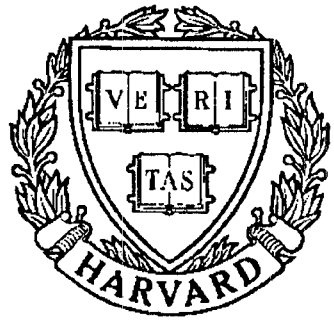
TECHNICAL RESEARCH REPORT

SYSTEMS RESEARCH CENTER



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey

by A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus

TR 91-109 r1

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
31 MAR 1992		**00-00-1992 to 00-00-1992**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Systems Research Center,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**70**	

DISCRETE-TIME CONTROLLED MARKOV PROCESSES WITH AVERAGE COST CRITERION: A SURVEY*

ARISTOTLE ARAPOSTATHIS†, VIVEK S. BORKAR‡,
EMMANUEL FERNÁNDEZ-GAUCHERAND§,
MRINAL K. GHOSH† AND STEVEN I. MARCUS¶

March 31, 1992

ABSTRACT. This work is a survey of the average cost control problem for discrete-time Markov processes. We have attempted to put together a comprehensive account of the considerable research on this problem over the past three decades. Our exposition ranges from finite to Borel state and action spaces and includes a variety of methodologies to find and characterize optimal policies. We have included a brief historical perspective of the research efforts in this area and have compiled a substantial yet not exhaustive bibliography. We have also identified several important questions which are still left open to investigation.

CONTENTS

AMS(MOS) subject classifications: 93E20, 60J70.
Key words: Controlled Markov processes, average cost, stationary policies, dynamic programming, optimal policies, ergodicity.

*This work was supported in part by the Texas Advanced Research Program (Advanced Technology Program) under Grants No. 003658-093 and No. 003658-186, in part by the Air Force Office of Scientific Research under Grants AFOSR-91-0033, F49620-92-J-0045, F49620-92-J-0083, and in part by the National Science Foundation under Grant CDR-8803012.

†Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712.

‡Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India.

§Systems and Industrial Engineering Department, University of Arizona, Tucson, Arizona 85721.

¶Department of Electrical Engineering and Systems Research Center, University of Maryland, College Park, Maryland 20742.

§1. Introduction

The average cost criterion (equivalently, the long run average or ergodic cost) is a popular criterion for optimization of stochastic dynamical systems over an infinite time horizon. It is a reasonable criterion to use when the anticipated time interval for optimization (which in practice is finite) is long compared to other time scales involved and there are no compelling reasons to prefer short term optimization over long term. Naturally, it is not favored in financial applications where money spent now is worth more than money spent later, but there are situations (communication networks being a prime example) where a 'steady state' operation is expected over intervals long compared to the time constants of the system. Then it makes sense to minimize the limiting time-averaged cost, i.e., the 'average cost.'

Mathematically, the criterion stands out as being much more difficult to analyze than others; while other classical criteria lead to reasonably complete solutions, the average cost does not. The finite state and action problem is well understood, but there are numerous counterexamples in which infinite state or action problems do not have a nice solution. In fact, it appears not as a single problem but a collection of problems, some of which do not have a nice solution (cf. [146]). Thus, a variety of approaches have been developed to handle different situations. Not surprisingly, this is one chapter of Markov decision theory which is anything but closed. At the same time, it has come of age, having been studied for over 30 years, with promises of significant advances on the horizon. This, in short, is the raison d'etre for this survey; we have attempted to put together a coherent account of what has been done, with an indication of what future advances may be.

Any such project has obvious limitations. Space constraints dictate a certain amount of selection, and not every relevant work can be covered in significant detail. We have included proofs where we felt they were essential to understanding the results or contained potentially useful novel ideas. In all cases, a serious attempt at objectivity has been made. For other complementary reading on the general subject of Markov decision theory, see [133], [177], [192], [203].

The paper is organized as follows: Section 2 describes the problem formulation in full detail. Section 3 gives a brief history. Sections 4, 5 and 6 treat extensively the finite state, the countable state and the Borel state space cases respectively, under complete observations. Section 7 treats the problem under partial observations. Section 8 describes

some recent results on multiobjective average cost control. Finally, we conclude with some relevant remarks.

§2. Preliminaries and Formulation of the Problem

In this section, the model and basic results concerning controlled Markov processes are given in the most general form needed for our presentation. In some subsequent sections, we specialize our presentation to situations in which measure-theoretic aspects are of no essential concern, as in the case for models with countable state space, allowing for a more transparent exposition. Before presenting the model we summarize our key notation:

- $\mathbb{R}$: set of real numbers.
- $\mathbf{N}$: set of positive integers.
- $\mathbf{N}_0$: set of nonnegative integers.
- $\mathcal{B}(W)$: Borel σ-algebra of a given topological space W.
- $\mathcal{P}(W)$: for a Borel space W (see [15]), the set of all probability measures on $\mathcal{B}(W)$, endowed with the topology of weak convergence (see [130]).

Function spaces on a topological space W:

- $C_b(W) := \{v : W \to \mathbb{R} \mid v \text{ is continuous and bounded}\}$.
- $\mathcal{M}(W) := \{v : W \to \mathbb{R} \mid v \text{ is Borel measurable}\}$.
- $\mathcal{M}_b(W) := \{v : W \to \mathbb{R} \mid v \text{ is Borel measurable and bounded}\}$.
- $\mathcal{L}(W) := \{v : W \to \mathbb{R} \mid v \text{ is lower semicontinuous and bounded below}\}$.
- $\mathcal{L}_b(W) := \mathcal{L}(W) \bigcap \mathcal{M}_b(W)$.

For $v \in \mathcal{M}_b(W)$ we let:

- $\|v\| := \sup_{w \in W} \{|v(w)|\}$.
- $\operatorname{span}(v) := \sup_{w,w' \in W} \{v(w) - v(w')\}$.
- $v^+ := v - \inf_{w \in W} \{v(w)\}, \quad v^- := v - \sup_{w \in W} \{v(w)\}$.

We refer to $\operatorname{span}(v)$ as the *span seminorm* of v.

The following is a list of the acronyms used in this paper (the subsection where each acronym is first introduced is indicated in parenthesis):

- AC average cost (Sect. 2.4)
- ACOE average cost optimality equation (Sect. 3)
- ACOI average cost optimality inequality (Sect. 5.2)
- CMP controlled Markov process (Sect. 2.1)
- CO completely observable (Sect. 3)
- DC ciscounted cost (Sect. 2.4)
- DCOE ciscounted cost optimality equation (Sect. 2.6)
- PO partially observable (Sect. 7.2)
- POCMP partially observable controlled Markov process (Sect. 3)
- TC total cost (Sect. 2.4)

2.1. The model. A discrete time, stationary controlled Markov process (CMP), or Markov decision process, is a stochastic dynamical system specified by the five-tuple (S, A, U, P, c),

where:

(a) $\boldsymbol{S}$ is a Borel space, called the *state space,* the elements of which are called *states.*

(b) $\boldsymbol{A}$ is a Borel space, called the *action* or *control* space.

(c) $U : \boldsymbol{S} \to \mathcal{B}(\boldsymbol{A})$ is a strict, measurable, compact-valued multifunction (see the Appendix). $U(x)$ represents the set of admissible actions (or control inputs) when the system is in state $x \in \boldsymbol{S}$. Accordingly, the set of admissible state/action pairs is $\boldsymbol{K} := \{(x,a) : x \in \boldsymbol{S},\, a \in U(x)\} = \mathrm{Graph}(U)$, and we have that $\boldsymbol{K} \in \mathcal{B}(\boldsymbol{S} \times \boldsymbol{A})$. This set is endowed with the subspace topology corresponding to $\mathcal{B}(\boldsymbol{S} \times \boldsymbol{A})$.

(d) P is a stochastic kernel on $\boldsymbol{S}$ given $\boldsymbol{K}$, called the *transition kernel.* It is assumed to be Borel measurable, i.e., $P(D \mid \cdot) : \boldsymbol{K} \to [0,1]$ is Borel measurable, for each $D \in \mathcal{B}(\boldsymbol{S})$.

(e) $c : \boldsymbol{K} \to \mathbb{R}$ is the (measurable) one-stage cost function.

The evolution of the system is as follows. Let X_t denote the state at time $t \in \mathbb{N}_0$, and A_t the action chosen at that time. If $X_t = x \in \boldsymbol{S}$ and $A_t = a \in U(x)$, then: (i) a cost $c(x,a)$ is incurred, and (ii) the system moves to the next state X_{t+1}, according to a probability distribution $P(\cdot \mid x,a)$. Once the transition into the next state has occurred, a new action is chosen, and the process is repeated.

The total period of time over which the system is to be observed is called the planning (or decision-making or control) horizon and is denoted by T. It can be a finite interval $\{0,\ldots,N-1\}$, with $N \in \mathbb{N}$, or an infinite horizon, e.g. $T = \mathbb{N}_0$.

Example 2.1. Let $\boldsymbol{S}$, $\boldsymbol{A}$, $\boldsymbol{W}$ be Borel spaces, and $F : \boldsymbol{S} \times \boldsymbol{A} \times \boldsymbol{W} \to \boldsymbol{S}$ a Borel function. Consider a nonlinear stochastic system described by the system equation

$$X_{t+1} = F(X_t, A_t, W_t), \qquad t \in T,$$

where the process $\{W_t\}$ is a sequence of independent and identically distributed (i.i.d.) $\boldsymbol{W}$-valued random variables, with common probability distribution $\mathcal{P}_{\boldsymbol{W}}$, often referred to as a stochastic state disturbance, or noise; $\{W_t\}$ is assumed to be independent of X_0. Suppose that a strict, measurable, compact-valued multifunction $U : \boldsymbol{S} \to \mathcal{B}(\boldsymbol{A})$ has been specified, giving the necessary constraints on the control actions, or that $U(x) = \boldsymbol{A}$, for all $x \in \boldsymbol{S}$, if there are no constraints. Then the evolution of the system is equivalently described in terms of the stochastic kernel P on $\boldsymbol{S}$ given $\boldsymbol{K}$ defined as

$$P(D \mid x,a) := \int_{\boldsymbol{W}} I\{F(x,a,w) \in D\}\mathcal{P}_{\boldsymbol{W}}(dw), \qquad (x,a) \in \boldsymbol{K}, \quad D \in \mathcal{B}(\boldsymbol{S}),$$

where $I\{A\}$ denotes the indicator function of the event A. The additional specification of a measurable cost function $c : \boldsymbol{K} \to \mathbb{R}$ would completely define a CMP $(\boldsymbol{S}, \boldsymbol{A}, U, P, c)$.

