{P}_x[T \le n] = 1$, and $\lim_n \mathbf{E}_x[T \wedge n] = T(x)$ by monotone convergence. Moreover, the uniform integrability of the rvs $\{h(X_t),\ t = 0, 1, \ldots\}$ under $\mathbb{P}_x$ implies $\lim_n \mathbb{E}_x \left[\mathbf{1}[T > n] h(X_n) \right] = 0$, and we have $R_\pm(x) = \lim_n \mathbb{E}_x \left[\sum_{t=0}^{T \wedge n-1} r_\pm(X_t) \right]$ by monotone convergence. The result (8.54) follows from these facts upon letting $n \uparrow \infty$ in (8.55).

To establish the second statement, we note that r being either bounded above or below implies that at least one of the quantities $R_+(x)$ and $R_-(x)$ is finite, whence both are necessarily finite in view of the relation (8.54). ∎

Corollary 8.2 states that under conditions **(R)** and **(I)**, any "uniformly integrable" solution (h, w) of the Poisson equation is necessarily given by (8.51) (up to an additive constant). In a sense, we can view (8.51) as the "minimal" solution to (8.9)–(8.10). However, as we next show, (8.51) does define a solution even when there may exist no uniformly integrable one.

Theorem 8.5 *Assume both the recurrence condition* **(R)** *and the integrability condition* **(I)** *to hold. Then the pair (h, w) given by*

$$w = \frac{R(z)}{T(z)} \quad \textit{and} \quad h(x) = R(x) - w \cdot T(x), \quad x \in \mathbb{X} \tag{8.56}$$

is a solution to the Poisson equation (8.9)–(8.10) with $h(z) = 0$.

When the state space $\mathbb{X}$ is finite and the chain has a single recurrent class, (8.56) provides a *probabilistic* interpretation for the solution described through purely *algebraic* means in [2, 41].

Although condition **(R)** may seem quite restrictive, it is in some sense close to being necessary. Indeed, as shown by Cavazos-Cadena [4, Cor. 2.1–2.2, p. 105] if the Poisson equation admits a *bounded* solution for *every forcing function* r *which vanishes at infinity*, then (i) P admits a single recurrent class, which is necessarily positive recurrent; and (ii) a condition stronger than **(R)** holds, namely $\sup_x T(x) < \infty$.

Proof. The algebraic manipulations below are validated through the following summability conditions

$$\sum\nolimits_{y\neq z} p_{xy} T(y) < \infty \quad \text{and} \quad \sum\nolimits_{y\neq z} p_{xy} |R(y)| < \infty, \quad x \in \mathbb{X}. \tag{8.57}$$

In view of the comment following (8.46), we only need to establish the second condition in (8.57) as the first one reduces to it when $r \equiv 1$. By the Markov property, we get

$$R_\star(x) = |r(x)| + \sum\nolimits_{y\neq z} p_{xy} R_\star(y), \quad x \in \mathbb{X} \tag{8.58}$$

and the second summability condition in (8.57) follows from the integrability condition **(I)** since $|R(x)| \leq R_\star(x)$ for all x in $\mathbb{X}$.

The arguments that lead to (8.58) also show that

$$R(x) \;=\; r(x) + \sum\nolimits_{y\neq z} p_{xy} R(y), \quad x \in \mathbb{X} \tag{8.59}$$

and

$$T(x) \;=\; 1 + \sum\nolimits_{y\neq z} p_{xy} T(y), \quad x \in \mathbb{X}. \tag{8.60}$$

Fix x in $\mathbb{X}$. For any scalar w, we use (8.59)–(8.60) to write

$$\begin{aligned} R(x) - w \cdot T(x) \;&=\; \Big[r(x) + \sum\nolimits_{y\neq z} p_{xy} R(y) \Big] \\ &\quad\; - w \cdot \Big[1 + \sum\nolimits_{y\neq z} p_{xy} T(y) \Big]. \end{aligned} \tag{8.61}$$

Now, with the choice $w = R(z)/T(z)$, (8.61) becomes

$$\begin{aligned} R(x) - w \cdot T(x) + w \;&=\; r(x) + \sum\nolimits_{y\neq z} p_{xy} [R(y) - w \cdot T(y)] \\ &=\; r(x) + \sum\nolimits_{y} p_{xy} \; [R(y) - w \cdot T(y)], \end{aligned} \tag{8.62}$$

and (h, w) is indeed the postulated solution of the Poisson equation. ∎

We conclude this section by showing in what sense uniform integrability comes close to being necessary to ensure uniqueness. This will follow from the next result which is a simple consequence of (8.47) once we observe that

$$R(x) = \lim\nolimits_n \mathbb{E}_x \left[\sum_{t=0}^{T\wedge n-1} r(X_t) \right]$$

whenever $R_\star(x)$ is finite.

Corollary 8.4 *Assume the recurrence condition* **(R)** *to hold and let* (h, w) *be a solution pair to the Poisson equation (8.9)–(8.10). If* h *is an element of* $\mathcal{I}_x$ *for some* x *in* $\mathbb{X}$ *and if* $R_\star(x)$ *is finite, then*

$$\lim_n \mathbb{E}_x \left[\mathbf{1}[n < T] h(X_n)\right] = h(x) - h(z) - \left[R(x) - wT(x)\right]. \tag{8.63}$$

We see from (8.63) that this solution h in $\mathcal{I}_x$ coincides with that given by (8.56) provided $\lim_n \mathbb{E}_x \left[\mathbf{1}[n < T] h(X_n)\right] = 0$, a condition reminiscent of uniform integrability (i.e. h in $\mathcal{U}_x$) and indeed implied by it.

8.7 AN EXAMPLE

In this section we specialize the results obtained so far to a simple reflected random walk. The solution given by the probabilistic representation is computed explicitly, and shown to belong to $\mathcal{B}_\gamma$ ($= \mathcal{U}_\gamma$ where γ is the invariant distribution) whenever the forcing function r is an element of $\mathcal{B}_\gamma$. In that case, we also identify a class of solutions which are not uniformly integrable; in fact, we calculate *all solutions* to the Poisson equation, thereby exhibiting non-uniqueness for a positive recurrent Markov chain. The calculations are carried out in Appendix 8.12.1.

The situation considered here is that of a random walk on the non-negative integers with reflection, i.e. $\mathbb{X} = \mathbb{N}$ and

$$p_{0,0} = p_{x+1,x} = 1 - p := q \quad \text{and} \quad p_{x,x+1} = p, \quad x = 0, 1, \ldots \tag{8.64}$$

for some $0 < p < 1$. Upon defining $\rho := p/q$, we note that this Markov chain is positive recurrent—and condition **(R)** holds—whenever $\rho < 1$ (or equivalently $0 < p < 1/2$). In that case, making use of the defining relation (8.24), we readily determine the invariant distribution γ to be

$$\gamma(x) = (1 - \rho)\, \rho^x, \quad x = 0, 1, \ldots \tag{8.65}$$

For *any* forcing function r, the Poisson equation (8.9)–(8.10) takes the form

$$ph(0) + w = ph(1) + r(0)$$

and

$$h(x+1) + w = qh(x) + ph(x+2) + r(x+1), \quad x = 0, 1, \ldots \tag{8.66}$$

Before addressing the existence of solutions to (8.66), we show that such solutions are *not unique*. Indeed, if (h_i, w_i), $i = 1, 2$, are two solution pairs to (8.66), then their difference $(\Delta h, \Delta w)$ (in the notation (8.19)) solves the homogeneous equation $\Delta h + \Delta w e = P \Delta h$, which can be rewritten as

$$p\left[\Delta h(1) - \Delta h(0)\right] = \Delta w \tag{8.67}$$

and

$$p\left[\Delta h(x+2) - \Delta h(x+1)\right] = \Delta w + q\left[\Delta h(x+1) - \Delta h(x)\right] \tag{8.68}$$

for all $x = 0, 1, \ldots$. For any given value of Δw it is a simple matter to show that all the solutions to (8.67)–(8.68) are given by

$$\Delta h(x) = \Delta h(0) + \frac{\Delta w}{p - q}\left[\frac{1 - \rho^{-x}}{1 - \rho} + x\right], \quad x \in \mathbb{X} \tag{8.69}$$

and parameterized by the initial condition $\Delta h(0)$. Therefore, if (h_1, w_1) is a solution to (8.66), so is $(h_1 + \Delta h, w_1 + \Delta w)$ for *any choice* of Δw (in $\mathbb{R}$) where Δh is given by (8.69) with that value of Δw. In other words, even when all solutions to (8.66) are required to have identical initial conditions—a normalizing condition which dictates $\Delta h(0) = 0$ in (8.69)—we conclude that the solution set to (8.66) must necessarily be *non-countable* provided it is *not* empty. This non-uniqueness is *independent* of the choice of r, and holds also when $\rho \geq 1$, i.e. the chain is null recurrent or transient.

When $0 < \rho < 1$, we observe that Δh given by (8.69) can never belong to $\mathcal{U}_\gamma$ *unless* $\Delta w = 0$, thereby confirming the uniqueness of solutions in $\mathcal{U}_\gamma$, a result that derives from Theorem 8.1 (and independently from Corollary 8.2). It now remains to determine conditions under which the solution in $\mathcal{U}_\gamma$ exists.

With the representation (8.56) in mind, we take $z = 0$ and use (8.60) to obtain

$$T(x) = \frac{q\delta(0, x) + x}{q - p}, \quad x = 0, 1, \ldots; \tag{8.70}$$

the calculations are outlined in Appendix 8.12.1.

Next, intent on using the Green decomposition technique of Section 8.4.2, we compute for each v in $\mathbb{X}$ the cost per cycle function R_v associated with the cost $r_v : \mathbb{X} \to \mathbb{R} : x \to \delta(v, x)$. Since $w_v = \gamma(v)$, we invoke (8.26) to get

$$R_v(0) = w_v T(0) = \gamma(v)\frac{1}{1 - \rho} = \rho^v. \tag{8.71}$$

In Appendix 8.12.1 we also show that

$$v = 0, 1 \quad R_v(x) = v/q, \ x = 1, 2, \ldots \tag{8.72}$$

$$v = 2 \quad R_v(1) = \rho^2/p, \ R_v(x) = 1/q^2, \ x = 2, 3, \ldots \tag{8.73}$$

$$v = 3, 4, \ldots \quad R_v(x) = \rho^v/p^x, \ x = 1, 2, v - 1 \tag{8.74}$$

$$v = 3, 4, \ldots \quad R_v(x) = \frac{1}{p}\sum_{j=0}^{x \wedge v - 1} \rho^{v-j}, \ x = 3, 4, \ldots \tag{8.75}$$

Substituting (8.70)–(8.75) into (8.56), we obtain the solution h_v to the Poisson equation with forcing function r_v in the form

$$h_v(x) = R_v(x) - w_v T(x) = R_v(x) - \frac{x}{q}\rho^v, \quad x = 1, 2, \ldots \tag{8.76}$$

with $h_v(0) = 0$ by virtue of (8.71). Inspection of (8.71)–(8.75) reveals that $R_v(x)$ is bounded in x, and the solution h_v thus grows linearly is x. Therefore, invoking Lemma 8.4 (in conjunction with (8.65)), we see that h_v is an element of $\mathcal{U}_\gamma$ and is therefore the unique solution in that class.

Using the Green decomposition technique of Section 8.4.2, we can identify a large class of forcing functions for which (8.66) will have a unique solution in $\mathcal{U}_\gamma$; details of the derivation are available in Appendix 8.12.1.

Theorem 8.6 *Consider the random walk with reflection at the origin defined through (8.64) with $0 < \rho < 1$. Let r be a forcing function $\mathbb{X} \to \mathbb{R}$ such that $|r(x)| \le K(1 + r^x)$ for all x in $\mathbb{X}$, for some positive constants r and K. If $r\rho < 1$, then the decomposition (8.35) (where (h_v, w_v) is given by (8.76) for all v in $\mathbb{X}$) provides a solution (h, w) to the Poisson equation (8.9)–(8.10), and this solution is (unique) in $\mathcal{U}_\gamma$.*

8.8 BOUNDS AND EXTENSIONS

In this section, we explore some of the advantages afforded by the probabilistic representation (8.56). We use it to develop various bounds on the solution to the Poisson equation and to obtain an existence result for *unbounded* costs under a *multichain* structure.

8.8.1 Bounds

The following growth estimate is an easy consequence of the probabilistic representation (8.56).

Theorem 8.7 *Assume the recurrence condition* **(R)** *to hold. If r is bounded, i.e. $A := \sup_x |r(x)| < \infty$, then the solution pair (h, w) given by (8.56) satisfies the growth estimate*

$$|h(x)| \le (A + w)T(x), \quad x \in \mathbb{X}. \tag{8.77}$$

In general Theorem 8.7 does not hold when r is not bounded. However, in many situations of interest, the underlying Markov chain is "skip-free to the left" with respect to z. For example, in discrete-time queueing systems it is often the case that the decrease per unit time in the total number of customers is bounded above by the maximal number of available servers, say K. As a result, if z represents the empty state, then T is now the time until the queue empties, and we obtain the relation $|X_t| \le KT$, $0 \le t \le T$, where $|X_t|$ denotes the total number of customers at time t, and T is here the time until the system empties. With this in mind, we introduce the following condition: There exists a positive constant K such that

$$\mathbb{P}_x[d(z, X_t) \le KT, \quad 0 \le t \le T] = 1, \quad x \in \mathbb{X} \tag{8.78}$$

for some metric d on $\mathbb{X}$. Under such a condition, the representation (8.56) implies the following bound.

Theorem 8.8 *Assume both the recurrence condition* **(R)** *and the integrability condition* **(I)** *to hold. If the Markov chain satisfies (8.78), and if r exhibits the growth condition*

$$|r(x)| \le A\,(1 + d(z, x)^\delta), \quad x \in \mathbb{X} \tag{8.79}$$

for positive constants A and δ, then the solution h given by (8.56) satisfies the growth estimate

$$|h(x)| \le B\left(T(x) + \mathbb{E}_x\left[T^{\delta+1}\right]\right), \quad x \in \mathbb{X} \tag{8.80}$$

where

$$B = \max\{A + w, AK^{\delta}\}.$$

In other words, the growth rate of h is determined by the growth rate of moments of T. In particular, Theorem 8.8 shows how moments of recurrence times can be used to check that the solution (8.56) indeed belongs to $\mathcal{B}_\mu$ or $\mathcal{U}_\mu$ for some μ. Such information is of interest when studying the a.s. convergence of stochastic approximations schemes driven by Markov chains [1, 22, 25].

Proof. Note that (8.80) is automatically satisfied for $x = z$ since then $h(z) = 0$. Now, fixing $x \neq z$ in $\mathbb{X}$, we observe from the definition of $R(x)$ that

$$\begin{aligned}
|R(x)| &\le \mathbb{E}_x\left[\sum_{t=0}^{T-1} |r(X_t)|\right] \\
&\le A\mathbb{E}_x\left[\sum_{t=0}^{T-1} \left(1 + d(z, X_t)^{\delta}\right)\right] \\
&\le A\left(T(x) + \mathbb{E}_x\left[\sum_{t=0}^{T-1} (KT)^{\delta}\right]\right) \\
&= A\left(T(x) + K^{\delta}\mathbb{E}_x\left[T^{\delta+1}\right]\right)
\end{aligned} \tag{8.81}$$

where the second and third inequalities were obtained by making use of (8.79) and (8.78), respectively. The form of (8.56) now yields (8.80). ∎

Bounds were also developed by Glynn and Meyn [10] in terms of Lyapunov functions. In [26] Meyn provides quadratic estimates for the solutions to the Poisson equation associated with a queueing network, when the forcing function is linear. In addition, he obtains properties of solutions for general state spaces. In particular, if the forcing function is large outside a small set, then the solution can be bounded below as follows.

Theorem 8.9 *Assume both the recurrence condition* **(R)** *and the integrability condition* **(I)** *to hold, and let (h, w) be a solution pair to the Poisson equation (8.9)–(8.10) such that h belongs to $\mathcal{U}_x$ for all x in $\mathbb{X}$. If the forcing function r is bounded below, and has the property that for some ε in $(0, 1)$, the set S given by*

$$S := \{x \in \mathbb{X} : \ (1 - \varepsilon)r(x) < w\} \tag{8.82}$$

is finite, then h is bounded below.

If r is "norm-like" in the sense that the set $\{x \in \mathbb{X} : \ r(x) \le M\}$ is finite for each M, then S in (8.82) is finite. However, the condition of Theorem 8.9 is much weaker.

Proof. We first consider the case when r is non-negative. Since S is finite, the claim that h is bounded below will follow if we can show that

$$h(x) \geq h(z) - w \sup_{y \in S} T(y), \quad x \in \mathbb{X}. \tag{8.83}$$

First, with S and ε as above, we see that $w - r(x) \leq -\varepsilon r(x)$ for x *not* in S, whence

$$w - r(x) \leq w\mathbf{1}[x \in S] - \varepsilon r(x), \quad x \in \mathbb{X} \tag{8.84}$$

by the non-negativity of r. The fact that (h, w) is a solution pair yields

$$\begin{aligned} \sum\nolimits_y p_{xy} h(y) &= h(x) + w - r(x) \\ &\leq h(x) + w\mathbf{1}[x \in S] - \varepsilon r(x), \quad x \in \mathbb{X}. \end{aligned} \tag{8.85}$$

Next fix x in $\mathbb{X}$ and proceed similarly to the proof of Lemma 8.1: Membership of h in $\mathcal{U}_x$ implies that of r in $\mathcal{U}_x$; the rvs $\{M_t,\ t = 0, 1, \dots\}$ defined by $M_0 := h(X_0)$ and

$$M_{t+1} := h(X_{t+1}) + \sum_{s=0}^{t} \left(\varepsilon r(X_s) - w\mathbf{1}[X_s \in S]\right), \quad t = 0, 1, \dots$$

are integrable under $\mathbb{P}_x$ and form a $(\mathbb{P}_x, \mathcal{F}_t)$-supermartingale sequence. By Doob's Optional Sampling Theorem [7, 52], the rvs $\{M_{T \wedge n},\ n = 0, 1, \dots\}$ form a $(\mathbb{P}_x, \mathcal{F}_t)$-supermartingale sequence, and we get $h(x) = \mathbb{E}_x[M_0] \geq \mathbb{E}_x[M_{T \wedge n}]$ for all $n = 0, 1, \dots$. Using conditions **(I)** and **(R)**, and the fact that h belongs to $\mathcal{U}_x$, we note that all terms are well defined. Moreover, upon letting $n \uparrow \infty$, these conditions also imply that

$$\begin{aligned} h(x) &\geq h(z) + \varepsilon \mathbb{E}_x \left[\sum_{s=0}^{T-1} r(X_s)\right] - w\mathbb{E}_x \left[\sum_{s=0}^{T-1} \mathbf{1}[X_s \in S]\right] \\ &\geq h(z) - w\mathbb{E}_x[T] \end{aligned} \tag{8.86}$$

by the non-negativity of r. It is now immediate that (8.83) holds for x *in* S. For x *not* in S, consider the first hitting time σ of S, i.e. $\sigma := \inf\{t = 1, 2, \dots : X_t \in S\}$ (with the usual convention). The definition of σ and the strong Markov property readily yield

$$\begin{aligned} \mathbb{E}_x \left[\sum_{s=0}^{T-1} \mathbf{1}[X_s \in S]\right] &\leq \mathbb{E}_x \left[\sum_{s=\sigma}^{T-1} \mathbf{1}[X_s \in S]\right] \\ &\leq \mathbb{E}_x \left[\mathbb{E}_{X_\sigma}[T]\right] \\ &= \mathbb{E}_x[T(X_\sigma)]. \end{aligned} \tag{8.87}$$

By definition, X_σ belongs to S so that $T(X_\sigma) \leq \sup_{y \in S} T(y)$. Substituting into (8.87) we conclude that (8.83) indeed holds throughout $\mathbb{X}$.

In general, we note that if r is bounded below and satisfies (8.82), then so does the function $x \to r(x) - \inf_y r(y)$ (with the same ε but with appropriately modified w), and the result follows from the first part of the proof. ∎

8.8.2 Multiple classes

When the state space contains several positive recurrent classes, it is convenient to use a decomposition of the state space $\mathbb{X}$ into its transient and recurrent components $\{Tr, R_\alpha,\ \alpha \in A\}$, and to partition the Poisson equation accordingly. The treatment is similar to the one sketched briefly in [41].

With the decomposition and notation of Section 8.4, the results of the previous section extend to the multiple class case. For every α in A, select a state z_α in R_α and write $Z := \{z_\alpha,\ \alpha \in A\}$. We define the first passage times to the states z_α, α in A, and to the set $Z := \{z_\alpha,\ \alpha \in A\}$ by

$$T_\alpha := \inf\{t > 0 : X_t = z_\alpha\}, \quad \alpha \in A \tag{8.88}$$

and

$$T := \inf\{t > 0 : X_t \in Z\}. \tag{8.89}$$

Since each recurrent class is closed under P, at most one of the rvs $\{T_\alpha,\ \alpha \in A\}$ is finite $\mathbb{P}_x$-a.s. for each x in $\mathbb{X}$, so that

$$T = \sum\nolimits_\alpha T_\alpha \mathbf{1}[T_\alpha < \infty] \quad \text{on } [T < \infty] \quad \mathbb{P}_x - a.s. \tag{8.90}$$

under the convention $0 \cdot \infty = 0$. For future use, we also define

$$T_\alpha(x) := \mathbb{E}_x[T_\alpha \mathbf{1}[T_\alpha < \infty]], \quad \alpha \in A,\ x \in \mathbb{X}. \tag{8.91}$$

The appropriate version of condition **(R)** for the multiple class case is the *finite mean* condition

$$(\mathbf{Rm}) \quad T(x) := \mathbb{E}_x[T] < \infty, \quad x \in \mathbb{X}. \tag{8.92}$$

Note that **(Rm)** is essentially **(R)** but with the first passage time T defined through (8.90) rather than by (8.43). Under **(Rm)**, it is plain that for each x in $\mathbb{X}$, we have $T < \infty$ $\mathbb{P}_x$-a.s. and that for each α in A, $T_\alpha(x) = \mathbb{E}_x[T_\alpha] < \infty$ whenever x lies in R_α with the implication that all recurrent states are positive recurrent. Condition **(Rm)** also implies that starting at any state x in $\mathbb{X}$, the process eventually reaches the recurrent classes and does so in finite expected time.

We now impose conditions **(Rm)** and **(I)** (with T defined through (8.90)). For every α in A, the following expressions

$$R_\alpha(x) = \mathbb{E}_x\left[\sum_{t=0}^{T_\alpha - 1} r(X_t)\right], \quad x \in R_\alpha \quad \text{and} \quad w_\alpha := \frac{R_\alpha(z_\alpha)}{T_\alpha(z_\alpha)} \tag{8.93}$$

are then well defined. Define $R(x)$ by (8.46) with T now given by (8.89).

Theorem 8.10 *Assume the recurrence condition* **(Rm)** *and the integrability conditions* **(I)** *to hold. If there exists a scalar w such that $w_\alpha = w$ for all α in A, then the pair (h, w) with $h : \mathbb{X} \to \mathbb{R}$ given by*

$$h(x) = R(x) - w \cdot T(x), \quad x \text{ in } \mathbb{X} \tag{8.94}$$

is a solution to the Poisson equation with the property that $h(z) = 0$ for every z in Z.

Proof. The proof proceeds in two steps.

Step 1: First assume the set Tr of transient states to be empty. In that case the result follows readily from Theorem 8.5 if it can be shown that for each α in A, the pair (h_{R_α}, w_α) is indeed a solution pair to the projected Poisson equation (8.27) on R_α. That this is indeed the case can be seen as follows. The recurrence condition **(Rm)** implies that the restriction of the Markov chain P to the recurrence class R_α satisfies the condition **(R)** imposed in the single recurrent case. Therefore, by Theorem 8.5 the projected Poisson equation (8.27) on R_α admits as solution the pair (h_α, w_α) given by

$$h_\alpha(x) = R_\alpha(x) - w_\alpha \cdot T_\alpha(x), \quad x \in R_\alpha \tag{8.95}$$

with w_α given by (8.93). However, under **(Rm)** note that for x in R_α, $T = T_\alpha < \infty$ $\mathbb{P}_x$-a.s., whence $T(x) = T_\alpha(x)$ and $R(x) = R_\alpha(x)$. As a result, we find that

$$w = w_\alpha = \frac{R_\alpha(z_\alpha)}{T_\alpha(z_\alpha)} = \frac{R(z_\alpha)}{T(z_\alpha)},$$

so that $h(x) = h_\alpha(x)$ for all x in R_α.

Step 2: When Tr is not empty, the difficulty in obtaining a solution to the Poisson equation is related to the existence of transient states from which more than one recurrent class can be reached. First observe however that now (8.59)–(8.60) have to be replaced by

$$T(x) = 1 + \sum\nolimits_{y \notin Z} p_{xy} T(y), \quad x \in \mathbb{X} \tag{8.96}$$

and

$$R(x) = r(x) + \sum\nolimits_{y \notin Z} p_{xy} R(y), \quad x \in \mathbb{X}. \tag{8.97}$$

Therefore, in the same way that (8.59)–(8.60) lead to (8.62), it is easy to see that (8.96)–(8.97) imply

$$\begin{aligned} [R(x) - w \cdot T(x)] + w &= r(x) + \sum\nolimits_{y \notin Z} p_{xy} [R(y) - wT(y)] \\ &= r(x) + \sum\nolimits_{y} p_{xy} [R(y) - w \cdot T(y)] \end{aligned} \tag{8.98}$$

for each x in $\mathbb{X}$, where the last step follows from the fact that $R(z) = w \cdot T(z)$ for every z in Z as was noted in the first part of the proof. This time algebraic manipulations are validated through the summability conditions

$$\sum\nolimits_{y \notin Z} p_{xy} T(y) < \infty \quad \text{and} \quad \sum\nolimits_{y \notin Z} p_{xy} |R(y)| < \infty, \quad x \in \mathbb{X} \tag{8.99}$$

which follow from (8.96)–(8.97) and the integrability condition **(I)**. ∎

In this case (8.30) also has a solution, as can easily be seen by using (8.29) and the fact that for all x in $\mathbb{X}$ and y in R_α, the n-step transition probabilities $p_{xy}^{(n)}$ each converge to $\mathbb{P}_x[T_\alpha < \infty] \cdot \nu_y^{(\alpha)}$ where $\nu^{(\alpha)}$ is the invariant distribution of the Markov chain P when restricted to R_α.

8.9 PARAMETRIC DEPENDENCE: CONTINUITY

In several applications, including stochastic adaptive control and stochastic approximations [1, 3, 18, 22, 24, 25], the analysis simultaneously deals with a parameterized family of Markov chains, rather than with a single Markov chain, and crucial to the arguments is the smoothness (in the parameter) of solutions to the associated Poisson equations. Of particular interest are conditions on the model data which guarantee that the solution to the Poisson equation is continuous, or even Lipschitz continuous in the parameter. In this and the next sections we show how the representation results of Sections 8.6 and 8.8 provide a natural vehicle to explore this question. Our intent is not to get the best possible results, but rather to suggest ways of attacking these parametric issues.

In order to simplify the notation, the discussion in Sections 8.9 and 8.10 is given in the following framework: Only the case of a scalar parameter set is discussed as similar arguments can be developed *mutatis mutandis* for more general situations. Let the parameter set Θ be an open subset of $\mathbb{R}$, and consider a family $\{P(\theta),\ \theta \in \Theta\}$ of one-step transition probability matrices on the countable set $\mathbb{X}$, with $P(\theta) \equiv (p_{xy}(\theta))$. For each θ in Θ and x in $\mathbb{X}$, let $\mathbb{P}_x^\theta$ and $\mathbb{E}_x^\theta$ denote the probability measure and corresponding expectation operator induced on $(\Omega, \mathbb{F})$ by $P(\theta)$ given that $X_0 = x$.

For every θ in Θ, a given mapping $r(\theta) : \mathbb{X} \to \mathbb{R} : x \to r(\theta, x)$ drives the Poisson equation (8.9)–(8.10) associated with $P(\theta)$, i.e.

$$h + w = r(\theta)e + P(\theta)h. \tag{8.100}$$

As was the case in Section 8.6, we assume the existence of a distinguished state z in $\mathbb{X}$, independent of θ, with respect to which the integrability conditions **(R)** and **(I)** both hold for the Markov chain induced by $P(\theta)$ for *all* θ in Θ. Hence, with the $\mathbb{F}_t$-stopping time T still given by (8.43), we assume the appropriate versions of (8.44) and (8.45) to hold for each θ in Θ, and set

$$T(\theta, x) := \mathbb{E}_x^\theta[T], \quad x \in \mathbb{X} \tag{8.101}$$

and

$$R(\theta, x) := \mathbb{E}_x^\theta \left[\sum_{t=0}^{T-1} r(X_t) \right], \quad x \in \mathbb{X}. \tag{8.102}$$

Under the enforced assumptions, we may invoke Theorem 8.5 to conclude that (8.100) admits at least one solution $(h(\theta), w(\theta))$ where $w(\theta)$ is a scalar and $h(\theta)$ is a mapping $\mathbb{X} \to \mathbb{R} : x \to h(\theta, x)$. With the requirement $h(\theta, z) = 0$, this solution $(h(\theta), w(\theta))$ has the representation

$$w(\theta) = \frac{R(\theta, z)}{T(\theta, z)} \quad \text{and} \quad h(\theta, x) = R(\theta, x) - w(\theta) \cdot T(\theta, x), \quad x \in \mathbb{X}. \tag{8.103}$$

The next result identifies a set of natural conditions for establishing continuity of solutions to (8.100). Such a regularity property was required, for example, in [3].

Theorem 8.11 *Under the foregoing conditions, suppose that for each x in $\mathbb{X}$,*

(i) the mapping $\theta \to r(\theta, x)$ is continuous on Θ;

(ii) the mapping $\theta \to p_{xy}(\theta)$ is continuous over Θ for all y in $\mathbb{X}$;

(iii) the family of probability measures $\{p_{x\cdot}(\theta),\ \theta \in \Theta\}$ on $\mathbb{X}$ is tight;

(iv) the rvs $\{(T, \mathbb{P}_x^\theta),\ \theta \in \Theta\}$ are uniformly integrable; and

(v) the rvs $\{(\sum_{t=0}^{T-1} |r(\theta, X_t)|, \mathbb{P}_x^\theta),\ \theta \in \Theta\}$ are uniformly integrable.

Then for every x in $\mathbb{X}$, the mappings $\theta \to T(\theta, x)$ and $\theta \to R(\theta, x)$ are continuous over Θ.

In many applications, $r(\theta, x) = r(x)$ for all x in $\mathbb{X}$ and θ in Θ so that (i) automatically holds, while (iii) is satisfied whenever one-step transitions have some uniform (in θ) nearest-neighbor properties. The conditions (iv)–(v) are usually checked by (stochastically) bounding the original system uniformly in θ by means of another system which is naturally suggested by the original system. This approach was taken by Rosberg and Makowski in [33].

The next two lemmas are needed in the proof of Theorem 8.11; their proof is elementary and is omitted in the interest of brevity.

Lemma 8.5 *Assume (ii)–(iii) of Theorem 8.11. For all x and y in $\mathbb{X}$, and $k = 1, 2, \ldots$, the mappings $\theta \to \mathbb{P}_x^\theta[T = k]$ and $\theta \to \mathbb{P}_x^\theta[X_t = y, T = k]$, $1 \le t < k$, are all continuous on Θ.*

Lemma 8.6 *Assume (iii) of Theorem 8.11. For each $t = 1, 2, \ldots$ and x in $\mathbb{X}$, the family of distributions $\{(X_t, P_x^\theta),\ \theta \in \Theta\}$ is tight.*

To prepare the proof of Theorem 8.11, we set

$$R_m(\theta, x) := \mathbb{E}_x^\theta \left[\mathbf{1}[T \le m] \sum_{t=0}^{T \wedge m - 1} r(\theta, X_t) \right], \quad x \in \mathbb{X},\ m = 1, 2, \ldots \tag{8.104}$$

A proof of Theorem 8.11. Let x be a fixed element in $\mathbb{X}$. By a standard decomposition argument, there is no loss of generality in assuming $r(\theta, x) \ge 0$ for all x in $\mathbb{X}$ and θ in Θ. Moreover the first claim follows from the second one upon using $r(\theta, x) \equiv 1$.

In the general case, standard facts from analysis [37] imply the desired continuity result if it can be established that the mappings $\theta \to R_m(\theta, x)$, $m = 1, 2, \ldots$, are continuous on Θ, and then that the convergence $\lim_m R_m(\theta, x) = R(\theta, x)$ is uniform in θ.

To establish the first step, it suffices to show that the mappings $\theta \to \mathbb{E}_x^\theta[\mathbf{1}[T = k] r(\theta, X_t)]$, $0 \le t < k$, are continuous for (8.104) can be written as

$$R_m(\theta, x) = \sum_{k=1}^{m} \sum_{t=0}^{k-1} \mathbb{E}_x^\theta \left[\mathbf{1}[T = k] r(\theta, X_t) \right], \quad m = 1, 2, \ldots \tag{8.105}$$

Fix $0 \leq t < k$. Because the rvs $\{(X_t, \mathbb{P}_x^\theta),\ \theta \in \Theta\}$ are tight by Lemma 8.6, for every $\delta > 0$ there exists a finite subset $G_x(\delta)$ of $\mathbb{X}$ such that $\sup_{\theta\in\Theta} \mathbb{P}_x^\theta[X_t \notin G_x(\delta)] < \delta$. The uniform integrability condition (v) and the easy bound

$$\mathbb{E}_x^\theta\left[\mathbf{1}[T=k]1[X_t \notin G_x(\delta)]r(\theta, X_t)\right] \leq \mathbb{E}_x^\theta\left[\mathbf{1}[X_t \notin G_x(\delta)]\sum_{s=0}^{T-1} r(\theta, X_s)\right]$$

together now imply that for every $\varepsilon > 0$ there exists some $\delta(\varepsilon) > 0$ such that

$$\sup_\theta\ \mathbb{E}_x^\theta\left[\mathbf{1}[T=k]1[X_t \notin G_x(\delta(\varepsilon))]r(\theta, X_t)\right] \leq \varepsilon. \tag{8.106}$$

On the other hand, the mapping $\theta \to \mathbb{E}_x^\theta\left[\mathbf{1}[T=k]1[X_t \in G_x(\delta(\varepsilon))]r(\theta, X_t)\right]$ is continuous by virtue of Lemma 8.5 since $G_x(\delta(\varepsilon))$ is finite. The desired continuity of the mapping $\theta \to \mathbb{E}_x^\theta[1[T=k]r(\theta, X_t)]$ readily follows from this remark and from (8.106) by using a standard decomposition argument. Details are left to the interested reader.

For the second step, start with the estimate

$$0 \leq R(\theta, x) - R_m(\theta, x) = \mathbb{E}_x^\theta\left[\mathbf{1}[m < T]\sum_{t=0}^{T-1} r(\theta, X_t)\right], \quad m = 1, 2, \ldots$$

and observe that the uniform integrability of the rvs $\{(T, \mathbb{P}_x^\theta),\ \theta \in \Theta\}$ yields $\lim_m \sup_\theta \mathbb{P}_x^\theta[T > m] = 0$. This fact and the uniform integrability condition (v) immediately imply the uniform convergence

$$\lim_m\ \sup_\theta\ \mathbb{E}_x^\theta\left[\mathbf{1}[m < T]\sum_{t=0}^{T-1} r(\theta, X_t)\right] = 0, \tag{8.107}$$

and the proof is now complete ∎

8.10 PARAMETRIC DEPENDENCE: LIPSCHITZ CONTINUITY

Métivier and Priouret [25] have shown that the a.s. convergence of stochastic approximations passes through the Lipschitz continuity of the solutions $(h(\theta), w(\theta))$ to the parameterized Poisson equation (8.100). Arguments for establishing this Lipschitz continuity are now outlined in a somewhat restricted set-up which nevertheless often occurs in applications [18, 22]. To that end, we postulate that for all x in $\mathbb{X}$, the probability measures $\{p_x.(\theta),\ \theta \in \Theta\}$ on $\mathbb{X}$ are *mutually absolutely continuous*, i.e. if $p_{xy}(\theta) = 0$ for some y in $\mathbb{X}$ and θ in Θ, then $p_{xy}(\theta') = 0$ for all θ' in Θ. As a result, for each $m = 1, 2, \ldots$, the probability measures $\{\mathbb{P}_x^\theta,\ \theta \in \Theta\}$ are mutually absolutely continuous on the σ-field $\mathcal{F}_m$. If $L_m^x(\theta, \theta')$ denotes the Radon-Nikodym derivative of $\mathbb{P}_x^{\theta'}$ with respect to $\mathbb{P}_x^\theta$ (on $\mathcal{F}_m$), then

$$L_m^x(\theta, \theta') = \prod_{i=0}^{m-1} \frac{p_{X_i X_{i+1}}(\theta')}{p_{X_i X_{i+1}}(\theta)}, \quad m = 1, 2, \ldots \tag{8.108}$$

where the convention $\frac{0}{0} = 0$ is adopted. With $L_0^x(\theta, \theta') \equiv 1$, the rvs $\{L_m^x(\theta, \theta'), m = 0, 1, \dots\}$ form a $(\mathbb{P}_x^\theta, \mathcal{F}_m)$-martingale, and for any non-negative $\mathcal{F}_{T\wedge m}$-measurable rv X,

$$\mathbb{E}_x^{\theta'}[X] = \mathbb{E}_x^\theta \left[L_{T\wedge m}^x(\theta, \theta') \cdot X\right], \quad m = 1, 2, \dots \tag{8.109}$$

by standard results on absolutely continuous changes of measures [7].

Theorem 8.12 *Under the foregoing conditions, suppose there exist a constant $K > 0$ and a mapping $\mathbb{X} \to (0, \infty) : x \to K(x)$ such that for all θ and θ' in Θ,*

$$|p_{xy}(\theta) - p_{xy}(\theta')| \leq K p_{xy}(\theta) \cdot |\theta - \theta'|, \quad x, y \in \mathbb{X} \tag{8.110}$$

and

$$|r(\theta, x) - r(\theta', x)| \leq K(x) \cdot |\theta - \theta'|, \quad x \in \mathbb{X}. \tag{8.111}$$

If the moment conditions

$$\tilde{K}(x) := \sup_\theta \mathbb{E}_x^\theta \left[\sum_{t=0}^{T-1} K(X_t)\right] < \infty, \quad x \in \mathbb{X} \tag{8.112}$$

and

$$\tilde{R}(x) := \sup_\theta \mathbb{E}_x^\theta \left[T(1+\delta)^T \sum_{t=0}^{T-1} |r(\theta, X_t)|\right] < \infty, \quad x \in \mathbb{X} \tag{8.113}$$

are satisfied for some $0 < \delta \leq 1$, then for every x in $\mathbb{X}$, the mappings $\theta \to R(\theta, x)$ are locally Lipschitz continuous over Θ. In fact, whenever $|\theta - \theta'| \leq \frac{\delta}{K}$, the Lipschitz estimates

$$|R(\theta, x) - R(\theta', x)| \leq L(x)|\theta - \theta'|, \quad x \in \mathbb{X} \tag{8.114}$$

hold with $L(x) := K\tilde{R}(x) + \tilde{K}(x)$ for all x in $\mathbb{X}$.

A few observations are in order before proving Theorem 8.12: A result on the Lipschitz continuity of the mappings $\theta \to T(\theta, x)$, x in $\mathbb{X}$, is readily obtained from Theorem 8.12 upon using $r(\theta, x) \equiv 1$, in which case conditions (8.111)–(8.112) are automatically satisfied (with $K(x) = 0$), and (8.113) reduces to

$$\tilde{T}(x) := \sup_\theta \mathbb{E}_x^\theta \left[T^2(1+\delta)^T\right] < \infty, \quad x \in \mathbb{X}. \tag{8.115}$$

In fact, (8.113) also reduces to (8.115) whenever the cost function is bounded, i.e. $|r(\theta, x)| \leq B$ for all x in $\mathbb{X}$ and θ in Θ.

When the Lipschitz constant in (8.111) does not depend on x, i.e. $K(x) = K$ for all x in $\mathbb{X}$, then (8.112) reduces to the condition $\sup_\theta \mathbb{E}_x^\theta[T] < \infty$ for all x in $\mathbb{X}$.

The uniform bounds (8.115) can be checked in a variety of ways. For instance, in [20, 21, 22, 39] the authors considered a particular model where the distribution of the first passage time T under $\mathbb{P}_x^\theta$ is independent of θ—of course a rare occurrence—so that (8.115) becomes a simple moment requirement. Some general methods are sketched in [22]. In other situations, specific

arguments have to be developed, as we now do under the assumption that for some distinguished $\theta^\star$ in Θ, there exists a constant $B > 0$ such that for all θ in Θ,

$$\frac{p_{xy}(\theta)}{p_{xy}(\theta^\star)} \le B \quad \text{whenever } p_{xy}(\theta^\star) > 0, \quad x, y \in \mathbb{X}. \tag{8.116}$$

In that case, fixing θ in Θ and x in $\mathbb{X}$, we observe from (8.109) that

$$\mathbb{E}_x^\theta \left[\mathbf{1}[T \le m]\,(T \wedge m)\right] \le \mathbb{E}_x^{\theta^\star} \left[\mathbf{1}[T \le m](T \wedge m) \cdot B^{T \wedge m}\right], \quad m = 1, 2, \ldots$$

because $0 \le L_{T\wedge m}^x(\theta^\star, \theta) \le B^{T\wedge m}$ by virtue of (8.116), whence $\mathbb{E}_x^\theta [T] \le \mathbb{E}_x^{\theta^\star} \left[T \cdot B^T\right]$ by a simple limiting argument. The same reasoning shows that $\mathbb{E}_x^\theta \left[T^2(1+\delta)^T\right] \le \mathbb{E}_x^{\theta^\star} \left[T^2((1+\delta)B)^T\right]$. Consequently (8.115) holds under the structural condition (8.116) whenever the more compact conditions

$$\mathbb{E}_x^{\theta^\star} \left[T^2((1+\delta)B)^T\right] < \infty$$

holds.

A proof of Theorem 8.12. Let x be a fixed element in $\mathbb{X}$. As in the proof of Theorem 8.11, there is no loss of generality in assuming $r(\theta, x) \ge 0$ for all x in $\mathbb{X}$ and θ in Θ.

Fix θ and θ' in Θ, and $m = 1, 2, \ldots$. It is easily seen from (8.104) and (8.109) that

$$R_m(\theta', x) := \mathbb{E}_x^\theta \left[1[T \le m] \cdot L_{T\wedge m}^x(\theta, \theta') \sum_{t=0}^{T\wedge m-1} r(\theta', X_t)\right].$$

With this relation in mind, we define

$$A_m(\theta, \theta') := \mathbb{E}_x^\theta \left[\mathbf{1}[T \le m]\,[1 - L_{T\wedge m}^x(\theta, \theta')] \cdot \sum_{t=0}^{T\wedge m-1} r(\theta, X_t)\right]$$

and

$$B_m(\theta, \theta') := \mathbb{E}_x^\theta \left[\mathbf{1}[T \le m] \cdot L_{T\wedge m}^x(\theta, \theta') \cdot \left[\sum_{t=0}^{T\wedge m-1} r(\theta, X_t) - \sum_{t=0}^{T\wedge m-1} r(\theta', X_t)\right]\right]$$

so that

$$R_m(\theta, x) - R_m(\theta', x) = A_m(\theta, \theta') + B_m(\theta, \theta').$$

Condition (8.110) implies

$$\left|1 - \frac{p_{xy}(\theta')}{p_{xy}(\theta)}\right| \le K \cdot |\theta - \theta'| \quad \text{whenever } p_{xy}(\theta) > 0, \quad x, y \in \mathbb{X}$$

so that on the event $[L_{T\wedge m}^x(\theta, \theta') > 0]$, provided $K|\theta - \theta'| < 1$, we have

$$(1 - K|\theta - \theta'|)^{T\wedge m} \le L_{T\wedge m}^x(\theta, \theta') \le (1 + K|\theta - \theta'|)^{T\wedge m}. \tag{8.117}$$

Now the easy identities

$$(1 \pm Kt)^m - 1 = \int_0^t (\pm mK) \cdot (1 \pm K\tau)^{m-1} d\tau, \quad t > 0$$

yield

$$\left|(1 \pm Kt)^{T \wedge m} - 1\right| \leq K(T \wedge m) \cdot (1+\delta)^{T \wedge m} \cdot t, \tag{8.118}$$

whenever $0 < t \leq \frac{\delta}{K}$ (where $0 < \delta \leq 1$). Therefore, upon combining (8.117) and (8.118), under the condition $K|\theta - \theta'| < \delta$ we find

$$\begin{aligned} & |A_m(\theta, \theta')| \\ \leq \; & K \cdot \mathbb{E}_x^\theta \left[\mathbf{1}[T \leq m] \cdot (T \wedge m) \cdot (1+\delta)^{(T \wedge m)} \cdot \sum_{t=0}^{T \wedge m - 1} r(\theta, X_t) \right] \cdot |\theta - \theta'| \end{aligned}$$

and a simple limiting argument gives

$$\overline{\lim}_m \, |A_m(\theta, \theta')| \leq K \cdot \mathbb{E}_x^\theta \left[T \cdot (1+\delta)^T \cdot \sum_{t=0}^{T-1} r(\theta, X_t) \right] \cdot |\theta - \theta'|. \tag{8.119}$$

On the other hand, for each $m = 1, 2, \ldots$, we have

$$\begin{aligned} |B_m(\theta, \theta')| &\leq \mathbb{E}_x^\theta \left[\mathbf{1}[T \leq m] \cdot L^x_{T \wedge m}(\theta, \theta') \cdot \sum_{t=0}^{T \wedge m - 1} |r(\theta, X_t) - r(\theta', X_t)| \right] \\ &\leq \mathbb{E}_x^\theta \left[\mathbf{1}[T \leq m] \cdot L^x_{T \wedge m}(\theta, \theta') \cdot \sum_{t=0}^{T \wedge m - 1} K(X_t) \right] \cdot |\theta - \theta'| \\ &= \mathbb{E}_x^{\theta'} \left[\mathbf{1}[T \leq m] \sum_{t=0}^{T \wedge m - 1} K(X_t) \right] \cdot |\theta - \theta'|, \end{aligned}$$

where the second inequality is a consequence of (8.111), and the final equality follows from (8.109). In the limit, we conclude

$$\overline{\lim}_m \, |B_m(\theta, \theta')| \leq \mathbb{E}_x^{\theta'} \left[\sum_{t=0}^{T} K(X_t) \right] \cdot |\theta - \theta'| \tag{8.120}$$

and the result now readily follows from (8.119) and (8.120). ∎

Condition (8.113) (without the supremum) is exactly the notion of r-geometric ergodicity used in [27]. Finally, note that is is possible to obtain continuity results through the operator-theoretic methods of Chapter 9.

8.11 ACKNOWLEDGMENT

We are indebted to an anonymous referee for pointing out reference [8]. The work of the first author was supported partially through NSF Grant NSFD CDR-88-03012, NASA Grant NAGW77S and the Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0002. The work of the second author was supported in part by the funds for the promotion of research, and the promotion of sponsored research, at the Technion.

8.12 APPENDICES

8.12.1 The example

To obtain (8.70) from (8.65), we apply (8.60) with $z = 0$ to get

$$T(x) = 1 + pT(x+1), \quad x = 0, 1 \tag{8.121}$$

and

$$T(x) = 1 + pT(x+1) + qT(x-1), \quad x = 2, 3, \ldots \tag{8.122}$$

Since $T(0) = 1/\gamma(0)$ by standard results on Markov chains, we can use (8.65) to obtain (8.70). Indeed, the validity of (8.70) can be seen by substituting $T(0)$ into (8.121)–(8.122), so that $T(3) - T(2) = T(2) - T(1)$. By virtue of (8.122), this last equality propagates by induction, i.e. $T(x+1) - T(x) = T(x) - T(x-1)$ for all $x = 2, \ldots$ and (8.70) readily follows.

Fixing v in $\mathbb{X}$, we now set out to compute the cost per cycle R_v associated with r_v. To do so, we use the system of equations (8.59) which here takes the form

$$R_v(x) = r_v(x) + pR_v(x+1) + qR_v(x-1), \quad x = 0, 1 \tag{8.123}$$

and

$$R_v(x) = r_v(x) + pR_v(x+1) + qR_v(x-1), \quad x = 2, 3, \ldots \tag{8.124}$$

For $v = 0, 1$ or $v = 2$, we use (8.123)–(8.124) to get (8.72)–(8.73) by straightforward calculations. The case $v \geq 3$ is more involved: We observe that $R_v(x) = R_v(x+1)$, $x = v, \ldots$, which is readily derived from the definition of R_v (which holds for $v \geq 1$). Moreover, as the relation (8.124) implies

$$p(R_v(x+1) - R_v(x)) = q(R_v(x) - R_v(x-1)), \quad x = 2, \ldots, v-1$$

we conclude that

$$\begin{aligned} R_v(x+1) &= (R_v(x+1) - R_v(x)) + (R_v(x) - R_v(x-1)) + \ldots \\ &\quad \ldots + (R_v(2) - R_v(1)) + R_v(1) \\ &= \sum_{j=0}^{x-1} \rho^{-j}(R_v(2) - R_v(1)) + R_v(1) \end{aligned} \tag{8.125}$$

for all $x = 2, \ldots, v-1$. Because $r_v(0) = r_v(1) = 0$, we obtain (8.74) from (8.71) and (8.123), and combining this last relationship with (8.125), we finally get (8.75) after some algebra.

A proof of Theorem 8.6. First, under the enforced assumptions, we conclude from (8.65) that

$$\sum_{v=0}^{\infty} |r(v)| w_v \leq K \sum_{v=0}^{\infty} (1 + r^v)(1-\rho)\rho^v < \infty \tag{8.126}$$

because $\rho < 1$ and $r\rho < 1$, and the quantity w given by (8.35) is therefore well defined. Next, using (8.65) and (8.70), and the fact $\rho < 1$, we see that

$$\sum_{x=0}^{\infty} \gamma(x)T(x) = 1 + \frac{\rho}{q(1-\rho)^2} < \infty. \tag{8.127}$$

Finally, we claim that

$$\sum_{x=0}^{\infty} \gamma(x) \sum_{v=0}^{\infty} |r(v)|R_v(x) < \infty. \tag{8.128}$$

Before giving a proof of (8.128), we conclude from it the finiteness of the series $\sum_{v=0}^{\infty} |r(v)|R_v(x)$ for each x in $\mathbb{X}$. This fact, when combined with (8.35) and (8.126), readily implies that for each x in X, the quantity $h(x)$ given by

$$\begin{aligned} h(x) := \sum_{v=0}^{\infty} r(v)h_v(x) &= \sum_{v=0}^{\infty} r(v)[R_v(x) - w_v T(x)] \\ &= \sum_{v=0}^{\infty} r(v)R_v(x) - wT(x) \end{aligned} \tag{8.129}$$

is well defined since all infinite series are absolutely convergent.

To establish (8.128), we interchange the order of summation (by a simple application of Tonelli's Theorem), and note that

$$\begin{aligned} \sum_{x=0}^{\infty} \gamma(x) \sum_{v=0}^{\infty} |r(v)|R_v(x) &\leq K \sum_{v=0}^{\infty} (1 + r^v) \sum_{x=0}^{\infty} \gamma(x)R_v(x) \\ &= K(1-\rho) \sum_{v=0}^{\infty} (1 + r^v) \sum_{x=0}^{\infty} \rho^x R_v(x). \end{aligned}$$

The desired conclusion (8.128) now follows from (8.72)–(8.75) once we observe that for $v = 3, 4, \ldots$, the bounds

$$R_v(x) = \begin{cases} \rho^v & x = 0 \\ R\rho^{v-x}(1-\rho^x) & x = 1, \ldots, v \\ R(1-\rho^v) & x = v, v+1, \ldots \end{cases} \tag{8.130}$$

hold for some positive constant R which depends only on p. The calculations are tedious and are omitted; the finiteness of the various infinite series follows from the constraints $\rho < 1$ and $r\rho < 1$.

Combining (8.127) and (8.128) with (8.129) we see that h defined by (8.129) belongs to $\mathcal{B}_\gamma = \mathcal{U}_\gamma$. As the Poisson equation (8.10) involves here only a finite sum, it is immediate by substitution that under the stated conditions, the pair (h, w) defined above is indeed a solution to (8.10) since for each v in $\mathbb{X}$, the pair (h_v, w_v) is a solution to the Poisson equation. ∎

References

[1] A. Benveniste, M. Métivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, NY (1990).

[2] D.P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, New York, NY (1976).

[3] V. Borkar and M.K. Ghosh, "Ergodic and adaptive control of nearest-neighbor motions," *Math. Control, Signals & Systems* **4** (1991), pp. 81–98.

[4] R. Cavazos-Cadena, "Necessary conditions for the optimality equation in average reward Markov decision processes," *Appl. Math. and Opt.* **19** (1989), pp. 97–112.

[5] H.F. Chen, "Stochastic Approximation and its new applications," *Proc. 1994 Hong Kong Intl. Workshop on New Directions in Control and Manufacturing*, pp. 2-12.

[6] K.L. Chung, *Markov Chains with Stationary Transition Probabilities*, Second Edition, Springer Verlag, New York, NY (1967).

[7] K.L. Chung, *A Course in Probability Theory*, Second Edition, Academic Press, New York, NY (1974).

[8] C. Derman and A.F. Veinott, Jr., "A solution to a countable system of equations arising in Markovian decision processes," *Ann. Math. Stat.* **38** (1967), pp. 582–584.

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

[10] P.W. Glynn and S.P. Meyn, "A Lyapunov bound for solutions of the Poisson equation," *Annals of Probability* **24** (1996), pp. 916–931.

[11] L.G. Gubenko and E.S. Shtatland, "On Markov discrete time processes decision," *Theory Prob. Math. Stat.* **7** (1972), pp. 51–64.

[12] D. Heyman and M. Sobel, *Stochastic Models in Operations Research, Volume II: Stochastic Optimization*, McGraw-Hill, New York, NY (1984).

[13] A. Hordijk and F.M. Spieksma, "A new formula for the deviation matrix," *Technical Report* **TW-93-08**, Leiden University (1994).

[14] S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, NY (1974).

[15] J.G. Kemeny, J.L. Snell and A.W. Knapp, *Denumerable Markov Chains*, Second Edition, Springer-Verlag, New York, NY (1976).

[16] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA (1984).

[17] H.J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York (1997).

[18] D.-J. Ma, A.M. Makowski and A. Shwartz, "Stochastic approximations for finite state Markov chains," *Stochastic Processes and Their Applications* **35** (1990), pp. 27–45.

[19] A.M. Makowski and A. Shwartz, "Implementation issues for Markov decision processes," pp. 323–337 in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, Eds. W. Fleming and P.-L. Lions, IMA Volume **10**, Springer-Verlag, New York, NY (1988).

[20] A.M. Makowski and A. Shwartz, "Recurrence properties of a system of competing queues, with applications," EE Pub. **627**, Technion, Israel (1987).

[21] A.M. Makowski and A. Shwartz, "Recurrence properties of a discrete-time single-server network with random routing," EE Pub. **718**, Technion, Israel (1989).

[22] A.M. Makowski and A. Shwartz, "Stochastic approximations and adaptive control of a discrete-time single-server network with random routing," *SIAM J. Control and Optimization* **30** (1992), pp. 1476–1506.

[23] P. Mandl, "Estimation and control in Markov chains," *Adv. Appl. Prob.* **6** (1974), pp. 40–60.

[24] M. Métivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Transactions on Information Theory* **IT-30** (1984), pp. 140–150.

[25] M. Métivier and P. Priouret, "Théorèmes de convergence presque sûre pour une classe d'algo rithmes stochastiques à pas décroissants," *Prob. Theory Related Fields* **74** (1987), pp. 403–428.

[26] S.P. Meyn, "The policy improvement algorithm for Markov decision processes with general state space," *IEEE Transactions on Automatic Control* **AC-42** (1997), pp. 1663–1680.

[27] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, Series in Control and Communication in Engineering, London (U.K.) (1993).

[28] E. Nummelin, *General Irreducible Markov Chains and Non-negative Operators*, Cambridge University Press, Cambridge (U.K.) (1984).

[29] E. Nummelin, "On the Poisson equation in the potential theory of a single kernel," *Math. Scand.* **68** (1991), pp. 59–82.

[30] S. Orey, *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold, London (U.K.) (1971).

[31] D.R. Robinson, "Markov decision chains with unbounded costs and applications to the control of queues," *Adv. Appl. Prob.* **8** (1976), pp. 159–197.

[32] D.R. Robinson, "Optimality conditions for a Markov decision chain with unbounded cost," *J. Appl. Prob.* **17** (1980), pp. 996–1003.

[33] Z. Rosberg and A.M. Makowski, "Optimal dispatching to parallel heterogeneous servers—Small arrival rates," *IEEE Transactions on Automatic Control* **AC–35** (1990), pp. 789–796.

[34] S.M. Ross, "Non discounted denumerable Markovian decision models," *Ann. Math. Stat.* **39** (1968), pp. 412–423.

[35] S.M. Ross, "Arbitrary state Markovian decision models," *Ann. Math. Stat.* **39** (1968), pp. 2118–2122.

[36] S.M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, NY (1984).

[37] W. Rudin, *Real and Complex Analysis*, Second Edition, McGraw-Hill, New York, NY (1974).

[38] A. Shwartz and A.M. Makowski, "On the Poisson equation for Markov chains," EE Pub. **646**, Technion, Israel, September 1987.

[39] A. Shwartz and A.M. Makowski, "An optimal adaptive scheme for two competing queues with constraints," pp. 515–532 in *Analysis and Optimization of Systems*, Eds. A. Bensoussan and J.-L. Lions, Lecture Notes in Control and Info. Sci. **83**, Springer-Verlag, New York, NY (1987).

[40] A. Shwartz and A.M. Makowski, "Comparing policies in Markov decision processes: Mandl's lemma revisited," *Mathematics of Operations Research* **15** (1990), pp. 155–174.

[41] P. Whittle, *Optimization Over Time (Volume II): Dynamic Programming and Stochastic Control*, Wiley & Sons, New York, NY (1983).

[42] A.A. Yushkevich, "On a class of strategies in general Markov decision models," *Theory Prob. Appl.* **18** (1973), pp. 777–779.

Armand M. Makowski
Department of Electrical and Computer Engineering
and Institute for Systems Research
University of Maryland
College Park, MD 20742, U.S.A.
armand@isr.umd.edu

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

9 STABILITY, PERFORMANCE EVALUATION, AND OPTIMIZATION

Sean P. Meyn

Abstract: The theme of this chapter is stability and performance approximation for MDPs on an infinite state space. The main results are centered around stochastic Lyapunov functions for verifying stability and bounding performance. An operator-theoretic framework is used to reduce the analytic arguments to the level of the finite state-space case.

9.1 INTRODUCTION

9.1.1 Models on a general state space

This chapter focuses on stability of Markov chain models. Our main interest is the various relationships between stability; the existence of Lyapunov functions; performance evaluation; and existence of optimal policies for controlled Markov chains. We also consider two classes of algorithms for constructing policies, the policy and value iteration algorithms, since they provide excellent examples of the application of Lyapunov function techniques for ψ-irreducible Markov chains on an uncountable state space.

Considering the importance of these topics, it is not surprising that considerable research has been done in each of these directions. In this chapter we do not attempt a survey of all existing literature, or present the most comprehensive results. In particular, only the average cost optimality criterion is treated, and the assumptions we impose imply that the average cost is independent of the starting point of the process. By restricting attention in this way we hope that we can make the methodology more transparent.

One sees in several chapters in this volume that the generalization from finite state spaces to countable state spaces can lead to considerable technicalities. In particular, invariant distributions may not exist, and the cost functions

of interest may not take on finite values. It would be reasonable to assume that the move from countable state spaces, to MDPs on a general state space should be at least as difficult. This assumption is probably valid if one desires a completely general theory.

However, the MDPs that we typically come across in practice exhibit structure which simplifies analysis, sometimes bringing us to the level of difficulty found in the countable, or even the finite state space case. For example, all of the specific models to be considered in this chapter, and most in this volume, have some degree of spatial homogeneity. The processes found in most applications will also exhibit some level of continuity in the sense that from similar starting points, and similar control sequences, two realizations of the model will have similar statistical descriptions. We do not require strong continuity conditions such as the strong Feller property, although this assumption is sometimes useful to establish existence and uniqueness of solutions to the various static optimization problems that arise in the analysis of controlled Markov chains. An assumption of ψ-irreducibility, to be described and developed below, allows one to lift much of the stability theory in the discrete state space setting to models on a completely general, non-countable state space. This is an exceptionally mild assumption on the model and, without this assumption, the theory of MDPs on a general state space is currently extremely weak.

9.1.2 An operator-theoretic framework

When $\boldsymbol{x}$ is ψ-irreducible it is possible to enlarge the state space to construct an *atom* $\boldsymbol{\theta} \in \mathbb{X}$ which is reachable from any initial condition (i.e. $\mathbb{P}_x\{\tau_\theta < \infty\} > 0$, $x \in \mathbb{X}$). When the atom is *recurrent*, that is, $\mathbb{P}_\theta\{\tau_\theta < \infty\} = 1$, then an invariant measure (see (9.9)) is given by

$$\mu\{Y\} = \mathbb{E}_\theta\Big[\sum_{t=1}^{\tau_\theta} \mathbf{1}_Y(x_t)\Big], \qquad Y \in \mathbb{F}, \tag{9.1}$$

where τ_θ is the first return time to $\boldsymbol{\theta}$ (see (9.6)), and $\mathbf{1}_Y$ is the indicator function of the set Y. This construction, and related results may be found in [45, 39]. In words, the quantity $\mu\{Y\}$ expresses the mean number of times that the chain visits the set Y before returning to $\boldsymbol{\theta}$. This expression assumes that τ_θ is almost surely finite. If the *mean* return time $\mathbb{E}_\theta[\tau_\theta]$ is finite then in fact the measure μ is finite, and it can then be normalized to give an invariant probability measure. Finiteness of the mean return time to some desirable state is the standard stability condition used for Markov chains, and for MDPs in which one is interested in the average cost optimality criterion.

Unfortunately, the split chain construction is cumbersome when developing a theory for controlled Markov chains. The sample path interpretation given in (9.1) is appealing, but it will be more convenient to work within an operator-theoretic framework, following [45]. To motivate this, suppose first that we remain in the previous setting with an uncontrolled Markov chain, and suppose that do have an state $\boldsymbol{\theta} \in \mathbb{X}$ satisfying $\mathbb{P}_\theta\{\tau_\theta < \infty\} = 1$. Denote by s the function which is equal to one at $\boldsymbol{\theta}$, and zero elsewhere: That is, $s = \mathbf{1}_\theta$. We let ν denote the probability measure on $\mathbb{X}$ given by $\nu(Y) = p(Y \mid \boldsymbol{\theta})$, $Y \in \mathbb{F}$,

and define the 'outer product' of s and ν by

$$s \otimes \nu\,(x, Y) := s(x)\nu(Y).$$

For example, in the finite state space case the measure ν can be interpreted as a row vector, the function s as a column vector, and $s \otimes \nu$ is the standard (outer) product of these two vectors. Hence $s \otimes \nu$ is an $N \times N$ matrix, where N is the number of states.

In general, the *kernel* $s \otimes \nu$ may be viewed as a rank-one operator which maps L_∞ to itself, where L_∞ is the set of bounded, measurable functions on $\mathbb{X}$. Several other bounded linear operators on L_∞ will be developed in this chapter. The most basic are the n-step transition kernels, defined for $n \geq 1$ by

$$P^n f\,(x) := \int_{\mathbb{X}} f(y) p^n(dy|x), \qquad f \in L_\infty,$$

where $p^n(\cdot \mid \cdot)$ is the n-step state transition function for the chain. We set $P = P^1$. We can then write, in operator-theoretic notation,

$$\mathbb{P}_\theta\{\tau_\theta \geq n, x_n \in Y\} = \nu(P - s \otimes \nu)^{n-1} \mathbf{1}_Y, \qquad n \geq 1,$$

and hence the invariant measure μ given in (9.1) is expressed in this notation as

$$\mu(Y) = \sum_{n=1}^{\infty} \nu(P - s \otimes \nu)^{n-1} \mathbf{1}_Y, \qquad Y \in \mathbb{F}. \tag{9.2}$$

It is this algebraic description of μ that will be generalized and exploited in this chapter.

How can we mimic this algebraic structure without constructing an atom $\boldsymbol{\theta}$? First, we require a function $s\colon \mathbb{X} \to \mathbb{R}_+$ and a probability measure ν on $\mathbb{F}$ satisfying the *minorization condition*,

$$p(Y \mid x) \geq s(x)\nu(Y), \qquad x \in \mathbb{X},\ Y \in \mathbb{F}.$$

In operator theoretic notation this is written $P \geq s \otimes \nu$, and in the countable state space case this means that the transition matrix P dominates an outer product of two vectors with non-negative entries.

Unfortunately, this 'one step' minorization assumption excludes a large class of models, even the simple linear models to be considered as examples below. One can however move to the resolvent kernel defined by

$$K = (1 - \beta) \sum_{t=0}^{\infty} \beta^t P^t, \tag{9.3}$$

where $\beta \in]0, 1[$ is some fixed constant. For a ψ-irreducible chain the required minorization always holds for the resolvent K [39, Theorem 5.2.3]. The move to the resolvent is useful since almost any object of interest can be mapped between the resolvent chain, and the original Markov chain. In particular, the

invariant measures for P and K coincide (see [39, Theorem 10.4.3], or consider the resolvent equation in (9.11) below).

Much of the analysis then will involve the *potential kernels*, defined via

$$G := \sum_{t=0}^{\infty} K^t. \tag{9.4}$$

$$H := \sum_{t=0}^{\infty} (K - s \otimes \nu)^t. \tag{9.5}$$

In Theorem 9.1 below we demonstrate invariance of the σ-finite measure μ defined by,

$$\mu(Y) = \int_{\mathbb{X}} \nu(dx) H(x, Y), \qquad Y \in \mathbb{F},$$

provided the chain is *recurrent*. The invariant measure can be written in the compact form $\mu = \nu H$.

The measure μ will be *finite*, rather than just σ-finite, provided appropriate stability conditions are satisfied. The most natural stability assumption is equivalent to the existence of a Lyapunov function, whose form is very similar to the Poisson equation found in the average cost optimality equation. The development of these connections is one of the main themes of this chapter.

9.1.3 Overview

We conclude with an outline of the topics to follow. In the next section we review a bit of the general theory of ψ-irreducible chains, and develop some stochastic Lyapunov theory for such chains following [39, Chapters 11-14]. Following this, in Section 9.3 we develop in some detail the computation of the average cost through the Poisson equation, and the construction of bounds on the average cost. All of these results are developed for time homogeneous chains without control.

In Section 9.4 this stability theory is applied to the analysis of the average cost optimality equation (ACOE). We explore the consequences of this equation, and derive criteria for the existence of a solution.

Section 9.5 concerns two recursive algorithms for generating solutions to the ACOE: value iteration and policy iteration. It is shown that (i) either algorithm generates stabilizing stationary policies; (ii) for any of these policies, the algorithms generate uniform bounds on steady state performance. However, such results hold only if the algorithms are properly initialized.

Convergence is established for the policy iteration algorithm: Under suitable conditions, and when properly initialized, the algorithm converges to a solution of the ACOE.

Section 9.6 illustrates the theory with a detailed application to linear models, and to network scheduling.

This chapter is concluded with a discussion of some extensions and open problems.

9.2 STABILITY

In this section we consider a Markov chain $\boldsymbol{x}$ with uncontrolled transition function $p(\cdot \mid \cdot)$. The state space $\mathbb{X}$ is assumed to be a locally compact, separable metric space, and we let $\mathbb{F}$ denote the (countably generated) Borel σ-field on $\mathbb{X}$. Unless other references are given, all of the results described here together with their derivations can be found in [39].

9.2.1 ψ-irreducibility

Throughout this chapter we assume that ψ is a σ-finite measure on $\mathbb{F}$.

Definition 9.1

(i) *The chain is called ψ-**irreducible** if the resolvent kernel defined in (9.3) satisfies*

$$K(x, Y) > 0, x \in \mathbb{X} \quad \Longleftrightarrow \quad \psi(Y) > 0.$$

*We then call ψ an **irreducibility measure**.*

(ii) *We let $\mathbb{F}^+$ denote the set of all measurable $h\colon \mathbb{X} \to \mathbb{R}_+$ satisfying*

$$\psi(h) := \int_{\mathbb{X}} h(x)\psi(dx) > 0.$$

For $Y \in \mathbb{F}$ we write $Y \in \mathbb{F}^+$ provided $\psi(Y) > 0$.

If the chain is ψ-irreducible, then from any initial condition x, the process has a chance of entering any set in $\mathbb{F}^+$ in the sense that $\mathbb{P}_x\{\tau_Y < \infty\} > 0$, where τ_Y is the first return time,

$$\tau_Y = \min\{t \geq 1 : x_t \in Y\}. \tag{9.6}$$

Definition 9.2

(i) *A function $s\colon \mathbb{X} \to \mathbb{R}_+$ and a probability measure ν on $\mathbb{F}$ are called **petite** if*

$$K(x, Y) \geq s(x)\nu(Y), \qquad x \in \mathbb{X},\ Y \in \mathbb{F}. \tag{9.7}$$

(ii) *A set $Z \in \mathbb{F}$ is called **petite** if for some probability measure ν, and a constant $\delta > 0$,*

$$K(x, Y) \geq \delta\nu(Y), \qquad x \in Z,\ Y \in \mathbb{F}.$$

(iii) *The Markov chain is called a T-**chain** if every compact set is petite.*

It is not difficult to show that, for a ψ-irreducible chain, the set Z is petite if for each $Y \in \mathbb{F}^+$, there exists $n \geq 1$, and $\delta > 0$ such that

$$\mathbb{P}_x(\tau_Y \leq n) \geq \delta \qquad \text{for any } x \in Z. \tag{9.8}$$

For a ψ-irreducible chain, there always exists a countable covering of the state space by petite sets. In virtually all examples these can be taken to be compact, so that $\boldsymbol{x}$ is a T-chain. The following result is taken from [39, Proposition 5.5.5]:

Proposition 9.1 *Suppose that the Markov chain $\boldsymbol{x}$ is ψ-irreducible. Then there is a non-negative function s and a probability measure ν satisfying (9.7). We can choose s so that it is strictly-positive valued, $s\colon \mathbb{X} \to]0,1[$, and ν can be chosen so that it is equivalent to ψ (that is, $\psi(A) = 0 \Leftrightarrow \nu(A) = 0$).* ∎

The bound (9.7) is the most powerful consequence of the ψ-irreducibility assumption since it allows the construction of the potential kernel H defined in (9.5). The following lemma will be useful below when constructing solutions to dynamic programming equations:

Lemma 9.1 *For any petite pair s, ν we have,*

$$Hs\,(x) := \sum_{i=0}^{\infty} (K - s \otimes \nu)^i s\,(x) \le 1, \qquad x \in \mathbb{X}.$$

Proof. Define for $n \ge 0$, the kernel

$$H_n := \sum_{i=0}^{n} (K - s \otimes \nu)^i.$$

We show by induction that $H_n s$ is uniformly bounded for each n. For $n = 0$ we have $H_n s = s$, which is assumed to be bounded by one.

Suppose that $\|H_n s\|_\infty \le 1$ for some arbitrary $n \ge 0$. We then have,

$$\begin{aligned} H_{n+1}s &= s + (K - s \otimes \nu) H_n s \\ &\le s + (K - s \otimes \nu)\mathbf{1} \\ &= s + K\mathbf{1} - s \cdot \nu(\mathbb{X}) = K\mathbf{1} = 1. \end{aligned}$$

This proves the result since $H_n s \to Hs$ as $n \to \infty$. ∎

9.2.2 Recurrence

The crudest form of stability for a Markov chain is the property that the state visit 'important' sets with probability one from any starting point.

Definition 9.3

(i) *A ψ-irreducible chain is called* ***recurrent*** *if*

$$\mathbb{E}_x\Big[\sum_{t=0}^{\infty} \mathbf{1}(x_t \in Y)\Big] = \infty, \qquad Y \in \mathbb{F}^+,\ x \in \mathbb{X}.$$

(ii) *A measure μ on $\mathbb{F}$ is* ***invariant*** *if*

$$\mu(Y) = \int p(Y \mid x)\mu(dx), \qquad Y \in \mathbb{F}. \tag{9.9}$$

(iii) *If $\boldsymbol{x}$ is recurrent, and if in addition the chain admits an invariant probability measure μ, then the chain is called* ***positive recurrent****.*

In terms of the potential kernel (9.4), recurrence is expressed,

$$G(x, Y) = \infty, \qquad Y \in \mathbb{F}^+,\ x \in \mathbb{X}.$$

There are several equivalent characterizations of recurrence which are easier to verify. A proof of the following equivalences can be found in [39, Chapter 8], or [45, Theorem 3.7].

Theorem 9.1 *The following are equivalent for a ψ-irreducible Markov chain $\boldsymbol{x}$:*

(i) *$\boldsymbol{x}$ is recurrent.*

(ii) *There exists a set $\mathbb{X}_0 \in \mathbb{F}$ satisfying $\psi\{\mathbb{X}_0^c\} = 0$, and*

$$\mathbb{P}_x\{\tau_Y < \infty\} = 1, \qquad Y \in \mathbb{F}^+,\ x \in \mathbb{X}_0.$$

(iii) *For one petite set Z,*

$$\mathbb{P}_x\{\tau_Z < \infty\} = 1, \qquad x \in Z.$$

(iv) *For some pair (s, ν) satisfying the minorization condition (9.7) with $s \in \mathbb{F}^+$,*

$$\nu Hs := \sum_{t=0}^{\infty} \nu(K - s \otimes \nu)^t s = 1.$$

If any of these four equivalent conditions hold, then there exists a σ-finite measure μ which is invariant for the kernel P. It is unique in the sense that any σ-finite invariant measure is a constant multiple of the measure given by

$$\mu_\circ(Y) = \nu H\{Y\} = \sum_{t=0}^{\infty} \nu(K - s \otimes \nu)^t\{Y\}, \qquad Y \in \mathbb{F}. \tag{9.10}$$

Proof. We ask the reader to consult [39, 45] for the equivalence of (i) and (ii). Proofs of the remaining equivalences are provided here to illustrate how the various operators come into play. In particular, we will show that $\mu_\circ$ defines an invariant measure.

The implication (iii) $\Longrightarrow$ (i) is proved in two steps. First, on letting $\mathbb{X}_0 := \{x : \mathbb{P}_x\{\tau_Z < \infty\} = 1\}$, we find that this set is *absorbing*. That is,

$$p(\mathbb{X}_0 \mid x) = 1, \qquad x \in \mathbb{X}_0.$$

It then follows that this set is *full*: $\psi(\mathbb{X}_0^c) = 0$. Hence, from ψ-a.e. initial condition, the set Z is visited infinitely often with probability one. It follows that $G(x, Z) = \infty$ for $x \in \mathbb{X}_0$, and ψ-irreducibility then requires that $G(x, Z) = \infty$ for *all* x.

We now establish a similar identity for arbitrary $Y \in \mathbb{F}^+$. Since Z is petite, for any such Y we can find $\epsilon > 0$ such that

$$K(x, Y) \geq \epsilon \mathbf{1}_Z(x), \qquad x \in \mathbb{X}.$$

This can also be written $K\mathbf{1}_Y \geq \epsilon \mathbf{1}_Z$, and hence, for $x \in \mathbb{X}$,

$$\begin{aligned} \infty &= \epsilon G\mathbf{1}_Z \\ &\leq GK\mathbf{1}_Y = -\mathbf{1}_Y + G\mathbf{1}_Y \leq G\mathbf{1}_Y. \end{aligned}$$

We have thus shown that (iii) implies recurrence.

Conversely, if the chain is recurrent, take any $Z_1 \in \mathbb{F}^+$, and define

$$\mathbb{X}_0 := \{x : \mathbb{P}_x\{\tau_{Z_1} < \infty\} = 1\}; \qquad Z := Z_1 \cap \mathbb{X}_0 .$$

The set $\mathbb{X}_0$ is absorbing, so we do find that, for $x \in Z$,

$$\mathbb{P}_x\{\tau_Z < \infty\} = \mathbb{P}_x\{\tau_{Z_1} < \infty\} = 1.$$

This shows that (iii) holds with this set Z, and establishes the implication (i) $\Longrightarrow$ (iii).

We now show that (iv) implies recurrence. To avoid dealing with potentially infinite sums, let $\lambda > 0$, and define the kernels

$$H_\lambda(x) = \sum_0^\infty \lambda^{-n-1}(K - s \otimes \nu)^n \qquad G_\lambda(x) = \sum_0^\infty \lambda^{-n-1} K^n.$$

Note that $H_1 = H$ and $G_1 = G$ are the potential kernels introduced in the introduction.

We denote $\alpha_\lambda = \nu(H_\lambda s)$, and $\beta_\lambda = \nu(G_\lambda s)$. It is obvious that β_λ is finite for each $\lambda > 1$. An application of Lemma 9.1 shows that $\alpha_\lambda \leq 1$ for all $\lambda \geq 1$.

Applying the kernel $\lambda^{-1}K$ to the function $H_\lambda s$ gives,

$$\begin{aligned} \lambda^{-1} K H_\lambda s &= \lambda^{-1}(K - s \otimes \nu) H_\lambda s + (s \otimes \nu) H_\lambda s \\ &= H_\lambda s - \lambda^{-1}(1 - \alpha_\lambda) s. \end{aligned}$$

Iterating this equation we find, for $\lambda > 1$,

$$H_\lambda s - \lambda^{-1}(1 - \alpha_\lambda) \sum_0^{n-1} \lambda^{-i} K^i s = \lambda^{-n} K^n H_\lambda s \to 0, \qquad n \to \infty.$$

This shows that $H_\lambda s = (1 - \alpha_\lambda) G_\lambda s$, and hence that $\alpha_\lambda = (1 - \alpha_\lambda)\beta_\lambda$ for $\lambda > 1$. Letting $\lambda \downarrow 1$ and applying the monotone convergence theorem we see that

$$\beta_1 = \frac{\alpha_1}{1 - \alpha_1}.$$

This shows that $\alpha_1 = 1$ if and only if $\beta_1 = \infty$.

It remains to show that an infinite value for β_1 is equivalent to recurrence. If $\beta_1 = \nu G s = \infty$, it then follows that

$$\int G(x, dy)s(y) = \infty,$$

for ψ-a.e. $x \in \mathbb{X}$ $[\nu]$, and we can extend this to all x by ψ-irreducibility. Let $Y \in \mathbb{F}^+$ be arbitrary, and let $\delta = \nu(Y)$. We can assume by Proposition 9.1 that $\delta > 0$. From the minorization condition we have, $K\mathbf{1}_Y \geq \delta s$, and hence the bound on G gives,

$$\infty = \delta^{-1} \int G(x, dy)s(y) \leq G(x, Y) - \mathbf{1}_Y(x).$$

We deduce that the chain is recurrent as required. The proof that recurrence implies (iv) is identical.

To establish invariance of $\mu_\circ$, first apply the kernel $(K - s \otimes \nu)$ to $\mu_\circ$ on the right to obtain,

$$\mu_\circ(K - s \otimes \nu) = \sum_{t=0}^{\infty} \nu(K - s \otimes \nu)^{t+1} = \mu_\circ - \nu.$$

Now by recurrence and (iv) we have $\mu_\circ(s) = 1$, which shows that $\mu_\circ$ is K-invariant: $\mu_\circ K = \mu_\circ$. Using the identity

$$PK = KP = \beta^{-1}K + (1 - \beta^{-1})I, \tag{9.11}$$

we conclude that $\mu_\circ$ is P-invariant. ∎

The invariant measure given in (9.10) will be finite, so the chain $\boldsymbol{x}$ is positive recurrent, provided that the *mean* return time to a petite set Z is bounded:

$$\sup_{x \in Z} \mathbb{E}_x[\tau_Z] < \infty. \tag{9.12}$$

In terms of the variables used in the previous proof, this is equivalent to requiring that $\alpha_1' < \infty$, where the prime denotes the left derivative of α with respect to λ (see discussion surrounding equation (5.6) of [45]).

While these definitions lead to an elegant theory, in practice one can typically take $\mathbb{X}_0 = \mathbb{X}$ in (ii). In this case the chain is called *Harris*, and it is called *positive Harris* if there is also an invariant probability measure. The chains we consider next exhibit a far stronger form of stability.

9.2.3 $\boldsymbol{c}$-Regularity and Lyapunov functions

The next level of stability that we consider is related to steady state performance, which moves us closer to the average cost optimality criterion. Suppose that $c\colon \mathbb{X} \to [1, \infty)$ is a measurable function on the state space, and suppose that the chain is ψ-irreducible.

Definition 9.4

(i) *A set $S \in \mathbb{F}$ is called* ***c-regular*** *if for any $Y \in \mathbb{F}^+$,*

$$\sup_{x \in S} \mathbb{E}_x \Big[\sum_{t=0}^{\tau_Y - 1} c(x_t) \Big] < \infty.$$

(ii) *The Markov chain is called c-regular if the state space $\mathbb{X}$ admits a countable covering by c-regular sets.*

A c-regular chain is automatically positive Harris, and using (9.10) we see that a c-regular chain possesses an invariant probability measure μ satisfying $\mu(c) := \int c(x)\, \mu(dx) < \infty$. The following result is a consequence of the f-Norm Ergodic Theorem of [39, Theorem 14.0.1].

Theorem 9.2 *Assume that $c\colon \mathbb{X} \to [1, \infty)$ and that $\boldsymbol{x}$ is c-regular. Then, for any measurable function g which satisfies*

$$\sup_{x \in \mathbb{X}} \Big(\frac{|g(x)|}{c(x)} \Big) < \infty,$$

the following ergodic theorems hold for any initial condition:

$$\textit{(i)} \qquad \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} g(x_t) \;=\; \mu(g), \qquad [\mathbb{P}_x] - a.s..$$

$$\textit{(ii)} \quad \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_x[g(x_t)] \;=\; \mu(g).$$

■

One approach to establishing c-regularity is through the following extension of Foster's criterion, or Lyapunov's second method. In general, such approaches involve the construction of a function V on the state space, taking positive values, such that $V(x_t)$ is in some sense decreasing whenever the state x_t is 'large'. In our context this decreasing property can be formulated as follows: Find a function $V : \mathbb{X} \to \mathbb{R}_+$ and a constant $\bar{J} \in \mathbb{R}_+$ such that

$$PV\,(x) := \mathbb{E}[V(x_{t+1}) \mid x_t = x] \le V(x) - c(x) + \bar{J}, \qquad x \in \mathbb{X}. \tag{9.13}$$

When this bound holds, we say that V is a *Lyapunov function*.

However, for this to imply any form of stability, the difference $c(x) - \bar{J}$ must be positive for 'large' x. We say that c is

Definition 9.5

near-monotone *if the sublevel set $Z_\eta := \{x \in \mathbb{X} : c(x) \le \eta\}$ is petite for any $\eta < \|c\|_\infty$. The supremum norm $\|c\|_\infty$ may be infinite.*

norm-like *if the sublevel set Z_η is a pre-compact subset of the metric space $\mathbb{X}$ for any η.*

Related assumptions on c are used in [1, 5, 39, 42].

Theorem 9.3 *Assume that $c\colon \mathbb{X} \to [1,\infty)$ is near-monotone, and suppose that $\bar{J} < \|c\|_\infty$. Then,*

(i) *If there exists a finite, positive-valued solution V to the inequality (9.13), then there exists $d_0 < \infty$ such that for each $Y \in \mathbb{F}^+$,*

$$\mathbb{E}_x\Big[\sum_{t=0}^{\tau_Y} c(x_t)\Big] \le d_0 V(x) + d(Y), \qquad x \in \mathbb{X}, \tag{9.14}$$

where $d(Y) < \infty$ is a constant. Hence, each of the sublevel sets $Z_n = \{x : V(x) \le n\}$ is c-regular, and the process itself is c-regular.

(ii) *If the chain is c-regular, then for any c-regular set $Z \in \mathbb{F}^+$, the function*

$$V^*(x) = \mathbb{E}_x\Big[\sum_{t=0}^{\tau_Z} c(x_t)\Big], \qquad x \in \mathbb{X}, \tag{9.15}$$

is a near-monotone solution to (9.13).

∎

Proof. First observe that the bound (9.13) is equivalent to the drift condition $PV_0 \le V_0 - c + b\mathbf{1}_Z$, where Z is petite: if (9.13) holds, we can take $V_0 = d_0 V$ and $b = d_0\bar{J}$, with d_0 sufficiently large. The result is then an immediate consequence of [39, Theorem 14.2.3]. ∎

9.3 PERFORMANCE

For a function $c\colon \mathbb{X} \to [1,\infty)$ we define the average cost by

$$J := \limsup_{n\to\infty} \frac{1}{n}\,\mathbb{E}_x\Big[\sum_{t=0}^{n-1} c(x_t)\Big].$$

We have seen that, for a c-regular chain, the average cost is finite and independent of x, with $J = \mu(c)$. Here we examine the relationship between c and J through Poisson's equation.

9.3.1 Poisson's equation

This functional equation originated in the analysis of partial differential equations: Assuming that f is some given function on $\mathbb{R}^n$, the equation is written

$$\Delta h = -f$$

where h is an unknown function on $\mathbb{R}^n$, and Δ is the Laplacian. The probabilistic interpretation of this equation becomes evident when one realizes that Δ is the generator for a Brownian motion on $\mathbb{R}^n$ - a similar equation can be

posed for any Markov process in continuous time. When time is discrete, we define the generator as $\Delta = P - I$, and the Poisson equation then takes on the exact same form. If we take $f = c - \mu(c)$, then Poisson's equation can be written,

$$Ph\,(x) := \mathbb{E}_x[h(x_{t+1}) \mid x_t = x] = h(x) - c(x) + J, \qquad x \in \mathbb{X}, \tag{9.16}$$

where $J = \mu(c)$.

Motivation for looking at this equation is provided by our prior stability analysis. First note that the drift inequality (9.13) suggests a simple approach to obtaining performance bounds. By iterating this equation one obtains,

$$0 \le \mathbb{E}_x[V(x_n)] \le V(x) - \sum_{t=0}^{n-1} \mathbb{E}_x[c(x_t)] + n\bar{J}. \tag{9.17}$$

Dividing by n and letting $n \to \infty$ then gives the upper bound, $J \le \bar{J}$. The question then is, by choosing V carefully can we get a tight upper bound? The answer is yes, provided the chain is c-regular, and in this case, the minimal upper bound $\bar{J}_*$ is evidently $\mu(c)$, with μ equal to the invariant probability for the chain. We see in Theorem 9.4 that this 'optimal Lyapunov function' is precisely the solution to Poisson's equation.

Equation (9.16) resembles a version of the Lyapunov drift inequality (9.13). However, the function h cannot play the role of a Lyapunov function unless it is positive-valued, or at least bounded from below. This cannot be expected in general, as the following example demonstrates.

Consider the Markov chain on $\mathbb{X} = \mathbb{N}$ with transition probabilities,

$$p(x+1 \mid x) = \begin{cases} \alpha & x > 0; \\ \beta & x < 0; \\ \frac{1}{2} & x = 0. \end{cases} \qquad p(x-1 \mid x) = \begin{cases} \beta & x > 0; \\ \alpha & x < 0; \\ \frac{1}{2} & x = 0. \end{cases}$$

We assume that $\alpha + \beta = 1$, and that $\alpha < \beta$. The latter condition ensures that $\boldsymbol{x}$ is a c-regular T-chain, where c can be taken as any polynomial function of x.

Define $h(x) = x/(\beta - \alpha)$. We can compute the conditional expectation,

$$Ph\,(x) = p(x+1 \mid x)h(x+1) + p(x-1 \mid x)h(x-1) = -\text{sign}\,(x).$$

The function $c(x) = 1 + \text{sign}(x)$, $x \in \mathbb{X}$, is bounded, non-negative, and has steady state mean equal to one. The function h is the associated solution to Poisson's equation,

$$Ph = h - c + 1.$$

It is not bounded from below.

A solution to Poisson's equation *will* be bounded from below under suitable conditions on the chain, and the function c. One such condition is the norm-like assumption, or the milder near-monotonicity condition for c. The following result surveys the relevant consequences of c-regularity, and introduces the form

of the Poisson equation that we will analyze in the remainder of this chapter. These results are taken from [42], following [46, 22].

Theorem 9.4 *Assume that* $c\colon \mathbb{X} \to [1,\infty)$ *and that* $\boldsymbol{x}$ *is* c*-regular. Then,*

(i) *There exists a measurable function* $h\colon \mathbb{X} \to \mathbb{R}$ *satisfying (9.16), where* $J = \mu(c)$.

(ii) *One solution to (9.16) may be expressed,*

$$h = \frac{\beta}{1-\beta}[H - I]\bar{c}, \tag{9.18}$$

where $\bar{c} = c - \mu(c)$, $\beta < 1$ *is used in the definition (9.3) of the kernel* K, *and the pair* (s,ν) *is petite with* $s \in \mathbb{F}^+$.

(iii) *Suppose moreover that the function* c *is near-monotone. Then the solution (9.18) is uniformly bounded from below,* $\inf_{x\in\mathbb{X}} h(x) > -\infty$. *It is essentially unique in the following sense: If* h' *is any function on* $\mathbb{X}$ *which is uniformly bounded from below, and solves the Poisson inequality*

$$Ph\,(x) \le h(x) - c(x) + J, \qquad x \in \mathbb{X},$$

with $J = \mu(c)$, *then there exists a constant* k' *such that*

$$\begin{aligned} h'(x) &\ge h(x) + k', \qquad x \in \mathbb{X}; \\ h'(x) &= h(x) + k', \qquad \psi - a.e.\ x \in \mathbb{X}. \end{aligned}$$

(iv) *If* V *is any solution to (9.13) with* $\bar{J} < \|c\|_\infty$ *and* c *near-monotone, then the solution (9.18) satisfies the uniform upper bound, for some* $d_0 < \infty$,

$$h(x) \le d_0(V(x) + 1), \qquad x \in \mathbb{X}.$$

Proof. To prove (i) and (ii), consider first the function

$$h_0(x) = H\bar{c} := \sum_{t=0}^{\infty} (K - s\otimes\nu)^t \bar{c}\,(x), \qquad x \in \mathbb{X}.$$

The construction of $\mu_\circ$ in (9.10) gives

$$\nu(h_0) = \mu_\circ(\bar{c}) = \mu_\circ(\mathbb{X})\mu(\bar{c}) = 0.$$

This immediately gives,

$$Kh_0\,(x) = (K - s\otimes\nu)h_0\,(x) = h_0(x) - \bar{c}(x).$$

That is, h_0 solves the Poisson equation for the kernel K. By applying the identity (9.11) we see that

$$h := \frac{\beta}{1-\beta}(h_0 - \bar{c}) = \frac{\beta}{1-\beta} Kh_0,$$

solves the Poisson equation for original transition kernel P.