Example 2.2. Consider a countable set $\boldsymbol{S}$ endowed with the discrete topology. With no loss in generality we can take $\boldsymbol{S} = \mathbb{N}_0$. Let $\boldsymbol{A}$ be a Borel space and $U(x) = \boldsymbol{A}$, for all $x \in \boldsymbol{S}$. In this case, every stochastic kernel on $\mathbb{N}_0$ given $\boldsymbol{K} := \mathbb{N}_0 \times \boldsymbol{A}$ reduces to a collection of discrete probability distributions parameterized by $(i,a) \in \boldsymbol{K}$. These can also be represented by

a collection of stochastic matrices $\{P(a) = [p_{ij}(a)]\}_{a \in \boldsymbol{A}}$, i.e., $P(a)$ is a state transition matrix, and $p_{ij}(a)$ is the probability that the state of the system makes a transition from i to j, under action a. Therefore, additionally specifying a cost function $c : \mathbf{N}_0 \times \boldsymbol{A} \to \mathbf{R}$ completely defines a CMP.

The (admissible) *history spaces* are defined as

$$\boldsymbol{H}_0 := \boldsymbol{S}\,, \qquad \boldsymbol{H}_t := \boldsymbol{H}_{t-1} \times \boldsymbol{K}\,, \quad t \in \mathbf{N}_0\,,$$

and the canonical sample space is defined as

$$\boldsymbol{\Omega} := (\boldsymbol{S} \times \boldsymbol{A})^{\infty}\,.$$

These spaces are endowed with their respective product topologies, and are therefore Borel spaces. A generic element $\omega \in \boldsymbol{\Omega}$ is of the form $\omega = (x_0, a_0, x_1, a_1, \dots)$, $x_i \in \boldsymbol{S}$, $a_i \in \boldsymbol{A}$; all random variables will be defined on the measurable space $\big(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}(\boldsymbol{\Omega})\big)$.

The state, action (or control) and information processes, denoted by $\{X_t\}_{t \in T}$, $\{A_t\}_{t \in T}$ and $\{H_t\}_{t \in T}$, respectively, are defined by the projections

$$X_t(\omega) := x_t\,, \quad A_t(\omega) := a_t\,, \quad H_t(\omega) := (x_0, \dots, a_{t-1}, x_t)\,, \qquad t \in T\,,$$

for each realization $\omega = (x_0, \dots, a_{t-1}, x_t, a_t, \dots) \in \boldsymbol{\Omega}$. Since $\boldsymbol{\mathcal{B}}(\boldsymbol{\Omega}) = \big(\boldsymbol{\mathcal{B}}(\boldsymbol{S}) \times \boldsymbol{\mathcal{B}}(\boldsymbol{A})\big)^{\infty}$, the above are well-defined random processes on $\big(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}(\boldsymbol{\Omega})\big)$. Note that $\boldsymbol{\mathcal{B}}(\boldsymbol{\Omega}) = \bigvee_{t=0}^{\infty} \mathfrak{F}_t$, where $\mathfrak{F}_t = \sigma(H_t)$, the σ-algebra generated by H_t.

2.2. Policies and performance criteria. An *admissible control strategy,* or *policy,* is a sequence $\pi = \{\pi_t\}_{t \in T}$ of Borel measurable stochastic kernels on $\boldsymbol{A}$ given $\boldsymbol{H}_t$, satisfying the constraint

$$\pi_t\big(U(x_t) \,\big|\, h_t\big) = 1\,, \qquad x_t \in \boldsymbol{S}\,, \quad h_t \in \boldsymbol{H}_t\,.$$

The set of all admissible policies will be denoted by Π.

If $\mu \in \boldsymbol{\mathcal{P}}(\boldsymbol{S})$ and $\pi \in \Pi$ are given, there exists a unique probability measure $\mathcal{P}_{\mu}^{\pi}$ on $\big(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}(\boldsymbol{\Omega})\big)$ satisfying the following [15, Prop. 7.28, pp. 140–144], [126, Prop. V.1.1, pp. 162–164], with $D \in \boldsymbol{\mathcal{B}}(\boldsymbol{S})$ and $C \in \boldsymbol{\mathcal{B}}(\boldsymbol{A})$,

$$\mathcal{P}_{\mu}^{\pi}(X_0 \in D) = \mu(D)\,, \tag{2.1}$$

$$\mathcal{P}_{\mu}^{\pi}(A_t \in C \mid H_t) = \pi_t(C \mid H_t)\,, \qquad \mathcal{P}_{\mu}^{\pi}\text{-a.s.} \tag{2.2}$$

$$\mathcal{P}_{\mu}^{\pi}(X_{t+1} \in D \mid H_t, A_t) = P(D \mid X_t, A_t)\,, \qquad \mathcal{P}_{\mu}^{\pi}\text{-a.s.} \tag{2.3}$$

Therefore, if μ is the distribution of the initial state X_0, and policy $\pi \in \Pi$ is used, the underlying probability space of all random variables of interest is $\big(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}(\boldsymbol{\Omega}), \mathcal{P}_{\mu}^{\pi}\big)$. The expectation operator with respect to $\mathcal{P}_{\mu}^{\pi}$ will be denoted by E_{μ}^{π}. Furthermore, if μ is a Dirac measure at $x \in \boldsymbol{S}$, we will simply write $\mathcal{P}_{x}^{\pi}$ and E_{x}^{π}.

Certain classes of admissible policies are of special interest. A policy π is called a *Markov randomized* policy if there exists a sequence of measurable maps $\{f_t\}_{t \in T}$, called *randomized decision rules,* where $f_t : \boldsymbol{S} \to \boldsymbol{\mathcal{P}}(\boldsymbol{A})$, for each $t \in T$, such that

$$\pi_t(\cdot \mid H_t) = f_t(X_t)(\cdot)\,, \qquad \mathcal{P}_{\mu}^{\pi}\text{-a.s.}$$

Conversely, every sequence of measurable maps $f_t : \boldsymbol{S} \to \boldsymbol{\mathcal{P}}(\boldsymbol{A})$, $t \in T$, satisfying $f_t(x)\big(U(x)\big) = 1$, defines a Markov randomized policy in an obvious way; with some abuse in notation the sequence itself will be referred to as the policy. The set of all Markov randomized policies will be denoted by Π_M. A policy $\{f_t\}_{t\in T} \in \Pi_M$ is called a *stationary* randomized policy if there is a randomized decision rule f such that, for all $t \in T$, $f_t = f$. The set of all stationary randomized policies will be denoted by Π_{SR}. A *nonrandomized, deterministic* or *pure* decision rule is a measurable map $f : \boldsymbol{S} \to \boldsymbol{A}$. A policy $\{f_t\}_{t\in T} \in \Pi_M$ is called a *nonrandomized, deterministic* or *pure* Markov policy if each f_t is deterministic. Hence, in this case $A_t = f_t(X_t)$ a.s. The set of deterministic Markov policies will be denoted by Π_{MD}. Stationary deterministic policies are defined in the obvious way. The set of all stationary deterministic policies is denoted by Π_{SD}, and for $\pi \in \Pi_{SD}$, $\pi(x)$ will denote the action chosen at $x \in \boldsymbol{S}$. Clearly $\Pi_{SD} \subseteq \Pi_{MD} \subseteq \Pi_M \subseteq \Pi$, and $\Pi_{SD} \subseteq \Pi_{SR} \subseteq \Pi_M$.

It is easily seen that under a policy $\pi = \{f_t\}_{t\in T} \in \Pi_M$, the state process $\{X_t\}_{t\in T}$ is a Markov process. That is, for $D \in \boldsymbol{\mathcal{B}}(\boldsymbol{S})$,

$$\begin{aligned}\mathcal{P}_\mu^\pi\big(X_{t+1} \in D \mid X_t, \ldots, X_0\big) &= \mathcal{P}_\mu^\pi\big(X_{t+1} \in D \mid X_t\big) \\ &= \int_{\boldsymbol{A}} P\big(D \mid X_t, a\big) f_t(X_t)(da)\,, \qquad \mathcal{P}_\mu^\pi\text{-a.s.},\end{aligned}$$

and under a policy $\pi' \in \Pi_{SR}$, $\{X_t\}_{t\in T}$ is a Markov process with stationary transition probabilities.

Each policy $\pi \in \Pi$ incurs a stream of random costs, e.g. $\big\{c\big(X_t, f_t(X_t)\big)\big\}_{t\in T}$, for $\{f_t\}_{t\in T} \in \Pi_{MD}$. Depending upon the problem requirements several cost evaluation criteria are studied. The following criteria are frequently used.