To prove (iii), define h_0' by

$$h_0' := \frac{1-\beta}{\beta} h' + \bar{c}.$$

The resolvent equation (9.11) combined with the bound in (iii) gives,

$$\frac{1-\beta}{\beta} Kh' \le \frac{1-\beta}{\beta} h' - K\bar{c},$$

from which it follows that $Kh_0' \le h_0' - \bar{c}$. This implies that the quantity $\nu(h_0')$ is finite:

$$s(x)\nu(h_0') \le Kh_0' \le h_0'(x) - \bar{c}(x), \qquad x \in \mathbb{X}.$$

By adding a constant to h' we can and will assume that $\nu(h_0') = 0$. We then have,

$$(K - s \otimes \nu) h_0' \le h_0' - \bar{c}$$

We have assumed that h' (and hence h_0') is bounded from below. By iterating the last inequality, it follows that for some $L < \infty$,

$$\begin{aligned} -L &\le (K - s \otimes \nu)^n h_0' \\ &\le h_0' - \sum_0^{n-1} (K - s \otimes \nu)^i \bar{c} \end{aligned}$$

We conclude that

$$h_0' \ge h_0 - L.$$

Letting $u = h_0' - h_0$, we see that u is bounded from below, and it is *superharmonic*: $Pu \le u$. These properties imply the desired result (see [39, p. 414] or [42]).

The proof of (iv) is similar to (iii). We first establish a bound of the form,

$$PV \le V - \epsilon c + bs,$$

with $\epsilon > 0$, $b < \infty$. We can move to the resolvent to obtain an analogous bound,

$$(K - s \otimes \nu)V \le KV \le V - \epsilon_1 c + b_1 s,$$

and on iterating we find that

$$\epsilon_1 Hc \le b_1 Hs + V \le b_1 + V.$$

In Lemma 9.1 we have shown that Hs is everywhere bounded by unity. The bound in (iv) immediately follows. ∎

9.3.2 Simulation

We have now seen that the Poisson equation has a direct role in performance evaluation since an approximation of the solution h will lead to an approximation of $J = \mu(c)$ using (9.17). With some structure imposed on the model this idea does lead to algorithms for computing bounds on J. For example, this is the essence of the main results in [35, 36], where performance bounds are obtained in the network scheduling problem. If the cost is linear, and if any of the linear programs constructed in these references admits a feasible solution, then the solution to Poisson's equation is approximated by a pure quadratic function.

Perhaps the most obvious approach to estimating J is through Monte-Carlo simulation via

$$\widehat{J}_n = \frac{1}{n} \sum_{0}^{n-1} c(x_t), \qquad n \in \mathbb{N}.$$

The Poisson equation again plays an important role in analysis, and in the generation of more efficient simulation approaches.

The effectiveness of the Monte Carlo method depends primarily on the magnitude of the Central Limit Theorem variance, also known as the time-average variance. Under suitably strong recurrence conditions on the Markov chain this can be expressed

$$\gamma_c^2 = \lim_{n\to\infty} \mathbb{E}_x \Big[\Big(\frac{1}{\sqrt{n}} \sum_{0}^{n-1} \bar{c}(x_t) \Big)^2 \Big] .$$

An alternative expression for the time-average variance is computed through the formula

$$\gamma_c^2 = \mu(h^2) - \mu((Ph)^2) = 2\mu(h\bar{c}) - \mu(\bar{c}^2), \tag{9.19}$$

with h any solution to Poisson's equation [39, eq. 17.50].

There are many variants of the simple Monte-Carlo estimate, some of which may have far smaller variance. After all, if $\{\Delta_t : t \geq 0\}$ is any sequence of random variables satisfying $\frac{1}{n}\sum_0^{n-1} \Delta_t \to 0$, $n \to \infty$, then the modified estimator,

$$\widehat{J}_n^{\Delta} = \frac{1}{n} \sum_{0}^{n-1} (c(x_t) + \Delta_t), \qquad n \in \mathbb{N},$$

is another consistent estimator of J. An *optimal* choice for Δ_t is computed using the solution h to Poisson's equation (9.16): by setting

$$\Delta_t^* = Ph\,(x_t) - h(x_t),$$

we obtain a time-average variance of *zero*. Of course, computing Δ_t^* involves a computation of J, so this approach is nonsensical! If however an approximation g to h can be found, then the choice $\Delta_t = Pg\,(x_t) - g(x_t)$ will lead to reduced variance if the approximation is sufficiently tight [24, 25].

We will discover such approximations when we attempt to solve some optimization problems below, and hence we will have many candidates for the function g.

9.3.3 Examples

In this chapter we develop two general examples: the linear state space model, and a family of network models. In this section we look at some special cases without control. Controlled linear systems, and controlled network models are considered as examples in the final section of this chapter.

The linear state space model is defined through the multi-dimensional recursion,

$$x_{t+1} = Ax_t + Fw_{t+1}, \qquad t \in \mathbb{N}, \tag{9.20}$$

where $x_t \in \mathbb{R}^d, w_t \in \mathbb{R}^q$, A is a $d \times d$ matrix, and F is a $d \times q$ matrix. We assume that $\boldsymbol{w}$ is i.i.d., and that $\boldsymbol{w}$ is a Gaussian process with mean zero, and covariance I. That is, $w_t \sim N(0, I)$, where I is the identity matrix.

The *controllability matrix* is the $d \times (dq)$ matrix

$$\mathcal{C} := [A^{d-1}F | A^{d-2}F | \cdots | AF | F],$$

where the bar denotes concatenation of matrices. The pair (A, F) is called *controllable* if the matrix $\mathcal{C}$ has rank d [37]. The process is ψ-irreducible with ψ equal to Lebesgue measure if the pair (A, F) is controllable. To see why, note that the state at time d can be written,

$$x_d = A^d x_0 + \mathcal{C} w_1^d$$

where $w_1^d = (w_1^{\mathrm{T}}, \ldots, w_d^{\mathrm{T}})^{\mathrm{T}}$. It follows that x_d itself is Gaussian with mean $A^d x_0$, and covariance given by

$$\Sigma_d = \mathcal{C}\mathcal{C}^{\mathrm{T}}.$$

The covariance is full rank if the model is controllable, and it follows that $P^t(x, \cdot\,)$ is equivalent to Lebesgue measure for any x, and any $t \geq d$. By continuity of the model it is easy to check that (9.7) holds with s continuous, and ν equal to normalized Lebesgue measure on an open ball in $\mathbb{R}^d$. We conclude that $\boldsymbol{x}$ is a T-chain if the controllability condition holds. A sample path from a particular two dimensional linear model is shown in Figure 9.1. When the state is large, the sample path behavior appears almost deterministic.

To find a stochastic Lyapunov function V with $c(x) = \frac{1}{2}x^T Q x$, first solve the *Lyapunov equation*,

$$A^T M A = M - Q. \tag{9.21}$$

If $M > 0$ (M is positive definite) then $V(x) = \frac{1}{2}x^T M x$ is a solution to (9.13). Direct calculations show that the function V is also the essentially unique solution to Poisson's equation, with $J = \frac{1}{2}\mathrm{trace}\,(F^T M F)$.

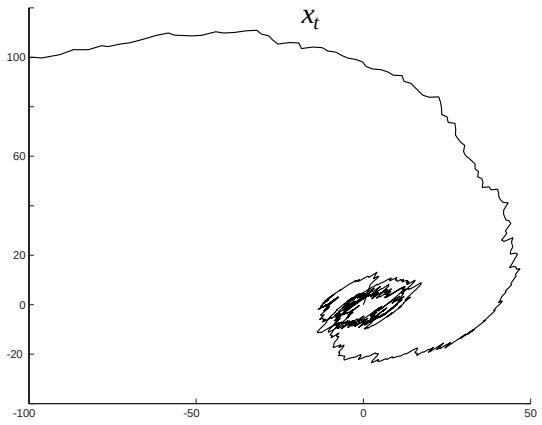

Figure 9.1 A sample path of the linear state space model with a 'large' initial condition $x_0 = \binom{-100}{100}$.

For the nonlinear state space model

$$x_{t+1} = F(x_t, w_{t+1}), \qquad t \in \mathbb{N},$$

the ψ-irreducibility condition can still be verified under a nonlinear controllability condition called *forward accessibility* [39, Chapter 7]. The construction of a Lyapunov function is however far more problem-specific.

Over the past five years there has been much research on algorithmic methods for constructing Lyapunov functions for *network models*. One is based upon linear programming methods, and is similar to the Lyapunov equation (9.21) used for linear state space models [35]. We describe here a recent approach based upon a fluid model [41, 42]. As an example we consider here the simplest case: An uncontrolled M/M/1 queue.

When the arrival stream is renewal, and the service times are i.i.d., then the waiting time for a simple queue can be modeled as a Markov chain with state space $\mathbb{X} = \mathbb{R}_+$. The dynamics take the form of a one dimensional linear state space model, where the state space is constrained to the positive half line. The queue length process is itself a Markov process in the special case where the service times and interarrival times are exponentially distributed. By applying *uniformization* (i.e. sampling the process appropriately - see [38]), the queue length process $\boldsymbol{x}$ obeys the recursion

$$x_{t+1} = x_t + (1 - I_{t+1})\pi(x_t) + I_{t+1}, \qquad t \in \mathbb{N},$$

where $\boldsymbol{I}$ is a Bernoulli, i.i.d. random process: $\lambda = \mathbb{P}(I_t = 1)$ is the arrival rate, and $\mu = \mathbb{P}(I_t = -1)$ is the service rate. The function π plays the role of a 'policy', where in this simple example we take $\pi(x) = \mathbf{1}(x > 0)$. Time has been normalized so that $\lambda + \mu = 1$.

To construct a Lyapunov function, first note that stability is a 'large state' property, so it may pay to consider the process starting from a large initial

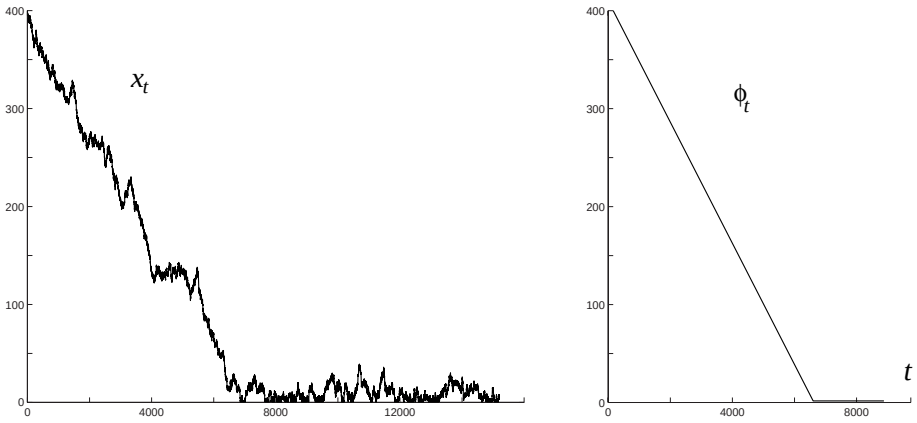

Figure 9.2 On the left is a sample path x_t of the M/M/1 queue with $\rho = \lambda/\mu = 0.9$, and $x_0 = 400$. On the right is a solution to the differential equation $\dot{\varphi} = (-\mu + \lambda)\pi(\varphi)$ starting from the same initial condition.

condition. In the left hand side of Figure 9.2 we see one such simulation. As was seen in the linear model, when the initial condition is large the behavior of the model is roughly deterministic.

Suppose we take the cost function $c(x) = 1 + x$. To construct a Lyapunov function we would ideally like to compute the expected sum given in (9.15), with S equal to some finite set, perhaps $S = \{0\}$. While this is computable for the M/M/1 queue, such computation can be formidable for more complex network models. However, consider the right hand side of Figure 9.2 which shows a sample path of the deterministic fluid, or leaky bucket model. This satisfies the differential equation $\dot{\varphi} = (-\mu + \lambda)\pi(\varphi)$, where π is again equal the indicator function of the strictly positive real axis. The behavior of the two processes look similar when viewed on this large spatial/temporal scale. It appears that a good approximation is

$$\begin{aligned} V(x) &:= \int_0^\infty \varphi(t)\, dt, \quad \varphi(0) = x, \\ &= \tfrac{1}{2}\frac{x^2}{\mu - \lambda}. \end{aligned} \tag{9.22}$$

If we apply the transition kernel P to V we find, for $x \geq 1$,

$$\begin{aligned} PV(x) &= \lambda V(x+1) + \mu V(x-1) \\ &= \frac{1}{2(\mu-\lambda)}\big(\lambda(x+1)^2 + \mu(x-1)^2\big) \\ &= V(x) - x + \frac{1}{2(\mu-\lambda)}, \end{aligned}$$

while for $x = 0$ we have,

$$PV(x) = \frac{\lambda}{2(\mu-\lambda)} \leq V(x) - x + \frac{1}{2(\mu-\lambda)}$$

That is, we see that this approach works: The stochastic Lyapunov criterion (9.13) does hold with this function V derived from the fluid model, where $\bar{J} = (2(\mu - \lambda))^{-1}$, under the stability condition that $\rho = \lambda/\mu < 1$. The actual steady state mean of $c(x) = x$ is given by $J = \lambda(\mu - \lambda)^{-1}$, which is indeed upper-bounded by $\bar{J}$.

What about the more exact Poisson's equation? Can the fluid model be used to approximate a solution?

With the cost function $c(x) = 1 + x$, the Poisson equation for the M/M/1 queue becomes

$$Ph = \lambda h(x+1) + \mu h((x-1)^+) = h(x) - c(x) + J.$$

One solution is given by

$$h(x) = \frac{x^2 + x}{2(\mu - \lambda)},$$

which is similar in form to the fluid value function given in (9.22).

For a general class of network models it can be shown that the value function for the fluid model and the solution to Poisson's equation are roughly equal for large x in the sense that

$$h(x) = V(x)(1 + o(1)),$$

where the term $o(1) \to 0$ as $x \to \infty$. Some results of this type are described in Section 9.6.2.

The M/M/1 queue illustrates the difficulties one faces in using simulation: Using (9.19) we can show that the time-average variance constant γ_c^2 is of order $(1-\rho)^{-4}$ in this example since pth moments for the M/M/1 queue are of order $(1-\rho)^{-p}$ [25].

With this background we are now ready to turn to MDP models.

9.4 THE AVERAGE COST OPTIMALITY EQUATION

We now assume that there is a control sequence taking values in the action space $\mathbb{A}$ which influences the behavior of $\boldsymbol{x}$. The state space $\mathbb{X}$ and the action space $\mathbb{A}$ are assumed to be locally compact, separable metric spaces, and we continue to let $\mathbb{F}$ denote the Borel σ-field on $\mathbb{X}$. Associated with each $x \in \mathbb{X}$ is a non-empty and closed subset $\mathbb{A}(x) \subseteq \mathbb{A}$ whose elements are admissible actions when the state process x_t takes the value x. The set of admissible state-action pairs $\{(x,a) : x \in \mathbb{X}, a \in \mathbb{A}(x)\}$ is assumed to be a measurable subset of the product space $\mathbb{X} \times \mathbb{A}$.

The transitions of $\boldsymbol{x}$ are governed by the conditional probability distributions $\{p(Y|x,a) : Y \in \mathbb{F}, x \in \mathbb{X}, a \in \mathbb{A}(x)\}$ which describe the probability that the next state is in Y, given that the current state is x, and the current action chosen is a. These are assumed to be probability measures on $\mathbb{F}$ for each state-action pair (x,a), and measurable functions of (x,a) for each $Y \in \mathbb{F}$.

We recall the following definitions:

Definition 9.6

(i) *A **nonrandomized policy** ϕ is a sequence of measurable functions ϕ_n, $n \in \mathbb{N}$, from H_n to $\mathbb{A}$ such that $\phi_n(x_0, a_0, \ldots x_{n-1}, a_{n-1}, x_n) \in \mathbb{A}(x_n)$.*

(ii) *If for each n this function ϕ_n depends only on x_n, then the policy ϕ is called **Markov**. The set of all Markov policies is denoted Π^M.*

(iii) *If ϕ is Markov, and if there is a fixed function π such that $\phi_n = \pi$ for all n, then the policy is called **stationary**. We denote by Π^S the set of all stationary policies.*

*For convenience, we extend the notation by writing $\pi \in \Pi^S$ when π is a measurable function from $\mathbb{X}$ to $\mathbb{A}$ which defines a stationary policy ϕ. The function π is called a **feedback law**.*

For any $\pi \in \Pi^S$, the state process $\boldsymbol{x}^\pi := \{x_t^\pi : t \geq 0\}$ is a Markov chain on $(\mathbb{X}, \mathbb{F})$ with stationary transition probabilities.

We do not consider randomized policies. This is without loss of generality since we can always redefine the MDP model so that the action space $\mathbb{A}$ is replaced with the *space of probability measures on* $\mathbb{A}$. An ordinary policy for the new MDP model is equivalent to a randomized one for the original model.

We shall write

$$P_\pi^t f = \int_{\mathbb{X}} f(y) p^t(dy|x, \pi(x)), \qquad f \in L_\infty, t \geq 1,$$

for the semigroup of kernels corresponding to a policy $\pi \in \Pi^S$, and we let K_π denote the corresponding resolvent kernel. We continue to use the operator-theoretic notation,

$$P_\pi^t h\,(x) := \mathbb{E}^\pi[h(x_t^\pi) \mid x_0 = x].$$

In the remainder of this section we describe consequences of the average cost optimality equation, and develop criteria for existence of solutions. These results are based on [42, 44]. More background may be found in [29, 5, 28, 1, 49].

9.4.1 Regular and optimal policies

We suppose that a one-step *cost function* $c\colon \mathbb{X} \times \mathbb{A} \to [1, \infty)$ is given. Other chapters in this volume consider a reward function r. Throughout this chapter we will take $c = -r$, so that the optimization problem becomes one of *cost minimization*. We assume below that c satisfies a near-monotone condition so that the results of Section 9.2 and 9.3 may be applied. We will use freely terminology that was introduced in these sections. In particular, we refer the reader to Definition 9.5 for a definition of near-monotonicity, and Definition 9.4 for c-regularity and related topics.

The steady state average cost is denoted

$$J(\phi, x) := \limsup_{n \to \infty} \frac{1}{n}\, \mathbb{E}_x \Big[\sum_{t=0}^{n-1} c(x_t^\phi, \phi_t) \Big].$$

For any $\phi \in \Pi^S$, defined by a state feedback law π, we define $c_\pi \colon \mathbb{X} \to \mathbb{R}$ as the function given by $c_\pi(x) = c(x, \pi(x))$, $x \in \mathbb{X}$. A policy ϕ will be called

Definition 9.7

regular *if $\phi \in \Pi^S$, and $\boldsymbol{x}^\phi$ is a c_π-regular Markov chain.*

s-optimal *if $\phi \in \Pi^S$ and*

$$J(\phi, x) \le J(\phi', x), \qquad \phi' \in \Pi^S, x \in \mathbb{X}.$$

m-optimal *if $\phi \in \Pi^M$ and*

$$J(\phi, x) \le J(\phi', x), \qquad \phi' \in \Pi^M, x \in \mathbb{X}.$$

Many of the results below concern conditions which guarantee the existence of a regular, s-optimal policy π_*. In this case $J_* = J(\pi_*, x)$ is independent of x.

When J_* is independent of x, and the optimization criterion is *cost minimization*, the associated *average cost optimality equation* (ACOE) is given as follows. The function h_* is known as the *relative value function.*.

$$J_* + h_*(x) = \min_{a \in \mathbb{A}(x)} [c(x,a) + P_a h_*(x)] \tag{9.23}$$

$$\pi_*(x) = \arg\min_{a \in \mathbb{A}(x)} [c(x,a) + P_a h_*(x)], \qquad x \in \mathbb{X}. \tag{9.24}$$

If a stationary policy π_*, a measurable function h_*, and a constant J_* exist which solve (9.23,9.24), then typically the policy π_* is optimal (see for example [1, 5, 27, 49, 55] for a proof of this and related results).

Theorem 9.5 *Suppose that (J_*, h_*, π_*) solve (9.23,9.24). Assume moreover that, for any $x \in \mathbb{X}$, and any $\pi \in \Pi^S$ satisfying $J(\pi, x) < \infty$,*

$$\frac{1}{n} P_\pi^n h_*(x) \to 0, \qquad n \to \infty. \tag{9.25}$$

Then π_* is an s-optimal control, and J_* is the optimal cost, in the sense that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_x[c_{\pi_*}(x_t^{\pi_*})] = J_*,$$

and $J(\pi, x) \ge J_*$ for all stationary policies π, and all initial states x. ∎

The assumption (9.25) is unfortunate, but examples show that some additional conditions on h_* are required (see e.g. page 87 of [5], Chapter 7 of [55], or the examples in [1, 15, 49, 51]). The following result gives a condition implying (9.25) which is often verifiable in practice, as we shall see in Section 9.6.1 and Section 9.6.2.

Suppose that $\phi \in \Pi^S$ is defined by a feedback law π, suppose that the controlled chain $\boldsymbol{x}^\pi$ is ψ_π-irreducible, and that $\mu_\pi(c_\pi) = \int c_\pi(x) \mu_\pi(dx)$ is finite.

Let $Z_\pi \in \mathbb{F}$ denote any fixed c_π-regular set for which $\mu_\pi(Z_\pi) > 0$. We then define

$$V_\pi(x) = \mathbb{E}_x\Big[\sum_{t=0}^{\tau_\pi - 1} c_\pi(x_t^\pi)\Big], \qquad x \in \mathbb{X}, \tag{9.26}$$

where $\tau_\pi = \tau_{Z_\pi}$ is the first entrance time to Z_π. Since $\mu_\pi(Z_\pi) > 0$, the function V_π is μ_π-a.e. finite-valued [39, Theorem 14.2.5]. Note that, by [39, Theorem 14.2.3], the particular c_π-regular set Z_π chosen is not important. If Z_π^1 and Z_π^2 give rise to functions V_π^1 and V_π^2 of the form (9.26), then for some constant $\gamma \geq 1$,

$$\gamma^{-1} V_\pi^1(x) \leq V_\pi^2(x) \leq \gamma V_\pi^1(x), \qquad x \in \mathbb{X}.$$

The following result is taken from [42]. We show how the assumptions are verified in Section 9.6.

Theorem 9.6 *Suppose that the optimality equation (9.23,9.24) holds for* (J_*, h_*, π_*)*, with* h_* *bounded from below. For any other* $\pi \in \Pi^S$ *assume that*

(a) *The function*

$$K_\pi c_\pi = (1-\beta)\sum_{t=0}^{\infty} \beta^t P_\pi^t c_\pi$$

is norm-like, and the Markov chain $\boldsymbol{x}^\pi$ *is a T-chain.*

(b) *There exists some constant* $d_0 = d_0(\pi) < \infty$ *such that,*

$$|h_*(x)| \leq d_0 V_\pi(x), \qquad x \in \mathbb{X}. \tag{9.27}$$

Then π_* *is a regular, s-optimal policy.* ∎

Theorem 9.6 asserts that a solution to the ACOE will yield an optimal policy provided h_* is 'small enough'. This motivates the construction of a solution through the formulation of a *minimal* relative value function. We consider this approach next.

9.4.2 Existence of solutions

The ACOE is a generalization of Poisson's equation. To formulate sufficient conditions for a solution, we generalize the operators H and K which formed the basis of analysis in Section 9.3. In an attempt to simplify the development, we assume here that the cost $c(\,\cdot\,,\,\cdot\,)$ is a function of x only. Analogous results hold for general cost functions.

Randomization is required to define the resolvent kernel for a Markov policy. Let $\{\xi_i : i \geq 1\}$ denote an i.i.d. sequence of geometrically distributed random variables which is independent of the controlled chain. We assume that for some $0 < \beta < 1$,

$$\mathbb{P}(\xi_i = k) = (1-\beta)\beta^k, \qquad k \geq 0,\ i \geq 1.$$

For any Markov policy, we then write,

$$K_\phi(x, Y) = \mathbb{P}^\phi\{x_{\xi_1} \in Y \mid x_0 = x\}, \qquad x \in \mathbb{X},\ Y \in \mathbb{F}.$$

This coincides with the previous definition (9.3) when the policy ϕ is stationary.

A potential kernel is then defined in analogy with (9.5) by sampling the chain at the renewal events $t_k = \sum_{i=1}^k \xi_i$. Let $\boldsymbol{y}$ denote the sampled, controlled chain $y_k = x_{t_k}$, $k \geq 0$. A Markov policy for $\boldsymbol{y}$ is expressed as a *sequence* of Markov policies for $\boldsymbol{x}$, via

$$\boldsymbol{\phi} := (\phi_1, \phi_2, \dots), \qquad \phi_i \in \Pi^M, i \geq 1.$$

We let $\boldsymbol{\Pi}^M$ denote the set of all such sequences, and we let $\boldsymbol{\Pi}^S \subset \boldsymbol{\Pi}^M$ denote the set of trivial sequences constructed from a *stationary* policy. That is, $\boldsymbol{\phi} \in \boldsymbol{\Pi}^S$ if $\boldsymbol{\phi} = (\phi, \phi, \dots)$, with $\phi \in \Pi^S$.

Given a policy $\boldsymbol{\phi} \in \boldsymbol{\Pi}^M$, the transition probabilities are given by

$$\mathbb{P}(y_{k+\ell} \in Y \mid y_k = x) = [K_{\phi_{k+1}} \cdots K_{\phi_{k+\ell}}](x, Y), \quad x \in \mathbb{X},\ Y \in \mathbb{F},\ k \geq 0,\ \ell \geq 1.$$

For a fixed $s \colon \mathbb{X} \to]0, 1[$ and a probability measure ν on $\mathbb{F}$ we define

$$\begin{aligned} M_n^{\boldsymbol{\phi}} &:= (K_{\phi_1} - s \otimes \nu) \cdots (K_{\phi_n} - s \otimes \nu), \qquad n \geq 1. \\ H^{\boldsymbol{\phi}} &:= I + \sum_{n=1}^{\infty} M_n^{\boldsymbol{\phi}}. \end{aligned}$$

This again agrees with our earlier definition of H when $\boldsymbol{\phi} \in \boldsymbol{\Pi}^S$. To ensure positivity of these kernels we impose a minorization condition below.

These operators immediately give formulae for a candidate solution to the ACOE. We first define a candidate average cost:

$$\eta(\boldsymbol{\phi}) := \inf\Big(\eta : \nu H^{\boldsymbol{\phi}}(c - \eta) \leq 0\Big), \qquad \boldsymbol{\phi} \in \boldsymbol{\Pi}^M; \tag{9.28}$$

$$\eta_* := \inf\Big(\eta(\boldsymbol{\phi}) : \boldsymbol{\phi} \in \boldsymbol{\Pi}^M\Big). \tag{9.29}$$

The construction of a solution to Poisson's equation in Theorem 9.4 leads to the following definitions: For any $x \in \mathbb{X}$,

$$h_*^0(x) := \inf\Big(H^{\boldsymbol{\phi}}(c - \eta_*)(x) : \boldsymbol{\phi} \in \boldsymbol{\Pi}^M\Big). \tag{9.30}$$

$$h_*(x) := \Big(\frac{\beta}{1-\beta}\Big) \inf\Big(K_\phi h_*^0(x) : \phi \in \Pi^M\Big). \tag{9.31}$$

A candidate s-optimal policy is then,

$$\pi_*(x) := \arg\min_{a \in \mathbb{A}} \int_{\mathbb{X}} p(dy \mid x, a) h_*(y), \qquad x \in \mathbb{X}. \tag{9.32}$$

The following assumptions ensure that these functions are well defined, and that the function h_* is bounded from below. Provided these assumptions hold,

and that there exists at least one 'stabilizing policy', we find that the triple (J_*, h_*, π_*) solves the ACOE with $J_* = \eta_*$.

(A1) The infimums in (9.30,9.31) exist, and admit measurable solutions h_*^0 and h_*. Moreover, the minimum in (9.32) exists point-wise to form a stationary (measurable) policy π_*.

(A2) There exists a norm-like function $\underline{c}\colon \mathbb{X} \to \mathbb{R}_+$ such that,

$$K_\phi c \geq \underline{c}, \qquad \phi \in \Pi^M.$$

(A3) There exists a continuous function $s\colon \mathbb{X} \to]0,1[$, and a probability measure ν on $\mathbb{F}$, such that

$$K_\phi(x, Y) \geq s(x)\nu(Y), \qquad x \in \mathbb{X},\ Y \in \mathbb{F},\ \phi \in \Pi^M.$$

The measurability assumption in (A1) is, surprisingly, the most subtle of these three conditions. A strong Feller assumption, that $p(h \mid x, a)$ is a continuous function of (x, a) for a sufficiently large class of functions h will imply that h_* is continuous, and hence measurable. The existence of a measurable solution π_* in (9.32) will require further conditions (such as compactness of $\mathbb{A}(x)$ for all x).

Lemma 9.2 *If (A1)-(A3) hold then, whenever the invariant probability μ_π exists,*

$$\eta_* \leq \mu_\pi(c), \qquad \pi \in \Pi^S.$$

Proof. This follows from the construction of $\mu(\,\cdot\,) = \mu_\circ(\,\cdot\,)/\mu_\circ(\mathbb{X})$ given in (9.10), the definition (9.29), and the minorization condition (A3). ∎

Thus, it is not surprising that π_* is an s-optimal policy:

Theorem 9.7 *Suppose that Assumptions (A1)-(A3) are satisfied, and suppose that there exists one regular policy $\pi_0 \in \Pi^S$ with average cost $\mu_{\pi_0}(c) < \infty$.*

Then the following hold:

(a) *The triple (J_*, h_*, π_*) solve the ACOE, where $J_* = \eta_*$, and the policy π_* is regular and s-optimal.*

(b) *The function h_* is uniformly bounded from below:*

$$\inf_{x \in \mathbb{X}} h_*(x) > -\infty.$$

(c) *If (h'_*, J_*) is any other solution to (9.23), with h_* uniformly bounded from below, then there exists a constant k' such that*

$$h'_*(x) \begin{cases} = h_*(x) + k' & \text{for almost every } x \\ \geq h_*(x) + k' & \text{for every } x. \end{cases}$$

Proof. We will just prove (a). This and the remaining parts are similar to the proof of Theorem 9.4. A complete proof in the countable state space setting is given in [44].

We first show that h_*^0 solves the ACOE for the resolvent. From the definition of η_* we find that

$$\nu(h_*^0) \le 0. \tag{9.33}$$

For any $\phi \in \Pi^M$, $\boldsymbol{\phi} \in \boldsymbol{\Pi}^M$, we can apply (9.33) and the definition of h_*^0 to give,

$$\begin{aligned} K_\phi h_*^0 &\le (K_\phi - s \otimes \nu) h_*^0 \\ &\le (K_\phi - s \otimes \nu) H^{\boldsymbol{\phi}}(c - \eta_*) \\ &= H^{\boldsymbol{\phi}^{[1]}}(c - \eta_*) - c + \eta_* \end{aligned}$$

where $\boldsymbol{\phi}^{[1]} := (\phi, \phi_1, \phi_2, \dots)$.

Infimizing over all $\phi \in \Pi^M$, $\boldsymbol{\phi} \in \boldsymbol{\Pi}^M$ gives the upper bound,

$$\inf_\phi \Big(K_\phi h_*^0(x) \Big) \le h_*^0(x) - c(x) + \eta_*, \qquad x \in \mathbb{X}.$$

We also know that h_*^0 is finite-valued by the regularity assumption imposed on π_0: With $\phi_0 \in \Pi^S$ equal to the stationary policy defined by π_0, the following bound follows from minimality of h_*^0,

$$h_*^0 \le H^{\boldsymbol{\phi}^0}(c - \eta_*), \qquad \boldsymbol{\phi}^0 := (\phi_0, \phi_0, \dots) \in \boldsymbol{\Pi}^S.$$

We now turn to the function h_*. Exactly as in the uncontrolled case (see the proof of Theorem 9.4), we can translate from the resolvent to the original chain to obtain,

$$\begin{aligned} P_{\pi_*} h_*(x) &:= \inf_{a \in \mathbb{A}(x)} P_a \Big(\inf_{\phi \in \Pi^M} K_\phi h_*^0(x) \Big) \\ &\le h_*(x) - c(x) + \eta_*, \qquad x \in \mathbb{X}. \end{aligned}$$

It follows that $\mu_{\pi_*}(c) \le \eta_*$, and then by minimality of η_* we must have equality. Hence by Theorem 9.4, the above is an equality for μ_{π_*}-a.e. x. By minimality of h_* and Theorem 9.4 (iv), it must be an equality for all x. ∎

The literature on average cost optimal control is filled with counter-examples. It is of some interest then to see why Theorem 9.7 does not fall into any of these traps. Consider first counter-examples 1 and 2 of [51, p. 142]. In each of these examples the MDP is completely non-irreducible in the sense that

$$\mathbb{P}(x_t^\pi < x_0^\pi) = 0, \qquad t \ge 1, \pi \in \Pi^S.$$

It is clear then from the cost structure that the bound (A3) on the resolvent cannot hold in this case.

Another example is given in the Appendix of [51] in which a version of (A3) is directly assumed! However, the cost is not unbounded, and is in fact designed to favor large states.

The assumptions (A2) and (A3) together imply that the center of the state space, as measured by the cost criterion, possesses some minimal amount of irreducibility. If either the unboundedness condition or the accessibility condition is relaxed, so that the process is non-irreducible on a set where the cost is low, then we see from these counter-examples that optimal stationary policies may not exist.

In the remainder of this chapter we preserve these three assumptions. They will be generalized sightly when we consider algorithms in the next section.

9.5 ALGORITHMS

Value iteration and policy iteration are two well-known algorithms for constructing optimal policies. The value iteration algorithm, or VIA, is a version of successive approximation. The policy iteration algorithm, or PIA, first proposed in [31], may be interpreted as a version of the Newton-Raphson method. We find that the PIA is more easily analyzed under the assumptions we impose even though the algorithm is considerably more complex than value iteration. The ease of analysis is a result of the hard work already taken care of in Section 9.3.

Although complex, the PIA may converge extremely quickly when properly normalized. See [20, 42] for application in the communication and network areas.

The results below are taken from [42, 10]. Related work on algorithms may be found in [30, 7, 54, 8].

9.5.1 Value iteration

The ACOE (9.23) can be viewed as a fixed point equation in the variables (h_*, J_*). By ignoring the constant term, and applying successive approximation to this fixed point equation, we obtain the VIA. Suppose that the positive-valued function V_n is given. Then the stationary policy π_n is defined as

$$\pi_n(x) = \arg\min_{a \in \mathbb{A}(x)} [P_a V_n(x) + c(x,a)], \qquad x \in \mathbb{X},$$

and one then defines

$$V_{n+1}(x) = c_{\pi_n}(x) + P_{\pi_n} V_n(x) = \min_{a \in \mathbb{A}(x)} (P_a V_n(x) + c_a(x)), \tag{9.34}$$

which then makes it possible to compute the next policy π_{n+1} by re-starting the algorithm.

This is in fact the standard dynamic programming approach to constructing a finite horizon optimal policy since for each n we may write,

$$V_n(x) = \min\Big(\mathbb{E}_x^{\phi} \Big[\sum_{t=0}^{n-1} c(x_t, a_t) + V_0(\Phi(n)) \Big] : \phi \in \Pi^M \Big). \tag{9.35}$$

We see in (9.35) that the initial condition V_0 plays the role of a terminal penalty function.

The initialization V_0 should be chosen with care. For a countable state space model, a poor choice (such as $V_0 \equiv 0$) can lead to policies for which the controlled chain is transient [10]. We assume in Theorem 9.8 below that at least one regular policy π_{-1} exists, and that the function V_0 serves as a Lyapunov function: for some constant $\bar{J} < \infty$,

$$P_{\pi_{-1}} V_0 \le V_0 - c_{\pi_{-1}} + \bar{J}. \tag{9.36}$$

The existence of a pair (V_0, π_{-1}) satisfying (9.36) is a natural stabilizability assumption on the model, and we find below that this initialization ensures that the VIA generates stabilizing policies.

To simplify notation we define $c_n = c_{\pi_n}$, $P_n = P_{\pi_n}$, and we define the resolvent for the nth policy by

$$K_n := (1-\beta) \sum_{t=0}^{\infty} \beta^t P_n^t, \qquad n \ge 0, \tag{9.37}$$

where $\beta \in]0,1[$ as before. We let $\mathbb{E}^n$ denote the expectation operator induced by the stationary policy π_n.

Let ν denote some fixed probability measure on $\mathbb{F}$. We define, for each n, the normalized value function, and the incremental cost,

$$h_n(x) = V_n(x) - \nu(V_n); \qquad \gamma_n(x) = V_{n+1}(x) - V_n(x), \qquad x \in \mathbb{X}, n \in \mathbb{N}. \tag{9.38}$$

From the definitions, for each n we have the familiar looking identity $P_n h_n = h_n - c_n + \gamma_n$.

Defining $\bar{J}_n = \sup_x \gamma_n(x) \le \infty$, we obtain the following solution to (9.13):

$$P_n V_n \le V_n - c_n + \bar{J}_n. \tag{9.39}$$

Under (A1)-(A3), and with an initial condition satisfying (9.36), we find that the $\{\bar{J}_n\}$ are finite valued, and non-increasing. The assumptions below are almost identical to (A1)-(A3) in Section 9.4.

(VIA1) For each n, if the VIA yields a value function $V_n \colon \mathbb{X} \to \mathbb{R}_+$, then for each $x \in \mathbb{X}$ the minimization

$$\pi_n(x) := \arg\min_{a \in \mathbb{A}(x)} \big[c(x,a) + P_a V_n(x) \big]$$

exists, and admits a measurable solution π_n.

(VIA2) There exists a norm-like function $\underline{c} \colon \mathbb{X} \to \mathbb{R}_+$ such that

$$K_n c_n(x) \ge \underline{c}(x), \qquad x \in \mathbb{X}, n \in \mathbb{N}.$$

(VIA3) There is a fixed probability ν on $\mathbb{F}$, a $\delta > 0$, and an initial value function V_0 with the following property: For each $n \ge 1$, if the VIA yields the value function V_n, then for any policy π_n given in (VIA1),

$$K_n(x, Y) \ge \delta \nu(Y) \qquad x \in S, Y \in \mathbb{F}, \tag{9.40}$$

where S denotes the pre-compact set

$$S = \{x : \underline{c}(x) \le 2\bar{J}\}. \tag{9.41}$$

The following result is largely taken from [10].

Theorem 9.8 *Suppose that (VIA1)–(VIA3) hold. Assume moreover that the initialization V_0 satisfies (9.36). Then,*

(i) *Each of the policies $\{\pi_i : i \in \mathbb{N}\}$ is regular.*

(ii) *The upper bounds $\{\bar{J}_n\}$ are decreasing:*

$$\bar{J}_0 \ge \bar{J}_1 \ge \cdots \ge \bar{J}_n \ge \cdots ;$$

(iii) *The sequence $\{h_n\}$ is uniformly bounded from below.*

Proof. The minimization in the value iteration algorithm immediately leads to the bound $P_n\gamma_n \ge \gamma_{n+1}$. From this we deduce by induction that the $\bar{J}_n$ are finite and decreasing: The initialization of the induction relies on the assumption that the initial condition V_0 satisfies (9.36). This then proves (ii).

To establish (i), note first that the following bound on the resolvent follows from (9.39):

$$K_nV_n \le V_n - \frac{\beta}{1-\beta}K_nc_n + \frac{\beta}{1-\beta}\bar{J}_n. \tag{9.42}$$

This inequality is a version of (9.13) since $V_n \ge 0$, and we have established that $\bar{J}_n$ is finite. Applying Theorem 9.3 and using (VIA2,VIA3) for the kernel K_n, we see that the Markov chain with transition kernel K_n is c-regular. This implies (i).

To prove (iii) note first of all that $h_n(x) \ge -\nu(V_n) > -\infty$ for all x. It remains to obtain a bound independent of n. For any n we have

$$K_nh_n \le h_n - K_nc_n + \bar{J} \le h_n + \bar{J}\mathbf{1}_S$$

Letting $s = \delta\mathbf{1}_S$ we then obtain,

$$(K_n - s\otimes\nu)h_n \le h_n + \bar{J}\delta^{-1}s,$$

and by iteration, for any N,

$$-\nu(V_n)(K_n - s\otimes\nu)^N\mathbf{1} \le (K_n - s\otimes\nu)^Nh_n \le h_n + \bar{J}\delta^{-1}\sum_{i=0}^{N-1}(K_n - s\otimes\nu)^is.$$

By c_n-regularity of the nth chain it follows that $(K_n - s\otimes\nu)^N\mathbf{1}(x) \to 0$ as $N\to\infty$ for any x. This and Lemma 9.1 then gives the bound

$$0 \le h_n + \bar{J}\delta^{-1}\sum_{i=0}^{\infty}(K_n - s\otimes\nu)^is \le h_n + \bar{J}\delta^{-1}.$$

■

Convergence of the algorithm is subtle. This is not surprising since it is rare in optimization to prove global convergence of successive approximation. The countable state space case is considered in [10] where it is shown that (VIA1), (VIA2), and a strengthening of (VIA3) do imply convergence of $\{h_n\}$ to a solution of the ACOE. To generalize this result to general state spaces it may be necessary to impose a blanket stability condition as in [29], or the stronger stability assumption imposed in [14, 56].

9.5.2 Policy iteration

The PIA, which is again a recursive algorithm for generating stationary policies, follows naturally as a refinement of the VIA. We saw that the value iteration algorithm generates regular policies because we have established in Section 9.5.1 the drift inequality,

$$P_{n-1}V_{n-1} \le V_{n-1} - c_{n-1} + \bar{J}_{n-1}.$$

From this bound we discovered easily that the next policy π_n has cost bounded by $J(\pi_n, x) \le \bar{J}_{n-1}$, $x \in \mathbb{X}$. We have seen that there are an infinite number of solutions to drift inequalities of this form, and some give better bounds than others. The *optimal* solution is the solution to Poisson's equation, since this gives the minimal possible value for $\bar{J}$. On replacing the function V_{n-1} by the solution to Poisson's equation in the VIA recursion (9.34) one obtains precisely the PIA.

To give a precise description of the algorithm, suppose that at the $(n-1)$th stage of the algorithm a stationary policy π_{n-1} is given, and assume that h_{n-1} satisfies the Poisson equation

$$P_{n-1}h_{n-1} = h_{n-1} - c_{n-1} + J_{n-1},$$

where $P_{n-1} = P_{\pi_{n-1}}$, $c_{n-1}(x) = c_{\pi_{n-1}}(x) = c(x, \pi_{n-1}(x))$, and J_{n-1} is a constant (equal to the steady state cost with this policy).

Given h_{n-1}, one then attempts to find an improved stationary policy π_n by choosing, for each x,

$$\pi_n(x) = \arg\min_{a \in \mathbb{A}(x)} \big[c(x,a) + P_a h_{n-1}(x) \big]. \tag{9.43}$$

Once π_n is found, stationary polices $\pi_{n+1}, \pi_{n+2}, \dots$ may be computed by induction, so long as the appropriate Poisson equation may be solved, and the minimization above has a solution.

Our analysis of the PIA is based on the pair of equations

$$P_n h_n = h_n - \bar{c}_n; \tag{9.44}$$

$$P_n h_{n-1} = h_{n-1} - \bar{c}_n + \gamma_n, \tag{9.45}$$

where $\bar{c}_n = c_n - J_n$, and γ_n is now *defined* through (9.45). From the minimization (9.43) we have

$$c_n + P_n h_{n-1} \le c_{n-1} + P_{n-1} h_{n-1},$$

and from Poisson's equation we have

$$c_{n-1} + P_{n-1}h_{n-1} = h_{n-1} + J_{n-1}.$$

Combining these two equations gives the upper bound $\gamma_n(x) \le J_{n-1} - J_n$, $x \in \mathbb{X}$, which shows that the PIA automatically generates solutions to (9.13).

As was the case with the value iteration algorithm, much of the analysis of [42] focuses on $\{K_n\}$ rather than $\{P_n\}$, as given in (9.37). To invoke the algorithm we must again ensure that the required minimum exists.

(PIA1) For each n, if the PIA yields a triplet $(J_{n-1}, h_{n-1}, \pi_{n-1})$ which solve Poisson's equation

$$P_{n-1}h_{n-1} = h_{n-1} - c_{n-1} + J_{n-1},$$

with h_{n-1} bounded from below, then for each $x \in \mathbb{X}$ the minimization

$$\pi_n(x) := \arg\min_{a \in \mathbb{A}(x)} \big[c(x,a) + P_a h_{n-1}\,(x)\big]$$

exists, and admits a measurable solution π_n.

(PIA2) There exists a norm-like function $\underline{c}\colon \mathbb{X} \to \mathbb{R}_+$ such that for the policies π_n obtained through the PIA,

$$K_n c_n\,(x) \ge \underline{c}(x), \qquad x \in \mathbb{X}, n \in \mathbb{N}.$$

(PIA3) There is a fixed probability ν on $\mathbb{F}$, a $\delta > 0$, and an initial regular policy π_0 with the following property: For each $n \ge 1$, if the PIA yields a triplet $(J_{n-1}, h_{n-1}, \pi_{n-1})$ with h_{n-1} bounded from below, then for any policy π_n given in (PIA1),

$$K_n(x, Y) \ge \delta\nu(Y) \qquad x \in S, Y \in \mathbb{F}, \tag{9.46}$$

where S denotes the pre-compact set

$$S = \{x : \underline{c}(x) \le 2J_0\}. \tag{9.47}$$

Under Assumptions (PIA1)-(PIA3), the algorithm produces stabilizing policies recursively. A proof of Theorem 9.9 may be found in [42].

Theorem 9.9 *Suppose that (PIA1)-(PIA3) hold, and that the initial policy π_0 is regular. Then for each n the PIA admits a solution (J_n, h_n, π_n) such that π_n is regular, and the sequence of relative value functions $\{h_n\}$ defined in (9.18) satisfy,*

(i) *For some constant $N < \infty$,*

$$\inf_{x \in \mathbb{X}, n \ge 0} h_n(x) > -N;$$

(ii) *There exists $h_\infty\colon \mathbb{X} \to \mathbb{R}$ such that*

$$\lim_{n\to\infty} h_n(x) = h_\infty(x), \qquad x \in \mathbb{X};$$

(iii) *There exists* $b_1, b_2 < \infty$ *such that*

$$-N \le h_\infty(x) \le b_1 h_0(x) + b_2, \qquad x \in \mathbb{X}.$$

■

Now that we know that $\{h_n\}$ is point-wise convergent to a function h_∞, we can show that the PIA yields a solution to the ACOE. We let π_∞ denote a solution to

$$\pi_\infty(x) = \arg\min_{a \in \mathbb{A}(x)} P_a h\,(x), \qquad x \in \mathbb{X}. \tag{9.48}$$

Theorem 9.10 is similar to Theorem 4.3 of [27] which requires a related continuity condition. Weaker conditions are surely possible for a specific application.

Theorem 9.10 *Suppose that (PIA1)-(PIA3) hold, and that the initial policy* π_0 *is regular. Assume in addition that*

(i) *The function* π_∞ *in (9.48) can be chosen to form a stationary policy.*

(ii) *The function* $c\colon \mathbb{X}\times\mathbb{A} \to [1,\infty)$ *is continuous, and the functions* $(P_a h_n\,(x) : n \ge 0)$ *and* $P_a h\,(x)$ *are continuous in* (a, x).

(iii) *For each* $x \in \mathbb{X}$, *the function* $c(x, \cdot\,)$ *is norm-like on* $\mathbb{A}$.

(iv) *The initial condition* h_0 *satisfies,*

$$\lim_{n\to\infty} \frac{1}{n} P_\pi^n h_0\,(x) = 0,$$

for any $\pi \in \Pi^S$, *and any* $x \in \mathbb{X}$ *for which* $J(\pi, x) < \infty$.

Then,

(a) *The PIA produces a sequence of solutions* (J_n, h_n, π_n) *such that* $\{J_n, h_n\}$ *is point-wise convergent to* (J_∞, h_∞). *The triple* $(J_\infty, h_\infty, \pi_\infty)$ *is a solution to the ACOE.*

(b) *The policy* π_∞ *is* c_{π_∞}*-regular, and s-optimal. Consequently, for any initial condition* $x \in \mathbb{X}$,

$$J(\pi_\infty, x) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_x^{\pi_\infty}[c_\pi(x_t^{\pi_\infty})] = \mu_{\pi_\infty}(c_\pi) = J_\infty.$$

■

We now illustrate the theory with some general examples.

9.6 EXAMPLES

9.6.1 Linear models

The controlled linear state space model is defined through the recursion,

$$x_{t+1} = Ax_t + Ba_t + Fw_{t+1}, \qquad t \in \mathbb{N}, \tag{9.49}$$

where $w_t \in \mathbb{R}^q$, $x_t \in \mathbb{R}^d$, and $a_t \in \mathbb{R}^p$. As before we assume that $w \sim N(0, I)$, and that (A, F) is controllable (see the discussion following (9.20)).

Since $\boldsymbol{w}$ is i.i.d., then this is a Markov decision process with transition function

$$p(Y \mid x, a) = \mathbb{P}(w_1 + Ax + Ba \in Y), \qquad x \in \mathbb{X}, Y \in \mathbb{F}, a \in \mathbb{A}.$$

The cost c is taken as the general quadratic,

$$c(x, a) = \tfrac{1}{2} x^T Q x + \tfrac{1}{2} a^T R a, \tag{9.50}$$

with $Q \geq 0$, and $R > 0$. The assumption $c \geq 1$ fails in this example. However, c is positive, so that we can add 1 to the cost function to satisfy the desired lower bound on c and the MDP is essentially unchanged.

The optimization of $J(\pi, x)$ is known as the LQG (linear-quadratic-Gaussian) problem. Under certain conditions on the model, it is known that one may obtain a solution (J_*, h_*, π_*) to the ACOE with h_* quadratic, and π_* linear in x, by solving a Riccati equation [37]. What conditions are required? Why should a solution give rise to an optimal policy?

Assumption (A1) in Section 9.4 requires the existence of measurable solutions to the static optimization problem arising in (9.23,9.24). Since the candidate relative value function is quadratic, this will hold under our assumption that $R > 0$.

The norm-like condition (A2) requires additional assumptions on Q. Let $\sqrt{Q}$ denote any $d \times d$ matrix for which $Q = \sqrt{Q}^{\mathrm{T}} \sqrt{Q}$, and suppose that $(A, \sqrt{Q})$ is *observable*. Algebraically, this means that $(A^T, \sqrt{Q}^T)$ is a *controllable* pair. Physically, it means that the cost will be large whenever the state is large. In [42] it is shown that observability implies (A2), and it is also shown that for any regular policy, the solution to Poisson's equation is bounded from below by a quadratic function of x.

Assumption (A3) will not hold for *any* stationary policy, but if one restricts to policies with bounded growth, say

$$\|\pi(x)\| \leq b_1(\|x\|^2 + 1) \qquad x \in \mathbb{X},$$

then this assumption will hold if F is a $d \times d$ matrix with rank d. This is stronger than the controllability assumption.

Under these conditions it follows from Theorem 9.6 that the linear/quadratic solution (π_*, h_*) to the ACOE does yield an optimal control over the class of all nonlinear feedback laws (i.e. all stationary policies).

Theorem 9.9 recovers known properties of the Newton-Raphson technique applied to the LQG problem, which is precisely the PIA. Suppose the initial

policy π_0 in the PIA is linear. One can verify that the solution to Poisson's equation is quadratic. Each subsequent policy is of the form $\pi_n(x) = -K_n x$, for some $p \times n$ matrix K_n, and each subsequent solution to Poisson's equation is quadratic,

$$h_n(x) = h_n(0) + \tfrac{1}{2} x^{\mathrm{T}} \Lambda_n x, \qquad x \in \mathbb{X}.$$

Under the observability condition, it follows from Theorem 9.9 that the matrices $\{\Lambda_n\}$, which are solutions to a *Riccati recursion*, are uniformly bounded in n.

The proof of Theorem 9.9 depends upon a bound of the form,

$$h_n(x) \le [1 + 2(J_{n-1} - J_n)] h_{n-1}(x) + b(J_{n-1} - J_n), \tag{9.51}$$

where b is a constant (see [42]). Letting $x \to \infty$, it follows that

$$\Lambda_n \le [1 + 2(J_{n-1} - J_n)] \Lambda_{n-1}, \qquad n \ge 1. \tag{9.52}$$

It is known that the matrices $\{\Lambda_n\}$ are *decreasing*, in the sense that $\Lambda_n - \Lambda_{n-1}$ is positive semidefinite for each $n \ge 1$ [18, 62]. Hence the bound (9.51) is not tight in the linear model. However, the semi-decreasing property (9.52) is sufficient to deduce convergence of the matrices $\{\Lambda_n\}$ to a finite limiting matrix.

There is no space here to consider the VIA in further detail. We note however that it is well known that the successive approximation procedure generates stabilizing policies for the linear state space model provided the initial policy is stabilizing and linear [18]. Theorem 9.8 shows that it is enough to assume only stability.

9.6.2 Network models

We now apply the general results of Section 9.4 to the scheduling problem for multiclass queuing networks. For simplicity we discuss here only a relatively simple class of network models which can be formulated through an extension of the M/M/1 model. A treatment of general network models is given in [41, 43].

Consider a network composed of d single server stations, indexed by $\sigma = 1, \ldots, d$. The network is populated by ℓ classes of customers: Class k customers require service at station $s(k)$. An exogenous stream of customers of class 1 arrive to machine $s(1)$, and subsequent routing of customers is deterministic. If the service times and interarrival times are assumed to be exponentially distributed, then after a suitable time scaling and sampling of the process, the dynamics of the network can be described by the random linear system,

$$x_{t+1} = x_t + \sum_{k=0}^{\ell} I_{t+1}(k)[e^{k+1} - e^k] a_t(k), \qquad t \ge 0, \tag{9.53}$$

where the state process $\boldsymbol{x}$ evolves on the countable state space $\mathbb{X} = \mathbb{N}^\ell$, and $x_t(k)$ denotes the number of class k customers in the system at time t. An example of a two station network is illustrated in Figure 9.3.

The random variables $\{I_t : t \ge 1\}$ are i.i.d. on $\{0,1\}^{\ell+1}$, with

$$\mathbb{P}\{\textstyle\sum_i I_t(k) = 1\} = 1, \text{ and } \mathbb{E}[I_t(k)] = \mu_k.$$

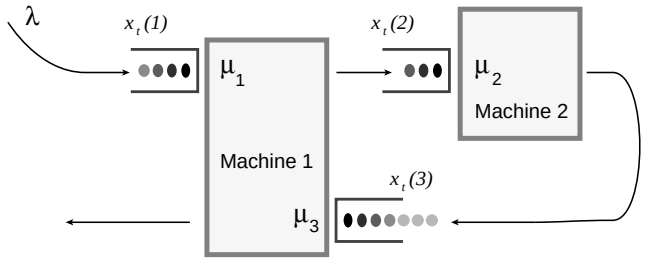

Figure 9.3 A multiclass network with $d = 2$ and $\ell = 3$.

For $1 \le k \le \ell$, μ_k denotes the service rate for class k customers. For $k = 0$, we let $\mu_0 := \lambda$ denote the arrival rate of customers of class 1. For $1 \le k \le \ell$ we let e^k denote the kth basis vector in $\mathbb{R}^\ell$, and we set $e^0 = e^{\ell+1} := 0$.

The sequence $\{a_t : t \ge 0\}$ is the control, which takes values in $\mathbb{A} := \{0,1\}^{\ell+1}$. We define $a_t(0) \equiv 1$. The set of admissible control actions $\mathbb{A}(x)$ is defined in an obvious manner: for $a \in \mathbb{A}(x)$,

(i) For any $1 \le k \le \ell$, $a(k) = 0$ or 1;

(ii) For any $1 \le k \le \ell$, $x_k = 0 \Rightarrow a(k) = 0$;

(iii) For any station σ, $0 \le \sum_{k:s(k)=\sigma} a(k) \le 1$;

(iv) For any station σ, $\sum_{k:s(k)=\sigma} a(k) = 1$ whenever $\sum_{k:s(k)=\sigma} x(k) > 0$.

If $a(k) = 1$, then buffer k is chosen for service. Condition (ii) then imposes the physical constraint that a customer cannot be serviced at a buffer if that buffer is empty. Condition (iii) means that only one customer may be served at a given instant at a single machine σ.

Since the control is bounded, a reasonable cost function is $c(x,a) = c^T x$, where $c \in \mathbb{R}^\ell$ is a vector with strictly positive entries. For concreteness, we take $c(x,a) = |x| := \sum_k x(k)$. The non-idling condition (iv) is satisfied by any optimal stationary policy with this cost criterion: An inductive proof is given in [41] based upon value iteration.

The controlled transition function has the simple form,

$$\begin{aligned} p(x + e^{k+1} - e^k \mid x, a) &= \mu_k a(k), \qquad 0 \le k \le \ell. \\ p(x \mid x, a) &= 1 - \sum_0^\ell \mu_k a(k) \end{aligned}$$

The accessibility condition (9.7) holds with s everywhere positive, and $\nu = \delta_\theta$, with $\boldsymbol{\theta}$ equal to the empty state $\boldsymbol{\theta} = (0, \dots, 0)^T \in \mathbb{X}$. This follows from the non-idling assumption (iv).

Associated with this network is a *fluid model*. For each initial condition $x_0 \ne 0$, we construct a continuous time process $\varphi^{x_0}(t)$ as follows. If $m = |x_0|$, and if tm is an integer, we set

$$\varphi^{x_0}(t) = \frac{1}{m} x_{mt}.$$

For all other $t \geq 0$, we define $\varphi^{x_0}(t)$ by linear interpolation, so that it is continuous and piecewise linear in t. Note that $|\varphi^{x_0}(0)| = 1$, and that φ^{x_0} is Lipschitz continuous. The collection of all "fluid limits" is defined by

$$\mathcal{L} := \bigcap_{n=1}^{\infty} \overline{\{\varphi^x : |x| > n\}}$$

where the overbar denotes weak closure in $C(\mathbb{R})$, the space of continuous functions, with the topology of uniform convergence on compact sets. This set of stochastic process of course depends on the particular policy π which has been applied.

Any $\varphi \in \mathcal{L}$ evolves on the state space $\mathbb{R}_+^\ell$ and, for a wide class of scheduling policies, satisfies a differential equation of the form

$$\frac{d}{dt}\varphi(t) = \sum_{k=0}^{\ell} \mu_k [e^{k+1} - e^k] u_t(k), \tag{9.54}$$

where the function u_t is analogous to the discrete control a_t, and satisfies similar constraints (see the M/M/1 queue model described earlier, or [12, 11] for more general examples).

Stability of (9.53) in terms of c-regularity is closely connected with the stability of the fluid model [12, 35, 13]. The fluid model $\mathcal{L}$ is called *L_p-stable* if

$$\lim_{t\to\infty} \sup_{\varphi\in\mathcal{L}} \mathbb{E}[|\varphi(t)|^p] = 0.$$

It is shown in [35] that L_2-stability of the fluid model is equivalent to a form of c-regularity for the network:

Theorem 9.11 *The following stability criteria are equivalent for the network under any non-idling, stationary policy.*

(i) The drift condition (9.13) holds for some function V. The function V is equivalent to a quadratic in the sense that, for some $\gamma > 0$,

$$1 + \gamma|x|^2 \leq V(x) \leq 1 + \gamma^{-1}|x|^2, \qquad x \in \mathbb{X}. \tag{9.55}$$

(ii) For some quadratic function V,

$$\mathbb{E}_x\Big[\sum_{n=0}^{\sigma_\theta} |x_n|\Big] \quad \leq \quad V(x), \qquad x \in \mathbb{X},$$

where σ_θ is the first entrance time to $\boldsymbol{\theta} = 0$.

(iii) For some quadratic function V and some $\bar{J} < \infty$,

$$\sum_{n=1}^{N} \mathbb{E}_x[|x_n|] \quad \leq \quad V(x) + N\bar{J}, \quad \text{for all } x \text{ and } N \geq 1.$$

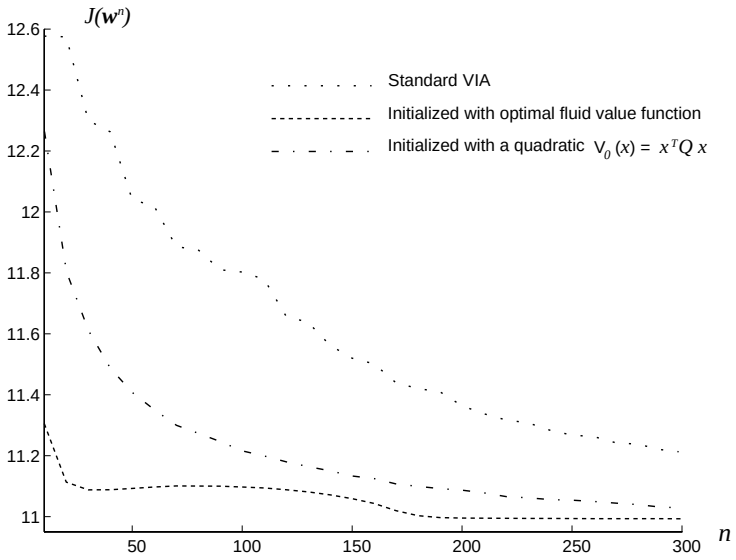

Figure 9.4 Convergence of the VIA with V_0 taken as the value function for the associated fluid control problem, or a pure quadratic function obtained through a linear program.

(iv) The fluid model $\mathcal{L}$ is L_2-stable.

■

Using this result it is shown in [42] that when applying policy iteration to a network model, on performing the fluid scaling one obtains a sequence of fluid models which are the solutions of a policy iteration scheme for the fluid model. Moreover, the algorithm convergence to yield a policy which is s-optimal for both the network and its fluid model. The minimal relative value function h_* is equivalent to a quadratic in the sense that, for some constant b_1,

$$b_1^{-1}\|x\|^2 - b_1 \le h_*(x) \le b_1\|x\|^2 + b_1$$

For any stationary policy, the solution to Poisson's equation is bounded from below by a quadratic.

Let us turn to the VIA: In view of Theorem 9.11, how should we initialize the algorithm? Two possibilities are suggested:

(i) Given the previous analysis of the M/M/1 queue it appears natural to set V_0 equal to the value function for a fluid model,

$$V_*(x) := \min \int_0^\infty |\varphi(t)|\, dt \qquad \varphi(0) = x,\ x \in \mathbb{X},$$

where the minimum is with respect to all policies for the fluid model. One can show that for large x, V_* does approximate the relative value function [41, 42, 25].

(ii) The conclusion that the relative value function is 'nearly quadratic' suggests that we search for a pure quadratic form satisfying (9.13),

$$V_0(x) = x^T Q x, \qquad x \in \mathbb{X}.$$

In [35] a linear program is constructed to compute a quadratic solution to (9.13) for network models, based on prior results of [36, 47].

We conclude with a numerical experiment to show how a careful initialization can dramatically speed convergence of the VIA. We consider the three buffer model illustrated in Figure 9.3 with the following parameters: $\lambda/\mu_2 = 9/10$; $\lambda/\mu_1 + \lambda/\mu_3 = 9/11$; and $\mu_1 = \mu_3$. The optimal value function V_* can be computed explicitly in this case, and a pure quadratic Lyapunov function can also be computed easily.

Two experiments were performed to compare the performance of the VIA initialized with these two value functions. To apply value iteration the buffer levels were truncated so that $x_i < 45$ for all i. This gives rise to a finite state space MDP with $45^3 = 91,125$ states. The results from two experiments are shown in Figure 9.4. For comparison, data from the standard VIA with $V_0 \equiv 0$ is also given. We have taken 300 steps of value iteration, saving data for $n = 10, \ldots, 300$. The convergence is exceptionally fast in both experiments. Note that the convergence of J_n is *not* monotone in the experiment shown using the fluid value function initialization. However, this initialization leads to fast convergence to the optimal cost $J_* \approx 10.9$.

9.7 EXTENSIONS AND OPEN PROBLEMS

It is hoped that the development in this chapter has suggested to the reader some interesting topics for further research. We list here some areas which have been of interest to the author.

Existence and structure of optimal policies. The results of Section 9.4 are fairly complete, but the setting is special. It appears that there is still much to be done to better understand the structure of optimal policies, and criteria for existence of optimal policies in this general setting.

Continuous time. In this chapter the analysis has been restricted to a resolvent kernel, and the same approach can be followed in continuous time where the resolvent becomes

$$K = \int_0^\infty \beta e^{-\beta t} P^t \, dt$$

with $\beta > 0$. Again one can show that any variable of interest (the invariant measures, solutions to Poisson's equation, or solutions to (9.13)) can be mapped between the resolvent and the continuous time process. Further discussion may be found in [42, 55].

Geometric ergodicity and risk sensitive control. The risk sensitive control criterion is given via

$$J_\gamma(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} \log\Big(\mathbb{E}_x\Big[\exp\Big(\gamma \sum_{t=0}^{n-1} c_\pi(x_t^\pi)\Big)\Big]\Big).$$

where the 'risk factor' γ is assumed to be a small, positive number in the *risk-averse* case.

Models of this sort were first considered in [3, p. 329] and in [32, 52]. This control problem has attracted more recent attention because of the interesting connections between risk sensitive control and game theory [33, 21, 62].

Under a norm-like condition on the model it can be shown that when this cost is finite valued, the Markov chain exhibits a strong form of stability known as geometric ergodicity [2, 6]. Conversely, such stability assumptions imply that the cost is finite, and ensure that an optimal policy does exist [23, 9].

Our present understanding of the optimization problem for Markov chains on an infinite state space is currently weak, and this appears to be an area worthy of further study.

Simulation. The use of simulation will become increasingly important in both evaluating and synthesizing policies. Much of the burden of finding an optimal policy surrounds the solution of Poisson's equation, for which now there are several simulation based algorithms such as temporal difference learning. There are also simulation based versions of both value and policy iteration (see [4, 57, 34]).

We have remarked that high variance can make simulation impractical. The use of the fluid value function is one promising approach to variance reduction for network models [25], and related techniques may prove useful in the development of simulation-based optimization algorithms.

Complexity. This has always been one of the most challenging issues in optimal control. Markovian models are frequently too 'fine-grained' to be useful in optimization. One solution then is to seek some form of aggregation. For general MDP models one can directly discretize the state space to obtain a finite state space model.

This is an area in which the most relevant research will most likely focus on a specific application. In the case of network models, either fluid models or Brownian motion models provide approaches to aggregation which deserve further study.

References

[1] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optim.*, 31:282–344, 1993.

[2] S. Balaji and S.P. Meyn. Multiplicative ergodicity and large deviations for an irreducible Markov chain. *Stochastic Process. Appl.*, 90:123–144, 2000.

[3] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

[4] Bertsekas, D., Tsitsiklis, J. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

[5] V. S. Borkar. *Topics in controlled Markov chains.* Pitman Research Notes in Mathematics Series # 240, Longman Scientific & Technical, UK, 1991.

[6] V. Borkar and S.P. Meyn. Risk Sensitive Optimal Control: Existence and Synthesis for Models with Unbounded Cost. To appear, *Mathematics of O.R.*, 2000.

[7] R. Cavazos-Cadena. Value iteration in a class of communicating Markov decision chains with the average cost criterion. *SIAM J. Control and Optimization*, 34:1848–1873, 1996.

[8] R. Cavazos-Cadena and E. Fernandez-Gaucherand. Value iteration in a class of average controlled Markov chains with unbounded costs: Necessary and sufficient conditions for pointwise convergence. *J. Applied Probability*, 33:986–1002, 1996.

[9] R. Cavazos-Cadena and E. Fernandez-Gaucherand. Controlled Markov chains with risk-sensitive criteria: Average cost, optimality equations, and optimal solutions. Mathematical Methods of Operations Research, 49:299-324, 1999.

[10] R-R. Chen and S.P. Meyn. Value iteration and optimization of multiclass queueing networks. *Queueing Systems*, 32:65–97, 1999.

[11] J. Dai and G. Weiss. Stability and instability of fluid models for certain re-entrant lines. *Mathematics of Operations Research*, 21(1):115–134, February 1996.

[12] J. G. Dai. On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.*, 5:49–77, 1995.

[13] J. G. Dai and S.P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Automat. Control*, 40:1889–1904, November 1995.

[14] R. Dekker. *Denumerable Markov Decision Chains: Optimal Policies for Small Interest Rates*, PhD thesis, University of Leiden, Leiden, the Netherlands, 1985.

[15] R. Dekker. Counterexamples for compact action Markov decision chains with average reward criteria. *Comm. Statist.-Stoch. Models*, 3:357–368, 1987.

[16] C. Derman. Denumerable state MDPs. *Ann. Amth. Statist.*, 37:1545–1554, 1966.

[17] D. Down, S. P. Meyn, and R. L. Tweedie. Geometric and uniform ergodicity of Markov processes. *Ann. Probab.*, 23(4):1671–1691, 1996.

[18] M. Duflo. *Méthodes Récursives Aléatoires.* Masson, 1990.

[19] E. B. Dynkin and A. A. Yushkevich. *Controlled Markov Processes*, volume Grundlehren der mathematischen Wissenschaften 235 of *A Series of*

Comprehensive Studies in Mathematics. Springer-Verlag, New York, NY, 1979.

[20] E. A. Feinberg, Ya. A. Kogan and A. N. Smirnov. Optimal Control by the Retransmission Probability in Slotted ALOHA Systems. *Performance Evaluation*, 5:85-96, 1985.

[21] W.H. Fleming and W.M. McEneaney. Risk-sensitive control and differential games. volume 84 of *Lecture Notes in Control and Info. Sciences*, pages 185–197. Springer-Verlag, Berlin; New York, 1992.

[22] P.W. Glynn and S.P. Meyn. A Liapunov bound for solutions of Poisson equation. *Annals of Prob.*, 24:916–931, 1996.

[23] D. Hernández-Hernández and S.I. Marcus. Risk sensitive control of Markov processes in countable state space. *Systems Control Lett.*, 29:147–155, July 1996. correction in *Systems and Control Letters*, 34:105-106, 1998.

[24] Henderson, S. G. *Variance Reduction Via an Approximating Markov Process*. Ph.D. thesis. Department of Operations Research, Stanford University. Stanford, California, USA, 1997.

[25] S.G. Henderson and S.P. Meyn. Variance reduction for simulation in multiclass queueing networks. *submitted to the IIE Transactions on Operations Engineering: special issue honoring Alan Pritsker on simulation in industrial engineering*, 1999.

[26] J. Humphrey D. Eng and S.P. Meyn. Fluid network models: Linear programs for control and performance bounds. In J. Cruz J. Gertler and M. Peshkin, editors, *Proceedings of the 13th IFAC World Congress*, volume B, pages 19–24, San Francisco, California, 1996.

[27] O. Hernández-Lerma and J. B. Lasserre. Discrete time Markov control processes I. Springer-Verlag, New York, 1996.

[28] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena. Recurrence conditions for Markov decision processes with Borel state space: A survey. *Ann. Operations Res.*, 28:29–46, 1991.

[29] A. Hordijk. *Dynamic Programming and Markov Potential Theory. Math. Centre Tracts, Mathematical Centrum, Amsterdam*, 2nd ed., 1977.

[30] A. Hordijk and M. L. Puterman. On the convergence of policy iteration. *Math. Op. Res.*, 12:163–176, 1987.

[31] R. A. Howard. *Dynamic Programming and Markov Processes*. John Wiley and Sons/MIT Press, New York, NY, 1960.

[32] R.A. Howard and J.E. Matheson. Risk-sensitive Markov decision processes. *Management Sci.*, 8:356–369, 1972.

[33] D. H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Automat. Control*, AC-18:124–131, 1973.

[34] V. R. Konda and V. S. Borkar. Actor-critic-type learning algorithms for Markov decision processes. *SIAM J. Control and Optimization*, 38:4-123, 1999.

[35] P.R. Kumar and S.P. Meyn. Duality and linear programs for stability and performance analysis queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 41(1):4–17, 1996.

[36] S. Kumar and P. R. Kumar. Performance bounds for queueing networks and scheduling policies. *IEEE Trans. Automat. Control*, AC-39:1600–1611, August 1994.

[37] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, NY, 1972.

[38] S. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23:687–710, 1975.

[39] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.

[40] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes. *Adv. Appl. Probab.*, 25:518–548, 1993.

[41] S.P. Meyn. Stability and optimization of multiclass queueing networks and their fluid models. In *Mathematics of Stochastic Manufacturing Systems*, Lectures in Applied Mathematics, Vol. 33. Proc. AMS-SIAM Summer Seminar in Applied Mathematics June 17-22, 1996, Williamsburg, Virginia. G. George Yin and Qing Zhang (Eds.). American Mathematical Society, Providence, 1997,

[42] S.P. Meyn. The policy improvement algorithm for Markov decision processes with general state space. *IEEE Trans. Automat. Control*, AC-42:1663–1680, 1997.

[43] S.P. Meyn. Feedback regulation for sequencing and routing in multiclass queueing networks. *SIAM J. Control and Optimization*, to appear, 1999.

[44] S.P. Meyn. Algorithms for optimization and stabilization of controlled Markov chains. *Sadhana*, 24:1-29, 1999.

[45] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge University Press, Cambridge, 1984.

[46] E. Nummelin. On the Poisson equation in the potential theory of a single kernel. *Math. Scand.*, 68:59–82, 1991.

[47] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, Optimization of Multiclass Queueing Networks: Polyhedral and Nonlinear Characterizations of Achievable Performance, *Annals of Applied Probability*, 4:43-75, 1994.

[48] J. Perkins. *Control of Push and Pull Manufacturing Systems.* PhD thesis, University of Illinois, Urbana, IL, September 1993. Technical report no. UILU-ENG-93-2237 (DC-155).

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

[50] R. K. Ritt and L. I. Sennott. Optimal stationary policies in general state space Markov decision chains with finite action set. *Mathematics of Operations Research*, 17(4):901–909, November 1993.

[51] S. M. Ross. Applied probability models with optimization applications. Dover books on advanced Mathematics, 1992. Republication of the work first published by Holden-Day, 1970.

[52] U.G. Rothblum. Multiplicative Markov decision chains. *Math. Operations Res.*, 9:6–24, 1984.

[53] L. I. Sennott. A new condition for the existence of optimal stationary policies in average cost Markov decision processes. *Operations Research Letters*, 5:17–23, 1986.

[54] L.I. Sennott. The convergence of value iteration in average cost Markov decision chains. *Operations Research Letters*, 19:11–16, 1996.

[55] L.I. Sennott. Stochastic Dynamic Programming and the Control of Queueing Systems. *Wiley*, 1999.

[56] F.M. Spieksma. *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, PhD thesis, University of Leiden, Leiden, the Netherlands, 1990.

[57] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. on Automatic Control*, 42:674-690, 1997.

[58] P. Tuominen and R.L. Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. Appl. Probab.*, 26:775–798, 1994.

[59] R. Weber and S. Stidham. Optimal control of service rates in networks of queues. *Adv. Appl. Probab.*, 19:202–218, 1987.

[60] G. Weiss. Optimal Draining of a Fluid Re-Entrant Line. In *Stochastic Networks*. Volume 71 of IMA volumes in Mathematics and its Applications, pp. 91-103. Frank Kelly and Ruth Williams, eds. Springer-Verlag, New York, 1995.

[61] G. Weiss. Optimal Draining of Fluid Re-Entrant Lines: Some Solved Examples. In *Stochastic Networks: Theory and Applications*. Volume 4 of Royal Statistical Society Lecture Notes Series , pp. 19-34, F.P. Kelly, S. Zachary and I. Ziedins eds. Oxford University Press, Oxford, 1996.

[62] P. Whittle. *Risk-Sensitive Optimal Control*. John Wiley and Sons, Chichester, NY, 1990.

Sean P. Meyn
Department of Electrical and Computer Engineering
and the Coordinated Sciences Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
meyn@newton.csl.uiuc.edu

10 CONVEX ANALYTIC METHODS IN MARKOV DECISION PROCESSES

Vivek S. Borkar

Abstract: This article describes the convex analytic approach to classical Markov decision processes wherein they are cast as a static convex programming problem in the space of measures. Applications to multiobjective problems are described.

10.1 INTRODUCTION

10.1.1 Background

Markov decision processes optimize phenomena evolving with time and thus are intrinsically dynamic optimization problems. Nevertheless, they can be cast as abstract 'static' optimization problems over a closed convex set of measures. They then become convex programming (in fact, infinite dimensional linear programming) problems for which the enormous machinery of the latter fields can be invoked and used to advantage. Logically, these are extensions of the linear programming approach to finite state finite action space problems due to Manne [43]. (Further references are given in the 'bibliographical note' at the end.) The attraction of this approach lies in the following:

(i) It leads to elegant alternative derivations of known results, sometimes under weaker hypotheses, from a novel perspective.

(ii) It brings to the fore the possibility of using convex/linear programming techniques for computing near-optimal strategies.

(iii) It allows one to handle certain unconventional problems (such as control under additional constraints on secondary 'costs') where traditional dynamic programming paradigm turns out to be infeasible or awkward.

While the convex analytic formulation is available for a variety of cost criteria and fairly general state spaces, our primary focus will be on the 'pathwise ergodic control' problem on a countable state space, for which the theory is the most elegant. This is done in the next section. Section 3 sketches extensions to other cost criteria and general state spaces, and the dual problem. Section 4 considers multiobjective problems, such as the problem of control under constraints. An 'Appendix' discusses stability of controlled Markov chains.

10.1.2 Notation

Initially we shall work with a denumerable state space, identified with the set $\{0, 1, 2, \cdots\}$ without any loss of generality. For a Polish space X, $\mathcal{P}(X)$ will denote the Polish space of probability measures on X with the Prohorov topology. (See, e.g. [15], Chapter 2.) Recall the various cost criteria commonly considered:

1. *Finite horizon cost*: $\mathbb{E}[\sum_{n=0}^{T} r(x_n, a_n)]$ for some finite $T > 0$.
2. *Discounted cost*: $\mathbb{E}[\sum_{n=0}^{\infty} \delta^n r(x_n, a_n)]$ for some discount factor $\delta \in (0, 1)$.
3. *Total cost*: Same as above for $\delta = 1$.
4. *Cost till exit time*: $\mathbb{E}[\sum_{n=0}^{\tau-1} r(x_n, a_n)]$ with $\tau = \min\{n \geq 0 : x_n \notin D\}$ for a prescribed $D \subset \mathbb{X}$.
5. *Expected ergodic cost*: $\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} \mathbb{E}[r(x_m, a_m)]$
6. *Pathwise ergodic cost*: Minimize 'almost surely'
$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} r(x_n, a_n). \tag{10.1}$$
7. *Risk-sensitive cost*: $\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}[e^{\alpha \sum_{m=0}^{n-1} r(x_n, a_n)}]$

All except the last one (which has per stage costs entering in a multiplicative, as opposed to additive, fashion) are amenable to a convex analytic formulation. Nevertheless, as already mentioned, we shall confine ourselves to the pathwise ergodic cost for the most part as a representative case. (See [49] for a traditional dynamic programming based perspective.) Section 3.1 will briefly mention the counterparts of these results for the other costs. Note that if $\pi \in \Pi^{RS}$ is used and x_0 is in the support of an ergodic distribution $\eta \in \mathcal{P}(\mathbb{X})$ for the corresponding time-homogeneous Markov chain $\{x_n\}$, then (10.1) will a.s. equal
$$\sum_i \eta(i) \bar{r}(i, \pi(i))$$
where $\bar{r} : \mathbb{X} \times \mathcal{P}(\mathbb{A}) \to \mathbb{R}$ is defined by
$$\bar{r}(i, u) = \int r(i, a) u(da), \ (i, u) \in \mathbb{X} \times \mathcal{P}(\mathbb{A}).$$
This will be the starting point of our convex analytic formulation of the pathwise ergodic control problem, which we take up next.

10.2 PATHWISE ERGODIC CONTROL

10.2.1 Ergodic occupation measures

If the chain controlled by $\pi \in \Pi^{RS}$ has an invariant probability measure $\eta \in \mathcal{P}(\mathbb{X})$, we associate with the pair (η, π) the *ergodic occupation measure* $\hat{\eta}_\pi$ defined by:

$$\int_{\mathbb{X}\times\mathbb{A}} g d\hat{\eta}_\pi = \sum_{i\in\mathbb{X}} \eta(i) \int_{\mathbb{A}} g(i,a)\pi(i,da),$$

for $g \in C_b(\mathbb{X}\times\mathbb{A})(\stackrel{\Delta}{=}$ space of bounded continuous maps $\mathbb{X}\times\mathbb{A}\to\mathbb{R}$. Analogous notation will be used throughout.) Let $\mathcal{C}$ be a countable subset of $C_b(\mathbb{X})$ satisfying: for any $f \in C_b(\mathbb{X})$, there exist $\{g_n\}$ in the linear span of $\mathcal{C}$ such that $g_n \to f$ in a bounded, pointwise fashion.

Lemma 10.1 *$\hat{\eta}_\pi$ satisfies: $\forall g \in C_b(\mathbb{X})$,*

$$\sum_{i\in\mathbb{X}} g(i)\hat{\eta}_\pi(\{i\},\mathbb{A}) = \int_{\mathbb{X}\times\mathbb{A}} \left(\sum_{j\in\mathbb{X}} p(j/i,a)g(j)\right) d\hat{\eta}_\pi.$$

Conversely, if $\nu \in \mathcal{P}(\mathbb{X}\times\mathbb{A})$ satisfies

$$\sum_{i\in\mathbb{X}} g(i)\nu(\{i\},\mathbb{A}) = \int_{\mathbb{X}\times\mathbb{A}} \left(\sum_{j\in\mathbb{X}} p(j/i,a)g(j)\right) d\nu \tag{10.2}$$

for $g \in \mathcal{C}$, then $\nu = \hat{\eta}_\pi$ for some (η,π).

Proof. The first claim is simply the invariance of η under π. For the second, note that (10.2) holds for all $g \in C_b(\mathbb{X})$ if it does so for $g \in \mathcal{C}$. Disintegrate ν as $\nu(\{i\},da) = \eta(i)\pi(i,dy)$ with $\eta \in \mathcal{P}(\mathbb{X})$ and $\pi : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ the appropriate regular conditional law defined η-a.s. uniquely. Identify π with an element of Π^{RS}, whence (10.2) for $g \in C_b(\mathbb{X})$ implies that η is invariant under π. The claim follows. ∎

Often $\mathcal{C} = \{I_{\{i\}}, i \in \mathbb{X}\}$ is a convenient choice for $\mathcal{C}$. Let G denote the set of all ergodic occupation measures.

Lemma 10.2 *G is closed convex in $\mathcal{P}(\mathbb{X}\times\mathbb{A})$.*

Proof. (10.2) is preserved under convergence in $\mathcal{P}(\mathbb{X}\times\mathbb{A})$. In view of Lemma 10.1, this implies that G is closed. Let $\nu_j \in G$, $1 \le j \le n$, with $\nu_j(\{i\},da) = \eta_j(i)\pi_j(i,da)$ for $\pi_j \in \Pi^{RS}$ and $\eta_j \in \mathcal{P}(\mathbb{X})$ invariant under respective π_j's. Let $b_j \in (0,1), 1 \le j \le n$, with $\sum_j b_j = 1$. Set $\nu = \sum_j b_j\nu_j$, $\eta = \sum_j b_j\eta_j$, and define $\pi \in \Pi^{RS}$ by:

$$\pi(i,da) = \sum_{j=1}^{n} \left[b_j\eta_j(i) \Big/ \left(\sum_{l=1}^{n} b_l\eta_l(i)\right)\right] \pi_j(i,dy),$$

for $i \in$ support (η), arbitrary otherwise. Then $\nu(\{i\}, da) = \eta(i)\pi(i, da),\ i \in \mathbb{X}$. Now write (10.2) separately for each ν_j, $1 \leq j \leq n$, multiply it through by b_j, and sum over j. Rearranging terms, one recovers (10.2) for ν. Thus $\nu \in G$. ∎

We now establish a technical lemma for later use. Let $\nu = \hat{\eta}_\pi \in G$. Now suppose that for some $j \in$ support (η) (say, $j = 1$), we have $\pi(j, da) = \pi(1, da) = b\varphi_1(da) + (1-b)\varphi_2(da)$ for some $b \in (0,1)$ and $\varphi_1 \neq \varphi_2$ in $\mathcal{P}(\mathbb{A})$. Define $\pi', \pi'' \in \Pi^{RS}$ by:

$$\begin{aligned} \pi'(i) &= \pi''(i) = \pi(i), \quad i \neq 1 \\ \pi'(1) &= \varphi_1,\ \pi''(1) = \varphi_2. \end{aligned}$$

Lemma 10.3 *Both π', π'' above admit invariant probability measures containing '1' in their support.*

Proof. Changing π to π' or π'' affects only the probabilities of transitions out of 1. Let T, T', T'' denote the mean return times to 1 under π, π', π'' resp. Then clearly $T = bT' + (1-b)T''$. Since $T < \infty$, $1 > b > 0$, we have $T', T'' < \infty$, implying the claim. ∎

Let $G_e \subset G$ denote the set of extreme points of G.

Lemma 10.4 *Every $\nu \in G_e$ is of the form $\nu = \hat{\eta}_\phi$ for some $\phi \in \Pi^S$ and η ergodic under ϕ.*

Proof. Let $\nu \in G_e$ with $\nu(\{i\}, da) = \eta(i)\pi(i, da)$. For $i \notin$ support (η), set $\pi(i) =$ some Dirac measure, without affecting ν. Let $i \in$ support (η), say, $i = 1$. Suppose $\pi(1, da) = b\varphi_1(da) + (1-b)\varphi_2(da)$ for some $b \in (0,1), \varphi_1 \neq \varphi_2$ in $\mathcal{P}(\mathbb{A})$. Define π', π'' as above and let η_1, η_2 be ergodic probability measures under π', π'' resp. containing 1 in their supports. Pick $c \in (0,1)$ such that

$$b = c\eta_1(1)/(c\eta_1(1) + (1-c)\eta_2(1)),$$

which is possible because $\eta_1(1), \eta_2(1) > 0$. Let $\eta' = c\eta_1 + (1-c)\eta_2, \nu'(\{i\}, da) = \eta'(i)\pi(i, dy)$. A computation similar to that in Lemma 10.2 shows that ν' satisfies (10.2) and hence $\nu' = \hat{\eta}'_\pi \in G$. Also, support (η') = support $(\eta_1) \cup$ support (η_2) and for $i, j \in \mathbb{X}$,

$$\begin{aligned} \int p(j/i, a)\pi(i, da) &= b \int p(j/i, a)\pi'(i, da) \\ &\quad + (1-b) \int p(j/i, a)\pi''(i, da). \end{aligned} \tag{10.3}$$

Since η_1 (resp., η_2) is ergodic under π' (resp., π''), any two states in its support communicate under π' (resp., π'') and therefore under π in view of (10.2) above. Since 1 is in support $(\eta_1) \cap$ support (η_2), it follows that support (η') is a single communicating class under π. Thus η' is an ergodic probability measure under π. If η is also ergodic, we must have $\eta = \eta'$ and $\nu = \nu'$. As in the proof of Lemma 10.2, one can then verify that $\nu' = c\nu_1 + (1-c)\nu_2$ where $\nu_1(\{i\}, da) = \eta_1(i)\pi'(i, da)$ and $\nu_2(\{i\}, da) = \eta_2(i)\pi''(i, dy)$. Since $0 < c < 1$

and $\nu_1 \neq \nu_2$, $\nu \notin G_e$, a contradiction. Suppose η is not ergodic under f. Then $\eta = s\eta^1 + (1-s)\eta^2$ for $s \in (0,1)$ and $\eta^1, \eta^2 \in \mathcal{P}(\mathbb{X})$ which are distinct invariant probability measures under π. Then $\nu = s\nu^1 + (1-s)\nu^2$ where $\nu^j(\{i\}, da) = \eta^j(i)\pi(i, da), i = 1, 2$ are clearly distinct elements of G. Thus $\nu \notin G_e$, a contradiction. So, (i) η must be ergodic under π, and (ii) (10.2) is impossible, i.e. $\pi(i, da)$ is Dirac for $i \in$ support (η). This completes the proof. ∎

Let $\bar{\mathbb{X}} = \mathbb{X} \cup \{\infty\}$ denote the one point compactification of $\mathbb{X}$ and view $\mathbb{X} \subset \bar{\mathbb{X}}$ via the natural embedding. Likewise, $\mathcal{P}(\mathbb{X} \times \mathbb{A}) \subset \mathcal{P}(\bar{\mathbb{X}} \times \mathbb{A})$ via the natural embedding. Let $\bar{G}$ be the closure of G in $\mathcal{P}(\bar{\mathbb{X}} \times \mathbb{A})$ and $\bar{G}_e$ the set of its extreme points. The statement of the next lemma requires familiarity with Choquet's theorem, which we recall here.

Let E be a Hausdorff locally convex topological vector space and $X \subset E$ a convex compact metrizable subset. Given $\mu \in \mathcal{P}(X)$, call x its barycenter if $f(x) = \int f d\mu$ for all continuous affine $f : X \to \mathbb{R}$.

Theorem 10.1 (Choquet) *Each $x \in X$ is the barycenter of some $\mu \in \mathcal{P}(X)$ supported on the extreme points of X.*

See [18], pp.140-141, for a proof. Metrizability of X ensures that the set of its extreme points is F_σ, hence measurable ([18], p.138), whereas compactness of X ensures that it is nonempty ([18], pp.105).