Total Cost (TC). The total cost incurred by the policy $\pi \in \Pi$ over the entire planning horizon is given by

$$J_T(\mu, \pi) := E_\mu^\pi \left[\sum_{t\in T} c(X_t, A_t)\right].$$

When the horizon is finite, i.e., $T = \{0, \ldots, N-1\}$, $N \in \mathbb{N}_0$, we denote the above more explicitly as $J_N(\mu, \pi)$. Furthermore, given a *terminal cost* function $h \in \mathcal{M}_b(\boldsymbol{S})$, we define

$$J_N(\mu, \pi, h) := E_\mu^\pi \left[\sum_{t=0}^{N-1} c(X_t, A_t) + h(X_N)\right].$$

Discounted Cost (DC). Let $0 < \beta < 1$, the *discount factor*, and $\pi \in \Pi$ be given. The total discounted cost incurred by π over the infinite planning horizon is given by

$$J_\beta(\mu, \pi) := E_\mu^\pi \left[\sum_{t=0}^{\infty} \beta^t c(X_t, A_t)\right].$$

Average Cost (AC). The expected long-run average cost incurred by $\pi \in \Pi$ is given by

$$J(\mu, \pi) := \limsup_{N\to\infty} E_\mu^\pi \left[\frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t)\right] = \limsup_{N\to\infty} \frac{1}{N} J_N(\mu, \pi)\,.$$

Sample Path Average Cost. This is a pathwise version of the AC and, for $X_0 = x$, it is given by

$$J_S(x,\pi) := \limsup_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t)\,,$$

where $\{X_t\}$ and $\{A_t\}$ are the state and control process induced by $\pi \in \Pi$. Here, $J_S(x,\pi)$ is to be regarded as an extended real-valued random variable on the canonical sample space.

For the AC criterion, the limit of the expected average cost may not exist for some or all policies $\pi \in \Pi$, and thus the limit superior is used. This is always well-defined and captures the *worst* possible asymptotic expected average performance under policy $\pi \in \Pi$, i.e., it gives a 'pessimistic' measure of performance. On the other hand, the limit inferior could also be used, which would yield an 'optimistic' measure of performance by capturing the *best* possible asymptotic expected average performance. The planning horizon for the TC criterion can be finite or infinite, whereas for the other criteria above it is always infinite. Under certain conditions, it can be shown that a problem with the DC criterion is equivalent to one with a TC criterion, with a random (finite) horizon, see [39, pp. 31–32]. Also, it can be shown that for each $\pi \in \Pi$, a policy $\pi' \in \Pi_M$ can be found such that $E_\mu^\pi\big[c(X_t,A_t)\big] = E_\mu^{\pi'}\big[c(X_t,A_t)\big]$, for each $t \in \mathbb{N}_0$ and any initial distribution $\mu \in \mathcal{P}(S)$ [41], [50, Sect. 3.8]. Thus, for criteria that are determined by these expected costs, such as the AC, DC, and TC criteria, it suffices to consider policies in Π_M.

For an infinite planning horizon, $J_T(\mu,\pi)$ need not be well-defined or may be infinite for all $\pi \in \Pi$, rendering this criterion useless for comparing the performance under different policies. Therefore, the DC or AC criteria are usually selected when the planning horizon is infinite. When the DC criterion is used, a rather complete theory is available for the corresponding dynamic programming formulation of the problem [14], [15], [50], [80], [100], [146], [196]. In this situation, future costs are discounted at a fixed rate $0 < \beta < 1$, and therefore, if β is not sufficiently close to 1, the asymptotic behavior of the state/cost process may not be important at all. Quite the opposite is the case with the AC criterion, under which all decision times are given equal weight and one takes the limit of time-averaged expected costs. The finite time evolution of the state/cost process is, in some sense, completely irrelevant in this case, and some sort of asymptotic stable behavior is desired, making this case mathematically much more involved than the previous one. Hence, the DC and AC can be seen as two opposite extremes in the spectrum of possible criteria that can be considered, in the sense that the first one captures primarily the performance of the process at the present and near future, and the second captures the performance at the distant future.

2.3. The optimal control problem. The *optimal control (or decision) problem* is that of selecting an admissible policy such that a given performance criterion is minimized over all admissible policies. For example, for the DC criterion, a policy $\pi^* \in \Pi$ is said to be (β)-discount ε-optimal for the initial distribution μ if

$$J_\beta(\mu,\pi^*) \le J_\beta(\mu,\pi) + \varepsilon\,, \qquad \forall\, \pi \in \Pi\,,$$

where $\varepsilon > 0$. If a policy is discount ε-optimal for all distributions $\mu \in \mathcal{P}(S)$, then it is simply called discount ε-optimal. If a policy is discount ε-optimal *for all* $\varepsilon > 0$, then it is called *discount optimal.* The (optimal) value function is given by

$$J_\beta^*(\mu) := \inf_{\pi \in \Pi} J_\beta(\mu, \pi)\,. \tag{2.4}$$

Also, if μ is concentrated at $x \in S$, we denote the value function by $J_\beta^*(x)$. Similar definitions apply to other criteria; $J_T^*(\mu)$ and $J^*(\mu)$ will denote the optimal value functions for the TC and AC criteria, respectively. For sample path AC, we define an optimal policy as follows: We say that a policy $\pi^* \in \Pi$ is *sample path AC optimal* (or *a.s. AC optimal*) if there exists a constant ρ^* such that for any initial law μ

$$J_S^*(\mu, \pi^*) = \rho^*, \qquad \mathcal{P}_\mu^{\pi^*}\text{-a.s.},$$

while for any other policy $\pi \in \Pi$ and any initial law μ'

$$J_S^*(\mu', \pi) \geq \rho^*, \qquad \mathcal{P}_{\mu'}^{\pi}\text{-a.s.}$$

The constant ρ^* is the sample path optimal average cost.

Having defined various optimality criteria, and the set of admissible policies Π, the obvious question at this point is: do there exist optimal policies? Without imposing further assumptions on our general model, the answer is no. One of the reasons behind this is that the Borel measurability assumption in the definition of admissible policies is too restrictive, in general, to be able to attain the infimum in (2.4). To circumvent this problem, either a broader sense of measurability is allowed, i.e., a larger set of admissible policies is used, or further assumptions are imposed. The first approach was taken by Shreve and Bertsekas [15], [160], [161], who considered *universally measurable* policies, a class properly containing the (Borel measurable) admissible policies defined previously; see also [50]. We will instead follow the second approach mentioned above and concentrate on the *semicontinuous model,* as studied in [15], [46], [50], [69], [86], [119], [148]–[150].

2.4. The semicontinuous model. In general, we consider the case when the one-stage cost function $c(\cdot,\cdot)$ is unbounded. Since for the most part the criteria considered in this paper are given by a sum of expected costs over the infinite horizon, then in order to avoid indeterminate situations, the following conditions will be assumed to hold throughout the rest of the paper, unless otherwise indicated.

(A2.1) $c(x,a) \geq 0$, for all $(x,a) \in K$.

(A2.2) The transition kernel $P(\cdot \mid x,a)$ is *weakly-continuous* in (x,a); that is, $v(\cdot) \in C_b(S)$ implies $\int_S v(y)P(dy \mid \cdot,\cdot) \in C_b(K)$.

(A2.3) (i) The multifunction $U(x)$ is upper semicontinuous;
(ii) $c(\cdot,\cdot) \in \mathcal{L}(K)$.

Remark 2.1. Concerning (A2.1), note that we only need to assume the cost is bounded below. The assumption that the cost is nonnegative is only made for convenience and does

not result in any loss of generality. Assumption (A2.2) is equivalent to $\int v(y)P(dy \mid \cdot,\cdot) \in \mathcal{L}(\boldsymbol{K})$, for each $v(\cdot) \in \mathcal{L}(\boldsymbol{S})$ [50, p. 52]. This property is crucial in our development.

Example 2.3. For the nonlinear stochastic system in Example 2.1, assume further that:

(i) $\boldsymbol{A}$ is compact,
(ii) for each $x \in \boldsymbol{S}$, $U(x)$ is closed (and therefore compact), and
(iii) the system function $F : \boldsymbol{K} \times \boldsymbol{W} \to \boldsymbol{S}$ is continuous.

If $c(\cdot,\cdot) \in \mathcal{L}(\boldsymbol{K})$, then, by Remark 2.1, (A2.2) will hold. Furthermore, the assumption on the compactness of $\boldsymbol{A}$ can be dispensed with if there are compact subsets $\boldsymbol{K}_1 \subseteq \boldsymbol{K}_2 \subseteq \ldots$ in $\boldsymbol{S} \times \boldsymbol{A}$, such that $\boldsymbol{K} = \bigcup_{n\in\mathbb{N}} \boldsymbol{K}_n$ and

$$\liminf_{n\to\infty} \{c(x,a) : (x,a) \in \boldsymbol{K}_n \setminus \boldsymbol{K}_{n-1}\} = +\infty\,,$$

since in this case $\boldsymbol{A}$ can be conveniently compactified, cf. [15, Cor. 8.6.1, p. 210]. Also, the case in which $\boldsymbol{S} = \mathbb{R}^n$, $\boldsymbol{A} = \mathbb{R}^m$, and $c(x,a) = x'Qx + a'Ra$, where Q and R are positive semidefinite and positive definite matrices, respectively, of appropriate dimensions can also be considered by a (one point) compactification of $\boldsymbol{A}$ [160, pp. 965–966].

Under (A2.1)–(A2.3), the *undiscounted dynamic programming map* T given by

$$T(v)(x) := \inf_{a\in U(x)} \left\{ c(x,a) + \int_{\boldsymbol{S}} v(y)P(dy \mid x,a) \right\}, \qquad \forall\, x \in \boldsymbol{S}\,, \tag{2.5}$$

maps $\mathcal{L}(\boldsymbol{S})$ into itself. Also, for $0 < \beta < 1$, the *discounted* dynamic programming map $T_\beta : \mathcal{L}(\boldsymbol{S}) \to \mathcal{L}(\boldsymbol{S})$ is given by

$$T_\beta(v) := T(\beta v)\,. \tag{2.6}$$

The following properties are easily verified.

Lemma 2.1. *Let $v, v' \in \mathcal{L}(\boldsymbol{S})$. Then*

(i) *for all $k \in \mathbb{R}$, $T(v + k) = T(v) + k$;*
(ii) *if $v \le v'$, then $T(v) \le T(v')$.*

Some key results for the stochastic control problem under a DC criterion are summarized in the following theorem.

Theorem 2.1. *Under the hypotheses (A2.1)–(A2.3),*

(i) *the following equation, which is called discounted cost optimality equation (DCOE), holds*

$$J^*_\beta(x) = T_\beta(J^*_\beta)(x) = \inf_{a\in U(x)} \left\{ c(x,a) + \beta \int_{\boldsymbol{S}} J^*_\beta(y)P(dy \mid x,a) \right\}, \qquad x \in \boldsymbol{S}\,; \tag{2.7}$$

(ii) *a policy $\pi^* \in \Pi_{SD}$ is discount optimal if and only if $\pi^*(x)$ attains the infimum in (2.7), for all $x \in \boldsymbol{S}$;*
(iii) *a discount optimal policy $\pi^* \in \Pi_{SD}$ exists;*

(iv) *define* $T_\beta^0 : \mathcal{L}(S) \to \mathcal{L}(S)$ *as the identity operator and* $T_\beta^k : \mathcal{L}(S) \to \mathcal{L}(S)$, $k \in \mathbf{N}$, *by* $T_\beta^k(f) := T_\beta(T_\beta^{k-1}(f))$. *Then for any* $f \in \mathcal{L}_b(S)$

$$T_\beta^k(f)(x) \xrightarrow[k\to\infty]{} J_\beta^*(x), \qquad \text{for all } x \in S;$$

(v) $J_\beta^*(\cdot)$ *is nonnegative and lower semicontinuous.*

Remark 2.2. The above results are essentially contained in [15], [50]. The usual approach is to prove the existence of a solution to the DCOE via a contraction mapping theorem [80]. The existence of a measurable selector which attains the infimum in (2.7), e.g. the result in (iii) of Theorem 2.1, follows from [15, Prop. 7.33, p. 153], [29], [46, pp. 35–38], [50, Sect. 2.6], [86], [135, Th. 4.1, p. 9], [180, Th. 9.1, p. 880]. The scheme used in (iv) of Theorem 2.1 to compute $J_\beta^*(\cdot)$ is called the *value iteration* (or successive approximations) algorithm. Note that $J_\beta^*(\cdot)$ is not necessarily the only fixed point of T_β; however, $J_\beta^*(\cdot)$ is the *minimal* fixed point of T_β among the class of nonnegative functions in $\mathcal{L}(S)$ [15, Chap. 5].

§3. A Sketch of Historical Development

We now present a brief historical sketch of the development of CMP, with an emphasis on the average cost criterion. The roots of CMP can be traced back to the pioneering work of Wald [182], [183] on sequential analysis and statistical decision functions. In the late 1940's and early 1950's, several investigators formulated the essential concepts of CMP, which are found in their work in sequential game models. A CMP can be viewed as a one-player game. Of particular interest is the work of Bellman and Blackwell [12], Bellman and LaSalle [13], and also Shapley, who formulated the essential mechanism of stochastic dynamic programming and used the theory of contraction mappings [156]. Using his famous heuristic 'minimum cost to go,' Bellman showed how powerful the dynamic programming technique was, by using it to solve problems in a myriad of settings [9]–[11]. Bellman studied mostly problems with a finite horizon, for which the backward induction approach of dynamic programming suffices to give a complete treatment. The situation is quite different in problems over an infinite horizon. Early work on CMP is also reported in econometrics [4], [48].

Howard [92] was apparently the first to study CMP with an average cost criterion. His *policy iteration* algorithm was the first major computational breakthrough, and his book helped establish CMP as an independent subject of investigation. For CMP with finite state and action spaces, Howard's policy iteration scheme established the existence of a stationary deterministic policy, optimal in this class only. Derman [37] and Viskov and Shiryaev [179] independently showed that this policy was optimal among all admissible policies. Other computational methods were later proposed. Manne [121] gave a linear programming formulation for the AC criterion, and Wagner [181] later characterized extreme-point optima of the linear program as stationary deterministic policies. White [193] introduced the value iteration (or successive approximations) technique. Excellent accounts of these and other computational methods are given in [14, Sect. 5.2] and [133].

On the theoretical side, Blackwell's seminal paper [18] gave considerable impetus to research in this area, motivating numerous other papers. In [18] Blackwell studied CMP with finite state and action spaces. He considered the DC criterion in great detail, and established many important results. In the same paper, he initiated an approach for the AC case which we will refer to as the *vanishing discount approach:* he treated the AC case as a limit of the DC case, when the discount factor goes to 1, i.e., the discounting effect vanishes. Blackwell established in [18] the existence of a stationary deterministic policy which is discount optimal, for all β sufficiently close to one. This type of optimality is now called *Blackwell optimality* [14, pp. 336–341]. The relation between the discounted and average case also becomes apparent via Tauberian theorems [85, Sect. 4.6]. This fact seems to have been observed first by Gillette [77], who used Tauberian theorems to establish the existence of optimal stationary policies in a stochastic game problem with an AC criterion. Also, using Tauberian theorems, Derman [37] showed that the Blackwell optimal policy found in [18] was also optimal for the AC criterion. Average cost CMP with finite state and arbitrary action spaces were studied under various conditions in the works of [35], [55]–[57], [97].

Blackwell optimal policies do not necessarily exist when the state space is countably infinite [118]. In fact, average optimal policies need not exist in this situation [117], [146]. Similar non-existence result holds when the state space is finite but the action space is an arbitrary compact metric space [8]. For such models the existence of an optimal policy has been proved by Bather [8], Martin-Löf [122] and Feinberg [56], under certain conditions. Derman [38] studied the problem with countable state space, finite action space and bounded cost. He studied the following equation which became known as the *average cost optimality equation* (ACOE)

$$\rho + h(i) = \min_{a \in U(i)} \left\{ c(i,a) + \sum_{j \in S} P(j \mid i,a) h(j) \right\}$$

where ρ is a scalar, $h : S \to \mathbb{R}$, $S = \mathbb{N}_0$, and we write $P(j \mid \cdot, \cdot)$ for $P(\{j\} \mid \cdot, \cdot)$. He showed that if the ACOE has a *bounded solution,* i.e, a solution (ρ, h) with $h(\cdot)$ a bounded function, then the stationary deterministic policy realizing the pointwise minimum on the right-hand side of the ACOE is average optimal, and ρ is the minimum average cost. Derman's paper, in conjunction with Derman-Veinott [42], showed that a sufficient condition for the existence of such a solution was that the expected hitting time of a fixed state under *any* stationary deterministic policy is bounded uniformly with respect to the choice of the policy and the initial state. Motivated by Blackwell's work, Taylor [173] extended the vanishing discount approach to obtain a bounded solution for a Markovian sequential replacement problem by studying the asymptotics of the *differential* discounted value function $h_\beta(\cdot) := J_\beta(\cdot) - J_\beta(0)$. Ross [143], [144] refined Taylor's procedure and showed that, under the Derman-Veinott [42] condition, $\{h_\beta(\cdot)\}_{\beta \in (0,1)}$ was uniformly bounded in β. By letting $\beta \uparrow 1$ Ross established that the ACOE had a bounded solution. This made the vanishing discount approach very popular. In subsequent works many variants of Derman-Veinott recurrence conditions appeared. See [51], [174] for a great variety of such conditions.

These conditions are difficult to remove, and counterexamples abound [146]. Actually, it has been shown in [62], in a very general setting, that the uniform boundedness of $\{h_\beta(\cdot)\}_{\beta\in(0,1)}$ in β is also a *necessary* condition for a bounded solution to the ACOE to exist. Cavazos-Cadena [30], [31], under some additional conditions, showed that the existence of bounded solutions to the ACOE necessarily impose a very strong recurrence structure on the model. Lippman [112] studied controlled semi-Markov processes with unbounded cost with both discounted and average cost criteria. Following the vanishing discount approach he derived results for the average cost case under several restrictive assumptions. Federgruen et al. [52] have explored the same approach.

Hordijk [89] extended many earlier results to countable state space and compact action spaces. He introduced the Lyapunov function method for CMP. He used this method to obtain a (possibly *unbounded*) solution to the ACOE, yielding an optimal policy. However, the Lyapunov function method necessarily imposes a blanket stability of the processes (in the sense of positive recurrence). Such stability is not always met in, e.g., many queueing model applications. In addition, he introduced some new concepts, particularly based on the relation of stochastic dynamic programming with Markov potential theory. There is a vast amount of literature devoted to CMP in several volumes of the Mathematisch Centrum tracts; see [177] and the references therein.

With Hordijk's work, it appeared that a shift away from the vanishing discount approach was necessary. Rosberg-Varaiya-Walrand [140] treated the average cost criterion as the limiting case of the finite horizon problem, but details of their arguments depend heavily on the specifics of the problem they consider, viz. the control of two queues in tandem with a linear cost structure. Federgruen and Tijms [54] initiated a direct study of the ACOE by a span semi-norm method, for bounded costs. This method allows one to obtain useful value iteration algorithms. Later Federgruen et al. [53] treated the problem with countable state space and unbounded costs. Assuming a recurrence condition on the model, they established the existence of a (possibly unbounded) solution to the ACOE, thereby establishing the existence of an optimal stationary deterministic policy.

In a series of papers [20]–[25] Borkar presented a convex analytic approach to treat the problem with countable state space, compact action space and unbounded cost. This approach can be seen as an extension of the ideas in Manne's [121] and Wagner's [181] work. Borkar stressed the existence of an optimal *stable* stationary deterministic policy, i.e., one that induces a positive recurrent process. While a blanket stability assumption (e.g. of Lyapunov type) may be too restrictive to cover many queueing applications, it nevertheless is desirable that the optimal policy be stable. Borkar showed that to obtain an optimal stable stationary deterministic policy either a blanket stability hypothesis or a condition on the cost that penalizes unstable behavior is necessary. He also emphasized the concept of almost sure optimality by a 'pathwise' treatment of the problem. A comprehensive account of the convex analytic approach to CMP is given in [26].