Lemma 10.5 *$G_e \subset \bar{G}_e$ and any $\nu \in G$ is the barycenter of a probability measure on G_e.*

Proof. Any $\nu \in G_e \backslash \bar{G}_e$ must be a convex combination of two distinct elements of $\bar{G}$ at least one of which must assign strictly positive probability to $\{\infty\} \times \mathbb{A}$. But then so will ν, a contradiction. Thus $G_e \subset \bar{G}_e$. Now, if G is compact, $G_e \neq \phi$ and $G_e = \bar{G}_e$, whence the second claim follows from Theorem 10.1. If not, apply the theorem to $\bar{G}$. Then $\bar{G}_e \neq \phi$ and ν is the barycenter of a probability measure Φ on $\bar{G}_e$. If $\Phi(\bar{G}_e \backslash G_e) > 0$, we must have $\nu(\{\infty\} \times \mathbb{A}) > 0$, a contradiction. Thus $\Phi(G_e) = 1$. ∎

Finally, we have:

Lemma 10.6 *Each $\nu \in \bar{G}$ is of the form: For $B \subset \bar{\mathbb{X}} \times \mathbb{A}$ Borel,*

$$\nu(B) = \delta\nu'(B \cap (\mathbb{X} \times \mathbb{A})) + (1-\delta)\nu''(B \cap (\{\infty\} \times \mathbb{A})) \tag{10.4}$$

with $\delta \in [0,1], \nu' \in G$ and $\nu'' \in \mathcal{P}(\{\infty\} \times \mathbb{A})$.

Proof. Clearly, (10.4) holds for some $\nu' \in \mathcal{P}(\mathbb{X} \times \mathbb{A})$. The claim is trivial for $\delta = 0$. For $\delta = 1, \nu' = \nu \in G$. Thus let $\delta \in (0,1)$. Let $\nu_n \in G, n \geq 1$, be such that $\nu_n \to \nu$ in $\bar{G}$. Suppose

$$\nu_n(i, da) = \eta_n(i)\pi_n(i, da), n \geq 1.$$

Then for $j \in \mathbb{X}$,

$$\int p(j/\cdot,\cdot)d\nu_n = \nu(\{j\} \times \mathbb{A}), \;\; n \geq 1.$$

Since $\{j\} \times \mathbb{A}$ is both open and closed in $\bar{\mathbb{X}} \times \mathbb{A}$, we have

$$\nu_n(\{j\} \times \mathbb{A}) \to \nu(\{j\} \times \mathbb{A}) = \delta\nu'(\{j\} \times \mathbb{A}).$$

Letting $\mathbb{X}(N) = \{0, 1, \ldots, N\} \subset \mathbb{X}$,

$$\begin{aligned}
\liminf_{n\to\infty} \int p(j/\cdot,\cdot)d\nu_n &\geq \lim_{n\to\infty} \int_{\mathbb{X}(N)\times\mathbb{A}} p(j/\cdot,\cdot)d\nu_n \\
&= \int_{\mathbb{X}(N)\times\mathbb{A}} p(j/\cdot,\cdot)d\nu \\
&= \delta \int_{\mathbb{X}(N)\times\mathbb{A}} p(j/\cdot,\cdot)d\nu' \\
&\to \delta \int p(j/\cdot,\cdot)d\nu'
\end{aligned}$$

as $N \to \infty$. Combining the two,

$$\int p(j/\cdot,\cdot)d\nu' \leq \nu'(\{j\} \times \mathbb{A}).$$

Both sides add up to one when summed over j, so equality must hold for all j. Therefore $\nu' \in G$ by Lemma 10.1. ∎

These lemmas form the backdrop for the convex programming problem we describe next.

10.2.2 The convex programming problem

Recall that if $\pi \in \Pi^{RS}$ has an ergodic probability measure η and $x_0 \in \text{support}(\eta)$ a.s., then the ergodic cost a.s. equals $\int r d\hat{\eta}_\pi$. This suggests the convex programming problem:

Minimize $\int r d\mu$ over $\mu \in G$.

Equivalently:

Minimize $\int r d\mu$ over $\{\mu \in \mathcal{P}(\mathbb{X} \times \mathbb{A})$: (10.2) holds $\}$.

This displays it as an infinite dimensional linear program.

Let $\mu^* \in G$ be such that

$$\alpha \stackrel{\Delta}{=} \inf_{\mu\in G} \int r d\mu \leq \int r d\mu^* \leq \alpha + \epsilon$$

for some $\epsilon \geq 0$. (Note that such a μ^* is guaranteed to exist for $\epsilon > 0$.) Such a μ^* is said to be ϵ-optimal (or optimal if $\epsilon = 0$).

Lemma 10.7 *For $\epsilon \geq 0$, if there is an ϵ-optimal $\mu^* \in G$, then there is an ϵ-optimal $\overline{\mu} \in G_e$.*

Proof. By Choquet's theorem, there exists a $\Phi \in \mathcal{P}(G_e)$ such that

$$\alpha \leq \int r d\mu^* = \int_{G_e} \left(\int r d\mu \right) \Phi(d\mu) \leq \alpha + \epsilon.$$

Thus for some $\overline{\mu} \in \text{support}(\Phi)$, $\alpha \leq \int r d\overline{\mu} \leq \alpha + \epsilon$. ∎

We next consider two conditions under which an optimal μ^* exists.

Case 1: The near-monotone case

Call $r(\cdot, \cdot)$ near-monotone if

$$\liminf_{i \to \infty} \min_u r(i, u) > \alpha. \tag{10.5}$$

(Note that this would automatically imply that r was bounded from below.)

Lemma 10.8 *Under (10.5), an optimal $\mu^* \in G_e$ exists.*

Proof. Let $\mu_n \in G, n \geq 1$, be such that $\int r d\mu_n \downarrow \alpha$. Viewing $G \subset \mathcal{P}(\overline{\mathbb{X}} \times \mathbb{A})$ as before, drop to a subsequence if necessary and suppose that $\mu_n \to \overline{\mu}$ in $\mathcal{P}(\overline{\mathbb{X}} \times \mathbb{A})$. Write $\overline{\mu} = \delta \mu' + (1 - \delta)\mu''$ as in (10.4). Pick $N \geq 1$, $\epsilon > 0$ such that $\inf_u r(i, u) \geq \alpha + \epsilon$ for $i \geq N$. For $n \geq 1$, define

$$r_n(i, u) = r(i, u) I\{i \leq N + n\} + (\alpha + \epsilon) I\{i > N + n\}.$$

Then $\forall\ j$,

$$\alpha \geq \liminf_{n \to \infty} \int r_j d\mu_n = \delta \int r_j d\mu' + (1 - \delta)(\alpha + \epsilon).$$

Let $j \uparrow \infty$ on the right to obtain

$$\alpha \geq \delta \int r d\mu' + (1 - \delta)(\alpha + \epsilon) \geq \delta \alpha + (1 - \delta)(\alpha + \epsilon).$$

Thus $\delta = 1$ and $\int r d\mu' = \alpha$. The claim now follows from Lemma 10.7. ∎

Case 2: The stable case

Here we assume G to be compact. (A 'Stochastic Lyapunov' condition that ensures this is given in the Appendix.) Since $\mu \to \int r d\mu$ is lower semicontinuous, it attains a minimum on G, hence, by Lemma 10.7, on G_e.

Note that lower semicontinuity of the above map is the key requirement for the above to hold. Thus we can relax our condition that r be bounded from below to, e.g.: There exists $h : \mathbb{X} \to \mathbb{R}^+$ such that $r(i, a) \geq -h(i)\ \forall i, a$ and $sup_{\nu \in G} \int h d\nu < \infty$.

We still need to show that the 'optimum' is an optimum with respect to all admissible policies. In 'Case 1', this is true:

Lemma 10.9 *In the near-monotone case, under any admissible policy,*

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} r(x_m, a_m) \geq \alpha \text{ a.s.} \tag{10.6}$$

To prove this, we introduce $\mathcal{P}(\mathbb{X} \times \mathbb{A})$-valued process $\{\nu_n\}$ of 'empirical measures', defined by

$$\int f d\nu_n = \frac{1}{n} \sum_{m=0}^{n-1} f(X_m, Z_m), \ f \in C_b(X \times U).$$

Lemma 10.10 *$\nu_n \to \overline{G}$ a.s.*

Proof. Let $\mathcal{C} = \{I_{\{i\}}, i \in \mathbb{X}\} \subset C_b(\mathbb{X})$. Then by the strong law of large number for martingales ([15], pp. 53-54),

$$\sum_i g(i)\nu_n(\{i\}, \mathbb{A}) - \int_{\mathbb{X} \times \mathbb{A}} \left(\sum_j p(j/i, a) g(j) \right) d\nu_n \to 0 \quad \text{a.s.} \tag{10.7}$$

Since $\mathcal{C}$ is countable, this holds for all sample points outside a common zero probability set. Fix one such sample point. Let $\overline{\nu}$ be a limit point of $\{\nu_n\}$ in $\mathcal{P}(\overline{\mathbb{X}} \times \mathbb{A})$ and write $\overline{\nu} = \delta\nu' + (1-\delta)\nu''$ with $\delta \in [0,1]$ and $\nu', \nu'' \in \mathcal{P}(\overline{\mathbb{X}} \times \mathbb{A})$ supported on $\mathbb{X} \times \mathbb{A}$ and $\{\infty\} \times \mathbb{A}$ resp. By (10.7), ν' satisfies (10.2) whenever $\delta > 0$. Identifying ν' with its restriction to $\mathbb{X} \times \mathbb{A}$ by abuse of notation, we have $\nu' \in G$, i.e. $\overline{\nu} \in \overline{G}$. ∎

Proof of Lemma 10.9 Consider a sample point for which Lemma 10.4 holds and let $\overline{\nu}$ be a limit point of $\{\nu_n\}$ as above, say, $\nu_{n(l)} \to \overline{\nu}$. Then for r_n's as in Lemma 10.8 above,

$$\liminf_{l \to \infty} \int r d\nu_{n(l)} \geq \int r_j d\overline{\nu} \geq \delta \int r_j d\nu' + (1-\delta)(\alpha + \epsilon), \ j \geq 1.$$

Let $j \uparrow \infty$ on the right to obtain

$$\liminf_{l \to \infty} \int r d\nu_{n(l)} \geq \delta \int r d\nu' + (1-\delta)(\alpha+\epsilon) \geq \delta\alpha + (1-\delta)(\alpha+\epsilon),$$

in view of the preceding lemma. The claim follows. ∎

Case 2 needs more work. For simplicity, we assume that there is a single communicating class under any $\pi \in \Pi^{RS}$. Let $\mathcal{F}_n = \sigma(x_m, a_m, m \leq n), n \geq 0$, for an arbitrary admissible control sequence $\{a_n\}$ and the corresponding chain $\{x_n\}$. Let $\tau(i) = \min\{n > 0 : x_n = i\}, i \in S$.

Lemma 10.11 *For any $\{\mathcal{F}_n\}$-stopping time τ, the regular conditional law of $x_{\tau+n}, n \geq 0$, on $\{\tau < \infty\}$ is a.s. the law of a controlled Markov chain on $\mathbb{X}$ with action space $\mathbb{A}$ and transition probabilities $p(\cdot/\cdot, \cdot)$.*

Proof. This is a straightforward consequence of the easily verified fact

$$\mathbb{E}\left[\left(I_{\{x_{\tau+n+1}=i\}} - p\left(i/x_{\tau+n}, a_{\tau+n}\right) \right) / \mathcal{F}_\tau \right] = 0 \text{ a.s.}$$

on $\{\tau < \infty\}$ for $i \in \mathbb{X}, n \geq 0$. ∎

Impose the additional condition:

$$(\dagger) \qquad \sup \mathbb{E}[\tau(0)^2/x_0 = 0] < \infty,$$

where the supremum is over all admissible control policies.

Remark 10.1 *To see that* (†) *is not implied by stability alone, consider a chain on* $\{(0,0),(0,1),(1,1),(0,2),(1,2),(2,2),(0,3),\dots\}$ *with transition probabilities: for* $j \geq 1$,

$$\begin{aligned} p((i,j)/(i+1,j)) &= 1, \quad 0 \leq i < j, \\ p((0,0)/(0,j)) &= 1, \\ p((j,j)/(0,0)) &= Cj^{-3}, \text{ where } C = \left(\sum_n n^{-3}\right)^{-1}. \end{aligned}$$

Then $\mathbb{E}[\tau(0)/X_0 = 0] < \infty$, *but* $\mathbb{E}[\tau(0)^2/X_0 = 0] = \infty$.

Lemma 10.12 *Under* (†), $\{\nu_n\}$ *are a.s. tight in* $\mathcal{P}(\mathbb{X} \times \mathbb{A})$.

Proof. Define stopping times $\tau_0 = 0, \tau_{n+1} = \min\{m > \tau_n : x_m = 0\}, n \geq 0$. By (†) and Lemma 10.11, it follows that

$$\sup_n \mathbb{E}[(\tau_{n+1} - \tau_n)^2] < \infty.$$

By the strong law of large numbers for martingales, we then have, for $N \geq 1$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left(\sum_{m=\tau_i}^{\tau_{i+1}-1} I_{\{x_m \geq N\}} - \mathbb{E}\left[\sum_{m=\tau_i}^{\tau_{i+1}-1} I_{\{x_m \geq N\}} / \mathcal{F}_{\tau_i} \right] \right) = 0 \text{ a.s.}$$

Thus

$$\begin{aligned} \limsup_{n\to\infty} \nu_n(\{N, N+1, \dots\} \times \mathbb{A}) &= \limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} I_{\{x_m \geq N\}} \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left(\sum_{m=\tau_i}^{\tau_{i+1}-1} I_{\{x_m \geq N\}} \right) \\ &= \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}\left[\sum_{m=\tau_i}^{\tau_{i+1}-1} I_{\{x_m \geq N\}} / \mathcal{F}_{\tau_i} \right] \\ &\leq \sup \mathbb{E}\left[\sum_{m=0}^{\tau(0)-1} I_{\{x_m \geq N\}} / x_0 = 0 \right], \end{aligned}$$

where the supremum is over all admissible policies and the last inequality holds by Lemma 10.11. Fix a sample point for which the foregoing holds true for all $N \geq 1$. A standard dynamic programming argument allows us to replace the above supremum by the supremum over Π^{RS}, which in turn equals

$$\sup_{\pi \in \Pi^{RS}} \left(\mathbb{E}[\tau(0)/X_0 = 0] \left(\sum_{i \geq N} \eta_\pi(i) \right) \right),$$

with η_π = the unique invariant probability under $\pi \in \Pi^{RS}$. By (†), this is bounded by a constant times $\sup_{\pi \in \Pi^{RS}} \left(\sum_{i \geq N} \eta_\pi(i) \right)$. This can be made arbitrarily small by increasing N in view of the compactness (hence tightness) of G. The claim follows. ∎

Corollary 10.1 *For the stable case, under (†), (10.6) holds.*

Proof. This follows as in Lemma 10.9 on observing that, outside a zero probability set, any limit point $\overline{\nu} = \delta\nu' + (1-\delta)\nu''$ of $\{\nu_n\}$ as in Lemma 10.10 must have $\delta = 1$ by virtue of the above lemma. ∎

The next theorem summarizes our main results, interpreted in terms of the original Markov decision process.

Theorem 10.2 (*i*) *In the near-monotone case, there exists a stationary policy that is optimal among all admissible policies.*
(*ii*) *In the stable case, there exists a stationary policy that is optimal in* Π^{RS}. *If* (†) *holds, it is also optimal among all admissible policies.*

It should be kept in mind that if $\hat{\eta}_\pi \in G_e$ is optimal, the support of η may not be the entire $\mathbb{X}$. It is, however, often reasonable to suppose that from any $i \in \mathbb{X}$, support(η) is reachable with probability one under *some* control policy. The optimal policy starting from i then is to use the above policy till support(η) is reached and then switch to π.

10.3 EXTENSIONS

10.3.1 Other cost criteria

Here we very briefly sketch the corresponding developments for a few other criteria. Notably, three of these: the finite horizon cost, the discounted cost and the cost up to exit time, defined in section 1.2, have been extensively treated in [12]. Therefore we shall give a bare sketch of the main results, referring the reader to [12] for further details.

Consider first the discounted cost on the infinite time horizon. Fix the initial distribution (i.e. the distribution of x_0) as, say, $\lambda \in \mathcal{P}(\mathbb{X})$. Under any admissible control policy, we may define the discounted occupation measure $\hat{\eta}$ on $\mathbb{X} \times \mathbb{A}$ by:

$$\int g d\hat{\eta} = \mathbb{E}[\sum_{n=0}^{\infty} \delta^n g(x_n, a_n)]$$

for $g \in C_b(\mathbb{X} \times \mathbb{A})$. In particular, when the control policy is $\pi \in \Pi^{RS}$, we may denote $\hat{\eta}$ by $\hat{\eta}_\pi$ to make the dependence explicit. The main results of the convex analytic approach in this context are then summarized as:

Theorem 10.3 *The discounted occupation measure $\hat{\eta}_\pi$ corresponding to $\pi \in \Pi^{RS}$ is characterized by:*

$$\sum_{i\in\mathbb{X}} g(i)\hat{\eta}_\pi(i,\mathbb{A}) = \sum_{i\in\mathbb{X}} \lambda(i)g(i) + \delta \int_{\mathbb{X}\times\mathbb{A}} \left(\sum_{j\in\mathbb{X}} p(j/i,a)g(j)\right) d\hat{\eta}_\pi \qquad (10.8)$$

for $g \in C_b(\mathbb{X}\times\mathbb{A})$.
(ii) The set G of discounted occupation measures under all admissible policies is the same as that under all $\pi \in \Pi^{RS}$ and is convex compact.
(iii) The extreme points of G correspond to stationary policies. In particular, an optimal stationary policy exists.

For the problem of cost until exit time τ, one may likewise define the associated occupation measure $\hat{\eta}$ on $D \times \mathbb{A}$ by:

$$\int g d\hat{\eta} = \mathbb{E}[\sum_{n=0}^{\tau-1} g(x_n, a_n)]$$

for $g \in C_b(D\times\mathbb{A})$. One usually imposes suitable conditions to ensure that $\mathbb{E}[\tau] < \infty$, which in turn ensures that the above defines a finite positive measure on $D \times \mathbb{A}$. An exact analog of the above theorem can then be proved, modulo two changes: The summations over $i \in \mathbb{X}$ in the equation (10.8) get replaced by summations over $i \in D$ and δ is set equal to 1.

The finite horizon control problem in turn can be reduced to the above by treating the time variable as an additional state variable, whereby the controlled Markov chain in question now is $\{(x_n, n)\}$, not just $\{x_n\}$. The problem then is to control this chain up to the first exit time from the set $\mathbb{X}\times\{0,1,\cdots,T\}$. One can then establish the appropriate counterpart of the above theorem. This, when translated back into the original set-up, calls for one important change: Define a Markov policy to be a control policy of the type $a_n = v(x_n, n),\ n \geq 0$, for some $v : \mathbb{X}\times\{0,1,2,\cdots\} \to \mathbb{A}$ and likewise, a randomized Markov policy to be one in which for each n, a_n is conditionally independent of $x_m, a_m, m < n$, given x_n, but its conditional law given the latter is not required to be independent of n. The above results then hold with 'stationary' (resp., 'randomized stationary') policy replaced by 'Markov' (resp., 'randomized Markov') policy throughout. Thus in particular, one is assured of an optimal policy that is Markov, but not necessarily stationary. This, of course, is to be expected, since in the finite horizon problem, the time count matters.

The expected ergodic cost criterion has a similar flavor to the preceding section, so we won't treat it separately. Some key references for this are given in the bibliographical note at the end.

The total cost criterion suggests the following definition for the associated occupation measure $\hat{\eta}$:

$$\int g d\hat{\eta} = \mathbb{E}[\sum_{n=0}^{\infty} g(x_n, a_n)]$$

for compactly supported $g \in C_b(\mathbb{X}\times\mathbb{A})$. But there can be problems with this: The r.h.s. may not be well-defined or finite for some or all policies. One

typically ensures that the total cost is well-defined and finite for some policies at least by imposing suitable restrictions. Even then, the analysis for this case is much harder and the results are correspondingly weaker. In particular, it is possible that for a given initial state and a given transient nonstationary policy, there is no randomized stationary policy that replicates the occupation measure [29]. An excellent account of this criterion appears in a companion article [27].

There is an important shortcoming of the convex analytic method when applied to these criteria. Theorem 10.3 above guarantees the existence of an optimal stationary policy $a_n = v(x_n),\ n \geq 0$, for a given initial law λ. It does not, however, guarantee that the *same* function $v(\cdot)$ will work for any other initial law, something which the classical dynamic programming approach delivers with no extra effort.

10.3.2 General state spaces

In this subsection, we briefly outline extensions to more general, viz., Polish $\mathbb{X}$. Denote by $p(dy/x,a) \in \mathcal{P}(\mathbb{X})$ the transition kernel for $\{x_n\}$ with $x \in \mathbb{X}, a \in \mathbb{A}$. This is assumed to be continuous in (x,a). A $\pi \in \Pi^{RS}$ will be given by a probability kernel $x \in \mathbb{X} \to \pi(x,da) \in \mathcal{P}(\mathbb{A})$, with the corresponding transition kernel given by

$$\overline{p}(dy/x,\pi) \stackrel{\Delta}{=} \int_U \pi(x,da)p(dy/x,a).$$

If $\eta(dx) \in \mathcal{P}(\mathbb{X})$ is invariant under π, define the corresponding ergodic occupation measure by

$$\nu(dx,da) = \eta(dx)\pi(x,da). \tag{10.9}$$

The set G of ergodic occupation measures in $\mathcal{P}(\mathbb{X} \times \mathbb{A})$ is then characterized by the fact that their disintegration (10.9) satisfies

$$\int_{\mathbb{X}} \int_{\mathbb{A}} \eta(dx)\pi(x,da)p(dy/x,a) = \eta(dy). \tag{10.10}$$

It is easy to deduce from this that G is closed. Also, if $\nu_i(dx,da) = \eta_i(dx)$ $\pi_i(x,da) \in G, i = 1,2$, then

$$\nu(dx,da) = b\nu_1(dx,da) + (1-b)\nu_2(dx,da), b \in (0,1),$$

satisfies (10.10) with $\eta(dx) = b\eta_1(dx) + (1-b)\eta_2(dx)$ and

$$\pi(x,da) = \Lambda(x)\pi_1(x,da) + (1-\Lambda(x))\pi_2(x,da)$$

for $\Lambda(x) = b\frac{d\eta_1}{d\eta}(x)$. Thus G is convex.

Assume as before that $r : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ is continuous and bounded from below. The definition of 'stable case' above extends to general $\mathbb{X}$ in the obvious manner. As for near-monotonicity, the definition can be retained if $\mathbb{X}$ is locally compact. If not, one must allow $r(\cdot,\cdot)$ to be extended real valued and suppose that the set $D_y = \{x : \inf_a r(x,a) \leq y\}$ is compact for each $y > 0$. The

reason $+\infty$ must be allowed as a possible value for $r(\cdot,\cdot)$ is as follows: By the Baire category theorem, a Polish space $\mathbb{X}$ that is not locally compact cannot be σ-compact and thus cannot equal $\cup_n D_n$ for compact $\{D_n\}$.

Under either condition, a cost-minimizing sequence in G can be shown to be relatively sequentially compact and one can mimic the arguments of section 2 to conclude that the map $\mu \in G \to \int r d\mu$ attains its minimum on G, hence on G_e.

If $\mathbb{X}$ is not compact but locally compact, we can embed it homeomorphically into its one point compactification $\overline{\mathbb{X}} = \mathbb{X} \cup \{\infty\}$. Define empirical measures $\nu_n \in \mathcal{P}(\overline{\mathbb{X}} \times \mathbb{A})$ by

$$\nu_n(C \times B) = \frac{1}{n} \sum_{m=0}^{n-1} I_{\{x_m \in C, a_m \in B\}}, \quad n \geq 1,$$

where C, B are Borel in $\overline{\mathbb{X}}, \mathbb{A}$ resp. Let $\overline{G} = \{\mu \in \mathcal{P}(\overline{\mathbb{X}} \times \mathbb{A}) : \mu = b\mu_1 + (1-b)\mu_2$ for some $b \in [0,1], \mu_1 \in G, \mu_2(\{\infty\} \times \mathbb{A}) = 1\}$.

Lemma 10.13 $\nu_n \to \overline{G}$ a.s.

This is proved as in Lemma 10.10, with $\mathcal{C}$ any countable convergence determining class in $C_b(\mathbb{X})$ (see [15], Chapter 2). For the near-monotone case, Lemma 10.2 now follows as before. The stable case is harder. Assume $\mathbb{X}$ to be locally compact as before and let B_1, B_2 be concentric closed balls of radii $r_1 < r_2$ resp. in $\mathbb{X}$. Define

$$\begin{aligned} \sigma_1 &= \min\{m \geq 0 : x_m \in \overline{B_2}^c\}, \\ \tau_1 &= \min\{m > \sigma_1 : x_m \in B_1\}, \\ \sigma_{n+1} &= \min\{m > \tau_n : x_m \in \overline{B_2}^c\}, \\ \tau_{n+1} &= \min\{m > \sigma_n : x_m \in B_1\}, \end{aligned}$$

for $n \geq 1$. Assume that for any open $B \subset S, P(\cup_{n \geq 1}\{x_n \in B\}) = 1$ regardless of the initial condition or control strategy. (This is a 'recurrence' condition.) Then $\tau_n, \sigma_n < \infty$ a.s. $\forall\ n$. The counterpart of (†) now is

$$\sup_{x \in B_1} \sup_{\Pi} \mathbb{E}[\tau_1^2 / X_0 = x] < \infty, \tag{10.11}$$

where the first supremum is over $\Pi \stackrel{\Delta}{=}$ the set of all admissible control policies.

Let $D_N, N \geq 1$, be a concentric family of open balls in $\mathbb{X}$ with radii $N \geq 1$ respectively. Now mimic the proof of Lemma 10.12 to conclude that

$$\begin{aligned} \limsup_{n \to \infty} \nu_n(D_N) &\leq \sup_{x \in B_1} \sup_{\Pi} \mathbb{E}\left[\sum_{m=0}^{\tau_1} I_{\{x_m \notin D_N\}} / x_0 = x\right] \text{ a.s.} \\ &= \sup_{x \in B_1} \sup_{\Pi^{RS}} E\left[\sum_{m=0}^{\tau_1} I_{\{x_m \notin D_N\}} / x_0 = x\right] \end{aligned} \tag{10.12}$$

where the equality follows by a standard dynamic programming argument. Now let $\pi_m \to \pi_\infty$ in $\Pi^{RS}, x_m \to x_\infty$ in B_1 and let $x_n^m, n \geq 1$, denote the chain

governed by π_m with $x_0^m = x_m, m = 1, 2, \cdots, \infty$. Standard arguments along the lines of [12], pp. 26-28, show that $\{x_n^m\} \to \{x_n^\infty\}$ in law. By Skorohod's theorem ([15], pp.23-24), we may view these processes as being defined on a common probability space and the convergence as being a.s. Define stopping times $\{\tau_m^n, n \geq 1\}, m \geq 1$, correspondingly.

Then

$$\liminf_{m\to\infty} \sum_{\ell=0}^{\tau_1^m} I_{\{x_l^m \notin D_N\}} \geq \sum_{\ell=0}^{\tau_1^\infty} I_{\{x_\ell^\infty \notin D_N\}}$$

a.s. and therefore

$$\liminf_{m\to\infty} \mathbb{E}\left[\sum_{\ell=0}^{\tau_1^m} I_{\{x_\ell^m \notin D_N\}}\right] \geq \mathbb{E}\left[\sum_{\ell=0}^{\tau_1^\infty} I_{\{x_\ell^\infty \notin D_N\}}\right].$$

Thus the map

$$(x, \pi) \to \mathbb{E}_\pi\left[\sum_{m=0}^{\tau_1} I_{\{x_\ell^m \notin D_N\}} / x_0 = x\right],$$

with $\mathbb{E}_\pi[\cdot]$ denoting the expectation under $\pi \in \Pi^{RS}$, is lower semicontinuous. As $N \uparrow \infty$, the r.h.s. $\downarrow 0$. By Dini's theorem, this convergence is uniform in $(x, \pi) \in B_1 \times \Pi^{RS}$. Thus the r.h.s. of (10.12) can be made arbitrarily small by choosing N sufficiently large. We have proved:

Lemma 10.14 *Under (10.11), $\{\nu_n\}$ remains tight a.s. under arbitrary control strategies.*

Corollary 10.2 *The conclusion of Corollary 10.1 holds in the present set-up.*

This follows exactly as for countable $\mathbb{X}$ in view of the foregoing. Given that the optimum, when it exists, is attained on G_e = the set of extreme points of G, we shall characterize G_e next, albeit for a special case. We assume in addition to local compactness the following conditions: (i) There exists a σ-finite nonnegative measure λ on $\mathbb{X}$ such that $p(dy/x, a) << \lambda(dy)\ \forall\, x, a$ and furthermore, (ii) if $\varphi(x, a, y)$ denote the corresponding density (i.e. $p(dy/x, a) = \varphi(x, a, y)\lambda(dy)$), then $\varphi(\cdot, \cdot, \cdot)$ is continuous, and $\{\varphi(x, a, \cdot) : x \in \mathbb{X},\ a \in \mathbb{A}\}$ is bounded equicontinuous and bounded away from zero from below on compacts, uniformly w.r.t. x, a. (These conditions are satisfied, e.g. for several stochastic systems in $\mathbb{R}^d$ with additive Gaussian white noise).

This has the following important consequence: If η is an invariant probability distribution under $\pi \in \Pi^{RS}$, then

$$\eta(dy) = \int_{\mathbb{X}} \eta(dx)\pi(x, da)\varphi(x, a, y)dy,$$

implying that η has a density w.r.t. λ given by

$$\psi(y) = \int_{\mathbb{X}} \eta(dx)\pi(x, da)\varphi(x, a, y) > 0.$$

Then any two ergodic probability distributions under π are mutually absolutely continuous, hence identical. That is, if π admits an invariant probability distribution, it is unique and the corresponding $\{x_n\}$ ergodic. Suppose now that G is compact.

Lemma 10.15 *$Q \triangleq \{\eta \in \mathcal{P}(\mathbb{X}) : \eta$ invariant under some $\pi \in \Pi^{RS}\}$ is compact in the total variation norm topology.*

Proof. Let $\eta_n \in Q$ be invariant under $\pi_n \in \Pi^{RS}, n \geq 1$. Then $\nu_n(dx, da) \triangleq \eta_n(dx)\pi_n(x, da) \in G \; \forall n$. Since G is compact, $\nu_n \to \nu_\infty$ (say) in $\mathcal{P}(\mathbb{X} \times \mathbb{A})$ along a subsequence, denoted by $\{\nu_n\}$ again by abuse of notation. Then $\eta_n \to \eta_\infty$ in $\mathcal{P}(\mathbb{X})$. Let $\psi_n = d\eta_n/d\lambda$ for $n = 1, 2, \ldots, \infty$. Under our assumptions, $\{\psi_n(\cdot)\}$ are equicontinuous and bounded. By the Arzela-Ascoli theorem, we may drop to a subsequence if necessary and suppose that $\psi_n(\cdot) \to \overline{\psi}(\cdot)$ in $C(S)$. Thus for compactly supported $f \in C(S)$,

$$\int f(y)\psi_n(y)\lambda(dy) \to \int f(y)\overline{\psi}(y)\lambda(dy).$$

But

$$\int f(y)\psi_n(y)\lambda(dy) = \int f d\eta_n \to \int f d\eta_\infty.$$

Thus $\overline{\psi} = \psi_\infty$ and $\psi_n \to \psi_\infty$ in $C(S)$. By Scheffe's theorem ([15], pp.26), $\eta_n \to \eta_\infty$ in the total variation norm. This completes the proof. ∎

We shall now characterize the set G_e for the stable case, i.e. when all $\pi \in \Pi^{RS}$ are stable and the corresponding G compact. Let $\pi_i \in \Pi^{RS}, i = 0, 1, 2$, be such that

$$\pi_0(x, da) = b\pi_1(x, da) + (1 - b)\pi_2(x, da), \pi_1 \neq \pi_2,$$

with $b \in (0, 1)$. Let $\eta_i, \nu_i, i = 0, 1, 2$, denote the corresponding invariant distributions and ergodic occupation measures resp., with $\psi_i = d\eta_i/d\lambda$, $i = 0, 1, 2$. Finally, let $\tilde{Q} = \{\psi : \mathbb{X} \to \mathbb{R} : \psi = d\eta/d\lambda$ for some $\eta \in Q\}$.

Lemma 10.16 *$\nu_0 \notin G_e$.*

Proof. By setting $\pi_i(x, da) = \pi_0(x, da), i = 1, 2$, for x outside a ball $B_R \subset \mathbb{X}$ of a sufficiently large radius $R > 0$, we may assume without loss of generality that $\pi_1(x, da) = \pi_2(x, da)$ for $x \notin B_R$. Let $\tilde{\pi} \in \Pi^{RS}$, with $\tilde{\eta}(dx) = \tilde{\psi}(x)\lambda(dx)$ the corresponding invariant measure and $\tilde{\nu}(dx, da) = \tilde{\eta}(dx)\tilde{\pi}(x, da)$ the corresponding ergodic occupation measure. Define $\hat{\pi} \in \Pi^{RS}$ by

$$\pi_0(x, da) = \frac{c\psi_1(x)\pi_1(x, da) + (1 - c)\tilde{\psi}(x)\hat{\pi}(x, da)}{c\psi_1(x) + (1 - c)\tilde{\psi}(x)} \tag{10.13}$$

with $c \in (0, 1)$. In order that this indeed define a $\hat{\pi} \in \Pi^{RS}, c$ must be chosen so that

$$\frac{c\psi_1(x)}{c\psi_1(x) + (1 - c)\tilde{\psi}(x)} \leq b \quad \forall x \in B_R.$$

(For $x \notin B_R, \pi_0(x, da) = \pi_1(x, da)$ and any choice of c will do.) For $x \in B_R$, such a choice of c is possible provided

$$\inf_{x \in B_R} \{\psi(x) : \psi \in \tilde{Q}\} > 0.$$

This is indeed so by our assumptions: If not, we have $\psi_n \in \tilde{Q}, x_n \in B_R, n \geq 1$, with $\psi_n(x_n) \to 0$. By dropping to a subsequence if necessary and using the Arzela-Ascoli theorem, let $\psi_n(\cdot) \to \psi_\infty(\cdot)$ in $C(\mathbb{X}), x_n \to x_\infty$ in B_R. Then $\psi_\infty \in \tilde{Q}$, implying $\psi_\infty(x_\infty) > 0$, a contradiction. Thus we can choose a $c \in (0, 1)$ as desired. Fix one such c. Then (10.13) defines a map $\tilde{\pi} \to \hat{\pi}$, or equivalently, $\tilde{\eta} \in Q \to \hat{\eta} \in Q, \hat{\eta}$ being the invariant distribution under $\hat{\pi}$. This map is continuous in the total variation norm: If $\eta_n(dx) = \psi_n(dx)\lambda(dx) \to \eta_\infty(dx) = \psi_\infty(x)\lambda(dx)$ in total variation, $\psi_n(\cdot) \to \psi_\infty(\cdot)$ in $C(\mathbb{X})$ as in the proof of Lemma 10.15. Letting $\hat{\psi}_n$ denote the image of ψ_n under the above map for $n = 1, 2, \ldots, \infty$, it is easily verified from (10.13) that $\hat{\psi}_n \to \hat{\psi}_\infty$ pointwise and hence, by Scheffe's theorem, $\hat{\psi}_n(x)\lambda(dx) \to \hat{\psi}_\infty(x)\lambda(dx)$ in total variation. Since Q is compact convex, Schauder fixed point theorem guarantees a fixed point for this map. That is, there exists a $\pi^* \in \Pi^{RS}$ with associated invariant distribution $\eta^*(dx) = \psi^*(x)\lambda(dx)$ such that

$$\pi_0(x, da) = \frac{c\psi_1(x)\pi_1(x, da) + (1 - c)\psi^*(x)\pi^*(x, da)}{c\psi_1(x) + (1 - c)\psi^*(x)}$$

Since $\pi_0 \neq \pi_1, \pi_0 \neq \pi^*$. It is easily deduced from this that for $\nu^*(dx, da) = \eta^*(dx)\pi^*(x, da)$,

$$\nu_0 = c\nu_1 + (1 - c)\nu^*, \ \nu_1 \neq \nu^*,$$

i.e. $\nu_0 \notin G_e$. ∎

If G is not compact, we need the following additional condition, called the 'stability under local perturbations': If $\pi(x, da), \pi'(x, da) \in \Pi^{RS}$ agree for x outside a bounded subset of $\mathbb{X}$ and one of them is stable, so is the other. Note that G is closed for the same reasons as before. Let $\pi_i \in \Pi^{RS}, i = 0, 1, 2$, be as before with $\nu_i, i = 0, 1, 2$, defined correspondingly as before.

Lemma 10.17 *$\nu_0 \notin G_e$.*

The proof is essentially the same as before: Recall from the proof of Lemma 10.16 that for any stable $\pi \in \Pi^{RS}$, it suffices to look at the subset of G corresponding to the ergodic occupation measures for $\tilde{\pi} \in \Pi^{RS}$ that agree with π outside a ball B_R of radius $R > 0$ sufficiently large. By our assumption of 'stability under local perturbations', all such $\tilde{\pi}$ are stable, so the same proof works.

Now argue as for the countable case to get the following:

Theorem 10.4 *(i) In the stable case, there is a stationary policy that is optimal in Π^{RS} for the pathwise ergodic cost. If (10.10) holds, it is also optimal among all admissible control policies.*
(ii) In the near-monotone case with stability under local perturbations, there exists a stationary policy that is optimal for the pathwise ergodic cost among all admissible policies.

Unlike in the preceding section, the initial law is no concern here because of ergodicity of the optimal Markov process.

10.3.3 Dual problems

The 'dual' linear program to the infinite dimensional linear program over occupation measures establishes a link with the traditional dynamic programming approach. To see how this comes by, we first recall the relevant results from the theory of infinite dimensional linear programming [5].

Two topological vector spaces X, Y are said to form a dual pair if there exists a bilinear form $\langle \cdot, \cdot \rangle : X \times Y \to R$ such that the functions $x \to \langle x, y \rangle$ for $y \in Y$ separate points of X and the functions $y \to \langle x, y \rangle$ for $x \in X$ separate points of Y. Endow X with the coarsest topology required to render the former family of maps continuous, and Y with the dual topology. Let P be the positive cone in X and $P^* \subset Y$ the dual cone: $P^* = \{y \in Y : \langle x, y \rangle \geq 0, x \in P\}$.

Let Z, W be another dual pair topologized in a similar manner and $F : X \to Z$ a continuous linear map. Define $F^* : W \to X^*$ by $\langle Fx, w \rangle = \langle x, F^* w \rangle, x \in X, w \in W$. The primal LP problem is:

$$\text{Minimize } \langle x, c \rangle \text{ s.t.} \quad Fx = b, \quad x \in P,$$

with $b \in Z, c \in Y$ prescribed. Let α denote the infimum of $\langle x, c \rangle$ subject to these constraints. The dual problem is

$$\text{Max } \langle b, w \rangle \text{ s.t.} \quad -F^* w + c \in P^*, \quad w \in W.$$

Let α' denote the supremum of $\langle b, w \rangle$ subject to these constraints. It is known that $\alpha \geq \alpha'$. Let $C = \{x \in P : Fx = b\}$, $D = \{(Fx, \langle x, c \rangle) : x \in P\}$. The following result gives conditions for the absence of a 'duality gap'.

Theorem 10.5 ([5], p.53) *If $C \neq \emptyset, D$ is closed and $x \to \langle x, c \rangle$ attains its minimum on C, then $\alpha = \alpha'$.*

Applying this to the linear programming formulation of ergodic control in section 2, the dual problem and Theorem 10.4 lead to the following characterization of the optimal cost α:

Corollary 10.3

$$\alpha = \sup \left\{ b : \min_a \left(r(i,a) + f(i) - \sum_j p(j/i,a) f(j) \right) \geq b, f : \mathbb{X} \to \mathbb{R} \right\}$$

.

Similar results are possible for general state spaces and other cost criteria. See [31], [32] for extensive accounts.

10.4 MULTIOBJECTIVE PROBLEMS

10.4.1 Constrained control: preliminaries

The first multiobjective problem we consider is that of minimizing one cost with other secondary costs being required to satisfy prescribed bounds. Again

we stick to the pathwise ergodic set up of section 2 and shall use the notation therein. We start with some convex analytic preliminaries.

Recall the set G of ergodic occupation measures. View it as a subset of the space of finite signed measures on S. Let H be its intersection with $m \geq 1$ closed half spaces thereof. Then H is closed convex. Let H_e be the set of its extreme points. The following is a special case of a result of Dubins [21].

Lemma 10.18 *Any $\nu_0 \in H_e$ can be expressed as a convex combination of k points in G_e for some $k \leq m+1$.*

Proof. Suppose G is compact. Suppose the claim is false and ν_0 is expressible as a convex combination of $k = m+2$ points in G, but not less. (A similar proof works for higher k.) Then ν_0 lies in the interior of an $(m+2)$-simplex B formed by these points in G_e. Let M = the $(m+1)$-dimensional affine space (i.e. translate of a linear subspace) generated by B, and C an open ball in M centered at ν_0 and contained in the relative interior of B. Then $C \subset B \subset G$. Now consider the intersections with M of the m hyperplanes that form the boundaries of the half spaces defining H. Since at most m of them can intersect at a time, any intersection with M of the intersections of these hyperplanes must have a co-dimension of at most m in M and thus cannot have a corner in the interior of C. Thus $\nu_0 \notin H_e$, a contradiction. For noncompact G, argue as above with $\bar{G}_e$ in place of G_e and observe that if ν_0 is a strict convex combination of points in $\bar{G}_e$, the latter must be in G_e, because ν_0 would otherwise assign a strictly positive probability to $\{\infty\} \times \mathbb{A}$, a contradiction. The rest is as before.

In case ν_0 cannot be expressed as a convex combination of finitely many extreme points of G, a simple adaptation of the above proof works. We claim that for any $j \geq 1$, we can find j linearly independent finite line segments in G which have ν_0 at their center. If this were not so for, say, $j = j_0 + 1$, then ν_0 would be in a j_0-dimensional face G' of G and therefore expressible as a convex combination of $j_0 + 1$ extreme points of G' by Caratheodory's theorem ([18], p.106), therefore of G. This goes against the hypothesis, proving the claim. Now take $j \geq m+2$, consider the polytope generated by the end points of these line segments, and argue as before. ∎

Write ν_0 as

$$\nu_0(\{i\}, da) = \eta_0(i)\pi(i, da), \ i \in \mathbb{X},$$

corresponding to $\pi \in \Pi^{RS}$. By the above lemma, ν_0 is a linear combination of $\nu_1, \cdots, \nu_k, k \leq m+1$, in G_e with strictly positive weights. Letting $\delta_x(du)$ denote the Dirac measure at x, we may disintegrate ν_j's as

$$\nu_j(\{i\}, du) = \eta_j(i)\delta_{\phi_{ij}}(du), i \in S, 1 \leq j \leq k.$$

By abuse of notation, let $\phi_j : i \to \phi_{ij}$ denote the corresponding stationary strategy. Also, by the ergodic decomposition of the chain, $\eta_0 = \sum_{i=1}^{n} a_i \tilde{\eta}_i$ where $a_i \in [0,1]$ with $\sum a_i = 1$ and $\{\tilde{\eta}_i\}$ are ergodic invariant probability measures under π. In particular, $\tilde{\eta}_i$'s have disjoint supports. We denote these by $\{\mathbb{X}_i\}$ respectively.

Lemma 10.19 *For $1 \leq j \leq k$, support $(\eta_j) \subset \mathbb{X}_i$ for some i.*

Proof. If not, two states in two distinct $\mathbb{X}_i$'s would communicate with each other under the stationary policy ϕ_j. But then they would also do so under π (cf. the proof of Lemma 10.4), a contradiction. The claim follows. ∎

For $i \in \mathbb{X}$ with $\eta_0(i) > 0$, let $N(i) = \{u \in \mathbb{A} : u = \phi_{ij}$ for some $j, 1 \leq j \leq k$, satisfying $\eta_j(i) > 0\}$. Let $n(i) = |N(i)| - 1$. Then for each i, $n(i)$ is the 'number of randomizations at i' for π. To understand this terminology, observe that π_i will be of the form

$$\pi(i, da) = \sum_{\ell=1}^{|N(i)|} \left(b'_\ell \eta'_\ell(i) \phi'_\ell(i, da) / \left[\sum_{n=1}^{|N(i)|} b'_n \eta'_n(i) \right] \right) \tag{10.14}$$

where $\phi'_l(i, da)$ are Dirac (cf. proof of Lemma 10.2) with ergodic distributions η'_ℓ and $\{b'_j\}$ satisfy $b'_j > 0$, $\sum_\ell b'_\ell = 1$. Thus $|N(i)| = 2$ corresponds to randomizing between two choices, i.e. a single randomization, and so on. By convention, $n(i) = 0$ if $N(i) = \phi$.

Pick $i \in \mathbb{X}$ (if any) such that $\eta_0(i) > 0, n(i) > 0$. Let $N(i) = \{u(1), u(2), \cdots, u(n(i)+1)\}$. Then $\pi(i, da)$ is a strict convex combination of the Dirac measures at $u(j)$'s. Define $\psi^{ij} \in \Pi^{RS}$ by:

$$\begin{aligned} \psi^{ij}(\ell, dy) &= \pi(\ell, dy), \ \ell \neq i, \\ &= \delta_{u(j)}(dy), \ \ell = i, \end{aligned}$$

Lemma 10.20 *ν_0 is a strict convex combination of distinct elements $\mu_1, \cdots, \mu_{n(i)+1}$ of G, where μ_j is an ergodic occupation measure under ψ^{ij} for each j. Furthermore, $\mu_1, \cdots, \mu_{n(i)+1}$ form the corners of an $(n(i)+1)$ simplex.*

Proof. If η_0 is ergodic, the first claim follows by iterating the argument used in the proof of Lemma 10.4. If not, pick the $\tilde{\eta}_\ell$ for which $\tilde{\eta}_\ell(i) > 0$ and apply the same argument. Suppose the second claim is false. Then ν_0 can be expressed as a strict convex combination of elements from $\{\mu_1, \cdots, \mu_{n(i)+1}\}$ in at least two distinct ways. But then (10.14) applied to the present situation allows us to write a finitely supported probability measure $\pi(i, da)$ as a strict convex combination of distinct Dirac measures in two different ways, an impossibility. The claim follows. ∎

Call this $(n(i)+1)$-simplex the 'perturbation simplex' at i, denoted by $Q(i)$.

Lemma 10.21 *Let $\nu_1 \neq \nu_2$ in $Q(i)$, with*

$$\nu_j(\{l\}, da) = \eta_j(l) \varphi_j(l, da), \quad l \in \mathbb{X}, j = 1, 2.$$

Let $\eta_j(l) > 0$ for $j = 1, 2$. Then $\varphi_1(l, da) = \varphi_2(l, da)$ for $l \neq i$ and $\varphi_1(i, da) \neq \varphi_2(i, da)$.

Proof. The claim for $l \neq i$ is immediate from (10.14) and our definition of $Q(i)$. That $\varphi_{1i} \neq \varphi_{2i}$ follows from (10.14) and the easily verifiable fact: for $n > 1, b_1, \ldots, b_n > 0$, the map

$$[a_1, \ldots, a_n] \in \{[x_1, \ldots, x_n] : 0 < x_i < 1 \ \forall \ i, \sum_i x_i = 1\} \to$$

$$[a_1b_1/c, a_2b_2/c, \ldots, a_nb_n/c] \in (0,1)^n,$$

with $c = \sum_{i=1}^n a_ib_i$, is one-one. ∎

Let $Y(i)$ = the $n(i)$-dimensional affine space, in the space of finite signed measures on $\mathbb{X} \times \mathbb{A}$, spanned by $Q(i)$ for i as above.

Lemma 10.22 *If $j \neq i$ satisfies $\eta_0(j) > 0$, $n(j) > 0$, then $Y(i) \cap Y(j) = \{\nu_0\}$.*

Proof. Suppose not. Then there must exist a $\overline{\nu} \neq \nu_0, \overline{\nu} \in Q(i) \cap Q(j)$. Let Z be the line segment joining $\nu_0, \overline{\nu}$. Then $Z \subset Q(i) \cap Q(j)$. Write a typical $\nu \in Z$ as

$$\nu(\{l\}, da) = \eta(l)\varphi_l(da), l \in S.$$

Note that for $\nu \neq \overline{\nu}$ in Z, support$(\eta) \subset$ support(η_0). As ν moves along Z, Lemma 10.21 and the fact that $Z \subset Q(i)$ implies $\varphi_j(\cdot) = \pi(j, \cdot)$ all along. Interchange the roles of i, j to conclude $\varphi_i(\cdot) = \pi(i, \cdot)$ all along, a contradiction to $\overline{\nu} \neq \nu_0$ in view of Lemma 10.21. The claim follows. ∎

Lemma 10.23 *Let $i_1, i_2, \ldots, i_{n+1}$ be different states in $\mathbb{X}$ with $n(i_j) > 0$ for all j and $\alpha_1, \alpha_2, \ldots, \alpha_n \in [0,1]$ with $\sum_{i=1}^n \alpha_i = 1$. Then*

$$\{\alpha_1 Q(i_1) + \alpha_2 Q(i_2) + \cdots + \alpha_1 Q(i_n)\} \cap Q(i_{n+1}) = \{\nu_0\}.$$

Proof. For $n = 1$, this reduces to the preceding lemma. The general claim follows by induction using an argument similar to that above at each state. ∎

Let $L(i) = Y(i) - \nu_0$ when $n(i) > 0$. (That is, $L(i)$ is $Y(i)$ translated by ν_0.)

Corollary 10.4 dim $(L(i_1) + \cdots + L(i_n)) = \sum_{l=1}^n$ dim $(L(i_l))$.

This is immediate from the preceding lemma. Thus $L(i_1) + L(i_2) + \cdots + L(i_n)$ is a direct sum.

Lemma 10.24 $\sum_i n(i) \leq m$.

Proof. If not, for some $\{i_1, \ldots, i_l\} \subset S, \sum_{j=1}^l n(i_j) \geq m+1$. By the foregoing, corners of $Q(i_1), \ldots, Q(i_l)$ together form a polytope with $\sum_{m=1}^l n(i_m) \geq m+1$ dimensional relative interior that contains ν_0. Now argue as in Lemma 10.18 to get a contradiction. ∎

This completes the convex analytic preliminaries for the constrained control problem.

10.4.2 Constrained control: main results

The constrained control problem is

$$\text{Minimize} \int g_0 d\nu \tag{10.15}$$

subject to

$$\nu \in G, B_i \leq \int g_i d\nu \leq C_i, 1 \leq i \leq K, \tag{10.16}$$

where $g_i : \mathbb{X} \times \mathbb{A} \to [0, \infty)$ are prescribed continuous functions and B_i, C_i prescribed nonnegative scalars for $0 \leq i \leq K$. Let $H = \{\nu \in G : B_i \leq \int g_i d\nu \leq C_i, 1 \leq i \leq K\}$, assumed nonempty. Also assume that

$$\alpha_0 \stackrel{\Delta}{=} \inf_{\nu \in H} \int g_0 d\nu < \infty.$$

As before, we consider two cases:

Case 1: Near-monotone case

$$B_i = 0 \ \forall \ i, \quad \text{and} \ \liminf_{j \to \infty} \inf_a g_0(j, a) > \alpha_0.$$

Case 2: Stable case

$$G \text{ is compact and } H \text{ closed, hence compact.}$$

Since the upper inequality in (10.16) is preserved anyway under convergence in $\mathcal{P}(\mathbb{X} \times \mathbb{A})$, the closedness of H is the requirement that the lower inequality also do so. This is the case, e.g. if B_i's are zero or if g_i's are bounded, $1 \leq i \leq K$.

Some relaxation of these conditions is possible along the lines of the remark preceding Lemma 10.9.

Lemma 10.25 *In either case, the minimization problem above has a solution in H_e.*

Proof. Existence of a minimizer in H follows by its compactness in Case 2 and as in Lemma 10.8 in Case 1: One repeats those arguments to conclude that if $\{\nu_n\} \subset H$ satisfy $\int g_0 d\nu_n \downarrow \alpha_0$, then $\{\nu_n\}$ is tight and every limit point ν thereof satisfies.

$$\int g_0 d\nu = \alpha_0, \quad \int g_j d\nu \leq C_j, \quad 1 \leq j \leq K.$$

The existence of a minimizer in H_e then follows from Choquet's theorem as in Lemma 10.7. ∎

Suppose we revert to our original 'almost sure' formulation of the ergodic control problem and consider the constrained version thereof. That is, we seek to minimize almost surely, over all admissible $\{a_n\}$, the quantity

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} g_0(x_m, a_m),$$

subject to: for $1 \leq i \leq K$,

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} g_i(X_m, Z_m) \leq C_i \ \text{ a.s.},$$

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} g_i(X_m, Z_m) \geq B_i \ \text{ a.s..}$$

As in section 2, this can be reduced to the problem (10.15)–(10.16) without any loss of generality by using our characterization of limit points of empirical measures in Lemma 10.10, with the proviso that we add condition (†) of section 2 to 'Case 2' above. We refer to this modification as 'Case 2''.

Theorem 10.6 *In both Case 1 and Case 2', there exists an optimal $\pi \in \Pi^{RS}$ that requires at most K randomizations. Furthermore, if H has a nonempty interior in G, then there exist $\lambda_1, \cdots, \lambda_{2K} \geq 0$ such that for all $\nu \in G$ and $\gamma_1, \cdots, \gamma_{2K} \geq 0$, we have: If $\nu_0 \in H_e$ is the optimal point, then*

$$\begin{aligned}
&\int g_0 d\nu - \sum_{i=1}^{K} \lambda_i \left(C_i - \int g_i d\nu \right) - \sum_{i=1}^{K} \lambda_{K+i} \left(\int g_i d\nu - B_i \right) \\
&\quad \geq \int g_0 d\nu_0 - \sum_{i=1}^{K} \lambda_i \left(C_i - \int g_i d\nu_0 \right) - \sum_{i=1}^{K} \lambda_{K+i} \left(\int g_i d\nu_0 - B_i \right) \\
&\quad \geq \int g_0 d\nu_0 - \sum_{i=1}^{K} \gamma_i \left(C_i - \int g_i d\nu_0 \right) - \sum_{i=1}^{K} \gamma_{K+i} \left(\int g_i d\nu_0 - B_i \right).
\end{aligned}$$

Proof. The existence of an optimal $\pi \in H_e$ is argued above. That it requires at most K randomizations follows from Lemma 4.7 and the observation that only one of the inequalities in (10.16) can be active at a time for each i (except in the degenerate case $C_i = B_i$, which is handled analogously). The last claim follows from standard Lagrange multiplier theory ([42], pp.216-219). ∎

A similar treatment is possible for constrained problems with other (e.g. discounted) cost criteria. As noted in section 3.1, an important difference there is the role played by the initial distribution, usually held fixed, which cannot be wished away. Any claims of optimality will be relative to a specific initial distribution (or more generally, a convex compact set thereof). A parallel treatment is possible for locally compact $\mathbb{X}$ under the conditions leading to Theorem 10.4.

It should also be noted that the lower bound constraints were defined in terms of *liminf* rather than *limsup*. This is essential, as the following example shows. Consider a chain with two states, 1 and 2, and two actions, 1, 2. Under action i, $i = 1, 2$, the next state is i with probability $1 - \epsilon$ and $j \neq i$ with probability ϵ, for some small $\epsilon \in (0, 0.2)$. Consider the two constraints

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} I\{x_m = i\} \geq 0.75, \ i = 1, 2.$$

Then no $\pi \in \Pi^{RS}$ is feasible, but there are nonstationary feasible policies. For example, one feasible policy is the one that uses action 1 at times $n \in I_k, k = 0, 2, 4, \cdots$, and action 2 otherwise, where the intervals I_k for $k = 1, 2, \cdots$, are

defined as follows: Let $\tau_0 = 0$ and

$$\begin{aligned}
\tau_1 &= \min\{n > 0 : \sum_{m=0}^{n-1} I\{x_m = 1\} > 0.75n\}, \\
\sigma_i &= \min\{n > \tau_i : \sum_{m=0}^{n-1} I\{x_m = 2\} > 0.75n\}, \\
\tau_{i+1} &= \min\{n > \sigma_i : \sum_{m=0}^{n-1} I\{x_m = 1\} > 0.75n\},
\end{aligned}$$

for $i \geq 1$. Define $I_k = [\tau_k, \sigma_k)$ for even k, $I_k = [\sigma_k, \tau_{k+1})$ for odd k.

10.4.3 A general multiobjective problem

The foregoing had one 'primary' cost function g_0, whose average was to be minimized subject to constraints on several 'secondary' costs $g_1, \cdots, g_K$. In traditional 'multiobjective' framework, one wants to treat them on a comparable footing. Since all of them cannot in general be minimized simultaneously, one seeks a 'Pareto point', defined next. We stick to the ergodic control framework of the preceding section.

Let $\hat{g}(\nu) = \left[\int g_0 d\nu, \cdots, \int g_K d\nu\right] \in \mathbb{R}^{K+1}$ denote the cost vector corresponding to $\nu \in G$ and define $W = \{\hat{g}(\nu) : \nu \in G\}$, which will be a closed convex subset of $\mathbb{R}^{K+1}$. On W, define a partial order '$<$' by: $x = [x_0, \cdots, x_K] < y = [y_0, \cdots, y_K]$ if $x_i \leq y_i$ for all i, with a strict inequality for at least one i. Call $x^* \in W$ a Pareto point if there is no $z \in W$ for which $z < x^*$. It is easy to see that a minimizer of $\sum_{i=0}^{K} \lambda_i x_i$ over $x = [x_0, \cdots, x_K] \in W$ for any choice of $\lambda_i > 0$, $0 \leq i \leq K$, will be a Pareto point. In fact, if W is a polytope (i.e. convex hull of finitely many extreme points), then all Pareto points can be obtained thus. By our characterization of G_e, it follows that if $\mathbb{X}$ and $\mathbb{A}$ are finite, then G and hence W will have finitely many extreme points and the foregoing remark applies. More generally, a weaker claim holds, viz., the Pareto points obtained thus are dense in the set of all Pareto points. All this is a consequence of the celebrated Arrow-Barankin-Blackwell theorem [6].

More generally, Pareto points can be obtained by minimizing a '*utility function*' $U : W \to \mathbb{R}$, which is any continuous function with the property: $U(y) < U(x)$ wherever $y < x$. The function $x \to \sum_i \lambda_i x_i$ with $\lambda_i > 0$ $\forall i$ is just a special case of this. Another important case is as follows: Let $g^* = [\min_G \int g_0 d\nu, \cdots, \min_G \int g_K d\nu] \in \mathbb{R}^{K+1}$, the so called 'ideal point'. The nontrivial case is $g^* \notin W$. Let $\tilde{g} \in W$ be the unique point in W such that $||g^* - \tilde{g}|| = \min_{g \in W} ||g^* - g||$, corresponding to $U(x) = ||x - g^*||$. It is easy to see that this will be a Pareto point. Finding $\tilde{g}$ is an abstract quadratic programming problem.

There is no reason why $\tilde{g}$ should correspond to any element of G_e. The following 'approximation theorem', however, justifies a search for good strategies among those that correspond to a randomization between finitely many points of G_e. Assume that $g_0, \cdots, g_K$ are bounded.

Theorem 10.7 *For any $\nu \in G, n \geq 1$, there exists a $\nu^n \in G$ such that ν^n is a convex combination of at most n points of G_e and*

$$||\hat{g}(\nu) - \hat{g}(\nu^n)|| = O\left(\frac{1}{n}\right).$$

Proof. Let ν be the barycenter of $\Phi \in \mathcal{P}(G_e)$ and $\{\xi(i)\}$ i.i.d. G_e-valued random variables with law Φ. Let $Y(i) = \left[\int g_0 d\xi(i), \ldots, \int g_K d\xi(i)\right], i \geq 1$. Then $\{Y(i)\}$ are i.i.d. with mean $\hat{g}(\nu)$ and

$$E\left[\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y(i) - \hat{g}(\nu)\right|\right|^2\right] = \frac{1}{n^2} \cdot n \cdot E[||Y(1) - \hat{g}(\nu)||^2] = \frac{C}{n}$$

for a constant $C > 0$. Thus for at least one sample point,

$$\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y(i) - \hat{g}(\nu)\right|\right|^2 < \frac{C}{n}.$$

The claim follows. ∎

Note, however, that this is not a constructive argument.

Appendix

We briefly recall here stochastic Lyapunov conditions for stability of controlled Markov chains. Let $\tau(0) = \min\{n > 0 : x_n = 0\}$. Also, let G_s denote the set of ergodic occupation measures corresponding to stationary policies. The main result is the following theorem from [16] (See also [22]):

Theorem A.1 *For an irreducible controlled Markov chain, the following are equivalent:*

1. *For $x_0 = 0$, $\tau(0)$ is uniformly integrable under all stationary controls.*
2. *G_s is tight.*
3. *G_s is compact.*
4. *For $x_0 = 0$, $\tau(0)$ is uniformly integrable under all randomized stationary controls.*
5. *G is tight.*
6. *G is compact.*
7. *There exists an $h : \mathbb{X} \to \mathbb{R}_+$ with $h(i) \to \infty$ as $i \to \infty$ and*

$$\sup_{\Pi^{RS}} \mathbb{E}\left[\sum_{m=0}^{\tau(0)} h(x_m)/x_0 = i\right] < \infty, \quad \forall i.$$

8. *There exists a* $V : \mathbb{X} \to \mathbb{R}_+$, *a function* $h : \mathbb{X} \to \mathbb{R}_+$ *as in 7. above, a constant* $B > 0$ *and a finite* $C \subset \mathbb{X}$ *such that under any stationary policy* ϕ,

$$\mathbb{E}[V(x_{n+1})/\mathcal{F}_n] \leq V(x_n) - h(x_n) + BI\{x_n \in C\}.$$

In 7. above, it suffices to state the condition only for $i = 0$. In this form, the above claim also extends to the more general hypothesis that state '0' be reachable from every other state under any stationary policy. This is easily verified by looking at the detailed proof of the above theorem in [16]. Recall the usual stochastic Lyapunov condition for stability (i.e. positive recurrence) of a single uncontrolled irreducible Markov chain: There exists a $V : \mathbb{X} \to \mathbb{R}_+$, constants B, $\epsilon > 0$, and a finite $C \subset \mathbb{X}$ such that

$$\mathbb{E}[V(x_{n+1})/\mathcal{F}_n] \leq V(x_n) - \epsilon + BI\{x_n \in C\}.$$

If we require this to hold for the controlled Markov chain uniformly under all $\pi \in \Pi^{RS}$, that is not sufficient to ensure compactness of G. A counterexample is given in [16], pp. 112-113.

As for (†), the following is proved in [16]:

Theorem A.2 *In the following,* $(i) \Longrightarrow (ii) \Longrightarrow (iii)$*:*
(i) *There exist* $V_1, V_2 : \mathbb{X} \to \mathbb{R}$, $\epsilon > 0$, $B > 0$ *and a finite* $C \subset \mathbb{X}$ *such that for* $n \geq 0$,

$$\begin{aligned} \mathbb{E}[V_1(x_{n+1})/\mathcal{F}_n] &\leq V_1(x_n) - \epsilon + BI\{x_n \in C\}, \\ \mathbb{E}[V_2(x_{n+1})/\mathcal{F}_n] &\leq V_2(x_n) - V_1(x_n) + BI\{x_n \in C\}. \end{aligned}$$

(ii) $g(i) \stackrel{\Delta}{=} sup\, \mathbb{E}[\tau(0)/x_0 = i] < \infty$ *for* $i \in \mathbb{X}$ *and*

$$\sup \mathbb{E}\left[\sum_{m=0}^{\tau(0)-1} g(x_m)/x_0 = 0\right] < \infty.$$

(iii) (†) *holds.*

See [51] for several interesting results in this vein, albeit for uncontrolled Markov chains.

Bibliographical note

The convex/linear programming approach, as already mentioned, goes back to Manne [43]. See [37] for a historical perspective. The approach to ergodic control taken in section 2 first appeared in [10] for the irreducible case. The fully general case presented here appears in [14]. Reference [45] is another survey on convex analytic methods.

For expected average cost criterion, see [1], [3], [34], [35], [38]. For other criteria (except total cost), see [9]. For total cost (or rather, reward), see [27] and the references therein. For the general state space case, the treatment that

appears here is new in principle, but closely mimics the corresponding theory for controlled diffusions developed in [13], [17]. In an even more general set-up than the one considered here, one is lead to some delicate measurability issues, see, e.g. [26]. Theorem 5.1 of this reference is an interesting result in the spirit of this article. References [23], [24], and [36] contain some early works with this flavor. Extensions to semi-Markov case appear in [7]. Continuous time processes were studied in [8], [39], [40], and [41].

For the duality theory of linear programming applied to Markov decision processes, extensive accounts appear in [31], [32]. See [33] for a short overview. Extensions to continuous time processes appear in [8].

The constrained control problem dates back to [19], [20]. The convex analytic approach presented here is from [14], which improves upon earlier results from [11], [12], [47]. The 'sample path' or 'a.s.' variant was studied in [48], though our treatment is different. See [28] for results in the discounted cost framework. Extensions to semi-Markov processes appear in [25] and to continuous time processes in [13], [17]. Comprehensive book length treatments appear in [1] and [44] where further references may be found. See [2] for applications to telecommunications.

The multiobjective problem studied in subsection 4.2 is from [30], except for Theorem 36 which is new. Reference [30] also considers computational issues. See [50] for another approach to the multiobjective problem.

References

[1] E. Altman, *Constrained Markov decision processes*, Chapman and Hall/ CRC, Boca Raton, Florida, 1999.

[2] E. Altman, "Applications of Markov decision processes in communication networks: a survey", Chapter 15, *this volume*.

[3] E. Altman and A. Shwartz, "Markov decision problems and state-action frequencies", *SIAM Journal of Control and Optimization* **29**, pp.786-809, 1991.

[4] E. Altman and F. Spieksma, "The linear program approach in Markov decision processes revisited", *ZOR - Methods and Models in Operations Research* **42**, Issue 2, 1995, pp. 169-188.

[5] E. J. Anderson and P. Nash, *Linear Programming in Infinite Dimensional Spaces*, John Wiley, Chichester, 1987.

[6] K. J. Arrow, E. W. Barankin and D. Blackwell, "Admissible points of convex sets", in *Contributions to the Theory of Games* (eds. H.W. Kuhn and A.W. Tucker), Princeton Uni. Press, Princeton, NJ, pp.87-91, 1950.

[7] S. Bhatnagar and V. S. Borkar, "A convex analytic framework for ergodic control of semi-Markov processes", *Mathematics of Operations Research* **20**, 1995, pp. 923-936.

[8] A. G. Bhatt and V. S. Borkar, "Occupation measures for controlled Markov processes: characterization and optimality", *The Annals of Probability* **24**, pp.1531-1562, 1996.

[9] V. S. Borkar, "A convex analytic approach to Markov decision processes", *Probability Theory and Related Fields* **78**, pp.583-602, 1988.

[10] V. S. Borkar, "Control of Markov chains with long-run average cost criterion: the dynamic programming equations", *SIAM Journal of Control and Optimization 27*, pp.642-657, 1989.

[11] V. S. Borkar, "Controlled Markov chains with constraints", *Sadhana: Indian Academy of Sciences Proceedings in Engineering Sciences 15*, pp.405-413, 1990.

[12] V. S. Borkar, *Topics in Controlled Markov Chains*, Pitman Research Notes in Maths. No.240, Longman Scientific and Technical, Harlow, England, 1991.

[13] V. S. Borkar, "Controlled diffusions with constraints II", *Journal of Mathematical Analysis and Applications* **176** pp.310-321, 1993.

[14] V. S. Borkar, "Ergodic control of Markov chains with constraints - the general case", *SIAM Journal of Control and Optimization* **32**, pp. 176-186, 1994.

[15] V. S. Borkar, *Probability Theory: An Advanced Course*, Springer Verlag, New York, 1995.

[16] V. S. Borkar, "Uniform stability of controlled Markov processes", in *System Theory: Modeling, Analysis and Control* (T. E. Djaferis and I. C. Schick, eds.), Kluwer Academic Publishers, Boston, pp. 106-120, 1999.

[17] V. S. Borkar and M. K. Ghosh, "Controlled diffusions with constraints", *Journal of Mathematical Analysis and Applications* **152**, pp.88-108, 1990.

[18] G. Choquet, *Lectures on Analysis, Vol.II: Representation Theory*, W.A. Benjamin, Inc., Reading, Mass., 1969.

[19] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

[20] C. Derman and M. Klein, "Some remarks on finite horizon Markovian decision models", *Operations Research* **13**, pp. 272-278, 1965.

[21] L. Dubins, "On extreme points of convex sets", *Journal of Mathematical Analysis and Applications* **5**, pp. 237-244, 1962.

[22] G. Fayolle, V. A. Malyshev and M. V. Menshikov, *Topics in the Constructive Theory of Countable Markov Chains*, Cambridge University Press, Cambridge, UK, 1995.

[23] E. A. Feinberg, "Nonrandomized Markov and semi-Markov strategies in dynamic programming", *Theory of Probability and Applications* **27**, pp. 116-126, 1982.

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

[25] E. A. Feinberg, "Constrained semi-Markov decision processes with average rewards", *ZOR - Methods and Models in Operations Research* **39**, pp. 257-288, 1995.

[26] E. A. Feinberg, "On measurability and representation of strategic measures in Markov decision processes", in *Statistics, Probability and Game*

Theory: Papers in Honour of David Blackwell (eds. T. S. Ferguson et al), IMS Lecture Notes - Monographs Series **30**, Hayward, pp. 29-43, 1996.

[27] E. A. Feinberg, "Total reward criteria", Chapter 5, *this volume*.

[28] E. A. Feinberg and A. Shwartz, "Constrained discounted dynamic programming", *Mathematics of Operations Research* **21**, pp. 922-945, 1996.

[29] E. A. Feinberg and I. M. Sonin, "Notes on equivalent stationary policies in Markov decision processes with total rewards", *Mathematical Methods of Operations Research* **44**, pp. 205-221, 1996.

[30] M. K. Ghosh,"Markov decision processes with multiple costs", *Operations Research Letters* **9**, pp. 257-260, 1990.

[31] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes*, Springer Verlag, New York, 1996.

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

[33] O. Hernández-Lerma and J. B. Lasserre, "The linear programming approach, Chapter 11, *this volume*.

[34] A. Hordijk and L. C. M. Kallenberg, "Linear programming and Markov decision chains", *Management Science* **25**, 1979, pp. 352-362.

[35] A. Hordijk and L. C. M. Kallenberg, "Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research* **9**, 1984, pp. 276-289.

[36] D. Kadelka, "On randomized policies and mixtures of deterministic policies in dynamic programming", *Methods of Operations Research* **46**, 1983, pp. 67-75.

[37] L. C. M. Kallenberg, "Finite state and action MDPs", Chapter 1, *this volume*.

[38] D. Krass, *Contributions to the theory and applications of Markov decision processes*, Ph.D. Thesis, Department of Mathematics, The Johns Hopkins University, Baltimore, Maryland.

[39] N.V. Krylov, "Once more about the connection between elliptic operators and Itô's stochastic equations" in *Statistics and Control of Stochastic Processes, Steklov Seminar*(eds. N.V. Krylov, R.Sh. Liptser, and A.A. Novikov), Optimization Software, New York, pp. 69–101, 1985

[40] N.V. Krylov, "An approach in the theory of controlled diffusion processes", *SIAM Theory Prob. Appl.* **31** pp. 604–626, 1987.

[41] T. G. Kurtz and R. Stockbridge, "Existence of Markov controls and characterization of optimal Markov controls", *SIAM Journal of Control and Optimization* **36**, pp. 609-653, 1998. Correction note in *ibid.* **37**, pp. 1310-1311, 1999.

[42] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[43] A. Manne, "Linear programming and sequential decisions", *Management Science* **6**, pp. 259-267, 1960.

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

[45] A. B. Piunovskiy, "Controlled random sequences: methods of convex analysis and problems with functional constraints", *Russian Math. Surveys* **6**, pp. 129-192, 1998.

[46] M. Puterman, *Markov Decision Processes*, John Wiley, New York, 1994.

[47] K. W. Ross, "Randomized and past-dependent policies for Markov decision problems with multiple constraints", *Operations Research* **37**, pp. 474-477, 1989.

[48] K. W. Ross and R. Varadarajan, "Markov decision processes with sample path constraints: the communicating case", *Operations Research* **37**, pp. 780-790, 1989.

[49] L. I. Sennott, "Average reward optimization theory for denumerable state spaces", Chapter 4, *this volume*.

[50] N. Shimkin and A. Shwartz, "Guaranteed performance regions in Markovian systems with competing decision makers", *IEEE Transactions on Automatic Control* **38**, pp. 84-95, 1993.

[51] P. Tuominen and R. L. Tweedie, "Subgeometric rates of convergence of f-ergodic Markov chains", *Advances in Applied Probability* **26**, pp. 775-798, 1994.

Vivek S. Borkar
School of Technology and Computer Science
Tata Institute of Fundamental Research
Homi Bhabha Road
Mumbai 400005, India
borkar@tifr.res.in

11 THE LINEAR PROGRAMMING APPROACH

Onésimo Hernández-Lerma
Jean B. Lasserre

Abstract: This chapter is concerned with the Linear Programming (LP) approach to MDPs in general Borel spaces, valid for several criteria, including the finite horizon, long run (ergodic) average cost, the long run expected discounted and average costs.

11.1 INTRODUCTION

In this chapter we study the *linear programming* (LP) approach to Markov decision problems and our ultimate goal is to show how a Markov decision problem (MDP) can be approximated by *finite* linear programs.

The LP approach to Markov decision problems dates back to the early sixties with the pioneering work of De Ghellinck [10], d'Epenoux [11] and Manne [30] for MDPs with finite state and action spaces. Among later contributions for finite or countable state and action MDPs, let us mention Altman [1], Borkar [8], [9], Denardo [12], Kallenberg [28], Hordijk and Kallenberg [25], Hordijk and Lasserre [26], Lasserre [29], and for MDPs in general Borel spaces and in discrete or continuous time, Bhatt and Borkar [7], Haneveld [13], Heilmann [14], [15], Hernández-Lerma and González-Hernández [16], Hernandez-Lerma and Lasserre [19], [21], Mendiondo and Stockbridge[31], Stockbridge [39], Yamada [42].

Among the nice features of the LP approach, the most evident is that it is valid in a very general context. For instance, for the long-run expected average cost (AC) problem, one does not need to assume that the Average Cost Optimality Equation (ACOE) holds, a restrictive assumption. Under weak assumptions, one obtains the existence of a stationary average-cost optimal

policy (possibly on a subset S of the state space). The LP approach permits to identify this set S which is an ergodic class of minimum expected average-cost. Getting an expected average-cost optimal policy for all initial states (for the unichain case as well as the multichain case) requires much stronger assumptions. However, the LP approach is still possible via the introduction of additional variables. Also, it permits to handle some constrained MDPs in a very natural form. Finally, it is possible to devise simple convergent numerical approximation schemes that require to solve finite LPs for which efficient codes are now available. However, if convergence to the optimal value is obtained, it remains to devise a convergent approximation scheme for policies, as done in alternative methods like for instance in Hernández-Lerma [17] or Sennott [37], [38] for control of queues.

Let us briefly outline one simple way to see how the LP approach can be naturally introduced, although it was not the idea underlying the first papers on the LP approach to MDPs. The starting point is to observe that given a policy $\pi \in \Pi$, an initial distribution ν and a one-step cost function $c : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$, the finite-horizon functional

$$J(\nu, \pi, N, 1, c) := N^{-1} E_\nu^\pi \sum_{t=0}^{N-1} c(x_t, a_t),$$

can be written as a *linear* functional $\int c d\mu_N^{\pi,\nu}$ with $\mu_N^{\pi,\nu}$ the expected (state-action) occupation measure

$$\mu_N^{\pi,\nu}(B) := N^{-1} E_\nu^\pi \sum_{t=0} N - 1\{(x_t, a_t) \in B\}, \qquad B \in \mathcal{B}(\mathbb{X} \times \mathbb{A}).$$

Under some conditions, and with some limiting arguments as $N \to \infty$, one may show that, for instance, minimizing the long-run expected average cost criterion (the AC problem) reduces to solving a linear program. More precisely, the AC problem reduces to minimize the linear criterion $\int c d\mu$ over a set of probability measures μ on $\mathbb{X} \times \mathbb{A}$ that satisfy some linear "invariance" constraints involving the transition kernel P. This approach for MDPs is of course related to the Birkhoff Individual Ergodic Theorem (for noncontrolled Markov chains) which states that given an homogeneous Markov chain X_t, $t = 0, 1, \ldots$ on $\mathbb{X}$, a cost function $c : \mathbb{X} \to \mathbb{R}$, and under some conditions,

$$\lim_{N \to \infty} N^{-1} E_\nu \sum_{t=0} N - 1 c(X_t) = \int c d\mu^\nu,$$

for some invariant probability measure μ^ν.

However, we should note that the first papers on the LP approach to MDPs used a different (in fact, dual) approach. Namely, the LP formulation was a rephrasing of the average (or discounted)-cost optimality equations. We briefly discuss this approach in Remark 11.5 that yields a dual linear program.

Although the LP approach is valid for several criteria, including the N-step expected total cost, the infinite-horizon expected discounted cost, the control up to exit time, the long run average cost the long-run expected average cost, the constrained discounted and average cost problems (see e.g. (11.1),

(11.71), (11.72)) we have chosen to illustrate the LP approach with the AC problem. With ad hoc suitable modifications and appropriate assumptions, the reader would easily deduce the corresponding linear programs associated with the other mentioned problems. For instance, and with respect to constrained MDPs, the reader is referred to Huang and Kurano [27], Altman [1], Piunovskiy [33] and Hernández-Lerma and Gonzalez-Hernández [18]. Similarly, for multiobjective MDPs, see for instance, Hernández-Lerma and Romera [23].

We shall first proceed to find a suitable linear program associated to the Markov decision problem. Here, by a "suitable" linear program we mean a linear program (P) that together with its dual (P*) satisfies that

$$\sup(\mathrm{P}^*) \leq (\mathrm{MDP})^* \leq \inf(\mathrm{P}), \tag{11.1}$$

where (using terminology specified in the following section)

$$\begin{aligned} \inf(\mathrm{P}) &:= \text{value of the primal program (P)}, \\ \sup(\mathrm{P}^*) &:= \text{value of the dual program (P}^*\text{)}, \\ (\mathrm{MDP})^* &:= \text{value function of the Markov decision problem.} \end{aligned}$$

In particular, if there is *no duality gap* for (P), so that

$$\sup(\mathrm{P}^*) = \inf(\mathrm{P}), \tag{11.2}$$

then of course the values of (P) and of (P*) yield the desired value function (MDP)*.

However, to find an *optimal policy* for the Markov decision problem, (11.1) and (11.2) are not sufficient because they do not guarantee that (P) or (P*) are *solvable*. If it can be ensured that, say, the primal (P) is solvable—in which case we write its value as min (P)—and that

$$\min(\mathrm{P}) = (\mathrm{MDP})^*, \tag{11.3}$$

then an optimal solution for (P) can be used to determine an optimal policy for the Markov decision problem. Likewise, if the dual (P*) is solvable and its value—which in this case is written as max (P*)—satisfies

$$\max(\mathrm{P}^*) = (\mathrm{MDP})^*, \tag{11.4}$$

then we can use an optimal solution for (P*) to find an optimal policy for the Markov decision problem. In fact, one of the main results in this chapter (Theorem 11.6) gives conditions under which (11.3) and (11.4) are both satisfied, so that in particular *strong duality* for (P) holds, that is,

$$\max(\mathrm{P}^*) = \min(\mathrm{P}). \tag{11.5}$$

Section 11.2 presents background material. It contains, in particular, a brief introduction to infinite LP. In Section 11.3 we introduce the program (P) associated to the AC problem, and we show that (P) is solvable and that there is no duality gap, so that (11.2) becomes

$$\sup(\mathrm{P}^*) = \min(\mathrm{P}).$$

Section 11.4 deals with approximating sequences for (P) and its dual (P*). In particular, it is shown that if a suitable maximizing sequence for (P*) exists, then the strong duality condition (11.5) is satisfied. Section 11.5 presents an approximation scheme for (P) using finite-dimensional programs. The scheme consists of three main steps. In step 1 we introduce an "increasing" sequence of *aggregations* of (P), each one with finitely many constraints. In step 2 each aggregation is *relaxed* (from an equality to an inequality), and, finally, in step 3, each aggregation-relaxation is combined with an *inner approximation* that has a finite number of decision variables. Thus the resulting aggregation-relaxation-inner approximation turns out to be a finite linear program, that is, a program with finitely many constraints and decision variables. The corresponding convergence theorems are stated without proof, and the reader is referred to [21] and [22] for proofs and further technical details. These approximation schemes can be extended to a very general class of infinite-dimensional linear programs (as in [20]), not necessarily related to MDPs.

11.2 LINEAR PROGRAMMING IN INFINITE-DIMENSIONAL SPACES

The material is divided into four subsections. The first two subsections review some basic definitions and facts related to dual pairs of vector spaces and linear operators whereas the last two subsections summarize the main results on infinite LP needed in later sections.

11.2.1 Dual pairs of vector spaces

Let $\mathcal{X}$ and $\mathcal{Y}$ be two arbitrary (real) vector spaces, and let $\langle \cdot, \cdot \rangle$ be a **bilinear form** on $\mathcal{X} \times \mathcal{Y}$, that is, a real-valued function on $\mathcal{X} \times \mathcal{Y}$ such that

- the map $x \mapsto \langle x, y \rangle$ is linear on $\mathcal{X}$ for every $y \in \mathcal{Y}$, and
- the map $y \mapsto \langle x, y \rangle$ is linear on $\mathcal{Y}$ for every $x \in \mathcal{X}$.

Then the pair $(\mathcal{X}, \mathcal{Y})$ is called a **dual pair** if the bilinear form "separates points" in x and y, that is,

- for each $x \neq 0$ in $\mathcal{X}$ there is some $y \in \mathcal{Y}$ with $\langle x, y \rangle \neq 0$, and
- for each $y \neq 0$ in $\mathcal{Y}$ there is some $x \in \mathcal{X}$ with $\langle x, y \rangle \neq 0$.

If $(\mathcal{X}, \mathcal{Y})$ is a dual pair, then so is $(\mathcal{Y}, \mathcal{X})$.

If $(\mathcal{X}_1, \mathcal{Y}_1)$ and $(\mathcal{X}_2, \mathcal{Y}_2)$ are two dual pairs of vector spaces with bilinear forms $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, respectively, then the product $(\mathcal{X}_1 \times \mathcal{X}_2, \mathcal{Y}_1 \times \mathcal{Y}_2)$ is endowed with the bilinear form

$$\langle (x_1, x_2), (y_1, y_2) \rangle := \langle x_1, y_1 \rangle_1 + \langle x_2, y_2 \rangle_2. \tag{11.6}$$

For MDPs, a typical dual pair of vector spaces is the following. Let S be a Borel space with Borel σ-algebra $\mathcal{B}(S)$, and let $\mathcal{X} := \mathbb{M}_w(S)$ be the normed linear space of finite signed measures μ on $\mathcal{B}(S)$, with finite *w-norm*

$$\|\mu\|_w := \int_S w d|\mu|, \tag{11.7}$$

for some weight function $w \geq 1$. Now, let $\mathcal{Y} := \mathbb{F}_w(S)$ be the normed linear space of real-valued measurable functions on S with finite w-norm

$$\|u\|_w := \sup_S |u(s)|/w(s). \tag{11.8}$$

Then, the dual pair $(\mathcal{X}, \mathcal{Y}) = (\mathbb{M}_w(S), \mathbb{F}_w(S))$ endowed with the bilinear form

$$\langle \mu, u \rangle := \int_S u d\mu \tag{11.9}$$

is easily seen to be a dual pair. Moreover, by (11.6), the bilinear form corresponding to the dual pair $(\mathbb{R}^n \times \mathbb{M}_w(S), \mathbb{R}^n \times \mathbb{F}_w(S))$ is

$$\langle (x, \mu), (y, u) \rangle = x \cdot y + \langle \mu, u \rangle \tag{11.10}$$

where $x \cdot y := \sum_{i=1}^n x_i y_i$ denotes the usual scalar product of n-vectors.

Given a dual pair $(\mathcal{X}, \mathcal{Y})$, we denote by $\sigma(\mathcal{X}, \mathcal{Y})$ the **weak topology** on $\mathcal{X}$ (also referred to as the **σ-topology** on $\mathcal{X}$), namely, the coarsest—or weakest—topology on $\mathcal{X}$ under which all the elements of $\mathcal{Y}$ are continuous when regarded as linear forms $\langle \cdot, y \rangle$ on $\mathcal{X}$. Equivalently, the base of neighborhoods of the origin of the σ-topology is the family of all sets of the form

$$N(I, \varepsilon) := \{x \in \mathcal{X} \mid |\langle x, y \rangle| \leq \varepsilon \quad \forall y \in I\}, \tag{11.11}$$

where $\varepsilon > 0$ and I is a *finite* subset of $\mathcal{Y}$. (See, for instance, Robertson and Robertson [35], p. 32.) In this case, if $\{x_n\}$ is a sequence or a net in $\mathcal{X}$, then x_n *converges to* x *in the weak topology* $\sigma(\mathcal{X}, \mathcal{Y})$ if

$$\langle x_n, y \rangle \to \langle x, y \rangle \quad \forall y \in \mathcal{Y}. \tag{11.12}$$

For instance, for the dual pair $(\mathbb{M}_w(S), \mathbb{F}_w(S))$, a sequence or a net of measures μ_n converges to μ in the weak topology $\sigma(\mathbb{M}_w(S), \mathbb{F}_w(S))$ if

$$\langle \mu_n, u \rangle \to \langle \mu, u \rangle \quad \forall u \in \mathbb{F}_w(S), \tag{11.13}$$

where $\langle \cdot, \cdot \rangle$ stands for the bilinear form in (11.9).

Remark 11.1 *(a) Let $(\mathcal{X}, \mathcal{Y})$ be a dual pair such that $\mathcal{Y}$ is a Banach space and $\mathcal{X} = \mathcal{Y}^*$ is the topological dual of $\mathcal{Y}$. In this case, the weak topology $\sigma(\mathcal{X}, \mathcal{Y})$ is called the* **weak*** *(weak-star)* **topology** *on $\mathcal{X}$, and so (11.12) is referred to as the* **weak* convergence** *of x_n to x.*

(b) For instance, with S a locally compact separable metric (LCSM) space, let $\mathcal{X} := \mathbb{M}(S)$ be the Banach space of finite signed measures on S, endowed with the total variation norm $\|\mu\|_{TV} = |\mu|(S)$, and let $\mathcal{Y} := C_0(S)$ be the (separable) Banach space of continuous functions that vanish at infinity, equipped with the sup-norm. By the Riesz Representation Theorem (see, for example, Rudin [36]), $\mathbb{M}(S)$ is the topological dual of $C_0(S)$, and so the weak topology $\sigma(\mathbb{M}(S), C_0(S))$ on $\mathbb{M}(S)$ is in fact the weak topology.*

Definition 11.1 *Let $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Z}, \mathcal{W})$ be two dual pairs of vector spaces, and $G : \mathcal{X} \to \mathcal{Z}$ a linear map.*

(a) G is said to be **weakly continuous** *if it is continuous with respect to the weak topologies $\sigma(\mathcal{X}, \mathcal{Y})$ and $\sigma(\mathcal{Z}, \mathcal{W})$; that is, if $\{x_n\}$ is a net in $\mathcal{X}$ such that $x_n \to x$ in the weak topology $\sigma(\mathcal{X}, \mathcal{Y})$ [see (11.12)], then $Gx_n \to Gx$ in the weak topology $\sigma(\mathcal{Z}, \mathcal{W})$, i.e.*

$$\langle Gx_n, v\rangle \to \langle Gx, v\rangle \quad \forall v \in \mathcal{W}. \tag{11.14}$$

(b) The **adjoint** *G^* of G is defined by the relation*

$$\langle Gx, v\rangle = \langle x, G^*v\rangle \quad \forall x \in \mathcal{X},\ v \in \mathcal{W}. \tag{11.15}$$

The following proposition gives a well-known (easy-to-use) criterion for the map G in Definition 11.1 to be weakly continuous—for a proof see, for instance, Robertson and Robertson [35, p. 38]

Proposition 11.1 *The linear map G is weakly continuous if and only if its adjoint G^* maps $\mathcal{W}$ into $\mathcal{Y}$, that is, $G^*(\mathcal{W}) \subset \mathcal{Y}$.*

Positive and dual cones. (a) Let $(\mathcal{X}, \mathcal{Y})$ be a dual pair of vector spaces, and K a *convex cone* in $\mathcal{X}$, that is, $x + x'$ and λx belong to K whenever x and x' are in K and $\lambda > 0$. Unless explicitly stated otherwise, we shall assume that K is not the whole space, that is, $K \neq \mathcal{X}$, and that the origin (that is, the zero vector, 0) is in K. In this case, K defines a partial order $\geq$ on X such that

$$x \geq x' \quad \Leftrightarrow \quad x - x' \in K,$$

and K will be referred to as a *positive cone* . The *dual cone* of K is the convex cone K^* in $\mathcal{Y}$ defined by

$$K^* := \{y \in \mathcal{Y} \mid \langle x, y\rangle \geq 0 \quad \forall x \in K\}. \tag{11.16}$$

(b) If $\mathcal{X} = \mathbb{M}_w(S)$, we will denote by $\mathbb{M}_w(S)_+$ the "natural" *positive cone in* $\mathbb{M}_w(S)$, which consists of all the *nonnegative* measures in $\mathbb{M}_w(S)$, that is,

$$\mathbb{M}_w(S)_+ := \{\mu \in \mathbb{M}_w(S) | \mu \geq 0\}.$$

The corresponding dual cone $\mathbb{M}_w(S)_+^*$ in $\mathbb{F}_w(S)$ coincides with the "natural" positive cone

$$\mathbb{F}_w(S)_+ := \{u \in \mathbb{F}_w(S) | u \geq 0\}.$$

11.2.2 Infinite linear programming

An infinite linear program requires the following components:

- two dual pairs $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Z}, \mathcal{W})$ of real vector spaces;
- a weakly continuous linear map $L : \mathcal{X} \to \mathcal{Z}$, with adjoint $L^* : \mathcal{W} \to \mathcal{Y}$;
- a positive cone K in $\mathcal{X}$, with dual cone K^* in $\mathcal{Y}$ [see (11.16)]; and

- vectors $b \in \mathcal{Z}$ and $c \in \mathcal{Y}$.

Then the **primal** linear program is

$$\begin{aligned} \mathbb{P}: \quad & \text{minimize } \langle x, c\rangle \\ & \text{subject to: } Lx = b,\ x \in K. \end{aligned} \tag{11.17}$$

The corresponding **dual** problem is

$$\begin{aligned} \mathbb{P}^*: \quad & \text{maximize } \langle b, w\rangle \\ & \text{subject to: } c - L^*w \in K^*,\ w \in \mathcal{W}. \end{aligned} \tag{11.18}$$

An element x of $\mathcal{X}$ is called **feasible** for $\mathbb{P}$ if it satisfies (11.17), and $\mathbb{P}$ is said to be **consistent** if it has a feasible solution. If $\mathbb{P}$ is consistent, then its **value** is defined as

$$\inf \mathbb{P} := \inf\{\langle x, c\rangle | x \text{ is feasible for } \mathbb{P}\}; \tag{11.19}$$

otherwise, $\inf \mathbb{P} := +\infty$. The program $\mathbb{P}$ is **solvable** if there is a feasible solution x^* that achieves the infimum in (11.19). In this case, x^* is called an **optimal solution** for $\mathbb{P}$ and, instead of $\inf \mathbb{P}$, the value of $\mathbb{P}$ is written as

$$\min \mathbb{P} = \langle x^*, c\rangle.$$

Similarly, $v \in \mathcal{W}$ is **feasible** for the dual program $\mathbb{P}^*$ if it satisfies (11.18), and $\mathbb{P}^*$ is said to be **consistent** if it has a feasible solution. If $\mathbb{P}^*$ is consistent, then its **value** is defined as

$$\sup \mathbb{P}^* := \sup\{\langle b, v\rangle \mid v \text{ is feasible for } \mathbb{P}^*\}; \tag{11.20}$$

otherwise, $\sup \mathbb{P}^* := -\infty$. The dual $\mathbb{P}^*$ is **solvable** if there is a feasible solution v^* that attains the supremum in (11.20), in which case we write the value of $\mathbb{P}^*$ as

$$\max \mathbb{P}^* = \langle b, w^*\rangle.$$

The next theorem can be proved as in elementary (finite-dimensional) LP.

Theorem 11.1 *(a) (***Weak duality.***) If $\mathbb{P}$ and $\mathbb{P}^*$ are both consistent, then their values are finite and satisfy*

$$\sup \mathbb{P}^* \le \inf \mathbb{P}. \tag{11.21}$$

*(b) (***Complementary slackness.***) If x is feasible for $\mathbb{P}$, v is feasible for $\mathbb{P}^*$, and*

$$\langle x, c - L^*v\rangle = 0, \tag{11.22}$$

then x is optimal for $\mathbb{P}$ and w is optimal for $\mathbb{P}^*$.

The converse of Theorem 11.1(b) does not hold in general. It does hold, however, if there is **no duality gap** for $\mathbb{P}$, which means that equality holds in (11.21), i.e.

$$\sup \mathbb{P}^* = \inf \mathbb{P}. \tag{11.23}$$

On the other hand, it is said that the **strong duality** condition for $\mathbb{P}$ holds if $\mathbb{P}$ and its dual $\mathbb{P}^*$ are both solvable and

$$\max \mathbb{P}^* = \min \mathbb{P}. \tag{11.24}$$

The following theorem gives conditions under which $\mathbb{P}$ is solvable and there is no duality gap—for a proof see Anderson and Nash [1, Theorem 3.9].

Theorem 11.2 *Let H be the set in $\mathcal{Z} \times \mathbb{R}$ defined as*

$$H := \{(Lx, \langle x, c\rangle + r) \mid x \in K,\ r \geq 0\}.$$

If $\mathbb{P}$ is consistent and H is weakly closed [that is, closed in the weak topology $\sigma(\mathcal{Z}\times\mathbb{R}, \mathcal{W}\times\mathbb{R})$], then $\mathbb{P}$ is solvable and there is no duality gap, so that (11.23) becomes

$$\sup \mathbb{P}^* = \min \mathbb{P}.$$

11.2.3 Approximation of linear programs

An important practical question is how to obtain—or at least estimate—the value of a linear program. In later sections we shall consider two approaches related to the following definitions.

Definition 11.2 (Minimizing and maximizing sequences)

(a) A sequence $\{x_n\}$ in $\mathcal{X}$ is called a **minimizing sequence** *for $\mathbb{P}$ if each x_n is feasible for $\mathbb{P}$ and $\langle x_n, c\rangle \downarrow \inf \mathbb{P}$.*

(b) A sequence $\{v_n\}$ in $\mathcal{W}$ is called a **maximizing sequence** *for the dual problem $\mathbb{P}^*$ if each v_n is feasible for $\mathbb{P}^*$ and $\langle b, v_n\rangle \uparrow \sup \mathbb{P}^*$.*

Note that if $\mathbb{P}$ is consistent with a finite value $\inf \mathbb{P}$, then [by definition (11.19) of $\inf \mathbb{P}$] there exists a minimizing sequence. A similar remark holds for $\mathbb{P}^*$.

The equality $Lx = b$ in (11.17) is of course equivalent to write $Lx - b = 0$, or $\langle Lx - b, v\rangle = 0$ for all $v \in \mathcal{W}$. If the latter equality is required to hold only in a subset W of $\mathcal{W}$, we then have an *aggregation* of constraints of $\mathbb{P}$. On the other hand, if (11.17) holds only in a subset K' of K, we obtain an *inner approximation* of $\mathbb{P}$. The corresponding linear programs become as follows.

Definition 11.3 (Aggregations and inner approximations.)

(a) Let W be a subset of $\mathcal{W}$. Then the linear program

$$\begin{aligned} \mathbb{P}(W)\ :\ & \textit{minimize } \langle x, c\rangle \\ & \textit{subject to: } \langle Lx - b, w\rangle = 0 \quad \forall w \in W,\ x \in K, \end{aligned} \tag{11.25}$$

is called an **aggregation** *(of constraints) of $\mathbb{P}$.*

(b) If $K' \subset K$ is a subset of the positive cone $K \subset \mathcal{X}$, then the program

$$\begin{aligned} \mathbb{P}(K')\ :\ & \textit{minimize } \langle x, c\rangle \\ & \textit{subject to: } Lx = b,\ x \in K', \end{aligned} \tag{11.26}$$

is called an **inner approximation** *of* $\mathbb{P}$.

As K' is contained in K, we have $\inf \mathbb{P} \leq \inf \mathbb{P}(K')$. On the other hand, if x satisfies (11.17), then it satisfies (11.25), and so $\inf \mathbb{P}(W) \leq \inf \mathbb{P}$. Hence

$$\inf \mathbb{P}(W) \leq \inf \mathbb{P} \leq \inf \mathbb{P}(K').$$

Thus, we can use an aggregation (of constraints) to approximate $\inf \mathbb{P}$ *from below*, whereas an inner approximation can be used to approximate $\inf \mathbb{P}$ *from above*. One can also easily get the following (for a proof see Hernández-Lerma and Lasserre [20]):

Proposition 11.2 *Suppose that* $\mathbb{P}$ *is solvable.*

(a) If W *is weakly dense in* $\mathcal{W}$*, then* $\mathbb{P}(W)$ *is equivalent to* $\mathbb{P}$ *in the sense that* $\mathbb{P}(W)$ *is also solvable and*

$$\min \mathbb{P}(W) = \min \mathbb{P}.$$

(b) If K' *is weakly dense in* K*, then there is a sequence* $\{x_n\}$ *in* K' *such that*

$$\langle x_n, c \rangle \to \min \mathbb{P}.$$

11.3 LINEAR PROGRAMMING FORMULATION OF THE AC-PROBLEM

Let $(\mathbb{X}, \mathbb{A}, \{\mathbb{A}(x), x \in \mathbb{X}\}, P, c)$ be an MDP with Borel state and action spaces $\mathbb{X}$, $\mathbb{A}$, one-step cost function $c : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$, transition kernel P, and with the long-run expected average cost criterion, that is, given a policy $\pi \in \Pi$ and an initial distribution $\nu \in \mathcal{P}(\mathbb{X})$ (the space of probability measures on $\mathcal{B}(\mathbb{X})$), its long-run expected average cost $J(\pi, \nu)$ is given by:

$$J(\pi, \nu) := \limsup_{N \to \infty} N^{-1} E_\nu^\pi \sum_{t=0}^{N-1} c(x_t, a_t).$$

We recall that to the state space $\mathbb{X}$ and the action space $\mathbb{A}$ is associated the space $\mathbb{K}$ of *feasible state-action pairs*, i.e.

$$\mathbb{K} := \{(x, a) \in \mathbb{X} \times \mathbb{A} \,|\, x \in \mathbb{X},\, a \in \mathbb{A}(x)\}. \tag{11.27}$$

It is assumed that $\mathbb{K}$ is a Borel subset of $\mathbb{X} \times \mathbb{A}$ and contains the graph of a measurable function from $\mathbb{X}$ to $\mathbb{A}$. This implies in particular that the set of stationary deterministic policies $\mathbb{F}$ is not empty. Let Φ be the set of stochastic kernels φ on $\mathbb{A}$ given $\mathbb{X}$ for which $\varphi(\mathbb{A}(x)|x) = 1$ for all $x \in \mathbb{X}$. In other words, Φ stands for the family of randomized stationary policies.

Remark 11.2 *Every p.m.* μ *on* $\mathbb{X} \times \mathbb{A}$ *concentrated on* $\mathbb{K}$ *can be "disintegrated" as*

$$\mu(B \times C) = \int_B \varphi(C \,|\, x)\, \widehat{\mu}(dx) \qquad \forall B \in \mathcal{B}(\mathbb{X}),\, C \in \mathcal{B}(\mathbb{A}), \tag{11.28}$$

for some $\varphi \in \Phi$, where $\widehat{\mu}$ is the marginal of μ on $\mathbb{X}$, that is, $\widehat{\mu}(B) := \mu(B \times \mathbb{A})$ for all $B \in \mathcal{B}(\mathbb{X})$. Sometimes we shall write the disintegration (11.28) of μ as $\mu = \widehat{\mu} \cdot \varphi$.

Throughout the rest of this chapter we suppose that the following assumption is satisfied.

Assumption A1. *(a) $J(\widehat{\pi}, \widehat{x}) < \infty$ for some policy $\widehat{\pi}$ and some initial state $\widehat{x}$.*

(b) The one-stage cost function $c(x, a)$ is nonnegative.

(c) $c(x, a)$ is inf-compact, that is, the set $C_r := \{(x, a) \in \mathbb{K} \,|\, c(x, a) \leq r\} \subset \mathbb{K}$ is compact for each $r \in \mathbb{R}$.

(d) The transition law P is weakly continuous, that is,

$$(x, a) \;\mapsto\; \int u(y) P(dy|x, a)$$

is a continuous bounded function on $\mathbb{K}$ for every continuous bounded function u on $\mathbb{X}$.

Remark 11.3 *(a) Assumption A1(c) clearly implies that the one-stage cost c is lower semicontinuous, that is, the set C_r is closed for each $r \in \mathbb{R}$. It also implies that the set $A_r(x) := \{a \in \mathbb{A}(x) \,|\, c(x, a) \leq r\}$ is compact for each $x \in \mathbb{X}$ and $r \in \mathbb{R}$. These facts, together with Assumption A1(b), yield, in particular, that the function $x \mapsto \min_{a \in \mathbb{A}(x)} c(x, a)$ is measurable (see Rieder [34]).*

(b) Another important consequence of Assumption A1(c) is as follows. Let $M = \{\mu_i, i \in I\}$ be an arbitrary family of p.m.'s on $\mathbb{X} \times \mathbb{A}$, concentrated on $\mathbb{K}$. If

$$k \;:=\; \sup_{i \in I} \langle \mu_i, c \rangle \;<\; \infty,$$

then M is tight, that is, for each $\epsilon > 0$ there is a compact subset $K = K_\epsilon$ of $\mathbb{K}$ such that

$$\sup_{i \in I} \mu_i(K^c) \;<\; \epsilon,$$

where K^c stands for the complement of K. Indeed, let C_r be as in Assumption A1(c), with $r > 0$, and note that, for all $i \in I$,

$$\begin{aligned} k \;\geq\; \langle \mu_i, c \rangle \;&\geq\; \int_{C_r^c} c(x, a)\, \mu_i(d(x, a)) \\ &\geq\; \mu_i(C_r^c) \cdot \inf \{c(x, a) \,|\, (x, a) \notin C_r\} \\ &\geq\; \mu_i(C_r^c) \cdot r. \end{aligned}$$

Hence $\sup_{i \in I} \mu_i(C_r^c) \leq k/r$ for all $r > 0$, which implies that M is tight. Note also that the above remains true if M is a bounded set of measures (that is, $\sup_{i \in I} \mu_i(\mathbb{K}) \leq m$ for some $m > 0$) rather than p.m.'s.

Let $\rho_{\min}$ be defined as

$$\rho_{\min} := \inf_{\mathcal{P}(\mathbb{X})} J^*(\nu) = \inf_{\mathcal{P}(\mathbb{X})} \inf_{\Pi} J(\pi, \nu). \tag{11.29}$$

A pair $(\pi^*, \nu^*) \in \Pi \times \mathcal{P}(\mathbb{X})$ that satisfies

$$J(\pi^*, \nu^*) \;=\; \rho_{\min} \tag{11.30}$$

is called a "minimum pair".

In this section we introduce a linear program (P) such that

$$\sup(\mathrm{P}^*) \le \rho_{\min} \le \inf(\mathrm{P}). \tag{11.31}$$

Then we will show that (P) is *solvable* and that there is *no duality gap*, so that instead of (11.31) we will have the stronger relation

$$\sup(\mathrm{P}^*) = \rho_{\min} = \min(\mathrm{P}). \tag{11.32}$$

Moreover, disintegrating any optimal solution μ^* of (P) as $\mu^* = \widehat{\mu^*} \cdot \varphi_*$ (see Remark 11.2) for some $\varphi_* \in \Phi$, yields that $(\varphi_*, \widehat{\mu^*})$ satisfies (11.30), that is, $(\varphi_*, \widehat{\mu^*})$ is a minimum pair, and, in addition,

$$J(\varphi_*, x) \;=\; \rho_{\min} \qquad \widehat{\mu^*}\text{-a.e.} \tag{11.33}$$

11.3.1 The linear programs

We first introduce the components of the linear program (P) as in §11.2.2.

The dual pairs. Let $\mathbb{K} \subset \mathbb{X} \times \mathbb{A}$ be the set defined in (11.27), and let $w(x,a)$ and $w_0(x)$ be the weight functions on $\mathbb{K}$ and $\mathbb{X}$, respectively, defined as

$$w(x,a) := 1 + c(x,a), \qquad w_0(x) := \min_{\mathbb{A}(x)} w(x,a). \tag{11.34}$$

By Remark 11.3(a), $w_0(x)$ is measurable. The dual pairs we are concerned with are

$$(\mathcal{X}, \mathcal{Y}) := (\mathbb{M}_w(\mathbb{K}), \mathbb{B}_w(\mathbb{K})) \tag{11.35}$$

and

$$(\mathcal{Z}, \mathcal{W}) := (\mathbb{R} \times \mathbb{M}_{w_0}(\mathbb{X}), \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X})). \tag{11.36}$$

In particular, the bilinear form on $(\mathbb{M}_w(\mathbb{K}), \mathbb{B}_w(\mathbb{K}))$ is as in (11.9), namely,

$$\langle \mu, u \rangle := \int_{\mathbb{K}} u \, d\mu, \tag{11.37}$$

and on $(\mathbb{R} \times \mathbb{M}_{w_0}(\mathbb{X}), \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X}))$ is

$$\langle (r, \nu), (\rho, v) \rangle := r \cdot \rho + \int_{\mathbb{X}} v \, d\nu. \tag{11.38}$$

Note that, since $c(x,a)$ is nonnegative [Assumption A1(b)], (11.34) yields

$$0 \le c(x,a) \le w(x,a) \quad \forall (x,a) \in \mathbb{K},$$

which implies that *the cost-per-stage function c is in* $\mathbb{B}_w(\mathbb{K})$, and, on the other hand,

$$1 \le w_0(x) \le w(x,a) \quad \forall (x,a) \in \mathbb{K}. \tag{11.39}$$

Moreover, the policy $\widehat{\pi}$ and the initial state $\widehat{x}$ in Assumption A1(a) satisfy

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} E_{\widehat{x}}^{\widehat{\pi}}[w(x_t,a_t)] = 1 + J(\widehat{\pi},\widehat{x}) < \infty. \tag{11.40}$$

We also need the following additional assumption.

Assumption A2. *There is a constant k such that*

$$\int_{\mathbb{X}} w_0(y) P(dy|x,a) \le k w(x,a) \quad \forall (x,a) \in \mathbb{K}.$$

In other words, Assumption A2 states that the function

$$(x,a) \mapsto \int_{\mathbb{X}} w_0(y) P(dy|x,a) \quad \text{is in} \quad \mathbb{B}_w(\mathbb{K}).$$

The linear maps. Let L_0 and L_1 be the linear maps

$$L_0 : \mathbb{M}_w(\mathbb{K}) \to \mathbb{R} \quad \text{and} \quad L_1 : \mathbb{M}_w(\mathbb{K}) \to \mathbb{M}_{w_0}(X),$$

with

$$L_0\mu := \langle \mu, 1 \rangle = \mu(\mathbb{K}) \tag{11.41}$$

and

$$(L_1\mu)(B) := \widehat{\mu}(B) - \int_{\mathbb{K}} P(B|x,a)\mu(d(x,a)) \quad \text{for} \quad B \in \mathcal{B}(\mathbb{X}), \tag{11.42}$$

where $\widehat{\mu}$ denotes the marginal of μ on $\mathbb{X}$. Finally, let

$$L : \mathbb{M}_w(\mathbb{K}) \to \mathbb{R} \times \mathbb{M}_{w_0}(\mathbb{X})$$

be the linear map

$$L\mu := (L_0\mu, L_1\mu) \quad \text{for} \quad \mu \in \mathbb{M}_w(\mathbb{K}), \tag{11.43}$$

with adjoint

$$L^* : \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X}) \to \mathbb{B}_w(\mathbb{K})$$

given by

$$L^*(\rho,u)(x,a) := \rho + u(x) - \int_{\mathbb{X}} u(y) P(dy|x,a) \tag{11.44}$$

for every pair (ρ,u) in $\mathbb{R} \times \mathbb{B}_{w_0}(X)$ and (x,a) in $\mathbb{K}$. Hence, (11.39), Assumption A2 and Proposition 11.1 yield that

the linear map L in (11.43) is weakly continuous, (11.45)

that is, continuous with respect to the weak topologies

$$\sigma(\mathbb{M}_w(\mathbb{K}), \mathbb{B}_w(\mathbb{K})) \text{ and } \sigma(\mathbb{R} \times \mathbb{M}_{w_0}(\mathbb{X}), \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X})).$$

The linear programs. Consider the vectors

$$b := (1,0) \text{ in } \mathbb{R} \times \mathbb{M}_{w_0}(\mathbb{X}), \text{ and } c \text{ in } \mathbb{B}_w(\mathbb{K}),$$

where c is the cost-per-stage function, as well as the positive cone

$$K := \mathbb{M}_w(\mathbb{K})_+, \tag{11.46}$$

whose dual cone is

$$K^* := \mathbb{B}_w(\mathbb{K})_+. \tag{11.47}$$

Then the **primal** linear program is

(P) minimize $\langle \mu, c \rangle$

$$\text{subject to: } L\mu = (1,0), \quad \mu \in \mathbb{M}_w(\mathbb{K})_+. \tag{11.48}$$

More explicitly, by (11.41)–(11.43), the constraint (11.48) is satisfied if

$$\mu(\mathbb{K}) = 1 \quad \text{with} \quad \mu \in \mathbb{M}_w(\mathbb{K})_+, \tag{11.49}$$

and $L_1\mu = 0$, i.e.

$$\widehat{\mu}(B) - \int_{\mathbb{K}} P(B|x,a)\mu(d(x,a)) = 0 \quad \forall B \in \mathcal{B}(\mathbb{X}), \tag{11.50}$$

with $\mu \in \mathbb{M}_w(\mathbb{K})_+$. Observe that (11.49) requires μ to be a *probability measure* (p.m.). Moreover, disintegrating μ into $\varphi \cdot \hat{\mu}$ (see Remark 11.2) for some $\varphi \in \Phi$, and using the notation

$$P_\varphi(\bullet|x) := \int_{\mathbb{A}} P(\bullet|x,a)\varphi(da|x),$$

(11.50) can be written as

$$\widehat{\mu}(B) = \int_{\mathbb{X}} P_\varphi(B|x)\widehat{\mu}(dx) \quad \forall B \in \mathcal{B}(\mathbb{X}),$$

which means that μ *is feasible for* (P) *if* μ *is a p.m. on* $\mathbb{K}$ *such that its marginal* $\widehat{\mu}$ *on* $\mathbb{X}$ *is an invariant p.m. (i.p.m.) for the transition kernel* $P_\varphi(\bullet|\bullet) = P(\varphi)$.

On the other hand, observe that

$$\langle b, v \rangle = \langle (1,0), (\rho, u) \rangle = \rho \quad \forall v = (\rho, u) \in \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X}).$$

Hence, by (11.47) and (11.44), the **dual** of (P) is

(P*) maximize ρ

$$\text{subject to : } \quad \rho + u(x) - \int_{\mathbb{X}} u(y)P(dy|x,a) \le c(x,a) \tag{11.51}$$
$$\forall (x,a) \in \mathbb{K}, \text{ with } (\rho, u) \in \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X}).$$

This completes the specification of the linear programs associated to the AC problem.

11.3.2 Solvability of (P)

Before proceeding to verify (11.31) and (11.32), let us note the following.

Remark 11.4 *We will use the following conventions:*

(a) A measure μ on $\mathbb{K} \subset \mathbb{X} \times \mathbb{A}$ may (and will) be viewed as a measure on all of $\mathbb{X} \times \mathbb{A}$ by defining $\mu(\mathbb{K}^c) := 0$, where $\mathbb{K}^c := \mathbb{X} \times \mathbb{A} \setminus \mathbb{K}$.

(b) We will regard $c : \mathbb{K} \to \mathbb{R}_+$ as a function on all of $\mathbb{X} \times \mathbb{A}$ with $c(x,a) := +\infty$ if (x,a) is in $\mathbb{K}^c$. Observe that this convention is consistent with Assumption A1(c), and, moreover, the weight function $w = +\infty$ on $\mathbb{K}^c$. Any other function u in $\mathbb{B}_w(\mathbb{K})$ can be arbitrarily extended to $\mathbb{X} \times \mathbb{A}$, for example, as $u := 0$ on $\mathbb{K}^c$.

(c) $0 \cdot (+\infty) := 0$

(d) A function u in $\mathbb{B}_{w_0}(\mathbb{X})$ will also be seen as the function in $\mathbb{B}_w(\mathbb{K})$ given by $u(x,a) := u(x)$ for all (x,a) in $\mathbb{K}$.

Then, in particular, we may write the bilinear form in (11.37) as

$$\langle \mu, u \rangle = \int_{\mathbb{X} \times \mathbb{A}} u \, d\mu$$

for any measure μ in $\mathbb{M}_w(\mathbb{K})$ and any function u in $\mathbb{B}_w(\mathbb{K})$ or in $\mathbb{B}_{w_0}(\mathbb{X})$.

We will next show that (P) is **solvable** and

$$\sup(\mathrm{P}^*) \leq \rho_{\min} = \min(\mathrm{P}). \tag{11.52}$$

To do this let us first recall that a randomized stationary policy $\varphi \in \Phi$ is said to be *stable* if there exists an invariant probability measure (i.p.m.) p_φ for the transition kernel $P_\varphi(B|x) := \int_{\mathbb{A}} P(B|x,a)\varphi(da|x)$, i.e.

$$p_\varphi(B) := \int P_\varphi(B|x) \, p_\varphi(dx),$$

and, in addition, the average cost $J(\varphi, p_\varphi)$ satisfies that

$$J(\varphi, p_\varphi) = \int c_\varphi(x) \, p_\varphi(dx),$$

where $c_\varphi(x) = \int_{\mathbb{A}} c(x,a)\varphi(da|x)$.

Theorem 11.3 *Suppose that Assumptions A1 and A2 are satisfied. Then:*

(a) **[Solvability of (P)].** *There exists an optimal solution μ^* for (P), and*

$$\min(P) = \rho_{\min} = \langle \mu^*, c \rangle. \tag{11.53}$$

The disintegration of μ^* as $\widehat{\mu^*} \cdot \varphi_*$ for some $\varphi_* \in \Phi$ yields that $(\varphi_*, \widehat{\mu^*})$ is a minimum pair and, in addition,

$$J(\varphi_*, x) = \rho_{\min} \quad \widehat{\mu^*}\text{-a.e.}. \tag{11.54}$$

(b) **[Consistency of (P*)].** *The dual problem (P*) is consistent and it satisfies the inequality in (11.52).*

Proof. (a) By Theorem 5.7.9(a) in [19], there exists a stable randomized stationary policy φ_* such that $(\varphi_*, p_{\varphi_*})$ is a minimum pair. That is, p_{φ_*} is an i.p.m. for the transition kernel

$$P_{\varphi_*}(B|x) := \int_{\mathbb{A}} P(B|x,a)\varphi_*(da|x),$$

and

$$J(\varphi^*, p_{\varphi_*}) = \int_{\mathbb{X}} c_{\varphi_*}(x) p_{\varphi_*}(dx) = \rho_{\min} < \infty, \tag{11.55}$$

where

$$c_{\varphi_*}(x) := \int_{\mathbb{A}} c(x,a)\varphi_*(da|x).$$

Furthermore, as p_{φ_*} is an i.p.m. for P_{φ_*}, for every B in $\mathcal{B}(\mathbb{X})$ we have

$$p_{\varphi_*}(B) = \int_{\mathbb{X}} P_{\varphi_*}(B|x) p_{\varphi_*}(dx),$$

i.e.

$$p_{\varphi_*}(B) = \int_{\mathbb{X}} \int_{\mathbb{A}} P(B|x,a)\varphi_*(da|x) p_{\varphi_*}(dx). \tag{11.56}$$

Now let μ^* be the measure on $\mathbb{X} \times \mathbb{A}$ defined as

$$\mu^*(B \times C) := \int_B \varphi_*(C|x) p_{\varphi_*}(dx) \quad \forall B \in \mathcal{B}(\mathbb{X}),\ C \in \mathcal{B}(\mathbb{A}).$$

Then, μ^* is a p.m. on $\mathbb{X} \times \mathbb{A}$, concentrated on $\mathbb{K}$, and its marginal on $\mathbb{X}$ coincides with p_{φ_*}:

$$\widehat{\mu}^*(B) := \mu^*(B \times \mathbb{A}) = p_{\varphi_*}(B) \quad \forall B \in \mathcal{B}(\mathbb{X}).$$

It follows that we may rewrite (11.56) and (11.55) as

$$\widehat{\mu}^*(B) - \int_{\mathbb{K}} P(B|x,a)\mu^*(d(x,a)) = 0 \quad \forall B \in \mathcal{B}(\mathbb{X}),$$

and

$$J(\varphi^*, p_{\varphi_*}) = \langle \mu^*, c \rangle = \rho_{\min} < \infty, \tag{11.57}$$

which means that we already have the second equality in (11.53), as well as the equalities $\mu^*(\mathbb{K}) = 1$ and $L_1\mu^* = 0$ in (11.49) and (11.50).

Therefore, to prove (11.53) in part (a) it suffices to show that

(i) μ^* is in $\mathbb{M}_w(\mathbb{K})$ so that μ^* is indeed feasible for (P); and

(ii) $\langle \mu, c \rangle \geq \rho_{\min}$ for any feasible solution μ for (P), which would yield $\inf(\mathrm{P}) \geq \rho_{\min}$.

In other words, (i), (ii) and (11.57) will give that μ^* is feasible for (P) and

$$\rho_{\min} = \langle \mu^*, c \rangle \geq \inf(\mathrm{P}) \geq \rho_{\min}, \quad \text{i.e.} \quad \langle \mu^*, c \rangle = \rho_{\min}.$$

Proof of (i). This is easy because, by (11.34) and (11.57),

$$\langle \mu^*, w \rangle = 1 + \langle \mu^*, c \rangle < \infty.$$

Proof of (ii). If μ satisfies (11.49) and (11.50), then, in particular, μ is a probability measure on $\mathbb{X} \times \mathbb{A}$ concentrated on $\mathbb{K}$. Thus, μ can be "disintegrated" as $\widehat{\mu} \cdot \varphi$ for some $\varphi \in \Phi$ (see Remark 11.2). Furthermore, taking $(\varphi, p_\varphi) := (\varphi, \widehat{\mu})$, (11.50) gives that φ is a stable randomized stationary policy, and, therefore, by the definition of $\rho_{\min}$,

$$\langle \mu, c \rangle = J(\varphi, \widehat{\mu}) \geq \rho_{\min}.$$

This proves (ii).

To complete the proof of part (a), observe that from $\langle c, \mu^* \rangle = \rho_{\min}$, it follows that $c_{\varphi_*} \in L_1(\widehat{\mu^*})$. Therefore, by Birkhoff's Individual Ergodic Theorem

$$\int_{\mathbb{X}} J(\varphi_*, x)\, d\widehat{\mu^*} \;=\; \int_{\mathbb{X}} c_{\varphi_*}\, d\widehat{\mu^*} \;=\; \rho_{\min},$$

and (11.54) follows from combining the above equality with $J(\varphi_*, x) \geq \rho_{\min}$.

(b) By (a) and the weak duality property (11.21), to prove (b) it suffices to show that (P*) is consistent. This, however, is obvious: for example, the pair (ρ, u) with $\rho = u(\cdot) \equiv 0$ satisfies (11.51). ∎

Hence, the stationary policy $\varphi_* \in \Phi$ in Theorem 11.3 is expected-average cost optimal for all initial states $x \in S$, where the absorbing set S is the support of the measure $\widehat{\mu^*}$.

Remark 11.5 *As mentioned in the Introduction, the LP formulation of MDPs began in the early 1960s as a way to solve the associated optimality (or dynamic programming) equation. In particular, for the average cost problem the question was to find a solution (ρ, u) to the Average Cost Optimality Equation (ACOE) studied in previous chapters, that is, a number ρ and a function u on $\mathbb{X}$ such that*

$$\rho + u(x) \;=\; \min_{a \in \mathbb{A}(x)} \left[c(x,a) + \int_{\mathbb{X}} u(y) P(dy|x,a) \right] \qquad \forall (x,a) \in \mathbb{K}. \tag{11.58}$$

The idea was the following. If (ρ, u) satisfies (11.58), then we obviously have

$$\rho + u(x) \;\leq\; c(x,a) + \int_{\mathbb{X}} u(y) P(dy|x,a) \qquad \forall (x,a) \in \mathbb{K},$$

or, equivalently,

$$\rho + u(x) - \int_{\mathbb{X}} u(y) P(dy|x,a) \;\leq\; c(x,a) \qquad \forall (x,a) \in \mathbb{K}, \tag{11.59}$$

which is exactly the same as (11.51). On the other hand, from (11.39) and Assumption A2, it is easy to check that

$$\lim_{n \to \infty} \frac{1}{n} E_\nu^\pi u(x_n) \;=\; 0$$

for any function $u \in \mathbb{B}_{w_0}(\mathbb{X})$, any policy π and any initial distribution ν for which $J(\pi,\nu) < \infty$, which in turn, by (11.59), yields that

$$\rho \leq J(\pi,\nu).$$

As this holds for any pair $(\rho,u) \in \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X})$ that satisfies (11.59), that is, for any feasible solution for (P^*), we obtain the inequality $\sup(P^*) \leq \rho_{\min}$, already obtained in Theorem 11.3(b) using standard LP arguments. Thus, the essence of the original LP approach to MDPs was to give conditions for the dual (P^*) to be solvable, and for the absence of a duality gap. We will now address these questions in Theorems 11.4 and 11.6.

11.3.3 Absence of duality gap

We now prove (11.32).

Theorem 11.4 (Absence of duality gap.) *If Assumptions A1 and A2 are satisfied, then (11.32) holds.*

Proof. We wish to use Theorem 11.2 with $\mathcal{Z}$ and L as in (11.36) and (11.43), respectively. Hence, we wish to show that the set

$$H := \{(L\mu, \langle \mu, c\rangle + r) | \mu \in \mathbb{M}_w(\mathbb{K})_+,\ r \geq 0\}$$

is closed in the weak topology

$$\sigma(\mathbb{R} \times \mathbb{M}_{w_0}(\mathbb{X}) \times \mathbb{R},\ \mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X}) \times \mathbb{R}).$$

Let $(D, \leq)$ be a directed set, and consider a net $\{(\mu_\alpha, r_\alpha), \alpha \in D\}$ in $\mathbb{M}_w(\mathbb{K})_+ \times \mathbb{R}_+$ such that

$$L_0\mu_\alpha := \mu_\alpha(\mathbb{K}) \to r_* \tag{11.60}$$

$$\langle L_1\mu_\alpha, u\rangle \to \langle \nu_*, u\rangle \qquad \forall u \in \mathbb{B}_{w_0}(\mathbb{X}),\ \text{and} \tag{11.61}$$

$$\langle \mu_\alpha, c\rangle + r_\alpha \to \rho_*. \tag{11.62}$$

We will show that $((r_*, \nu_*), \rho_*)$ is in H; that is, there exists a measure μ in $\mathbb{M}_w(\mathbb{K})_+$ and a number $r \geq 0$ such that

$$r_* = L_0\mu := \mu(\mathbb{K}), \tag{11.63}$$

$$\nu_* = L_1\mu,\ \text{and} \tag{11.64}$$

$$\rho_* = \langle \mu, c\rangle + r. \tag{11.65}$$

We shall consider two cases, $r_* = 0$ and $r_* > 0$.

Case 1: $r_* = 0$. By definition (11.41) of L_0,

$$L_0\mu_\alpha = \mu_\alpha(\mathbb{K}). \tag{11.66}$$

Therefore, if $r_* = 0$ in (11.60), it follows easily that (11.63)–(11.65) hold with $\mu(\cdot) = 0$ and $r = \rho_*$.

Case 2: $r_* > 0$. By (11.60) [together with (11.66)] and (11.62), there exists α_0 in D such that

$$0 < \mu_\alpha(\mathbb{K}) \leq 2r_* \text{ and } \langle \mu_\alpha, c\rangle \leq 2\rho_* \quad \forall \alpha \geq \alpha_0. \tag{11.67}$$

Hence, as $\langle \mu_\alpha, c+1\rangle = \langle \mu_\alpha, c\rangle + \mu_\alpha(\mathbb{K})$, we get that $\Gamma := \{\mu_\alpha, \alpha \geq \alpha_0\}$ is a *bounded* set of measures, which combined with Assumption A1(c) yields that Γ is *tight* (see Remark 11.3(b)). Moreover, if $\mu_\alpha(\mathbb{K}) > 0$, we may "normalize" μ_α rewriting it as $\mu_\alpha(\cdot)/\mu_\alpha(\mathbb{K})$, and so we may assume that Γ is a (tight) family of probability measures. Then, by Prohorov's Theorem (see e.g. [4]), for each sequence $\{\mu_n\}$ in Γ there is a subsequence $\{\mu_m\}$ and a p.m. μ on $\mathbb{K}$ such that

$$\langle \mu_m, v\rangle \to \langle \mu, v\rangle \quad \forall v \in C_b(\mathbb{K}), \tag{11.68}$$

where $C_b(\mathbb{K})$ denotes the space of continuous bounded functions on $\mathbb{K}$.

In particular, taking $v(\cdot) \equiv 1$, (11.60) yields that μ *satisfies (11.63)*. We will next show that

(i) μ is in $\mathbb{M}_w(\mathbb{K})_+$, that is, $\|\mu\|_w := \langle \mu, w\rangle < \infty$ [see (11.7)], and

(ii) μ satisfies (11.64).

Proof of (i). As $w := 1 + c$, to prove (i) we need to show that $\langle \mu, c\rangle$ is finite. We will prove the latter by showing that

$$[(11.68),\ c \geq 0 \text{ and l.s.c.}] \Rightarrow \liminf \langle \mu_m, c\rangle \geq \langle \mu, c\rangle. \tag{11.69}$$

Indeed, if $c \geq 0$ and l.s.c. [as in Assumption A1(b)], then there exists an increasing sequence of functions v_k in $C_b(\mathbb{K})$ such that $v_k \uparrow c$. It follows from (11.68) that for each k

$$\liminf_{m\to\infty} \langle \mu_m, c\rangle \geq \liminf_{m\to\infty} \langle \mu_m, v_k\rangle = \langle \mu, v_k\rangle.$$

Thus, letting $k \to \infty$, the Monotone Convergence Theorem gives (11.69).

Proof of (ii). The weak continuity condition on P [Assumption A1(d)] implies that the adjoint of L_1, namely,

$$(L_1^* u)(x,a) := u(x) - \int_{\mathbb{X}} u(y) P(dy|x,a),$$

maps $C_b(\mathbb{X})$ into $C_b(\mathbb{K})$. Therefore, (11.68) and (11.61) yield that for any function u in $C_b(\mathbb{X})$

$$\begin{aligned} \langle L_1\mu, u\rangle = \langle \mu, L_1^* u\rangle &= \lim_{m\to\infty} \langle \mu_m, L_1^* u\rangle \text{ [by (11.68)]} \\ &= \lim_{m\to\infty} \langle L_1\mu_m, u\rangle \\ &= \langle \nu_*, u\rangle \text{ [by (11.61)]}. \end{aligned}$$

That is, $\langle L_1\mu, u\rangle = \langle \nu_*, u\rangle$ for any function u in $C_b(\mathbb{X})$, which implies (11.64). This proves (ii).

Summarizing, we have shown that μ is a measure in $\mathbb{M}_w(\mathbb{K})_+$ that satisfies (11.63) and (11.64). Finally, from (11.69) and (11.62) we see that

$$\rho_* \geq \langle \mu, c \rangle + \liminf_{m \to \infty} r_m \geq \langle \mu, c \rangle \quad \text{as} \quad r_m \geq 0 \ \forall m.$$

Thus, defining $r := \rho_* - \langle \mu, c \rangle$ (≥ 0), we conclude that μ and r satisfy (11.63), (11.64) and (11.65). This shows that H is indeed weakly closed, and so (11.32) follows. ∎

Having (11.32), in the following sections we consider conditions for the solvability of the dual problem (P*) and for the convergence of approximations to the optimal values $\max(\mathrm{P}^*)$ and $\min(\mathrm{P})$.

11.4 APPROXIMATING SEQUENCES AND STRONG DUALITY

In the rest of this chapter we are mainly interested in the approximation of the AC-related linear program (P) and its dual (P*). In this section we first study minimizing sequences for (P), and then maximizing sequences for (P*).

11.4.1 Minimizing sequences for (P)

By Definition 11.2(a), a sequence of measures μ_n in $\mathbb{M}_w(\mathbb{K})_+$ is a **minimizing sequence** for (P) if each μ_n is feasible for (P), that is, it satisfies (11.48), and in addition

$$\langle \mu_n, c \rangle \downarrow \min(\mathrm{P}), \tag{11.70}$$

where we have used that (P) is *solvable* [Theorem 11.3(a)] to write its value as min (P) rather than inf (P).

Theorem 11.5 *Suppose that Assumptions A1 and A2 are satisfied. If $\{\mu_n\}$ is a minimizing sequence for* (P)*, then there exists a subsequence $\{j\}$ of $\{n\}$ such that $\{\mu_j\}$ converges in the weak topology $\sigma(\mathbb{M}(\mathbb{K}), C_b(\mathbb{K}))$ to an optimal solution for* (P).

Proof. Let $\{\mu_n\}$ be a minimizing sequence for (P); that is [by (11.48)],

$$\langle \mu_n, 1 \rangle = 1 \quad \text{and} \quad L_1 \mu_n = 0 \quad \forall n, \tag{11.71}$$

and (11.70) holds. In particular, (11.70) implies that for any given $\varepsilon > 0$ there exists $n(\varepsilon)$ such that

$$\min(\mathrm{P}) \leq \langle \mu_n, c \rangle \leq \min(\mathrm{P}) + \varepsilon \quad \forall n \geq n(\varepsilon). \tag{11.72}$$

By the second inequality [together with Assumption A1(c)], the sequence $\{\mu_n\}$ is tight (see Remark 11.3(b)), so that there exists a p.m. μ^* on $\mathbb{K}$ and a subsequence $\{j\}$ of $\{n\}$ such that

$$\langle \mu_j, v \rangle \to \langle \mu^*, v \rangle \quad \forall v \in C_b(\mathbb{K}). \tag{11.73}$$

Moreover, by (11.69),

$$\langle \mu^*, c \rangle \leq \liminf_{j \to \infty} \langle \mu_j, c \rangle \leq \min(\mathrm{P}) + \varepsilon. \tag{11.74}$$

Thus, as ε was arbitrary, the latter inequality and (11.72) yield

$$\min(\mathrm{P}) = \langle \mu^*, c \rangle. \tag{11.75}$$

This will prove that μ^* is optimal for (P) provided that μ^* is *feasible* for (P); in other words, provided that μ^* is a measure in $\mathbb{M}_w(\mathbb{K})_+$ and that

$$L\mu^* = (L_0\mu^*, L_1\mu^*) = (1, 0). \tag{11.76}$$

This, however, is obvious because (11.74) yields $\langle \mu^*, w \rangle = 1 + \langle \mu^*, c \rangle < \infty$, whereas (11.76) follows from (11.71) and (11.73). ∎

11.4.2 Maximizing sequences for (P^*)

By Definition 11.2(b) and the definition of the dual program (P*), a sequence (ρ_n, u_n) in $\mathbb{R} \times \mathbb{B}_{w_0}(\mathbb{X})$ is a maximizing sequence for (P*) if

$$\rho_n + u_n(x) \le c(x,a) + \int_{\mathbb{X}} u_n(y) Q(dy|x,a) \tag{11.77}$$

for all n and $(x,a) \in \mathbb{K}$, and, in addition,

$$\rho_n = \langle (1,0), (\rho_n, u_n) \rangle \uparrow \sup(\mathrm{P}^*). \tag{11.78}$$

The following theorem shows that the existence of a suitable maximizing sequence for (P*) implies, in particular, that the *strong duality* condition for (P) holds [see (11.24)].

Theorem 11.6 [Solvability of (P*), strong duality and the ACOE.] *Suppose that Assumptions A1 and A2 are satisfied, and, furthermore, there exists a maximizing sequence* (ρ_n, u_n) *for* (P*) *with* $\{u_n\}$ *bounded in the* w_0*-norm, that is,*

$$\|u_n\|_{w_0} \le k \quad \forall n, \tag{11.79}$$

for some constant k*. Then:*

(a) The dual problem (P*) *is solvable.*

(b) The strong duality condition holds, that is, $\max(\mathrm{P}^*) = \min(\mathrm{P})$.

(c) If μ^* *is an optimal solution for the primal program* (P)*, then the ACOE holds* $\widehat{\mu}^*$*-a.e., where* $\widehat{\mu}^*$ *is the marginal of* μ^* *on* $\mathbb{X}$*; in fact, there is a function* h^* *in* $\mathbb{B}_{w_0}(\mathbb{X})$ *and a deterministic stationary policy* f_* *such that*

$$\begin{aligned} \rho^* + h^*(x) &= \min_{\mathbb{A}(x)} \left[c(x,a) + \int_{\mathbb{X}} h^*(y) P(dy|x,a) \right] \\ &= c(x, f_*) + \textstyle\int_{\mathbb{X}} h^*(y) P(dy|x, f_*) \end{aligned} \tag{11.80}$$

for $\widehat{\mu}^*$*-almost all* $x \in \mathbb{X}$.

Proof. (a) By Theorem 11.4 we have

$$\sup(\mathrm{P}^*) = \rho^* = \min(\mathrm{P}) \tag{11.81}$$

and, moreover, we can write (11.78) as

$$\rho_n \uparrow \rho^*. \tag{11.82}$$

Now define the function

$$h^*(x) := \limsup_{n\to\infty} u_n(x),$$

which belongs to $\mathbb{B}_{w_0}(\mathbb{X})$, by (11.79). Therefore [by (11.82) and Fatou's Lemma], taking $\limsup_n$ in (11.77) we obtain

$$\rho^* + h^*(x) \le c(x,a) + \int_{\mathbb{X}} h^*(y)P(dy|x,a) \quad \forall (x,a) \in \mathbb{K}.$$

This yields that (ρ^*, h^*) is feasible for (P^*) [see (11.51)], which together with the first equality in (11.81) shows that (ρ^*, h^*) is in fact *optimal for* (P^*).

(b) This part follows from (a) and (11.81).

(c) Let us first note that if μ is feasible for (P) and (ρ, u) is feasible for (P^*), then

$$\langle L\mu, (\rho, u)\rangle = \langle (1,0), (\rho, u)\rangle = \rho,$$

or, equivalently,

$$\langle \mu, L^*(\rho, u)\rangle = \rho, \tag{11.83}$$

where L^* is the adjoint of L, in (11.44). Now let μ^* be an optimal solution for (P), and (ρ^*, h^*) an optimal solution for (P^*). By part (b) we have

$$\langle \mu^*, c\rangle = \rho^*,$$

whereas (11.83) gives

$$\langle \mu^*, L^*(\rho^*, h^*)\rangle = \rho^*.$$

Thus, subtracting the last two equalities we get

$$\langle \mu^*, c - L^*(\rho^*, h^*)\rangle = 0,$$

i.e.

$$\int_{\mathbb{X}\times\mathbb{A}} [c(x,a) - L^*(\rho^*, h^*)(x,a)]\mu^*(d(x,a)) = 0. \tag{11.84}$$

We may disintegrate μ^* as $\mu^*(d(x,a)) = \varphi(da|x)\widehat{\mu}^*(dx)$ for some stochastic kernel $\varphi \in \Phi$, and then [using (11.44)] we can rewrite (11.84) as

$$\int_{\mathbb{X}} \left[c(x,\varphi) - \rho^* - h^*(x) + \int_{\mathbb{X}} h^*(y)P(dy|x,\varphi) \right] \widehat{\mu}^*(dx) = 0.$$

Therefore, as the integrand is *nonnegative* [by (11.51)], we get that for $\widehat{\mu}^*$-a.a. (almost all) x in $\mathbb{X}$

$$\begin{aligned}\rho^* + h^*(x) &= c(x,\varphi) + \int_{\mathbb{X}} h^*(y) P(dy|x,\varphi) \\ &= \int_{\mathbb{A}} \left[c(x,a) + \int_{\mathbb{X}} h^*(y) P(dy|x,a) \right] \varphi(da|x),\end{aligned}$$

and so

$$\rho^* + h^*(x) \geq c(x,f_*) + \int_{\mathbb{X}} h^*(y) P(dy|x,f_*) \quad \widehat{\mu}^* - a.a.\ x \in \mathbb{X} \tag{11.85}$$

for some decision function $f_* \in \mathbb{F}$ whose existence is guaranteed by a measurable selection result (see Lemma 15.1 in Hinderer [24]). Finally, as (11.51) implies

$$\rho^* + h^*(x) \leq \min_{\mathbb{A}(x)} \left[c(x,a) + \int_{\mathbb{X}} h^*(y) P(dy|x,a) \right] \text{ for all } (x,a) \in \mathbb{K},$$

we get that, by (11.85), for $\widehat{\mu}^*$-a.a. $x \in \mathbb{X}$

$$\begin{aligned}\rho^* + h^*(x) &\geq c(x,f_*) + \int_{\mathbb{X}} h^*(y) P(dy|x,f_*) \\ &\geq \min_{\mathbb{A}(x)} \left[c(x,a) + \int_{\mathbb{X}} h^*(y) P(dy|x,a) \right] \\ &\geq \rho^* + h^*(x),\end{aligned}$$

and (11.80) follows. ∎

Theorem 11.6 refines Theorem 11.3. Indeed, from Theorem 11.3, we already had the existence of an average-cost optimal policy $\varphi_* \in \Phi$ for all initial states $x \in S$, where the absorbing set S is the support of the p.m. $\widehat{\mu^*}$. Now, we also get that the ACOE holds for all $x \in S$. In general, getting the same result for all initial states $x \in \mathbb{X}$, requires much stronger assumptions, not always realistic in practical applications. The LP approach in multichain MDPs to get a stationary expected average-cost optimal policy for all initial states (with optimal cost depending on the initial state) is still possible, but under stronger assumptions, and via the introduction of additional variables.

For MDPs with *finite* state and action spaces it is well known that the *policy iteration* (or Howard's) algorithm is equivalent to solving the primal program (P) by the simplex method. For non-finite MDPs there is nothing similar. In fact, for general infinite-dimensional linear programs it is not even known what the "simplex method" is! However, every algorithm that would produce a minimizing sequence for (P) can be interpreted as a "policy iteration" method since every feasible point μ of (P), when disintegrated as $\widehat{\mu} \cdot \varphi$ for some $\varphi \in \Phi$, can be associated with the stationary randomized policy φ. Similarly, every "policy iteration" algorithm moving in the space of stationary policies with an i.p.m. would produce a minimizing sequence for (P).

On the other hand, by "duality", one would expect that *value iteration* should be somehow related to solving the dual program (P*). This relation can be seen as follows.

Remark 11.6 [($\mathbf{P}^*$) vs. value iteration.] *Consider the n-step cost*

$$E_x^\pi \,[\sum_{t=0}^{n-1} c(x_t, a_t)\,]$$

and the corresponding value function $v_n(x)$, which can be computed recursively by

$$v_n(x) \;=\; \min_{a\in\mathbb{A}(x)}\,[c(x,a) + \int_{\mathbb{X}} v_{n-1}(y)P(dy|x,a)] \quad \textit{for } n = 1,2,\ldots, \tag{11.86}$$

with $v_0(.) \equiv 0$. Define $m_0 := 0$, and for $n = 1, 2, \ldots,$

$$\begin{aligned} m_n &:= \inf_{\mathbb{X}}\,[v_n(x) - v_{n-1}(x)] + m_{n-1}, \\ \rho_n &:= m_n - m_{n-1} \;=\; \inf_{\mathbb{X}}\,[v_n(x) - v_{n-1}(x)], \\ u_n(\cdot) &:= v_n(\cdot) - m_n. \end{aligned}$$

Then (11.86) can be rewritten as

$$\rho_n + u_n(x) \;=\; \min_{a\in\mathbb{A}(x)}\,[c(x,a) + \int_{\mathbb{X}} u_{n-1}(y)P(dy|x,a)], \tag{11.87}$$

which yields

$$\rho_n + u_n(x) \;\le\; c(x,a) + \int_{\mathbb{X}} u_{n-1}(y)P(dy|x,a).$$

Moreover, as

$$v_n(\cdot) - v_{n-1}(\cdot) \;\ge\; \rho_n,$$

the sequence $\{u_n(\cdot)\}$ is nondecreasing, and so from (11.87) we obtain

$$\rho_n + u_n(x) \;\le\; c(x,a) + \int_{\mathbb{X}} u_n(y)P(dy|x,a),$$

which means that the pairs (ρ_n, u_n) are feasible for (P^*); see (11.77). In addition, the sequence $\{\rho_n\}$ is nondecreasing because (using the inequality $\inf u(.) - \inf v(.) \ge \inf[u(.) - v(.)]$*) (11.86) yields*

$$v_n(x) - v_{n-1}(x) \;\ge\; \min_{a\in\mathbb{A}(x)} \int_{\mathbb{X}} [v_n(y) - v_{n-1}(y)]\; P(dy|x,a) \;\ge\; \rho_{n-1} \quad \forall x \in \mathbb{X};$$

hence, $\rho_n \ge \rho_{n-1}$. Therefore, there exists a number $\widehat{\rho} \le \sup(P^*)$ such that

$$\rho_n \;=\; \langle (1,0), (\rho_n, u_n)\rangle \;\uparrow\; \widehat{\rho}. \tag{11.88}$$

Thus, comparing (11.88) and (11.78) we conclude that the pairs (ρ_n, u_n) form a maximizing sequence for (P^*) provided that $\widehat{\rho} = \rho_{\min}$ (see (11.32)). In turn,

we can get the condition $\widehat{\rho} = \rho_{\min}$ *as follows. Suppose that the nonnegative function*

$$h(x) := \lim_{n\to\infty} u_n(x) \qquad \textit{for } x \in \mathbb{X}$$

is finite-valued. Then, letting $n \to \infty$ *in (11.87) and using Fatou 's Lemma, we get that*

$$\widehat{\rho} + h(x) \geq \inf_{a\in\mathbb{A}(x)} \left[c(x,a) + \int_{\mathbb{X}} h(y) P(dy|x,a) \right] \quad \forall x \in \mathbb{X},$$

which the so-called Average Cost Optimality Inequality (ACOI). Finally, if there exists a stationary policy $f \in \mathbb{F}$ *that attains the minimum on the right-hand-side of the ACOI, a standard argument shows that* $\widehat{\rho} \geq J(f,x)$ *for all* $x \in \mathbb{X}$*, which together with (11.29), gives* $\widehat{\rho} \geq \rho_{\min}$*. That is,* $\widehat{\rho} = \rho_{\min}$ *and so the pairs* (ρ_n, u_n) *form a maximizing sequence for* $(\mathbb{P}^*)$*.*

11.5 FINITE LP APPROXIMATIONS

We will now show a procedure to approximate the AC-related primal linear program (P) by *finite-dimensional* linear programs. We will work in essentially the same setting of the previous sections except that now we shall require the spaces $\mathbb{X}$ and $\mathbb{K}$ to be *locally compact separable metric* (LCSM) spaces. Hence throughout the following we suppose:

Assumption A3. Assumptions A1 and A2 are satisfied, and in addition $\mathbb{X}$ and $\mathbb{K}$ are LCSM spaces.

A sufficient condition for $\mathbb{K}$ to be LCSM is that $\mathbb{X}$ and $\mathbb{A}$ are LCSM spaces and that $\mathbb{K}$ is either open or closed in $\mathbb{X} \times \mathbb{A}$. On the other hand, the hypothesis that $\mathbb{X}$ and $\mathbb{K}$ are LCSM spaces ensures that $C_0(\mathbb{X})$ and $C_0(\mathbb{K})$ are both *separable* Banach spaces [See Remark 11.1(b).] In particular, $C_0(\mathbb{X})$ contains a *countable* subset $\mathcal{C}(\mathbb{X})$ which is dense in $C_0(\mathbb{X})$. This is a key fact to proceed with the first step of our approximation procedure.

11.5.1 Aggregation

Let $\mathcal{P}_w(\mathbb{K})$ be the family of *probability measures* (p.m.'s) in $\mathbb{M}_w(\mathbb{K})_+$, that is, the family of measures μ that satisfy (11.49). Thus, by (11.48), we may rewrite (P) as:

(P) minimize $\langle \mu, c \rangle$

$$\text{subject to: } L_1\mu = 0, \quad \mu \in \mathcal{P}_w(\mathbb{K}), \tag{11.89}$$

where $L_1\mu$ is the signed measure in $\mathbb{M}_{w_0}(\mathbb{X}) \subset \mathbb{M}(\mathbb{X})$ defined by (11.42). We also have:

Lemma 11.1 *Let* $\mathcal{C}(\mathbb{X}) \subset C_0(\mathbb{X})$ *be a countable dense subset of* $C_0(\mathbb{X})$*. Then the following are equivalent conditions for* μ *in* $\mathcal{P}_w(\mathbb{K})$*:*

(a) $L_1\mu = 0$.

(b) $\langle L_1\mu, u\rangle = 0 \quad \forall u \in C_0(\mathbb{X})$.

(c) $\langle L_1\mu, u\rangle = 0 \quad \forall u \in \mathcal{C}(\mathbb{X})$.

Proof. The equivalence of (a) and (b) is due to the fact that $(\mathbb{M}(\mathbb{X}), C_0(\mathbb{X}))$ is a dual pair—in fact, $\mathbb{M}(\mathbb{X})$ is the topological dual of $C_0(\mathbb{X})$ [Remark 11.1(b)]. Finally, the implication (b) $\Rightarrow$ (c) is obvious, whereas the converse follows from the denseness of $\mathcal{C}(\mathbb{X})$ in $C_0(\mathbb{X})$. ∎

By (11.89) and Lemma 11.1, we may further rewrite (P) in the equivalent form:

(P) minimize $\langle \mu, c\rangle$

$$\text{subject to: } \langle L_1\mu, u\rangle = 0 \quad \forall u \in \mathcal{C}(X);\ \mu \in \mathcal{P}_w(\mathbb{K}). \tag{11.90}$$

Observe that (11.90) defines an *aggregation* (of constraints) of (P); see Definition 11.3(a). In other words, the constraint $L_1\mu = 0$ in (11.89) is "aggregated" into *countably* many constraints $\langle L_1\mu, u\rangle = 0$ with u in $\mathcal{C}(\mathbb{X})$. We next reaggregate (11.90) into *finitely* many constraints as follows.

Let $\{\mathcal{C}_k\}$ be an increasing sequence of *finite* sets $\mathcal{C}_k \uparrow \mathcal{C}(\mathbb{X})$. For each k, consider the aggregation

$\mathbb{P}(\mathcal{C}_k)$ minimize $\langle \mu, c\rangle$

$$\text{subject to: } \langle L_1\mu, u\rangle = 0 \quad \forall u \in \mathcal{C}_k;\ \mu \in \mathcal{P}_w(\mathbb{K}). \tag{11.91}$$

This linear program has indeed a *finite number of constraints*, namely, the cardinality $|\mathcal{C}_k|$ of $\mathcal{C}_k$. We also have our first approximation result:

Theorem 11.7 *Suppose that Assumption A3 is satisfied. Then*

(a) $\mathbb{P}(\mathcal{C}_k)$ *is solvable for each* $k = 1, 2, \ldots$; *in fact, the aggregation* $\mathbb{P}(W)$ *is solvable for any subset* W *of* $C_0(\mathbb{X})$.

(b) For each $k = 1, 2, \ldots$, *let* μ_k *be an optimal solution for* $\mathbb{P}(\mathcal{C}_k)$, *i.e.*

$$\langle \mu_k, c\rangle = \min \mathbb{P}(\mathcal{C}_k).$$

Then

$$\langle \mu_k, c\rangle \uparrow \min(P) = \rho_{\min}, \tag{11.92}$$

where the equality is due to Theorem 11.3(a).

Furthermore, there is a subsequence $\{\mu_m\}$ *of* $\{\mu_k\}$ *that converges in the weak topology* $\sigma(\mathbb{M}(\mathbb{K}), C_b(\mathbb{K}))$ *to an optimal solution* μ^* *for* (P), *i.e.*

$$\langle \mu_m, v\rangle \to \langle \mu^*, v\rangle \quad \forall v \in C_b(\mathbb{K}); \tag{11.93}$$

in fact, any weak-$\sigma(\mathbb{M}(\mathbb{K}), C_b(\mathbb{K}))$ *accumulation point of* $\{\mu_k\}$ *is an optimal solution for* (P).

11.5.2 Aggregation-relaxation

The *equality* constraint $\langle L_1\mu, u\rangle = 0$ in (11.91) will now be "relaxed" to inequalities of the form $|\langle L_1\mu, u\rangle| \leq \varepsilon$ with $\varepsilon > 0$.

Let $\mathcal{C}_k \uparrow \mathcal{C}(\mathbb{X})$ be as in (11.91), and let $\{\varepsilon_k\}$ be a sequence of numbers $\varepsilon_k \downarrow 0$. For each $k = 1, 2, \ldots$, consider the linear program

$$\mathbb{P}(\mathcal{C}_k, \varepsilon_k) \qquad \text{minimize } \langle \mu, c\rangle$$

$$\text{subject to: } |\langle L_1\mu, u\rangle| \leq \varepsilon_k \quad \forall u \in \mathcal{C}_k;\ \mu \in \mathcal{P}_w(\mathbb{K}). \tag{11.94}$$

Remark 11.7 *If $\varepsilon > 0$ and $I \subset C_0(\mathbb{X})$ is a finite subset of $C_0(\mathbb{X})$, then [by (11.11)] the set*

$$N(I, \varepsilon) := \{\nu \in \mathbb{M}(\mathbb{X}) \,|\, |\langle \nu, u\rangle| \leq \varepsilon \ \ \forall u \in I\}$$

defines a (closed) weak—actually weak—neighborhood of the "origin" (that is, the null measure) in $\mathbb{M}(\mathbb{X})$. In particular, if we take ε and I as ε_k and $\mathcal{C}_k$, respectively, then the constraint (11.94) states that $L_1\mu$ is in the weak* neighborhood $N(\mathcal{C}_k, \varepsilon_k)$, i.e.*

$$L_1\mu \in N(\mathcal{C}_k, \varepsilon_k). \tag{11.95}$$

This provides a natural interpretation of $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$ as an approximation of the original program (P) in the weak topology $\sigma(\mathbb{M}(\mathbb{X}), C_0(\mathbb{X}))$.*

The following result states that Theorem 11.7 remains basically unchanged when $\mathbb{P}(\mathcal{C}_k)$ is replaced by $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$.

Theorem 11.8 *Suppose that Assumption A3 is satisfied. Then*

(a) $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$ is solvable for each $k = 1, 2, \ldots$.

(b) If μ_k is an optimal solution for $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$, i.e.

$$\langle \mu_k, c\rangle = \min \mathbb{P}(\mathcal{C}_k, \varepsilon_k) \ \text{ for } \ k = 1, 2, \ldots,$$

then $\{\mu_k\}$ satisfies the same conclusion of Theorem 11.7(b); in particular,

$$\langle \mu_k, c\rangle \uparrow \min(P) = \rho_{\min}. \tag{11.96}$$

11.5.3 Aggregation-relaxation-inner approximations

The programs $\mathbb{P}(\mathcal{C}_k)$ and $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$ have a *finite number of constraints* and give "nice" approximation results—Theorems 11.7 and 11.8. However, they are still not good enough for our present purpose because the "decision variable" μ lies in the *infinite-dimensional* space $\mathbb{M}_w(\mathbb{K}) \subset \mathbb{M}(\mathbb{K})$. (For the latter spaces to be finite-dimensional we would need the state and action sets, $\mathbb{X}$ and $\mathbb{A}$, to be both finite sets.) Now, to obtain *finite-dimensional* approximations of (P) we will combine $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$ with a suitable sequence of *inner approximations* [see Definition 11.3(b)]. These are based on the following well-known result (for a proof see, for instance, Theorem 4, p. 237, in Billingsley [4], or Theorem 6.3, p. 44, in Parthasarathy [32]).

Proposition 11.3 [Existence of a weakly dense set in $\mathcal{P}(S)$.] *Let S be a separable metric space and $D \subset S$ a countable dense subset of S. Then the family of p.m.'s whose supports are finite subsets of D is dense in $\mathcal{P}(S)$ in the weak topology $\sigma(\mathbb{M}(S), C_b(S))$.*

We will now apply Proposition 11.3 to the space $S := \mathbb{K}$. Let $D \subset \mathbb{K}$ be a *countable* dense subset of $\mathbb{K}$, and let $\{D_n\}$ be an increasing sequence of *finite* sets $D_n \uparrow D$. For each $n = 1, 2, \ldots$, let $\Delta_n := \mathcal{P}(D_n)$ be the family of p.m.'s on D_n; that is, *an element of Δ_n is a convex combination of the Dirac measures concentrated at points of D_n*. Then, as $D_n \uparrow D$, the sets Δ_n for an *increasing sequence* (of sets of p.m.'s) whose limit

$$\Delta := \bigcup_{n=1}^{\infty} \Delta_n \tag{11.97}$$

is dense in $\mathcal{P}(\mathbb{K})$ in the weak topology $\sigma(\mathbb{M}(\mathbb{K}), C_b(\mathbb{K}))$; that is, for each p.m. μ in $\mathcal{P}(\mathbb{K})$, there is a sequence $\{\nu_k\}$ in Δ such that

$$\langle \nu_k, v \rangle \to \langle \mu, v \rangle \quad \forall v \in C_b(\mathbb{K}). \tag{11.98}$$

Let us now consider a linear program as $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$ except that the p.m.'s μ in (11.94) are replaced by p.m.'s in $\Delta_n \cap \mathcal{P}_w(\mathbb{K})$. That is, instead of $\mathbb{P}(\mathcal{C}_k, \varepsilon_k)$ consider the *finite* program

$$\begin{aligned} \mathbb{P}(\mathcal{C}_k, \varepsilon_k, \Delta_n): \qquad & \text{minimize } \langle \mu, c \rangle \\ & \text{subject to: } |\langle L_1 \mu, u \rangle| \le \varepsilon_k \quad \forall u \in \mathcal{C}_k,\ \mu \in \Delta_n \cap \mathcal{P}_w(\mathbb{K}). \end{aligned} \tag{11.99}$$

This is indeed a *finite* linear program because it has a finite number $|\mathcal{C}_k|$ of constraints, and a finite number $|D_n|$ of "decision variables", namely, the coefficients of a measure in $\Delta_n \cap \mathcal{P}_w(\mathbb{K})$.

The corresponding approximation result is as follows.

Theorem 11.9 [Finite approximations for (P).] *If Assumption A3 is satisfied then:*

(a) For each $k = 1, 2, \ldots$, there exists $n(k)$ such that, for all $n \ge n(k)$, the finite linear program $\mathbb{P}(\mathcal{C}_k, \varepsilon_k, \Delta_n)$ is solvable and

$$\min \mathbb{P}(\mathcal{C}_k, \varepsilon_k) \le \min \mathbb{P}(\mathcal{C}_k, \varepsilon_k, \Delta_n). \tag{11.100}$$

(b) Suppose that, in addition, the cost-per-stage function $c(x, a)$ is continuous. Then for each $k = 1, 2, \ldots$ there exists $n^*(k)$ such that

$$\min \mathbb{P}(\mathcal{C}_k, \varepsilon_k, \Delta_n) \le \min(P) + \varepsilon_k \quad \forall n \ge n^*(k); \tag{11.101}$$

hence [by (11.100) and (11.96)]

$$\min \mathbb{P}(\mathcal{C}_k, \varepsilon_k, \Delta_n) \to \min(P) = \rho_{\min} \ \text{ as } \ k \to \infty, \tag{11.102}$$

where of course the limit is taken over values of $n \ge n^*(k)$. Moreover, if μ_{kn} [for $k \ge 1$ and $n \ge n^*(k)$] is an optimal solution for $\mathbb{P}(\mathcal{C}_k, \varepsilon_k, \Delta_n)$,

then every weak accumulation point of $\{\mu_{kn}\}$ *is an optimal solution for* (P).

The approximation results in this section are based on Hernández-Lerma and Lasserre [21]. A similar approach, combining aggregations, relaxations and inner approximations, can be used to approximate general (not necessarily MDP-related) infinite linear programs, as in Hernández-Lerma and Lasserre [20]. These two papers provide many related references.

The approximation schemes in Section 11.5 are somewhat similar in spirit to schemes proposed by Vershik [40] and Vershik and Temel't [41], but with a basic difference. Namely, we use *weak* and *weak** topologies (see Remark 11.7 and Lemma 11.3), whereas Vershik and Temel't use stronger—for instance, normed—topologies. This is a key fact because we only need "reasonable" things, whereas their context would require convergence in the *total variation norm*, which is obviously too restrictive. For instance, for an uncountable metric space, the density result in Proposition 11.3—with finitely supported measures—is, in general, virtually impossible to get in the total variation norm.

Finally, it is worth noting that the approach in this section can be used to approximately compute an i.p.m. for a noncontrolled Markov chain on a LCSM space whose transition kernel satisfies the (weak) Feller condition in Assumption A1(d). The idea would be to introduce an "artificial" MDP with a *singleton* control set $\mathbb{A}$ and with a continuous "cost" function that satisfies the hypothesis of Theorem 11.9.

11.6 CONCLUSION

In this chapter we have developed an LP approach for MDPs with the average cost criterion. As mentioned in the introduction, this LP approach can be adapted to other optimality criteria, and to constrained MDPS with ad hoc modifications left to the reader.

From an approximation point of view, it was shown how finite LP approximations schemes can be designed to approximate the optimal value. Validating such numerical schemes on a significant sample of problems remains to be done. In addition, an important computational issue is to construct control *policies* that are ϵ-optimal, i.e. whose cost is within ϵ of the optimal value. The numerical schemes for the optimal value might provide a valuable tool. Indeed, from the converging sequence of finite LPs, one may construct (incomplete) stationary "policies" defined only at some points of the state space $\mathbb{X}$. Extending such policies to the whole space $\mathbb{X}$ and proving their ϵ-optimality is a topic for further research. Another interesting issue is to compare the LP approach with others, notably those that approximate from the beginning the original problem with a finite or countable state and action model.

Finally, as already noted, such numerical schemes might prove to be a valuable tool to compute invariant probability distributions for (noncontrolled) Markov chains.

References

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

[2] E.J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, Wiley, Chichester, U.K., 1987.

[3] R.B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972.

[4] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[5] N. Bourbaki, *Integration*, Chap. IX. Hermann, Paris, 1969.

[6] H. Brézis, *Analyse Fonctionnelle: Théorie and Applications*, 4th Ed., Masson, Paris, 1993.

[7] A. Bhatt and V. Borkar, "Occupation measures for controlled Markov processes: Characterization and optimality," *Ann. Probab.* **24**, 1531–1562, 1996.

[8] V. Borkar, "A convex analytic approach to Markov decision processes," *Probab. Theory Relat. Fields* **78**, 583–602, 1988.

[9] V. Borkar, "Ergodic control of Markov chains with constraints – the general case," *SIAM J. Control Optimization* **32**, 176–186, 1994.

[10] G.T. De Ghellinck, "Les problèmes de décisions séquentielles," *Cahiers du Centre d'Etudes de Recherche Opérationnelle* **2**, 161–179, 1960.

[11] F. D'Epenoux, "Sur un problème de production et de stockage dans l'aléatoire," *Revue Francaise de Recherche Opérationnelle* **14**, 3–16, 1960.

[12] E.V. Denardo, "On linear programming in a Markov decision problem," *Management Science* **16**, 281–288, 1970.

[13] W.K. Haneveld, *Duality in Stochastic Linear and Dynamic Programming*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, 1986.

[14] W.R. Heilmann, "Solving stochastic dynamic programming problems by linear programming - an annotated bibliography," *Z. Oper. Res. Ser. A* **22**, 43–53, 1978.

[15] W.R. Heilmann, "Solving a general discounted dynamic program by linear programming," *Z. Wahrsch. Gebiete* **48**, 339–346, 1979.

[16] O. Hernández-Lerma and J. González–Hernández, "Infinite linear programming and multichain Markov control processes in uncountable spaces," *SIAM J. Contr. Optim.* **36**, 313–335, 1998.

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

[18] O. Hernández-Lerma and J. González-Hernández, "Constrained Markov control processes in Borel spaces: the discounted case", *Math. Meth. Oper. Res.* **52**, 271–285, 2000.

[19] O. Hernández-Lerma and J.B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[20] O. Hernández-Lerma and J.B. Lasserre, "Approximation schemes for infinite linear programs," *SIAM J. Optim.* **8**, 973–988, 1998.

[21] O. Hernández-Lerma and J.B. Lasserre, 'Linear programming approximations for Markov control processes in metric spaces," *Acta Appl. Math.* **51**, 123–139, 1998.

[22] O. Hernández-Lerma and J.B. Lasserre, *Further Topics on Discrete-Time Markov Control processes*, Springer-Verlag, New York, 1999.

[23] O. Hernández-Lerma and R. Romera, "Pareto optimality in multiobjective Markov control processes ", *Reporte interno #278*, Depto. de Matemáticas, CINVESTAV-IPN, Mexico, (submitted).

[24] K. Hinderer, *Foundations of Non-stationary Dynamic Programming with Discrete-Time Parameter*, Lecture Notes in Oper. Res. and Math. Syst. **33**, Springer-Verlag, Berlin, 1970.

[25] A. Hordijk and L.C.M. Kallenberg, "Linear programming and Markov decision chains," *Management Science* **25**, 352–362, 1979.

[26] A. Hordijk and J.B. Lasserre, "Linear programming formulation of MDPs in countable state space: the multichain case," *Zeitschrift fur Oper. Res.* **40**, 91–108, 1994.

[27] Y. Huang and M. Kurano, "The LP approach in average rewards MDPs with multiple cost constraints: the countable state case ", *J. Infor. Optim. Sci.* **18**, 33–47, 1997.

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

[29] J.B. Lasserre, "Average optimal stationary policies and linear programming in countable space Markov decision processes," *J. Math. Anal. Appl.* **183**, 233–249, 1994.

[30] A.S. Manne, "Linear programming and sequential decisions," *Management Science* **6**, 259-267, 1960.

[31] M.S. Mendiondo and R. Stockbridge, "Approximation of infinite-dimensional linear programming problems which arise in stochastic control," *SIAM J. Contr. Optim.* **36**, 1448-1472, 1998.

[32] K.R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

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

[34] U. Rieder, "Measurable selection theorems for optimization problems," *Manuscripta Math.* **24**, 115–131, 1978.

[35] A.P. Robertson and W. Robertson, *Topological Vector Spaces*, Cambridge University Press, Cambridge, U.K., 1964.

[36] W. Rudin, *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York, 1986.

[37] L. Sennott, "On computing average optimal policies with application to routing to parallel queues ", *Math. Meth. Oper. Res.* **45**, 45–62, 1997.

[38] L. Sennott, "Stochastic dynamic programming and the control of queueing systems ", New York, 1999.

[39] R. Stockbridge, "Time-average control of martingale problems: A linear programming formulation," *Ann. Prob.* **18**, 190–205, 1990.

[40] A.M. Vershik, "Some remarks on the infinite-dimensional problems of linear programming," *Russian Math. Surveys* **29**, 117–124, 1970.

[41] A.M. Vershik and V. Temel't, "Some questions concerning the approximation of the optimal value of infinite-dimensional problems in linear programming," *Siberian Math. J.* **9**, 591-601, 1968.

[42] K. Yamada, "Duality theorem in Markovian decision problems," *J. Math. Anal. Appl.* **50**, 579–595, 1975.

Onésimo Hernández-Lerma
Depto. de Matemáticas
CINVESTAV-IPN
Apdo. Postal 14-740
D.F. 07000, Mexico
ohernand@math.cinvestav.mx

Jean B. Lasserre
LAAS-CNRS
7 Avenue du Colonel Roche
31077 Toulouse Cédex 4
France
lasserre@laas.fr

12 INVARIANT GAMBLING PROBLEMS AND MARKOV DECISION PROCESSES

Lester E. Dubins
Ashok P. Maitra
William D. Sudderth

Abstract: Markov decision problems can be viewed as gambling problems that are invariant under the action of a group or semi-group. It is shown that invariant stationary plans are almost surely adequate for a leavable, measurable, invariant gambling problem with a nonnegative utility function and a finite optimal reward function. This generalizes results about stationary plans for positive Markov decision models as well as measurable gambling problems.

12.1 INTRODUCTION

This paper introduces the notion of a gambling problem that is invariant, in a sense to be specified below, under the action of a group or a semigroup of transformations.

Our primary stimulus has been to understand more fully the relationship of Markov decision processes to gambling theory. It has long been known that these two theories are closely related (cf. Chapter 12 of Dubins and Savage (1965)) and perhaps each contains the other. However, it has not been possible to translate theorems about stationary plans, for example, directly from one theory to the other. It will be explained below how Markov decision processes may be viewed as invariant gambling problems. Subsequent sections will show how stationarity results in gambling theory (Dubins and Sudderth (1979)) are extended to invariant gambling problems and, in particular, to Markov decision problems. Our main interest here is in the reward structures. From a different

point of view Schäl (1989) has also studied the reward structures of gambling and Markov decision theory. A comparative study of the measurable structures of the two theories was made by Blackwell (1976).

A secondary stimulus to us is the fact that there are a number of gambling problems which possess a natural group theoretic structure. Group invariance techniques have found many applications in statistical decision theory (cf. Eaton (1989) and the references therein) and could prove useful in Markov decision theory as well.

We will begin with a review of measurable gambling theory, and then introduce invariance.

12.2 MEASURABLE GAMBLING PROBLEMS

Let F be a *Borel set*, that is, a Borel subset of a complete separable metric space. Let $\mathbb{P}(F)$ be the set of probability measures on the Borel sigma-field of subsets of F. Then $\mathbb{P}(F)$, equipped with its customary weak topology, is again a Borel set. A *gambling house* on F is a subset Γ of $F \times \mathbb{P}(F)$ such that each section $\Gamma(x)$ of Γ at $x \in F$ is nonempty. A *strategy* σ is a sequence $\sigma_0, \sigma_1, \ldots$ such that $\sigma_0 \in \mathbb{P}(F)$, and , for each $n \geq 1$, σ_n is a universally measurable function from F^n into $\mathbb{P}(F)$. A strategy is *available in* Γ *at* x if $\sigma_0 \in \Gamma(x)$ and $\sigma_n(x_1, x_2, \ldots, x_n) \in \Gamma(x_n)$ for every $n \geq 1$ and $x_1, x_2, \ldots, x_n \in F$.

Each strategy σ determines a unique probability measure, also denoted by σ, on the Borel subsets of the *history space* $H = F^N$, where N is the set of positive integers and H is given the product topology. Let $X_1, X_2, \ldots$ be the coordinate process on H; then, under the probability measure σ, X_1 has distribution σ_0 and , for $n \geq 1$, X_{n+1} has conditional distribution $\sigma_n(x_1, x_2, \ldots, x_n)$ given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

A *measurable gambling problem* is a triple (F, Γ, u), where F, the *fortune space*, is a nonempty Borel set, Γ is a gambling house on F which is an analytic subset of $F \times \mathbb{P}(F)$, and $u : F \mapsto \mathbb{R}$, the *utility function*, is upper analytic, which means that $\{x : u(x) > a\}$ is an analytic subset of F for every real a. Such structures with Γ and u both Borel were introduced by Strauch (1967); the extension to analytic Γ and upper analytic u is due to Meyer and Traki (1973).

In the theory of gambling (Dubins and Savage (1965)) there are two natural approaches to a gambling problem. In *leavable* problems, the gambler is allowed to stop playing at any time, whereas in *nonleavable* problems, the gambler is compelled to continue playing forever.

Consider first a leavable problem and define a *stop rule* t to be a universally measurable function from H into $\{0, 1, \ldots\}$ such that whenever $t(h) = k$ and h and h' agree in their first k coordinates, then $t(h') = k$. (In particular, if $t(h) = 0$ for some h, then t is identically 0.) A gambler with initial fortune x selects a strategy σ available at x and a stop rule t. The pair $\pi = (\sigma, t)$ is a *policy* available at x. The expected reward to a gambler who selects the policy π is

$$u(\pi) = \int u(X_t)\, d\sigma,$$

where $X_0 = x$. We will usually assume in this paper that u is nonnegative and we will always assume $u(\pi)$ is well-defined and finite for all available π. The *optimal reward function* for this leavable problem is

$$U(x) = \sup u(\pi),$$

where the supremum is taken over all policies π available at x.

It is also of interest to consider how well the gambler can do when playing time is restricted to at most n days, and we define $U_n(x) = \sup u(\pi)$ over all policies $\pi = (\sigma, t)$ available at x such that $t \le n$ for positive integers n and set $U_0 = u$. The functions U_n can be calculated by backward induction as we will now explain. First define the operator S by

$$(S\phi)(x) = \sup \left\{ \int \phi \, d\gamma \, : \, \gamma \in \Gamma(x) \right\} \tag{12.1}$$

for universally measurable functions ϕ on X such that $\int \phi \, d\gamma$ is well-defined for all available γ. Then

$$U_{n+1} = (SU_n) \vee u \, , \qquad n = 0, 1, \ldots \, , \tag{12.2}$$

where $a \vee b$ is the maximum of a and b. For ϕ upper analytic, $S\phi$ is also (Meyer and Traki (1973); see also Dubins and Sudderth (1979)). Thus every U_n is upper analytic. Furthermore,

$$U = \lim U_n. \tag{12.3}$$

(See Section 2.15 of Dubins and Savage (1965) and also Maitra, Purves, and Sudderth (1990).) Hence, U is upper analytic as well.

The utility of a strategy σ is defined to be

$$u(\sigma) = \limsup_t \int u(X_t) \, d\sigma \, ,$$

where the lim sup is over the directed set of stop rules t. In many cases, for example when u is bounded or when the sequence $u(X_n)$ is monotone increasing or decreasing,

$$u(\sigma) = \int \{\limsup_n u(X_n)\} \, d\sigma. \tag{12.4}$$

(cf. Theorems 4.2.2 and 4.2.7 in Maitra and Sudderth (1996).) The optimal reward function for the nonleavable problem is

$$V(x) = \sup\{u(\sigma) \, : \, \sigma \text{ available at } x\}.$$

The function V can be calculated by a transfinite recursive scheme based on the operator T defined for bounded upper analytic functions u by

$$(Tu)(x) = \sup \left\{ \int u(X_t) \, d\sigma \, : \, \sigma \text{ available at } x, t \ge 1 \right\}.$$

It turns out that $Tu = SU$, where U is the optimal reward function for the leavable problem defined above (Maitra and Sudderth (1996, Theorem 4.8.12)). Now let $V_0 = Tu$ and , for each countable ordinal $\alpha > 0$, let

$$V_\alpha = \begin{cases} T(u \wedge V_{\alpha-1}), & \text{if } \alpha \text{ is a successor,} \\ \inf_{\beta<\alpha} V_\beta, & \text{if } \alpha \text{ is a limit ordinal.} \end{cases}$$

(Here $a \wedge b$ is the minimum of a and b.) Then

$$V = \inf_\alpha V_\alpha \ ,$$

where the infimum is over all countable ordinals α, and V is also upper analytic (Dubins, Maitra, Purves, and Sudderth (1989), Theorem 2.1 and Lemma 7.3).

A gambling house Γ is said to be *leavable* if the point-mass $\delta(x) \in \Gamma(x)$ for every $x \in X$. If Γ is leavable, then U and V are the same (Corollary 3.2.2, Dubins and Savage (1965)) and this holds even for unbounded u. So, in particular, V is upper analytic if Γ is leavable because U is upper analytic, as we already observed.

12.3 INVARIANT GAMBLING PROBLEMS

Let (F, Γ, u) be a measurable gambling problem and suppose that G is a topological group or semigroup with identity. (A *semigroup* G is a set of elements $\{a, b, c, \dots\}$ equipped with a binary operation $(a, b) \mapsto ab$ such that $(ab)c = a(bc)$ for all a, b, c in G. An *identity* for G is an element e such that $ea = ae = a$ for all a in G. A *group* is a semigroup G with an identity e such that, for every element a in G, there is an element a^{-1} such that $aa^{-1} = a^{-1}a = e$. A *topological group* is a group with a topology such that the group operation and the inverse function are continuous; a *topological semigroup* is a semigroup with a topology such that the semigroup operation is continuous.) We assume that G acts on the fortune space F. This means that there is a Borel mapping $\alpha : G \times F \to F$, which we write $\alpha(g, x) = gx$, such that $ex = x$ for e the identity element of G and $g_1(g_2 x) = (g_1 g_2)x$. For a group the mapping $x \to gx$ is necessarily one-to-one for every $g \in G$. We assume that these mappings are one-to-one in the case when G is only a semi-group. We also assume that G is a Polish space; that is, the topology on G has a countable base and is induced by a complete metric. The gambling problem will be called *invariant* (under G) if Conditions 12.1 and 12.2 below are satisfied.

To formulate the first condition, we associate to each probability measure $\gamma \in \mathbb{P}(F)$ and each $g \in G$ a measure $g\gamma \in \mathbb{P}(F)$ defined by

$$g\gamma(B) = \gamma(g^{-1}(B))$$

for all Borel subsets B of F. An equivalent condition is that

$$\int \phi(x)\,(g\gamma)(dx) = \int \phi(gx)\,\gamma(dx)$$

for bounded measurable functions ϕ on F. For a set $\Delta \subseteq \mathbb{P}(F)$ and $g \in G$, let

$$g\Delta = \{g\gamma : \gamma \in \Delta\}.$$

Condition 12.1 *For all $x \in F$ and $g \in G$, $g\Gamma(x) = \Gamma(gx)$.*

To state the second condition, let A be the group of positive, affine transformations of the real line. That is, A consists of all mappings $r \to ar + b$ for some $a > 0$ and $b \in \mathbb{R}$. Give A the topology it inherits when considered as a subset of the plane.

Condition 12.2 *There is a Borel measurable mapping $w : G \to A$ such that $u(gx) = w(g)(u(x))$ for every $x \in F$ and $g \in G$.*

This condition implies that the problem of maximizing $\int u(x)\,\gamma(dx)$ over a set of probability measures γ is equivalent to that of maximizing $\int u(gx)\,\gamma(dx) = w(g)(\int u(x)\,\gamma(dx))$ over the same set. (According to Savage's theory of utility (see Theorems 5.3.2 and 5.3.3 in Savage (1954)), if u is a utility function, then $\tilde{u}$ is also if and only if $\tilde{u}$ is a positive affine transformation of u. Thus Condition 12.2 says that for every $g \in G$, the function $u_g(x) = u(gx)$ is a utility function in the sense of Savage if u is.)

Lemma 12.1 *If the function u is not constant, then w is a homomorphism from G into A.*

Proof. We must show that $w(g_1g_2) = w(g_1)w(g_2)$ for all $g_1, g_2 \in G$. But for each $x \in F$, $w(g_1g_2)(u(x)) = u(g_1g_2x) = w(g_1)(w(g_2)(u(x)))$. Now use the fact that if two affine transformations agree on two points, then they agree everywhere. ∎

As the case of a constant utility function u is uninteresting, we will assume from now on that w is a homomorphism.

Two notions of an invariant function will be used below. Let ψ be a function with domain F and suppose that G acts on the range of ψ. Then ψ will be called simply *invariant* if $\psi(gx) = g\psi(x)$ for all $g \in G, x \in F$. Thus Condition 12.1 says that the house Γ is invariant. To define the other notion, let w be the mapping given by Condition 12.2 and let ϕ be a mapping from F to the real numbers. Call ϕ *w-invariant* (under G) if $\phi(gx) = w(g)(\phi(x))$ for all $x \in F$ and $g \in G$. So Condition 12.2 says that the utility function u is w-invariant.

Lemma 12.2 *If ϕ is w-invariant and $S\phi$ is well-defined, then $S\phi$ is w-invariant also.*

Proof. For $x \in F$ and $g \in G$,

$$\begin{aligned}
(S\phi)(gx) &= \sup\left\{\int \phi(x)\,\gamma(dx) \,:\, \gamma \in \Gamma(gx)\right\} \\
&= \sup\left\{\int \phi(x)\,(g\gamma)(dx) \,:\, \gamma \in \Gamma(x)\right\} \\
&= \sup\left\{\int \phi(gx)\,\gamma(dx) \,:\, \gamma \in \Gamma(x)\right\} \\
&= \sup\left\{\int w(g)(\phi(x))\,\gamma(dx) \,:\, \gamma \in \Gamma(x)\right\} \\
&= w(g)(\sup\left\{\int \phi(x)\,\gamma(dx) \,:\, \gamma \in \Gamma(x)\right\}) \\
&= w(g)(S\phi(x)).
\end{aligned}$$

■

Corollary 12.1 *The functions $U_n, n \geq 1$, and U are w-invariant, as is the function V when u is bounded.*

Proof. It is easy to check that that suprema, infima, and pointwise limits of w-invariant functions are again w-invariant. Hence the corollary is a consequence of the lemma and the recursive schemes for calculating U_n, U, and V sketched in the previous section. ■

Example 12.1 *Positive Markov Decision Models. Suppose that S is a nonempty Borel set, $\mathbb{R}$ is the set of real numbers, and that $F = S \times \mathbb{R}$. Thus a fortune x is a pair (s, c) whose first coordinate s is regarded as the* state *and whose second coordinate c as the* cash *accumulated up to the present time. The additive group G_0 of real numbers acts on F thus*

$$g(s, c) = (s, c + g)\,, \qquad g \in G_0.$$

Assume that Condition 12.1 holds. Then, in particular,

$$\Gamma(s, c) = c\Gamma(s, 0)\,, \qquad c \in \mathbb{R}.$$

Thus, if $\gamma \in \Gamma(s, 0)$ and the random pair (s_1, r_1) has distribution γ, then $c\gamma \in \Gamma(s, c)$ and $(s_1, r_1 + c)$ has distribution $c\gamma$. Let the utility function be $u(s, c) = c$. Condition 12.2 is immediate and, for each g, the affine transformation $w(g)$ is translation by g.

Suppose that a gambler has initial fortune $x = (s, 0)$. The gambler's successive fortunes can be written as $x_1 = (s_1, r_1), x_2 = (s_2, r_1 + r_2), \ldots, x_n = (s_n, r_1 + r_2 + \cdots + r_n), \ldots$, and $u(x_n) = r_1 + r_2 + \cdots + r_n$. The utility of a strategy σ is

$$u(\sigma) = \limsup_t \int \left(\sum_{i=1}^{t} r_i\right) d\sigma.$$

The positive model corresponds to the special case where, for every $s \in S$ *and* $\gamma \in \Gamma(s,0)$, $\gamma\{(s_1, r_1) : r_1 \geq 0\} = 1$. *In this case, the gambler's fortune is almost sure to increase in utility at every stage and the utility of* σ *becomes*

$$u(\sigma) = \int \left(\sum_{i=1}^{\infty} r_i \right) d\sigma.$$

It is natural in the case of a positive model to restrict the fortune space to be $S \times [0, \infty)$ *and to take* G *to be the semigroup of nonnegative real numbers under addition.*

To see that these models include positive MDP's, consider an MDP with state space S, *action space* A, *daily reward function* r, *and law of motion* q *with* $r \geq 0$. *Define the gambling problem* F, Γ, u *by setting* $F = S \times [0, \infty)$,

$$\Gamma(s,c) = \{q(\cdot|s,a) \times \delta(r(s,a) + c) \,|\, a \in A\}$$

$u(s,c) = c$ *for all* $(s,c) \in F$. *The semigroup* G *of nonnegative reals acts on* F *as above* ($g(s,c) = (s, c+g)$), *and Conditions 12.1 and 12.2 are satisfied. For each* $s \in S$, *the optimal reward function* V *for the gambling problem has the same value at the fortune* $(s,0)$ *as does the optimal reward function for MDP at the state* s.

Negative models can be formulated by analogy with Example 12.1. However, the situation is more complicated in the case of discounted models and we are indebted to David Heath for the crucial idea in the formulation below.

Example 12.2 *Discounted Markov Decision Models. Let* $0 < \beta < 1$ *be the discount factor and let* $F = S \times \mathbb{R} \times \mathbb{N}$, *where* S *is a nonempty Borel set,* $\mathbb{R}$ *the set of real numbers, and* $\mathbb{N}$ *the set of nonnegative integers. A fortune* x *is now a triple* (s,c,n), *where the new coordinate* n *represents the number of days of play. A gamble* $\gamma \in \Gamma(x)$ *determines the distribution of the next fortune* $x_1 = (s_1, \beta^{-1}c + r, n+1)$ *where* s_1 *is the new state and* r *is the reward earned at the current stage of play. The utility function is* $u(x) = u(s,c,n) = \beta^n c$.