After the extensive works of Hordijk, Federgruen et al., and Borkar, it seemed that the vanishing discount approach was not appropriate for many classes of problems with unbounded costs. However, this approach has been revived and generalized to a great

extent in [17], [59], [61], [72], [74], [75], [81], [83], [151], [152], [163], [168], [186]. In some of these references, an *inequality* version of the ACOE is studied. In view of the results of [30], [31], [62], it is clear that a bounded solution to the ACOE is too restrictive, in general. A natural candidate solution is one that is bounded below [28], [74], [83], [151], [152], [168], [186], or one having suitable growth properties [28], or satisfying other conditions [163]. Weber and Stidham [168], [186], studied the problem for queueing systems. Under a penalizing condition on the cost and some structural assumptions, they established the existence of a (possibly unbounded) solution to the ACOE, and showed the existence of an optimal stationary deterministic policy. Sennott proceeded along similar lines. She identified very general conditions on the discounted value function so that the vanishing discount approach could successfully be pursued. We refer to [151]–[153], [168], [186] for many interesting examples of queueing systems, and to [34] for a comparison of different sets of assumptions. Extensions of these techniques to semi-Markov decisions processes with applications to queueing systems have been reported in [153].

The first attempt to give a description of CMP with more general state and actions spaces was carried out by Karlin [95]. Blackwell [19], Maitra [119] and Strauch [169] studied CMP with a general state space and the discounted cost criterion. Their work was significantly extended by Shreve and Bertsekas in [15], [160], [161]. Feinberg [58] studied CMP with Borel state space and with arbitrary numerical criteria, which include TC, AC, DC as particular cases. By establishing the convexity of the set of strategic measures (measures of the type $\mathcal{P}_\mu^\pi$ on the canonical space) he established the existence of an ε-optimal $f \in \Pi_{SD}$ for these criteria. De Leve [109], [110], [111] considered general state and action space CMP in continuous time with an AC criterion, with an emphasis on the ergodic behavior of the processes. Ross [144] used the vanishing discount approach to study CMP with an AC criterion, general state space, finite action space, and bounded cost function. He showed that if the family of differential discounted value functions $\{h_\beta(\cdot)\}_{\beta\in(0,1)}$ is equicontinuous and uniformly bounded, then the ACOE admits a bounded solution, yielding an optimal stationary deterministic policy. Ross also introduced the concept of minorant. He showed that if there exists a state $x_0 \in S$ and $\alpha > 0$ such that

$$P(x_0 \mid x, a) > \alpha, \quad \text{for all } a \in U(x), \quad x \in S$$

then the average cost case could be reduced to a discounted one. This was greatly extended in the work of Gubenko and Statland [78] (see also [42]). They showed that under similar minorant (or majorant) conditions a contraction map, with respect to the sup norm, could be defined on $\mathcal{M}_b(S)$, which would yield a bounded solution to the ACOE. They also obtained bounded solutions to the ACOE under continuity and boundedness conditions which guarantee that $\{h_{\beta_n}(\cdot)\}$, with $\beta_n \uparrow 1$, is uniformly bounded and equicontinuous, and thus a similar approach as in [144] can be followed. Georgin [70], [71] also explored this approach, under some ergodicity conditions. Tijms [175] and Hübner [93] directly studied the ACOE, under some ergodicity assumptions, by showing that the undiscounted dynamic programming map is a contraction on $\mathcal{M}_b(S)$, with respect to the span seminorm. For an excellent presentation of these methods, and the type of ergodicity conditions used,

see [80, Sect. 3.3]. Wijngaard [197], [198] and Kumar [98] studied the problem under Doeblin's condition using an operator theoretic method. Under several conditions, Kurano [101] obtained solutions to the ACOE and also showed the existence of an average optimal stationary deterministic policy. Also, in [102]–[104], he obtained the existence of an optimal stationary deterministic policy under Doeblin's condition. For a comprehensive presentation of the different recurrence conditions used for the above purposes, see [84].

The study of *partially observable controlled Markov processes* (POCMP) was initiated independently by various authors [5], [45], [49], [157], [158]. The reduction to models with complete information was done in various cases in [5], [134], [147], [201]. The study of finite state space POCMP with an AC criterion was initiated by Sondik [166]. Transforming the problem into an equivalent *completely observable* (CO) problem with Borel state space, Sondik tried to cast the problem in the framework of Ross [144], but did not show equicontinuity of $\{h_\beta(\cdot)\}_{\beta\in(0,1)}$. Ross [146], Wang [185], and White [187] showed this for specific, scalar replacement problems. Ohnishi et al. [128] studied a multi-state replacement problem by using concavity properties of $h_\beta(\cdot)$. Platzman studied the general problem of finite state and action space POCMP, also by using concavity properties of the functions $h_\beta(\cdot)$. Under certain reachability conditions he proved that the family $\{h_\beta(\cdot)\}_{\beta\in(0,1)}$ is uniformly bounded. However, even though this family may not be equicontinuous with respect to the Euclidean metric, he showed that it is equi-Lipschitzian with respect to some other appropriate metric, thus putting the problem within the framework of Ross [144]. Fernández-Gaucherand et al. [60], [61] followed a different approach to the problem, using the concepts of invariant sets of a CMP and controlled sub-Markov processes. This approach allows one to consider POCMP with countable state and observation spaces. Borkar [26] also studied the problem via his convex analytic approach.

§4. Finite State Space

In this section we will consider models with a finite state space. Initially, we restrict our attention to the case when $\boldsymbol{A}$ is a finite set; models with compact action space will be discussed at the end of the section.

4.1. Finite action spaces. Let $S = \{1,\dots,k\}$. In this case Π_{SD} is finite. This fact plays a crucial role in the analysis for the average cost problem. For a policy $\pi \in \Pi$, let $J_\beta(\pi)$ denote the vector $\big(J_\beta(1,\pi),\dots,J_\beta(k,\pi)\big)^T$; similarly we define $J_N(\pi)$, $J(\pi)$, J^*_β, J^* and J^*_N. For a stationary deterministic policy $f \in \Pi_{SD}$, let $P(f)$ denote the transition matrix of the corresponding process and

$$c(f) := \big(c(1,f(1)),\dots,c(k,f(k))\big)^T.$$

Also, the (i,j)-th entry in the n-th power of the transition matrix $P(f)$ will be denoted by $P^n_{ij}(f)$ or $P^n(f)(i,j)$. It is well known that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}P^n(f) := P^*(f)$$

exists [18], [85, Chap. 4], [133], where $P^0(f) = I$ (the $k \times k$ identity matrix). Using the theory of stochastic matrices the following results can be proved. For details see [8], [18], [85], [133].

Theorem 4.1. *For each $f \in \Pi_{SD}$,*

(i) $J(f) = P^*(f)c(f)$

(ii) *The number of linearly independent equations in*

$$\big(I - P(f)\big)w = c(f) - J(f)$$

is k minus the number of communicating classes in $P(f)$.

(iii) *The equations*

$$\big(I - P(f)\big)w = c(f) - v \tag{4.1}$$

$$P^*(f)w = 0 \tag{4.2}$$

have solutions $v = J(f)$ and $w = w(f)$ where

$$w(f) := \big(I - P(f) + P^*(f)\big)^{-1}\big(I - P^*(f)\big)c(f)\,.$$

(iv) *$v = J(f)$ and $w = w(f)$ are the unique solutions to (4.1) and (4.2) for which $v(s) = v(s')$ if s and s' are in the same communicating class of $P(f)$, and $v(s) = J(s, f)$ if state s is transient in $P(f)$.*

Remark 4.1.

(a) It is easily seen from the above theorem that if under an $f \in \Pi_{SD}$, the process is irreducible or unichain (see [85]), then $J(\cdot\,, f)$ is constant.

(b) The matrix

$$H(f) := \big(I - P(f) + P^*(f)\big)^{-1}\big(I - P^*(f)\big)$$

is called the *deviation matrix*. It plays a fundamental role in the analysis. For the discounted case $J_\beta(f) = \big(I - \beta P(f)\big)^{-1}c(f)$. Analogous results can be developed for the average cost case using $H(f)$. The following result, due to Miller and Veinott [123] and Lamond and Puterman [107], can be proved using the spectral theory of stochastic matrices.

Theorem 4.2. *Let $\beta \in [0,1)$ and $\lambda = (1-\beta)\beta^{-1}$. Let $f \in \Pi_{SD}$ and ν the eigenvalue of $P(f)$ less than one with largest modulus. If $0 \le \lambda \le 1 - |\nu|$, then*

$$(\lambda I + I - P)^{-1} = \lambda^{-1}P^*(f) + \sum_{n=0}^{\infty}(-\lambda)^n H^{n+1}(f) \tag{4.3}$$

and

$$J_\beta(f) = (1+\lambda)\Big[\lambda^{-1}P^*(f)c(f) + \sum_{n=0}^{\infty}(-\lambda^n)H^{n+1}(f)c(f)\Big]. \tag{4.4}$$

Remark 4.2.

(a) The quantity $h(f) := H(f)c(f)$ plays a crucial role in the analysis of the problem. It is called the *bias* or *transient cost.* It can be easily seen from the Neumann series expansion of $\big(I - P(f) + P^*(f)\big)^{-1}$ [133] that for $s \in S$

$$h(f)(s) = E_s^f \left[\sum_{t=0}^{\infty} \Big(c\big(X_t, f(X_t)\big) - J(X_t, f) \Big) \right].$$

From the above representation $h(f)$ can be interpreted as the expected total cost for a CMP with cost $c - J$. If $P(f)$ is aperiodic, the distribution of X_t converges to a limiting distribution, so eventually $c\big(X_t, f(X_t)\big)$ and $J(X_t, f)$ will differ very little. Thus, $h(f)$ can be thought of as the expected total cost 'until convergence' or the expected total cost during the 'transient' phase of the evolution of the process [133].