To get the idea, consider an initial fortune $x = (s,c,0)$. *The successive fortunes of a gambler can be written as* $x_1 = (s_1, \beta^{-1}c + r, 1), \ldots, x_n = (s_n, \beta^{-n}c + \beta^{-(n-1)}r_1 + \cdots + r_n, n), \ldots$. *The utility of* x_n *is* $u(x_n) = c + \beta r_1 + \cdots + \beta^n r_n$.

Suppose that the daily reward is bounded by a constant b *in the sense that, for every* $s \in S$ *and* $\gamma \in \Gamma(s,0,0)$, $\gamma\{(s_1, r_1, 1) : |r_1| \leq b\} = 1$. *Then the utility of a strategy* σ *available at* $(s,0,0)$ *can be written*

$$u(\sigma) = \int \left(\sum_{i=1}^{\infty} \beta^i r_i \right) d\sigma.$$

The additive group of reals $\mathbb{R}$ *and the additive semigroup* $\mathbb{N}$ *both act on* F *as follows:*

$$g(s,c,n) = (s, c + \beta^{-n} g, n), \qquad g \in \mathbb{R}$$

$$m(s,c,n) = (s, c, n+m), \qquad m \in \mathbb{N}.$$

Condition 12.2 holds with $w(g)(y) = y+g$ *and* $w(m)(y) = \beta^m y$. *Condition 12.1 is assumed to hold for every* g *and* m, *or, equivalently, for the semigroup* G *of transformations on* F *that they generate.*

These models subsume the traditional discounted MDP's as can be seen by imitating the argument for positive MDP's at the end of Example 12.1.

Here is a simple class of examples on the real line.

Example 12.3 *Proportional Houses. Let* $F = [0, \infty)$ *and let* G *be the group of strictly positive real numbers under multiplication. The action of* G *on* F *is just the usual multiplication. Condition 12.1 now says that a gambler's opportunities are proportional to the size of his fortune. In particular,* $\Gamma(x) = x\Gamma(1)$ *for* $x \geq 0$. *(A famous special case is the red-and-black casino where* $\Gamma(x) = \{w\delta(x+s) + (1-w)\delta(x-s) : 0 \leq s \leq x\}$.*) A utility function that satisfies Condition 12.2 is* $u(x) = \log x$. *Now* $u(gx) = \log g + \log x$ *so that the affine function* $w(g)$ *is translation by* $\log g$. *By Corollary 12.1, the optimal* n*-day return is* w*-invariant and, hence,* $U_n(x) = U_n(x \cdot 1) = w(x)(U_n(1)) = U_n(1) + \log x$. *In fact, it is easy to show, by backward induction, that* $U_n(x) = nU_1(1) + \log x$. *It therefore follows from (1.3) that either* $U(x) = \log x$ *or* $U(x) = \infty$ *according to whether* $U_1(1) = 0$ *or* $U_1(1) > 0$.

12.4 INVARIANT SELECTORS

A Γ-*selector* is a function $\gamma : F \to \mathbb{P}(F)$ such that $\gamma(x) \in \Gamma(x)$ for every $x \in F$. The von Neumann selection theorem guarantees the existence of a universally measurable Γ-selector γ. Such a selector determines a *stationary* family of strategies γ^∞ which uses the gamble $\gamma(x)$ whenever the current fortune is x. A selector γ is *invariant* if $\gamma(gx) = g\gamma(x)$ for all $x \in F$, $g \in G$ and, in this case, the corresponding stationary family is also called *invariant.* In the next section, we address the question of when invariant stationary families are adequate. This section is devoted to the preliminary problem of the construction of measurable, invariant selectors. For brevity, we often write "u.m." for "universally measurable" below.

Suppose now that γ is a u.m. invariant rule in the special case of Example 12.1 that corresponds to a positive MDP. Then invariance ensures that γ picks out the same distribution of the next day's state at $(s, 0)$ as it does at (s, c) for any $c > 0$. This gives rise directly to a u.m. function from S to A in the MDP. Thus, an invariant stationary family of strategies γ^∞ in the gambling problem defines a stationary plan in the MDP with exactly the same return when the gambling problem starts in $(s, 0)$ and the MDP starts in s.

Let us now return to the general framework. Assume first that G is a group. (We will consider the case of a semigroup at the end of the section.) Define R to be the equivalence relation

$$R = \{(x, y) \in F \times F : (\exists g \in G)(gx = y)\}.$$

The *orbit* of an $x \in F$ is the equivalence class containing x and is denoted $[x]$. An *orbit selector* is a function $\phi : F \to F$ such that (i) $\phi(x) \in [x]$ for every x, and (ii) $\phi(x) = \phi(y)$ whenever xRy.

For the construction of a measurable, invariant Γ-selector, we will need a measurable orbit selector. The following condition guarantees its existence.

Condition 12.3 *There is a sequence $\{E_n\}$ of Borel subsets of F such that $xRy \leftrightarrow (\forall n)(x \in E_n \leftrightarrow y \in E_n)$.*

The equivalence relation R is said to be *smooth* when Condition 12.3 holds.

Lemma 12.3 *If R is smooth, then R admits a Borel measurable orbit selector.*

Proof. By Theorem 5.2.1 in Becker and Kechris (1996), there is a Polish topology on F with the same Borel sets as the original topology and such that the action $(g, x) \to gx$ is continuous. So we can assume without loss of generality that the action is continuous. The existence of a Borel orbit selector then follows from a result of Burgess (1979), an exposition of which can be found in Srivastava (1998), Theorem 5.6.1. ∎

It can be shown that if R is Borel and admits even a universally measurable orbit selector, then R is smooth (Harrington et al, 1990). Thus Condition 12.3 is necessary as well as sufficient for the existence of a Borel orbit selector.

One additional condition is needed. For each $x \in F$, the *stabilizer subgroup* of x is defined to be $G_x = \{g \in G : gx = x\}$.

Condition 12.4 *For every $x \in F$, the stabilizer subgroup G_x is either the singleton $\{e\}$ or the whole group G.*

Notice that this condition is satisfied by the proportional houses of Example 12.3. Indeed, for these houses, $G_0 = G$ and $G_x = \{1\}$, for $x \neq 0$.

Lemma 12.4 *Under Condition 12.4, there is a unique Borel function $f : R \to G$ such that, for all $x, y \in F$, (i)$f(x, x) = e$ and (ii)$f(x, y)x = y$ whenever $y \in [x]$.*

Proof. If $G_x = \{e\}$ and $y \in [x]$, then there is a unique $g \in G$ such that $gx = y$ and we take $f(x, y) = g$. If $G_x = G$, then $[x]$ is a singleton and we take $f(x, x) = e$. Then f is Borel because its graph is

$$\{(x, y, g) \in R \times G : gx = y, x \neq y\} \cup \{(x, x, e) \in R \times G : x \in F\},$$

a Borel subset of $R \times G$ ∎

Here is the basic lemma on the construction of invariant Γ-selectors.

Lemma 12.5 *Suppose that G is a group and that Conditions 12.3 and 12.4 hold. Let η be a u.m. Γ-selector such that $\eta(x) = \delta(x)$ if $G_x = G$, let ϕ be a Borel measurable orbit selector, and let f be the function of Lemma 12.4. Then the mapping $\gamma : F \to \mathbb{P}(F)$ defined by*

$$\gamma(x) = f(\phi(x), x)\eta(\phi(x)), \qquad x \in F$$

is a u.m. invariant Γ-selector.

Proof. Clearly γ is a u.m. Γ-selector. To see that γ is invariant, consider first an $x \in F$ for which $G_x = \{e\}$. Then for each $g \in G$,

$$\begin{aligned}\gamma(gx) &= f(\phi(gx), gx)\eta(\phi(gx)) \\ &= f(\phi(x), gx)\eta(\phi(x)) \\ &= gf(\phi(x), x)\eta(\phi(x)) \\ &= g\gamma(x).\end{aligned}$$

Now consider an $x \in F$ such that $G_x = G$. Then, for $g \in G$, $gx = x = \phi(x)$, and

$$\gamma(gx) = f(x, x)\eta(x) = e\delta(x) = \delta(gx) = g\delta(x) = g\gamma(x).$$

■

If the house Γ is leavable, it is not difficult to see that there is a u.m. selector η as hypothesized in Lemma 12.5. Start with any u.m. Γ-selector λ and set $\eta(x) = \lambda(x)$ if $G_x = \{e\}$ and $\eta(x) = \delta(x)$ if $G_x = G$. This η works because the set $A = \{x : G_x = G\}$ is Borel. To see that A is Borel, assume that the mapping $(g, x) \mapsto gx$ is continuous (as in the proof of Lemma 12.3, let C be a countable dense subset of G, and observe that $A = \{x : gx = x \text{ for all } g \in C\}$.

Suppose now that G is a semigroup with identity, but not necessarily a group. Call a subset C of F *closed under* G if (i) $x \in C$ and $g \in G$ imply that $gx \in C$ and (ii) $g \in G$ and $gx \in C$ imply that $x \in C$. The *orbit* $[x]$ of $x \in F$ is defined to be the smallest set closed under G and containing x. It is easy to check that

$$R = \{(x, y) \in F \times F : y \in [x]\}$$

is an equivalence relation on F which is the same as the one previously defined when G is a group. We will not attempt here to develop a general theory for semigroups, but will instead limit ourselves to a special class of invariant problems which includes dynamic programming models and for which invariant selectors are readily available.

The semigroup G is called *special* if (i) there exists a Borel measurable orbit selector ϕ, (ii) for every $x \in F$ and $y \in [x]$, there is a unique $g = f(x, y)$ such that $f(x, y)\phi(x) = y$, and (iii) the function $f : R \to G$ is Borel measurable. Under Conditions 12.3 and 12.4 with $G_x = \{e\}$ for all x, a group G is a special semigroup.

Lemma 12.6 *If G is special and η is a u.m. Γ-selector, then*

$$\gamma(x) = f(\phi(x), x)\eta(\phi(x)), \qquad x \in F$$

is a u.m. invariant Γ-selector.

Proof. The proof is the same as for the first half of Lemma 12.5. ■

For Markov decision models as in Example 12.1, the orbit of any fortune $x = (s, c)$ is the set of all fortunes with the same first coordinate and we can take $\phi(x) = (s, 0)$. Then $f(\phi(x), x) = c$ and $\gamma(x) = c\eta(\phi(x))$.

12.5 INVARIANT STATIONARY FAMILIES OF STRATEGIES

Dubins and Savage (1965) first posed the problem as to the existence of good stationary families. They obtained a positive result for F finite and Γ leavable which was extended to the countable case by Ornstein (1969). A number of authors (cf. Barbosa-Dantas (1966), Dellacherie and Meyer (1983), Dubins and Sudderth (1979), Frid (1976), Schäl and Sudderth (1987), and Sudderth (1969)) used the techniques of Ornstein to get results about good stationary families in Borel measurable settings. Perhaps all or most of these results can be extended to give results about good invariant stationary families for invariant problems. We present two such extensions in this section and mention some questions that remain open.

In the gambling theoretic formulations of Markov Decision Models presented in Examples 12.1 and 12.2 of Section 3, invariant stationary families correspond to the usual stationary plans studied in Markov Decision Theory. Thus theorems about the existence of good invariant stationary families may provide common generalizations of results in gambling and Markov Decision Theory. In addition, such theorems apply in situations like Example 12.3, where the problem is invariant under the multiplicative group of real numbers, to say that there exist good stationary families that are invariant under the same group.

Assume that the gambling problem (F, Γ, u) is invariant under a special semi-group G. Assume also that Γ is leavable so that, in particular, the two optimal reward functions U and V are the same.

The assumption that Γ is leavable is necessary for the existence of good stationary families except in the case when F is finite and every set of gambles $\Gamma(x)$, $x \in F$, is finite (Dubins and Savage (1965), see also Maitra and Sudderth (1996)). This assumption can be made without loss of generality for positive Markov decision models because, with positive daily rewards, a player gains nothing by terminating the game. This is not the case for negative models where, as is well-known, good stationary plans need not exist even when the state space is finite (Strauch (1966)).

Recall that to each $g \in G$ is associated an affine mapping $w(g)$ which we can write as

$$w(g)(r) = a_g r + b_g, \qquad r \in \mathbb{R},$$

for some $a_g > 0, b_g \in \mathbb{R}$. We say that the action of G is *additive* if $a_g = 1$ for every $g \in G$ and we call the action *multiplicative* if $b_g = 0$ for every $g \in G$.

Here is an extension of Theorem 2.3 of Sudderth (1969) to the invariant case.

Theorem 12.1 *Assume that the gambling problem (F, Γ, u) is leavable and invariant under a special semigroup G. Suppose also that the function $U - u$ is bounded and that the action of G is additive. Then, for each $\epsilon > 0$ and probability measure $\alpha \in \mathbb{P}(F)$, there is a u.m. invariant Γ-selector γ such that*

$$\alpha\{x \in F : u(\gamma^\infty(x)) \geq U(x) - \epsilon\} = 1.$$

Here is a similar extension of Proposition 7.1 of Dubins and Sudderth (1979).

Theorem 12.2 *Assume that the gambling problem (F, Γ, u) is leavable and invariant under a special semigroup G. If $u \geq 0$ and the action of G is multiplicative, then, given $0 < \epsilon < 1$, there is a u.m. invariant Γ-selector γ such that*

$$\alpha\{x \in F : u(\gamma^{\infty}(x)) \geq (1-\epsilon)U(x)\} = 1.$$

We will explain in the next section how the proof of Dubins and Sudderth (1979) can be modified for Theorem 12.2. A similar modification of the proof in Sudderth (1969) works for Theorem 12.1.

Theorem 12.1, in the special case of a positive Markov decision model, corresponds to a result of Barbosa-Dantas (1966). However, neither Theorem 12.1 nor Theorem 12.2 includes the result of Frid (1976) for positive models with unbounded optimal reward functions. (Theorem 12.2 does not apply because the action of the semi-group G in Example 12.1 is additive rather than multiplicative.) It would be interesting to have an extension of Frid's result to a more general invariant setting.

Dubins and Savage (1965) showed that an optimal stationary family of strategies exists for any gambling problem with F finite and $\Gamma(x)$ finite for all x. The analogous result holds for Markov decision models with finite state space and finite action sets. A common generalization would be that there is an optimal invariant stationary family for invariant gambling problems with a finite set of orbits and finite $\Gamma(x)\,, x \in F$. We do not know whether this is true.

12.6 THE PROOF OF THEOREM 12.2

For a proof of Theorem 12.2, we will show how to make the necessary changes in the arguments given for Proposition 7.1 in Dubins and Sudderth (1979). We will make reference to results in that paper by adding an asterisk. For example, Proposition 7.1* will denote Proposition 7.1 of Dubins and Sudderth (1979).

Throughout this section we fix a Borel measurable orbit selector ϕ as constructed in Section 4. Also, for $\gamma \in \mathbb{P}(F)$, and ψ a γ-integrable function, we will often write $\gamma\psi$ for the integral $\int \psi \, d\gamma$.

Lemma 12.7 *(cf. Lemma 6.4*) For each $\epsilon > 0$, there is a u.m. invariant Γ-selector γ such that*

$$\gamma(x)u \geq (1-\epsilon)U_1(x) \text{ for all } x \tag{12.5}$$

and

$$\gamma(x) = \delta(x) \Rightarrow u(x) = U_1(x). \tag{12.6}$$

Proof. By Lemma 6.4*, there is a u.m. Γ-selector η that satisfies (1.5) and (1.6), but η need not be invariant. So define

$$\gamma(x) = f(\phi(x), x)\eta(\phi(x))$$

as in Lemma 12.6. Then γ is u.m. and invariant by those lemmas.

To verify (1.5), let $y = \phi(x)$ and $g = f(y, x)$ so that $x = gy$. Then

$$\gamma(x)u = \gamma(gy)u = (g\gamma(y))u = \int u(gz)\,\gamma(y)(dz) = w(g)(\int u(z)\,\gamma(y)(dz))$$
$$\geq w(g)((1-\epsilon)U_1(y)) = (1-\epsilon)w(g)(U_1(y)) = (1-\epsilon)U_1(gy) = (1-\epsilon)U_1(x).$$

The second equality is by the invariance of γ and the next to last is by the w-invariance of U_1 as in Corollary 12.1.

For (1.6), suppose $\gamma(x) = \delta(x)$, where $x = gy$ as above. Then

$$g\delta(y) = \delta(gy) = \gamma(gy) = g\gamma(y).$$

Hence, $\delta(y) = \gamma(y)$ because the mapping $x \to gx$ is one-to-one by assumption. But $y = \phi(x)$ and $\gamma(y) = \eta(y)$. Now η satisfies condition (1.6). So we have $u(y) = U_1(y)$ and

$$u(x) = w(g)(u(y)) = w(g)(U_1(y)) = U_1(x).$$

■

Define the set $Y = \phi(X)$. Then Y is Borel because $Y = \{x \in F : \phi(x) = x\}$. Also Y intersects each orbit $[x]$ in the singleton $\{\phi(x)\}$.

Corollary 12.2 *(cf. Corollary 6.1*) For each $\epsilon > 0$ and $\alpha \in \mathbb{P}(F)$, there is an invariant Borel Γ-selector γ such that*

$$\gamma(x)u \geq (1-\epsilon)U_1(x) \tag{12.7}$$

for α-almost every x.

Proof. Use Lemma 12.7 to get an invariant u.m. Γ-selector $\bar{\gamma}$ such that (1.5) holds when γ is replaced by $\bar{\gamma}$. Define α' to be the probability measure $\alpha\phi^{-1}$ so that, in particular, $\alpha'(Y) = \alpha(\phi^{-1}(Y)) = 1$. Because $\bar{\gamma}$ is u.m. and Γ is leavable, there is a Borel Γ-selector γ' such that $\alpha'\{y \in Y : \gamma'(y) = \bar{\gamma}(y)\} = 1$. Define

$$\gamma(x) = f(\phi(x), x)\gamma'(\phi(x)), \qquad x \in F.$$

Then γ is Borel measurable because f, ϕ, and γ' are. Also γ is an invariant Γ-selector by Lemma 12.6. Note that $\gamma'(\phi(x)) = \bar{\gamma}(\phi(x)) \Rightarrow \gamma(x) = \bar{\gamma}(x)$ as $\bar{\gamma}$ is invariant. Hence,

$$\begin{aligned}\alpha(\{x \in F : \gamma(x) = \bar{\gamma}(x)\}) &\geq \alpha(\{x \in F : \gamma'(\phi(x)) = \bar{\gamma}(\phi(x))\} \\ &= \alpha'(\{y \in Y : \gamma'(y) = \bar{\gamma}(y)\} \\ &= 1.\end{aligned}$$

Thus (1.7) holds for γ α-almost surely in x since it does for $\bar{\gamma}$.

■

The next result is that a gambler can get uniformly close (in the multiplicative sense) to the optimal n-day return U_n using an invariant stationary family of strategies. The same is not true for the optimal reward function U

even if the gambler uses a stationary family that is not invariant (Blackwell and Ramakrishnan, 1988).

Here is a technical remark that we need for the next few results.

Remark 12.1 *There is a technical difficulty that arises in the application of results from Dubins and Sudderth (1979) to the present context. Namely, in that paper gambles are assumed to be defined as finitely additive integrals on the set of all nonnegative, extended-real-valued functions with domain the state space F. Non-measurable strategies and stop rules are allowed and the utility of a strategy σ, say $\tilde{u}(\sigma)$, is defined to be the lim sup of $\int u(X_t)\, d\sigma$ taken over the directed set of* all *stop rules, whereas we have defined $u(\sigma)$ to be the lim sup taken over the universally measurable stop rules. As is shown in an appendix,*

$$\tilde{u}(\sigma) \leq u(\sigma) \tag{12.8}$$

for σ measurable and u nonnegative, upper analytic. Thus, though the results from Dubins and Sudderth (1979) that we use have been proved for $\tilde{u}(\sigma)$, a fortiori, they hold for $u(\sigma)$.

Lemma 12.8 *(cf. Proposition 6.1*) For $n \geq 1$ and $\epsilon > 0$, there is a u.m. invariant Γ-selector η such that*

$$u(\eta^\infty(x)) \geq (1-\epsilon)U_n(x)$$

for all $x \in F$.

Proof. Choose β so that $0 < \beta < 1$ and $\beta^n > 1 - \epsilon$. By Lemma 12.7, we can find u.m. invariant Γ-selectors $\gamma_k, k = 1, \ldots, n$ such that

$$\gamma_k(x)U_{k-1} \geq \beta^{1/2}U_k(x), \qquad x \in F.$$

Let $k(x)$ be the least natural number $k \leq n$ such that

$$\beta^k U_k(x) = \max_{0 \leq j \leq n} \beta^j U_j(x).$$

Since the U_k's are w-invariant and u.m., it follows that $k(x)$ is u.m. and invariant. Consequently, η defined by

$$\eta(x) = \begin{cases} \gamma_{k(x)}(x), & k(x) = 1, \ldots, n \\ \delta(x), & k(x) = 0 \end{cases}$$

is a u.m. invariant Γ-selector. Proposition 5.1*, together with (1.8) now yield the desired result. ∎

To overcome certain measurability difficulties, Dubins and Sudderth found it useful to replace the original gambling problem in their paper by a simpler Borel problem. The next two lemmas enable us to do the same.

Lemma 12.9 *Given $u \geq 0$, u.m., and w-invariant, and $\alpha \in \mathbb{P}(F)$, there is a Borel w-invariant function v on F such that $0 \leq v \leq u$ and $v = u$ α-almost surely.*

Proof. As in the proof of Corollary 12.2, let $\alpha' = \alpha\phi^{-1}$ so that $\alpha'(Y) = 1$. Choose a Borel function $v' : Y \to [0, \infty)$ such that $0 \le v' \le u$ on Y and $\alpha'(\{y \in Y : v'(y) = u(y)\}) = 1$. Recall from Section 4 that each $x \in F$ can be written as $x = f(\phi(x), x)\phi(x)$ for a unique group element $f(\phi(x), x)$ and set

$$v(x) = w(f(\phi(x), x))(v'(\phi(x)), \qquad x \in F.$$

Then v is Borel measurable because all of the functions appearing in its definition are Borel. To check that v is w-invariant, let $x \in F, g \in G$ and calculate:

$$\begin{aligned} v(gx) &= w(f(\phi(gx), gx))(v'(\phi(gx))) \\ &= w(f(\phi(x), gx))(v'(\phi(x))) \\ &= w(gf(\phi(x), x))(v'(\phi(x))) \\ &= w(g)(w(f(\phi(x), x))(v'(\phi(x)))) \\ &= w(g)(v(x)). \end{aligned}$$

Here the definition of ϕ is used for the second equality and Lemma 12.1 for the next to last equality.

Next note that $v'(\phi(x)) = u(\phi(x)) \Rightarrow v(x) = u(x)$, since u is w-invariant. Now $\alpha'(\{x \in F : v'(\phi(x)) = u(\phi(x))\}) = 1$. So it follows that $v = u$ α-almost surely. Finally, $v(x) = w(f(\phi(x), x))(v'(\phi(x)) \le w(f(\phi(x), x))(u(\phi(x))) = u(x)$. ∎

A leavable house Γ' is *(Borel) countably parameterized* if there exist Borel mappings $\gamma_1, \gamma_2, \dots$ from F to $\mathbb{P}(F)$ such that $\Gamma'(x) = \{\gamma_1(x), \gamma_2(x), \dots\}$ for every $x \in F$. The house Γ' is clearly invariant if each of the functions γ_k is invariant.

Lemma 12.10 *(cf. Lemma 7.1*) For each $\alpha \in \mathbb{P}(F)$, there is an invariant, countably parameterized sub house Γ' of Γ and a nonnegative Borel function u' on F such that*

(i) $u' \le u$ on F, and

(ii) $U' \ge U$ α-almost surely,

where U' is the optimal reward function for (F, Γ', u').

Proof. Repeat the proof of Lemma 7.1* using Corollary 12.2 instead of Corollary 6.1* so that the Γ-selectors γ_k turn out to be both invariant and Borel measurable. To get the function u', use Lemma 12.9. ∎

Consider now a leavable, Borel, *stop-or-go house* Σ, which means that, for some Borel mapping $\eta : F \mapsto \mathbb{P}(F)$ and all x, $\Sigma(x) = \{\eta(x), \delta(x)\}$. Assume also that η is invariant and let W be the optimal reward function for the invariant gambling problem (F, Σ, u).

Lemma 12.11 *The stationary family γ^∞, where*

$$\gamma(x) = \begin{cases} \eta(x), & u(x) < W(x), \\ \delta(x), & u(x) = W(x), \end{cases}$$

is Borel measurable, invariant, and optimal for (F, Σ, u). *Moreover,* $u(\gamma^\infty(\cdot)) = \tilde{u}(\gamma^\infty(\cdot)) = W(\cdot)$ *is Borel measurable and w-invariant.*

Proof. By Lemma 7.2*, W is Borel and, by Corollary 12.1, W is w-invariant. Also, by Corollary 4.1*, $\tilde{u}(\gamma^\infty(\cdot)) = W(\cdot)$. But, by (1.8), $W(\cdot) \geq u(\gamma^\infty(\cdot)) \geq \tilde{u}(\gamma^\infty(\cdot))$. ∎

In view of Lemma 12.10 it now suffices to prove Theorem 12.2 under the additional assumptions that Γ is countably parameterized and u is Borel. These assumptions will be in force for the remainder of the proof.

Lemma 12.12 *(cf. Lemma 7.2*) The functions* $U_1, U_2, \ldots, U$ *are Borel measurable. For each* $\epsilon > 0$ *and* $n \geq 1$, *there is an invariant, Borel* Γ*-selector* γ *such that* $u(\gamma^\infty(\cdot))$ *is w-invariant, Borel measurable, and*

$$u(\gamma^\infty(x)) \geq (1 - \epsilon/2)U_n(x) \text{ for all } x. \tag{12.9}$$

Hence, for each $\alpha \in \mathbb{P}(F)$, *there exists such a Borel* Γ*-selector* γ *for which*

$$u(\gamma^\infty(x)) \geq (1 - \epsilon)U(x) \tag{12.10}$$

with α*-probability at least* $1 - \epsilon$.

Proof. The first assertion is easy to prove because Γ is countably parameterized (see Theorem 4.1 of Sudderth (1969)). Next choose $\beta \in (0, 1)$ such that $\beta^n > 1 - \epsilon/2$. Then the Γ-selector η given by Lemma 12.8 is invariant and satisfies (1.9). Also, because the functions $U_1, U_2, \ldots$ are Borel, it is clear from the construction in the proof of Lemma 12.8 that η is Borel. However, it is not clear that the function $u(\eta^\infty(\cdot))$ is Borel measurable. To sidestep this difficulty, we consider the Borel, invariant, stop-or-go house Σ where $\Sigma(x) = \{\eta(x), \delta(x)\}$ for all x. Then, for γ and W defined as in Lemma 12.11, we have $u(\gamma^\infty(\cdot)) = W(\cdot) \geq u(\eta^\infty(\cdot))$, and, by Lemma 12.11 again, $u(\gamma^\infty(\cdot))$ is Borel and w-invariant. Since $U_n \uparrow U$, the final assertion follows from (1.9). ∎

The idea of the remainder of the proof, based on ideas of Ornstein (1969), is to find a Borel invariant stop-or-go subhouse of the original Γ whose optimal reward function is almost surely as large as $(1 - \epsilon)U$. Theorem 12.2 will then follow from Lemma 12.11.

The construction of the stop-or-go house is in a countable number of steps and each step will use the following operation.

To each stop-or-go house Σ, associate the house $\Gamma \cdot \Sigma$ defined by

$$(\Gamma \cdot \Sigma)(x) = \begin{cases} \Sigma(x), & \text{if } \Sigma(x) \text{ contains two elements} \\ \Gamma(x), & \text{otherwise.} \end{cases}$$

Plainly, $\Sigma \subseteq \Gamma \Rightarrow \Sigma \subseteq \Gamma \cdot \Sigma \subseteq \Gamma$ and $\Sigma \subseteq \Sigma' \Rightarrow \Gamma \cdot \Sigma' \subseteq \Gamma \cdot \Sigma$.

If η is Borel measurable and $\Sigma(x) = \{\eta(x), \delta(x)\}$ for all x, then Σ is Borel and countably parameterized, as is $\Gamma \cdot \Sigma$. So, by Lemma 12.12, the optimal return function W for Σ is Borel measurable, as is the optimal return function R for $\Gamma \cdot \Sigma$.

Lemma 12.13 *(cf. Lemma 7.5*) Suppose Σ is a leavable, invariant, Borel, stop-or-go subhouse of Γ, $\alpha \in \mathbb{P}(F)$, and $\epsilon > 0$. Then there is a leavable, invariant, Borel, stop-or-go house Σ' such that*

(i) $\Sigma \subseteq \Sigma' \subseteq \Gamma$,

(ii) $\alpha[W' \geq (1-\epsilon)R] \geq 1-\epsilon$, *and*

(iii) $R' \geq (1-\epsilon)R$,

where W' and R' are the optimal return functions for Σ' and $\Gamma \cdot \Sigma'$, respectively.

Proof. The house $\Gamma \cdot \Sigma$ is invariant, and hence, by Lemma 12.12, there is an invariant, Borel $\Gamma \cdot \Sigma$-selector γ such that $u(\gamma^\infty(\cdot))$ is Borel, w-invariant, and $\alpha(S) > 1-\epsilon$, where $S = \{x : u(\gamma^\infty(x)) \geq (1-\epsilon^2/2)R(x)\}$. Set $T = \{x : u(\gamma^\infty(x)) \geq (1-\epsilon)R(x)\}$. Then the set T is Borel and, because the functions u and R are w-invariant and the action of G is multiplicative, T also has the property that $x \in T \Leftrightarrow gx \in T$ for all $x \in F, g \in G$. Now define the house Σ' by

$$\Sigma'(x) = \begin{cases} \{\eta(x), \delta(x)\}, \text{ if } x \in T \text{ and } \Sigma(x) = \{\delta(x)\}, \\ \Sigma(x), \text{ if } \Sigma(x) \text{ contains two elements}, \\ \{\delta(x)\}, \text{ otherwise.} \end{cases}$$

The proof that Σ' has the desired properties, with the exception of invariance, is exactly the same as in the proof of Lemma 7.5*. The proof that Σ' is invariant is straightforward if one uses the property of T mentioned above and our standing assumption that the mappings $x \to gx$ are one-to-one. ∎

Lemma 12.14 *(cf. Lemma 7.6*) Let $\alpha \in \mathbb{P}(F)$ and $\epsilon > 0$. There is a sequence $\Sigma_0 \subseteq \Sigma_1 \subseteq \cdots$ of leavable, invariant, Borel stop-or-go subhouses of Γ whose optimal return functions $W_0, W_1, \ldots$ satisfy*

$$\alpha[W_n \geq (1-\epsilon)U] \uparrow 1.$$

Proof. The proof is the same as that of Lemma 7.6* except that Lemma 12.13 is used instead of Lemma 7.5*. ∎

Define the house Σ by setting

$$\Sigma(x) = \cup_n \Sigma_n(x), \qquad x \in F,$$

where the Σ_n are from Lemma 12.14. Then Σ is a leavable, invariant, stop-or-go subhouse of Γ whose optimal return function W satisfies $\alpha[W \geq (1-\epsilon)U] = 1$. Theorem 12.2 now follows from Lemma 12.11.

12.7 APPENDIX

The aim of this appendix is to sketch a proof of the following fact:

Theorem 12.3 *If σ is a measurable strategy available in a gambling house Γ at x and u is a nonnegative, upper analytic function on F, then*

$$u(\sigma) \geq \tilde{u}(\sigma),$$

where $\tilde{u}(\sigma) = \limsup \int u(X_t)\, d\sigma$, the lim sup being taken over all stop rules measurable or not.

Consider first an analytic gambling house $\bar{\Gamma}$ such that $\bar{\Gamma}(x)$ is a singleton for each x. Fix a state x_0. If $\gamma(\cdot)$ is a $\bar{\Gamma}$-selector, then it is necessarily Borel measurable and $\gamma^\infty(x_0)$ is the only strategy available in $\bar{\Gamma}$ at x_0. Fix a measurable stop rule s and denote by γ^* the distribution of the terminal state when the the policy $(\gamma^\infty(x_0), s)$ is used starting at x_0.

Enlarge the state space F by adding a new element x^* and define a new gambling house Γ^* on $F^* = F \cup \{x^*\}$ as follows:

$$\Gamma^*(x^*) = \{\gamma^*\} \text{ and } \Gamma^*(x) = \bar{\Gamma}(x) \text{ if } x \neq x^*.$$

Extend the selector $\gamma(\cdot)$ to F^* by setting $\gamma(x^*) = \gamma^*$. Then $\gamma^\infty(x^*)$ is the only strategy available in Γ^* at x^*.

It follows from Theorem 4.8 in Maitra, Purves, and Sudderth (1990) that

$$\sup\{E_{x^*}(u(X_t)) \,:\, t \geq 1, t \text{ a measurable stop rule}\}$$

$$= \sup\{E_{x^*}(u(X_t)) \,:\, t \geq 1, t \text{ a stop rule}\},$$

where E_{x^*} is the expectation under the unique strategy available in Γ^* at x^*. In terms of the gambling house $\bar{\Gamma}$ and initial state x_0, this equality becomes

$$\sup\{E_{x_0}(u(X_t)) \,:\, t \geq s, t \text{ a measurable stop rule}\}$$

$$= \sup\{E_{x_0}(u(X_t)) \,:\, t \geq s, t \text{ a stop rule}\},$$

where E_{x_0} is the expectation under the unique strategy available in $\bar{\Gamma}$ at x_0. Hence

$$\inf_{s \text{ measurable}} \sup\{E_{x_0}(u(X_t)) \,:\, t \geq s, t \text{ a measurable stop rule}\}$$

$$= \inf_{s \text{ measurable}} \sup\{E_{x_0}(u(X_t)) \,:\, t \geq s, t \text{ a stop rule}\}$$

$$\geq \inf_s \sup\{E_{x_0}(u(X_t)) \,:\, t \geq s, t \text{ a stop rule}\}.$$

We have thus proved that $u(\gamma^\infty(x_0)) \geq \tilde{u}(\gamma^\infty(x_0))$, which is the special case of Theorem 12.3 corresponding to a stationary strategy σ.

For the general case, assume without loss of generality that σ is Borel measurable. Define a gambling house $\bar{\Gamma}$ on the set of partial histories so that the only gamble available in $\bar{\Gamma}((x_1, x_2, \ldots, x_n))$ is $\sigma_n(x_1, x_2, \ldots, x_n)p^{-1}$ where $p(x) = (x_1, x_2, \ldots, x_n, x)$. To complete the proof, invoke the result just proved for stationary strategies.

12.8 ACKNOWLEDGMENT

The research of Sudderth was partially supported by National Science Foundation Grant DMS 97-03285.

References

[1] C.A. Barbosa-Dantas, *The Existence of Stationary Optimal Plans*, Ph.D. Dissertation, University of California, Berkeley, 1966.

[2] H. Becker and A.S. Kechris, *The Descriptive Set Theory of Polish Group Actions*, LMS Lecture Notes, Cambridge University Press, 1996.

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

[4] D. Blackwell and S. Ramakrishnan, "Stationary plans need not be uniformly optimal for leavable, Borel gambling problems," *Proceedings of the American Mathematical Society* **102**, 1024-1027, 1988.

[5] J.P. Burgess, "A selection theorem for group actions," *Pacific Journal of Mathematics* **80**, 333-336, 1979.

[6] C. Dellacherie and P.A. Meyer, *Probabilités et Potentiel*, Chapitres IX à XI, Hermann, Paris, 1983.

[7] L. Dubins, A. Maitra, R. Purves, and W. Sudderth, "Measurable, nonleavable gambling problems," *Israel Journal of Mathematics* **67**, 257-271, 1989.

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

[9] L.E. Dubins and W.D. Sudderth, "On stationary strategies for absolutely continuous houses," *Annals of Probability* **7**, 461-476, 1979.

[10] M. Eaton, *Invariance Applications in Statistics*, Regional Conference Series in Probability and Statistics I, Institute of Mathematical Statistics, Hayward, California, 1989.

[11] E.B. Frid, "On a problem of D. Blackwell from the theory of dynamic programming," *Theory of Probability and Applications* **15**, 719-722, 1976.

[12] L. Harrington, A.S. Kechris, and A. Louveau, "A Glimm-Effros dichotomy for Borel equivalence relations," *Journal of the American Mathematical Society* **3** 903-928, 1990.

[13] A. Maitra, R. Purves, and W. Sudderth, "Leavable gambling problems with unbounded utilities," *Transactions of the American Mathematical Society* **333**, 543-567, 1990.

[14] A. Maitra and W. Sudderth, *Discrete Gambling and Stochastic Games*, Springer-Verlag, New York, 1996.

[15] P.A. Meyer and M. Traki, "Réduites et jeux de hasard," *Séminaire de Probabilités XII*, Lecture Notes in Mathematics 321, Springer-Verlag, Berlin, 155-171, 1973.

[16] D. Ornstein, "On the existence of stationary optimal strategies," *Proceedings of the American Mathematical Society* **20**, 563-569, 1969.

[17] L.J. Savage, *The Foundations of Statistics*, Wiley, New York, 1954.

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

[19] M. Schäl and W. Sudderth, "Stationary policies and Markov policies in Borel dynamic programming," *Probability Theory and Related Fields* **74**, 91-111, 1987.

[20] S.M. Srivastava, *A Course on Borel Sets*, Springer-Verlag, Berlin, 1998.

[21] R.E. Strauch, "Negative dynamic programming," *Annals of Mathematical Statistics* **37**, 871-890, 1966.

[22] R.E. Strauch, "Measurable gambling houses," *Transactions of the American Mathematical Society* **126**, 64-72, 1967.

[23] W.D. Sudderth, "On the existence of good stationary strategies," *Transactions of the American Mathematical Society* **135**, 399-414, 1969.

Lester E. Dubins
Departments of Statistics and Mathematics
University of California
Berkeley, CA 94720, USA
lester@stat.berkeley.edu

Ashok P. Maitra
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA
maitr001@tc.umn.edu

William D. Sudderth
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA
bill@stat.umn.edu

III Applications

13 NEURO-DYNAMIC PROGRAMMING: OVERVIEW AND RECENT TRENDS

Benjamin Van Roy

Abstract: Neuro-dynamic programming is comprised of algorithms for solving large-scale stochastic control problems. Many ideas underlying these algorithms originated in the field of artificial intelligence and were motivated to some extent by descriptive models of animal behavior. This chapter provides an overview of the history and state-of-the-art in neuro-dynamic programming, as well as a review of recent results involving two classes of algorithms that have been the subject of much recent research activity: temporal-difference learning and actor-critic methods.

13.1 INTRODUCTION

In the study of decision-making, there is a dividing line between those who seek an understanding of how decisions *are made* and those who analyze how decisions *ought to be made* in the light of clear objectives. Among the former group are psychologists and economists who examine participants of physical systems in their full complexity. This often entails the consideration of both "rational" and "irrational" behavior. The latter group—those concerned with *rational decision-making*—includes engineers and management scientists who focus on the strategic behavior of sophisticated agents with definite purposes. The intent is to devise strategies that optimize certain criteria and/or meet specific demands. The problems here are well-defined and the goal is to find a "correct" way to make decisions, if one exists.

The self-contained character of rational decision problems has provided a ground for the development of much mathematical theory. Results of this work—as exemplified by previous chapters of this volume—provide an under-

standing of various possible models of dynamics, uncertainties, and objectives, as well as characterizations of optimal decision strategies in these settings. In cases where optimal strategies do exist, the theory is complemented by computational methods that deliver them.

In contrast to rational decision-making, there is no clear-cut mathematical theory about decisions made by participants of natural systems. Scientists are forced to propose speculative theories, and to refine their ideas through experimentation. In this context, one approach has involved the hypothesis that behavior is in some sense rational. Ideas from the study of rational decision-making are then used to characterize such behavior. In financial economics, this avenue has lead to utility and equilibrium theory. To this day, models arising from this school of economic thought—though far from perfect—are employed as mainstream interpretations of the dynamics of capital markets. The study of animal behavior presents another interesting case. Here, evolutionary theory and its popular precept—"survival of the fittest"—support the possibility that behavior to some extent concurs with that of a rational agent.

There is also room for reciprocal contributions from the study of natural systems to the science of rational decision-making. The need arises primarily due to the computational complexity of decision problems and the lack of systematic approaches for dealing with it. For example, practical problems addressed by the theory of dynamic programming can rarely be solved using dynamic programming algorithms because the computational time required for the generation of optimal strategies typically grows exponentially in the number of variables involved—a phenomenon known as the *curse of dimensionality*. This deficiency calls for an understanding of suboptimal decision-making in the presence of computational constraints. Unfortunately, no satisfactory theory has been developed to this end.

It is interesting to note that similar computational complexities arise in attempts to automate decision tasks that are naturally performed by humans or animals. The fact that biological mechanisms facilitate the efficient synthesis of adequate strategies motivates the possibility that understanding such mechanisms can inspire new and computationally feasible methodologies for strategic decision-making.

Over the past two decades, algorithms of *reinforcement learning*—originally conceived as descriptive models for phenomena observed in animal behavior– have grown out of the field of artificial intelligence and been applied to solving complex sequential decision problems. The success of reinforcement learning algorithms in solving large-scale problems has generated excitement and intrigue among operations researchers and control theorists, and much subsequent research has been devoted to understanding such methods and their potential. Developments have focused on a normative view, and to acknowledge the relative disconnect from descriptive models of animal behavior, some operations researchers and control theorists have come to refer to this area of research as *neuro-dynamic programming*, instead of *reinforcement learning*.

In this chapter, we provide a sample of recent developments and open issues at the frontier of research in neuro-dynamic programming. Our two points of focus are temporal-difference learning and actor-critic methods—two algo-

rithmic ideas that have found greatest use in applications of neuro-dynamic programming and for which there has been significant theoretical progress in recent years. We begin, though, with three sections providing some background and perspective on the methodology and problems that may address.

13.2 STOCHASTIC CONTROL

As a problem formulation, let us consider a discrete-time dynamic system that, at each time t, takes on a state x_t and evolves according to

$$x_{t+1} = f(x_t, a_t, w_t),$$

where w_t is a disturbance and a_t is a control decision. Though more general (infinite/continuous) state spaces can be treated, to keep the exposition simple, we restrict attention to finite state, disturbance, and control spaces, denoted by $\mathbb{X}$, $\mathbb{W}$, and $\mathbb{A}$, respectively. Each disturbance $w_t \in \mathbb{W}$ is independently sampled from some fixed distribution.

A function $r : \mathbb{X} \times \mathbb{A} \mapsto \mathbb{R}$ associates a reward $r(x_t, a_t)$ with a decision a_t made at state x_t. A *stationary policy* is a mapping $\phi : \mathbb{X} \mapsto \mathbb{A}$ that generates state-contingent decisions. For each stationary policy ϕ, we define a value function $v(\cdot, \phi) : \mathbb{X} \mapsto \mathbb{R}$ by

$$v(x, \phi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t r(x_t, \phi(x_t)) \Big| x_0 = x\right],$$

where $\beta \in [0, 1)$ is a discount factor and the state sequence is generated according to $x_0 = x$ and $x_{t+1} = f(x_t, \phi(x_t), w_t)$. Each $v(x, \phi)$ can be interpreted as an assessment of long term rewards given that we start in state x and control the system using a stationary policy ϕ. The optimal value function V is defined by

$$V(x) = \max_{\phi} v(x, \phi).$$

A standard result in dynamic programming states that any stationary policy ϕ^* given by

$$\phi^*(x) = \operatorname*{argmax}_{a \in \mathbb{A}} \mathop{\mathbb{E}}_{w}\Big[r(x, a) + \beta V(f(x, a, w))\Big],$$

where $\mathop{\mathbb{E}}_{w}[\cdot]$ denotes expectation with respect to the distribution of disturbances, is optimal in the sense that

$$V(x) = v(x, \phi^*),$$

for every state x (see, e.g. [8]).

For illustrative purposes, let us provide one example of a stochastic control problem.

Example 13.1 *The video arcade game of Tetris can be viewed as an instance of stochastic control (we assume that the reader is familiar with this popular game). In particular, we can view the state x_t as an encoding of the current "wall of bricks" and the shape of the current "falling piece." The decision a_t*

identifies an orientation and horizontal position for placement of the falling piece onto the wall. Though the arcade game employs a more complicated scoring system, consider for simplicity a reward $r(x_t, a_t)$ equal to the number of rows eliminated by placing the piece in the position described by a_t. Then, a stationary policy ϕ that maximizes the value

$$v(x, \phi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t r(x_t, \phi(x_t)) \Big| x_0 = x\right],$$

essentially optimizes a combination of present and future row elimination, with decreasing emphasis placed on rows to be eliminated at times farther into the future.

Classical dynamic programming algorithms compute the optimal value function V. The result is stored in a "look-up" table with one entry $V(x)$ per state $x \in \mathbb{X}$. When the need arises, the value function is used to generate optimal decisions. In particular, given a current state $x_t \in \mathbb{X}$, a decision a_t is selected according to

$$a_t = \underset{a \in \mathbb{A}}{\operatorname{argmax}} \underset{w}{\mathbb{E}} \Big[r(x_t, a) + \beta V(f(x_t, a, w)) \Big].$$

13.3 CONTROL OF COMPLEX SYSTEMS

Our primary interest is in the development of a methodology for the control of "complex systems." It is difficult to provide a precise definition for this term, but let us mention two characteristics that are common to such systems: an intractable state space and severe nonlinearities. Intractable state spaces preclude the use of classical dynamic programming algorithms, which compute and store one numerical value per state. At the same time, methods of traditional linear control, which are applicable even when state spaces are large, are ruled out by severe nonlinearities. To give a better feel for the types of problems we have in mind, let us provide a few examples.

1. **Call Admission and Routing**
   With rising demand in telecommunication network resources, effective management is as important as ever. Admission (deciding which calls to accept/reject) and routing (allocating links in the network to particular calls) are examples of decisions that must be made at any point in time. The objective is to make the "best" use of limited network resources. In principle, such sequential decision problems can be addressed by dynamic programming. Unfortunately, the enormous state spaces involved render dynamic programming algorithms inapplicable, and heuristic control strategies are used in lieu.

2. **Strategic Asset Allocation**
   Strategic asset allocation is the problem of distributing an investor's wealth among assets in the market in order to take on a combination of risk and expected return that best suits the investor's preferences. In general, the optimal strategy involves dynamic rebalancing of wealth

among assets over time. If each asset offers a fixed rate of risk and return, and some additional simplifying assumptions are made, the only state variable is wealth, and the problem can be solved efficiently by dynamic programming algorithms. There are even closed form solutions in cases involving certain types of investor preferences [39]. However, in the more realistic setting involving risks and returns that fluctuate with economic conditions (see, e.g. [11]), economic indicators must be taken into account as state variables, and this quickly leads to an intractable state space. The design of effective strategies in such situations constitutes an important challenge in the growing field of financial engineering.

3. **Supply-Chain Management**
   With today's tight vertical integration, increased production complexity, and diversification, the inventory flow within and among corporations can be viewed as a complex network—called a *supply chain*—consisting of storage, production, and distribution sites. In a supply chain, raw materials and parts from external vendors are processed through several stages to produce finished goods. Finished goods are then transported to distributors, then to wholesalers, and finally retailers, before reaching customers. The goal in supply-chain management is to achieve a particular level of product availability while minimizing costs. The solution is a policy that decides how much to order or produce at various sites given the present state of the company and the operating environment. See [34] and references therein for further discussion of this problem.

4. **Emissions Reductions**
   The threat of global warming that may result from accumulation of carbon dioxide and other "greenhouse gasses" poses a serious dilemma. In particular, cuts in emission levels bear a detrimental short-term impact on economic growth. At the same time, a depleting environment can severely hurt the economy—especially the agricultural sector–in the longer term. To complicate the matter further, scientific evidence on the relationship between emission levels and global warming is inconclusive, leading to uncertainty about the benefits of various cuts. One systematic approach to considering these conflicting goals involves the formulation of a dynamic system model that describes our understanding of economic growth and environmental science, as is done in [40]. Given such a model, the design of environmental policy amounts to dynamic programming. Unfortunately, classical algorithms are inapplicable due to the size of the state space.

5. **Semiconductor Wafer Fabrication**
   The manufacturing floor at a semiconductor wafer fabrication facility is organized into service stations, each equipped with specialized machinery. There is a single stream of jobs arriving on a production floor. Each job follows a deterministic route that revisits the same station multiple times. This leads to a scheduling problem where, at any time, each station must select a job to service such that (long term) production capacity is maximized (see, e.g. [33]). Such a system can be viewed as a special class of

queueing networks, which are models suitable for a variety of applications in manufacturing, telecommunications, and computer systems. Optimal control of queueing networks is notoriously difficult, and this reputation is strengthened by formal characterizations of computational complexity in [41].

For complex systems as those we have described, state spaces are intractable. This is a consequence of the "curse of dimensionality"—that is, the fact that state spaces generally grow exponentially in the number of state variables. For example, in a queueing network, every possible configuration of queues corresponds to a different state, and therefore, the number of states increases exponentially with the number of queues involved. For this reason, it is essentially impossible to compute (or even store) one value per state, as is required by classical dynamic programming algorithms.

There is an additional shortcoming of classical dynamic programming algorithms that is worth mentioning here—that the computations they carry out require use of transition probabilities. For many complex systems, such probabilities are not readily accessible. On the other hand, it is often easier to develop a simulator for the system that generates sample trajectories, as is commonly done to test performance of particular decision policies.

Neuro-dynamic programming algorithms aim at overcoming both deficiencies of classical algorithms. The curse of dimensionality is conquered through use of parameterized function approximators that approximate the value function in a spirit similar to statistical regression. At the same time, these algorithms rely on output generated by simulators, rather than explicit transition probabilities, in their computation.

13.4 VALUE FUNCTION APPROXIMATION

The intractability of state spaces calls for value function approximation. There are two important preconditions for the development of an effective approximation. First, we need to choose a parameterization $\tilde{v} : \mathbb{X} \times \mathbb{R}^K \mapsto \mathbb{R}$ that yields a good approximation

$$\tilde{v}(x, u) \approx V(x),$$

for some setting of the parameter vector $u \in \mathbb{R}^K$. In this respect, the choice of a suitable parameterization requires some practical experience or theoretical analysis that provides rough information about the shape of the function to be approximated. Second, we need algorithms for computing appropriate parameter values, such as those studied in neuro-dynamic programming.

Though more general classes of parameterizations have been used in neuro-dynamic programming, to keep the exposition simple, let us focus on linear parameterizations, which take the form

$$\tilde{v}(x, u) = \sum_{k=1}^{K} u(k)\psi_k(x),$$

where $\psi_1, \ldots, \psi_K$ are "basis functions" mapping $\mathbb{X}$ to $\mathbb{R}$ and $u = (u(1), \ldots, u(K))'$ is a vector of scalar weights. In a spirit similar to that of statistical

regression, the basis functions $\psi_1, \ldots, \psi_K$ are selected by a human user based on intuition or analysis specific to the problem at hand. One interpretation that is useful for the construction of basis functions involves viewing each function ψ_k as a "feature"—that is, a numerical value capturing a salient characteristic of the state that may be pertinent to effective decision making. This general idea is probably best illustrated by a concrete example.

Example 13.2 *In our stochastic control formulation of Tetris, the state is an encoding of the current wall configuration and the current falling piece. There are clearly too many states for exact dynamic programming algorithms to be applicable. However, we may believe that most information relevant to game-playing decisions can be captured by a few intuitive features. In particular, one feature, say ψ_1, may map states to the height of the wall. Another, say ψ_2, could map states to a measure of "jaggedness" of the wall. A third might provide a scalar encoding of the type of the current falling piece (there are seven different shapes in the arcade game). Given a collection of such features, the next task is to select weights $u(1), \ldots, u(K)$ such that*

$$\sum_{k=1}^{K} u(k)\psi_k(x) \approx V(x),$$

for all states x. This approximation could then be used to generate a game-playing strategy. Such an approach to Tetris has been developed in [58] and [9]. In the latter reference, with 22 features, the authors are able to generate a strategy that eliminates an average of 3554 rows per game, reflecting performance comparable to that of an expert player.

13.5 TEMPORAL-DIFFERENCE LEARNING

In this section, we introduce temporal-difference learning as applied to tuning basis function weights in autonomous and controlled systems. Our presentation is not mathematically rigorous. Instead, emphasis is placed on conveying ideas and results at an intuitive level. More detailed discussions and mathematical analyses can be found in cited references.

13.5.1 Autonomous systems

Let us begin by considering an autonomous process

$$x_{t+1} = f(x_t, w_t),$$

and aim at approximating a value function

$$V(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t r(x_t) \Big| x_0 = x\right],$$

where $r(x)$ is a scalar reward associated with state x and $\beta \in [0, 1)$ is a discount factor. Note that this setting is equivalent to one where we are dealing with a controlled system and wish to approximate the value function $v(\cdot, \phi)$ corresponding to a fixed stationary policy ϕ.

Let $\psi_1, \ldots, \psi_K$ be a collection of basis functions, and let $\tilde{v} : \mathbb{X} \times \mathbb{R}^K \mapsto \mathbb{R}$ be defined by

$$\tilde{v}(x, u) = \sum_{k=1}^{K} u(k)\psi_k(x).$$

Suppose that we observe a sequence of states $x_0, x_1, x_2, \ldots$ and that at time t the weight vector has been set to some value u_t. We define the *temporal difference* d_t corresponding to the transition from x_t to x_{t+1} by

$$d_t = r(x_t) + \beta\tilde{v}(x_{t+1}, u_t) - \tilde{v}(x_t, u_t).$$

Then, given an arbitrary initial weight vector u_0, the temporal-difference learning algorithm generates subsequent weight vectors according to

$$u_{t+1} = u_t + \gamma_t d_t z_t,$$

where γ_t is a scalar step size, and $z_t \in \mathbb{R}^K$ is an *eligibility vector* defined by

$$z_t = \sum_{\tau=0}^{t} (\beta\lambda)^{t-\tau}\psi(x_\tau),$$

where $\psi(x) = (\psi_1(x), \ldots, \psi_K(x))'$. The parameter λ takes on values in $[0, 1]$, and to emphasize its presence, the temporal-difference learning is often referred to as TD(λ). Note that the eligibility vectors can be recursively updated according to

$$z_{t+1} = \beta\lambda z_t + \psi(x_{t+1}).$$

Let us provide one (heuristic) interpretation of the algorithm. Note that the temporal difference d_t can be viewed as a difference between two predictions of future rewards:

1. $\tilde{v}(x_t, u_t)$ is a prediction of $\sum_{\tau=t}^{\infty} \beta^{\tau-t} r(x_\tau)$ given our current approximation $\tilde{v}(\cdot, u_t)$ to the value function.

2. $r(x_t) + \beta\tilde{v}(x_{t+1}, u_t)$ is an "improved prediction" that incorporates knowledge of the reward $r(x_t)$ and the next state x_{t+1}.

Roughly speaking, the learning process tries to make predictions $\tilde{v}(x_t, u_t)$ consistent with their improved versions. Note that $\psi(x_t) = \nabla_u \tilde{v}(x_t, u)$. (We define the gradient ∇g of a function $g : \mathbb{R}^m \mapsto \mathbb{R}$ to be a vector-valued function mapping $\mathbb{R}^m$ to $\mathbb{R}^m$ with the each ith component function equal to the partial derivative of g with respect to the ith component of its domain.) Consequently, when $\lambda = 0$, the update can be rewritten as

$$u_{t+1} = u_t + \gamma_t \nabla_u \tilde{v}(x_t, u_t)\Big(r(x_t) + \beta\tilde{v}(x_{t+1}, u_t) - \tilde{v}(x_t, u_t)\Big).$$

The gradient can be viewed as providing a direction for the adjustment of u_t such that $\tilde{v}(x_t, u_t)$ moves towards the improved prediction. In the more general case of $\lambda \in [0, 1]$, the direction of the adjustment is determined by the eligibility vector $z_t = \sum_{\tau=0}^{t} (\beta\lambda)^{t-\tau} \nabla_u \tilde{v}(x_\tau, u_t)$. Here, each gradient term

in the summation corresponds to one of the previous states, and the temporal difference can be viewed as "triggering" adjustments of all previous predictions. The powers of β account for discounting effects inherent to the problem, while the powers of λ influence the "credit assignment"—that is, the amounts by which previous predictions are to be adjusted based on the current temporal difference.

A sizable literature addresses the dynamics of temporal-difference methods in the context of an autonomous process. Examples include [49, 18, 19, 25, 43, 58, 62]. The most recent of these results [58, 62] state that, under appropriate technical conditions:

1. For any $\lambda \in [0,1]$, there exists a vector $u^{(\lambda)}$ such that the sequence u_t generated by the algorithm converges (with probability one) to $u^{(\lambda)}$.

2. The limit of convergence $u^{(\lambda)}$ satisfies

$$\|V - \Psi u^{(\lambda)}\|_\mu \leq \frac{1}{\sqrt{1-\kappa^2}} \|\Pi V - V\|_\mu,$$

where

$$\kappa = \frac{\beta(1-\lambda)}{1-\lambda\beta} \leq \beta,$$

the norm $\|\cdot\|_\mu$ is defined by

$$\|v\|_\mu = \left(\sum_{x \in \mathbb{X}} \mu(x) v^2(x) \right)^{1/2},$$

with μ being the invariant distribution of the process, and the matrix Π projects onto the span of $\psi_1, \ldots, \psi_K$ with respect to $\|\cdot\|_\mu$.

These results imply that the iterates u_t converge to some $u^{(\lambda)}$. Furthermore, $\Psi u^{(\lambda)}$ provides an approximation to V in a sense that we will now describe. The term $\|\Pi V - V\|_\mu$ represents the error associated with the projection ΠV. By the projection theorem, this error is minimal (if we are constrained to selecting approximations within the span of $\psi_1, \ldots, \psi_K$). The bound stated above therefore establishes that the error associated with $\Psi u^{(\lambda)}$ is within a constant factor of the best possible.

13.5.2 Controlled systems

The algorithm described in the previous section involves simulating a system and updating weights of an approximate value function based on observed state transitions. Unlike an autonomous system, a controlled system cannot be passively simulated and observed. Control decisions are required and influence the system's dynamics. In this section, we discuss extensions of temporal-difference learning to this context. The objective is to approximate the optimal value function of a controlled system.

Approximate policy iteration. A well-known result in dynamic programming is that, given a value function $v(\cdot, \phi)$ corresponding to a stationary policy ϕ, an improved policy $\overline{\phi}$ can be defined by

$$\overline{\phi}(x) = \operatorname*{argmax}_{a \in \mathbb{A}} \mathop{\mathbb{E}}_{w} \Big[r(x,a) + \beta v(f(x,a,w), \phi) \Big].$$

In particular, $v(x, \overline{\phi}) \geq v(x, \phi)$ for all $x \in \mathbb{X}$. Furthermore, a sequence of policies $\{\phi_m | m = 0, 1, 2, \dots\}$ initialized with some arbitrary ϕ_0 and updated according to

$$\phi_{m+1}(x) = \operatorname*{argmax}_{a \in \mathbb{A}} \mathop{\mathbb{E}}_{w} \Big[r(x,a) + \beta v(f(x,a,w), \phi_m) \Big],$$

converges to an optimal policy ϕ^*. This iterative method for generating an optimal policy constitutes *policy iteration*, a classical dynamic programming algorithm due to Howard [28].

As with other dynamic programming algorithms, policy iteration suffers from the curse of dimensionality. In particular, each value function $v(\cdot, \phi_m)$ generated during the course of the algorithm can not be efficiently computed or stored. A possible approach to overcoming such limitations involves approximating each iterate $v(\cdot, \phi_m)$ in terms of a weighted combination of basis functions. For instance, letting $\psi_1, \dots, \psi_K$ be a set of basis functions and letting $\tilde{v}(x,u) = \sum_{k=1}^{K} u(k)\psi_k(x)$, consider generating a sequence of weight vectors $u^1, u^2, \dots$ by selecting each u^{m+1} such that

$$\tilde{v}(x, u^{m+1}) \approx v(x, \tilde{\phi}_m),$$

where $\tilde{\phi}_0$ is an arbitrary initial stationary policy and for $m = 1, 2, 3, \dots,$

$$\tilde{\phi}_m(x) = \operatorname*{argmax}_{a \in \mathbb{A}} \mathop{\mathbb{E}}_{w} \Big[r(x,a) + \beta \tilde{v}(f(x,a,w), u^m) \Big].$$

We will refer to such an algorithm as *approximate policy iteration.*

There is one key component missing in our description of approximate policy iteration—a method for generating each iterate u^m. The possibility we have in mind is, of course, temporal-difference learning. In particular, we can apply the temporal-difference learning algorithm to the autonomous system resulting from simulation of the controlled system under a fixed stationary policy $\tilde{\phi}_m$. (The dynamics are described by $x_{t+1} = f(x_t, \tilde{\phi}_m(x_t), w_t)$.) Initializing with $u_0^{m+1} = u^m$, the algorithm would generate a sequence of vectors $u_1^{m+1}, u_2^{m+1}, u_3^{m+1}, \dots$ that converges. The limiting vector provides the subsequent iterate u^{m+1}.

To clarify the interplay between the two types of iterations involved in approximate policy iteration, let us note that we have nested sequences:

- An "external" sequence is given by $u^0, u^1, u^2, \dots$
- For each $m = 1, 2, 3, \dots,$ an "internal" sequence is given by $u_0^m, u_1^m, u_2^m, \dots$

For each m, the internal sequence is initialized with $u_0^{m+1} = u^m$ and the limit of convergence becomes the next element u^{m+1} of the external sequence.

The dynamics of approximate policy iteration are not very well understood. However a result from [10] to some extent motivates its use. The result states that, if there exists some $\epsilon > 0$ such that

$$\max_{x \in \mathbb{X}} |(\Psi u^m)(x) - v(x, \phi_{m-1})| \leq \epsilon,$$

for all m, then

$$\limsup_{m \to \infty} \max_{x \in \mathbb{X}} |(\Psi u^m)(x) - V(x)| \leq \frac{2\beta\epsilon}{(1-\beta)^2}.$$

In other words, if each of the policy evaluations errs by no more than ϵ per component, approximate value iteration eventually comes within a constant factor of ϵ from the optimal value function. However, as illustrated by examples in [10], the external sequence u^m does not always converge.

Controlled TD. Any function $v : \mathbb{X} \mapsto \mathbb{R}$ can be used to generate a stationary policy

$$\phi(x) = \operatorname*{argmax}_{a \in \mathbb{A}} \mathop{\mathbb{E}}_{w} \Big[r(x,a) + \beta v(f(x,a,w)) \Big].$$

In this respect, one can view v as a guide for decision-making. The value functions $v(\cdot, \phi_0), v(\cdot, \phi_1), v(\cdot, \phi_2), \ldots$ generated by (exact) policy iteration can then be viewed as a monotonically improving sequence of guides.

Recall that given a stationary policy ϕ, the value function $v(\cdot, \phi)$ generates an improved policy. It therefore seems reasonable to hope that the approximation $\tilde{v}(\cdot, u^{m+1})$ to $v(\cdot, \tilde{\phi}_m)$ similarly generates a policy $\tilde{\phi}_{m+1}$ that improves on $\tilde{\phi}_m$. Now recall that, approximate policy iteration employs temporal-difference learning to compute u^{m+1} given u^m. This is done by simulating the system under the control policy $\tilde{\phi}_m$, initializing a sequence with $u_0^{m+1} = u^m$, and generating $u_1^{m+1}, u_2^{m+1}, u_3^{m+1}, \ldots$ according to the temporal-difference learning iteration. Since the corresponding sequence of functions $\tilde{v}(\cdot, u^1), \tilde{v}(\cdot, u^2), \tilde{v}(\cdot, u^3), \ldots$ converges to $\tilde{v}(\cdot, u^{m+1})$, one might speculate that these intermediate functions themselves provide improving guides to decision-making, each of which can be used to control the system. This possibility motivates an alternative algorithm, which we refer to as *controlled TD*.

Controlled TD simulates a state trajectory $x_0, x_1, x_2, \ldots$ and then generates weight vectors $u_0, u_1, u_2, \ldots$. The initial state x_0 and weight vector u_0 can be arbitrary. Given a state x_t and a weight vector u_t, a decision a_t is generated according to

$$a_t = \operatorname*{argmax}_{a \in \mathbb{A}} \mathop{\mathbb{E}}_{w} \Big[r(x_t,a) + \beta \tilde{v}(f(x_t,a,w), u_t) \Big].$$

The next state x_{t+1} is then given by

$$x_{t+1} = f(x_t, a_t, w_t).$$

Analogously with the autonomous case, let the temporal difference d_t be defined by

$$d_t = r(x_t, a_t) + \beta \tilde{v}(x_{t+1}, u_t) - \tilde{v}(x_t, u_t).$$

Then, the weight vector is updated according to

$$u_{t+1} = u_t + \gamma_t d_t z_t,$$

where γ_t is a scalar step size and the eligibility vector $z_t \in \mathbb{R}^K$ is once again defined by

$$z_t = \sum_{\tau=0}^{t} (\beta\lambda)^{t-\tau} \psi(x_\tau).$$

There is little theory providing understanding of controlled TD. For the case of a "look-up table" representation—i.e. where we store one value per state, as is done by classical dynamic programming algorithms—existing results indicate that controlled TD with λ set to 0 converges so long as every state is visited infinitely often in the course of simulation [57, 29]. There are also results involving very restrictive types of parameterizations such as those arising from state aggregation [47, 58, 24]. These establish convergence in the context of such parameterizations for variants of controlled TD that sample states with fixed relative frequencies.

Another special case for which fairly comprehensive results are available involves a version of controlled TD tailored for solving optimal stopping problems —a quite limited albeit practically relevant class of stochastic control problems. This theory establishes convergence of the algorithm to a unique limit that offers a desirable approximation to the value function [62, 61].

In practice, controlled TD often suffers from getting "stuck" in "deadlock" situations. In particular, viewing the procedure in an anthropomorphic light, the state x_t constitutes an animal's operating environment and a_t is the action it takes. The action is selected based on an approximate value function $\tilde{v}(\cdot, u_t)$, and the weight vector u_t is improved based on experience. If the animal always selects actions in terms of a deterministic function of x_t and $\tilde{v}(\cdot, u_t)$, there is a possibility that only a small subset of the state space will ever be visited and that the animal will never "learn" the value of states outside that region. A modification that has been found to be useful in practical applications involves adding "exploration noise" to the controls. One approach to this end involves randomizing decisions by choosing at each time t a decision $a_t = \overline{a}$, for $\overline{a} \in \mathbb{A}$, with probability

$$\frac{\exp\left(\left(\mathop{\mathbb{E}}_{w}\left[r(x_t, \overline{a}) + \beta\tilde{v}(f(x_t, \overline{a}, w), u_t)\right]\right)/\delta\right)}{\sum_{a\in\mathbb{A}} \exp\left(\left(\mathop{\mathbb{E}}_{w}\left[r(x_t, a) + \beta\tilde{v}(f(x_t, a, w), u_t)\right]\right)/\delta\right)},$$

where $\delta > 0$ is a small scalar. Note that at any state, each decision is selected with positive probability upon each visit, and that as δ approaches 0, the probability that a_t is a decision that maximizes

$$\mathop{\mathbb{E}}_{w}\left[r(x_t, a) + \beta\tilde{v}(f(x_t, a, w), u_t)\right],$$

becomes 1.

Recent theoretical results have pointed to an additional reason for exploration. The trajectory of weight vectors u_t generated by controlled TD can be

viewed as an approximation to a trajectory of an ordinary differential equation. Limits of convergence of the algorithm correspond to stationary points of the ordinary differential equation. Some recent work has studied such stationary points, showing that—in the absence of exploration—there need not exist any stationary points [20]. This work also shows that, with the incorporation of exploration of the type described above, controlled TD is guaranteed to possess at least one stationary point. One might also hope that this stationary point is unique. However, as illustrated by an example in [20], this is not necessarily the case.

Due to the current absence of adequate theory, there is no streamlined and widely accepted version of controlled TD. Instead, there is a conglomeration of variants, and each one is parameterized by values that must be selected by a user. It is unclear which algorithms and parameter settings will work on a particular problem, and when a method does work, it is still unclear which ingredients are actually necessary for success. As a result, applications often require trial and error in a long process of parameter tweaking and experimentation.

Approximating the Q-function. Given the optimal value function V, the generation of optimal control decisions

$$a_t = \operatorname*{argmax}_{a \in \mathbb{A}} \mathop{\mathbb{E}}_{w} \Big[r(x_t, a) + \beta V\big(f(x_t, a, w)\big) \Big],$$

requires computing one expectation per element of the decision space $\mathbb{A}$, which requires in turn repeated evaluation of the system function f. One approach to avoiding this computation involves obtaining a "Q-function," as originally introduced by Watkins [64], which maps $\mathbb{X} \times \mathbb{A}$ to $\mathbb{R}$ and is defined by

$$Q(x, a) = \mathop{\mathbb{E}}_{w} \Big[r(x, a) + \beta V\big(f(x, a, w)\big) \Big].$$

Given this function, optimal decisions can be computed according to

$$a_t = \operatorname*{argmax}_{a \in \mathbb{A}} Q(x_t, a),$$

which no longer involves taking expectations or evaluating the system function.

Q-learning [64, 65] is a variant of temporal-difference learning that approximates Q functions rather than value functions. The basis functions $\psi_1, \ldots, \psi_K$ now map $\mathbb{X} \times \mathbb{A}$ to $\mathbb{R}$, and the objective is to obtain a weight vector $u = (u(1), \ldots, u(K))'$ such that

$$Q(x, a) \approx \tilde{q}(x, a, u) = \sum_{k=1}^{K} u(k) \psi_k(x, a).$$

Like in controlled TD, Q-learning simulates a state trajectory $x_0, x_1, x_2, \ldots$ and then generates weight vectors $u_0, u_1, u_2, \ldots$. Given a state x_t and a weight vector u_t, a decision a_t is generated according to

$$a_t = \operatorname*{argmax}_{a \in \mathbb{A}} \tilde{q}(x_t, a, u_t).$$

The next state x_{t+1} is then given by

$$x_{t+1} = f(x_t, a_t, w_t).$$

The temporal difference d_t is defined by

$$d_t = r(x_t, a_t) + \beta \tilde{q}(x_{t+1}, a_{t+1}, u_t) - \tilde{q}(x_t, a_t, u_t),$$

and the weight vector is updated according to

$$u_{t+1} = u_t + \gamma_t d_t z_t,$$

where γ_t is a scalar step size and the eligibility vector $z_t \in \mathbb{R}^K$ is defined by

$$z_t = \sum_{\tau=0}^{t} (\beta\lambda)^{t-\tau} \psi(x_\tau, a_\tau).$$

Like in the case of controlled TD, it is often desirable to incorporate exploration, for example, by selecting decisions according to $a_t = \overline{a}$ with probability

$$\frac{\exp\left(\tilde{q}(x_t, \overline{a}, u_t)/\delta\right)}{\sum_{a \in \mathbb{A}} \exp\left(\tilde{q}(x_t, a, u_t)/\delta\right)},$$

for some small parameter $\delta > 0$.

The analysis of Q-learning bears many similarities with that of controlled TD, and results that apply to one can often be generalized in a straightforward way to accommodate the other. For example, results applying to the "look-up table" case apply when $\lambda = 0$ to both controlled TD and Q-learning [57, 29]. Similarly, results on the relevance of exploration to the existence of stationary points for controlled TD [20] can also be extended to the case of Q-learning.

13.5.3 Relationship with approximate value iteration

The classical value iteration algorithm can be described compactly in terms of the "dynamic programming operator" T, defined by

$$(Tv)(x) = \max_{a \in \mathbb{A}} E_w \left[r(x, w) + \beta v(f(x, a, w)) \right],$$

for any v. In particular, value iteration generates a sequence of functions according to $v_{k+1} = Tv_k$, each mapping states to real numbers. This sequence converges to the optimal value function V, which is the unique fixed point of T and can be used to generate an optimal policy.

Approximate value iteration—which dates all the way back to 1959 [7]—aims at approximating each iterate v_k by a linear combination of prespecified basis functions $\psi_1, \ldots, \psi_K$. In rough terms, iterates $\tilde{v}_k$ are generated according to $\tilde{v}(\cdot, u_{k+1}) = \Pi T \tilde{v}(\cdot, u_k)$, where Π is a projection operator that produces a function that is in the span of $\psi_1, \ldots, \psi_K$ and close to $T\tilde{v}_k$. The hope is that $\tilde{v}_k$ converges to a good approximation of V.

Recent work points out that the approximate value iteration need not possess fixed points [20], and therefore should not be expected to converge. In fact, even

in cases where a fixed point exists, and even when the system is autonomous, the algorithm can generate a diverging sequence of weight vectors [62].

Controlled TD can be thought of as a stochastic approximation algorithm designed to converge on fixed points of approximate value iteration [62]. One advantage of controlled TD is its use of simulation to effectively bypass the need to explicitly compute projections required for approximate value iteration. But two features of controlled TD also come to the rescue where approximate value iteration can fail.

One advantage can be fully appreciated in the context of autonomous systems. In this case, through use of simulation, controlled TD visits states with relative frequencies equal to the steady-state distribution of underlying Markov chain. This effectively induces a projection onto the subspace spanned by basis functions with respect to a weighted quadratic norm, with weights given by the relative frequencies. It turns out that the use of such a projection, which is related to the dynamics of the underlying Markov chain, ensures in the autonomous case that approximate value iteration converges to a unique fixed point and that controlled TD converges to the same point. Without the use of simulation, it is generally difficult to implement approximate value iteration with such a projection. This is important since the use of alternative norms in projections can lead to divergence. Results along these lines are proved in [59, 62].

A second advantage, realized in the context of controlled systems, involves the possible introduction of exploration. Without exploration, controlled TD, and related versions of approximate value iteration need not possess fixed points [20].

13.5.4 Historical notes

There is a long history behind the algorithms discussed in the preceding sections. We will attempt to provide a brief historical account of items that are particularly relevant to what we have presented.

The line of research originated in an area of artificial intelligence known as *reinforcement learning*. Temporal-difference—originally proposed by Sutton [49]—comprises a major development in this area, but draws on earlier work by Barto and Sutton [50, 5] on models for classical conditioning phenomena observed in animal behavior and by Barto, Sutton, and Anderson on "actor-critic methods," which will be further discussed in the next section. In the look-up table case, the algorithm also bears similarities with one proposed a decade earlier by Witten [71]. Another major development came with the thesis of Watkins [64], in which "Q-learning" was proposed, and the study of temporal-difference learning was integrated with classical ideas from dynamic programming and stochastic approximation theory. The work of Werbos [66, 67, 68] and Barto, Bradtke, and Singh [4] also contributed to this integration.

In addition to advancing the understanding of temporal-difference learning, the marriage with classical engineering ideas furthered the view of the algorithm as one for addressing complex engineering problems and lead to a number of applications. The practical potential was first demonstrated by Tesauro [54, 55, 56], who used a variant of controlled TD to produce a world-class

Backgammon playing program. Several case studies involving problems such as channel allocation in cellular communication networks [46], elevator dispatching [16, 17], inventory management [63], and job-shop scheduling [72], followed to demonstrate additional signs of promise.

Since the completion of Watkin's thesis, there has been a growing literature involving the application of ideas from dynamic programming and stochastic approximation to the analysis of temporal-difference learning and its variants. However, the existing theory does not provide sufficient support for applications, as we will now explain. In controlled TD, approximation accuracy is limited by the choice of a parameterization. The hope, however, is that the iterative computation of parameters should lead to a good approximation relative to other possibilities allowed by this choice. Unfortunately, there is a shortage of theory that ensures desirable behavior of this kind.

13.6 ACTORS AND CRITICS

The methods described thus far make use of a parameterized representation of the value function. An alternative that has been studied in other research communities as well as within the vein of neuro-dynamic programming involves parameterization of control policies and tuning of parameters via stochastic gradient methods (see, e.g. [23, 26, 69, 15, 37, 36]). Such methods simulate the system of interest and directly adapt the parameters of a controller as performance is observed.

A parameterized controller can be thought of as an *actor*, since it makes decisions and acts on them, thereby influencing the dynamics of the system. On the other hand, an approximate value function can be thought of as a *critic*, assessing alternative decisions to provide guidance. The interpretation as a critic may be particularly suitable in the context of approximate policy iteration, as described in Section 5.2.1. Here, given a stationary policy, one generates an approximate value function (a critic) that provides an evaluation of each state when the system is controlled by this policy. Given feedback from this critic, one can select improved decisions (greedy decisions with respect to the value function).

A current area of active research in neuro-dynamic programming involves the combination of actors and critics, working in tandem to improve system performance. Recent results suggest a possibility that critics can accelerate the computation of gradients that are used to improve actor performance. This line of research provides an interesting interface between value function approximation methods of neuro-dynamic programming and Monte Carlo gradient estimation methods studied in operations research.

In this section, we will overview actor-critic methods and discuss some recent results. Before doing so, however, we will introduce an average reward formulation of stochastic control, involving the maximization of time-averaged rewards rather than discounted rewards. This formulation provides a more natural setting for the developments discussed in this section.

13.6.1 Averaged rewards

As before, we consider a discrete-time stochastic system that, at each time t, evolves according to

$$x_{t+1} = f(x_t, a_t, w_t),$$

where x_t is a state in $\mathbb{X}$, w_t is a disturbance drawn from $\mathbb{W}$, and a is a decision selected from $\mathbb{A}$. We will assume for convenience that for any stationary policy $\phi : \mathbb{X} \mapsto \mathbb{A}$, the Markov chain following

$$x_{t+1} = f(x_t, a_t, \phi(x_t)),$$

is aperiodic and irreducible. The average reward of the system when operated by a stationary policy ϕ is defined by

$$\overline{r}_\phi = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{t=0}^{N} r(x_t, \phi(x_t))\right],$$

and the optimal average reward is

$$r^* = \max_\phi \overline{r}_\phi.$$

Analogous to the value functions employed in a discounted setting, for any stationary policy ϕ, we define a *differential value function* (also known as the *relative value function* or *bias*) by

$$h(x,\phi) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{t=0}^{N} \left(r(x_t, \phi(x_t)) - \overline{r}_\phi\right) \Big| x_0 = x\right],$$

where $x_{t+1} = f(x_t, \phi(x_t), w_t)$. The optimal differential value function is defined by

$$H(x) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{t=0}^{N} \left(r(x_t, \phi^*(x_t)) - r^*\right) \Big| x_0 = x\right],$$

where $x_{t+1} = f(x_t, \phi^*(x_t), w_t)$, and ϕ^* is an optimal policy. Similarly with the discounted case, an optimal policy can be generated by greedy decisions with respect to H:

$$\phi^*(x) = \operatorname*{argmax}_{a\in\mathbb{A}} \mathbb{E}_w \left[r(x,a) + H(f(x,a,w))\right].$$

There is a substantial literature extending temporal-difference learning and related theory to average reward formulations (see, e.g. [44, 45, 35, 53, 1, 2, 62, 60]). To provide a feel for the nature of such generalizations, let us describe a variant of temporal-difference learning that approximates the differential value function of an autonomous system. For this variant, the temporal difference d_t corresponding to a transition from x_t to x_{t+1} is taken to be

$$d_t = r(x_t, a_t) - \overline{r}_t + \tilde{h}(x_{t+1}, u_t) - \tilde{h}(x_t, u_t),$$

where $\overline{r}_t$ represents an approximation to the average reward. These approximations are updated according to

$$\overline{r}_{t+1} = (1 - \gamma_t)\overline{r}_t + \gamma_t r(x_t, a_t)$$

The weights of an approximate value function are simultaneously adapted according to

$$u_{t+1} = u_t + \gamma_t d_t z_t,$$

with the eligibility vector z_t defined by

$$z_{t+1} = \sum_{\tau=0}^{t} \lambda^{t-\tau} \psi(x_\tau).$$

It is shown in [62, 60] that for any $\lambda \in [0, 1)$ and under appropriate technical conditions, $\overline{r}_t$ converges to the true average reward and the weights u_t converge to values that offer an approximation to the desired differential value function.

13.6.2 Independent actors

An actor is a parameterized class $\{\phi_\theta | \theta \in \mathbb{R}^l\}$ of policies. If it were possible to compute gradients $\nabla_\theta \overline{r}_{\phi_\theta}$ of the performance with respect to parameter settings, one could improve performance via a gradient method. However, because the space of decisions is often discrete the gradient is not generally well-defined.

It is convenient to expand the class of policies under consideration to include those that select decisions randomly. By doing this, the discrete space of decisions can effectively be transformed into a continuous one. In particular, let us define a *randomized stationary policy* to be a function $\pi : \mathbb{A} \times \mathbb{X} \mapsto [0, 1]$ with $\sum_{a \in \mathbb{A}} \pi(a|x) = 1$ for all $x \in \mathbb{X}$. Each $\pi(a|x)$ represents the probability with which decision a is selected when at state x.

Consider now a parameterized class $\{\pi_\theta | \theta \in \mathbb{R}^l\}$ of randomized policies for which the probabilities $\pi_\theta(a|x)$ are continuously differentiable functions of θ. Each randomized stationary policy π generates an average reward $\overline{r}_\pi$, defined by

$$\overline{r}_\pi = \lim_{N \to \infty} \frac{1}{N} \mathbb{E}\left[\sum_{t=0}^{N} r(x_t, a_t)\right],$$

where the system is controlled by decisions a_t sampled at each time step according to probabilities $\pi_\theta(\cdot|x_t)$. It is well-known that there exists a deterministic stationary policy that attains the optimal reward, and hence, $\max_\pi \overline{r}_\pi = r^*$, even when the maximum is taken over all randomized policies. We define a differential value function $h(\cdot, \pi)$ associated with each randomized stationary policy by letting

$$h(x, \pi) = \lim_{N \to \infty} \mathbb{E}\left[\sum_{t=0}^{N} \left(r(x_t, a_t) - \overline{r}_\pi\right) \middle| x_0 = x\right].$$

It is also convenient to define Q values associated with each randomized stationary policy:

$$q(x,a,\pi) = \mathop{\mathbb{E}}_{w}\left[r(x,a) - \overline{r} + h(f(x,a,w),\pi)\right].$$

If $\pi_\theta(a|x) > 0$ for all x and a, one can imagine employing a steepest ascent method of the form

$$\theta_{m+1} = \theta_m + \gamma_m \nabla_\theta \overline{r}_{\pi_{\theta_m}}.$$

Unfortunately, obtaining the gradient $\nabla_\theta \overline{r}_{\pi_\theta}$ poses a computational challenge. Monte Carlo simulation techniques, however, can generate estimates, which in turn can be employed in a stochastic gradient iteration. Such an iteration adapts the parameters according to

$$\theta_{m+1} = \theta_m + \gamma_m \chi_m,$$

where each χ_m is an estimate of the gradient given parameters θ_m. Gradient estimation techniques have been studied extensively in the infinitesimal perturbation analysis literature [15, 26].

Let us discuss one stochastic gradient method, which was recently proposed by Marbach and Tsitsiklis [37, 36]. The algorithm simulates a single endless trajectory of the system, updating the parameters of an actor upon visits to some distinguished state $\overline{x} \in \mathbb{X}$. Let t_m be the m^{th} time at which state $\overline{x}$ is visited. For times $t = t_m, \ldots, t_{m+1} - 1$, controls a_t are sampled at each state x_t according to probabilities $\pi_{\theta_m}(\cdot|x_t)$. The algorithm generates a sequence of estimates to the average reward according to

$$\overline{r}_{m+1} = \overline{r}_m + \gamma_m \sum_{t=t_m}^{t_{m+1}-1} (r(x_t,a_t) - \overline{r}_m).$$

Each gradient estimate is given by

$$\chi_m = \sum_{t=t_m}^{t_{m+1}-1} \hat{q}_t \frac{\nabla_\theta \pi_{\theta_m}(a_t|x_t)}{\pi_{\theta_m}(a_t|x_t)},$$

where

$$\hat{q}_t = \sum_{\tau=t}^{t_{m+1}-1} (r(x_\tau,a_\tau) - \overline{r}_m).$$

It is shown in [37, 36] that under appropriate technical conditions, the average rewards $\overline{r}_{\pi_{\theta_m}}$ associated with the sequence of parameter vectors θ_m converges, and that

$$\lim_{m\to\infty} \nabla_\theta \overline{r}_{\pi_{\theta_m}} = 0,$$

with probability one. Hence, one should expect the asymptotic behavior of this stochastic gradient method to mimic that of its deterministic counterpart.

To provide some motivation for the structure of this algorithm and also to set the stage for integration of actors and critics in the next section, let us

discuss the algorithm's relation to a characterization of the gradient provided in [37, 36]:

$$\nabla_\theta \overline{r}_{\pi_\theta} = \sum_{x \in \mathbb{X}, a \in \mathbb{A}} \eta_\theta(x,a) q(x,a,\pi_\theta) \frac{\nabla_\theta \pi_\theta(a|x)}{\pi_\theta(a|x)},$$

where $\eta_\theta(x,a) = \mu_\theta(x)\pi_\theta(a|x)$ and $\mu_\theta(x)$ denotes the steady-state probability of state x when the system is controlled by the stationary policy π_θ. (Similar characterizations have also been employed in earlier work [14, 22, 30].) Note that if we have access to $q(x,a,\pi_\theta)$ for every θ, x, and a, we could generate noisy estimates $\chi_m^\dagger$ according to

$$\chi_m^\dagger = \sum_{t=t_m}^{t_{m+1}-1} q(x_t, a_t, \pi_\theta) \frac{\nabla_\theta \pi_{\theta_m}(a_t|x_t)}{\pi_{\theta_m}(a_t|x_t)},$$

and it would turn out that

$$\mathbb{E}[\chi_m^\dagger] = \nabla_\theta \overline{r}_{\pi_{\theta_m}}.$$

This is a consequence of the fact that state-decision probabilities $\eta_{\theta_m}(x,a)$ are equal to the relative frequencies with which state-decision pairs are sampled during a trajectory $x_{t_m}, x_{t_m+1}, \ldots, x_{t_{m+1}-1}$.

Since we do not have access to the desired values $q(x_t, a_t, \pi_{\theta_m})$, an estimate $\hat{q}_t$ is employed in computing χ_m. To see why $\hat{q}_t$ may constitute a suitable estimate, note that if $\overline{r}_m = \overline{r}_{\theta_m}$, we have

$$\mathbb{E}\left[\hat{q}_t | x_t, a_t\right] = q(x_t, a_t, \pi_{\theta_m}) - h(\overline{x}, \pi_{\theta_m}).$$

It turns out that the constant term $h(\overline{x}, \pi_{\theta_m})$ bears no consequence on computation of the gradient, because for any θ,

$$\begin{aligned}
\sum_{x \in \mathbb{X}, u \in \mathbb{A}} \eta_\theta(x,a) \frac{\nabla_\theta \pi_\theta(a|x)}{\pi_\theta(a|x)} &= \sum_{x \in \mathbb{X}, a \in \mathbb{A}} \mu_\theta(x) \nabla_\theta \pi_\theta(a|x) \\
&= \sum_{x \in \mathbb{X}} \mu_\theta(x) \nabla_\theta \left(\sum_{a \in \mathbb{A}} \pi_\theta(a|x) \right) \\
&= \sum_{x \in \mathbb{X}} \mu_\theta(x) \nabla_\theta \, (1) \\
&= 0,
\end{aligned}$$

and therefore

$$\nabla_\theta \overline{r}_{\pi_\theta} = \sum_{x \in \mathbb{X}, a \in \mathbb{A}} \eta_{\theta_m}(x,a) (q(x_t, a_t, \pi_\theta) - h(\overline{x}, \pi_{\theta_m})) \frac{\nabla_\theta \pi_{\theta_m}(a|x)}{\pi_{\theta_m}(a|x)}.$$

Hence,

$$\chi_m = \sum_{t=t_m}^{t_{m+1}-1} \hat{q}_t \frac{\nabla_\theta \pi_{\theta_m}(a_t|x_t)}{\pi_{\theta_m}(a_t|x_t)},$$

serves as a noisy estimate of $q(x_t, a_t, \pi_{\theta_m})$ for the purpose of gradient estimation.

13.6.3 Using critic feedback

The stochastic gradient algorithm described in the previous section made use of estimates in place of $q(x, a, \pi)$. Each estimate was generated based on a single sample trajectory from the state x to a distinguished state $\overline{x}$. High variance associated with such estimates can impede progress in stochastic gradient algorithms. In some sense, being based solely on a single sample, the estimate should not be expected to provide a close approximation to the expectation.

One possible motivation for the introduction of a critic in conjunction with an actor is as a mechanism for variance reduction in stochastic gradient algorithms. In particular, if a critic is able to offer accurate approximations of $q(x, a, \pi)$, their use in place of single-sample estimates, may dramatically accelerate stochastic gradient methods.

As a concrete example, let us introduce an actor-critic algorithm that is similar in spirit to those proposed in [32, 52]. The algorithm involves an actor π_θ and a critic that approximates Q-functions via a parameterization of the form

$$\tilde{q}(x, a, u) = \sum_{k=1}^{K} u(k)\psi_k(x, a).$$

Parameters θ_t and u_t and an average reward estimate $\overline{r}_t$ are adapted during simulation of the system. At each time t, a decision a_t is sampled according to probabilities $\pi(\cdot|x_t)$, and the weight vector u_t is updated by temporal-difference learning. In particular, defining a temporal difference by

$$d_t = r(x_t, a_t) - \overline{r}_t + \tilde{q}(x_{t+1}, a_{t+1}, u_t) - \tilde{q}(x_t, a_t, u_t),$$

the weights at the next time step are given by

$$u_{t+1} = u_t + \gamma_t d_t z_t,$$

with the eligibility vector z_t defined by

$$z_{t+1} = \sum_{\tau=0}^{t} \lambda^{t-\tau}\psi(x_\tau, a_\tau).$$

The average reward estimate follows

$$\overline{r}_{t+1} = (1 - \gamma_t)\overline{r}_t + \gamma_t r(x_t, a_t).$$

At the same time, to improve performance, the actor's parameters are adjusted according to

$$\theta_{t+1} = \theta_t + \nu_t \chi_t,$$

where ν_t is a step size (possibly different in value from γ_t) and the noisy estimate χ_t of the gradient is given by

$$\chi_t = \tilde{q}(x_t, a_t, u_t)\frac{\nabla_\theta \pi_{\theta_t}(a|x)}{\pi_{\theta_t}(a|x)}.$$

One interpretation of this algorithm involves viewing the parameters of the critic as evolving much faster, and therefore converging faster, than those of

the actor. In practice, this is achieved by keeping the step sizes ν_t of the actor extremely small relative to β_t, the step sizes of the critic. In the extreme case, one might imagine that the parameters of the critic converge so fast that at every point in time $\tilde{q}(x_t, a_t, u_t)$ looks to the actor as though it has already converged to an approximation of $q(\cdot, \cdot, \pi_{\theta_t})$. We would then have

$$\nabla_\theta \overline{r}_{\pi_\theta} \approx \sum_{x \in \mathbb{X}, a \in \mathbb{A}} \eta(x, a) \tilde{q}(x_t, a_t, u_t) \frac{\nabla_\theta \pi_\theta(a|x)}{\pi_\theta(a|x)},$$

and estimates of the Q values would no longer be noisy, possibly leading to a significant reduction in variance of gradient estimates.

In general, the reduction in variance brought about by using $\tilde{q}(x_t, a_t, u_t)$ in place of a single sample estimate as was done in the previous section may come at a cost incurred by bias in the estimate. In particular, if $q(\cdot, \cdot, \pi_{\theta_t})$ is not in the span of the basis functions $\psi_1, \ldots, \psi_K$, we should not expect to generate a good approximation. As discussed earlier in the context of critic-only methods, one might try to select basis functions based on engineering insights. In the context of the actor-critic algorithm we have described, however, there is another approach that leads to appropriate basis functions [32, 52]. To understand this approach, note that each component of the gradient

$$(\nabla_\theta \overline{r}_{\pi_\theta})_k = \sum_{x \in \mathbb{X}, u \in \mathbb{A}} \eta_\theta(x, a) q(x, a, \pi_\theta) \frac{\partial \pi_\theta(a|x)/\partial \theta_i}{\pi_\theta(a|x)},$$

can be interpreted as an inner product between functions $q(\cdot, \cdot, \pi_\theta)$ and

$$\frac{\partial \pi_\theta(\cdot|\cdot)/\partial \theta_i}{\pi_\theta(\cdot|\cdot)},$$

where the inner product is defined by

$$\langle f, h \rangle = \sum_{x \in \mathbb{X}, u \in \mathbb{A}} \eta_\theta(x, a) f(x, a) h(x, a),$$

for any scalar functions f and h with domain $\mathbb{X} \times \mathbb{A}$. Let $K = l$ (recall that l is the dimension of θ), and suppose we select basis functions

$$\psi_k(x, a) = \frac{\partial \pi_\theta(a|x)/\partial \theta_i}{\pi_\theta(a|x)},$$

for $k = 1, \ldots, K$. Then, given a projection $\overline{q}$ of $q(\cdot, \cdot, \pi_\theta)$ onto the span of the basis functions, with projection defined with respect to the inner product space under consideration, we have

$$(\nabla_\theta \overline{r}_{\pi_\theta})_k = \left\langle q(\cdot, \cdot, \pi_\theta), \frac{\partial \pi_\theta(\cdot|\cdot)/\partial \theta_i}{\pi_\theta(\cdot|\cdot)} \right\rangle = \left\langle \overline{q}, \frac{\partial \pi_\theta(\cdot|\cdot)/\partial \theta_i}{\pi_\theta(\cdot|\cdot)} \right\rangle.$$

This—together with the fact that temporal-difference learning does indeed approximate such a projection [59, 62]—suggests that the selection of basis functions is an appropriate one.

The approach we have described for basis function selection is appealing because it is automated. That is, given an actor with a current policy π_{θ_t}, one can generate a small collection of basis functions that is sufficient for approximating $q(\cdot, \cdot, \pi_{\theta_t})$ to an extent that facilitates improvement of actor performance just as the actual function $q(\cdot, \cdot, \pi_{\theta_t})$ would. Note, however, that the selection of basis functions is contingent on the value of θ_t. As a consequence, the choice should probably change over time as θ_t evolves. Understanding the dynamics of actor-critic algorithms coupled with basis functions that change with the actor's evolving policy is an interesting open issue.

In our motivation of actor-critic algorithms, as well as existing analyses [32, 52], the critic is viewed as converging much faster than the actor. Essentially, the actor "waits" until the critic converges before computing the desired gradient. The potential speed-up brought about by the critic's evaluation hopefully reduces variance in gradient estimation, thereby speeding up the actor's convergence. However, it is not known whether the delays brought about by "waiting" for the critic to converge end up slowing down the actor's dynamics to an extent that negates potential improvements in gradient computation. A related research topic of interest involves understanding the dynamics of actor-critic systems where both actor and critic evolve on the same "time scale;" that is, where the actor does not "wait" for the critic to converge before attempting to compute gradients.

13.6.4 Historical notes

Actor-critic methods have as long a history as does temporal-difference learning, and their stories are intertwined. Some of the earliest research in artificial intelligence on reinforcement learning involved interacting actors and critics, in which critics adapt according to temporal-difference learning. This includes the work of Barto, Sutton, and Anderson [6, 48]. However, algorithms with some similar ingredients had been proposed earlier on in control theory [71]. In artificial intelligence, actor-critic models were inspired in part as models of brain activity, and their role in neuroscience has been explored by Barto, Houk, and Adams [3, 27].

On the technical side, Williams and Baird [70] developed some of the earliest theoretical results about deterministic variants of actor-critic methods that employ exhaustive representations both of policies and of value functions. Again in a context involving exhaustive representations, convergence of a stochastic simulation-based method where the actor and critic operate on separate time scales was established by Borkar and Konda [31].

13.7 CLOSING REMARKS

The "curse of dimensionality" can be viewed as the primary obstacle prohibiting effective solution methods for stochastic control problems. It is interesting to note that an analogous impediment arises in statistical regression. In particular, given an ability to collect data pairs of the form $(x, v(x))$, the problem of producing an accurate approximation $\tilde{v}$ to the underlying function v becomes computationally intractable as the dimension of the domain increases.

Similarly with the context of stochastic control, difficulties arise due to the curse of dimensionality. In the setting of statistical regression, a common approach to dealing with this limitation involves selecting a set of basis functions $\psi_1, \ldots, \psi_K$, collecting a set of input-output pairs $\{(x_1, v(x_1)), \ldots, (x_m, v(x_m))\}$, and using least-squares algorithm to compute weights $u(1), \ldots, u(K)$ that minimize

$$\sum_{i=1}^{m} \left(v(x_i) - \sum_{k=1}^{K} u(k)\psi_k(x_i) \right)^2.$$

The result is an approximation of the form

$$\tilde{v}(x) = \sum_{k=1}^{K} u(k)\psi_k(x).$$

Though there is no systematic and generally applicable method for choosing basis functions, a combination of intuition, analysis, guesswork, and experimentation often leads to a useful selection. In fact, the combination of basis function selection and least-squares is a valuable tool that has met prevalent application.

The utility of least-squares statistical regression provides inspiration for neuro-dynamic programming. In particular, neuro-dynamic programming algorithms can be viewed as analogs to least-squares algorithms that are applicable to stochastic control rather than statistical regression. Given a stochastic control problem and a parameterized representation of the value function and/or policy, the intent is to compute parameters that lead to an effective approximation.

Though existing results provide a starting point, the development of streamlined methods and analyses applicable to general classes of stochastic control problems remains largely open. Our hope, however, is that the range of problems that can be addressed in such a manner will broaden with future research. A goal might be to eventually produce neuro-dynamic programming algorithms that are as useful and widely accessible in the context of stochastic control as is least-squares in the context of statistical regression.

Our brief account of ongoing research in neuro-dynamic programming is by no means exhaustive. We have chosen to focus discussion on temporal-difference learning and actor-critic methods, two research directions for which substantive results have been generated in recent years and in which many problems remain open. Among other interesting areas of neuro-dynamic programming research is the study of how problem structure should influence choices of parameterized value functions and/or policies. One current thrust here, for example, aims at exploiting hierarchical structure of complex decision problems in defining abstractions, which often can be viewed as parameterizations of value functions and/or policies (see, e.g. [21, 42, 38] and references therein). Researchers have also dedicated effort towards extending algorithms and results to encompass alternative dynamic programming models such as those with risk-sensitive optimality criteria [12] and continuous state spaces [13]. We refer the reader to books by Bertsekas and Tsitsiklis [10] and Sutton and Barto [51] for excellent coverage of many additional topics.

Acknowledgments

The author's understanding of neuro-dynamic programming was developed over several years of work with his dissertation advisor John Tsitsiklis. He has also enjoyed collaborations with Dimitri Bertsekas and Daniela Pucci de Farias on topics in this field. The author thanks Eugene Feinberg and an anonymous reviewer for useful comments on an earlier draft of this chapter.

References

[1] J. Abounadi, *Stochastic Approximation for Non-Expansive Maps: Application to Q-Learning Algorithms.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1998.

[2] J. Abounadi, D. P. Bertsekas, and V. S. Borkar, Learning Algorithms for Markov Decision Processes. Technical Report LIDS-P-2434, MIT Laboratory for Information and Decision Systems, 1998.

[3] A. G. Barto, Adaptive Critics and the Basal Ganglia. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 215–232, Cambridge, MA, 1995. MIT Press.

[4] A. G. Barto, S. J. Bradtke, and S. P. Singh, Real-Time Learning and Control Using Asynchronous Dynamic Programming. *Artificial Intelligence*, 72:81–138, 1995.

[5] A. G. Barto and R. S. Sutton, Simulation of Anticipatory Responses in Classical Conditioning by a Neuron-Like Adaptive Element. *Behavioural Brain Research*, 4:221–235, 1982.

[6] A. G. Barto, R. S. Sutton, and C. W. Anderson, Neuron-Like Elements That Can Solve Difficult Learning Control Problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:835–846, 1983.

[7] R. E. Bellman and S. E. Dreyfus, Functional Approximation and Dynamic Programming. *Math. Tables and Other Aids Comp.*, 13:247–251, 1959.

[8] D. P. Bertsekas, *Dynamic Programming and Optimal Control.* Athena Scientific, Bellmont, MA, 1995.

[9] D. P. Bertsekas and S. Ioffe, Temporal Differences-Based Policy Iteration and Applications in Neuro-Dynamic Programming. Technical Report LIDS–P–2349, MIT Laboratory for Information and Decision Systems, 1996.

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

[11] M. J. Brennan, E. S. Shwartz, and R. Lagnado, Strategic Asset Allocation. *Journal of Economic Dynamics and Control*, 21:1377–1403, 1997.

[12] V. S. Borkar, Q-Learning for Risk Sensitive Control. Submitted to *Mathematics of Operations Research*, 2000.

[13] V. S. Borkar, A Learning Algorithm for Discrete Time Stochastic Control. *Probability in Engineering and Informational Sciences*, 14(2):243–248, 2000.

[14] X. R. Cao and H. F. Chen, Perturbation Realization, Potentials, and Sensitivity Analysis of Markov Processes. *IEEE Transactions on Automatic Control*, 42:1382–1393, 1997.

[15] E. K. P. Chong and P. J. Ramadge, Stochastic Optimization of Regenerative Systems Using Infinitesimal Perturbation Analysis. *IEEE Transactions on Automatic Control*, 39:1400–1410, 1994.

[16] R. H. Crites, *Large-Scale Dynamic Optimization Using Teams of Reinforcement Learning Agents.* PhD thesis, University of Massachusetts, Amherst, MA, 1996.

[17] R. H. Crites and A. G. Barto, Improving Elevator Performance Using Reinforcement Learning. In *Advances in Neural Information Processing Systems 8*, Cambridge, MA, 1995. MIT Press.

[18] P. D. Dayan, The Convergence of TD(λ) for General λ. *Machine Learning*, 8:341–362, 1992.

[19] P. D. Dayan and T. J. Sejnowski, TD(λ) Converges with Probability 1. *Machine Learning*, 14:295–301, 1994.

[20] D. P. de Farias and B. Van Roy, On the Existence of Fixed Points for Approximate Value Iteration and Temporal-Difference Learning. *Journal of Optimization Theory and Applications*, vol. 105. no. 3, June, 2000.

[21] T. G. Dietterich, Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition. to appear in the Journal of Artificial Intelligence Research, 1997.

[22] M. C. Fu and J. Hu, Smoothed Perturbation Analysis Derivative Estimation for Markov Chains. *Operations Research Letters*, 15:241–251, 1994.

[23] P. W. Glynn, Stochastic Approximation for Monte-Carlo Optimization. In *Proceedings of the 1986 Winter Simulation Conference*, pages 285–289, 1986.

[24] G. J. Gordon, Stable Function Approximation in Dynamic Programming. Technical Report CMU-CS-95-103, Carnegie Mellon University, 1995.

[25] L. Gurvits, L. J. Lin, and S. J. Hanson, Incremental Learning of Evaluation Functions for Absorbing Markov Chains: New Methods and Theorems. unpublished manuscript, 1994.

[26] Y. C. Ho and X. R. Cao, *Perturbation Analysis of Discrete Event Systems*, volume 60. Kluwer Academic Publisher, Boston, MA, 1991.

[27] J. C. Houk, J. L. Adams, and A. G. Barto, A Model of how the Basal Ganglia Generates and Uses Neural Signals that Predict Reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 249–270, Cambridge, MA, 1995. MIT Press.

[28] R. Howard, *Dynamic Programming and Markov Processes.* M.I.T. Press, Cambridge, MA, 1960.

[29] T. Jaakkola, M. I. Jordan, and S. P. Singh, On the Convergence of Stochastic Iterative Dynamic Programming Algorithms. *Neural Computation*, 6:1185–1201, 1994.

[30] T. Jaakkola, S. P. Singh, and M. I. Jordan, Reinforcement Learning Algorithms for Partially Observable Markov Decision Problems. In *Advances in Neural Information Processing Systems 7*, pages 345–352, San Francisco, CA, 1995. Morgan Kauffman.

[31] V. R. Konda and V. S. Borkar, Actor-Critic Like Learning Algorithms for Markov Decision Problems. *SIAM Journal of Control and Optimization*, 38(1):94–123, 1999.

[32] V. R. Konda and J. N. Tsitsiklis, Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA, 2000. MIT Press.

[33] P. R. Kumar, Re-Entrant Lines. *Queueing Systems: Theory and Applications*, 13:87–110, 1993.

[34] H. L. Lee and C. Billington, Material Management in Decentralized Supply Chains. *Operations Research*, 41(5):835–847, 1993.

[35] S. Mahadevan, Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results. *Machine Learning*, 22:1–38, 1996.

[36] P. Marbach, *Simulation-Based Optimization of Markov Reward Processes*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1998.

[37] P. Marbach and J.N. Tsitsiklis, Simulation-Based Optimization of Markov Reward Processes. To appear in the *IEEE Transactions on Automatic Control*, 1998.

[38] A. McGovern, D. Precup, S. P. Singh, and R. S. Sutton, Hierarchical Optimal Control of MDPs. In *Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems*, pages 186–191, 1998.

[39] R. C. Merton, *Continuous-Time Finance*. Basil Blackwell, Oxford, UK, 1992.

[40] W. D. Nordhaus, *Managing the Global Commons: the Economics of Climate Change*. MIT Press, Cambridge, MA, 1994.

[41] C. H. Papadimitriou and J. N. Tsitsiklis, The Complexity of Optimal Queueing Network Control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

[42] R. E. Parr, *Hierarchical Control and Learning for Markov Decision Processes*. PhD thesis, University of California, Berkeley, Berkeley, CA, 1998.

[43] R. E. Schapire and M. K. Warmuth, On the Worst-Case Analysis of Temporal-Difference Learning Algorithms. *Machine Learning*, 22:95–122, 1996.

[44] A. Schwartz, A Reinforcement Learning Method for Maximizing Undiscounted Rewards. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 298–305, 1993.

[45] S. P. Singh, Reinforcement Learning Algorithms for Average Payoff Markovian Decision Processes. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 202–207, 1994.

[46] S. P. Singh and D. P. Bertsekas, Reinforcement Learning for Dynamic Channel Allocation in Cellular Telephone Systems. In *Advances in Neural Information Processing Systems 10*, Cambridge, MA, 1997. MIT Press.

[47] S. P. Singh, T. Jaakkola, and M. I. Jordan, Reinforcement Learning with Soft State Aggregation. In *Advances in Neural Information Processing Systems 7*, Cambridge, MA, 1994. MIT Press.

[48] R. S. Sutton, *Temporal Credic Assignment in Reinforcement Learning.* PhD thesis, University of Massachusetts, Amherst, Amherst, MA, 1984.

[49] R. S. Sutton, Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 3:9–44, 1988.

[50] R. S. Sutton and A. G. Barto, Toward a Modern Theory of Adaptive Networks: Expectation and Prediction. *Psychological Review*, 88:135–170, 1981.

[51] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[52] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA, 2000. MIT Press.

[53] P. Tadepalli and D. Ok, Model-Based Average Reward Reinforcement Learning. *Artificial Intelligence*, 100:177–224, 1998.

[54] G. J. Tesauro, Practical Issues in Temporal Difference Learning. *Machine Learning*, 8:257–277, 1992.

[55] G. J. Tesauro, TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. *Neural Computation*, 6(2):215–219, 1994.

[56] G. J. Tesauro, Temporal Difference Learning and TD-Gammon. *Communications of the ACM*, 38:58–68, 1995.

[57] J. N. Tsitsiklis, Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning*, 16:185–202, 1994.

[58] J. N. Tsitsiklis and B. Van Roy, Feature-Based Methods for Large Scale Dynamic Programming. *Machine Learning*, 22:59–94, 1996.

[59] J. N. Tsitsiklis and B. Van Roy, An Analysis of Temporal-Difference Learning with Function Approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[60] J. N. Tsitsiklis and B. Van Roy, Average-Cost Temporal-Difference Learning. *Automatica*, 35(11):1799–1808, 1999.

[61] J. N. Tsitsiklis and B. Van Roy, Optimal Stopping of Markov Processes: Hilbert Space Theory, Approximation Algorithms, and an Application to Pricing High-Dimensional Financial Derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840–1851, 1999.

[62] B. Van Roy, *Learning and Value Function Approximation in Complex Decision Processes.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1998.

[63] B. Van Roy, D. P. Bertsekas, Y. Lee, and J. N. Tsitsiklis, A Neuro-Dynamic Programming Approach to Retailer Inventory Management. In *Proceedings of the IEEE Transactions on Automatic Control*, 1997.

[64] C. J. C. H. Watkins, *Learning From Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, UK, 1989.

[65] C. J. C. H. Watkins and P. Dayan, Q-learning. *Machine Learning*, 8:279–292, 1992.

[66] P. J. Werbos, Building and Understanding Adaptive Systems: a Statistical/Numerical Approach to Factory Automation and Brain Research. *IEEE Transactions on Systems, Man, and Cybernetics*, 17:7–20, 1987.

[67] P. J. Werbos, Approximate Dynamic Programming for Real-Time Control and Neural Modeling. In D. A. White and D. A. Sofge, editors, *Handbook of Intelligent Control*, 1992.

[68] P. J. Werbos, Neurocontrol and Supervised Learning: An Overview and Evaluation. In D. A. White and D. A. Sofge, editors, *Handbook of Intelligent Control*, 1992.

[69] R. J. Williams, Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8:229–256, 1992.

[70] R. J. Williams and L. C. Baird, Analysis of Some Incremental Versions of Policy Iteration: First Steps Toward Understanding Actor-Critic Learning Systems. Technical Report NU–CCS–93–11, College of Computer science, Northeastern University, 1993.

[71] I. H. Witten, An Adaptive Optimal Controller for Discrete-Time Markov Environments. *Information and Control*, 34:286-295, 1977.

[72] W. Zhang and T. G. Dietterich, A Reinforcement Learning Approach to Job Shop Scheduling. In *Proceeding of the IJCAI*, 1995.

Benjamin Van Roy
Department of Management and Science Engineering
Stanford University
CA 94305–4026, USA
bvr@stanford.edu

14 MARKOV DECISION PROCESSES IN FINANCE AND DYNAMIC OPTIONS

Manfred Schäl

Dedicated to Professor Dr. Dr.h.c. Karl Hinderer
on the occasion of his seventieth birthday

Abstract: In this paper a discrete-time Markovian model for a financial market is chosen. The fundamental theorem of asset pricing relates the existence of a martingale measure to the no-arbitrage condition. It is explained how to prove the theorem by stochastic dynamic programming via portfolio optimization. The approach singles out certain martingale measures with additional interesting properties. Furthermore, it is shown how to use dynamic programming to study the smallest initial wealth x^* that allows for super-hedging a contingent claim by some dynamic portfolio. There, a joint property of the set of policies in a Markov decision model and the set of martingale measures is exploited. The approach extends to dynamic options which are introduced here and are generalizations of American options.

14.1 INTRODUCTION AND SUMMARY

In discrete time $n = 0, 1, \cdots, N$ a financial market is studied which is free of arbitrage opportunities but incomplete. In the market $1 + d$ assets can be traded. One of them with price process $\{B_n,\, 0 \le n \le N\}$ is called the bond or savings account and is assumed to be nonrisky. The other d assets are called stocks and are described by the d-dimensional price process $\{S_n,\, 0 \le n \le N\}$. An investor is considered whose attitude towards risk is specified in terms of a utility function U. A dynamic portfolio is specified by a policy ϕ. The investor's objective is to maximize the expected utility of the discounted terminal wealth $X_N^\phi(x)$ when starting with an initial wealth x .

Utility optimization is now a classical subject (Bertsekas [1], Hakansson [16]). The present paper is intended as a bridge between Markov decision processes (MDPs) and modern concepts of finance. It is assumed the reader knows some basic facts about MDPs, but knowledge about Mathematical Finance is not needed. In particular, we here are interested in using the no-arbitrage condition for the construction of a particular martingale measure by use of the optimal policy ϕ^*. A condition close to the no-arbitrage condition was already used by Hakansson [16]. Martingale measures are used for option pricing. The paper makes use of an approach how to base option pricing on the optimal solution ϕ^* to the portfolio optimization problem. This approach is also explained by Davis [4].

In a one-period model, the construction of a martingale measure by use of an optimization problem is easily explained. Let us assume $d = 1$. If the action (portfolio) $a \in \mathbb{R}$ is chosen then the discounted terminal wealth of the portfolio has the form $x + a \cdot R$ where R can be interpreted as a return. If $a = a^*$ is optimal then one has $\partial E[U(x+a\cdot R)]/\partial a = 0 = E[U'(x+a\cdot R)\cdot R]$ for $a = a^*$. Upon defining a new measure Q by $dQ = \text{const} \cdot U'(x + a^* \cdot R)dP$, one obtains $\int R\, dQ = E_Q[R] = 0$ which is the martingale property. Since the martingale property is a local one both in time and in space, martingale measures can be constructed by local optimization problems (see Rogers 1994) whereas we will apply global (dynamic) optimization. This has the advantage that the resulting martingale measures have interesting interpretations. The case where the utility function U is only defined for positive values is treated in [39, 40].

A second interesting problem concerns values x for the initial wealth that allow for super-hedging discounted contingent claims $\breve{X}$ by some policy ϕ, i.e. $X_N^{\phi}(x) \geq \breve{X}$. The smallest value x^* coincides with the maximal expectation of $\breve{X}$ under (equivalent) martingale measures. The proof can make use of dynamic programming by exploiting an analogy between the set of all policies in a stochastic dynamic programming model and the set of martingale measures. A similar problem can be considered for American options where an optimal stopping time has to be chosen. It is well-known that an optimal stopping problem can be considered as a special stochastic dynamic programming problem. Therefore it is natural from the point of view of Markov decision theory to generalize the concept of an American option. This is done in the present paper and the generalization is called a dynamic option which has interesting applications.

14.2 THE FINANCIAL MARKET

On the market an investor can observe the prices of $1+d$ securities at the dates $n = 0, 1, \cdots, N$ where N is the *time horizon*. One of the securities is a *bond* (or savings account) with *interest rates* r_n, $1 \leq n \leq N$. It is essential for the theory that the interest rates for borrowing and lending are assumed to be the same. The *bond price process* is defined by

$$B_n := (1 + r_1) \cdots (1 + r_n), 0 \leq n \leq N, \text{ where } B_0 = 1\,. \tag{14.1}$$

Here we assume that $\{B_n\}$ is a deterministic process. If $\{B_n\}$ is given as the initial term structure, then the interest rates r_n can be computed by (14.1).

We will have $r_n \geq 0$, but we only need that $1 + r_n > 0$. The other d securities are called stocks. The evolution of the *stock prices* will be modeled by a d-dimensional stochastic process $\{S_n, n = 0, 1, \cdots, N\}$ where S_0 is deterministic. There the components S_n^k of S_n, $1 \leq k \leq d$, are assumed to be positive. One may also think of a foreign exchange market where S_n^k is the exchange rate for a foreign currency. Besides the savings account, the investor has d accounts for different foreign exchanges. The value of these accounts as well as that of the savings account may be negative which can be interpreted as a loan.

The information about the market at time n including the observed stock prices will be represented by a Markov chain $(I_n, 0 \leq n \leq N)$ on some probability space (Ω, P) where I_n takes on values in some space E_n, $0 \leq n \leq N$, where $I_0 = i_0$ is a given constant and hence $E_0 = \{i_0\}$. In order to avoid integrability and measurability problems we assume here:

$$E_n \text{ is finite, } 1 \leq n \leq N, \text{ hence (w.l.o.g.) } \Omega \text{ is finite and } P[\{\omega\}] > 0, \omega \in \Omega\,. \tag{14.2}$$

For any vector-valued process $\{Z_n\}$, we define the backward increment by $\Delta Z_n := Z_n - Z_{n-1}$. Further, we write $\zeta^\top \cdot \xi$ for the inner product of $\xi, \zeta \in \mathbb{R}^d$.

The representation often becomes easier (see (14.6) below) if one considers the *discounted stock price process* $\breve{S}_n = (\breve{S}_n^1, \cdots, \breve{S}_n^d)$ defined by

$$\breve{S}_n^k := S_n^k / \, B_n, \qquad k = 1, \cdots, d, \qquad n = 0, \cdots, N. \tag{14.3}$$

The *relative risk process* $\{R_n = (R_n^1, \cdots, R_n^d),\ 1 \leq n \leq N\}$ (Karatzas & Kou [23]) is defined by

$$1 + R_n^k := 1 + \Delta \breve{S}_n^k / \breve{S}_{n-1}^k = \frac{1}{1 + r_n} \{1 + \Delta S_n^k / \, S_{n-1}^k\} \tag{14.4}$$

where $\{\Delta S_n^k / S_{n-1}^k,\ 1 \leq n \leq N\}$ is the *return process* corresponding to $\{S_n^k, 0 \leq n \leq N\}$ (Pliska [32] § 3.2). Then we get

$$\breve{S}_n^k = \breve{S}_{n-1}^k \cdot (1 + R_n^k) = S_0^k \cdot (1 + R_1^k) \cdots (1 + R_n^k)\,. \tag{14.5}$$

Since the investor can observe S_n and R_n, it is natural to assume that S_n and R_n are known if the history $I_0, \cdots, I_n$ is known [i.e. that S_n and R_n are adapted in the sense explained in section 3]. In order to get a Markovian structure, we assume that R_n is a function $R_n = \rho_n(I_{n-1}, I_n)$ for some function ρ_n on $E_{n-1} \times E_n$. Important examples are $I_n = S_n$ or $I_n = R_n$. The latter example $I_n = R_n$ is sometimes convenient for a model where $R_1, \cdots, R_N$ are independent random variables as in many papers on portfolio optimization. An example for independent random variables $R_1, \cdots, R_N$ is the so-called *Binomial model*; there one has $d = 1$ and $r_n = r$ (independent of n). It is defined by two numbers δ and u such that $0 < \delta < 1 + r < u$ and $S_n = (1 + r) \cdot (1 + R_n) \cdot S_{n-1}$ is either equal to $u \cdot S_{n-1}$ or $\delta \cdot S_{n-1}$. If we choose $I_n = R_n$, then we get $E_n = \{\iota/(1+r) - 1;\ \iota = \delta, u\}$ (which is independent of n). However, sometimes [when considering so-called contingent claims $X = f(S_N)$ in section 3] one should choose $I_n = S_n$ and then $E_n = \{u \cdot i,\ i \in E_{n-1}\} \cup \{\delta \cdot i,\ i \in E_{n-1}\}$. This example shows that it is useful to let E_n depend on n.

At each time n, the investor will obtain the market information $I_n = i$ and observe the discounted value x_n of his portfolio. Then he will decide about the amount ϕ_n^k invested in stock k during $(n, n+1]$, i.e. ϕ_n^k/S_n^k denotes the number of shares the investor holds during $(n, n+1]$. The decision may depend on the present information (i_n, x_n). A dynamic portfolio is here described by a (Markov) *policy* ϕ which is given by a sequence of functions $\phi = \{\phi_n, 0 \le n < N\}$ where ϕ_n is a mapping from $E_n \times \mathbb{R}$ to $\mathbb{R}^d$. In particular, one allows for negative amounts ϕ_n^k, i.e. one allows for short selling of stocks. In the case of a foreign exchange market one can think of a negative ϕ_n^k as a loan. Given the *initial wealth* x, the amount η_n invested in the bond in $[n, n+1)$ is then specified by ϕ according to the following *budget equation* where we write $\underset{\sim}{1}$ for the d-dimensional vector with every component equal to 1.

$$\begin{aligned}
&\eta_0 + \phi_0^\top \cdot \underset{\sim}{1} = x\,,\\
&\eta_n + \phi_n^\top \cdot \underset{\sim}{1} = \eta_{n-1} \cdot (1 + r_n) + \sum_{k=1}^{d} \phi_{n-1}^k \cdot S_n^k / S_{n-1}^k\\
&\qquad\qquad = (1 + r_n) \cdot \Big[\eta_{n-1} + \phi_{n-1}^\top \cdot (\underset{\sim}{1} + R_n)\Big], \quad 1 \le n < N\,. \qquad (14.6)
\end{aligned}$$

The investor can choose any new portfolio ϕ_n at time n for the stocks. This decision is then compensated by the savings account. In particular, this means that no funds are added to or withdrawn from the wealth of the portfolio at any time. All changes in wealth are due to capital gains (appreciation of stocks and interest from the bond). Then the policy is called *self-financing*.

Again it is convenient to work with the discounted wealth of the portfolio in place of the wealth itself. The *discounted wealth process* $\{X_n^\phi(x)\}$ is given through

$$\begin{aligned}
X_n^\phi(x) &:= \Big[\eta_n + \phi_n^\top \cdot \underset{\sim}{1}\Big]/B_n = [\eta_{n-1} + \phi_{n-1}^\top \cdot (\underset{\sim}{1} + R_n)] \;/\; B_{n-1}, 0 \le n < N,\\
X_N^\phi(x) &:= \Big[\eta_{N-1} + \phi_{N-1}^\top \cdot (\underset{\sim}{1} + R_N)\Big] \;/\; B_{N-1}\,. \qquad (14.7)
\end{aligned}$$

At time N there is no rebalancing of the portfolio. Obviously we have

$$X_n^\phi(x) = X_{n-1}^\phi(x) + \phi_{n-1}^\top \cdot R_n\,. \qquad (14.8)$$

It is also usual to describe a portfolio by the numbers $\xi_n^k := \phi_n^k/S_n^k$ of shares the investor holds. In the case where the wealth is positive it can be useful to describe a portfolio by the proportions $\pi_n^k := \phi_n^k/X_n^\phi(x)$ of the wealth which are invested in the stocks. But here it will be convenient to use the description of a portfolio by the amounts ϕ_n^k.

In the present approach we have $E_n \times \mathbb{R}$ as state space of the underlying Markov decision process and we tried to explain why it is useful to let E_n depend on n. It is well-known that one can transform the present non-stationary Markovian model to a stationary model by including the time-parameter in the state (Feinberg [9]); but we will stick to the non-stationary setting. The action space is $\mathbb{R}^d$ and all actions $a \in \mathbb{R}^d$ are admissible at each time and state. Given

the initial wealth x, the history $I_0, \cdots, I_n$, and the decisions $\phi_0, \cdots, \phi_{n-1}$, we know the discounted value $X_n^\phi(x) = x_n$ of the portfolio. Therefore, one could dispense with x_n as part of the state. However, the present choice (i_n, x_n) of the state makes the model Markovian. In the case (mostly considered in the literature) when $R_n = I_n$ and the $R_1, \cdots, R_N$ are independent, one can even dispense with i_n as part of the state and just choose x_n as the state at time n.

As underlying basic process we choose the process $\{I_n\}$ representing the informations about the market. There are transition probabilities $p_{ij}(n)$, $i \in E_{n-1}$, $j \in E_n$, where $p_{ij}(n)$ specifies the conditional probability $P[I_{n+1} = j | I_n = i]$ given the present value $I_n = i$ [and the past $(i_0, \cdots, i_{n-1})$]. It is a particular feature that the probability measure P describing the dynamics does not depend on the policy. This will be important when considering other artificial probability measures Q on Ω in the next section. As explained above, the state process for the Markov decision process is $\{(I_n, X_n^\phi(x))\}$. The distribution of the state process does depend on the policy as usual. Therefore, the measure P describing the dynamics of $\{I_n\}$ is here more appropriate than the policy depending (strategic) measure describing the dynamics of $\{(I_n, X_n^\phi(x))\}$ on the canonical space of trajectories of length N. If $\{I_n\}$ is a non-Markovian process, it can be transformed to the present non-stationary Markovian model by choosing the history space $E_0 \times \cdots \times E_n$ as new space $\tilde{E}_n$.

14.3 NO-ARBITRAGE AND MARTINGALE MEASURES

In order to use the important concept of martingales w.r.t. to the given information structure, we have to use conditional expectations $\zeta(i_1, \cdots, i_n) := E[Z | I_1 = i_1, \cdots, I_n = i_n]$ defined on $E_1 \times \cdots \times E_n$ and $E[Z|\sigma(I_1, \cdots, I_n)]$ defined on Ω for a random variable Z. As usual, we use the convention that $E[Z|\sigma(I_1, \cdots, I_n)](\omega) := \zeta(I_1(\omega), \cdots, I_n(\omega))$, $\omega \in \Omega$.

Then a real-valued stochastic process $\{Z_n, 0 \leq n \leq N\}$ is a *martingale* if Z_n is *adapted* to the information structure, i.e. Z_n is a function of $(I_1, \cdots, I_n)$ where Z_0 is a constant and if

$$E[\Delta Z_n | \sigma(I_1, \cdots, I_{n-1})] = 0, \qquad 1 \leq n \leq N\,.$$

It is necessary to consider further probability measures Q on Ω which are *equivalent* (to the given physical probability measure P); i.e. $Q[\{\omega\}] > 0$, $\omega \in \Omega$. We write $E_Q[Z] = \int Z\, dQ$ for the expectation of the random variable Z under Q whereas $E[Z]$ is the expectation of the random variable Z under P as usual.

Definition 14.1 *Q is called a* **martingale measure** *iff $\{\breve{S}_n\}$ forms a martingale under Q, i.e.*

$$E_Q\left[\Delta \breve{S}_n^k | \sigma(I_1, \cdots, I_{n-1})\right] = 0, \quad 1 \leq k \leq d, \quad 1 \leq n \leq N\,. \tag{14.9a}$$

Obviously (14.9a) is equivalent to

$$E_Q\left[R_n^k | \sigma(I_1, \cdots, I_{n-1})\right] = 0, \quad 1 \leq k \leq d, \quad 1 \leq n \leq N\,. \tag{14.9b}$$

Then we set

$$\mathfrak{Q} := \{Q; Q \text{ is an equivalent martingale measure}\}. \tag{14.10}$$

Equivalent martingale measures are used for the valuation of contingent claims (Harrison & Kreps [17]). A *contingent claim* is a random variable X that represents the time N payoff from a 'seller' to a 'buyer'. In most instances the random variable X can be taken to be some function of an underlying stock price, and so contingent claims are examples of what are called *derivative securities*. In the present framework, it is natural to assume that X is a function of I_N and we recall that we may choose $I_n = S_n$ as an example. A typical example is given by a *European call option*. At time $n = 0$ the buyer signs a contract which gives him the option to buy, at a specified time N (called the *maturity date* or *expiration date*), one share of stock 1 at a specified price χ, called the *exercise price*. At maturity, if the price S_N^1 of stock 1 is below the exercise price, the contract is worthless to the buyer; on the other hand, if $S_N^1 > \chi$, the buyer can exercise his option (i.e. buy one share at the preassigned price χ) and then sell the share immediately in the market for S_N^1. This contract is thus equivalent to a payment of $X = (S_N^1 - \chi)^+$ (the contingent claim) at maturity. Given the price dynamics of the securities, one tries to determine the prices of such a contingent claim X. A classical answer in the sense of Huygens and Bernoulli to the question of a *fair price* π for the option would rely on the concept of a fair game and would then be $\pi = E[X/B_N]$. There is a different answer in the present situation where the seller can hedge X by investing in the stocks. For example if S_N^1 is high, then the buyer's and the seller's gain is high.

A contingent claim X is said to be *attainable* if there exists a policy ϕ and some $x \in \mathbb{R}$ such that $\breve{X} = X_N^\phi(x)$ for $\breve{X} := X/B_N$. The corresponding policy ϕ is said to replicate X, since the (nondiscounted) wealth $B_N \cdot X_N^\phi(x)$ of the dynamic portfolio ϕ at time N replicates X. This is the case if $\chi = 0$, i.e. $X = S_N^1$, in the example above; then one can choose $x = S_0^1$ and $\phi_n^1 \equiv S_n^1$, $\phi_n^k \equiv 0$, $k \neq 1$, for all n. It is clear that for attainable contingent claims with $\breve{X} = X_N^\phi(x)$ a fair price is x. The buyer and the seller could instead have invested the wealth x in such a way as to replicate the payoff of the contingent claim. Now from the martingale property one immediately gets

$$E_Q[X_N^\phi(x)] = x \quad \text{for} \quad Q \in \mathfrak{Q}. \tag{14.11}$$

Since a price system should be a linear functional on the space of all contingent claims X, i.e. of all random variables X on Ω (Harrison & Kreps [17]), it becomes clear that

$$\pi_Q(X) := E_Q[\breve{X}] \quad \text{for} \quad Q \in \mathfrak{Q} \tag{14.12}$$

is a candidate for a fair price. The market is said to be *complete* if each contingent claim is attainable. An example is the Binomial model for the case $d = 1$. For complete markets option pricing is no problem. However, there are only very few examples of complete discrete-time markets (Harrison & Pliska [18], Jacod & Shiryaev [21]). In continuous time, the Black-Scholes model is

an important example of a complete market. The question of the existence of some $Q \in \mathfrak{Q}$ is strongly connected with the following *no-arbitrage condition*:

$$\text{for } 1 \le n \le N \text{ and any portfolio } \phi_{n-1} \text{ depending on } (I_1, \cdots, I_{n-1}): \quad \text{(NA)}$$
$$\phi_{n-1}^\top \cdot R_n \ge 0 \text{ implies } \phi_{n-1}^\top \cdot R_n = 0.$$

There (NA) means that if there is a chance that $\phi_{n-1}^\top \cdot R_n > 0$, then there is also a chance that $\phi_{n-1}^\top \cdot R_n < 0$. A portfolio ϕ_{n-1} with "$\phi_{n-1}^\top \cdot R_n \ge 0$" and "$\phi_{n-1}^\top \cdot R_n > 0$ for some ω" is called an *arbitrage opportunity*. This definition of arbitrage is standard in the literature where we here need not care about null sets. Such an arbitrage opportunity represents a riskless source of generating profit, strictly greater than the profit from the bond. In order to present (NA) in full generality, we used a non-Markovian dynamic portfolio ϕ. Now we can present one of the most important results for financial markets:

Theorem 14.1 (Fundamental Theorem of Asset Pricing) *There exists an equivalent martingale measure, i.e. $\mathfrak{Q}$ is not empty, if and only if the no-arbitrage condition (NA) holds.*

Proofs for general spaces Ω can be found in Dalang et al [3], Schachermayer [36], Rogers [33], Jacod & Shiryaev [21]. It is easy to see that (NA) is necessary for the existence of some $Q \in \mathfrak{Q}$; therefore we can give the proof here: Let be $Q \in \mathfrak{Q}$. From $\phi_{n-1}^\top \cdot R_n \ge 0$ we conclude that $\phi_{n-1}^\top \cdot R_n = 0 \;\; Q-$a.s. since $E_Q[\phi_{n-1}^\top \cdot R_n] = 0$. As Q is equivalent we have $\phi_{n-1}^\top \cdot R_n = 0$ everywhere. In this paper it will be proved in §6 by use of dynamic programming that (NA) is also sufficient for the existence of some $Q \in \mathfrak{Q}$. Since the martingale property is a local one both in time and in space, martingale measures can also be constructed by local optimization problems. This was done by Rogers [33] whereas we will apply global (dynamic) optimization. From now on we use the following assumption:

Assumption 14.1 *(NA) holds.*

Since in discrete time the assumption of completeness is a severe restriction, we are forced to consider general incomplete markets. From the so-called second fundamental theorem of asset pricing it is known (Harrison & Pliska [18], Jacod & Shiryaev [21]) that in incomplete markets with the (NA)-condition one has several choices of equivalent martingale measures (from the convex set $\mathfrak{Q}$). Thus in incomplete markets, no preference independent pricing of contingent claims is possible. The approach of this paper, which makes use of dynamic programming, has another interesting consequence about identifying a certain price. There option pricing is based on an optimal solution to the portfolio optimization problem. This approach is also explained by Davis [4]. Suppose a utility function U on $\mathbb{R} \times E_N$ is given where $U(x,i)$ is strictly concave and differentiable in x with partial derivative $U'(x,i)$. The investor's objective is to maximize the expected utility of terminal wealth and the maximum utility is given by

$$V_0(x) = \sup_\phi E\Big[U(X_N^\phi(x), I_N)\Big] = E\Big[U(X_N^{\phi^*}(x), I_N)\Big] \tag{14.13}$$

where ϕ^* is a solution to the optimization problem. In §6 it will be explained that it may be useful to let the utility depend on I_N. Consider a contingent claim X made available for trading with purchase price π. One can ask the question whether the maximum utility in (14.13) can be increased by the purchase (or short-selling) of an option described by X. In order to find a fair price $\hat{\pi}$ for X one can use the following argument: $\hat{\pi}$ is a fair price for the contingent claim if diverting a little of the investor's initial wealth into the option at time zero has a neutral effect on the investor's achievable utility. For $N = 1$ this concept was also explained by Merton [28, §7.7]. More precisely, consider the discounted terminal wealth at time N of an investor who follows the dynamic portfolio ϕ^* which is optimal for his initial wealth x and who then diverts the amount $\xi \in \mathbb{R}$ to purchase ξ/π shares of the option with price π and discounted contingent claim $\breve{X} = X/B_N$:

$$X_N^{\phi^*}(x-\xi) + \frac{\xi}{\pi} \cdot \breve{X} = x - \xi + X_N^{\phi^*}(0) + \frac{\xi}{\pi} \cdot \breve{X} = X_N^{\phi^*}(x) + \xi \cdot \left\{ \frac{1}{\pi} \breve{X} - 1 \right\} .$$

Then the expected utility is

$$g(\phi^*, \xi, \pi, x) = E\left[U(X_N^{\phi^*}(x) + \xi \cdot \left\{ \frac{1}{\pi} \breve{X} - 1 \right\}, I_N) \right] \tag{14.14}$$

where $\sup_\phi g(\phi, 0, \pi, x) = V_0(x) = g(\phi^*, 0, \pi, x)$. Then one obtains

$$\frac{\partial}{\partial \xi} g(\phi^*, \xi, \pi, x) \Big|_{\xi=0} = E\left[U'(X_N^{\phi^*}(x), I_N) \cdot \left\{ \frac{1}{\pi} \breve{X} - 1 \right\} \right] . \tag{14.15}$$

Further, $g(\phi^*, \xi, \pi, x)$ is concave in ξ and strictly concave except for the less interesting case where $\breve{X}$ agrees with π and hence X is attainable.

Now choose $\pi = \hat{\pi}$ such that $\frac{\partial}{\partial \xi} g(\phi^*, \xi, \hat{\pi}, x)|_{\xi=0} = 0$ and $g(\phi^*, \xi, \hat{\pi}, x)$ is hence maximal for $\xi = 0$. Thus $\hat{\pi}$ is not too high otherwise the investor would like to sell short the option; and $\hat{\pi}$ is not too low otherwise the investor would like to purchase the option. This leads to the formula

$$\hat{\pi} = E\Big[U'(X_N^{\phi^*}(x), I_N) \cdot \breve{X} \Big] \; / E\Big[U'(X_T^{\phi^*}(x), I_N) \Big] = \int \breve{X} dQ_x^U . \tag{14.16}$$

The relation (14.16) leads to the interesting result that in our situation the option pricing formula of Davis is given by the martingale measure Q_x^U which is constructed in §5. A similar result was found by Karatzas & Kou [23, Theorem 7.4], who studied a diffusion model extensively. The case where U is defined only for positive values is treated in [39, 40].

Again consider an option with discounted contingent claim $\breve{X} := X/B_N$. In the second part of the paper we are interested in

$$x^* := \inf\Big\{ x \in \mathbb{R}; X_N^\phi(x) \geq \breve{X} \;\; \text{for some } \phi \Big\} . \tag{14.17}$$

Thus we look for values x of the initial wealth that allow for *super-hedging* X by some dynamic portfolio ϕ. Then a duality result holds: The smallest

value x^* will be shown to coincide with the maximal expectation of $\breve{X}$ under equivalent martingale measures.

Theorem 14.2 $x^* = \sup_{Q \in \mathcal{Q}} E_Q[\breve{X}]$.

The quantity x^* is called upper price, hedging price, upper bound for the fair price, selling price, or arbitrage upper bound. It is known (see e.g. [38, 1.16]) that $E_Q[\breve{X}] < x^*$ for each $Q \in \mathfrak{Q}$ unless X is attainable. In view of Theorem 14.2, a price $\pi = E_Q[\breve{X}]$ (for some $Q \in \mathfrak{Q}$) thus offers no arbitrage opportunity to the seller in the sense that $X_N^\phi(\pi) \geq \breve{X}$, $X_N^\phi(\pi) \neq \breve{X}$ for some ϕ.

A quantity similar to x^* will also be considered for *American options* (Karatzas [22]) where an optimal stopping time can be chosen by the buyer. Moreover, a generalization of American options is introduced here which is called *dynamic option*. This is a natural generalization from the point of stochastic dynamic programming where it is known how to embed an optimal stopping problem in a Markov decision problem.

14.4 THE ONE-PERIOD MODEL

In this section we will study the case $N = 1$. We write $R := R_1$ and

$$\sum := \Big\{ R(\omega); \omega \in \Omega \Big\} \tag{14.18}$$

for the *support* of R. Furthermore, $\mathcal{L}$ is the smallest linear space in $\mathbb{R}^d$ containing Σ, i.e. $\mathcal{L}$ is the smallest linear space L in $\mathbb{R}^d$ such that $P[R \in L] = 1$. Then it is easy to show that (NA) is equivalent to:

$$\text{for all } a \in \mathcal{L} \setminus \{0\} : P[a^\top \cdot R < 0] > 0\,. \tag{NA}_1$$

We start from a utility function $U(x, i)$ depending on two variables and we use the following properties:

Assumption 14.2 $U \in \mathbb{F}_N$ *where* $\mathbb{F}_n := \{f : \mathbb{R} \times E_n \mapsto \mathbb{R}$ *such that* $x \mapsto f(x, i)$ *is strictly concave and differentiable with derivative* $f'(x, i)$*; furthermore,* $f'(-\infty, i) > 0$ *and* $f'(+\infty, i) \leq 0\}$.

From the concavity we know that $U'(x, i)$ is decreasing in x. In a one-period model, a policy ϕ specifies just one portfolio $a \in \mathbb{R}^d$. We obtain from (14.6) and (14.7) that then

$$X_1^\phi(x) := x + a^\top \cdot R\,. \tag{14.19}$$

We want to maximize the expected utility:

$$\begin{aligned} v(x, a) &:= E\Big[U(x + a^\top \cdot R, I_1)\Big] \text{for } a \in \mathbb{R}^d; \\ V(x) &:= \sup_{a \in \mathbb{R}^d} v(x, a) \text{ for } x \in \mathbb{R}^d\,. \end{aligned} \tag{14.20}$$

Lemma 14.1

(a) If Γ denotes the orthogonal projection on $\mathcal{L}$, then $a^\top \cdot R = (\Gamma a)^\top \cdot R$;

(b) $(x,a) \mapsto v(x,a)$ is continuous;

(c) for each $x \in \mathbb{R}$, $a \mapsto v(x,a)$ attains the maximum on $\mathbb{R}^d$ where the maximum point can be chosen in $\mathcal{L}$.

Proof. The parts (a) and (b) are obvious. For part (c) one can use the methods of Leland [27], Bertsekas [1] or Rogers [33]. By (a) we can restrict attention to $a \in \mathcal{L}$. We have

$$\begin{aligned}&\frac{1}{\lambda} \cdot \Big[v(x, \lambda \cdot a) - v(x,0)\Big] \\ &= E\Big[\, D(\lambda, a^\top \cdot R, x, I_1) \cdot \mathbf{1}_{\{a^\top \cdot R > 0\}}\Big] + E\Big[D(\lambda, a^\top \cdot R, x, I_1) \cdot \mathbf{1}_{\{a^\top \cdot R < 0\}}\Big]\end{aligned}$$

where $D(\lambda, y, x, i) := \frac{1}{\lambda} \cdot \{U(x + \lambda\, y, i) - U(x,i)\}$ is decreasing in λ both for $y > 0$ and for $y < 0$ because of the concavity of $U(\cdot, i)$. Further $D(\infty, y, x, i) := \lim_{\lambda \uparrow \infty} D(\lambda, y, x, i) \leq 0$ for $y > 0$ and $D(\infty, y, x, i) < 0$ for $y < 0$ by our assumption on U. Now, we get by use of the monotone convergence theorem:

$$\begin{aligned}&\lim_{\lambda \uparrow \infty} \frac{1}{\lambda} \cdot \Big[v(x, \lambda \cdot a) - v(x,0)\Big] \\ &= E\Big[\, D(\infty, a^\top \cdot R, x, I_1) \cdot \mathbf{1}_{\{a^\top \cdot R > 0\}}\Big] + E\Big[D(\infty, a^\top \cdot R, x, I_1) \cdot \mathbf{1}_{\{a^\top \cdot R < 0\}}\Big].\end{aligned}$$

By (NA) we have for $a \in \mathcal{L} \setminus \{0\} : P[a^\top \cdot R < 0] > 0$ and thus:

$$\lim_{\lambda \uparrow \infty} \frac{1}{\lambda} \cdot \Big[v(x, \lambda \cdot a) - v(x,0)\Big] < 0 \text{ for } a \in \mathcal{L} \setminus \{0\}\,.$$

In particular, we have $\lim_{\lambda \uparrow \infty} v(x, \lambda \cdot a) = -\infty$ for $a \in \mathcal{L} \setminus \{0\}$. Now the result follows (see Rockafellar [37, Theorems 27.1, 27.3], Bertsekas [1, Proposition 1], Rogers [36, Proposition 2.2]). ∎

Lemma 14.2

(a) $v(x,a)$ is strictly concave in x for fixed a;

(b) $V(x)$ is strictly concave in x and the maximum point of $a \mapsto v(x,a)$ in $\mathcal{L}$ is unique.

Proof. Part (a) is obvious. For the proof of (b) choose $x, \tilde{x} \in \mathbb{R}$ and $\lambda, \tilde{\lambda} > 0$ such that $\lambda + \tilde{\lambda} = 1$. Further choose $a \in \mathcal{L}$ such that $V(x) = v(x,a)$ and similarly $\tilde{a}$ for $\tilde{x}$. We consider the cases (i) $\tilde{x} > x$ and (ii) $\tilde{x} = x$ and $a \neq \tilde{a}$. Then $P[x + a^\top \cdot R \neq \tilde{x} + \tilde{a}^\top \cdot R] > 0$; otherwise $(a - \tilde{a})^\top \cdot R = \tilde{x} - x$ which contradicts (NA) in case (i) and contradicts $a - \tilde{a} \in \mathcal{L} \setminus \{0\}$ in case (ii). Now we obtain from the strict concavity of $U(\cdot, i) : \lambda \cdot V(x) + \tilde{\lambda} \cdot V(\tilde{x}) = \lambda \cdot v(x,a) + \tilde{\lambda} \cdot v(\tilde{x}, \tilde{a}) < v(\lambda \cdot x + \tilde{\lambda} \cdot \tilde{x}\ ,\ \lambda \cdot a + \tilde{\lambda} \cdot \tilde{a}) \leq V(\lambda \cdot x + \tilde{\lambda} \cdot \tilde{x})$. In case (ii) we have a contradiction. ∎

Lemma 14.3 *For $a \in \mathbb{R}^d$, $v(x,a)$ is differentiable in x with derivative*

$$v'(x,a) = E\Big[U'(x + a^\top \cdot R, I_1)\Big] .$$

The proof is obvious since Ω is finite.

Theorem 14.3 *$V(x)$ is differentiable in x and $V'(x) = v'(x,a^*)$ where $a^* = a^*(x)$ is the maximum point in $\mathcal{L}$ of the function $\mathbb{R}^d \ni a \mapsto v(x,a)$.*

Proof. From Lemmata 14.1 and 14.2 we know that a unique maximum point a^* exists in $\mathcal{L}$. Now $V(x \pm \epsilon) - V(x) \geq v(x \pm \epsilon, a^*) - v(x,a^*)$. Hence $v'_+(x,a^*) \leq V'_+(x) \leq V'_-(x) \leq v'_-(x,a^*)$ where $V'_\pm(x)$ and $v'_\pm(x,a^*)$ denote the right and left derivatives, respectively. Now Lemma 14.3 applies. ∎

Lemma 14.4 *For $x > 0$, $v(x,a)$ is partially differentiable in $a \in \mathbb{R}^d$ with partial derivatives*

$$\partial_k v(x,a) = E\Big[U'(x + a^\top \cdot R, I_1) \cdot R^k\Big] .$$

Again, the proof is obvious since Ω is finite.

Theorem 14.4 *Let be $x \in \mathbb{R}$, then $E[U'(x + {a^*}^\top \cdot R,\, I_1) \cdot R^k] = 0$ for $1 \leq k \leq d$ where a^* is the unique maximum point in $\mathcal{L}$ of the function $\mathbb{R}^d \ni a \mapsto v(x,a)$.*

Proof. The function $a \mapsto v(x,a)$ is partially differentiable by Lemma 14.4. Therefore we know for the maximum point a^* that $\partial_k v(x,a^*) = 0$ for $1 \leq k \leq d$. ∎

Corollary 14.1 *If $U'(x,i) > 0$ for all x and i or more generally if $U'(x + {a^*}^\top \cdot R,\, I_1) > 0$ on Ω, then the probability measure Q defined by $dQ = const \cdot U'(x + {a^*}^\top \cdot R,\, I_1) dP$ is an equivalent martingale measure.*

14.5 THE MULTI-PERIOD MODEL

We remind the reader that $R_n = \rho_n(I_{n-1}, I_n)$ for some function ρ_n. Now we need the support of R_n given $I_{n-1} = i$ defined by $\Sigma_n(i) := \{\rho_n(i,j);\ j \in E_n,\ p_{ij}(n) > 0\}$, $i \in E_{n-1}$, $n \geq 1$. $\mathcal{L}_n(i)$ is the smallest linear space in $\mathbb{R}^d$ containing $\Sigma_n(i)$, i.e. $\mathcal{L}_n(i)$ is the smallest linear space L in $\mathbb{R}^d$ such that $P[\, R_n \in L | I_{n-1} = i] = 1$. Then for Σ and $\mathcal{L}$ as defined in §4 we have $\Sigma = \Sigma_1(i_0)$ and $\mathcal{L} = \mathcal{L}_1(i_0)$.

We may assume (w.l.o.g.) that the no-arbitrage condition also holds locally (see Dalang et al. [3, Lemma 2.3], Pliska [32, (3.22)], Schäl [38, §2]), i.e.

$$\begin{aligned} & \text{“}a^\top \cdot \sigma \geq 0 \ \forall \sigma \in \Sigma_n(i)\text{”} \\ \text{implies} \quad & \text{“}a^\top \cdot \sigma = 0 \ \forall \sigma \in \Sigma_n(i)\text{”}, \quad i \in E_{n-1}, \quad a \in \mathbb{R}^d . \qquad \text{(NA)}^* \end{aligned}$$

The condition (NA)* just means:

$$\begin{aligned} & \text{“}P\Big[a^\top \cdot R_n \geq 0 | I_{n-1} = i\Big] = 1\text{”} \\ \text{implies} \quad & \text{“}P\Big[a^\top \cdot R_n = 0 | I_{n-1} = i\Big] = 1\text{”}, \quad i \in E_{n-1}, \quad a \in \mathbb{R}^d . \end{aligned}$$

As in §4, we will use the assumption $U \in \mathbb{F}_N$ for the utility function U. We recall that by (14.8): $X_N^{\phi}(x) = X_n^{\phi}(x) + \sum_{m=n+1}^{N} \phi_{m-1}^{\top} \cdot R_m$ and we now define:

$$\begin{aligned} v_n(x,i,\phi) &= E\left[U\left(x + \sum_{m=n+1}^{N} \phi_{m-1}^{\top} \cdot R_m, I_N\right) | I_n = i\right], \\ V_n(x,i) &= \sup_{\phi} v_n(x,i,\phi), \\ v_N(x,i,\phi) &= V_N(x,i) = U(x,i) \text{ for all } \phi. \end{aligned} \tag{14.21}$$

There $v_n(x,i,\phi)[V_n(x,i)]$ is the [maximal] expected utility of the terminal wealth given the market information i and the discounted wealth x at time n. Since $I_0 = i_0$ is fixed, we can set:

$$v_0(x,\phi) = v_0(x,i_0,\phi), \quad V_0(x) = V_0(x,i_0). \tag{14.22}$$

Obviously we have an N-stage Markov decision model with no running costs and terminal reward U. Let us now introduce the well-known *reward operator*:

$$\begin{aligned} T_n^a f(x,i) := E\Big[f(x + a^{\top} \cdot R_{n+1}, I_{n+1}) | I_n = i\Big] \\ \text{for any } f : \mathbb{R} \times E_{n+1} \mapsto \mathbb{R}, a \in \mathbb{R}^d, i \in E_n\,. \end{aligned} \tag{14.23}$$

We can express $T_n^a f(x,i)$ in (14.23) by the transition matrix $(p_{ij}(n))$ according to

$$T_n^a f(x,i) = \sum_{j \in E_{n+1}} p_{ij}(n) \cdot f\Big(x + a^{\top} \cdot \rho_{n+1}(i,j), j\Big) \tag{14.24}$$

Thus for fixed i, $T_n^a f(x,i)$ can be expressed by an ordinary expectation and we can use the results of §4. The following equation is sometimes called *fundamental equation* (Dynkin & Yushkevich [6]) and follows from Fubini's theorem (Hinderer [19, Lemma 11.1]).

$$v_n(x,i,\phi) = T_n^{\phi_n(x,i)} v_{n+1}(x,i,\phi), n \geq 0\,. \tag{14.25}$$

Now we can give the well-known *optimality equation* (see Hinderer 1970 [19, Theorems 14.4, 18.4]):

$$V_n(x,i) = \sup_{a \in \mathbb{R}^d} T_n^a V_{n+1}(x,i) = \sup_{a \in \mathcal{L}_{n+1}(i)} T_n^a V_{n+1}(x,i), n \geq 0\,. \tag{14.26}$$

Proposition 14.1 *Let be $f \in \mathbb{F}_{n+1}$ and $0 \leq n < N$. Then:*

(a) *$(x,a) \mapsto T_n^a f(x,i)$ is continuous;*

(b) *there exists a unique measurable function $\delta : \mathbb{R} \times E_n \mapsto \mathbb{R}^d$ such that $\delta(x,i) \in \mathcal{L}_{n+1}(i)$ and $T_n^{\delta(x,i)} f(x,i) = \max_{a \in \mathbb{R}} T_n^a f(x,i) =: F(x,i)$;*

(c) *$F(x,i)$ is strictly concave in x.*

Proof. Part (a) and part (b) follow from Lemma 14.1. The remaining properties follow from Lemma 14.2. ∎

Finally, we will use another assumption which will be discussed in §6 for special cases.

Assumption 14.3 *For* $V_n'(x,i) = \partial V_n(x,i)/\partial x$ *we have:* $V_n'(-\infty,i) > 0$ *and* $V_n'(+\infty,i) \le 0$, $0 \le n < N$.

The existence of $V_n'(x,i)$ follows from:

Lemma 14.5

(a) $V_n(x,i)$ *is differentiable in* x *with derivative* $V_n'(x,i)$.

(b) $V_n(x,i) \in \mathbb{F}_n$ *for* $0 \le n \le N$.

Proof. Part (a) can be proved as Theorem 14.3. By assumption we know that $V_T = U$ is in $\mathbb{F}_N$. Upon using the optimality equation (14.26), this statement follows by backward induction from Assumption 14.3 and Proposition 14.1c. ∎

Now we obtain from (14.25), (14.26), Proposition 14.1, and Lemma 14.5 by the usual arguments of dynamic programming (Feinberg [9]):

Theorem 14.5 *There exists a unique policy* $\phi^* = (\phi_n^*)$ *such that for* $0 \le n < N$*:*

$$\begin{aligned}
&\phi_n^*(x,i) \in \mathcal{L}_{n+1}(i) \text{ and} \\
&T_n^{\phi_n^*(x,i)} V_{n+1}(x,i) = \max_{a \in \mathbb{R}^d} T_n^a V_{n+1}(x,i) = V_n(x,i)\,, \\
&v_n(x,i,\phi^*) = V_n(x,i)\,.
\end{aligned}$$

Proof. We just have to define $\phi_n = \delta$ as in Proposition 14.1b where $f = V_{n+1}$. ∎

Proposition 14.2 *For some* $x > 0$, *let* ϕ^* *be the optimal policy of Theorem 14.5. Then*

$$V_0'(x) = E\Big[U'(X_N^{\phi^*}(x), I_N)\Big].$$

Proof. By induction we are going to prove

$$V_0'(x) = E\Big[V_n'(X_n^{\phi^*}(x), I_n)\Big], \quad 1 \le n \le N\,. \tag{14.27}$$

For $n = 1$ we have

$$V_0(x) = \max_{a \in \mathcal{L}_1(i_0)} E\Big[V_1(x + a^\top \cdot R_1, I_1)\Big]$$

where $\phi_0(x, i_0)$ is a maximum point.

In view of Lemma 14.5, $V_1(x,i)$ satisfies the Assumption 14.2 for $U(x,i)$ (for $N=1$). With the help of Lemma 14.3 and Theorem 14.3, we can conclude:

$$\begin{aligned} V_0'(x) = v_0'(x, \phi_0^*(x,i_0)) =& E\Big[V_1'(x + \phi_0^*(x,i_0)^\top \cdot R_1, I_1)\Big] \\ =& E\Big[V_1'(X_1^{\phi^*}(x), I_1)\Big]. \end{aligned}$$

Hence (14.27) holds for $n=1$. Now assume that (14.27) holds for $n-1$. For fixed $i \in E_{n-1}$ we have

$$V_{n-1}(x,i) = \max_{a \in \mathcal{L}_n(i)} E\Big[V_n(x + a^\top \cdot R_n, I_n) | I_{n-1} = i\Big]$$

where $\phi_{n-1}^*(x,i)$ is a maximum point. In view of Lemma 14.5, $V_n(x,i)$ satisfies the Assumption 14.2 for $U(x,i)$ (for $N=n$). As above we obtain:

$$V_{n-1}'(x,i) = E\Big[V_n'(x + \phi_{n-1}^*(x,i)^\top \cdot R_n, I_n) | I_{n-1} = i\Big].$$

Then by (14.8): $X_n^{\phi^*}(x) = X_{n-1}^{\phi^*}(x) + \phi_{n-1}^{*\top} \cdot R_n$

Since $X_{n-1}^{\phi^*}(x)$ is a function of $(I_1, \cdots, I_{n-1})$, it follows that

$$\begin{aligned} V_{n-1}'(X_{n-1}^{\phi^*}(x), I_{n-1}) &= E\Big[V_n'(X_{n-1}^{\phi^*} + \phi_{n-1}^*(x,i)^\top \cdot R_n, I_n) | \sigma(I_1, \cdots, I_{n-1})\Big] \\ &= E\Big[V_n'(X_n^{\phi^*}, I_n) | \sigma(I_1, \cdots, I_{n-1})\Big]. \end{aligned} \tag{14.28}$$

By assumption we get

$$\begin{aligned} V_0'(x) &= E\Big[V_{n-1}'(X_{n-1}^{\phi^*}(x), I_{n-1})\Big] \\ &= E\Big[E[V_n'(X_n^{\phi^*}, I_n) | \sigma(I_1, \cdots, I_{n-1})]\Big] = E\Big[V_n'(X_n^{\phi^*}(x)), I_n)\Big] \end{aligned}$$

and (14.27) follows. Now we can use (14.27) for $n=N$ and finally obtain the result. ∎

Theorem 14.6 *Assume that U' is positive or more generally that $U'(X_N^{\phi^*}(x), I_N)$ is positive where ϕ^* is the optimal policy of Theorem 14.5. Define for some x the process $\{Z_n,\ 0 \le n \le N\}$ by*

$Z_n := V_n'(X_n^{\phi^*}(x), I_n)$, *in particular* $Z_0 = V_0'(x)$, $Z_N = U'(X_N^{\phi^*}(x), I_N)$, *then:*

(a) One obtains an equivalent martingale measure Q_x^U on Ω by

$$Q_x^U[\{\omega\}] = \frac{1}{V_0'(x)} Z_N(\omega) P[\{\omega\}] = \Big(Z_N(\omega)/Z_0\Big) P[\{\omega\}].$$

(b) $\{Z_n/Z_0, 0 \le n \le N\}$ is a martingale under P and is called the density process of dQ_x^U/dP .

(c) If we write $E_x^U[\cdots]$ for the expectation under Q_x^U, then for any function f_n

$$\begin{aligned} &E_x^U\Big[f_n(I_1, \cdots, I_n) | \sigma(I_1, \cdots, I_{n-1})\Big] \\ &\qquad = E\Big[f_n(I_1, \cdots, I_n) \cdot Z_n | \sigma(I_1, \cdots, I_{n-1})\Big] / Z_{n-1}. \end{aligned}$$

Proof. If $U'(X_N^{\phi^*}(x), I_N)$ is positive then we know from Proposition 14.2 that Q_x^U is indeed a probability measure which is equivalent to P. Now part (b) immediately follows from (14.28) and part (c) is a consequence of Bayes' rule. From Theorem 14.4 we conclude that

$$E\Big[V_n'(x+\phi_{n-1}^*(x,i)^\top \cdot R_n, I_n) \cdot R_n^k | I_{n-1} = i\Big] = 0 \text{ for } 1 \le k \le d \qquad (14.29)$$

since $\phi_{n-1}^*(x,i)$ is the unique maximum point in $\mathcal{L}_n(i)$ of the function $\mathbb{R}^d \ni a \mapsto E[V_n'(x+a^\top \cdot R_n, I_n)|I_{n-1}=i]$. From (14.29) we obviously obtain as in proof of (14.28):

$$E\Big[R_n^k \cdot Z_n | \sigma(I_1, \cdots, I_{n-1})\Big] = 0, \quad 1 \le k \le d\,.$$

Finally from (c) we get

$$E_x^U\Big[R_n^k | \sigma(I_1, \cdots, I_{n-1})\Big] = 0, \quad 1 \le n \le N, \quad 1 \le k \le d\,. \qquad (14.30)$$

■

14.6 APPLICATIONS

In this section we will show that all assumptions are satisfied for the case where

$$U(x,i) = -e^{-\gamma x} \cdot L(i) \text{ for some } \gamma > 0 \text{ and some function } L : E_N \mapsto (0,\infty). \qquad (14.31)$$

A classical example is $U(x,i) = -\frac{1}{\gamma}e^{-\gamma x}$ (or $U(x,i) = \frac{1}{\gamma}(1-e^{-\gamma x})$). If one wants to consider the utility of the terminal wealth itself rather than the discounted terminal wealth, then one can choose $U_B(x,i) := U(B_N \cdot x, i)$ which has the same form (14.31). A further interesting example was studied by Grandits & Rheinländer [15]. They look for a policy ϕ^0 such that

$$E\Big[\exp\{\gamma \cdot (\breve{X} - X_N^{\phi^0}(x))\}\Big] = \inf_\phi\,. \qquad (14.32)$$

Instead of super-hedging the contingent claim X one tries to choose ϕ^0 for the given initial wealth x such that the expected loss is minimized. There the positive values of $\breve{X} - X_N^{\phi^0}(x)$ are given more weight than the negative values. Of course, this is the view of the seller. If one chooses $I_N = S_N$ and $\breve{X} = f(S_N)$ as in the case of a European call option, then this problem is included in the framework of (14.31) by choosing $L(i) := \frac{1}{\gamma}\exp\{\gamma \cdot f(i)\}$. There are some recent papers treating the problem $E[\ell(\{\breve{X} - X_N^{\phi^0}(x)\}^+)] = \inf_\phi$ (see Runggaldier & Zaccaria [35] for further references and an approach via dynamic programming in the case of transaction costs). For convenience, we only will consider the case $\gamma = 1$.

Lemma 14.6 *In the situation (14.31) with $\gamma = 1$, one has:*

(a) Assumption 14.2 is satisfied and $U'(x,i) = e^{-x} \cdot L(i)$ is positive;

(b) $V_n(x,i) = -e^{-x} \cdot L_n(i)$ *for some function* $L_n : E_n \mapsto (0,\infty)$*;*

(c) for the optimal policy ϕ^* *of Theorem 14.5,* $\phi_n^*(x,i)$ *does not depend on* x*;*

(d) Assumption 14.3 is satisfied.

Proof. Part (a) is obvious and part (d) follows from part (b). Part (b) is obvious for $n = N$. Now assume that (b) holds for $n+1$. Then according to Theorem 14.5

$$V_n(x,i) = \max_{a \in \mathcal{L}_{n+1}(i)} - \sum_{j \in E_{n+1}} p_{ij}(n) \exp\{-x - a^\top \cdot \rho_{n+1}(i,j)\} \cdot L_{n+1}(j) =: -e^{-x} \cdot L_n(i)$$

where the maximum point $\phi_n(i)$ is indeed independent of x and

$$L_n(i) = \min_{a \in \mathcal{L}_{n+1}(i)} \sum_{j \in E_{n+1}} p_{ij}(n) \exp\{-a^\top \cdot \rho_{n+1}(i,j)\} \cdot L_{n+1}(j)\,. \tag{14.33}$$

■

Consider the case where $R_1, \cdots, R_N$ are independent and $I_n = R_n$. Then both $p_{ij}(n)$ and $\rho_{n+1}(i,j)$ are independent of i. From the proof of Lemma 14.6 we conclude that in that case $\phi_n^*(x,i) =: \phi_n^*$ neither depends on x nor on i.

Definition 14.2 *In the situation (14.31) with* $\gamma = 1$*, let* ϕ^* *be the optimal policy of Theorem 14.5 which is independent of the initial wealth* x *according to Lemma 14.6. Then* $X_n^* := X_n^{\phi^*}(0)$*,* $n = 0, \cdots, N$ *is called the* **optimal wealth process**.

Lemma 14.7 *In the situation (14.31) with* $\gamma = 1$*, the density process of Theorem 14.6 is given by*

$$Z_n/Z_0 = \exp\{-X_n^*\} \cdot L_n(I_n)/L_0(i_0), 0 \le n \le N,$$

where

$$Z_N/Z_0 = \exp\{-X_N^*\} \cdot L(I_N)/L_0(i_0).$$

Proof. We have $Z_n = V_n'(x + X_n^*, I_n) = e^{-x} \cdot \exp\{-X_n^*\} \cdot L_n(I_n)$, thus $Z_0 = e^{-x} \cdot L_0(i_0)$ and $Z_N = e^{-x} \cdot \exp\{-X_N^*\} \cdot L(I_N)$. ■

Now we obtain from Theorem 14.6 the following result:

Theorem 14.7 *Let* $\{X_n^*\}$ *be the optimal wealth process in the situation (14.31) with* $\gamma = 1$ *and let* L_n *be defined by (14.33) where* $L_N = L$*. Then* Q^U *defined by* $Q^U[\{\omega\}] := \exp\{-X_N^*(\omega)\} \cdot L(I_N(\omega))/L_0(i_0) P[\{\omega\}]$ *is a martingale measure.*

Upon choosing $L(i) = 1$, for example, we obtain the existence of a special martingale measure and the Fundamental Theorem of §3 is proved where (NA) was used in the proof of Lemma 14.1c. Then the resulting martingale measure Q^E has the form:

$$Q^E[\{\omega\}] = \text{const} \cdot \exp\{-X_N^*(\omega)\} P[\{\omega\}]\,. \tag{14.34}$$

There Q^E is the unique solution of the minimum problem where the relative entropy $I(Q,P)$ of Q w.r.t. P has to be minimized (Fritelli [13], Grandits [14]). Q^E may also be considered as a multi-stage *Esscher transform* of P whereas in Bühlmann et al. [2] the Esscher transform is used locally for each time and each history in order to construct a martingale measure.

Another important example is the following:

$$U(x,i) = L^{(0)}(i) + 2\ L^{(1)}(i)\cdot x - L^{(2)}(i)\cdot x^2 \tag{14.35}$$

for some function $L^{(k)} : E_N \mapsto \mathbb{R}, k = 0,1,2$ where $L^{(2)}$ is positive.

An interesting special case is the best hedging policy for hedging the contingent claim X. One looks for a policy ϕ^o such that

$$E\Big[\{\breve{X} - X_N^{\phi^o}(x)\}^2\Big] = \inf_{\phi}\ . \tag{14.36}$$

Thus, instead of super-hedging X one tries to choose ϕ^o for the given initial wealth x such that the expected quadratic loss is minimized. Now, in contrast to (14.32), the positive values of $\breve{X} - X_N^{\phi^o}(x)$ have the same weight as the negative values. This is fair both from the point of view of the seller and the buyer. If one again chooses $I_N = S_N$ and $\breve{X} = f(S_N)$, then this problem is included in the framework of (14.35) by choosing $L^{(0)}(i) = f(i)^2$, $L^{(1)}(i) = f(i)$, and $L^{(2)}(i) := 1$.

Now, one can look for that initial wealth $\hat{x}$ such

$$\inf_{\phi} E\Big[\{\breve{X} - X_N^{\phi}(\hat{x})\}^2\Big] = \min_{x}\inf_{\phi} E\Big[\{\breve{X} - X_N^{\phi}(x)\}^2\Big]. \tag{14.37}$$

The optimal value $\hat{x}$ is called *fair hedging price* in [37] and *approximation price* by Schweizer [41] for the contingent claim X. Whereas [37] relies on stochastic dynamic programming, Schweizer [41] uses martingale methods.

Lemma 14.8 *In the situation (14.35) one has:*

(a) Assumption 14.2 is satisfied ;

(b) $V_n(x,i) = L_n^{(0)}(i) + 2\ L_n^{(1)}(i)\cdot x - L_n^{(2)}(i)\cdot x^2$ *for some function* $L_n^{(k)} : E_n \mapsto \mathbb{R}$, $k = 0,1,2$, *where* $L_n^{(2)}$ *is positive;*

(c) Assumption 14.3 is satisfied.

Proof. Part (a) is obvious and part (c) follows from part (b). Part (b) is obvious for $n = N$. Now assume that (b) holds for $n+1$. Then according to Theorem 14.5

$$\begin{aligned} V_n(x,i) &= \max_{a\in\mathbb{R}^d} \sum_{j\in E_{n+1}} p_{ij}(n+1)\Big[L_{n+1}^{(0)}(j) + 2\ L_{n+1}^{(1)}(j)\cdot\{x + a^\top\cdot\rho_{n+1}(i,j)\} \\ &\qquad - L_{n+1}^{(2)}(j)\{x + a^\top\cdot\rho_{n+1}(i,j)\}^2\Big] \\ &= \tilde{L}^{(0)}(i) + 2\tilde{L}^{(1)}(i)x - \tilde{L}^{(2)}(i)x^2 + L^*(x,i) - \hat{L}(x,i) \end{aligned}$$

where

$$\tilde{L}^{(k)}(i) := \sum_j p_{ij}(n+1) L^{(k)}_{n+1}(j), k = 0, 1, 2,$$

$$L^*(x,i) := -\min_{a \in \mathbb{R}^d} \sum_j p_{ij}(n+1) \cdot L^{(2)}_{n+1}(j) \{ L^{(1)}_{n+1}(j)/L^{(2)}_{n+1}(j) - x - a^\top \cdot \rho_{n+1}(i,j) \}^2,$$

$$\hat{L}(x,i) := \sum_j p_{ij}(n+1) L^{(2)}_{n+1}(j) \{ L^{(1)}_{n+1}(j)/L^{(2)}_{n+1}(j) - x \}^2.$$

Now $\hat{L}$ has the desired form. This is also true for L^* as follows from [37, Proposition 7.3] for the case $d = 1$ and is easily extended to the case $d \geq 1$. ∎

The problem with this example (14.35) is that $U'(x,i) = 2\, L^{(1)}(i) - 2\, L^{(2)}(i) \cdot x$ is not positive everywhere and also $U'(X_N^{\phi^*}(x), I_N)$ need not to be positive everywhere. Therefore the measure Q_x^U is in general only a signed measure and is called signed martingale measure because it has all other properties of a martingale measure. In Schweizer [41, 42] the relations to the *mean-variance frontier*, the *variance-optimal* and the *minimal martingale measures* are explained. Schweizer [42] also considers the general d-dimensional case in a general semi-martingale framework which includes the discrete-time case as special case. Motoczynski [29] studies the discrete-time d-dimensional setting. Grandits [14] constructs the *p-optimal martingale measure* as a generalization of the variance-optimal martingale measure by replacing L^2 by L^p for some $p > 1$. For the construction Grandits also uses methods of dynamic programming similar to those explained in § 5.

Example. The multi-dimensional Binomial model

For convenience we assume $T = 1$ and write $R^k := R^k_1$, $1 \leq k \leq d$. We consider the model where

$$R^k \text{ takes on values in } \{-\alpha^k, \beta^k\} \text{ where } \alpha^k, \beta^k > 0.$$

We assume w.l.o.g. that $I_1 = R_1 = R$. For $d = 1$ the model is complete and is also called Cox-Ross-Rubinstein model. The unique martingale measure P^* is given by $P^*[R^1 = i] = q_1^*(i)$ where

$$q_k^*(-\alpha^k) := \beta^k/(\alpha^k + \beta^k), \qquad q_k^*(\beta^k) := \alpha^k/(\alpha^k + \beta^k).$$

For $d > 1$, the model is no longer complete since R takes on 2^d values where $2^d > 1 + d$ (Jacod & Shiryaev [21]). It is easy to see that $Q^*[R = (i^1, \cdots, i^d)] = \Pi_{k=1}^d q_k^*(i^k)$ defines a martingale measure for $d \geq 1$ which seems to be a natural one if the components R^k of R are independent under P. In fact in that case this measure coincides with Q^E as defined in (14.34) but is different from the so-called *minimal martingale measure* which is obtained by the (discrete-time) *Girsanov transformation* and which coincides with the *variance-optimal martingale measure* for $T = 1$ (Schweizer [41]).

Q^* also differs from the martingale measure obtained by the so-called *numéraire portfolio* (Korn & Schäl [24]). We want to prove:

$$Q^* = Q^E \text{ if the components } R^k \text{of } R \text{ are independent under } P\,.$$

We first observe that $P^* = Q^* = Q^E$ if $d = 1$ since then there is only one martingale measure. Hence for $\inf_{a\in\mathbb{R}} E[\exp\{a \cdot R^k\}] = E[\exp\{a_k^* \cdot R^k\}] =: A_k$ we know that $q_k^*(i) = \exp\{a_k^* \cdot i\}/\ A_k$. Now

$$\begin{aligned}\inf_{a\in\mathbb{R}^d} E[\exp\{x + a^\top R\}] &= e^x \cdot \inf_{a\in\mathbb{R}^d} \Pi_{k=1}^d E[\exp\{a^k \cdot R^k\}] \\ &= e^x \cdot \Pi_{k=1}^d \inf_{a\in\mathbb{R}} E[\exp\{a \cdot R^k\}] = e^x \cdot \Pi_{k=1}^d A_k\end{aligned}$$

by independence. Therefore we obtain

$$\begin{aligned}Q^E\Big[\{(i^1,\cdots,i^d)\}\Big] \ \Big/\ P\Big[\{(i^1,\cdots,i^d)\}\Big] \\ = \exp\left\{x + \sum_{k=1}^d a_k^* \cdot i^k\right\} \Big/ \left(e^x \cdot \prod_{k=1}^d A_k\right) \\ = \prod_{k=1}^d \exp\{a_k^* \cdot i^k\} \Big/\ A_k = \prod_{k=1}^d q_k^*(i^k).\end{aligned}$$

The result immediately extends to $T > 1$ if one assumes that $R_1^1, \cdots, R_1^d, \cdots, R_T^1, \cdots, R_T^d$ are independent. Motoczynski & Stettner [30] consider super-hedging (see next sections) in the multi-dimensional multi-period Cox-Ross-Rubinstein model. ∎

14.7 SUPER-HEDGING, THE ONE-PERIOD MODEL

In this section we restrict attention to the case $N = 1$ and study values x for the initial wealth that allow for super-hedging a given discounted contingent claims $\breve{X}$ by some policy ϕ such that $X_N^\phi(x) \geq \breve{X}$. As in §4 we set $R := R_1$. Then we have to solve the problem:

$$x^* = \inf\Big\{x \in \mathbb{R};\quad \exists\, a \in \mathbb{R}^d \text{s.th. } x + a^\top \cdot R(\omega) \geq \breve{X}(\omega), \omega \in \Omega\Big\}. \tag{14.38}$$

Since $\omega \in \Omega$ is finite, one obtains a linear program by introducing:

$$\mathfrak{V} := \left\{(x,a) \in \mathbb{R}^{1+d}; 1 \cdot x + \sum_k a^k \cdot R^k(\omega) \geq \breve{X}(\omega), \omega \in \Omega\right\}. \tag{14.39}$$

Then (14.38) can be written as the linear program:

$$\min\{x; (x,a) \in \mathfrak{V}\}. \tag{P}$$

Then the dual program is:

$$\max\left\{\sum_{\omega\in\Omega} q(\omega) \cdot \breve{X}(\omega); q \in \mathfrak{Q}_1\right\} \tag{D}$$

where

$$\hat{\mathfrak{Q}}_1 := \left\{ q \in \mathbb{R}^{|\Omega|}; q(\omega) \geq 0, \omega \in \Omega, \sum_{\omega} q(\omega) = 1, \sum_{\omega} q(\omega) \cdot R^k(\omega) = 0, 1 \leq k \leq d \right\}. \tag{14.40}$$

$\hat{\mathfrak{Q}}_1$ can be identified with set of martingale measures in the case $N = 1$. A useful characterization is given in Pliska [32, p.59]. The set of equivalent martingale measures is

$$\mathfrak{Q}_1 := \{q \in \hat{\mathfrak{Q}}; q(\omega) > 0, \omega \in \Omega\}. \tag{14.41}$$

Obviously, $\mathfrak{V}$ is not empty; one only has to choose x large enough. From the Fundamental Theorem in §3 we know that $\mathfrak{Q}_1$ and hence $\hat{\mathfrak{Q}}_1$ is not empty under the no-arbitrage condition. Now the duality theorem of linear programming applies and we obtain (Pliska [32, p.27]:

Proposition 14.3

$$\begin{aligned} x^* &= \min\Big\{ x \in \mathbb{R};\ \exists\, a \in \mathbb{R}^d s.th.\ x + a^\top \cdot R(\omega) \geq \breve{X}(\omega), \omega \in \Omega \Big\} \\ &= \max_{q \in \hat{\mathfrak{Q}}_1} \sum_{\omega \in \Omega} q(\omega) \cdot \breve{X}(\omega) = \sup_{q \in \mathfrak{Q}_1} \sum_{\omega \in \Omega} q(\omega) \cdot \breve{X}(\omega) =: \sup_{q \in \mathfrak{Q}_1} E_q[\breve{X}]. \end{aligned}$$

The latter equality holds because $\mathfrak{Q}_1$ is dense in $\hat{\mathfrak{Q}}_1$. This is Theorem 14.2 in the case $N = 1$. An immediate consequence is:

Corollary 14.2 *$E_q[\breve{X}] \leq 0\ \forall q \in \mathfrak{Q}_1$ if and only if there exists some $a \in \mathbb{R}^d$ such that $a^\top \cdot R \geq \breve{X}$.*

14.8 SUPER-HEDGING, THE MULTI-PERIOD MODEL

In this section we consider the general case $N \geq 1$ and we may assume (NA)* as in §5. It will be convenient to work with a canonical probability space:

$$\begin{aligned} &\Omega := \Omega_N \text{ where } \Omega_n := E_0 \times \cdots \times E_n, \\ &I_n(\omega) = i_n, \quad \omega_n := (i_0, \cdots, i_n) \text{ for } \omega = (i_0, \cdots, i_N) \in \Omega. \end{aligned} \tag{14.42}$$

We need the following characterization of $\mathfrak{Q}$ ([34], Korn & Schäl [24]).

Proposition 14.4 *$Q \in \mathfrak{Q}$ if and only if*

$$\begin{aligned} Q[\{\omega\}] = q_0(i_0; i_1) \cdot q_1(i_0, i_1; i_2) \cdot \cdots \cdot q_{N-1}(\omega_{N-1}; i_N) \\ \text{where } q_n(\omega_n; \cdot) \in \mathfrak{Q}_n(i_n) \text{ for } \omega = (i_0, \cdots, i_N) \text{ and} \end{aligned} \tag{14.43}$$

$$\mathfrak{Q}_n(i) := q; q : E_{n+1} \mapsto (0,1); \sum_{j \in E_{n+1}} q(j) = 1, \sum_{j \in E_{n+1}} q(j) \cdot \rho(i,j) = 0\, i \in E_n. \tag{14.44}$$

There is now an interesting relation to dynamic programming. One can look upon $\mathfrak{Q}_n(i)$ as a set of actions available at time n given the information i about the market at time n. Moreover, a function q_n can be considered as a function which selects for each given history ω_n an action $q_n(\omega_n;\cdot) \in \mathfrak{Q}_n(i_n)$. Then a policy is given by $(q_0, q_1, \cdots, q_{N-1})$ and defines a martingale measure through (14.43). In models with a more general space Ω, the situation is more complicated. But even then the set $\mathfrak{Q}$ enjoys a property ([38, Lemma 1.8]) which is known for the set of policies in dynamic programming and is called "to admit needle-like variation" by Fakeev [8], "stability" by Hinderer [19], and "product property" by Hordijk [20]. This relation to dynamic programming was used by El Karoui & Quenez [7] for diffusion models and by Naik & Uppal [31].

As mentioned above, it is natural from the point of view of Markov decision theory to generalize the concept of an American option. We call the generalization a *dynamic option* which has interesting applications. The buyer has to pay a premium x to the seller and then he can choose any policy δ according to a certain dynamic program. Then the policy δ implies some (non- discounted) cost or claim $X^\delta := B_N \cdot \breve{X}^\delta : \Omega \mapsto \mathbb{R}$ which the seller has to pay to the buyer at N. On the other side, the seller can invest according to a policy ϕ in the market, i.e. the seller is the investor. Now, we look for some initial wealth x such that

for each policy δ of the buyer there is some policy $\phi(\delta)$ of the seller such that

$$X_N^{\phi(\delta)}(x) \geq \breve{X}^\delta \text{ on } \Omega, \tag{14.45}$$

i.e. such that $\breve{X}^\delta$ can be super-hedged by x and $\phi(\delta)$. In particular, we are again interested in the smallest value x^* such that (14.45) holds for $x = x^*$.

We obtain the usual European option if there exists only one policy for the buyer. One obtains an American call option if the buyer may choose a stopping time $\tau : \Omega \mapsto \{0, \cdots, N\}$ and may exercise his option at time τ . Such a contract is equivalent to a payment $X^\tau = (S_\tau^1 - \chi)^+$. Then one is interested in finding some initial wealth x and some policy ϕ such that for the discounted claim:

$$X_\tau^\phi(x) \geq \breve{X}^\tau = (\breve{S}_\tau^1 - \chi/B_\tau)^+ = (S_\tau^1 - \chi)^+/B_\tau \text{ for all stopping times } \tau\,. \tag{14.46}$$

We can write $X_\tau^\phi(x) = X_N^{\phi(\tau)}(x)$ where $\phi_n(\tau) := \phi_n$ for $n \leq \tau$ and $\phi_n(\tau) := 0$ for $n > \tau$. There, first the buyer decides to stop and then the seller decides to stop investing. Thus, the American option fits into the framework of a dynamic option and the investment policy of the seller indeed depends on τ which is not clear from (14.46) at first glance.

A necessary condition for (14.45) can be deduced from (14.11) (i.e. $E_Q[X_N^\phi(x)] = x$ for any $Q \in \mathfrak{Q}$). We obtain:

$$\bar{x} := \sup_{Q,\delta} E_Q[\breve{X}^\delta] \leq x^* := \inf\{x;\, (14.45) \text{ holds }\}. \tag{14.47}$$

Now our goal is to show that

$$\forall \delta\ \exists\ \phi(\delta) \text{ such that } \quad X_N^{\phi(\delta)}(\bar{x}) \geq \breve{X}^\delta. \tag{14.48}$$

Then $\bar{x} = x^*$ and the infimum in the definition of x^* is attained.

First we describe a dynamic program for the buyer by the usual set up. Here it is useful to choose a model with no running costs and a terminal reward (claim) $\breve{X}$ which depends on the whole history at time N. [In such a framework one can model restrictions on the admissible actions by the choice of the terminal reward.] In order to describe general stopping times of the buyer by policies of the buyer we have to allow for non-Markovian policies.

(1) $\mathbb{A}'_n$ is the *action space* of the buyer;

(2) $H_n := E_0 \times \mathbb{A}'_0 \times \cdots \times E_n$ is the *space of histories* of the buyer up to epoch n;

(3) a *policy* $\delta = (\delta_0, \cdots, \delta_{N-1})$ of the buyer is given by functions $\delta_n : \Omega_n \mapsto \mathbb{A}'_n$;

(4) $\breve{X} : H_N \mapsto \mathbb{R}$ is a *discounted claim function* where we write $\breve{X}^\delta(\omega) := \breve{X}(h_N^\delta(\omega))$ and $h_n^\delta(\omega_n) := (i_0, \delta_0(i_0), i_1, \cdots, \delta_{n-1}(\omega_{n-1}), i_n) \in H_n$.