(b) Howard [92] has shown that

$$J_N(f) = NJ(f) + h(f) + o(1).$$

Therefore, as N becomes large, for each $s \in S$, $J_N(f)$ approaches a straight line with slope $J(f)$ and intercept $h(f)$. When $J(f)(s)$ is constant, $J_N(s) - J_N(s')$ approaches $h(f)(s) - h(f)(s')$, so that $h(f)$ is the asymptotic relative difference of starting the process in two states s and s'. That is why $h(f)$ is often referred to as the relative value.

(c) The expansion (4.4) extends Blackwell's expansion [18].

(d) Using the expansion (4.4), the following important result is immediate.

Corollary 4.1. *For $f \in \Pi_{SD}$, $J(f) = \lim_{\beta \uparrow 1} (1 - \beta) J_\beta(f)$.*

Following Blackwell [18] and Derman [39] we now prove the following existence results.

Theorem 4.3. *There exists an $f \in \Pi_{SD}$ which is discount optimal for all β sufficiently close to 1 and is also optimal for the average cost criterion.*

Proof. For each $f \in \Pi_{SD}$ and $s \in S$, $J_\beta(s, f)$ is obviously an analytic function of β. Let $\{\beta_n\}$, $0 < \beta_n < 1$ be a sequence such that $\beta_n \uparrow 1$. For a fixed n, let $f_n \in \Pi_{SD}$ be β_n-discount optimal (see Theorem 2.1). Since Π_{SD} is a finite set, the sequence $\{f_n\}$ must contain at least one $f^* \in \Pi_{SD}$ which occurs infinitely often. Let $\{\beta_{n_k}\}$ be a subsequence of $\{\beta_n\}$ such that $\beta_{n_k} \uparrow 1$ and $f^* = f_{n_1} = f_{n_2} = \cdots$. Then for every $g \in \Pi$, $J_{\beta_{n_k}}(f^*) \leq J_{\beta_{n_k}}(g)$. Since all coordinates of $J_\beta(f^*)$ and $J_\beta(g)$ are analytic functions of β, it follows that

$$J_\beta(f^*) \leq J_\beta(g)$$

for all β near 1. Since this holds for all $g \in \Pi$, it follows that f^* is β-discount optimal for all β near 1. We next show that f^* is average optimal. Let $\pi \in \Pi$. Then

$$(1 - \beta_{n_k}) J_{\beta_{n_k}}(f^*) \leq (1 - \beta_{n_k}) J_{\beta_{n_k}}(\pi), \qquad k = 1, 2, \ldots.$$

Therefore, letting $k \to \infty$ and using Theorem 4.1 and a standard Tauberian theorem (Theorem A.2 in the Appendix), it follows that

$$\begin{aligned} J(f^*) &= \lim_{\beta \uparrow 1} (1-\beta) J_\beta(f^*) \\ &= \lim_{k \to \infty} (1-\beta_{n_k}) J_{\beta_{n_k}}(f^*) \\ &\leq \limsup_{k \to \infty} (1-\beta_{n_k}) J_{\beta_{n_k}}(\pi) \leq J(\pi) \end{aligned}$$

and the proof is complete. □

We now briefly mention three numerical approaches. For details we refer to [14], [133], [176], etc. Our presentation follows [133].

Value iteration. We assume that under any $f \in \Pi_{SR}$, the corresponding chain is unichain and aperiodic. For any positive integer N, the finite horizon value function J_N^* satisfies the equation

$$J_{N+1}^* = \min_{f \in \Pi_{SD}} \{c(f) + P(f) J_N^*\}. \tag{4.5}$$

The equation (4.5) can act as an iteration equation with $J_0^* \equiv 0$ as the initial condition. Let $f_{N+1}^* \in \Pi_{SD}$ realize the minimum in (4.5). We can treat $\frac{1}{N} J_N^*$ and f_N^* as our guesses for J^* and an average optimal policy. Then $J_N^* - NJ^*$ converges as $N \to \infty$. Also, there exists and integer N_0 such that for any $N \geq N_0$, any $f \in \Pi_{SD}$ which attains the minimum in (4.5) is average optimal. However, this property does not yield an error estimate and hence fails to provide a stopping rule for the iteration scheme. To this end, with $h = \big(h(1), \ldots, h(k)\big)$, we let

$$\begin{aligned} L(h) &:= \min_{x \in S} \big\{Th(x) - h(x)\big\} \\ U(h) &:= \max_{x \in S} \big\{Th(x) - h(x)\big\}. \end{aligned}$$

It can be shown that [133]

$$\min_{x \in S} \big\{J_N^*(x) - J_{N-1}^*(x)\big\} \leq J^* \leq \max_{x \in S} \big\{J_N^*(x) - J_{N-1}^*(x)\big\}$$

and

$$L(J_{N-1}^*) \leq L(J_N^*) \leq J^* \leq U(J_N^*) \leq U(J_{N-1}^*).$$

Furthermore,

$$\lim_{N \to \infty} \big\{U(J_N^*) - L(J_N^*)\big\} = 0\,.$$

Thus, an average ε-optimal policy can be found by stopping the value iteration when

$$U(J_N^*) - L(J_N^*) < \varepsilon\,.$$

There are other variants of this approach, e.g., the relative value iteration [193] and the span contraction method [54].

Linear programming. To simplify our presentation, we will assume that under any $f \in \Pi_{SR}$, the corresponding process is irreducible. Let $P(f)$ denote the transition matrix of the process and $\eta(f) \in \mathcal{P}(S)$ its invariant measure. Then for any $s \in S$, $J(s,f) = J(f)$, a constant, and

$$J(f) = \sum_{s \in S} \sum_{a \in U(s)} c(s,a) f(s,a) \eta(f)(s) .$$

Therefore, the average cost problem can be reduced to the following linear programming problem:

(4.6a) $$\text{minimize} \quad \sum_{s \in S} \sum_{a \in U(s)} c(s,a) x(s,a)$$

subject to

(4.6b) $$x(s,a) \geq 0 , \qquad s \in S , \quad a \in U(s) ,$$

(4.6c) $$\sum_{s \in S} \sum_{a \in U(s)} x(s,a) = 1 ,$$

(4.6d) $$\sum_{a \in U(s)} x(s,a) = \sum_{s' \in S} \sum_{a \in U(s')} x(s',a) P(s' \mid s,a) .$$

Under the irreducibility assumption the simplex method can be employed to find an optimal stationary deterministic policy. This formulation is due to Manne [121].

Policy improvement. We work under the irreducibility assumption. The dual to the linear program (4.6a)–(4.6d) is the problem of finding variables g and $h(s)$, $s \in S$, in order to

(4.7a) $$\text{maximize} \quad g$$

subject to

(4.7b) $$g + \sum_{s' \in S} \big(\delta(s,s') - P(s' \mid s,a)\big) h(s) \leq c(s,a)$$

$(s,a) \in S \times U(s)$ where $\delta(s,s')$ is the Kronecker delta.

The functional equation

(4.8) $$g + h(s) = \min_{a \in U(s)} \left\{ c(s,a) + \sum_{s' \in S} P(s' \mid s,a) h(s) \right\}$$

is equivalent to (4.7a)–(4.7b) under the irreducibility assumption and is the average cost optimality equation [85]. We will discuss this equation in detail in the next section. It will

be shown that an $f \in \Pi_{SD}$ is optimal if and only if f realizes the pointwise minimum in (4.8) and then g is the optimal average cost. This suggests the following iteration algorithm.

(i) Let $n = 1$. Choose $f_n \in \Pi_{SD}$. Let $h_n(s) \equiv 0$ for all $s \in \boldsymbol{S}$.

(ii) Find a solution g_n and $h_n(s)$ of the following equation:

$$g_n + h_n(s) = c\big(s, f_n(s)\big) + \sum_{s' \in \boldsymbol{S}} P\big(s' \mid s, f_n(s)\big) h_n(s)\,.$$

(iii) For each $s \in \boldsymbol{S}$, compute

$$\phi_n(s) = \min_{a \in U(s) \setminus \{f_n(s)\}} \left\{ c(s,a) + \sum_{s' \in \boldsymbol{S}} P(s' \mid s, a) h_n(s') \right\} - g_n - h_n(s)\,.$$

If $\phi_n(s) \geq 0$ for all $s \in \boldsymbol{S}$, then f_n is average optimal and g_n is the optimal average cost. If $\phi_n(s) < 0$ for some $s \in \boldsymbol{S}$, then pick $a \in U(s)$ such that

$$c(s,a) + \sum_{s' \in \boldsymbol{S}} P(s' \mid s, a) h_n(s') - g_n - h_n(s) < 0\,.$$

Define $f_{n+1} \in \Pi_{SD}$ as $f_{n+1}(s) = a$ and $f_{n+1}(\cdot) = f_n(\cdot)$ otherwise. Then f_{n+1} yields a lower average cost. Since Π_{SD} is finite, the policy improvement scheme converges in a finite number of steps.

4.2. Compact action spaces. We now consider the problem where the action set $\boldsymbol{A}$ is not finite but a compact metric space. In this situation an optimal policy may not exist; see [50, p. 178, Example 1]. Note that here Π_{SD} is no longer finite. Under certain ergodicity assumptions Martin-Löf [122] and Feinberg [55] have proved the existence of an optimal $f \in \Pi_{SD}$. We will discuss various ergodicity assumptions on a countable state space in detail in the next section. Here we focus on ε-optimal policies established by Chitashvili [35] and Feinberg [56]; see [50, Chap. 7].

Theorem 4.4. *Under (A2.1)–(A2.3), for every $\varepsilon > 0$ there exists an ε-optimal $f \in \Pi_{SD}$.*

Proof (Sketch). For $f \in \Pi_{SD}$, let $J(f)$ be as in Theorem 4.1. For $i \in \boldsymbol{S}$, let

$$\tilde{J}(i) = \inf_{f \in \Pi_{SD}} J(f)(i)\,. \tag{4.9}$$

Clearly $J^*(i) \leq \tilde{J}(i)$, for each $i \in \boldsymbol{S}$. Corresponding to $i \in \boldsymbol{S}$, select an $f_i \in \Pi_{SD}$ such that

$$J(f_i)(i) \leq \tilde{J}(i) + \varepsilon\,. \tag{4.10}$$

The set $\widetilde{\boldsymbol{A}} = \big\{f_i(j) : i, j \in \boldsymbol{S}\big\}$ is obviously finite. Taking the action set to be $\widetilde{\boldsymbol{A}}$, the preceding results can be applied to the finite CMP $\big(\boldsymbol{S}, \widetilde{\boldsymbol{A}}, P, c\big)$. For this model there exists a stationary deterministic policy, say f^*, which is average optimal. Thus

$$J(f^*)(i) \leq J(f_i)(i) \leq \tilde{J}(i) + \varepsilon\,, \qquad \text{for each } i \in \boldsymbol{S}\,. \tag{4.11}$$

Let

$$\rho^*(i) := \limsup_{\beta \uparrow 1} (1-\beta) J^*_\beta(i).$$

Then by Theorem A.2

$$\rho^*(i) \le J^*(i), \qquad \text{for each } i \in S. \tag{4.12}$$

By (4.11), it suffices to show that $J^*(i) = \tilde{J}(i)$, for each $i \in S$. From (4.12), it then suffices to show that $\rho^*(i) \ge \tilde{J}(i)$. For each $\beta \in (0,1)$, let $f_\beta \in \Pi_{SD}$ be β-discount optimal. Let f be a limit point of f_β as $\beta \uparrow 1$. Then using (4.4) (which is valid in this case as well) and (A2.1)–(A2.3), it can be shown that $\rho^*(i) \ge J(f)(i) \ge \tilde{J}(i)$. □

§5. Countable State Space

The average cost problem becomes much more complicated when the state space is countable. Maitra [117] has given a counterexample which shows that there need not exist an optimal policy. In [118] Maitra has studied a particular problem in which there does not exist any policy which is β-discount optimal for all β sufficiently close to 1. Flynn [66] has constructed a more dramatic counterexample. In his example, there exists an average optimal policy in Π_{SD}. Nevertheless he exhibits an $f \in \Pi_{SD}$ and a $\beta_0 \in (0,1)$ such that f is β-discount optimal for all $\beta \in (\beta_0, 1)$, but it is not average optimal. Fisher and Ross [65] have presented a counterexample which shows that the optimal policy need not be stationary or deterministic. We refer to [146] for several other counterexamples. It is apparent that the average cost problem is closely related to the ergodic behavior of the process and it is well known that the ergodic theory of Markov processes on a countable state space is much more involved than on a finite state space; for example, a Markov process on a finite state space cannot be null recurrent. Another vital difference in this case is that the number of stationary deterministic policies is no longer finite. To study the ergodic theory some recurrence conditions are necessary. There are many such conditions available in the literature [26], [51], [174]; we will survey a few representative ones.

In what follows the state space $S = \{0, 1, 2, \dots\}$. For each $i \in S$, the action space $U(i)$ is a prescribed compact metric space. We will always assume that for fixed $i, j \in S$, $c(i,\cdot)$, $P(i \mid j, \cdot)$, are continuous. These conditions can be weakened or dropped in several places, as will be clear from the specific context.

Derman [37] studied the ACOE which, with ρ a scalar and $h : S \to \mathbb{R}$, takes the form:

$$\rho + h(i) = \min_{a \in U(i)} \left\{ c(i,a) + \sum_{j \in S} P(j \mid i, a) h(j) \right\}. \tag{5.1}$$

A solution to (5.1) is a pair (ρ, h) satisfying it.

Suppose $f \in \Pi_{SD}$ is a minimizing selector in (5.1). Then (5.1) becomes

$$\rho + h(i) = c(i, f(i)) + \sum_{j \in S} P(j \mid i, f(i)) h(j). \tag{5.1'}$$

Equation (5.1′) asserts that, apart from ρ, the cost if the process stops now equals the expected cost if it continues under the policy f for just one more period. We can give a similar interpretation to (5.1). Hence, we may think that ρ is the average cost under f and no other $f \in \Pi_{SD}$ has a smaller average cost. Thus, the function h in (5.1) is roughly a measure of how much we are prepared to pay to stop the process, though continuing to pay an average cost ρ in the future [137] (cf. Remark 4.2, (a)). Therefore, the function h may be viewed as a cost potential. Also, by a stochastic representation of h, using (5.1) and (5.1′), h is indeed a potential. Hordijk [89] has pursued this line of thought in great detail, which we will discuss later.

We start with a characterization of optimal policies.

Theorem 5.1. *If the ACOE has a solution (ρ, h) satisfying*

$$\lim_{t\to\infty} \frac{1}{t} E_i^{\pi} h(X_t) = 0\,, \qquad \forall\, \pi \in \Pi_{SD}\,, \quad \forall\, i \in S\,, \tag{5.2}$$

then there exists an $f \in \Pi_{SD}$ such that

$$\rho = J(i,f) = J^*(i)\,, \qquad \forall\, i \in S\,.$$

Moreover, an $f \in \Pi_{SD}$ is average optimal if for each $i \in S$

$$c\big(i, f(i)\big) + \sum_{j\in S} P\big(j \mid i, f(i)\big) h(j) = \min_{a\in U(i)} \left\{ c(i,a) + \sum_{j\in S} P(j \mid i,a) h(j) \right\}, \tag{5.3}$$

and, conversely, if an $f \in \Pi_{SD}$ is average optimal and the corresponding chain is irreducible and positive recurrent then (5.3) holds.

Proof. Let $f \in \Pi_{SD}$ satisfy (5.3). Then since

$$E_i^f\big[h(X_{t+1}) \mid \mathfrak{F}_t\big] = \sum_{j\in S} P\big(j \mid X_t, f(X_t)\big) h(j)\,,$$

it follows from (5.1) and (5.3) that

$$\rho + h(X_t) = c\big(X_t, f(X_t)\big) + E_i^f\big[h(X_{t+1}) \mid \mathfrak{F}_t\big]. \tag{5.4}$$

Summing (5.4) from $t = 0$ to $N-1$, dividing by N and taking expectations, we obtain

$$\rho = \frac{1}{N} E_i^f \left[\sum_{t=0}^{N-1} c\big(X_t, f(X_t)\big) \right] + \frac{E_i^f\big[h(X_N)\big] - h(i)}{N}\,.$$

Next, letting $N \to \infty$ and using (5.2), yields

$$\rho = \lim_{N\to\infty} \frac{1}{N} E_i^f \left[\sum_{t=0}^{N-1} c\big(X_t, f(X_t)\big) \right].$$

On the other hand if π is any other policy, we can show using the same arguments that

$$\rho \leq \limsup_{N\to\infty} \frac{1}{N} E_i^{\pi}\left[\sum_{t=0}^{N-1} c(X_t, A_t)\right].$$

Hence, f is average optimal. Conversely, let $f \in \Pi_{SD}$ be average optimal and suppose that the corresponding chain is irreducible and positive recurrent. If f does not satisfy (5.3) then there exist $i_0 \in S$, $a_0 \in U(i_0)$ and $\delta > 0$ such that

$$c(i_0, f(i_0)) + \sum_{j\in S} P(j \mid i_0, f(i_0))h(j) \tag{5.5}$$
$$= c(i_0, a_0) + \sum_{j\in S} P(j \mid i_0, a_0)h(j) + \delta\,.$$

Let $f' \in \Pi_{SD}$ be defined as follows:

$$f'(i) = \begin{cases} f(i) & \text{if } i \neq i_0, \\ a_0 & \text{if } i = i_0. \end{cases}$$

Then, using (5.5) along with irreducibility and positive recurrence, it is easily seen that

$$J(i_0, f') < J(i_0, f)$$

which contradicts the average optimality of f. □

Remark 5.1.

(a) We say that (5.1) admits a bounded solution if $h(\cdot)$ is bounded. If the ACOE has a bounded solution, then (5.2) is clearly satisfied; moreover, using the martingale stability theorem [114, p. 53] it can be shown that the $f \in \Pi_{SD}$ selecting the minimum in (5.3) is sample path average optimal [70].

(b) Various extensions of last assertion of Theorem 5.3 have been obtained by Sennott [154].

Derman and Veinott [42] have prescribed a certain recurrence condition which ensures that (5.1) admits a bounded solution. We will discuss it later in this section. The ACOE resembles the dynamic programming equation, and Theorem 5.1 is analogous to a dynamic programming characterization of an optimal policy. However, the dynamic programming heuristic does not lead directly to the ACOE. Taylor [173] developed a vanishing discount approach for a particular problem which was extended for the general case by Ross [143]–[146]. Our presentation here follows Ross [146]. As noted earlier, the average case can in some sense be treated as the limiting case of the discounted problem as the discount factor approaches one. The discounted value function $J_\beta^*(\cdot)$ satisfies the DCOE (cf. Theorem 2.1)

$$J_\beta^*(i) = \min_{a\in U(i)} \left\{ c(i,a) + \beta \sum_{j\in S} P(j \mid i, a) J_\beta^*(j) \right\}$$

and a β-discounted optimal policy selects a minimizing action. One possible way of finding an average optimal policy might be to choose the actions minimizing

$$\lim_{\beta \to 1} \left\{ c(i,a) + \beta \sum_{j \in S} P(j \mid i,a) J_\beta^*(j) \right\}.$$

However, this limit need not exist and indeed would often be infinite for all actions. The situation can nevertheless be salvaged by considering a 'differential' discounted value function, i.e., $h_\beta(i) := J_\beta^*(i) - J_\beta^*(0)$, where $0 \in S$ is an arbitrary, fixed state. The function $h_\beta(\cdot)$ satisfies

$$(1-\beta) J_\beta^*(0) + h_\beta(i) = \min_{a \in U(i)} \left\{ c(i,a) + \beta \sum_{j \in S} P(j \mid i,a) h_\beta(j) \right\}. \tag{5.6}$$

From (5.6) it is now apparent that (5.1) can be derived under certain conditions by letting $\beta \to 1$. Indeed we have the following result [146].