At each time n the buyer can choose an action a'_n according to some policy δ and the seller will immediately be informed about the choice of a'_n at time n. The claim of the dynamic option is executed at time N; at the end the seller has to pay the amount $B_N \cdot \breve{X}^\delta(\omega)$ to the buyer.

For an American call option, one will have $\mathbb{A}'_n := \{0, 1\}$ where '1' means 'to stop' and '0' stands for 'to do nothing'. Moreover, one can choose

$$\breve{X}(i_0, a'_0, i_1, \cdots, a'_{N-1}, i_N) = (i_\tau - \chi)^+ / B_\tau$$

where $\tau := \inf\{n;\ a_n = 1\}$ with $\inf \emptyset = N$. One can imagine that the claim $(i_\tau - \chi)^+$ at time τ is invested in the savings account up to time N; at N the amount $(i_\tau - \chi)^+ \cdot B_N / B_\tau$ is paid to the buyer including interest for the time between τ and N. Discounting then leads to $(i_\tau - \chi)^+ / B_\tau$. Thus the American option can also be described by a claim executed at time N.
We will present another example.

Example from the bond market

Assume that at time $n = 0$ the buyer gives some capital y_0 to the seller. In that situation, an action a' consists in taking back the amount a'. Then

$$\mathbb{A}'_n = [0, \bar{a}_n] \text{ for some } \bar{a}_n \geq 0 \text{ and for } n \geq 0.$$

Actually admissible actions a'_n at time n are only those with $a'_n \leq \alpha_n$ where

$$\alpha_n := \bar{a}_n \wedge (y_0 - a'_1 - \cdots - a'_{n-1}), n < N.$$

This can be modeled by either defining $\breve{X}(i_0, a'_0, i_1, \cdots, a'_{N-1}, i_N) = -\infty$ if some a'_n is not admissible or by setting $\breve{X}(i_0, a'_0, i_1, \cdots, a'_{N-1}, i_N) = \breve{X}(i_0, a'_0 \wedge \alpha_0, i_1, \cdots, a'_{N-1} \wedge \alpha_{N-1}, i_N)$. In the later case the actions a'_n and $a'_n \wedge \alpha_n$ are identified. ∎

A further modern example is provided by a *passport option* (see Delbaen & Yor [5], Shreve & Vecer [43] which allows the buyer to take (long or short) positions in a stock. If at N the buyer makes a benefit, he can keep it. Otherwise he does not have to pay for the losses.

The idea of our approach consists in writing $\bar{x} := \sup_{Q,\delta} E_Q[\breve{X}^\delta]$ as value of a dynamic program in a super-model with a super-player combining the buyer choosing δ (controlling the claim) and the market choosing the martingale measure Q (controlling the law of motion).

As before, the initial state i_0 is fixed. The data are

$(\hat{1})$ $\hat{\mathbb{A}}_n := \mathbb{A}'_n \times \{q : E_{n+1} \mapsto (0,1), \sum_{i\in E_{n+1}} q(i) = 1\}, n \geq 0,$

$(\hat{2})$ $\hat{\mathbb{A}}_n(i_n) = \hat{\mathbb{A}}'_n \times \mathfrak{Q}_n(i_n)$ for $n \geq 0$,

$(\hat{3})$ $\hat{\delta} := (\delta, Q)$ where $\delta = (\delta_0, \cdots, \delta_{N-1})$, $Q = (q_0, \cdots, q_{N-1})$,

$(\hat{4})$ $\hat{X} := \breve{X}$, $\hat{X}^\delta := \breve{X}^\delta$;

$(\hat{5})$ the law of motion is given through Q as in Proposition 14.4.

A policy δ is called admissible at $h_n = (i_0, a'_0, \cdots, i_n)$ if $\delta_m(i_0, \cdots i_m) = a'_m$ for $m < n$. Then the value functions of the super-model are given by

$$\hat{V}_n(h_n) := \sup\{E_Q[\breve{X}^\delta | I_0 = i_0, \cdots, I_n = i_n]; \delta \text{ is admissible at } h_n, Q \in \mathfrak{Q}\},$$
$$\text{for } h_n = (i_0, a'_0, \cdots, i_n) \in H_n. \tag{14.49}$$

Then $\hat{V}_0(i_0) := \bar{x}$.

An important tool will be the optimality equation (Hinderer [19, Theorem 14.4)]:

$$\hat{V}_n(h_n) = \sup_{a'\in\mathbb{A}'_n, q\in\mathfrak{Q}_n(i_n)} \sum_{i\in E_{n+1}} q(i)\hat{V}_{n+1}(h_n, a', i)\,, \tag{14.50}$$

which implies:

Lemma 14.9

$$\sum_{i\in E_{n+1}} q(i)\Big[\hat{V}_{n+1}(h_n, a', i) - \hat{V}_n(h_n)\Big] \leq 0 \;\forall a' \in \mathbb{A}'_n, q \in \mathfrak{Q}_n(\omega_n). \tag{14.51}$$

Now we can use the results for the one-period model by use of the *standard reduction method* of Hinderer [19, p.24]. The idea consists in treating time n as new origin for fixed (h_n, a'). Then $\mathfrak{Q}_n(i_n)$ is the set of equivalent martingale measures for the new one-period model. Now Corollary 14.2 applies and we obtain:

Proposition 14.5 *For fixed* $(h_n, a') = (i_0, a'_0, \cdots, i_n, a')$ *there exists some* $\bar{\phi}_n(h_n, a')$ *such that*

$$\hat{V}_{n+1}(h_n, a', i) - \hat{V}_n(h_n) \leq \bar{\phi}_n(h_n, a')^\top \cdot \rho_{n+1}(i_n, i)\; \forall i \in E_{n+1}.$$

One can interpret $\bar{\phi}_n(h_n, a')$ as the decision of the seller at time n about the investment in the market after the buyer chose action a'. Therefore the supermodel can be looked upon as a stochastic game with complete information in the sense of Küenle [26]. Now we can construct the desired investment policy, which will be non-Markovian, as follows:

$$\bar{\phi}_n^{\delta}(\omega_n) := \bar{\phi}_n(h_n^{\delta}(\omega_n), \delta_n(\omega_n)), n \geq 0, \text{ where } h_0^{\delta}(\omega_0) = \omega_0 = i_0. \tag{14.52}$$

There $h_n^{\delta}(\omega_n)$ is defined above. Then one has by Proposition 14.5:

$$\hat{V}_{n+1}(h_n^{\delta}(\omega_n), \delta_n(\omega_n), i_{n+1}) - \hat{V}_n(h_n^{\delta}(\omega_n)) \leq \bar{\phi}_n^{\delta}(\omega_n)^{\top} \cdot \rho_{n+1}(i_n, i_{n+1}). \tag{14.53}$$

Now we can prove our goal (14.48):

Theorem 14.8 *$X_N^{\bar{\phi}(\delta)}(\bar{x}) \geq \breve{X}^{\delta}$ where $\bar{\phi}(\delta) := (\bar{\phi}_n^{\delta})$.*

Proof. Choose some $h = h_N = (i_1, a_1', \cdots, i_N)$ with histories $h_n = (i_1, a_1', \cdots, i_n)$, $\omega = (i_1, \cdots, i_N)$, $\omega_n = (i_1, \cdots, i_n)$. Then by Proposition 14.5

$$\begin{aligned}
\hat{V}_N(h) - \hat{V}_0(i_0) &= \sum_{n=0}^{N-1} [\hat{V}_{n+1}(h_{n+1}) - \hat{V}_n(h_n)] \\
&\leq \sum_{n=0}^{N-1} \bar{\phi}_n(h_n, a_n')^{\top} \cdot \rho_{n+1}(i_n, i_{n+1}) \text{ where} \\
\hat{V}_N(h) &= \breve{X}(h), \hat{V}_0(i_0) = \bar{x} \text{ by (14.47) and (14.49).}
\end{aligned}$$

Thus we have

$$\breve{X}(h) \leq \bar{x} + \sum_{n=0}^{N-1} \bar{\phi}_n(h_n, a_n')^{\top} \cdot \rho_{n+1}(i_n, i_{n+1}) \text{ for } h \in H_N$$

which implies

$$X^{\delta}(\omega) \leq \bar{x} + \sum_{n=0}^{N-1} \bar{\phi}_n^{\delta}(\omega_n)^{\top} \cdot \rho_{n+1}(i_n, i_{n+1}) = X_N^{\phi(\delta)}(\bar{x}). \tag{14.54}$$

■

Lemma 14.9 can be looked upon as a supermartingale property under each martingale measure which was derived by means of dynamic programming in the same sense as in El Karoui & Quenez [7]. One can write (14.54) as

$$X^{\delta}(\omega) = X_N^{\phi(\delta)}(\bar{x}) + \sum_{n=1}^{N} c_n^{\delta} \text{ where} \tag{14.55}$$

$$c_{n+1}^{\delta} := \left[\hat{V}_{n+1}(h_{n+1}^{\delta}(\omega_{n+1})) - \hat{V}_n(h_n^{\delta}(\omega_n))\right] - \bar{\phi}_n^{\delta}(\omega_n)^{\top} \cdot \rho_{n+1}(i_n, i_{n+1}) \geq 0.$$

Then c_n^{δ} can be interpreted as a possible consumption at time n when hedging $\breve{X}$.

Theorem 14.8 can be embedded in the framework of a general optional decomposition theorem for supermartingales under each martingale measure. (Föllmer & Kabanov [10], Föllmer & Kramkov [11], Kramkov [25], Schäl [38]).

The analogy between the set of all policies in Markov decision model and the set of martingale measures can also be used to show that Markovian martingale measures are sufficient in the same sense as Markovian policies are sufficient in a Markovian environment ([38]).

Acknowledgment

I would like to thank the referee for very useful remarks. Moreover, it is a pleasure to acknowledge the influence of my teacher Karl Hinderer who laid much of the foundation of the general non-Markovian theory of dynamic programming. Modern finance is another important area where the non-Markovian setting is natural. In particular, contingent claims of so-called exotic options provide interesting examples of general history-dependent reward and cost functions, respectively.

References

[1] D. Bertsekas, "Necessary and sufficient conditions for existence of an optimal portfolio", *J. Econ. Theory* **8**, 235–247, 1974.

[2] H. Bühlmann, F. Delbaen, P. Embrechts and A. Shiryaev, "On Esscher transforms in discrete finance models", *Astin Bulletin* **28**, 171–186, 1998.

[3] R.C. Dalang, A. Morton and W. Willinger, "Equivalent martingale measures and no-arbitrage in stochastic securities market models", *Stochastics and Stochastic Reports* **29**, 185–201, 1990.

[4] M.H.A. Davis, "Option pricing in incomplete markets", in: *Mathematics of Derivative Securities*, ed: M.A.H. Dempster, S.R. Pliska, Cambridge Univ. Press, 216–226, 1997.

[5] F. Delbaen, and M. Yor, "Passport options", Working paper, ETH Zürich, 1999.

[6] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes*, Springer, Berlin, 1997.

[7] N. El Karoui and M.C. Quenez, "Dynamic programming and pricing of contingent claims in an incomplete market", *SIAM J. Control Optim.* **33**, 29–66, 1995.

[8] A.G. Fakeev, "Optimal stopping rules for processes with continuous parameter", *Theory Probab. Appl.* **15**, 324–331, 1970.

[9] E.A. Feinberg, "Total reward criteria", This volume, 2000.

[10] H. Föllmer and Yu.M. Kabanov, "Optional decomposition and Lagrange multipliers", *Finance Stochast.* **2**, 69–81, 1998.

[11] H. Föllmer and D. Kramkov, "Optional decompositions under constraints", *Probab. Theory Relat. Fields* **109**, 1–25, 1997.

[12] H. Föllmer and M. Schweizer, "Hedging of contingent claims under incomplete information", in: *Applied Stochastic Analysis*, eds. M.H.A. Davis and R.J. Elliot, Gordon and Breach, London, 1990.

[13] M. Fritelli, "The minimal entropy martingale measure and the valuation problem in incomplete markets", *Mathematical Finance*. To appear.

[14] P. Grandits, "The p-optimal martingale measure and its asymptotic relation with the minimal-entropy martingale measure", *Bernoulli* **5**, 225–248, 1999.

[15] P. Grandits and Th. Rheinlaender "On the minimal entropy martingale measure", Preprint, Techn. Univ. Berlin, Fb. Mathem. 1999.

[16] N. H. Hakansson, "Optimal entrepreneurial decisions in a completely stochastic environment", *Management Science* **17**, 427–449, 1971.

[17] J.M. Harrison and D.M. Kreps, "Martingales and arbitrage in multiperiod securities markets", *J. Economic Theory* **20**, 381–408, 1979.

[18] J.M. Harrison and S.R. Pliska, "Martingales and stochastic integrals in the theory of continuous trading", *Stoch. Processes & Appl.* **11**, 215–260, 1981.

[19] K. Hinderer, *Foundations of non-stationary dynamic programming with discrete time-parameter*, Lecture Notes in Operations Research and Mathematical Systems **33** Berlin-Heidelberg-New York: Springer, 1970.

[20] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Amsterdam: Mathematical Centre Tracts **51**, 1974.

[21] J. Jacod and A. N. Shiryaev, "Local martingales and the fundamental asset pricing theorems in the discrete-time case", *Finance Stochast.* **3**, 259–273, 1998.

[22] I. Karatzas, "On the pricing of American Options", *Appl.Math. Optim.* **17**, 37–60, 1988.

[23] I. Karatzas and S.G. Kou, "On the pricing of contingent claims under constraints", *Ann. Appl. Probab.* **6**, 321–369, 1996.

[24] R. Korn and M. Schäl, "On value preserving and growth optimal portfolios", *Math. Methods Op. Res.* **50**, 189–218, 1999.

[25] D.O. Kramkov, "Optional decomposition of supermartingales and hedging of contingent claims in incomplete security markets", *Probab. Theory Relat. Fields* **105**, 459–479, 1996.

[26] H.-U. Küenle, *Stochastische Spiele und Entscheidungsmodelle*, Teubner-Texte zur Mathematik, 89. Teubner, Leibzig, 1986.

[27] H. Leland, "On the existence of optimal policies under uncertainty", *J. Econ. Theory* **4** , 35–44, 1972.

[28] R.C. Merton, *Continuous-time Finance*, Blackwell, Cambridge, MA, 1990.

[29] M. Motoczynski, "Multidimensional variance-optimal hedging in discrete time model—a general approach", *Mathematical Finance* **10**, 243–258, 2000.

[30] M. Motoczynski and L. Stettner, "On option pricing in the multidimensional Cox-Ross- Rubinstein model", *Applicationes Mathematicae* **25**, 55–72, 1998.

[31] V. Naik and R. Uppal, "Minimum cost hedging in incomplete Markets", Univ. of British Columbia, Working paper, 1992.

[32] S.R. Pliska, *Introduction to Mathematical Finance*, Blackwell publisher, Malden, USA, Oxford, UK, 1997.

[33] L.C.G. Rogers, "Equivalent martingale measures and no-arbitrage", *Stochastics and Stochastic Reports* **51**, 41–49, 1994.

[34] R.T. Rockafeller, *Convex Analysis*, Princeton Univ. Press, Princeton, N.J, 1970.

[35] W. J. Runggaldier and A. Zaccaria, "A stochastic control approach to risk management under restricted information", *Mathematical Finance* **10**, 277–288, 2000.

[36] W. Schachermayer, "A Hilbert space proof of the fundamental theorem of asset pricing in finite discrete time", *Insurance: Mathematics and Economics* **11**, 249–257, 1992.

[37] M. Schäl, "On quadratic cost criteria for option hedging", *Math. Oper. Res.* **19**, 121–131, 1994.

[38] M. Schäl, "Martingale measures and hedging for discrete-time financial markets", *Math. Oper. Res.* **24**, 509–528, 1999.

[39] M. Schäl, "Portfolio optimization and martingale measures", *Mathematical Finance* **10**, 289–304, 2000.

[40] M. Schäl, "Price systems constructed by optimal dynamic portfolios", To be published in *Math. Methods Op. Res*, 2000.

[41] M. Schweizer, "Variance-optimal hedging in discrete time", *Math. Oper. Res.* **20**, 1–32, 1995.

[42] M. Schweizer, "Approximation pricing and the variance-optimal martingale measure", *Ann. Prob.* **24**, 206–236, 1996.

[43] S. E. Shreve and J. *Večeř*, "Options on a traded account: vacation calls, vacation puts and passport options", *Finance Stochast.* **4**, 255–274, 2000.

Manfred Schäl
Inst. Angew. Math.
University of Bonn
Wegelerstr.6
D-53115 Bonn, Germany
schael@wiener.iam.uni-bonn.de

15 APPLICATIONS OF MARKOV DECISION PROCESSES IN COMMUNICATION NETWORKS

Eitan Altman

Abstract: We present in this chapter a survey on applications of MDPs to communication networks. We survey both the different application areas in communication networks as well as the theoretical tools that have been developed to model and to solve the resulting control problems.

15.1 INTRODUCTION

Various traditional communication networks have long coexisted providing disjoint specific services: telephony, data networks and cable TV. Their operation has involved decision making that can be modeled within the stochastic control framework. Their decisions include the choice of routes (for example, if a direct route is not available then a decision has to be taken which alternative route can be taken) and call admission control; if a direct route is not available, it might be wise at some situations not to admit a call even if some alternative route exists.

In contrast to these traditional networks, dedicated to a single application, today's networks are designed to integrate heterogeneous traffic types (voice, video, data) into one single network. As a result, new challenging control problems arise, such as congestion and flow control and dynamic bandwidth allocation. Moreover, control problems that had already appeared in traditional networks reappear here with a higher complexity. For example, calls corresponding to different applications require typically different amount of network resources (e.g. bandwidth) and different performance bounds (delays, loss probabilities, throughputs). Admission control then becomes much more complex than it was in telephony, in which all calls required the same perfor-

mance characteristics and the same type of resources (same throughput, bounds on loss rates and on delay variation).

We do not aim at a complete survey of the area, since several other surveys [85, 89, 142, 113, 143, 145, 243, 240, 242, 245] on related issues already exist, see also [213]. Other references that focus on the methodology that allows to use MDPs to communication networks can be found in the following books [4, 50, 106, 231, 263, 275] as well as in the survey paper [185].

We have two goals in this survey. First, we wish to present to researchers who specialize in MDPs a central application area which provides a vast field of challenging problems. We would like to familiarize these researchers with special complex features in control problems that arise in communications: complex information structure, problems with multiobjective and multiagents. A second objective is to familiarize researchers in communications with tools that have been developed for modeling and for solving control problems in networks.

Problems that are described in this survey are MDPs in a general sense: they are described as Markov chains whose transition probabilities are controlled. This control can be done by a single or several controllers, having the same or having different objectives. It is often tools other than the standard dynamic programming that are used to solve these control problems. For completeness, we occasionally mention approaches that have been used for control of communication systems which are not based on MDPs, and then relate or compare these to MDPs.

15.2 THE CONTROL THEORETICAL FRAMEWORK

The most popular telecommunication network architectures today are the Internet and the ATM (Asynchronous Transfer Mode) networks. The Internet offers today a "best effort" type service, i.e. the resources in the network are shared among all users, and when the number of users increases, the quality of service (in terms of delay, throughput, losses) per user decreases. In contrast, ATM networks provide mostly guaranteed services: if a session needs a given Quality Of Services (QOS) it may establish a contract with the network that guarantees that all along the duration of the session, some required bounds would hold on given performance measures (loss probabilities, delays, delay variation, throughput), as long as the source respects the terms of the contract (in terms of the average and the maximum throughput it sends, as well as some constraints on its bursty behavior). Two guaranteed service classes are defined in ATM: the CBR (Constant Bit Rate) and VBR (Variable Bit Rate) service classes. ATM contains also two best effort type services: the Available Bit Rate (ABR) service and the Unspecified Bit Rate (UBR) service. In the ABR service, the network determines the allowed transmission rate of each source by sending to them periodically appropriate control signals. As long as a source abides to those commands, the network guarantees some given bounds on its loss rates. In the UBR service no guarantees are given on performance measures.

From a control theoretical point of view, an important classification of the control in networks is according to who is controlling and what the objectives are. Three general frameworks are possible:

1. Centralized control, or a single controller. This is the case in problems such as call admission control: a request for a new connection arrives and the network has to decide whether to accept it or not.

2. Team theory [48]: several controllers but a single objective. This is the case when the control decisions are taken by different elements (or agents) in the network rather than by the users. The common objective(s) might be the efficient use of the network and providing a good service to the network users.

3. Game theory [48]: there is more than one selfish decision makers (players) and each has its own objective. The goal of each player is to maximize its own performance, and the decision of each player has an impact on the performance of other players. This framework models the case of several users who can control their own flow, or the routes of their own traffic.

From a control theoretical point of view, ATM networks can be viewed generally as *team control problems*: there are several agents within the network, typically situated in the routers or in the access points to the network, that take decisions. Their objectives is to guarantee the performances that they wish to provide to the various applications and, if possible, to optimize them. In addition, an important objective is to use network resources efficiently (such as memory, bandwidth). (Often different controllers have different information, as will be discussed in Section 15.3.) The literature on problems in telecommunications (or related models) that fall into this category is very rich, see [11, 13, 107, 230, 278].

The appropriate theoretical framework to consider the Internet is more involved. It could be considered as in a *game* framework since essential control decisions (such as flow control) are taken by the users which are a-priori non-cooperative. The sources may have different objectives and their dynamic behavior may have an influence on other users. In practice, however, the actual design of controllers of data transfer (TCP/IP [135]) is frequently done (through the software that comes with computers handling Internet connections) in an unselfish way. The controllers can still be changed by the users or by applications to improve their individual performance, but this is seldom done. On the other hand, rate controllers are frequently applied in the Internet to interactive voice and video (by changing the compression rate and thus the quality of service). In these applications there is much more variety of controllers and these are often designed in a selfish way.

Two optimality concepts are used in the non-cooperative games arising in networks. The first is the Nash equilibrium: it arises when the performances are determined by a finite number of controllers and it constitutes of a strategy set for each of these controllers such that no user can benefit from deviating from its own strategy as long as the other controllers stick to their strategy. A second type of equilibrium concept arises when the number of users is infinite

(which may model in practice a finite but large number of users) and the influence of a single user on the performances of other users is negligible. This is called an atomic game. In the context of routing problems in networks, the corresponding equilibrium is called a *Wardrop equilibrium* [264]. This concept turns out to be useful also in cases in which both the flow and the routing are controlled [219].

If we consider a very large number of sessions in a network, the routing decisions for a single session will typically have a negligible influence on the performance of other sessions. This can be viewed as an atomic game. If however, the routing decisions are not made by the application that initiated the session but, rather, by the service provider to which the user is subscribed, then the analysis will be in terms of a Nash equilibrium, given that the routing decisions of any service provider can have a non-negligible impact on the performance of users of other service providers as well. Another example of where the Nash equilibrium is the proper theoretical concept is flow control on the Internet initiated by Web connections. A single request to access a Web page that contains many pictures can result in opening *simultaneously* several connections each of which is used to fetch the data of another pictures. In such a case, a single application can have an important impact on many other applications.

Most work that studied telecommunication networks within the the game theoretic framework did not use a dynamic setting and restricted to determining the size and the routes of average flows. We present some examples that did use stochastic dynamic models (Markov games—which are the extension of MDPs to a game situation of selfish controllers).

Korilis and Lazar established in [161] the existence of an equilibrium in a distributed flow control problem in a network in a setting of Nash equilibrium. They further characterized its solution using coupled linear programs.

The references [42, 34, 49, 73] treat a problem related to the choice of connection between a best-effort type service and a dedicated guaranteed service. Consider a request that arrives for a connection for a data transfer. The user that initiates the connection can either use a guaranteed service and request a fixed dedicated amount of resources (such as bandwidth), or it could use a typically cheaper best-effort service in which the resources are shared between all best-effort ongoing connections. The information available for taking the decision is the number of ongoing best-effort connections. The user wishes to minimize the expected time it takes for the data transfer. Note that this time depends also on the decisions that will be made by future users that will wish to connect: if more users in the future will decide to use the best-effort service too, the expected transfer time will become longer, since the available resources are shared between all ongoing connections. However, the decisions of future arrivals are unknown. This explains the need for the Markov game approach and the equilibrium concept that is used in these references. The existence and uniqueness of an equilibrium is established in [34] and it is further computed in the case of a Poisson arrival process of sessions.

Other examples of Markov games which study models related to telecommunication applications appear in the references [3, 8, 117].

Before ending the discussion on game models, we mention zero-sum Markov games in which there are only two players with opposite objectives. Several models within this special framework have been used for the study of problems in telecommunications with a single *central controller* in which some parameters are unknown. The unknown parameters were assumed to vary in time in a way which is unpredictable to the central controller (called player 1). The goal of player 1 is to guarantee the best performance under the worst possible (unknown) dynamic choice of the unknown parameters. To solve such problems, one models the unknown parameters as if they were chosen by a second player with opposite objective. This gives rise to a zero-sum game. Admission control, routing, flow control and service assignment problems were considered in the references [1, 2, 5, 23, 28, 22].

In all three theoretical frameworks that we discussed above, the objective of a controller may be in fact a vector. Such objective is said to be *multicriteria*. In the case of a central controller one may wish to minimize delays, minimize losses, minimize rejection rate of sessions, minimize waste of resources. One could add up the objectives and consider minimizing a weighted sum of objectives instead. Alternative formulations are

1. Minimizing the vector of objective in the Pareto sense; a Pareto solution is a policy such that no strict improvement in the performance of one component can be achieved without strictly decreasing another component.

2. Minimizing one of the components of the objective subject to constraints on other components.

We state some example of the multicriteria approach in the modeling of telecommunication problems:

(1) *The maximization of the throughput of some traffic, subject to constraints on its delays* (this problem has been studied along with its dual problem: the minimization of expected delay subject to a lower bound constraint on the throughput). A rich research literature in this direction was started up by Lazar [179] and has been pursued and developed together with other researchers; some examples are the references [66, 130, 131, 161, 162, 258]. In all these cases, limit-type optimal policies were obtained (known as window flow control). Koole [152] and Hordijk and Spieksma [129] considered the problem introduced in [179] as well as other admission control problems within the framework of MDPs, and discovered that for some problems, optimal policies are not of a limit-type (the so called "thinning policies" were shown to be optimal under some conditions).

In [4], the author considers a discrete time model that extends the framework of the above problems and also includes service control. The latter control can model bandwidth assignment or control of quality of service. The flow control has the form of the control of the probability of arrivals at a time slot. The control of service is modeled by choosing the service rate, or more precisely, by assigning the probability of service completion within a time slot. A tradeoff exists between achieving high throughput, on the one hand, and low expected delays on the other. It is further assumed that there are costs on

the service rates. The problem is formulated as a constrained MDP, where we wish to minimize the costs related to the delay subject to constraints on the throughputs and on the costs for service.

(2) *Dynamic control of access of different traffic types.* A pioneering work by Nain and Ross considered in [205] the problem where several different traffic types compete for some resource; some weighted sum of average delays of some traffic types is to be minimized, whereas for some other traffic types, a weighted sum of average delays should be bounded by some given limit. This research stimulated further investigations; for example, Altman and Shwartz [35] who considered several constraints and Ross and Chen [226] who analyzed the control of a whole network. The typical optimal policies for these types of models requires some randomization or some time-sharing between several fixed priority policies.

(3) *Control of admission and routing in networks.* Feinberg and Reiman have solved in [95] the problem of optimal admission of calls of two types into a multichannel system with finite capacity. They established the optimality of a randomized trunk reservation policy.

Other problems in telecommunications which have been solved by constrained MDPs are reported in [74, 56, 192]. A study of a constrained MDP in a queueing model with a removable server, with possible applications in telecommunications or in production, was done in [93].

15.3 INFORMATION ISSUES AND ACTION DELAYS

When attempting to model control protocols in telecommunication networks within the optimal stochastic control framework, we note that many non standard features arise in the information structure.

We recall that a "full information" framework is a setting in which each controller knows at each time instant all previous states as well as the previous actions taken by all controllers, plus the current state. In addition, it is assumed that all controllers know the initial state (or the initial distribution over the initial state). We list below some complex aspects of the information available to the decision maker.

15.3.1 MDPs with incomplete information and MDPs with partial information

The decision maker does not have full information about the network state, but only some observations of the history. For example, the flow control in ABR services in ATM is performed by routers within the networks which have information only on their own congestion state (the number of cells queued at the router) and may have some rough information on congestion experienced by connections that use this router; but they have no information on that of other connections. There are various models that handle MDPs without full state information. By "MDP with incomplete information" we refer to a Baysian-type framework in which the decision maker(s) can take actions according to the observed history, and knows the initial state, or an initial distribution over the state.

In contrast, in the framework called "MDPs with partial information", the initial distribution or the initial state need not be available (see [127, 128, 167, 188].) To illustrate this framework, we mention the problem of two service stations in tandem. Customers of two different types arrive at a single server facility. The control decisions of the server are which type of customer to serve at each time. Customers that have arrived and have not yet been served are queued and wait their turn. Once a customer is served in the first station it is routed to a second service station that has the same structure, in which a second server has to decide which customer to serve at each time. There again queueing may occur. The decision of each server may depend only on the number of customers of each type queued in that center. This is a very realistic assumption in practice. This problem is solved in [128].

Another example of MDPs with partial information is window-based flow control in networks. We consider a network with several sources and destinations. Each source of packets has to take decisions concerning the transmission of new packets. When a packet reaches the destination the source receives an acknowledgment. The only information that a source has on the state of the network is through these acknowledgments, from which it can infer how many packets have not yet reached the destination (or more precisely, for how many packets acknowledgments have not yet come back). Thus each controller has some different local information. This problem has been fully solved in [130, 131] by exploiting the so called *Norton Equivalence*, which allows each source to consider the rest of the network as an equivalent single queue (see also [258]).

Below we review information structures that are, on one hand, suitable for modeling telecommunication systems and on the other hand, lead to solutions to the stochastic control problem (within a reasonable complexity).

15.3.2 Quantized information

This is a special case of incomplete information; due to the discrete nature of data networks, information that is originally represented by real numbers has to be transformed into numerical date using a finite number of bits. This causes loss of precision in the information. For example, congestion information in networks is often transmitted only by one or two information bits, so that the transmitted information takes a small number of possible values. This is the case in the TCP flow control on the Internet, where the way to infer that congestion exists is by a binary information on whether a packet is lost or not.

15.3.3 Delayed information

The information available to decision makers on the state often suffers from delays. A one-way propagation delay can exceed 250ms in a network that contains a Geostationary satellite link. Around 20 ms of propagation delay is incurred in a communication between the west and east coast of the USA. In addition to propagation delays, large random time varying delays are often incurred due to queueing. These components depend on the congestion state of the network and are sometime unknown to the controllers. In a network with several

controllers, the delays further vary from one controller to another. The delays could be neglected in analysis of networks in which throughputs are small and transmission delays large; in such cases the time scales related to events in the networks are larger than those involving the delays. Today high speed networks achieve very large throughputs, and the time between the transmission of two consecutive information bits can be smaller than 10^{-9}sec (when considering Gbit/s networks). Hence information delays become a crucial practical (and theoretical) problem.

We briefly review work on control problems with delayed information in telecommunications. Flow control with delayed information has been studied in [33, 36, 169] by transforming the problem into an equivalent MDP with full information. The first paper has been extended to noisy delayed information in [29]. Two types of flow control have been studied. The first type is a rate-base flow control, in which the rate of transmission of packets is directly controlled. The second type is a window-based flow control, in which the controller adjusts its window dynamically; a window stands for the number of packets that can be sent before acknowledgments to the source arrive from the destination. Work on rate-based flow control with delay in the framework of linear quadratic control (linear dynamics and quadratic cost) has appeared in [7, 9, 13]. The impact of delay on window-based flow control in the framework of Jackson network is analyzed in [66]. Routing with delayed information has been investigated in [40, 168, 251]. Finally, a problem of optimal priority assignment for access to a single channel with delay has been investigated in [27].

15.3.4 Sampled Information

Sampled information means that the decision maker (either within the network or at the source) does not get the information on the network state continuously, but it get some occasional updates. For example, the information used for flow control in the Internet are the acknowledgment that return from the destination to indicate the receipt of a packet. Acknowledgments return however at discrete times, may be lost, and their rate further depends on the transmission rate of the original transmitted data packets. To our knowledge there has been almost no work on MDPs with sampled information in the context of telecommunication networks. We should mention that modeling the flow control within the linear quadratic control framework (in particular—when considering Gaussian noise or the H^∞ approach, see [6, 9]) allows one to handle sampled information, using the theory in [47, Sec. 5.3]. But this has not yet been done. In [18] an alternative framework is used to handle sampled information; properties of optimal policies are established whenever the value function is multimodular in the controls.

15.3.5 Asynchronous Information

This is a special case of sampled information in which several controllers in a network receive information at different times, possibly independent of each other; a controller may not know when another controller receives an information update. For example, routing information for establishing the shortest

path in the network are typically gathered in different nodes (routers) in which local routing decisions are taken. Such routers exchange occasionally information to determine shortest paths in an asynchronous way.

15.3.6 Delayed Sharing Information

This is a special case of information structure that arises in team or game problems, see [132, 170, 171, 224, 260]. The state is given as a product of several local components where each component corresponds to one controller (or player). Time is discrete. All controllers have a one step delayed information about the global state of the system. However, each controller gets immediate information about the local component of the state that corresponds to it.

We present some examples of applications of this type of information structure in telecommunications. In [230], the authors have studied decentralized control in packet switched satellite communication and a decentralized control problem for multiaccess broadcast networks have been studied in [107]. In both examples, each controller has to decide whether to transmit or not, without knowing if packets have arrived in the current time unit to other nodes. If they did, then packets from other nodes could be scheduled for transmission at the same time and collisions could occur.

15.3.7 No information

When transferring short files, the information about the state of the network may come back to the source after the whole file is transferred. This situation occurs when transmission delays are negligible with respect to information delays. This illustrates the fact that controllers often have to make decisions with no available state information. There have been several approaches to solving MDPs with no state information. Within the partial information framework, the algorithm of Hordijk and Loeve [188] has been used in a problem of routing into parallel queues [127]. An alternative approach has been proved to be useful in solving routing control, admission control and polling [134, 14, 150]. This approach applies to cases in which instead of keeping the whole history of previous actions, only some finite (bounded) number of events are recorded. In all above papers, the events which have to be recorded when considering routing to several queues are simply how long ago a packet was routed to each one of the queues. This allows one to transform problems with no information to equivalent MDPs with full information, to establish the optimality of periodic policies, and to obtain other characteristics of optimal policies. It is in particular remarkable that models which are not Markovian (arrival of customers may be general stationary ergodic processes) can be handled by MDPs with finite spaces of states and actions [14]. A third approach consists of using special structure of queueing systems (queueing systems are often used to model telecommunication networks) which leads to the optimality of policies that are regular in some sense. In particular, whenever the value function can be shown to be multimodular in the controls, a rich theory exists for obtaining optimal policies with no state information. We refer to [111, 19] for the definition and properties of multimodular functions, and refer to [111, 14, 21, 19, 16, 17, 20] for

applications in routing, admission control and polling with no state information. Weaker notions of regularity of policies are presented in [21, 15, 68, 83, 121, 134]. In particular, the Golden-ratio approach is used in [121, 134] for an optimal channel assignment problem with no state information to obtain simple policies which are close to optimal. This approach is adapted to a context of optimal scheduling of search engines on the Web in [83] and to optimal polling in [68].

15.3.8 Nested information

In the case of several controllers, say N, we say that information is nested if we can order the controllers in a way that information available to a controller i is a subset of the information available to controller $i+1$ for all $1 \leq i < N$. This structure is again a special case of the incomplete information setting. Note that the case where several controllers receive all information after some fixed delay (that may depend on the controller) is a special case of the nested information structure. The controllers are then ordered with decreasing delays. An example of a flow control problem that gives rise to this information structure appears in [13].

15.3.9 On the tractability of complex information structure

There is one appealing feature in the nested information as well as in the one-delayed sharing pattern. In both cases it is possible to transform the problem into an MDP with full information with somewhat larger state and action spaces. The way to transform an MDP with incomplete information to an equivalent one with full information can be found in [120, 132, 149, 260] and references therein. For the case of several controllers, we cite a general condition that allows one to transform an incomplete information MDP with several controllers into an equivalent MDP that has a tractable solution [48, p. 369]: "An agent's information at a particular stage n can depend on the control of some other agent at some stage $k < n$ only if he also has access to his information available to that agent at that stage k". This condition includes the one-delayed sharing pattern as well as the nested information. It is called in [48] a "classical information pattern".

We note that some complex information structures seem natural for modeling telecommunication systems but have not been actually used in the literature since their solution is either hard, or unknown, or requires policies which are complex to realize (require many computation and/or memory). An example is team or game problems with no state information, with full information on the actions of the agents, but in which the different agents have different knowledge on the initial states. (This problem is related to the one in [97].)

An alternative conservative approach for handling various information structures is to consider noisy information, where the noise is allowed to be a general, possibly state dependent disturbance. A worst case design can then be used, see [6, 9]. This approach may be used for example to handle the case of quantized information.

15.3.10 Action delays

Another important implication of the delays in high speed networks is the so called *action delay*: even in absence of information delay, a large delay may elapse between the moment that a decision is taken by a controller till this decision has an impact on the network. For example, in ABR service of ATM networks, routers issue commands on flow control that are forwarded to the transmitting sources through special information cells. A router has immediate information on the local state of the queues in that router. Based on this information it sends commands to the sources. But by the time the sources react to these commands, and by the time the reaction influences the queues at the router, a large *action delay* elapses.

It has been shown in [13] that MDPs with both information and action delays can be transformed into a equivalent MDPs with only information delay.

In the following sections we summarize some central control issues that arise in telecommunications.

15.4 CALL ADMISSION CONTROL

The decision to accept another call to the system may influence both the performance of that call as well as that of ongoing calls in the system. This decision changes the state of the system and thus has also an impact on whether future calls will be accepted. Call admission control (CAC) is not applied today on the Internet, but is implemented in ATM networks. Whereas many protocols and control policies have been standardized in the ATM, the implementation of CAC in ATM network is left to the constructor and will probably not be standardized. We describe below several type of admission control problems that arise in telecommunication. Admission control is often combined with routing; we discuss this in the subsection on routing. Other issues related to admission control in wireless communication will be discussed in Section 15.10.

15.4.1 Admission control for CBR sources

If the network guarantees a fixed bandwidth per accepted call (e.g. in telephony or in CBR connections in ATM), then the performances of accepted calls is no more influenced by future decisions. In particular, in an ATM environment, delays and throughputs are guaranteed along the entire duration of the call once it is admitted. The main performance measure of interest in that context is then the average rejection rate of a session. Dynamic decisions are required in that framework if there are different classes of calls, each with its own requirement of network resources. This type of admission control problems is frequently modeled at a session level: the state is taken to be the number of sessions of each class, and the MDP is formulated by using the arrival rate of sessions, their average bandwidth requirement and the probability distribution of the duration of a session.

In the case of a single node, a given bandwidth has to be shared between sessions. This problem is known as the *stochastic knapsack problem*, see [24, 183, 218, 228] and references therein. In this setting, there may be cases in which if we accept a call that requires a small amount of bandwidth then there

may not be sufficient room for accepting a large call. This type of dilemma explains the fact that optimal acceptance policies are quite complex and are often non-monotone, see [24, 218].

In special cases, the acceptance policies turn out to be monotone and the acceptance region is convex [95, 183, 218, 228]. In the case of two classes with the same bandwidth requirements per class, the optimal policy is known as a trunk reservation policy, which means that we reserve some bandwidth for a high priority class and calls of the other classes are rejected when the remaining bandwidth for the priority class goes below that level. We mention here that in [95] the problem is posed as a constrained MDP and the trunk reservation policy requires randomization.

In the case of jobs with different requirements for bandwidth, an elegant solution of the optimal call admission control is presented in [102, 227] (for two classes) when restricting to coordinate convex policies, see also [204, 228]. The solution is based on the observation that for such policies, the steady state probabilities have a product form. A fluid approximation approach is used in [24] when relaxing the above restriction. The solution shows, in fact, that in the appropriate scaling, the limit fluid model has a trunk reservation solution which is a special case of a coordinate convex policy.

In the above discussion it is assumed that if a call is not accepted then it disappears. There are, however, cases in which the call can be put into some finite waiting queue instead of being completely rejected. An MDP is used in [215] to solve this problem and its performance is compared to other methods.

15.4.2 Admission control of sources with variable throughput

A very rich literature exists on admission control of calls with variable transmission rate (VBR sources). Assume that a single node with a given bandwidth is to be shared between several sources of this type. A conservative approach would be to consider the sum of peak rates of all existing connections and to check whether the overall bandwidth would be exceeded if we further added to that the peak rate of the new connection. If it is then the call is rejected. Instead, one tries to make use of *statistical multiplexing*, i.e. of the fact that rarely all sources transmit at peak rate. The typical question that is posed then is: how many calls can we accept so that the probability that a packet is lost is below some threshold? note that a packet is lost when the sum of the instantaneous transmission rates exceeds the available bandwidth. A popular approach is to characterize sources by a parameter called the *effective bandwidth* (see [87, 141, 186] and references therein) which is, roughly speaking, the amount of bandwidth the source needs so that the average loss rate is below some threshold. It is typically larger than the average transmission rate, but is still smaller than the peak rate. There are methods to estimate or to compute effective bandwidth [87, 141, 186].

The effective bandwidth approach itself is not related to MDPs, but it can be combined with MDPs. Indeed, once we know the effective bandwidth of a source with a variable transmission rate, we can use methods from the two previous sections for admitting calls and consider that the bandwidth they require is the effective bandwidth (instead of the peak rate bandwidth). This

allows one to accept more calls and thus use the network more efficiently at a price of some additional loss probability of packets in each connection. If the effective bandwidth of existing calls is unknown to the network, adaptive mechanisms can be used to estimate it and combining that with call admission control. This approach is known as *measurement based admission control*, see [108].

The effective bandwidth is not the only approach that can be used to combine session level considerations with packet level phenomena (i.e. the actual process of transmission of sources). In [202] an MDP is formulated to obtain a call admission control that takes both into account, and a solution is obtained through linear programming.

15.4.3 Call admission in an integrated service environment

The control problems that we mentioned in the previous subsections were related to "non-elastic" calls, i.e. calls that have some given bandwidth requirements and cannot adjust their transmission rate to the available bandwidth. On the other hand, there are applications that can adapt their transmission rate to the available bandwidth (for example, data transfer). In this case it may seem that call admission is not required; the ATM forum (a standardization institution for ATM) has decided [253] that ABR sessions will not be subject to CAC, unless they require from the network a guarantee on minimum cell rate. (An ABR with minimum cell rate can adapt its transmission to the available bandwidth: it will use more bandwidth if available, but it still requests a minimum guaranteed bandwidth.)

The following questions then arise:

- How to control acceptance of ABR sessions when they do have a minimum cell rate requirement?
- How to handle call acceptance of CBR or VBR traffic in the presence of ABR traffic?

In [31], the authors study the second problem using diffusion approximations under general distribution of interarrival times. They obtain numerically policies that have a monotone switching curve structure and show that a substantial improvement in the performance of ABR traffic classes can be obtained at the price of a slight increase in the rejection rate of CBR and VBR traffic classes. In [136] the monotone switching curve structure is proved to be optimal in the case of exponential distributions.

In [213, 212, 216], routing and call admission of combined CBR, VBR and ABR is considered. Here ABR traffic is also subject to admission control as it is assumed to have a minimum cell rate requirement.

15.5 BUFFER MANAGEMENT AND PACKET ADMISSION CONTROL

Admission control may occur not only at a session level, but also at a packet level. In the absence of packet admission control, packets that arrive to routers queue in dedicated buffers; if the buffer overflows then the packets are lost. In both ATM as well as in the Internet environment [166, 96] it has been

recognized that performance may be improved if the network rejects arriving packets even before the buffer overflows. This is especially the case when we wish to provide different performances to packets of different priorities. Yet, even in the case of a single priority, if real time interactive applications are used (such as video or voice) then it may be useful to discard packets before buffers overflow in order to avoid large queueing delays.

There has been a rich literature on optimal acceptance of customers into a single queue (or a network of queues) using MDPs, either directly related or with potential application to the above queueing model, starting from the late 1960s and beginning of the 1970s [209, 238, 280]. Later references are [63, 46, 137, 239, 241, 245], see also [232, 233].

In general, the type of results obtained in these papers is the existence of threshold optimal policies that reject packets if and only if the queue size exceeds some level.

Surprising structural results have are obtained in [129] when considering multiobjective control problems. They consider the problem of minimizing the expected average delay under some lower bound constraint on the expected throughput, or alternatively, the problem of maximizing the expected throughput subject to an upper bound constraint on the expected average delay. Optimal policies are shown to be of threshold type, but in addition to that, there is one state in which randomization (between acceptance and rejection) is required. Surprisingly, this randomization is performed not necessarily at the threshold, so the optimal policy is not monotone.

Admission control with *no state information* has been studied in [111] using the concept of multi modularity. This has been extended to a framework of a network in [19, 16]. The case of delayed (and other) information was studied in [18, 36, 169].

Remark 15.1 In flow control problems the source has to decide at which rate to send packets. The admission control problem can thus be viewed as a special case of a flow control problem, in which only two actions exist: that of transmitting at rate zero (corresponding to rejection) and that of transmitting at maximum rate (corresponding to acceptance). Thus much of the literature on flow control can be of potential use for admission control.

We would finally like to mention [10] in which combined rate-based flow control and admission control are handled, and the reference [3] in which the control of admission and of service (in a non-zero sum stochastic game setting) are considered.

Other issues in buffer management in which MDPs have been used are given in [82, 235]. These concern optimization of policies for sharing a buffer between several classes of arrival streams.

15.6 FLOW AND CONGESTION CONTROL

In many applications, the users have to adapt the transmission rate of the packets they transmit to the instantaneous state of the network. There are two main approaches to do so: the rate-based flow control, in which the rate is directly controlled, and the window-based control, in which the window size

is to be determined dynamically. More precisely, when a packet reaches the destination the source receives an acknowledgment. The window is defined to be the number of packets that are allowed to be transmitted before being acknowledged. If this number is fixed to, say, one hundred, then the source can first send one hundred packets, and then later send one packet per each acknowledgment that returns from the destination. Depending on some parameters that can be measured (as delays, or losses), the source can dynamically change the window size. The rate-based flow control is used in ATM networks [253], whereas the window-based flow control is used in the Internet [135].

Many papers have been devoted to flow control into a single queue. The network has often been modeled as a single bottleneck queue in which congestion occurs, and the rest of the network just adds to additional propagation delay. This has been justified by both experimentation and theoretical analysis [64, 67, 130, 131, 258]. In the last three references, optimal window-based flow control is obtained in the case of several sources and in a whole network. The problem is solved by reducing it to an equivalent single queue models.

A key research issue has been to establish the optimality of flow control policies that are monotone in the state, under different information structure. The monotonicity in the case of a single queue has been studied in [18, 29, 33, 36, 232, 233, 246, 245]. In the case of two actions, these monotone policies become threshold policies. The monotonicity in the case of a whole network is treated in [105, 106, 265] (see also [30, 148] for extensions), based on the submodularity concept [255] (see also [272]). In general, the proof of the monotone structure uses value iteration and it requires that the cost be convex nondecreasing. The convexity is, however, not necessary in the case of expected average cost, see [246] and [3].

Note that the flow control in a network often has the form of the control of service rate of a queue since by controlling it we determine the flow rate into the downstream nodes. Typical objectives are to minimize expected delays or other monotone functions of the workload or number of queued packets. In the case of finite queues, one often considers the minimization of losses as an objective. In that context, a tandem queueing system with finite buffer capacities is analyzed in [157], where both queues are fed by exogenous arrivals. The first server can be stopped so as to avoid congestion at the downstream queue. It is shown that the optimal control policy for the minimization of the total loss rate (in both queues) has a monotone switching curve. The authors then compare the optimal policy to other policies that are simpler to implement.

Monotone structure for flow control has also been obtained in the framework of constrained MDPs and in stochastic games. In the setting of constrained MDPs, the expected delay is to be minimized subject to a lower bound on the throughput, or vice versa: the throughput is to be maximized subject to an upper bound on the expected delay [4, 65, 66, 130, 131, 179]. Adaptive implementations of this type of constrained problems can be found in [189, 190]. In the framework of zero-sum stochastic games, the service rate may change in time in a way which is unpredictable to the flow controller [1, 2, 22]. The server is then modeled as an adversary player and the flow controller seeks to guarantee the best performance under the worst case behavior of the server.

(The opposite case, in which the flow rate is controlled by an adversary player, is studied in [23].) Finally, a model that combines non-zero sum stochastic games with constraints has been analyzed in [161].

An alternative framework for the study of flow control is the LQ (linear dynamics, quadratic cost) model. A plausible objective is to keep queue lengths around some desired level. This objective is derived from the fact that when queues are large then the risk of overflow increases, which results in undesirable loss of packets. On the other hand, when queues become empty then the output rate of the queue cannot exceed the input rate (which is required to be, on the average, lower than the service rate, in order to avoid instabilities); an empty queue thus results in loss of throughput. By letting the queues' size track in an appropriate way the desirable level, optimal performance (in terms of throughput and losses) can be achieved. When the queue seldom becomes empty, the dynamics of the queue can be well approximated as being linear in the control (the input rate). The well known linear quadratic framework is obtained by setting the immediate cost to be quadratic in the deviation from the desired queue level. Other objectives in terms of (undesirable) transmission rate variation, or in terms of (desirable) good tracking of the available bandwidth can be included in that framework, see [6, 9]. The available bandwidth may be described as an ARMA model, which is also suitable for the description of additional noise in the information available to the controller.

One advantage of this setting is that delayed information is not difficult to handle [7, 9]. Moreover, it allows one to handle the case of several controllers (modeled as a team problem) with different action as well as information delay [11, 12, 13]. Finally, this setting allows one to obtain an explicit solution of the problem: both the optimal policy as well as the optimal value can be explicitly computed.

15.7 ROUTING

Routing consists of determining the route of each packet from a given source to one or more destinations. There are networks in which each packet may use a different route (this is frequently the case in the Internet), such networks are called packed-switched networks. In such cases dynamic routing control is performed on a very granular scale. In contrast, in networks based on a circuit-switched architecture (such as traditional telephony or ATM networks), routing decisions are made for each connection, and all packets of the connection use the same path.

Among the mostly used routing algorithms are the Bellman-Ford algorithm, the Dijkstra Algorithm and the Floyd-Marshall Algorithm [51, Sec. 5.2]. In all these algorithms, the dynamic programming principle plays a central role. The objective of routing algorithms is to find the shortest path between nodes, either in terms of number of hops, or in terms of the link length (delay). The latter may change in time either due to link failures and repairs, or due to changing traffic conditions in the network. Routing algorithms should therefore adapt to such changes. Routing is further decentralized: finding shortest paths involves computations that occur in parallel in various nodes of the network. Finally, it is typically asynchronous, in the sense that updates in different nodes do not

occur at the same time. Some examples of such algorithms are the one used in the original 1969 ARPANET [51, p. 327], and the RIP (Routing Information Protocol) [193] (the latter reference can be considered to be the standard of the Internet routing).

The above characteristics of routing algorithms, namely adaptivity to time changes, asynchronicity and decentralization may cause oscillations and stability problems. Discussions on these problems can be found in [51, Sec. 5.2.5]. Therefore an important research issue on these asynchronous dynamic programming algorithms is the study of their convergence and correctness. An example of such a proof can be found in [51, Sec. 5.2.4].

The general approach of routing along shortest paths is well suited to routing of packets, since the routing decision for one individual packet has a negligible impact on the delay along that path and the computed delay will be approximately the one to be experienced by that packet. Thus from a theoretical point of view, the routing decisions result in solutions that is related to the concept of Wardrop equilibrium (Section 15.2). An important feature in Wardrop equilibrium is that all routes from a source to a destination that are actually used have the same delay. This property is appealing in applications in telecommunications, since it reduces the overhead of resequencing in case different packets take different routes (see discussion in [109]).

Below we further describe research on routing in packet-switched and circuit-switched networks. We mention additional issues that arise in routing in mobile networks in Section 15.10.

15.7.1 Routing and admission control in circuit-switched networks

When routing whole sessions rather than individual packets along fixed paths (as is typically the case of telephony or of ATM networks), the routing decisions will have a non negligible impact on the delay on that path, and thus on the delay experienced by the session to be routed as well as by future sessions. Other types of dynamic programming formulations have frequently been used in those cases.

In case that all packets of a call have the same route, the Call Admission Control is coupled with the routing problem: the question is whether there exists a route along which we should accept a connection. This is the *call setup problem*. Sometimes there is the possibility of taking alternative long routes in case that a direct route is occupied by other calls, and the call admission controller has to decide whether or not to use the alternative route. To illustrate this, consider three nodes: A, B and C, and assume that between each two nodes there is a direct link with a capacity to handle one hundred calls. If all the capacity between A and B is used, one can still attempt to route a call from A to B through the point C. This path is called an *alternative route*.

In networks that use alternative routing whenever a direct route is not available, it has been observed that bi-equilibria behavior occur: the network spends a long time in an uncongested mode and a long time in a congested mode. In the uncongested equilibrium, many direct links are available. In the congested mode, many ongoing calls in the network use alternative routes and the blocking probability of a new call is high. To understand this, note that an alternative

route uses more resources than a call established on a direct link, and it can increase congestion in the network: the danger with using alternative routes is that if a large number of connections are actually routed through alternative routes then the amount of resources used is probably high, and the chances that a new call will find a direct route is small. If accepted, it would probably also require an alternative route which will further increase congestion and would further decrease the chances that the next call will be accepted or will find a free direct link.

To avoid the congested mode, trunk reservation is often used: it is a policy that does not allow a connection to use an alternative route on a link if the free capacity on the link is below some threshold. The remaining capacity of the link (the trunk) is reserved for direct calls. This means that new calls can be rejected even if there are available resources to handle them (see e.g. [39, 144, 133]).

In the 1980s, dynamic and adaptive routing has been introduced into telephony, which resulted in substantial improvement in network performance and reliability [86]. A number of methods have been proposed for adaptive routing: methods based on decentralized adaptive schemes [210], centralized time-variable schemes and adaptive methods based on the least loaded path (for more details and references, see [86]). Several papers use MDPs to compute off line state-dependent routing and admission control (see e.g. [86, 164, 165] and references therein). Since the state space is huge, it is impossible to obtain an optimal policy except for small networks. However, policies with good performance are obtained as follows. First, some independence assumption is used which allows us to decompose the network problem into a set of MDPs each related to another link. The optimal policy for the decomposed problem is easy to compute, but it is not optimal for the original problem for which the decomposition assumption does not hold. But based upon this policy, one obtains an improved policy using a one step policy improvement iteration [180, 163, 164, 165, 229] or several steps of policy improvement [86, 256]. The low complexity of this approach allows us to obtain good policies based on real-time measurements, see [86]. Another method that uses a similar decomposition approach as a starting point is given in [39].

In some cases of regular topologies (for example, a fully connected networks), it becomes easier to obtain good policies as the network becomes very large. In [133] the policy that uses the least loaded alternative route with trunk reservation is shown to be asymptotically optimal as the number of nodes grows to infinity. For more references on routing in circuit-switched networks using MDPs, see [118, 114].

Diffusion approximations in routing of calls has been studied in [176].

15.7.2 Routing and Admission in packet-switched networks

In addition to the question of decentralized asynchronous dynamic routing in networks, other theoretical questions have attracted much attention. In particular, many researchers have considered the case of routing into parallel queues. Two types of structural results have been obtained: the optimality of monotone switching-curve policies, and the optimality of policies that send a

packet to the shorter queue, if it has a faster server. These characterizations of optimal policies may seem quite trivial (although it is not at all straightforward to prove their optimality). Yet, they turn out to hold under quite specific conditions. In fact, several counterexamples have been presented in which this structure does not hold once we deviate from these conditions, see e.g. [273] and [41].

A routing policy into N queues is described by partitioning the space state into disjoint sets S_i, such that in set i it is optimal to route to queue i, $i = 1, ..., N$.

Monotone switching curves. In [110, 112, 257, 277], an optimal policy is shown to have the following structure under various assumptions, for routing into two queues: the two sets S_i are separated by a monotone curve; for each given state x_i of queue i there is a threshold L^i such that it is optimal to route packet to queue $j \neq i$ at state (x_i, x_j) if the number of packets x_j at queue j is smaller than $L^i(x_i)$.

This type of results is extended to different information structure and to more than one controller. The delayed information is treated in [18, 40, 168], and other information patterns (including the sampled information and the case of no information) are handled in [18].

Joining the shortest queue with fastest server. In [90, 124, 138, 195, 139, 140, 276], an optimal policy is shown to have the following structure under various assumptions when routing into N queues. If the available information is the number of packets in the queues then a packet should be routed to a queue if it has the smallest number of customers. If the workload in the queue is known, then the routing is done to the queue with the shortest workload.

This type of result is extended to different information structure and to more than one controller. The delayed information is treated in [168].

An extension to a game setting is given in [3, 22] and references therein.

Diffusion approximations for routing have been used in [98, 99, 142] and references therein.

15.7.3 Routing with no state information

An active area of research in packet-switched network is the routing to N queues, or N servers, or N networks with no state information.

In the particular case of symmetric queues, the optimality of a round robin policy has been established in [20, 187, 263] (for the case of average costs). The case of finite horizon with no state information, but with a given prior distribution has been studied in [200, 201].

The case of routing to N networks (not necessarily identical) that are linear in the so called "$max+$" algebra, is studied in [20]. Based on the theory of multimodularity policies with some regular properties are shown to be optimal in many cases, under very general assumptions on the service and arrival distributions. The objective is to minimize expected average waiting times or workloads (or convex functions of the latter). This framework includes as special cases the routing into N parallel queues.

In [14, 17, 150] similar structural results have also been obtained in the case of routing to N parallel servers *with no buffering*, with exponential service times, where the throughput is to be maximized. It has been shown in [14] that even for *non Markovian inter-arrival times*, the control problem can be transformed into an equivalent MDP with full information where all but a finite number of states are recurrent. For practical purposes, this means that an MDP with a finite number of states and actions can be used to solve this problem. Moreover, it is shown that there exist optimal deterministic *periodic* policies. In the particular case of two queues, the optimal period is of the form $(1, 2, 2, ..., 2)$, where 2 corresponds to the faster server.

15.7.4 Routing after queueing

A special routing problem that is somewhat different than the previous one is that of routing after queueing: there are several servers with different speeds. In general it is optimal to send a packet to the faster server if it is not busy; the question is then whether a a packet should be sent to a slow server (or should we wait until the fast server is free). In the case of two servers it has been shown in [182] that there is some threshold on the number of queued packets: an arriving packet should be routed to the slow server if and only if the number of packets in the is below the threshold. The original lengthy proof is based on a policy iteration argument. Since then a simpler sample-path proof has been presented in [262], and an even simpler proof based on dynamic programming has appeared in [154].

Surprisingly, this intuitive structure of the optimal policy does not extend to more than two queues, see [41].

15.8 SCHEDULING OF SERVICE

Optimal service scheduling models many scenarios in telecommunications: access control to a communication channel, dynamic priority assignment between different traffic types and dynamic bandwidth allocation in ATM networks.

15.8.1 Infinite queues and linear costs: the $c\mu$ rule

One of the most studied problem in stochastic optimal scheduling of service is the one in which there are K parallel infinite buffer queues, and the average weighted expected sum of queue lengths (or of work load) in the queue is to be minimized. The weight factors are given by some positive constants c_i, $i = 1, ..., K$. The service times in queue i are assumed to be exponentially distributed with parameter μ_i whereas inter-arrival times are generally distributed. It has been known already from the early sixties that the optimal policy is a fixed priority policy; it is the so called "$c\mu$" rule [84, p. 84-85]: the different queues are ordered according to the decreasing order of the product of the weight c_i times the service rate μ_i, and a queue is served only if those queues with a higher product of $\mu_i c_i$ are empty.

These results have been adapted to other frameworks, in particular to the discrete time setting, see [44, 43, 75].

Other extensions and generalizations both in the model as well as in the proof techniques can be found in [78, 79, 126, 153, 156, 203, 206, 259, 263] (many related references are presented in [263]). In particular, the case in which packets can be rerouted and change class after they terminate their service is analyzed in [146] see also [55, 268]. We should mention that the first proofs have used the theory of Bandits for which it was known that optimal index (Gittins index) rule exists, see [104, 261, 268, 274, 275].

This problem received particular attention in the setting of MDPs with additional constraints. As an example, consider the problem where different interactive and non-interactive traffic compete for the access to a single channel. One may wish to minimize the expected delay of the non-interactive traffic, but yet to impose bounds on the average delays of interactive traffic. The case of a single constraint has been studied in [205], who proposed a randomized mixture at each step between two fixed-priority policies; they show the existence of an optimal policy within this class. The case of several constraints has been solved in [35] who use a finite linear program to determine an optimal policy within the class of so called "time sharing policies": the controller switches between different fixed priority policies when the queues all empty. The linear programming determines the relative frequency at which each one of the finite number of fixed priority policies should be used. The solutions in both references [35] and [205] use a Lagrangian argument that transforms the control problem into a nonconstrained one, for which the $c\mu$ rule is known to be optimal.

As a side result of the structural results of optimal policies in [35], the authors characterize the achievable set of performance measures attained by *any* policy. When restricting to policies that do not idle when the system is non-empty, this region is shown to be a polytope whose extreme points correspond to fixed priority policies.

Later, a reverse approach can be used to solve the constrained problem: first the achievable region is characterized, see [91, 92], which then allows one to obtain the solution of constrained control problem, see [55, 52, 234]. This approach turned out to be quite powerful in more complex control problems as well, see [53, 54, 211]. The achievable region is often referred to as *conservation laws*, see e.g. [263, p. 258].

The type of results for the constrained control problem has been extended to a network in [226]. Yet one should be careful when considering networks: there is a counterexample that shows that the $c\mu$ rule is not optimal in the second node in a tandem network, see [125].

The $c\mu$ rule has been shown in [28] to be optimal for a problem of scheduling service in parallel queues in a setting of a stochastic zero-sum game (that corresponds to a worst case control) where the arrivals are MADP (Markov Decision Arrival Processes), an extension of MAP (Markov Arrival Processes) that allows the arrival process to be controlled and to depend on the state of the system.

15.8.2 Other scheduling problems

We briefly mention other scheduling problems. Scheduling with deadline constraints has been studied in [57].

In the case of finite queues, it has been observed in [236] that the control model is equivalent to that of optimal routing. This observation has been used in [14, 21] in which stochastic scheduling of service with no state information is studied. Other references on the case of no information are [121, 134].

A special class of scheduling is polling problems, which we describe in the next section.

15.9 POLLING

The term polling is used when a single server moves between several queues. Upon arrival to a queue, some customers in the queue are served and then the server moves to another queue. The number to be served may depend on the number of customers in the queue. An important feature of polling systems is that when the server moves from one queue to another then switching times or switching costs are incurred.

Polling systems are used to model LANs—local area networks (these are the networks that interconnect computers and printers within a department or a university) or MANs—metropolitan area networks (these are deployed over an area of up to 100km^2 and serve to interconnect local area networks). Examples of LANs that can be modeled using polling systems are the IEEE 802.4 token bus and the IEEE 802.5 token ring (Chapters 4.5.3–4.5.4 in [51]). Examples of MANs modeled as polling systems are given in [225].

Polling systems can also be used to model the scheduling of transmission of packets in the output of an ATM switch.

Interesting control problems arise in the context of polling; the *schedule problem:* in what order should queues be visited and the *service regime:* how many packets are to be served upon a visit to a queue.

In [70, 71, 281], a polling system with infinite queues is considered. The problem of which queue to visit next (for a fixed given service regime) is formulated as an MDP. The objective is to minimize the expected weighted waiting time accruing to the system per unit time. Another control problem is then formulated: the choice of a permutation of the set of queues $\{1, ..., N\}$. The server then visits these queues according to that order (using a fixed service regime in each queue). Once all the queues are visited (we call this a cycle) a new choice is made for the next cycle. The objective is to minimize the expected cycle time. This problem is completely solved in [70] leading to a simple rule which is optimal for both gated[1] as well as exhaustive service regimes[2]: the visit is according to the increasing order of the values of n_i/λ_i, where n_i is the number of packets at queue i in the beginning of the cycle, and λ_i is the arrival rate to that queue. The optimality of the same criterion under other service disciplines is established in [281].

[1] In the gated regime, a packet is served in queue i if and only if it is present there upon the arrival of the service to that queue.
[2] The server stays in a queue till it becomes empty.

A similar formulation of an MDP is derived in [72] for polling systems with a single buffer, and some optimality results are obtained under particular assumptions.

Several papers study optimal dynamic polling under various information structure. The maximization of the throughputs for polling system with a single buffer is studied in [21] in the case of no state information, whereas the corresponding infinite buffer case (in which the weighted sum of expected workload is minimized) is studied in [20]. Other information patterns are handled in [208, 184].

Other papers that consider the objective of minimization of expected (weighted) workload or weighted average queue lengths in polling systems are [26, 32, 38, 69, 68, 101, 122, 159, 181, 178, 282] and references therein. In general, the problem of minimizing the expected (weighted) workload is a difficult one and standard dynamic programming techniques have been seldom used in the references above, see [122]. Characterization and structure of optimal policies have been derived using sophisticated sample path approaches [181, 184], techniques based on multimodularity [20], fluid approximations [178] and diffusion approximations [32]. Some papers restrict to a given rich class of policies and consider the optimization among this class [37, 59, 60, 61, 62].

15.10 WIRELESS AND SATELLITE COMMUNICATIONS

In wireless and mobile satellite networks there are other dynamic control problems that are due to the mobility.

15.10.1 Control of handover and admission

If a call is accepted and established then the mobile terminal transmits and receives information from a local base station that covers the area in which it is located. This area is called a cell. If a mobile moves from one cell to another, then a handover of the communication to the new base station has to take place in order for the call to continue. Usually one puts more importance to maintain an existing call than to accept a new one. Thus when accepting a call at a cell, one should have in mind that this decreases the number of calls that can be handed-over to that cell from neighboring ones. Several papers have used MDPs to solve the call admission control for mobile networks, see [119, 221].

Another related problem is how to decide whether and when to perform a handover. The decision may take into account the movements of the mobile and the strength of the signal. A solution of this problem using an MDP can be found in [223].

15.10.2 Routing in wireless networks

More complex routing problems occur in the case where users are mobile. Moreover, the nodes themselves may be mobile, as is the case in satellite communications using Low Earth Orbit satellites [100]), or in so called "Ad-hoc Networks" [109, 220]. In Ad-hoc networks, adaptive decentralized routing is again used,

but it should be very sensitive to changes in the topology. In satellite communications one can often use simpler routing schemes since the movement of the routers (the satellites) is predictable. Routing in such networks is an active area of research. To the best of our knowledge, dynamic programming or MDP tools have not yet been applied to routing through moving satellites, whereas in Ad-hoc networks, dynamic programming is still the basic tool for determining shortest paths.

15.10.3 Scheduling transmission opportunities

The scheduling of transmission opportunities between the mobile terminals is determined by the base station. An important factor that is specific to wireless communication is that there may be disconnectivity problems due to the fact that the communication link is more vulnerable to physical obstacles, fading and noise. Some papers that have addressed this control problem are [32, 77, 151, 250, 252].

Another related problem is that of random access to a common channel in a distributed setting, i.e. in the absence of a base station. In that case, users send their packets independently of each other; simultaneous transmissions by more than a single user result in collisions and in the need of retransmission. This type of random access to a common wireless channel was first used in 1970 to interconnect the islands of Hawaii and is still quite common today in satellite communications. Control is needed to determine the retransmission probabilities so as to avoid a large number of collisions during that phase. Finding an optimal retransmission probability in slotted ALOHA systems was modeled in [94] as a finite state MDP with compact action spaces. It was shown that this MDP is unichained and several properties such as monotonicity and estimates for optimal policies were derived. By using natural initial approximations for an optimal policy, a policy iteration algorithm was implemented. The algorithm computed optimal retransmission probabilities after few iterations.

15.10.4 Other control problems

Some other control problems in both wireless and satellite networks are mobility tracking [45, 191], and energy control [279]; for the latter problem we do not know of an MDP solution approaches, but we believe that it could be used.

15.11 MDPS IN APPLICATIONS OF THE WORLD WIDE WEB

Most of today's traffic on the Internet is World Wide Web traffic, which makes it an important field of application of optimal control techniques. The World Wide Web offers search engines, such as Altavista, Lycos, Infoseek and Yahoo, that serve as a database that allows one to search information on the Web. These search engines often use robots that periodically traverse a part of the Web structure in order to keep the database up-to-date by copying the Web pages from around the world into the database. An important control problem is the efficient design of these search engines. In particular, the question that arises is how often should pages be fetched in order for the information in

the database to be updated. It is required to minimize the probability that a request for an information on a page finds that page in the database out-of-date, i.e. that the page has since been modified but the new version has not yet been updated in the database. Fetching pages is a costly operation, and efficient updates (taking into account the frequency that a page is requested) is crucial.

Several papers have studied this problem using MDPs. In [83], the authors consider a problem where there is a fixed number of M Web-pages. The contents of page i is modified at time epochs that follow a Poisson process with parameter μ_i. The time a page is considered up-to-date by the search engine is the time since the last visit by the robot until the next time instant of modification; at this point the Web-page is considered out-of-date until the next time it is visited by the robot. The times between updates by the robot are given by a sequence τ_n, assumed to be i.i.d. A simple policy based on the golden-ratio approach is shown in [121, 134] to perform close to optimum. In [14] the i.i.d. assumption is relaxed and the τ_n sequence is only assumed to be stationary. The problem is then solved using a finite MDP, and the existence of periodic optimal policies is established. The solution, using MDPs of some related problems can be found in http://www.path.berkeley.edu/~guptar/webtp/index.html.

Other applications of MDPs to the control of search engines in the Web can be found in [247, 248].

15.12 SOLUTION METHODOLOGIES

We shall survey in this section some solution methodologies that were successful in problems in telecommunications. Some of these are classical and are related directly to general solution methods of dynamic programming equations. Some other solution methods, however, make use of special properties of queueing networks which are useful in modeling problems in telecommunications.

We have already mentioned in the previous sections the use of Bandits together with the Gittins index for scheduling problems, and the use of conservation laws (which are especially used in scheduling as well). We further discussed in the part on flow control the advantage of linear quadratic framework, which could be also used in other contexts (e.g. in dynamic bandwidth allocation). Below we present several other techniques and methodologies which are very useful in telecommunication applications.

15.12.1 Structural characteristics of optimal policies

Due to the curse of dimensionality of dynamic programming, researchers have been interested in inferring the structure of optimal policies and/or of the value function. In some cases, when one knows the structure of optimal policies, the original optimal control problem can be reduced to that of an optimization problem over a small subclass of policies that possess the required properties.

A very popular method for obtaining the structure of optimal policies goes through value iteration as follows. Under quite general conditions, if one knows the value function then one can compute the optimal policies as the policy that chooses the argument of the optimization (maximization or minimization) of

the corresponding dynamic programming. The question is then how does the argument of the optimization behave as a function of the state. This of course depends on the properties of the immediate rewards (or costs), the transition probabilities, and the value of the dynamic programming. For a given set of transition probabilities, one can often show that if the immediate rewards and value have some properties (such as convexity or concavity or submodularity) then this implies the required structure of the argument of the optimization and thus of the optimal policy (see e.g. [147]).

For example, in many one-dimensional queueing systems, if the value function is convex then the policy that minimizes the costs is of threshold type [106, 148]. In two dimensional problems, submodularity often implies the optimality of monotone switching curve policies [148]. In routing problems, weak Schur convexity of the value function implies typically the optimality of the policy that routes a customer to the shortest queue [124].

In order to obtain the required properties of the value function, one proceeds by value iteration. One first checks that the terminating cost has this property; in case of infinite horizon (discounted cost) one can choose some arbitrary terminating cost (typically a cost that is everywhere zero) that has the required property (which will not have an influence on the value for the infinite horizon). Then one checks by iterating the dynamic programming operator that the value for the n-step horizon also possesses the required structure for any integer n. For the infinite horizon case, one then has to establish that the property also holds for the limit (as n tends to infinity), which coincides with the value for the infinite horizon (under fairly general conditions).

The above approach can be used not only to establish properties of an optimal policy or the optimal value, but also to compare the values of different policies or the value corresponding to different statistical assumptions. For example, this method is used in [158] to show the advantage of stochastic multiplexing of many small sources, by comparing the performance of one big source to that of several small sources.

We finally note that other techniques can be used to establish structural properties of optimal policies, and in particular sample path methods.

15.12.2 Sample path methods

Several sample path techniques have been used for solving MDPs in queueing applications, see e.g. [203, 207, 206, 237, 263]. A thorough survey on these methodologies can be found in [185]. The most frequently used are interchange arguments, in which one can show that by interchanging the order in which actions are taken, the policy can be improved. This technique has been used for both routing as well as scheduling problems, see e.g. [263]. This approach is based on coupling: one constructs on the same probability space the evolution of a two stochastic systems that differ only in the control, but not in the driving sequence (such as interarrival times or service times). An opposite approach is, on the contrary, to change the probability space and solve the control problem in the new space instead, which if chosen appropriately is simpler to solve. In some cases, although the probability spaces are different, the costs depend only on some marginal (rather than joint) distributions and is thus the same

for the two models. An example is routing to parallel symmetric queues with no state information (for appropriate costs). One constructs a new model for which (potential) service times are the same in all queues, which allows one to establish the optimality of a round robin policy for the new model. It then turns out that this policy is also optimal for the original model [263, p. 264].

15.12.3 Stability analysis

There is a whole class of MDPs that arise in telecommunications for which the solution goes through stability analysis, and typically through Lyapunov function techniques.

As an example, consider the assignment or polling problem when the connections between each queue and the server is broken at random times and for random durations. The possibility of such interruptions complicates the control problem considerably, since the possibility that any queue might not be available to the server at any future time needs to be accounted for in choosing the current server allocation. This problem is typical for wireless communications in which noise can cause disconnectivity problems. The goal is to maximize the throughput. The solution approach is to first find stability conditions that are *necessarily* for an arbitrary policy, and then show that a the same stability condition is a sufficient condition under some given candidate policy. This approach has been used in [250, 252] where it was shown that the policy that serves the longest connected queue is optimal. Lyapunov functions that are quadratic in the state were used to obtain the stability condition. A related problem in satellite communications was solved with a similar technique in [77]. The same problem as in [250] was later solved under more refined criteria [32] using diffusion approximations.

In [249, 251] state dependent routing and flow control is considered in a queueing network with arbitrary topology. The routing is based on local state information. In addition, the rate of a server is controlled based on local information (which means that the outflow is controlled). A distributed policy is shown to achieve maximum throughput in the case of delayed state information [251].

Optimal scheduling in another type of network topology is considered in [103], namely ring networks. Again, a scheduling policy with maximal stability region is obtained.

A further discussion of stability issues in our context can be found in Chapter 9 of this volume.

15.12.4 Fluid limits and diffusion approximations

Telecommunication systems are usually described as discrete event systems, where discrete units of information (bits, cells, packets or sessions) are transmitted at discrete times. Due to the curse of dimensionality of dynamic programming it is often impossible to solve optimal control problems in networks within this framework. Two less granular approaches are the fluid limits and the diffusion approximations, in which the transmission of discrete information units is approximated by the transmission of a deterministic fluid and

of a stochastic fluid, respectively. In both approximations, the fluid represents the expected workload or the number of customers that arrive, that are present, and that leave the system. Using functional laws of large numbers for the fluid approximation and functional central limit theorems for the diffusion approximation, one can typically show that the approximating processes become tight at high loads. The limiting processes allows us to approximate the value and to obtain policies that are almost optimal for the original system. Public domain software exists which is specially adapted to the use of diffusion approximations for telecommunication problems, see http://www.dam.brown.edu/lcds/software.html. Fluid and diffusion approximations often have the advantage of a collapse of the dimension of the state space: in many cases, a high dimensional problem reduces to a lower dimensional one in these approximations, see e.g. [24] for the case of fluid approximations and [31] for the case of diffusion approximations.

Some other papers using fluid limits in control of queues in this line are [24, 39, 81, 80, 88, 178, 198, 196, 197, 199, 269, 270, 271]. A public domain animations of fluid control corresponding to the papers [269, 270, 271] is available on the Web in the home page http://rstat.haifa.ac.il/ gweiss of Weiss. Some references on diffusion models are [115, 116, 142, 172, 173, 174, 175, 176, 177, 217, 222, 266, 267].

In yet another type of fluid models, one consider two time scales. The distribution of the arrival and/or service in a network may vary in some stochastic way (whose distribution is possibly controlled) on a time scale which is much slower than the transmission time of packets. In that case one uses some environment state to describe the slow variation of the distributions of the parameters of the network, and then uses a fluid approximation only to replace the granular arrivals and service in each given environment by a fluid. In that context, recent work [25] shows that fluid limits are not only approximations that become asymptotically tight (in some appropriate scaling), but also that they give in fact *optimistic bounds* on the performance.

Using this approach, the authors in [10] consider combined admission and flow control. The state of the network depends on the number of sessions of different traffic types, which varies at a time scale much slower than the controlled transmission of packets. Termination of sessions (and thus the decrease in their number) occurs according to an exponential distribution. The arrival of new sessions of different types occur according to a Poisson process, and the admission control can decide to accept or reject an arrival of a session. Once in the system, the rate of transmission of packets (modeled as fluid) of each session, is controlled. Another example of this approach is the reference [244] where the authors consider a fluid model for the optimal flow control.

15.12.5 Power series algorithm

The power series algorithm, developed originally in [123] and further in [58, 155] in the context of non-controlled Markov chains, allows one to obtain the performance measures by a recursive numerical method that is based on the expansion of the value function in the load parameter. Recently this method has been extended to MDPs and optimization of Markov chains in [59, 160] and

applied to several problems in optimization and control of queueing systems, which can be used to model scenarios in telecommunications [61, 62].

15.12.6 Neuro-dynamic programming

Neuro-dynamic programming (NDP), also known as reinforcement learning, is a recent class of methods that can be used to solve very large and complex dynamic optimization problems [50]. NDP combines simulation, learning, neural networks or other approximation architectures and heuristics, with the central ideas in dynamic programming. It provides a rigorous framework for addressing challenging and often intractable problems from a broad variety of fields. As such, it is a promising tool for solving large scale control problems in telecommunications. Applications of this methodology in telecommunications can be found in [50, Sec. 8.5], in [76, 194, 214, 254] and in references therein. A detailed survey on NDP can be found in Chapter 13 of this volume.

15.13 CONCLUDING REMARKS

We provided a survey of the areas in telecommunications in which MDPs have been applied and have potential application. In addition, we surveyed some modeling issues (multiagent and information issues) as well as solution techniques (other than the standard dynamic programming) that are special to telecommunication applications of MDPs.

In preparing this survey we have interviewed around forty researchers working on MDPs and on telecommunications. Our general impression from these interviews and from the preparation of the survey are the following.

(i) Communication networks is a very rich area of application that has an impact on MDPs, including the development of theoretical tools that seem adapted to problems encountered in telecommunications.

(ii) Optimal control and applied mathematics are not central in the development of communications and network technology, as opposed perhaps to areas such as aeronautics, robotics. Some people interviewed regretted that persons that venture into today's communication problems do not have a solid background of the fundamentals issues addressed by control theory.

(iii) Dynamic programming techniques have an important impact in some areas, and in particular on routing (Bellman-Ford and other algorithms). Theoretical work using MDPs that were mostly cited in the interviews as having an impact in telecommunications is that of [86, 165] on admission control and routing. For the theoretical work on state dependent routing which included the MDP approach [86], as well as for implementation of state dependent routing, in the Bell Canada and STENTOR networks, H. Cameron, Z. Dziong, A. Girard and L. Mason and J. Regnier received the prestigious Stentor award for Industry-University Collaborative Research in Telecommunications.

(iv) We believe that MDPs have an important potential for applications in telecommunications, in particular in the areas of scheduling in ATM switches, in buffer management schemes (admission of packets of different priorities to buffers) and in flow control.

References

[1] E. Altman. Flow control using the theory of zero-sum Markov games. *IEEE Transactions on Automatic Control*, 39:814–818, 1994.

[2] E. Altman. Monotonicity of optimal policies in a zero sum game: A flow control model. *Advances of Dynamic Games and Applications*, 1:269–286, 1994.

[3] E. Altman. Non zero-sum stochastic games in admission, service and routing control in queueing systems. *Queueing Systems*, 23:259–279, 1996.

[4] E. Altman. *Constrained Markov Decision processes.* Chapman and Hall/CRC, 1999.

[5] E. Altman. A Markov game approach for optimal routing into a queueing network. *Anals of Dynamic Games Vol 5: Stochastic and Differential Games, Theory and Numerical Methods, M. Bardi, T.E.S. Raghavan and T. Parthasarathy (Editors), Birkhauser Boston, Basel, Berlin*, pages 359–376, 1999.

[6] E. Altman and T. Başar. Optimal rate control for high speed telecommunication networks. In *Proc. of the 34th IEEE Conference on Decision and Control*, New Orleans, Louisiana, USA, Dec. 1995.

[7] E. Altman and T. Başar. Optimal rate control for high speed telecommunication networks: the case of delayed information. In *First Workshop on ATM Traffic Management, WATM, IFIP, WG.6.2 Broadband Communication, Paris*, pages 115–122, Dec 1995.

[8] E. Altman and T. Başar. Multi-user rate-based flow control. *IEEE Trans. on Communications*, pages 940–949, 1998.

[9] E. Altman, T. Başar, and N. Hovakimian. Worst-case rate-based flow control with an arma model of the available bandwidth. *Anals of Dynamic Games*, 6, 1999.

[10] E. Altman, T. Başar, and Z. Pan. Piecewise-deterministic differential games and dynamic teams with hybrid controls. *Anals of Dynamic Games*, 6, 1999.

[11] E. Altman, T. Başar, and R. Srikant. Multi-user rate-based flow control with action delays: a team-theoretic approach. In *Proc. of the 36th IEEE Conference on Decision and Control*, San Diego, California, Dec. 1997.

[12] E. Altman, T. Başar, and R. Srikant. Robust rate control for abr sources. In *IEEE INFOCOM, San-Francisco, California, USA*, 1998.

[13] E. Altman, T. Başar, and R. Srikant. Congestion control as a stochastic control problem with action delays. *Automatica*, 1999.

[14] E. Altman, S. Bhulai, B. Gaujal, and A. Hordijk. Optimal routing to M parallel servers with no buffers. *Journal of Applied Probability*, 37(3), 2000.

[15] E. Altman, B. Gaujal, and A. Hordijk. Regularity for admission control comparisons. In *Proceedings of the 37th IEEE Conference on Decision and Control*, Dec. 1998.

[16] E. Altman, B. Gaujal, and A. Hordijk. Admission control in stochastic event graphs. *to appear in JACM*, 2000.

[17] E. Altman, B. Gaujal, and A. Hordijk. Balanced sequences and optimal routing. *to appear in IEEE Trans. Automatic Control*, 2000.

[18] E. Altman, B. Gaujal, and A. Hordijk. Multimodular value functions: monotonicity of feedback control. *in preparation*, 2000.

[19] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity and optimization properties. *Mathematics of Operations Research*, 25:324–347, 2000.

[20] E. Altman, B. Gaujal, and A. Hordijk. Optimal open-loop control of vacations, polling and service assignment. *to appear in Queuing Systems*, 2000.

[21] E. Altman, B. Gaujal, A. Hordijk, and G. Koole. Optimal admission, routing and service assignment control: the case of single buffer queues. In *Proceedings of the 37th IEEE Conference on Decision and Control*, Tampa, Florida, USA, Dec. 1998.

[22] E. Altman and A. Hordijk. Zero-sum Markov games and worst-case optimal control of queueing systems. *Queueing Systems*, 21:415–447, 1995.

[23] E. Altman, A. Hordijk, and F.M. Spieksma. Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards. *Mathematics of Operations Research*, 22 No. 3:588–618, 1997.

[24] E. Altman, T. Jimenez, and G. Koole. On optimal call admission control. In *Proceedings of the 37th IEEE Conference on Decision and Control*, pages 569–574, Tampa, Florida, USA, Dec. 1998.

[25] E. Altman, T. Jimenez, and G. Koole. Comparing tandem queueing systems and their fluid limits. *Probability in the Engineering and Informational Sciences*, 2000.

[26] E. Altman, A. Khamisy, and U. Yechiali. On elevator polling with globally gated regime. *Queueing Systems*, 11:85–90, 1992.

[27] E. Altman, D. Kofman, and U. Yechiali. Discrete time queues with delayed information. *Queueing Systems*, 19:361–376, 1995.

[28] E. Altman and G. M. Koole. Stochastic scheduling games and Markov decision arrival processes. *Computers and Mathematics with Applications*, 26(6):141–148, 1993.

[29] E. Altman and G.M. Koole. Control of a random walk with noisy delayed information. *Systems and Control Letters*, 24:207–213, 1995.

[30] E. Altman and G.M. Koole. On submodular value functions and complex dynamic programming. *Stochastic Models*, 14:1051–1072, 1998.

[31] E. Altman and H. Kushner. Admission control for combined guaranteed performance and best effort communications systems under heavy traffic. *SIAM J. Control and Optimization*, 37(6):1780–1807, 1999.

[32] E. Altman and H. Kushner. Control of polling in presence of vacations in heavy traffic with applications to satellite and mobile radio systems. In

Proceedings of the 37rd Allerton Conference on Communication, Control, and Computing, Illinois, USA, Sept. 1999.

[33] E. Altman and P. Nain. Closed-loop control with delayed information. *Performance Evaluation Review*, 20:193–204, 1992.

[34] E. Altman and N. Shimkin. Individual equilibrium and learning in processor sharing systems. *Operations Research*, 46:776–784, 1998.

[35] E. Altman and A. Shwartz. Optimal priority assignment: a time sharing approach. *IEEE Transactions on Automatic Control*, AC-34:1089–1102, 1989.

[36] E. Altman and S. Stidham, Jr. Optimality of monotonic policies for two-action Markovian decision processes, including information and action delays. *Queueing Systems*, 12 No. 2:307–328, 1996.

[37] E. Altman and U. Yechiali. Cyclic Bernoulli polling. *ZOR - Mathematical Methods of Operations Research*, 38:55–76, 1993.

[38] E. Altman and U. Yechiali. Polling in a closed network. *Probability in the Engineering and Informational Sciences*, 8:327–343, 1994.

[39] V. Anantharam and M. Benchekroun. Trunk reservation based control of circuit switched networks with dynamic routing. In *Proceedings of the 29th Conference on Decision and Control*, pages 2102–2105, Honolulu, Hawaii, Dec. 1990.

[40] D. Artiges. Optimal routing into two heterogeneous service stations with delayed information. *IEEE Transactions on Automatic Control*, 40(7):1234–1236, 1995.

[41] D. Artiges. *Controle et evaluation des reseaux de telecommunication (in french)*. PhD thesis, INRIA, Sophia Antipolis, France, 1996.

[42] D. Assaf and M. Haviv. Reneeging from time sharing and random queues. *Mathematics of Operations Research*, 15:129–138, 1990.

[43] J. S. Baras, A. J. Dorsey, and A. M. Makowski. K competing queues with geometric service requirements and linear costs: the μc rule is always optimal. *Systems and Control Letters*, 6:173–180, 1985.

[44] J. S. Baras, A. J. Dorsey, and A. M. Makowski. Two competing queues with linear costs and geometric service requirements : the μc rule is often optimal. *Advances in Applied Probability*, 17:186–209, 1985.

[45] A. Bar-Noy, I. Kessler, and M. Sidi. Mobile users: To update or not to update? *Wireless Networks journal*, 1:175–186, 1995.

[46] M. Bartroli. On the structure of optimal control policies for networks of queues. Ph.D. dissertation, Department of Operations Research, University of North Carolina at Chapel Hill, 1989.

[47] T. Başar and P. Bernhard. *H^∞-Optimal Control and Relaxed Minimax Design Problems: A Dynamic Game Approach*. Birkhauser, Boston, MA, USA, 1991 (2nd edition, 1995).

[48] T. Başar and J. B. Cruz. Concepts and methods in multi-person coordination and control. In S. G. Tzafestas, editor, *Optimization and Control of*

Dynamic Operational Research Methods, pages 351–394. North-Holland Publishing Company, 1982.

[49] I. Ben-Shahar, A. Orda, and Nahum Shimkin. Dynamic service sharing with heterogeneous preferences. *submitted to QUESTA*, 1999.

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

[51] D.P. Bertsekas and R.G. Gallager. *Data Networks.* Prentice-Hall, 1987.

[52] D. Bertsimas and Jose Nino-Mora. Conservation laws, extended polymatroids and multi-armed bandit problems; a unified polyhedral approach. *Mathematics of Operations Research*, 21:257–306, 1996.

[53] D. Bertsimas and Jose Nino-Mora. Optimization of multiclass queueing networks with changeover times via the achievable region method: Part i, the single-station case. *Mathematics of Operations Research*, 24:306–330, 1999.

[54] D. Bertsimas and Jose Nino-Mora. Optimization of multiclass queueing networks with changeover times via the achievable region method: Part ii, the multi-station case. *Mathematics of Operations Research*, 24:331–361, 1999.

[55] D. Bertsimas, I. Paschalidis, and J. N. Tsitsiklis. Branching bandits and Klimov's problem: Achievable region and side constraints. *IEEE Transactions on Automatic Control*, 40:2063–2075, 1995.

[56] F.J. Beutler and K.W. Ross. Time-average optimal constrained semi-Markov decision processes. *Advances of Applied Probability*, 18:341–359, 1986.

[57] P.P. Bhattacharya, L. Tassiulas, and A. Ephremides. Optimal scheduling with deadline constraints in tree networks. *IEEE Transactions on Automatic Control*, 42(12):1703–1705, 1997.

[58] J.P.C. Blanc. On a numerical method for calculating state probabilities for queueing systems with more than one waiting line. *Journal of Computational and Applied Mathematics*, 20:119–125, 1987.

[59] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In L. Donatiello and R. Nelson, editors, *Performance Evaluation of Computer and Communication Systems*, pages 53–80. Springer-Verlag, 1993. Lecture Notes in Computer Science 729.

[60] J.P.C. Blanc and R. D. van der Mei. The power series algorithm applied to polling systems with a dormant server. (9346), 1993.

[61] J.P.C. Blanc and R. D. van der Mei. Optimization of polling systems with Bernoulli schedules. *Performance Evaluation*, 22:139–158, 1995.

[62] J.P.C. Blanc and R. D. van der Mei. Computation of derivatives by means of the power-series algorithm. *INFORMS J. Comput.*, 2:45–54, 1996.

[63] J.P.C. Blanc, P.R. de Waal, P. Nain, and D. Towsley. Optimal control of admission to a multiserver queue with two arrival streams. *IEEE Transactions on Automatic Control*, pages 785–797, 1992.

[64] J.-C. Bolot. End-to-end delay and loss behavior in the internet. In *Proceedings of ACM Sigcomm '93, San Francisco, CA, USA*, pages 289–298, Sept. 1993.

[65] A. D. Bovopoulos and A. A. Lazar. Optimal load balancing algorithms for Jacksonian networks with acknowledgment delays. *IEEE Transactions on Communications*, pages 144–151, 1988.

[66] A. D. Bovopoulos and A. A. Lazar. The effect of delayed feedback information on network performance. *Annals of Operations Res.*, pages 581–588, 1991.

[67] O.J. Boxma. Sojourn times in cyclic queues – the influence of the slowest server. In G. Iazeolla, P. J. Courtois, and O. J. Boxma, editors, *Computer Performance and Reliability*, pages 84–88. Elsevier Science Publishers B.V. (North-Holland), 1988.

[68] O.J. Boxma, H. Levy, and J.A. Weststrate. Efficient visit frequencies for polling tables: Minimization of waiting costs. *Queueing Systems*, 9:133–162, 1991.

[69] O.J. Boxma, H. Levy, and U. Yechiali. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research*, 31:187–208, 1991.

[70] S. Browne and U. Yechiali. Dynamic priority rules for cyclic-type queues. *Advances in Applied Probability*, 21:432–450, 1989.

[71] S. Browne and U. Yechiali. Dynamic routing in polling systems. In M. Bonati, editor, *Teletraffic Science*, pages 1455–1466. Elsevier Science Pub. (North-Holland), 1989.

[72] S. Browne and U. Yechiali. Dynamic scheduling in single-server multiclass service systems with unit buffers. *Naval Research Logistics*, 38:383–396, 1991.

[73] R. Buche and H. J. Kushner. Stochastic approximation and user adaptation in a competitive resource sharing system. In *Proceedings of the 37th Conference on Decision and Control*, Tampa, Florida, USA, Dec. 1998.

[74] E. B. N. Bui. Contrôle de l'allocation dynamique de trame dans un multiplexeur integrant voix et données (in french). Master Thesis 89 E 005, TELECOM, Département Réseaux, Paris, June 1989.

[75] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Advances in Applied Probability*, 17:237–238, 1985.

[76] J. Carlstrom and E. Nordstrom. Control of self-similar ATM call traffic by a reinforcement learning. In J. Alspector, R. Goodman, and T. X. Brown, editors, *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunication 3, IWANNT'97*, pages 54–62, Melbourne, Australia, 1997. Lawrence Erlbaum.

[77] M. Carr and B. Hajek. Scheduling with asynchronous service opportunities with applications to multiple satellite systems. *IEEE Transactions on Automatic Control*, 38(12):1820–1833, 1993.

[78] C.S. Chang, X. Chao, M. Pinedo, and R.R. Weber. On the optimality of LEPT and $c\mu$-rules for machines in parallel. *Journal of Applied Probability*, 29:667–681, 1992.

[79] C.S. Chang and R. Righter. The optimality of LEPT in parallel machine scheduling. Working paper, 1993.

[80] H. Chen and D. Yao. Dynamic scheduling of a multi-class fluid network. *Operations Research*, 41(6):1104–1115, 1993.

[81] R. R. Chen and S. P. Meyn. Value iteration and optimization of multiclass queueing network, invited paper. 32:65–97, 1999.

[82] I. Cidon, L. Georgiadis, R. Guerin, and A. Khamisy. Optimal buffer sharing. In *IEEE INFOCOM*, Boston, MA, USA, Apr. 1995.

[83] E.G. Coffman Jr, Z. Liu, and R.R. Weber. Optimal robot scheduling for web search engines. *Journal of Scheduling*, 1, 1994.

[84] D.R. Cox and W.L. Smith. *Queues*. John Wiley, New York, 1961.

[85] T. Crabill, D. Gross, and M. Magazine. A classified bibliography of research on optimal design and control of queues. *Operations Research*, 25:219–232, 1977.

[86] Z. Dziong and L. G. Mason. Call admission and routing in multi-service loss networks. *IEEE Transactions on Communications*, 42:2011–2022, 1994.

[87] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. on Networking*, 1:329–343, 1993.

[88] D. Eng, J. Humphrey, and S. Meyn. Fluid network models: Linear programs for control and performance bounds. In J. Cruz J. Gertler and Eds. M. Peshkin, editors, *Proceedings of the 13th World Congress of International Federation of Automatic Control*, volume B, pages 19–24, San Francisco, California, June 30 to July 5 1996.

[89] A. Ephremides and S. Verdu. Control and optimization methods in communication network problems. *IEEE Transactions on Automatic Control*, 34:930–942, 1989.

[90] T.M. Farrar. Generalizations of the join the shortest queue rule for symmetric queues - a sample-path proof. 1993.

[91] A. Federgruen and H. Groenevelt. Characterization and control of achievable performance in general queueing systems. *Operations Research*, 36:733–741, 1988.

[92] A. Federgruen and H. Groenevelt. M/g/c queueing systems with multiple customer classes: character- ization and control of achievable performance. *Management Science*, 34, 1988.

[93] E.A. Feinberg and D.J. Kim. Bicriterion optimization of an $M|G|1$ queue with a removable server. *Probability in the Engineering and Informational Sciences*, 10:57–73, 1996.

[94] E.A. Feinberg, Y.A. Kogan, and A.N. Smirnov. Optimal control by the retransmission probability in slotted ALOHA systems. *Performance Evaluation*, 5:85–96, 1985.

[95] E.A. Feinberg and M.I. Reiman. Optimality of randomized trunk reservation. *Probability in the Engineering and Informational Sciences*, 8:463–489, 1994.

[96] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1:397–413, 1993.

[97] F. Forges. Repeated games of incomplete information: Non-zero-sum. In R. J. Aumann and S. Hart, editors, *Handbook of Game Theory*, volume 1. Elsevier, North-Holland.

[98] G. Foschini. On heavy traffic diffusion analysis and dynamic routing in packet-switched networks. *Computer Performance*, pages 499–513, 1977. Chandy, K.M. and Reiser, M. (eds.), North-Holland.

[99] G. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communication*, 26:320–327, 1978.

[100] J. Galtier. Geographical reservation for guaranteed handover and routing in low earth orbit constellations. In *WCSF'99*, 1999.

[101] A.S. Gandhi and C.G. Cassandras. Optimal control of polling models for transportation applications. *Journal of Mathematical and Computer Modeling*, 23(11-12):1–23, 1996.

[102] M. Gavious and Z. Rosberg. A restricted complete sharing policy for a stochastic knapsack problem in B-ISDN. *IEEE Transactions on Communications*, 42, No. 7:2375–2379, 1994.

[103] L. Georgiadis, W. Szpankowski, and L. Tassiulas. A scheduling policy with maximal stability region for ring networks with spatial reuse. *Queueing Systems*, 19:131–148, October 1995.

[104] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal Royal Statistical Society*, B41:148–164, 1979.

[105] P. Glasserman and D. Yao. Monotone optimal control of permutable GSMPs. *Mathematics of Operations Research*, 19:449–476, 1994.

[106] P. Glasserman and D. Yao. *Monotone Structure in Discrete-Event Systems*. Wiley, New York, 1994.

[107] J. W. Grizzle, S. I. Marcus, and K. Hsu. A decentralized control strategy for multiaccess broadcast networks. *Large Scale Systems*, 3:75–88, 1982.

[108] M. Grossglauser and D. Tse. A framework for robust measurement-based admission control. *IEEE/ACM Trans. on Networking*, 7:293–309, 1999.

[109] P. Gupta and P. R. Kumar. A system and traffic dependent adaptive routing algorithm for Ad Hoc networks. In *Proceedings of the 36th IEEE Conference on Decision and Control*, pages 2375–2380, San Diego, USA, Dec. 1997.

[110] B. Hajek. Optimal control of two interacting service stations. *IEEE Transactions on Automatic Control*, 29:491–499, 1984.

[111] B. Hajek. Extremal splittings of point processes. *Mathematics of Operations Research*, 10:543–556, 1985.

[112] R. Hariharan V. G., Kulkarni, and S. Stidham. Optimal control of admission and routing to two parallel infinite- server queues. *Proceedings of 29th IEEE Conference on Decision and Control*, 1990.

[113] R. Hariharan, V. G. Kulkarni, and S. Stidham. A survey of research relevant to virtual-circuit routing in telecommunication networks. preprint, Department of Operations Research, University of North Carolina at Chapel Hill, 1990.

[114] R. Hariharan, V.G. Kulkarni, and S. Stidham, Jr. A survey of research relevant to virtual-circuit routing in telecommunication networks. Technical Report UNC/OR/TR90-13, University of N.C. at Chapel Hill, 1990.

[115] J. M. Harrison and L. M. Wein. Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Systems*, 5(4):265–279, 1989.

[116] J. M. Harrison and L. M. Wein. Scheduling networks of queues: heavy traffic analysis of a two-station closed network. *Operations Research*, 38(6):1052–1064, 1990.

[117] R. Hassin and M. Haviv. Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying. *Stochastic Models*, 10:415–435, 1994.

[118] M. Herzberg. An optimal decision process for routing circuit-switched calls originated by users of a private distribution network. In A. Jensen and V. B. Iversen, editors, *Teletraffic and Datatraffic in a period of change, ITC-13*, pages 453–458. Elsevier Science Publisher B. V. (North-Holland), 1991.

[119] M. Herzberg and D. McMillan. State-dependent control of call arrivals in layered cellular mobile networks. *Telecommunication Systems*, 1:365–378, 1993.

[120] K. F. Hinderer. *Foundation of Non-stationary Dynamic Programming with Discrete Time Parameter, Lecture Notes in Operations Research and Mathematical Systems No. 33.* Springer-Verlag, Berlin, Heidelberg, New York, 1970.

[121] M. Hofri and Z. Rosberg. packet delay under the golden ratio weighed tdm policy in a multiple-access channel. *IEEE Trans. Inform. Theory*, 33:341–349, 1987.

[122] M. Hofri and K.W. Ross. On the optimal control of two queues with server setup times and its analysis. *SIAM Journal on Computing*, 16:399–420, 1987.

[123] G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM J. Appl. Math.*, 48(5):1159–1166, 1988.

[124] A. Hordijk and G. M. Koole. On the assignment of customers to parallel queues. *Probability in the Engineering and Informational Sciences*, 6:495–511, 1992.

[125] A. Hordijk and G.M. Koole. The μc-rule is not optimal in the second node of the tandem queue: A counterexample. *Advances in Applied Probability*, 24:234–237, 1992.

[126] A. Hordijk and G. M. Koole. On the optimality of LEPT and μc rules for parallel processors and dependent arrival processes. *Advances in Applied Probability*, 25:979–996, 1993.

[127] A. Hordijk, G.M. Koole, and J.A. Loeve. Analysis of a customer assignment model with no state information. *Probability in the Engineering and Informational Sciences*, 8:419–429, 1994.

[128] A. Hordijk and J.A. Loeve. Undiscounted Markov decision chains with partial information; an algorithm for computing a locally optimal periodic policy. *ZOR - Mathematical Methods of Operations Research*, 40:163–181, 1994.

[129] A. Hordijk and F. Spieksma. Constrained admission control to a queueing system. *Advances in Applied Probability*, 21:409–431, 1989.

[130] M. T. Hsiao and A. A. Lazar. Optimal flow control of multiclass queueing networks with partial information. *IEEE Transactions on Automatic Control*, 35 No. 7:855–860, 1990.

[131] M. T. Hsiao and A. A. Lazar. Optimal decentralized flow control of Markovian queueing networks with multiple controllers. *Performance Evaluation*, 13:181–204, 1991.

[132] K. Hsu and S.I. Marcus. Decentralized control of finite state Markov processes. *IEEE Transactions on Automatic Control*, 27 No. 2:426–431, 1982.

[133] P. J. Hunt and C. N. Laws. Asymptotically optimal loss network control. *Mathematics of Operations Research*, 18(4):880–900, 1993.

[134] A. Itai and Z. Rosberg. A golden ratio control policy for a multiple-access channel. *IEEE Transactions on Automatic Control*, 29:712–718, 1984.

[135] V. Jacobson. Congestion avoidance and control. In *ACM SIGCOMM 88*, pages 273–288, 1988.

[136] T. Jimenez. Optimal admission control for high-speed networks: A dynamic programming approach. In *Proceedings of the 39th IEEE Conference on Decision and Control, Sidney, Australia*, Dec. 2000.

[137] S. G. Johansen and S. Stidham. Control of arrivals to a stochastic input-output system. *Advances in Applied Probability*, 12:972–999, 1980.

[138] P. K. Johri. Optimality of the shortest line discipline with state-dependent service times. *European Journal of Operational Research*, 41:157–161, 1990.

[139] T. Kämpke. On the optimality of static priority policies in stochastic scheduling on parallel machines. *Journal of Applied Probability*, 24:430–448, 1987.

[140] T. Kämpke. Optimal scheduling of jobs with exponential service times on identical parallel processors. *Operations Research*, 37:126–133, 1989.

[141] F. P. Kelly. Effective bandwidth at multi-class queues. *Queueing Systems*, 9:5–16, 1991.

[142] F. P. Kelly and C. N. Laws. Dynamic routing in open queueing networks: Brownian models, cust constraints and resource pooling. *Queueing Systems*, 13:47–86, 1993.

[143] P. B. Key. Some control issues in telecommunications networks. In F. P. Kelly, editor, *Probability, Statistics and Optimization A tribute to Peter Whittle*, pages 383–395. Wiley, 1994.

[144] P.B. Key. Optimal control and trunk reservation in loss networks. *Probability in the Engineering and Informational Sciences*, 4:203–242, 1990.

[145] M. Kitaev and V. Rykov. *Controlled Queueing Systems*. CRC Press, 1995.

[146] G. P. Klimov. Time sharing systems. *Theory of Probability and Applications*, 9:532–551, 1974.

[147] G. M. Koole. Dynamic programming tools for control of telecommunication systems. In *Proceedings of the 35th IEEE CDC*, Dec. 1996.

[148] G. M. Koole. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems*, 20:323–339, 1998.

[149] G. M. Koole. A transformation method for stochastic control problems with partial observations. *Systems and Control Letters*, 35:301–308, 1998.

[150] G. M. Koole. On the static assignment to parallel servers. *IEEE Transactions on Automatic Control*, 44:1588–1592, 1999.

[151] G. M. Koole. Optimal transmission policies for noisy channels. Research Report WS-515, Vrije Universiteit, Amsterdam, 1999.

[152] G.M. Koole. Stochastische dynamische programmering met bijvoorwaarden (translation: Stochastic dynamic programming with additional constraints). Master's thesis, Leiden University, 1990.

[153] G.M. Koole. Optimal server assignment in the case of service times with monotone failure rates. *Systems and Control Letters*, 20:233–238, 1993.

[154] G.M. Koole. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems and Control Letters*, 26:301–303, 1995.

[155] G.M. Koole. On the use of the power series algorithm for general Markov processes, with an application to a Petri net. *INFORMS Journal on Computing*, 9:51–56, 1997.

[156] G.M. Koole. Stochastic scheduling with event-based dynamic programming. *ZOR - Mathematical Methods of Operations Research*, 51, 2000.

[157] G.M. Koole and Z. Liu. Nonconservative service for minimizing cell loss in ATM networks. In *Proceedings of the 33rd Allerton Conference on Communication, Control, and Computing*, pages 736–745, Illinois, USA, 1995.

[158] G. M. Koole and Z. Liu. Stochastic bounds for queueing systems with multiple on-off sources. *Probability in the Engineering and Informational Sciences*, 12:25–48, 1998.

[159] G. M. Koole and P. Nain. On the value function of a priority queue with an application to a controlled polling model. *to appear in QUESTA*, 1999.

[160] G. M. Koole and O. Passchier. Optimal control in light traffic Markov decision processes. *ZOR - Mathematical Methods of Operations Research*, 45:63–79, 1997.

[161] Y.A. Korilis and A. Lazar. On the existence of equilibria in noncooperative optimal flow control. *Journal of the ACM*, 42 No. 3:584–613, 1995.

[162] Y.A. Korilis and A. Lazar. Why is flow control hard: optimality, fairness, partial and delayed information. *preprint*, 1995.

[163] K. R. Krishnan and F. Hubner-Szabo de Bucs. Admission control and state-dependent routing for multirate circuit-switched traffic. In *Proceedings of the 15th ITC*, pages 1043–1055. Elsevier Science B. V., 1997.

[164] K.R. Krishnan and T.J. Ott. State-dependent routing for telephone-traffic: Theory and results. In *Proceedings of the 25th IEEE Conference on Decision and Control*, pages 2124–2128, 1986.

[165] K.R. Krishnan and T.J. Ott. Separable routing: A scheme for state-dependent routing of circuit switched telephone traffic. *Annals of Operations Research*, 35:43–68, 1992.

[166] H. Kroner, G. Hebuterne, P. Boyer, and A. Gravey. Priority management in ATM switching nodes. *IEEE J. Selected Areas in Communications*, pages 418–427, Apr. 1991.

[167] V.G. Kulkarni and Y. Serin. Optimal implementable policies: Discounted cost case. In W. J. Stewart, editor, *Proceeding of the International Meeting on Computations with Markov Chains*, pages 283–306, Raleigh, NC, USA, 1995. Kluwer Academic Publishers.

[168] J. Kuri and A. Kumar. Optimal control of arrivals to queues with delayed queue length information. In *Proceedings of the 31th IEEE Conference on Decision and Control*, 1992.

[169] J. Kuri and A. Kumar. On the optimal control of arrivals to a single queue with arbitrary feedback delay. *Queueing Systems*, 27(1-2):1–16, 1997.

[170] B. Kurtaran. Decentralized stochastic control with delayed sharing information pattern. *IEEE Transactions on Automatic Control*, 24:656–657, 1976.

[171] B. Kurtaran. Corrections and extensions to "Decentralized stochastic control with delayed sharing information pattern". *IEEE Transactions on Automatic Control*, 24:656–657, 1979.

[172] H. J. Kushner. Control of trunk line systems in heavy traffic. *SIAM J. Control Optim.*, 33:765–803, 1995.

[173] H. J. Kushner. Heavy traffic analysis of controlled multiplexing systems. *Queueing Systems*, 28:79–107, 1998.

[174] H.J. Kushner and L.F. Martins. Heavy traffic analysis of a data transmission system with many independent sources. *SIAM J. Appl. Math.*, 53:1095–1122, 1993.

[175] H.J. Kushner and L.F. Martins. Heavy traffic analysis of a controlled multi class queueing network via weak convergence theory. *SIAM J. on Control and Optimization*, 34:1781–1797, 1996.

[176] H.J. Kushner and J. Yang. Numerical methods for controlled routing in large trunk line systems via stochastic control theory. *ORSA J. Computing*, 6:300–316, 1994.

[177] H.J. Kushner, J. Yang, and D. Jarvis. Controlled and optimally controlled multiplexing systems: A numerical exploration. *Queueing Systems*, 20(3-4):255–291, 1995.

[178] W.M. Lan and T. Lennon Olsen. A lower bound for dynamic scheduling of single machine multi-product systems with setups. manuscript, 1999.

[179] A. A. Lazar. Optimal flow control of a class of queuing networks in equilibrium. *IEEE Transactions on Automatic Control*, 28:1001–1007, 1983.

[180] W. G. Lazarev and S. M. Starobinets. The use of dynamic programming for optimization of control in networks of communications of channels. *Engineering Cybernetics (Academy of Sciences, USSR)*, (3), 1997.

[181] H. Levy, M. Sidi, and O.J.Boxma. Dominance relations in polling systems. *Queueing Systems*, 6:155–172, 1990.

[182] W. Lin and P.R. Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic Control*, 29:696–703, 1984.

[183] S.A. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23:687–710, 1975.

[184] Z. Liu, P. Nain, and D. Towsley. On optimal polling policies. *Queueing Systems*, 11:59–83, 1992.

[185] Z. Liu, P. Nain, and D. Towsley. Sample path methods in the control of queues. *Queueing Systems*, 21:293–336, 1995.

[186] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with application to call admission. *Journal of the ACM*, 44:366–394, 1997.

[187] Z. Liu and D. Towsley. Optimality of the round-robin routing policy. *Journal of Applied Probability*, 31:466–475, 1994.

[188] J.A. Loeve. *Markov Decision Chains with Partial Information*. PhD thesis, Leiden University, 1995.

[189] D.-J. Ma and A. M. Makowski. A class of steering policies under a recurrence condition. In *Proceedings of the 27th IEEE Conference on Decision and Control*, pages 1192–1197, Austin, TX, USA, Dec. 1988.

[190] D.-J. Ma and A. M. Makowski. A class of two-dimensional stochastic approximations and steering policies for Markov decision processes. In *Proceedings of the 31st IEEE Conference on Decision and Control*, pages 3344–3349, Tucson, Arizona, USA, Dec. 1992.

[191] U. Madhow, M.L. Honing, and K. Steiglitz. Optimization of wireless resources for personal communications mobility tracking. In *Proceedings of IEEE Infocom '94*, pages 577–584, 1994.

[192] B. Maglaris and M. Schwartz. Optimal fixed frame multiplexing in integrated line- and packet-switched communication networks. *IEEE Transactions on Information Theory*, 28:263–273, 1982.

[193] G. Malkin. Rip version 2 carrying additional information. In *IETF RFC 1388*, 1994.

[194] P. Marbach, O. Mihatsch, and J. N. Tsitsiklis. Call admission control and routing in integrated service networks using neuro-dynamic programming. *IEEE Journal on Selected Areas in Communications*, 18(2):197–208, 2000.

[195] R. Menich and R. Serfozo. Optimality of routing and servicing in dependent parallel processing systems. *Queueing Systems*, 9(4):403–418, 1991.

[196] S. P. Meyn. The policy improvement algorithm: General theory with applications to queueing networks and their fluid models. In *Proceedings of the 35th IEEE Conference on Decision and Control*, Kobe, Japan, Dec. 1996.

[197] S. P. Meyn. The policy improvement algorithm for Markov Decision Processes with general state space. *IEEE Transactions on Automatic Control*, 42:191–196, 1997.

[198] S. P. Meyn. Stability and optimization of queueing networks and their fluid models. In *Proceedings of the Summer Seminar on The Mathematics of Stochastic Manufacturing Systems, VA, June 17–21, 1996, Williamsburg, VA. In Lectures in Applied Mathematics,* **33***, American Mathematical Society*, volume 33, 1997.

[199] S. P. Meyn. Feedback regulation for sequencing and routing in multiclass queueing networks. In *2000 IEEE International Symposium on Information Theory*, Sorrento, Italy, June 25 - June 30 2000.

[200] R. Milito and E. Fernández-Gaucherand. Open-loop routing of n arrivals to m parallel queues. *IEEE Transactions on Automatic Control*, 40:2108–2114, 1995.

[201] R. Milito and E. Fernández-Gaucherand. Routing arrivals to queues in parallel. In *Proc. 34th IEEE Conference on Decision and Control*, pages 1415–1420, New Orleans, LA, 1995.

[202] D. Mitra, M.I. Reiman, and J. Wang. Robust dynamic admission control for unified cell and call qos in statistical multiplexers. *IEEE J. Sel. Areas in Commun.*, 16, No.5:692–707, June 1998.

[203] P. Nain. Interchange arguments for classical scheduling problems in queues. *Systems and Control Letters*, 12:177–184, 1989.

[204] P. Nain. Qualitative properties of the Erlang blocking model with heterogeneous user requirements. *Queueing Systems*, 6:189–206, 1990.

[205] P. Nain and K. W. Ross. Optimal priority assignment with hard constraint. *IEEE Transactions on Automatic Control*, 31:883–888, 1989.

[206] P. Nain, P. Tsoucas, and J. Walrand. Interchange arguments in stochastic scheduling. *Journal of Applied Probability*, 27:815–826, 1989.

[207] P. Nain, P. Tsoucas, and J. Walrand. Interchange arguments in stochastic scheduling. *Journal of Applied Probability*, 27:815–826, 1989.

[208] K. Nakade, M. Ohnishi, T. Ibaraki, and T. Ohno. On the average optimality of circular assignment policy. *Queueing Systems*, 11:241–254, 1992.

[209] P. Naor. On the regulation of queueing size by levying tolls. *Econometrica*, 37:15–24, 1969.

[210] K. S. Narendra, E. A. Wright, and L. G. Mason. Application of learning automata to telephone traffic routing and control. *IEEE Trans. Systems, Man and Cybernetics*, 7(11), 1977.

[211] J. Nino-Mora. On the throughput-wip trade-off in queueing systems, diminishing returns and the threshold property: A linear programming approach. *Submitted to Mathematics of Operations Research*, 1998.

[212] E. Nordstrom. Call admission control for preemptive and partially blocking service integration schemes in ATM networks. submitted, 1998.

[213] E. Nordstrom. *Markov Decision Problems in ATM Traffic Control.* PhD thesis, Department of Computer Systems, Upsala University, available as report DoCS 98/100, 1998.

[214] E. Nordstrom and J. Carlstrom. A reinforcement learning scheme for adaptive link allocation in ATM networks. In J. Alspector, R. Goodman, and T. X. Brown, editors, *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunication 2, IWANNT'95*, pages 88–95, Stockholm, Sweden, 1995. Lawrence Erlbaum.

[215] E. Nordstrom and J. Carlstrom. Near-optimal link allocation of blockable narrow-band and queueable wide-band call traffic in ATM networks. In V. Ramasawami and P. E. Wirth, editors, *Proceedings of the 15th International Teletraffic Congress, ITC'15*, pages 987–996, Washington D. C, USA, 1997. Elsevier Science.

[216] E. Nordstrom and J. Carlstrom. Call admission control and routing for integrated CBR/VBR and ABR services: A Markov decision approach. In *Proceedings of the ATM99 workshop*, Kochi, Japan, May 1999.

[217] J. Ou and L. M. Wein. On the improvement from scheduling a two-station queueing network in heavy traffic. *Operations Research Letters*, 11(4):225–232, 1992.

[218] J. D. Papastavrou, S. Rajagopalan, and A. J. Kleywegt. The dynamic and stochastic knapsack problem with deadlines. *Management Science*, 42, No. 12:1706–1718, 1996.

[219] M. Patriksson. *The Traffic Assignment Problem: Models and Methods.* VSP BV, P.O. Box 346, 3700 AH Zeist, The Netherlands, 1994.

[220] C. E. Perkins and E. M. Royer. Ad-hoc on-demand distance vector routing. In *Proceedings of the 2nd IEEE Workshop on Mobile computer systems and Applications*, pages 90–100, 1999.

[221] R. Ramjee, R. Nagarajan, , and D. Towsley. On optimal call admission control in cellular networks. In *Proceedings of IEEE INFOCOM 96*, pages 43–50, 1996.

[222] M.I. Reiman and L.M. Wein. Dynamic scheduling of a two-class queue with setups. *Operations Research*, 46(4):532–547, 1998.

[223] R. Rezaiifar, A. M. Makowski, and S. P. Kumar. Stochastic control of handoffs in cellular networks. *IEEE Journal of Selected Areas in Communications*, 13(7):1348–13162, September 1995.

[224] U. Rieder. Decentralized Markov decision problems with delayed information. In *Abstracts of the 10th INFORMS Applied Probability Conference*, University of Ulm, Germany, July 1999.

[225] F. E. Ross. An overview of fddi: the fiber distributed data interface. *IEEE J. Select. Areas Commun.*, 7(7):1043–1051, Sept. 1989.

[226] K. Ross and B. Chen. Optimal scheduling of interactive and non-interactive traffic in telecommunications systems. *IEEE Transactions on Automatic Control*, 33:261–267, 1988.

[227] K. Ross and D. H. K. Tsang. The stochastic knapsack problem. *IEEE Transactions on Communications*, 37:740–747, 1989.

[228] K. Ross and D. D. Yao. Monotonicity properties for the stochastic knapsack. *IEEE Transactions on Information Theory*, 36:1173–1179, 1990.

[229] H. Rummukainen and J. Virtamo. Polynomial cost approximations in Markov decision theory based least cost routing. *submitted, available in http://www.tct.hut.fi/tutkimus/com2/*, 2000.

[230] F.C. Schoute. Decentralized control in packet switched satellite communication. *IEEE Transactions on Automatic Control*, 23:362–371, 1978.

[231] L. I. Sennott. *Stochastic dynamic programming and the control of queueing systems*. Wiley, New York, 1999.

[232] R. Serfozo. Monotone optimal policies for Markov decision processes. In R. Wets, editor, *Stochastic Systems, II: Optimization*, volume 6, pages 202–215, Amsterdam, New York, 1976. North-Holland. Mathematical Programming Studies.

[233] R. Serfozo. Optimal control of random walks, birth and death processes, and queues. *Advances in Applied Probability*, 13:61–83, 1981.

[234] J.G. Shanthikumar and D.D. Yao. Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research*, 40:S293–S299, 1992.

[235] S. Sharma and Y. Viniotis. Optimal buffer management policies for shared-buffer ATM switches. *IEEE/ACM transactions on networking*, 7(4):575–587, August 1999.

[236] P.D. Sparaggis, C.G. Cassandras, and D.F. Towsley. On the duality between routing and scheduling systems with finite buffer spaces. *IEEE Transactions on Automatic Control*, 38:1440–1446, 1993.

[237] P.D. Sparaggis. Routing and scheduling in heterogeneous systems: A sample path approach. *IEEE Transactions on Automatic Control*, 40:156–161, 1995.

[238] S. Stidham. Socially and individually optimal control of arrivals to a $GI|M|1$ queue. *Management Science*, 24:1598–1610, 1970.

[239] S. Stidham. Optimal control of arrivals to queues and networks of queues. In *Proceedings of the 21th IEEE Conference on Decision and Control*, 1982.

[240] S. Stidham. Optimal control of admission, routing, and service in queues and networks of queues: a tutorial review. *Proceedings ARO Workshop: Analytic and Computational Issues in Logistics R and D*, pages 330–377, 1984. George Washington University.

[241] S. Stidham. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, 30:705–713, 1985.

[242] S. Stidham. Scheduling, routing, and flow control in stochastic networks. *Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA-10:529–561, 1988. W. Fleming and P.L. Lions, eds.

[243] S. Stidham and N. U. Prabhu. Optimal control of queueing systems. In A.B. Clarke, editor, *Mathematical Methods in Queueing Theory*, volume 98, pages 263–294, Berlin, 1974. Springer-Verlag. Lecture Notes in Economics and Mathematical Systems.

[244] S. Stidham, S. Rajagopal, and V. G. Kulkarni. Optimal flow control of a stochastic fluid-flow system. *IEEE Journal on Selected Areas in Communications*, 13:1219–1228, 1995.

[245] S. Stidham and R. Weber. A survey of Markov decision models for control of networks of queues. *Queueing Systems*, 13:291–314, 1993.

[246] S. Stidham and R.R. Weber. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations Research*, 37:611–625, 1989.

[247] J. Talim, Z. Liu, P. Nain, and E. G. Coffman. Controlling robots in web search engines. *Submitted to Performance Evaluation*, 1999.

[248] J. Talim, Z. Liu, P. Nain, and E. G. Coffman. Optimizing the number of robots for web search engines. *To appear in Telecommunication Systems*, 1999.

[249] L. Tassiulas and A. Ephremides. Jointly optimal routing and scheduling in packet radio networks. *IEEE Transactions on Information Theory*, 38(1):165–168, January 1992.

[250] L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, 39(2):466–478, March 1993.

[251] L. Tassiulas and A. Ephremides. Throughput properties of a queueing network with distributed dynamic routing and flow control. *Advances in Applied Probability*, 28:285–307, March 1996.

[252] L. Tassiulas and S. Papavassiliou. Optimal anticipative scheduling with asynchronous transmission opportunities. *IEEE Transactions on Automatic Control*, 40(12):2052–2062, Dec. 1995.

[253] The ATM Forum Technical Committee. *Traffic Management Specification*, Version 4.0, af-tm-0056, April 1996.

[254] H. Tong and T. X. Brown. Adaptive call admission control under quality of service constraints: A reinforcement learning solution. *IEEE Journal on Selected Areas in Communications*, 18(2):209–221, Feb. 2000.

[255] D. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26:305–321, 1978.

[256] D. Towsley R.-H. Hwang, J.F. Kurose. MDP routing for multirate loss networks. *Computer Networks and ISDN*, 2000.

[257] D. Towsley, P.D. Sparaggis, and C.G. Cassandras. Optimal routing and buffer allocation for a class of finite capacity queueing systems. *IEEE Transactions on Automatic Control*, 37:1446–1451, 1992.

[258] F. Vakil and A. A. Lazar. Flow control protocols for integrated networks with partially observed voice traffic. *IEEE Transactions on Automatic Control*, 32:2–14, 1987.

[259] M. P. van Oyen, D.G. Pandelis, and D. Teneketzis. Optimality of index policies for stochastic scheduling with switching costs. *Journal of Applied Probability*, 29:957–966, 1992.

[260] P. Varaiya and J. Walrand. On delayed sharing patterns. *IEEE Transactions on Automatic Control*, 23:443–445, 1978.

[261] P. Varaiya, J. Walrand, and C. Buyukkoc. Extensions of the multiarmed bandit problem: the discounted case. *IEEE Transactions on Automatic Control*, 30:426–439, 1985.

[262] J. Walrand. A note on 'Optimal control of a queueing system with heterogeneous servers'. *System Control Letters*, 4:131–134, 1984.

[263] J. Walrand. *An Introduction to Queueing Networks.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[264] J. G. Wardrop. Some theoretical aspects of road traffic research communication networks. *Proc. Inst. Civ. Eng.*, Part 2, 1:325–378, 1952.

[265] R. Weber and S. Stidham. Control of service rates in networks of queues. *Advances in Applied Probability*, 24:202–218, 1987.

[266] L. M. Wein. Optimal control of a two-station Brownian network. *Mathematics of Operations Research*, 5(2):215–242, 1990.

[267] L. M. Wein. Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs. *Operations Research*, 38(6):1065–1078, 1990.

[268] G. Weiss. Branching bandit processes. *Probability in the Engineering and Informational Sciences*, 2:269–278, 1988.

[269] G. Weiss. Optimal draining of a fluid re-entrant line. In Frank Kelly and Ruth Williams, editors, *Stochastic Networks*, volume 71 of IMA volumes in Mathematics and its Applications, pages 91–103. Springer-Verlag, New York, 1995.

[270] G. Weiss. Optimal draining of fluid re-entrant lines: Some solved examples. In *Stochastic Networks: Theory and Applications*, volume 4 of Royal Statistical Society Lecture Notes Series, pages 19–34. Oxford University Press, Oxford, 1996.

[271] G. Weiss. Scheduling and control of manufacturing systems – a fluid approach. In *Proceedings of the 37rd Allerton Conference on Communication, Control, and Computing*, Illinois, USA, Sept. 1999.

[272] C. C. White. Monotone control laws for noisy, countable-state Markov chains. *European Journal of Operational Research*, 5:124–132, 1980.

[273] W. Whitt. Deciding which queue to join; some counterexamples. *Operations Research*, 34:55–62, 1986.

[274] P. Whittle. Multi-armed bandits and the Gittins index. *Journal Royal Statistical Society*, B42:143–149, 1980.

[275] P. Whittle. *Optimal control: basics and beyond.* John Wiley and sons, 1996.

[276] W. Winston. Optimality of the shortest-line discipline. *Journal of Applied Probability*, 14:181–189, 1977.

[277] S.H. Xu and H. Chen. A note on the optimal control of two interacting service stations. Working paper, 1991.

[278] D. D. Yao and Z. Schechner. Decentralized control of service rates in a closed Jackson network. *IEEE Transactions on Automatic Control*, 34 No. 2:236–240, 1989.

[279] R. D. Yates. A framework for uplink power control in cellular radio systems. *IEEE Journal on Selected Areas in Communications*, 13(7):1341–1347, September 1995.

[280] U. Yechiali. On optimal balking rules and toll charges in a $GI|M|1$ queueing process. *Operations Research*, 19:349–370, 1971.

[281] U. Yechiali. Optimal dynamic control of polling systems. In J.W. Cohen and C.D. Pack, editors, *Queueing, Performance and Control in ATM*, pages 205–217. North-Holland, 1991.

[282] U. Yechiali. Analysis and control of polling systems. In L. Donatiello & R. Nelson, editor, *Performance Evaluation of Computer and Communication Systems*, pages 630–650. Springer-Verlag, 1993.

Eitan Altman
INRIA B.P. 93
06902 Sophia Antipolis Cedex
France
altman@sophia.inria.fr

16 WATER RESERVOIR APPLICATIONS OF MARKOV DECISION PROCESSES

Bernard F. Lamond
Abdeslem Boukhtouta

Abstract: Decision problems in water resources management are usually stochastic, dynamic and multidimensional. MDP models have been used since the early fifties for the planning and operation of reservoir systems because the natural water inflows can be modeled using Markovian stochastic processes and the transition equations of mass conservation for the reservoir storages are akin to those found in inventory theory. However, the "curse of dimensionality" has been a major obstacle to the numerical solution of MDP models for systems with several reservoirs. Also, the use of optimization models for the operation of multipurpose reservoir systems is not so widespread, due to the need for negotiations between different users, with dam operators often relying on operating rules obtained by simulation models.

In this chapter, we present the basic concepts of reservoir management and we give a brief survey of stochastic inflow models based on statistical hydrology. We also present a stochastic dynamic programming model for the planning and operation of a system of hydroelectric reservoirs, and we discuss some applications and computational issues. We feel many research opportunities exist both in the enhancement of computational methods and in the modeling of reservoir applications.

16.1 INTRODUCTION

Dams and reservoirs have long been used for storing surplus water during rainy seasons to provide irrigation and drinking water during dry periods. They prevent flooding during periods of thaw or unusually high rainfall. They also

serve to regulate flow and depth of water in lakes and rivers for navigational purposes, and to move ships up and down locks as in the Panama canal and the Saint-Lawrence seaway. Throughout the twentieth century, hydroelectric production has become a major economic benefit of dams, reservoirs and water resources.

The sequential nature of the reservoir management decisions, together with the inherent randomness of natural water inflows, explains the frequent modeling of reservoir management problems as Markov decision processes (MDPs), and their optimization by stochastic dynamic programming (SDP). The first discussion of reservoir management in this framework is usually credited to Pierre Massé [29] in 1946. Optimization results for the hydroelectric production of a single reservoir were published a decade later, with the numerical computation of an optimal policy [25] and the analytic structure of optimal policies for hydrothermal systems [12]. These results paralleled similar developments that occurred in inventory theory at the same epoch. There is an extensive literature on models and methods for reservoir optimization. Surveys can be found in [22, 53, 54].

Nonetheless, large reservoir systems have been in operation for decades before the development of optimization models. Reservoir operators have thus relied on rule curves and other agreed upon operating rules, as well as their own judgment and experience in making reservoir release decisions [27]. While optimization models are now often used in practice for planning purposes, their use in real-time multiple-reservoir operation is not so widespread. According to [32],

> "The need for comprehensive negotiations and subsequent agreements on how to operate a reservoir system seems to be a main reason why most reservoir systems are still managed based on fixed predefined rules. [...] Optimization models can help define these predefined rules, rules that satisfy various constraints on system operation while minimizing future spills or maximizing energy production or minimizing expected future undesired deviations from various water release, storage volume and/or energy production targets."

The "optimization models" referred to in the above citation are usually based on linear programming (LP) or nonlinear programming (NLP), with the random variables of future inflows replaced by their most recent forecasts. These (deterministic) models must be solved every period with updated forecasts and their solutions provide an *open loop control*. By contrast, an optimal policy of an MDP gives a *closed loop control*, or *feedback solution*, which is more in the form of traditional operating rules.

On the other hand, for reservoir systems whose main purpose is hydroelectric generation, the use of solutions from optimization models is widespread. In [45, 49], for instance, MDP models are presented for the long term planning of the aggregated system, to obtain optimal policies for monthly release and storage targets. Then a hierarchy of deterministic models [13] are used for medium term (NLP) and short term scheduling (LP). See also [16]. A comparison of optimal MDP solutions with traditional rule curves solutions was made in [45] for the Brazilian system, where optimal MDP solutions were shown to have the same reliability as rule curve solutions, but with significantly increased profits.

Stochastic optimization models of hydroelectric production are usually needed when the planning horizon has a length of one or several years, with a time step of one month or longer. The long term scheduling of hydroelectric production is mainly concerned with the larger (annual or multiannual) reservoirs managed by a utility. Typical problems consider twenty or more such reservoirs and are therefore multidimensional. In general, an optimal decision rule for a given period would thus consist of twenty functions of at least twenty variables, each function giving a release target for one reservoir depending on the stored volumes at every reservoir in the system. Such functions are obviously impossible to tabulate numerically (curse of dimensionality). Hence research in this area has attempted to develop (1) aggregation-disaggregation methods, (2) numerical approximation and optimization methods, and (3) analytical solutions [22].

Research also addresses important modeling issues in statistical hydrology (stochastic processes of natural inflows with adequate representation of serial and spatial correlations) and energy economics (such as evaluating marginal production costs in the context of deregulated markets).

The sequel is organized as follows. A brief review of the basic reservoir management concepts and traditional operating policies, are given in §16.2. A survey of several models describing the stochastic processes of natural inflows is presented in §16.3 and a dynamic programming optimization model for a multi-reservoir hydroelectric system is given in §16.4. Different applications of MDP models are presented in §16.5. A survey of recent research on MDP solution methods is presented in §16.6. Finally, §16.7 contains directions and open problems in water reservoir applications of MDPs, and some concluding remarks.

16.2 RESERVOIR MANAGEMENT CONCEPTS

More than 5000 years ago the Egyptians measured fluctuations of the Nile river and built dams and other hydraulic structures to divert water to agricultural fields. Since then, practical water knowledge has proliferated among dam operators, farmers, and other users. But the concept of a water cycle (hydrological cycle) became firmly established in the scientific literature only in the seventeenth century [38]. The *hydrological cycle* is the continuous circulation of water from the sea to the atmosphere, then to the land and back again to the sea. The water exchanges involved at the various stages of the cycle are evaporation, water-vapor transport, condensation, precipitation, and runoff. *Runoff* from land surface is the residual water of the hydrological cycle, which has not been evaporated by plants and has not infiltrated the ground surface, so it is available for use. The collection of land whose surface waters drain into a river valley forms the *hydrographic basin* of that river.

A *dam* is a barrier built across a watercourse for impounding water. By erecting dams, humans can obstruct and control the flow of water in a basin. A *reservoir* is a (possibly artificial) lake, usually the result of a dam, where water is collected and stored in quantity for use. Reservoirs must occupy the best available sites in the hydrographic basin because their development requires unique geological, hydrological, topographical and geographical characteristics.

Controlled inflows into a reservoir include all releases from adjacent upstream reservoirs on the same river or its tributaries. Uncontrolled or *natural* inflows include all other inflows from surface runoffs, streams, and undammed rivers. Water may flow out of a reservoir through various outlets such as derivations (to draw water for irrigation or other consumption), spillways (for flood protection), and penstocks (to produce electricity). Also, there may be water losses due to evaporation and seepage into the ground.

One of the most important uses of reservoirs is to produce electricity. In this case a hydro plant is provided near the reservoir. The quantity of energy produced by a hydro plant depends both on the flow through the turbines and the water head. The *water head* is the difference between forebay elevation and tailwater elevation, which are the reservoir levels in front of the intake and at the exit of the draft tube respectively.

Reservoir systems are also used for a variety of other purposes such as flow control, depth regulation, flood control, water storage for irrigation or supply of drinking water, recreation, navigation, and fish and wildlife enhancement. The management of a multi-purpose reservoir system is a complicated process that must comply with public laws and regulations (public safety, environmental protection, and so forth). It usually attempts to find an effective compromise between the conflicting needs of different uses. For example, flood control requires depletion of reservoirs in advance of floods, and the maximum volume of unused storage has to be maintained until all danger of flooding is past. But energy needs require a full reservoir to allow greater turbine efficiencies for power generation. Moreover, recreational uses require a full pool during the vacation season, which coincides with the need to lower the pool to supply irrigation. Reservoir operators thus rely on *reservoir operating policies* to make release decisions that satisfy all conditions and allocate water equitably between the different uses.

Rule curve and storage allocation zones were the first operating policies used to manage multi-purpose reservoirs. Operating policies associated to *rule curves* define the ideal storage pool level and discharges at different times of the year for each reservoir. The rule curve is based on historical operating practice. In operating policies based on multiple *zones*, the total storage volume of the reservoir is divided into several zones as in Figure 16.1, based on the placement of outlet structures and operational assignments. The *inactive zone* or *dead storage zone* represents the lower part of the reservoir that is not normally used. The *buffer zone* is above the inactive zone. Only essential needs are satisfied when the storage volumes are within this zone, usually as a result of a dry period. The *conservation* or *active zone* represents the volume of water that can be used to satisfy various beneficial uses including recreational and environmental needs. The *flood control zone* is above the conservation zone and it is reserved for flood detention especially during periods of abnormally high runoff. The *spill zone* is the upper portion of the pool, in which the downstream flows are at or near their maximum. See Loucks and Sigvaldason [27] for details about traditional operating policies.

The management and planning of multi-reservoir systems are often supported by a hierarchy of mathematical models. Stochastic optimization models

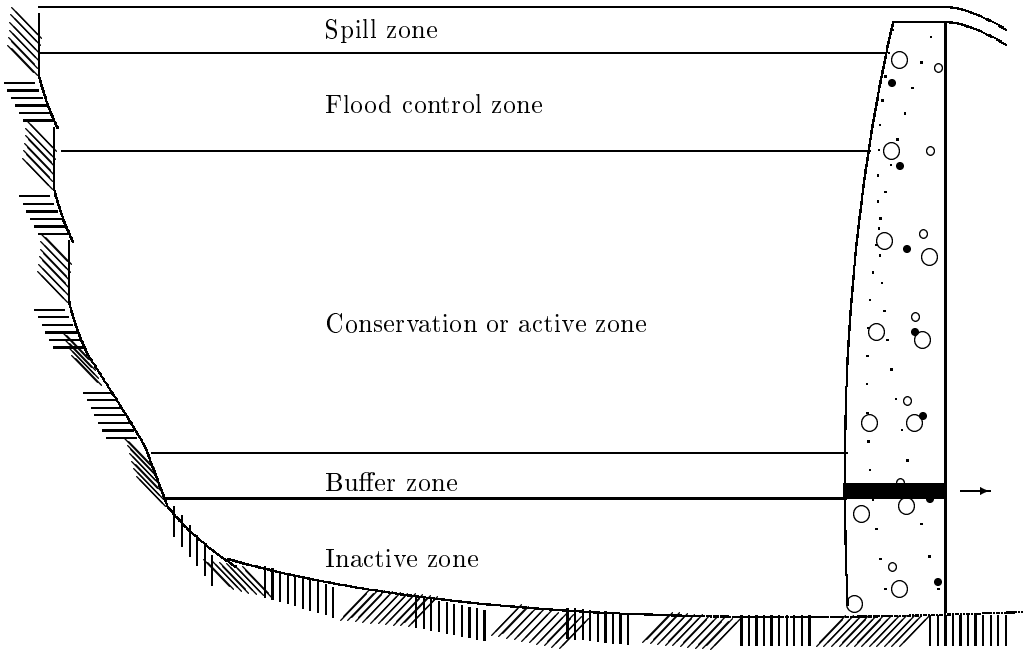

Figure 16.1 Storage allocation zones within a reservoir

usually lie in the top layer of this hierarchy, in which long term planning is performed. The planning horizon in these models exceeds one year, and reaches typically five to twenty years or more. Although reservoir operation happens continuously, the long term planning exercise normally separates the planning horizon into a number of time intervals, or periods, with a fixed time step of one month to one year. The purpose of the planning exercise is to assign storage and production targets in every period. These targets, in turn, are passed on to the next layer (medium term scheduling) in which a deterministic optimization model is applied to a shorter horizon with a smaller time step [13, 16]. Discrete time MDP models are well suited for the long term planning of reservoir systems, especially for hydro-power production, because of the Markovian nature of the stochastic processes governing the natural water inflows.

16.3 STOCHASTIC PROCESSES OF NATURAL INFLOWS

The natural phenomena governing rainfall, runoff, river flows, and flood and drought characteristics are complex and largely unpredictable. The design and operation of dams and reservoirs must therefore take into account the high level of randomness present in these physical processes. The statistical properties of hydrological phenomena are usually obtained by analyzing the time series based on historical records of river flows. In particular, the largest annual flood intensities in successive years have been found to be independent random

variables, while the total annual flows tend to be autocorrelated and are often modeled as autoregressive or moving average processes [28].

Let the random variable D_{it} be the volume of natural inflows received in the i^{th} reservoir during period t. We denote by D_t the column vector of natural inflows in period t for all sites. The sequence $\{D_{it},\ t \in \mathbb{Z}\}$ is the stochastic process of natural inflows at reservoir i. Under this book's convention, the time index $t = 0$ corresponds to the first period of the planning horizon, and past periods have a negative time index. Similarly, the sequence $\{D_t,\ t \in \mathbb{Z}\}$ is the (vector) stochastic process of natural inflows at all sites in the system.

Statistical hydrology

Statistical inflow data are usually available from historical records for a finite number of past periods, and most modelers also make a stationarity assumption about the stochastic process of natural inflows. Various properties of the stochastic process can be inferred from the historical time series using appropriate statistical methods, such as the Box-Jenkins methods [4, 14, 31] for time series analysis. For example, McCleod et al. [31] analyzed the time series of average annual river flows from 1860 to 1957 for the Saint-Lawrence at Ogdensburg, New York. They obtained the following autoregressive model of order 3 (denoted AR(3)):

$$Z_t = 0.6219 Z_{t-1} + 0.1771 Z_{t-3} + E_t,$$

where $Z_t = D_t - \overline{D}$ are the centered inflows ($\overline{D}$ is the historic mean) and E_t are i.i.d.[1] Gaussian errors (innovations) with mean zero and coefficient of variation equal to six percent (relative to $\overline{D}$). Reservoir inflows can also be modeled as moving average (MA) processes or, more generally, as autoregressive moving average (ARMA) processes. Multivariate models have also been proposed for the statistical analysis of natural inflows into multiple reservoir systems. For example, Salas et al. [36] examined the bivariate time series of annual river flows from 1932 to 1963 for the Skykomish and Green rivers located in Washington (USA). They obtained the following autoregressive process:

$$\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} = \begin{pmatrix} 0.407 & 0 \\ 0 & 0.345 \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \begin{pmatrix} E_{1t} \\ E_{2t} \end{pmatrix},$$

where X_{it}, the standardized flow into site i in period t, is given by

$$X_{it} = \frac{D_{it} - \mu_i}{\sigma_i},$$

with μ_i the historical mean of the flows D_{it} at site i, and σ_i their standard deviation. The errors E_{1t} and E_{2t} are independent of previous periods and follow a multivariate normal distribution with mean zero and variance-covariance matrix

$$\Sigma = \begin{pmatrix} 0.837 & 0.823 \\ 0.823 & 0.881 \end{pmatrix}.$$

[1] independent, identically distributed

Similarly, the annual flows of the Wolfe and Fox rivers, located in Wisconsin (USA), were modeled in [5] using records from 1899 to 1965. After a logarithmic transformation of data, the following moving average process was obtained:

$$\begin{pmatrix} Z_{1t} \\ Z_{2t} \end{pmatrix} = E_t - \begin{pmatrix} -0.626 & 0 \\ 0 & -0.543 \end{pmatrix} E_{t-1},$$

where $Z_{it} = \log(D_{it}) - \mu_i$ with μ_i the mean of $\log(D_{it})$, and the vector E_t of innovations is i.i.d., having a multivariate normal distribution with mean zero and variance-covariance matrix

$$\Sigma = \begin{pmatrix} 0.0552 & 0.0502 \\ 0.0502 & 0.075 \end{pmatrix}.$$

The general multivariate ARMA(p,q) model can be written as

$$Z_t = \sum_{i=1}^{p} \Phi_i Z_{t-i} - \sum_{j=1}^{q} \Theta_j E_{t-j} + E_t,$$

where Z_t are m-vectors of standardized (possibly transformed) inflows, E_t are m-vectors of i.i.d. multinormal innovations, Φ_i are $m \times m$ matrices of autoregressive coefficients, and Θ_j are $m \times m$ matrices of moving-average coefficients [36]. When all the matrices Φ_i and Θ_j are lower triangular or upper triangular, the model is said to be a *transfer-function noise* model. When these matrices are diagonal, as in the numerical examples above, the model is said to be *contemporaneous* because the correlations between flows at different sites in the *same* period are accounted for in the variance-covariance matrix, while the serial correlations (with *previous* periods) between flows at different sites are neglected. Contemporaneous ARMA models are much easier to estimate than the general multivariate ARMA models, and they were found to reproduce well the main statistical characteristics of the time series analyzed. Moreover, their selection is often justified a priori by physical considerations [5, 36, 40].

Stedinger et al. [40] examined different ways of fitting the following ARMA(1,1) model to bivariate hydrologic time series by building upon simple univariate procedures:

$$Z_t = \Phi Z_{t-1} - \Theta E_{t-1} + E_t,$$

where $Z_t = D_t - \mu$ is the 2×1 vector of centered flows in period t, Φ and Θ are the 2×2 coefficient matrices and E_t is the 2×1 vector of time-independent normally distributed random fluctuations. Three models were compared: a contemporaneous ARMA(1,1) model where the matrices Φ and Θ are diagonal, a univariate ARMA(1,1) model of aggregate flows on both rivers with a simple disaggregation procedure described in [42], and a nondiagonal ARMA(1,1) model. The study concluded that the first two models performed as well as the multivariate nondiagonal ARMA(1,1) model.

Aggregation/disaggregation models for multivariate water resources time series were proposed in [42, 50]. These models reproduce (relevant) appropriate statistics at several time intervals. Aggregation/disaggregation can be used to

obtain flows at different sites and also to model seasonality. Seasonal inflow models can also be modeled as periodic ARMA processes but their estimation is sometimes difficult especially in the multivariate case [36].

Sophisticated ARMA or aggregation/disaggregation models have been used mostly for forecasting purposes or for synthetic streamflow generation. Inflow forecasts are then used in deterministic optimization models to obtain open loop solutions. On the other hand, synthetic inflow sequences are generated at random for use in Monte Carlo simulation studies to evaluate the design of a reservoir system, or to compare various operating policies.

A review of stochastic models in hydrology is given in [55]. However, the stochastic structure of natural inflow phenomena is not fully understood and is the object of intense research. For instance, recent studies indicate that regional average streamflow statistics contain more information about the variability and persistence of streamflow at a particular site than does the individual streamflow record at this site only. See [51].

Discretized inflow models

Most stochastic optimization models for reservoir operations use discrete random variables to model the natural inflows. This assumption is fundamental when a discrete dynamic programming model is used in which the reservoir levels and discharges are discretized. However, in more realistic models with continuous state and action variables, computation of the expected value of a state-action pair often requires the discretization of the natural inflow distribution, which corresponds to a quadrature rule for numerical integration [9, 19]. A numerical study of discretization error was also presented in [21]. Exceptions to this rule are the linear quadratic Gaussian controller [3] and the myopic-affine dynamic model [24], in which inflow discretization is avoided. Here, we review briefly some discrete inflow models used in reservoir applications of MDPs.

In the long term reservoir scheduling model of [48], the year is broken down in two periods of six months, representing the winter and summer seasons, respectively. The random variable D_t expresses the inflows in terms of potential energy added into the aggregate reservoir system during period t. The energy inflows in period t are assumed correlated with the previous summer's inflows through a seasonal autoregressive model with lag 1 dependence if t is the winter season, and lag 2 dependence if t is the summer season:

$$D_t = \begin{cases} \alpha_{1,w} + \alpha_{2,w} D_{t-1} + \alpha_{3,w} \xi_t, & \text{in winter,} \\ \alpha_{1,s} + \alpha_{2,s} D_{t-2} + \alpha_{3,s} \xi_t, & \text{in summer,} \end{cases}$$

where the coefficients $\alpha_{i,w}$ (for winter season) and $\alpha_{i,s}$ (for summer season), $i = 1, 2, 3$, are constants and the ξ_t are i.i.d. standard normal. Moreover, the random variable D_t is discretized into eleven levels.

In the aggregate reservoir model of [45], the inflows are also expressed in terms of energy. The seasonal autoregressive lag-one model, used to represent the stochastic inflow process, has the form

$$\frac{(D_t - \mu_t)}{\sigma_t} = \rho_t \frac{(D_{t-1} - \mu_{t-1})}{\sigma_{t-1}} + (1 - \rho_t^2)^{0.5} U_t$$

where D_t is the energy inflow during period t (a month), with mean μ_t and standard deviation σ_t, ρ_t is the correlation coefficient between inflows in stage t and stage $t-1$, and U_t is a 3-parameter (mean, standard deviation, and shift) lognormal random variable. The energy inflows are discretized into ten levels.

A different approach, called *sampling* stochastic dynamic programming, was presented in [18]. With this approach, the serial and spatial structures of the streamflow process are captured by using a large number of randomly generated 12 month streamflow sequences. A conditional distribution, developed using a historical time series of streamflow forecasts, is then assigned to various streamflow scenarios. The approach was applied to the North Fork Feather River hydroelectric system composed of nine reservoirs and located in California. Similarly, the DP models developed in [41] derived a reservoir release policy by using the best forecast of the current period's inflow as hydrologic state variable instead of the previous period's inflow. The potential advantage of the proposed approach was illustrated using the Nile river basin as a case study.

In [44], the performance of operating policies derived using stochastic DP models with different sets of hydrologic state variables were compared. Various choices for hydrologic state variables were considered: current period flow, previous period flow, and current period or seasonal flow forecasts. The stochastic process representing the hydrologic variables was described by a month-to-month Markov chain and was discretized to allow the calculation of transition probability matrices for the hydrologic state variables. Based on numerical results, the authors found that for a benefit function stressing energy maximization, all policies did nearly as well. However, for benefit functions involving large water and firm power targets and severe penalties for shortages, the policies that employed more complete hydrologic information performed significantly better.

16.4 DYNAMIC PROGRAMMING MODEL

Typically, the objective of stochastic optimization models for long term reservoir planning is to maximize the expected total discounted reward over a finite horizon. In MDP notation, these models seek a policy π that maximizes the function $v(x, \pi, \dots)$ of equation (0.2), Chapter 0, given an initial state x. The rewards are the revenues from sales of electricity minus the cost of fuel (used for thermal production) and other penalties (such as failure to supply contracted demand). Dynamic programming is an attractive tool for modeling such problems because the stochastic nature of natural inflows and the nonlinear functions associated with energy generation can be modeled explicitly.

We present a generic MDP model for the long term optimization of a multireservoir system composed of m reservoirs over a planning horizon of T periods $t = 0, \dots, T-1$[2]. The purpose of the generic model is to give an overview of the various components of reservoir management problems and to illustrate how they can be represented mathematically. The reader must be warned,

[2]Unlike (0.2), Chapter 0, we use T periods indexed by t instead of N and n.

however, that numerical solution of this generic multi-reservoir model seems well beyond the reach of present computational capabilities, unless stringent simplifications are made. Detailed definitions of state variables, decision variables, constraints, and objective function for the general multi-reservoir system, whose main purpose is hydro-power production, are provided below.

State variables

The state vector for period t includes the volume of water in storage in each reservoir and also some information about recent hydrological activity. Including hydrologic variables in the state vector allows consideration of the serial correlation of natural inflows. Additional hydrologic state variables can be added to model spatial correlation for multi-reservoir systems. Incorporating more hydrologic information in the model improves reservoir operation, but it increases the dimension of the model.

In our generic model, we make the fairly broad assumption that the volume D_{it} of water inflow at site i during period t is related to the inflows at all sites in the K previous periods by a multivariate autoregressive model AR(K). Let S_{it} denote the volume of water in storage in reservoir i at the beginning of period t. Then the state X_t of the system at period t is represented by $K+1$ column m-vectors (m sites) and may be written as $X_t = (D'_{t-K}, \ldots, D'_{t-1}, S'_t)'$, where prime denotes the vector transpose. The special case with $K = 0$ occurs when there is no serial correlation, in which case the state vector is simply $X_t = S_t$. The energy inflow models of [45, 48] are also special cases with $K = 1$ and $K = 2$, respectively.

Decision variables

The decision to be taken in period t is the quantity of water Z_{it} to release through the turbines and the quantity Y_{it} to evacuate through the spillways, for every site i. The model considers both the release Z_{it} and the spill Y_{it} as if they flowed at a constant rate over the period. Similarly, it is convenient to assume the natural inflows D_{it} are collected at a constant flow rate over the duration of period t. Therefore, it is reasonable to assume, as in [48], that the decision depends not only on the state X_t, but also on the inflow D_t. It might then be argued that D_t should be included in the state vector. Instead, we prefer to view D_t as a random perturbation occurring during period t, and to restrict its role in the state information to that of a predictor of future hydrologic activity when serial inflow correlations are present.

The action in period t, denoted A_t, is represented by the pair of m-vectors $(Z'_t, Y'_t)'$. Usually, in reservoir applications, the state X_t and the action A_t are assumed to be random vectors resulting from random variables with continuous joint distributions, except possibly for a small number of mass points (e.g. droughts, storage capacities, etc.). However, to simplify the analysis and for the purpose of numerical computations, many studies assume that X_t and A_t have finite, discrete distributions instead. We denote the realizations of X_t and A_t respectively by x_t and a_t. A decision rule for period t is a function

ϕ_t specifying the action $a_t = \phi_t(x_t, d_t)$ given the state $X_t = x_t$ and inflow $D_t = d_t$. A policy is a sequence of decision rules $\phi_0, \ldots, \phi_{T-1}$.

Transition equations

The first transition equation describes the dynamic behavior of the reservoir system. It is the usual water conservation equation

$$S_{t+1} = S_t - BZ_t - CY_t + D_t, \tag{16.1}$$

where B and C are the $m \times m$ connectivity matrices (or network incidence matrices), used to allocate releases and spills from upstream reservoirs. For systems in which the spilled water is routed on the same river as the turbine releases, we have $C = B$. On other systems, the spilled water is expelled from the system, so that $C = I$. Evaporation could also be taken into account by adding an extra term in the right hand side of equation (16.1), noting that the evaporation rate of a reservoir is proportional to its surface area [11], with the area, in turn, being a function of the storage S_t.

The second transition equation represents the evolution of the stochastic process of the natural inflows. Our assumption of a multivariate AR(K) model leads to the transition equation

$$D_t = \mu_t + \sum_{\ell=1}^{K} \Phi_{\ell t}(D_{t-\ell} - \mu_{t-\ell}) + E_t, \tag{16.2}$$

where $\Phi_{\ell t}$ is an $m \times m$ matrix of "lag ℓ" autoregressive coefficients and E_t is a random vector of innovations, assumed independent of all prior states and actions and with a known, arbitrary joint distribution. Verification of the Markovian property is immediate by inspection of (16.1) and (16.2).

Economic structure

The reward in period t is the sum of net benefits from sales of electricity and possibly some benefit function associated with the amount of water in storage. Strictly speaking, both terms are complicated functions of the state and action variables S_t, Z_t and Y_t as well as the natural inflows D_t. In practice, however, some simplifications are usually assumed. In our generic model, we suppose the reward in period t is given by

$$R_t = r_t(X_t, D_t, A_t) = g_t\left(E_t(S_t, D_t, Z_t)\right) + h_t(S_t), \tag{16.3}$$

where the function $g_t\left(E_t(S_t, D_t, Z_t)\right)$ is the net benefit of the hydropower system in period t and the function $h_t(S_t)$ represents the sum of revenues from other uses of the water in the reservoirs. For many utilities, the function g_t is a concave piecewise linear function of the energy produced $E_t(S_t, D_t, Z_t)$.

The total amount of energy produced by the whole system in period t is

$$E_t(S_t, D_t, Z_t) = \sum_{i=1}^{m} E_{it}(S_{it}, D_{it}, Z_{it}),$$

where the quantity of electricity E_{it} produced at the hydro-plant i in period t is a nonlinear function of the volume of water in storage S_{it} and the volume released through the turbines Z_{it}. For example, [39] neglects the inflows D_{it} and expresses the power generation function as

$$E_{it}(S_{it}, Z_{it}) = K[H_i^{up}(S_{it}) - H_i^{dw}(Z_{it})]Z_{it},$$

where K is a constant, H^{up} gives the height of water at the reservoir's surface (as a function of storage), and H^{dw} gives the height of water in the river at the turbines outlet (as a function of flow). The difference $H_i^{up} - H_i^{dw}$ is the water head factor, which influences the efficiency of electric generation turbines because the power generated by a turbine is proportional to the flow multiplied by the water pressure. We note that the spill variables Y_t do not appear in these expressions. Nonetheless, they should be treated as decision variables because of their presence in the flow equation (16.1).

Inequality constraints

The inequality constraints (16.4–16.6) below account for physical limitations on the state and decision variables. The set $\mathbb{A}_t(x_t, d_t)$ of allowable actions in period t comprises all pairs of vectors z_t and y_t satisfying these inequalities as well as the balance equation (16.1). These limits may include legal restrictions (such as navigational safety, flood control or scheduled maintenance) as well as reservoir capacity constraints. They are represented by constraints on the stored volumes at epoch $t+1$:

$$\underline{s}_{i,t+1} \leq s_{i,t+1} \leq \overline{s}_{i,t+1}. \tag{16.4}$$

There may also be lower an upper limits on the allowable volumes that can flow through the turbines during the period. These limits incorporate nonnegativity, legal requirements and turbine capacities:

$$\underline{z}_{it} \leq z_{it} \leq \overline{z}_{it}. \tag{16.5}$$

Finally, there are nonnegativity and (possibly) upper bound constraints on the spill variables,

$$0 \leq y_{it} \leq \overline{y}_{it}, \tag{16.6}$$

where $\overline{y}_{it}$ is the spillway capacity at site i.

The set $\mathbb{A}_t(x_t, d_t)$ of feasible actions may be empty for certain state-inflow combinations, in which case there is no feasible solution satisfying the above constraints. For example, after a long drought sequence, it may be impossible to satisfy some of the lower storage limits. One way to account for such situations in the model is to add deviation variables in some of the constraints and to penalize these deviations in the objective function. With this device, the model can be used to keep the probability of such events below a tolerable level. See, e.g. [45].

Optimization

The expected future consequence of choosing an action a_t is given by the "cost to go" function that can be written as

$$V_t(x_t) = \mathbb{E}[V_t(X_t; D_t) \mid X_t = x_t], \tag{16.7}$$

where the expectation $\mathbb{E}$ is taken according to equation (16.2). The dynamic programming optimal value function $V_t(x_t; d_t)$ represents the expected operating reward from stage t to the end of the planning horizon, given the inflows for period t. This function results from the recursive DP equation that can be written, for period $t = T-1, \ldots, 0$, as

$$V_t(x_t; d_t) = \max_{a_t \in \mathbb{A}_t(x_t, d_t)} [r_t(x_t, d_t, a_t) + \beta V_{t+1}(x_{t+1})], \tag{16.8}$$

where $r_t(x_t, d_t, a_t)$ is the immediate reward given by (16.3), and β is the discount factor. Then $V_0(x_0)$ gives the expected total discounted reward of the optimal policy in equation (0.2), Chapter 0.

The functions $V_t(x_t)$ and $V_t(x_t; d_t)$ are obtained by backward induction, starting with $t = T-1$, using (16.7) and (16.8), and the (given) terminal reward function $V_T(x_T) = \mathbb{E}[f(x_T)]$. At each epoch t, we need to solve Bellman's equation (16.8) for all possible states $x_t \in \mathbb{X}$. Compactness of the set $\mathbb{A}_t(x_t, d_t)$ of feasible actions insures the existence of optimal actions. The result is an optimal decision rule ϕ_t.

The multivariate expectation in equation (16.7), and hence the "cost to go" function V_t, cannot be calculated exactly in general. Often, an approximate solution of V_t can be obtained after discretization of D_t and X_t. Then the continuous state space is replaced by a discrete grid and the function V_t is evaluated at the grid points only. See [9, 19, 20, 23]. See also [22] for a survey of other approximation methods that have been used in reservoir management models.

16.5 APPLICATIONS

The model of [45], adopted officially in 1979 to manage the Braziliannational electrical generating system, determines the optimal hydro and thermal generations in the system. In addition, this model is used to calculate the expected incremental costs of producing the thermal generation. These costs represent the increase of the expected future operation cost if hydro generation is increased by one MWh (megawatthour). They are used to make decisions about selling or purchasing energy. The model belongs to a chain of generation expansion planning models used to establish when and where to build the new plants. It has also been used to determine the reliability indices of trial expansion plans for 10 to 30 years.

The PERESE model for the long term scheduling of reservoirs, developed especially for the Hydro-Québec system, is presented in [48]. The model is useful to study different generation expansion scenarios in order to determine if the demand can be satisfied, for each scenario, with the desired reliability and to calculate the expected benefits and costs. The model can be used to

decide about construction of the generation expansion plan and to determine guidelines for middle-term horizon studies. It can also be used to determine if additional energy can be sold on the spot markets.

The optimal operating policy developed in [49] gives not only the hydroelectric energy to produce in a month but also the expected marginal cost of the hydroelectric energy produced. This last information is very important in the current context of deregulated markets. To deal in this deregulated market, certain hydroelectric producers purchase energy in spot markets when the price is low. They store it in their reservoirs (by letting them fill with water) and sell it in the spot markets later when the prices are higher. To realize a profit it is important for the producers to determine the exact marginal production cost. The model can also be used to determine the volume of energy that can be sold in the spot markets without endangering the reliability of the system, and to study different scenarios of electrical energy demand and system expansion.

A model of the Shasta-Trinity subsystem, which is a part of the Central Valley Project located in Northern California and operated by the United States Bureau of Reclamation, is formulated in [44]. The benefit function of the model seeks to maximize energy production and to meet reliably water and firm power targets. In addition, the Shasta-Trinity subsystem must produce a part of the energy that the Central Valley Project contracted to provide. Other applications on this system are reported in [43] using other approaches with the objective of meeting water and energy targets.

The control method presented in [11] can be used for real-time operation of a reservoir system as well as for developing policy-making guidelines. It was implemented for the High Aswan dam on Lake Nasser in Egypt. The application seeks to determine the optimal release sequence over a 36 months horizon. In the case when storage exceeds the reservoir capacity, the spilled water is diverted to a depression area where it evaporates. The objective is to maximize the expected energy generation while satisfying downstream water supply requirements, and taking into account the monthly evaporation rates. In spite of the fact that the High Aswan dam suffers heavy evaporation losses, its storage capacity is adequate for current water supply purposes. The tradeoff sought in the Aswan dam application is to maintain the lowest reservoir elevation necessary for meeting the water supply requirements at the expense of energy losses due to the lower hydraulic head. This policy tends to minimize water losses due to evaporation, since evaporation depends on the reservoir's volume and surface. This policy has practical interest since water rather than hydropower availability is the limiting factor in the development of the Egyptian economy. Another benefit of operating the High Aswan dam at low reservoir levels is safety in case of seismic activity. The High Aswan dam system has also been modeled in [41] to define a reservoir release policy and to calculate the expected benefits from future operations. The application used alternative formulations of SDP models and numerical examples on the High Aswan dam system to demonstrate that the approach used can identify more efficient reservoir operations policies.

A real-time optimal control approach was applied in [30], to a system of hydropower reservoirs in the Caroni river basin in Venezuela. The system is

composed of one very large and one moderately sized reservoir. The objective of the application is to track an optimal trajectory that provides a reliable power output to satisfy contractual obligations. Inflows records, collected since 1960, allow consideration of the hydrologic seasonality inputs for the application. The results of the Caroni system application illustrate that there is a trade-off between operation strategies which sacrifice hydrologic complexity in exchange for a more nearly optimal solution and those which sacrifice theoretical optimality in exchange for more accurate hydrologic predictions. For this specific application, the authors believe that improvements in the model that predict reservoir inflows have more impact on the performance of the system than improvements in the optimization algorithm. The method used in this application is expected by its authors to work bestfor systems with large reservoirs.

The control approach of [8] was applied to the Great Lakes levels regulation problem. The Great Lakes, forming a chain of natural reservoirs situated between USA and Canada, consist of Lakes Superior, Michigan, Huron, Erie, and Ontario. The larger lakes, Superior, Michigan, and Huron are modeled in the application as infinite capacity reservoirs with no upper or lower bounds. The management of the Great Lakes system should not cause any disagreement between its five major groups of interest: commercial navigation, riparian or shore property, hydro-power, recreational boating, and environment. Thus, to provide a Great Lakes levels regulation plan, the application is performed to minimize the variations of lakes levels over the planning horizon, knowing the storage and release targets prescribed over a twelve month time horizon. The application can also provide the estimation of reliability and failure probability. Such informations are useful to determine the risks associated with investment decisions involving lake resources and they are useful also in the context of reservoir operation and design.

16.6 COMPUTATIONAL ISSUES

The optimization of operating policies for a multi-reservoir system is a stochastic nonlinear dynamic programming problem. Its main drawback is the need to discretize the state variables. This may lead to a very large number of combinations even with a small number of variables. For small systems the optimization problem can be solved by classical discrete dynamic programming, but for large systems the usefulness of this approach is limited, because computational complexity increases exponentially with the number of reservoirs in the system. This is the well known *curse of dimensionality* (see, e.g. [22] for more details).

Early water reservoir applications of MDPs were made about fifty years ago. They were concentrated on single reservoir and on dynamic programming algorithms for solving these problems. In the sixties and seventies, models with two reservoirs were solved by SDP [53, 54]. Then models with three or four reservoirs were solved in the eighties [9]. Researchers have thus had to resort to other methodologies by combining SDP with other sophisticated approaches to extend the studies to models with more than four reservoirs.

The solution of models based on MDPs usually requires some form of state discretization. Assuming each of the m reservoirs has M discrete storage levels, and assuming further that the hydrologic information can be summarized with e discrete values, then the discrete DP model has eM^m states. Finding an optimal policy thus requires the computation of TeM^m actions. Consequently, the computational complexity of MDP based models increases exponentially with the number of reservoirs in the system, but only linearly with the number of planning periods.

Multistage stochastic programming has also been used in recent years for water reservoir applications. In this approach, the stochastic process of the natural inflows is discretized in the form of a scenario tree. Assuming I possible inflow variations in each period, the number of branches in the scenario tree is $1 + I + I^2 + I^3 + \ldots + I^{T-1}$. Each branch has a corresponding probability and is associated with a set of $3m$ continuous decision variables (storages, turbine releases, and spills). With more than $3mI^{T-1}$ variables, the computational complexity of finding an optimal solution to the stochastic LP increases linearly with the number of reservoirs, but exponentially in the number of planning periods. Recent advances in stochastic linear programming have made this approach practical for multi-reservoir systems planning when the number of periods is relatively small.

Decomposition techniques are often used to solve the multistage stochastic programming problem. In particular, models based on Benders decomposition seem to perform well for the linear and piecewise linear problems. The approach of Pereira and Pinto [33, 34], called SDDP (stochastic Dual Dynamic Programming), is based on Benders decomposition and it uses duality theory to approximate the value function by a piecewise linear function. A comparison of the Benders and dynamic programming approaches is given in [1]. An approach using duality theory and parametric linear programming, but not Benders decomposition, was developed in [35]. This approach was applied to the stochastic problem of scheduling hydro and thermal power generation of a system with two reservoirs. The procedure presented in [15] is based on the same concepts as the algorithm of Pereira and Pinto and it was used for scheduling a large hydroelectric generation problem with a 24 month planning horizon. Another algorithm, based on Benders decomposition, but for nonlinear convex multistage stochastic programming problem was developed in [37]. This algorithm is specialized for a hydropower scheduling application and was applied to a problem with 176 powerhouses on 94 reservoirs with 174 additional controlled water spillways over a 24 month planning horizon. The weakness of these approaches is that the head effects were neglected.

Other multi-reservoir methods with continuous state and action variables have been developed using an approximation of the expected value function. In the parameter iteration method, developed in [10], the value function (or cost to go function) is approximated, at each stage of the dynamic program, by a simple functional form with a small number of parameters. The control is a function of the state characterized by a set of parameters that are improved at each iteration by least squares minimization. In their gradient dynamic programming method, Foufoula-Georgiou and Kitanidis [9] use Hermite inter-

polation to approximate the cost to go function and they reduce dimensionality by using a coarser state discretization. A similar approach, but using tensor-product cubic splines, has been developed in [17], with a parallel computing extension in [7]. The approach was further extended in [6] to higher dimensions using multivariate adaptive regression splines on orthogonal arrays.

Another way to avoid the curse of dimensionality consists of formulating an MDP problem with a small number of discrete state and action variables by using composite reservoirs that can be solved by discrete DP. Models in which the storage capacities of the many reservoirs in the system are aggregated into a single composite reservoir of potential energy are proposed in [45, 48]. The mains drawbacks of such single composite reservoir models is that it is difficult to derive optimal operating policies for individual reservoirs in the system, from the aggregate policy. To address this difficulty, an approach for the optimal operation was proposed in [46] for a system composed of reservoirs in parallel and in [47] in the case where reservoirs are in series. The two approaches consist of using aggregation and decomposition to break up the original model into subproblems of two reservoirs. In the case of a general arborescent multi-reservoir system, an aggregate stochastic dynamic programming model for determining an operating policy was proposed in [2].

Another set of models, belonging to the optimal control approach, and involving no discretization of the system variables, have been used in water reservoirs applications. In these models, based on the linear quadratic Gaussian controller [3], the optimal solution of the cost to go function is deduced analytically, with parameters that can be determined numerically by solution of a "static" optimization problem. An extended linear quadratic Gaussian (ELQG) control technique was applied to a three reservoir system in [11]. In this approach the state variables are replaced by the mean and variance of the storages and an assumption on the probability distribution of the storage state variables is required. The ELQG technique is an iterative refinement of the linear-quadratic control approach, used in [26, 52] in the context of operation of a multi-reservoir system for flood control. A similar approach was used in [30] for the real-time control of a hydropower system with two reservoirs. The main drawback of the linear quadratic Gaussian controller is that it is limited to unbounded reservoir systems. The stochastic control approach presented recently in [8] remedy to this problem by considering the storage bounds explicitly in the expressions for the reservoir systems dynamics by modifying the dynamic equation. As in ELQG, the state variables are the mean and variance of the storages but no assumption on the probability distribution of the storage state variables is required. The limitation of this method lies in the fact that Taylor series approximations are used in the derivation.

16.7 RESEARCH PERSPECTIVES

Rising demands for water for different uses are forcing stiff competition over allocation of scarce water resources among different users. Decision makers face the challenge of having to arbitrate between two conflicting objectives: on the one hand to manage and conserve water supplies, and on the other hand, to satisfy all needs in face of growing demand from population growth and

industries. There is a clear need for more research in the development and refinement of models and methods to help deriving efficient operating policies for reservoir systems, and we feel that the potential of the MDP tools has not yet been fully exploited.

Recent technological developments in digital computing have permitted an evolution in the size of models that can be solved by discrete DP, spline approximations, and stochastic programming. But despite this progress, more research is still needed to solve large stochastic reservoir systems in a precise, detailed manner, under more realistic assumptions. Methods still have to be developed that consider a realistic representation of the natural nonlinearities of hydroelectric generation, by taking head effect into account, and reservoir operations. These methods should also consider a more detailed description of the physical system and uncertainty on inflows, demands, and the price of energy and fuel.

References

[1] Archibald, T.W., Buchanan, C.S., McKinnon, K.I.M and Thomas, L.C. (1999) Nested Benders decomposition and dynamic programming for reservoir optimization. *Journal of the Operational Research Society* **50**, pp. 468–479.

[2] Archibald, T.W., McKinnon, K.I.M and Thomas, L.C. (1997). An aggregate stochastic dynamic programming model of multiple reservoir systems. *Water Resources Research* **33**, pp. 333–340.

[3] Bertsekas, D.P. (1976). *Dynamic Programming and Stochastic Control.* Academic Press, New York.

[4] Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco.

[5] Camacho, F., McLeod, A.I. and Hipel, W.H. (1985). Contemporaneous autoregressive-moving average (CARMA) modeling in water resources. *Water Resources Bulletin* **21**, pp. 709–720.

[6] Chen, V.C.P., Ruppert, D. and Shoemaker, C.A. (1999). Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming. *Operations Research* **47**, pp. 38–53.

[7] Eschenbach, E.A., Shoemaker, C.A. and Caffey, H.M. (1995). Parallel algorithms for stochastic dynamic programming with continuous state and control variables. *ORSA J. on Computing* **7**, pp. 386–401.

[8] Fletcher, S.G. and Ponnambalam, K. (1998). A Constrained state formulation for the stochastic control of multi-reservoir systems. *Water Resources Research* **34**, pp. 257–270.

[9] Foufoula-Georgiou, E. and Kitanidis, P.K., (1988). Gradient dynamic programming for stochastic optimal control of multidimensional water resources systems. *Water Resources Research* **24**, pp. 1345–1359.

[10] Gal, S. (1979). Optimal management of a multi-reservoir water supply system. *Water Resources Research* **15**, pp. 737–749.

[11] Georgakakos, A.P., (1989). Extended linear quadratic Gaussian control for the real time operation of reservoir systems. In *Dynamic Programming for Optimal Water Resources Systems Analysis*, Prentice Hall, Esogbue Editor, pp. 329–360.

[12] Gessford, J. and Karlin, S. (1958). Optimal policy for hydroelectric operations. In *Studies on the Mathematical Theory of Inventory and Production*, Stanford University Press, Stanford, Calif., pp. 179–200.

[13] Hanscom, M. A., Lafond, L., Lasdon, L. and Pronovost, G. (1980). Modelling and resolution of the medium term energy generation planning problem for a large hydro-electric system. *Management Science* **26**, pp. 659–668.

[14] Hipel, K.W., McLeod, A.I. and Lennox, W.C. (1977). Advances in Box-Jenkins modeling 1: model construction. *Water Resources Research* **13**, pp. 567–575.

[15] Jacobs, J.J., Freeman, G., Grygier, J., Morton, D., Schultz, G.L., Staschus, K., and Stedinger, J.R. (1995). Socrates: a system for scheduling hydroelectric generation under uncertainty. *Annals of Operations Research* **59**, pp. 99–133.

[16] Johannesen, A. and Flatabo, N. (1989). Scheduling methods in operation planning of hydro-dominated power production system. *Electrical Power & Energy Systems* **11** pp. 189–199.

[17] Johnson, S.A., Stedinger, J.R., Shoemaker C.A., Li, Y. and Tejada-Guibert, J.A. (1993). Numerical solution of continuous-state dynamic programs using linear and spline interpolation. *Operations research* **41** pp. 484–500.

[18] Kelman, J., Stedinger, J.R., Cooper, L.A., Hsu, E. and Tuan, S. (1990). Sampling stochastic dynamic programming applied to reservoir operation. *Water Resources Research* **26**, pp. 447–454.

[19] Kitanidis, P.K. and Foufoula-Georgiou, E. (1987). Error analysis of conventional discrete and gradient dynamic programming. *Water Resources Research* **23**, pp. 845–856.

[20] Lamond, B. F. (1997) Stochastic reservoir optimization using piecewise polynomial approximations. Working paper, Faculté des sciences de l'administration, Université Laval, Québec, Canada.

[21] Lamond, B. F. and Bachar, B. (1998). Une étude numérique de la discrétisation des apports aléatoires pour un réservoir non saisonnier. *INFOR* **36**, pp. 247–260.

[22] Lamond, B. F. and Boukhtouta, A. (1996). Optimizing long-term hydropower production using Markov decision processes. *Int. Trans. Opl. Res.* **3**, pp. 223–241.

[23] Lamond, B. F. and Lang, P. (1996). Lower bounding aggregation and direct computation for an infinite horizon one-reservoir model. *European Journal of Operational Research* **95**, pp. 404–410.

[24] Lamond, B.F. and Sobel, M.J. (1995). Exact and approximate solutions of affine reservoir models. *Operations Research* **43**, pp. 771–780.

[25] Little, J. D. C. (1955). The use of storage water in a hydroelectric system. *Operations Research* **3**, pp. 187–197.

[26] Loaiciga, H.A. and Mariño, M.A. (1985) An approach to parameter estimation and stochastic control in water resources with an application to reservoir operation. *Water Resources Research* **21**, pp. 1575–1584.

[27] Loucks, D. P. and Sigvaldason, O. T. (1982). Multiple-reservoir operation in North America. In Z. Kaczmarek and J. Kindler (Eds), *The Operation of Multiple Reservoir Systems*, pp. 2–103. IIASA, Laxenburg, Austria.

[28] Maidment, D. R. (1993). *Handbook of Hydrology.* McGraw-Hill, New-York.

[29] Massé, P. B. D. (1946). *Les Réserves et la Régulation de l'Avenir dans la Vie Économique.* Hermann, Paris.

[30] McLaughlin, D. and Velasco, H.L. (1990). Real-time control of a system of large hydropower reservoirs. *Water Resources Research* **26**, pp. 623–635.

[31] McLeod, A.I., Hipel, K.W. and Lennox, W.C. (1977). Advances in Box-Jenkins modeling 2: Applications. *Water Resources Research* **13**, pp. 577–586.

[32] Oliveira, R. and Loucks, D. P. (1997). Operating rules for multi-reservoir systems. *Water Resources Research* **33**, pp. 839–852.

[33] Pereira, M.V.F. and Pinto, L.M.V.G. (1985). Stochastic optimization of a multi-reservoir hydroelectric system: A decomposition approach. *Water Resources Research* **21**, pp. 779–792.

[34] Pereira, M.V.F. and Pinto, L.M.V.G. (1991). Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming* **52**, pp. 359–375.

[35] Read, E. G. (1989). A dual approach to stochastic programming for reservoir release scheduling. In *Dynamic Programming for Optimal Water Resources Systems Analysis*, Prentice Hall, Esogbue Editor, pp. 361–372.

[36] Salas, J.D., Tabios III, G.Q. and Bartolini, P. (1985). Approaches to multivariate modeling of water resources time series. *Water Resources Bulletin* **21**, pp. 683–708.

[37] Salinger, D.H. (1997). *A Splitting Algorithm for Multistage Stochastic Programming with Application to Hydropower Scheduling.* Ph.D. thesis, University of Washington.

[38] Shaw, E. M. (1988). *Hydrology in Practice.* Van Nostrand Reinhold, England.

[39] Soares, S. and Carneiro, A.A.F.M. (1990). Optimal operation of reservoirs for electric generation. *Proceedings of the IEEE Power Engineering Society 1990 Summer Meeting*, Minneapolis, Minnesota.

[40] Stedinger, J.R., Lettenmaier, D.P. and Vogel, R.M. (1985). Multisite ARMA(1,1) and disaggregation models for annual streamflow generation. *Water Resources Research* **21**, pp. 497–509.

[41] Stedinger, J.R., Sule, B.F. and Loucks, D.P. (1984). Stochastic dynamic programming models for reservoir operation optimization. *Water Resources Research* **20**, pp. 1499–1505.

[42] Stedinger, J.R. and Vogel, R. M. (1984). Disaggregation procedures for generating serially correlated flow vectors. *Water Resources Research* **20**, pp. 47–56.

[43] Tejada-Guibert, J.A., Johnson, S.A. and Stedinger, J.R. (1993). Comparison of two approaches for implementing multi-reservoir operating policies derived using stochastic dynamic programming. *Water Resources Research* **29**, pp. 3969–3980.

[44] Tejada-Guibert, J.A., Johnson, S.A. and Stedinger, J.R. (1995). The value of hydrologic information in stochastic dynamic programming models of a multi-reservoir system. *Water Resources Research* **31**, pp. 2571–2579.

[45] Terry, L. A., Pereira, M. V. F., Araripe Neto, T. A., Silva, L. F. C. A. and Sales, P. R. H. (1986). Coordinating the energy generation of the Brazilian national hydrothermal electrical generating system. *Interfaces* **16**, pp. 16–38.

[46] Turgeon, A. (1980). Optimal operation of multi-reservoir power systems with stochastic inflows. *Water Resources Research* **16**, pp. 275–283.

[47] Turgeon, A. (1981) A decomposition method for the long-term scheduling of reservoirs in series. *Water Resources Research* **17**, pp. 1565–1570.

[48] Turgeon, A. (1992). PERESE: The New Hydro-Québec optimization model for the long-term scheduling of reservoirs. In *Hydropower '92: Proceedings of the 2nd International Conference on Hydropower*, Held in Lillehamer, Norway Ed. Broch, E. and Lysne, D.K., Balkema publisher, Rotterdam, pp. 603–612.

[49] Turgeon, A. and Charbonneau, R. (1998). An aggregation - disaggregation approach to long-term reservoir management. *Water Resources Research* **34**, pp. 3585–3594.

[50] Valencia, R.D. and Schaake, J.C. (1973). Disaggregation processes in stochastic hydrology. *Water Resources Research* **9**, pp. 580–585.

[51] Vogel, R.M., Tsai, Y. and Limbrunner, J.F. (1998). The regional persistence and variability of annual steamflow in the United States. *Water Resources Research* **34**, pp. 3445–3459.

[52] Wasimi, S.A. and Kitanidis, P.K. (1983). Real-time forecasting and daily operation of a multi-reservoir system during floods by linear quadratic Gaussian control. *Water Resources Research* **19**, pp. 1511–1522.

[53] Yakowitz, S. (1982). Dynamic programming applications in water resources. *Water Resources Research* **18**, pp. 7673–7696.

[54] Yeh, W.W.G. (1985). Reservoir management and operations models: a state-of-the-art review. *Water Resources Research* **21**, pp. 1797–1818.

[55] Yevjevich, V. (1987). Stochastic models in hydrology. *Stochastic Hydrological Hydraulic* **1**, pp. 17–36.

Bernard F. Lamond
Département Opérations et Systèmes de Décision
Faculté des Sciences de l'Administration
Université Laval
Québec (QC), Canada G1K 7P4
Bernard.Lamond@fsa.ulaval.ca

Abdeslem Boukhtouta
Département Opérations et Systèmes de Décision
Faculté des Sciences de l'Administration
Université Laval
Québec (QC), Canada G1K 7P4