Theorem 5.2. *Suppose there exists a constant $K > 0$ such that $|h_\beta(i)| \leq K$, for all $\beta \in (0,1)$ and $i \in S$. Then*

(i) *the ACOE admits a bounded solution (ρ, h).*
(ii) *For some sequence $\beta_n \to 1$, $h(i) = \lim_{n \to \infty} h_{\beta_n}(i)$, $i \in S$.*
(iii) $\lim_{\beta \to 1} (1-\beta) J_\beta^*(i) = \rho$ *for any $i \in S$.*

Proof. Let $\beta_n \uparrow 1$ be given. By the uniform boundedness of $h_\beta(\cdot)$, using a diagonalization procedure, we can find a subsequence, which for simplicity we also denote by β_n, such that $h_{\beta_n}(i) \to h(i)$ for each $i \in S$, where $h(\cdot)$ is a bounded function. Again, since $(1-\beta_n) J_{\beta_n}^*(0)$ is bounded, there is a further subsequence $\beta_{n_k} \uparrow 1$ such that

$$\lim_{k \to \infty} (1-\beta_{n_k}) J_{\beta_{n_k}}^*(0)$$

exists. (i) then follows from (5.6) and an application of the dominated convergence theorem. Furthermore, by Theorem 5.1, ρ is the minimum average cost. Since the above results are independent of the sequence chosen, then (iii) follows. □

Remark 5.2. It has been shown [62] that if the ACOE has a bounded solution then there exists a constant $K > 0$ such that $|h_\beta(i)| \leq K$ for all $\beta \in (0,1)$, $i \in S$.

5.1. Bounded costs. In this subsection we assume that $c(\cdot\,,\cdot)$ is bounded. Ross [146] has proved that under a Derman-Veinott [42] type recurrence condition (see (5.7) below), the uniform boundedness hypothesis of Theorem 5.2 is satisfied.

Theorem 5.3. *Let $f \in \Pi_{SD}$ and let $\{X_t\}$ be the corresponding state process. Let*

$$\tau = \min\{t \geq 1 : X_t = 0\}.$$

If there exist a $K > 0$ such that

$$E_i^f[\tau] < K \tag{5.7}$$

for all $f \in \Pi_{SD}$ and all $i \in \boldsymbol{S}$, then $h_\beta(i)$ is bounded uniformly in $\beta \in (0,1)$ and $i \in \boldsymbol{S}$.

Proof. Let $\beta \in (0,1)$ and $f_\beta \in \Pi_{SD}$ be β-discount optimal. We have

$$
\begin{aligned}
J^*_\beta(i) &= E_i^{f_\beta}\left[\sum_{t=0}^{\infty} \beta^t c\big(X_t, f_\beta(X_t)\big)\right] \\
&= E_i^{f_\beta}\left[\sum_{t=0}^{\tau-1} \beta^n c\big(X_t, f_\beta(X_t)\big)\right] + E_i^{f_\beta}\left[\sum_{t=\tau}^{\infty} \beta^t c\big(X_t, f_\beta(X_t)\big)\right] \\
&\le M E_i^{f_\beta}[\tau] + J^*_\beta(0) E_i^{f_\beta}[\beta^\tau].
\end{aligned} \tag{5.8}
$$

where M is a bound on $c(\cdot,\cdot)$. From (5.7) and (5.8) it follows that

$$
J^*_\beta(i) - \beta J^*_\beta(0) \le MK\,. \tag{5.9}
$$

Also, from (5.8) and applying Jensen's inequality, we get

$$
J^*_\beta(i) \ge J^*_\beta(0) E_i^{f_\beta}[\beta^\tau] \ge J^*_\beta(0)\beta^K.
$$

Therefore,

$$
\begin{aligned}
J^*_\beta(0) - J^*_\beta(i) &\le (1-\beta^K) J^*_\beta(0) \\
&\le (1-\beta^K)\frac{M}{1-\beta} \le MK\,.
\end{aligned} \tag{5.10}
$$

The desired result follows from (5.9) and (5.10). □

After the work of Derman [37], Derman-Veinott [42], and Ross [143], [144], several recurrence conditions have appeared [174]. We will look into a few representative ones.

Let $f \in \Pi_{SD}$. For a finite set $A \subset \boldsymbol{S}$, let

$$
\tau_A = \min\{t \ge 1 : X_t \in A\}. \tag{5.11}
$$

(A5.1) There is a finite $A \subset \boldsymbol{S}$ and a constant $K > 0$ such that

$$
E_i^f[\tau_A] < K
$$

for all $i \in \boldsymbol{S}$ and $f \in \Pi_{SD}$. Further, for any $f \in \Pi_{SD}$ the corresponding process does not have two disjoint invariant sets.

(A5.2) There exists a constant $K > 0$, and for every $f \in \Pi_{SD}$ there is a state $j(f) \in \boldsymbol{S}$ such that

$$
E_i^f[\tau_{\{j(f)\}}] < K\,, \qquad \forall\, i \in \boldsymbol{S}\,.
$$

(A5.3) (Simultaneous Doeblin) There is a finite set $\boldsymbol{A}$, an integer $n \ge 1$ and a scalar $\alpha > 0$ such that

$$
\sum_{j \in A} P\big(j \mid i, f(i)\big) \ge \alpha
$$

for all $i \in S$ and all $f \in \Pi_{SD}$. Further, for any $f \in \Pi_{SD}$ the corresponding process does not have two disjoint invariant sets.

(A5.4) (Scrambling) There is an integer $n \geq 1$ and a scalar $\alpha > 0$ such that for any $f \in \Pi_{SD}$

$$\sum_{j \in S} \min\{P^n_{i_1,j}(f), P^n_{i_2,j}(f)\} \geq \alpha, \qquad \forall i_1, i_2 \in S.$$

(A5.5) (Ergodicity) There is an integer $n \geq 1$ and a scalar $\rho > 0$ such that for each $f \in \Pi_{SD}$ there exists an $\eta(f) \in \mathcal{P}(S)$ for which

$$\sum_j |P^m_{ij}(f) - \eta(f)(j)| \leq 2(1-\rho)^{\lfloor m/n \rfloor}$$

for all $i \in S$ and $m \geq 1$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x.

Remark 5.3. Clearly (A5.1) and (A5.2) are generalizations of the Derman-Veinott condition. Hordijk [89] has proved the existence of a bounded solution to the ACOE using (A5.1). Under (A5.5), for each $f \in \Pi_{SD}$, $\eta(f)$ is the unique invariant measure of the corresponding process.

Federgruen et al. [51] have established the following.

Theorem 5.4. *The conditions (A5.1)–(A5.3) are equivalent. Also, if for any $f \in \Pi_{SD}$ the corresponding process is aperiodic, then (A5.1)–(A5.5) are equivalent.*

Remark 5.4. Under any one of the conditions (A5.1)–(A5.5), Federgruen et al. [51] have established the existence of a bounded solution to the ACOE by extending the vanishing discount approach of Taylor and Ross.

We have thus far seen several recurrence conditions which are sufficient for the ACOE to admit a bounded solution. Cavazos-Cadena [30], [31] has dealt with the converse question of what are the necessary recurrence conditions for the ACOE to have a bounded solution. He has obtained the following result. Consider the condition:

(A5.6) There exists a constant $K > 0$ such that for each bounded and measurable $c : S \times A \to \mathbb{R}$ and every collection $\{U(i) : i \in S\}$, $U(i) \subset A$, there exist $\rho \in \mathbb{R}$ and $h : S \to \mathbb{R}$ bounded which solve (5.1) and satisfy $\|h\| \leq K\|c\|$, where $\|\cdot\|$ is the sup norm.

Theorem 5.5. *The conditions (A5.2) and (A5.6) are equivalent.*

The proof follows by an application of the uniform boundedness principle. For details and other variants we refer to [30], [31]. Thus, an assumption on the existence of a bounded solution to the ACOE necessarily imposes a strong recurrence structure on the system. Also, note that (A5.6) involves not just one CMP but a family of CMP (one for each c and $\{U(i)\}$). Since it is equivalent to (A5.1)–(A5.3) and under aperiodicity conditions to (A5.1)–(A5.5), it follows that (A5.1)–(A5.5) are too strong for many important applications. In fact there are interesting situations [20] in which these conditions are not satisfied but for which one can find average optimal stationary deterministic policies.

Ross [144] has proved that under the following recurrence condition the AC can be reduced to an appropriate DC. Therefore, in view of Theorem 2.1, the problem can be resolved in this case.

Theorem 5.6. *If there exists a constant $\alpha > 0$ such that*

$$P(0 \mid i,a) \geq \alpha > 0$$

for all $i \in S$, $a \in U(i)$, then the AC can be reduced to an appropriate DC.

Proof. Let

$$\widetilde{P}(j \mid i,\cdot) = \begin{cases} (1-\alpha)^{-1}P(j \mid i,\cdot) & \text{for } j \neq 0 \\ (1-\alpha)^{-1}(P(0 \mid i,\cdot) - \alpha) & \text{for } j = 0. \end{cases}$$

Let $\widetilde{J}^*_\beta(\cdot)$ denote the β-discounted value function for the CMP with cost $c(\cdot,\cdot)$ and transition law $\widetilde{P}(\cdot \mid \cdot,\cdot)$. Then it is easily verified that for each $i \in S$

$$\alpha\widetilde{J}^*_{1-\alpha}(0) + \widetilde{J}^*_{1-\alpha}(i) = \min_{a \in U(i)} \left\{ c(i,a) + \sum_{j \in S} P(j \mid i,a)\widetilde{J}^*_{1-\alpha}(j) \right\}.$$

Let $f \in \Pi_{SD}$ be $(1-\alpha)$-discount optimal for the modified CMP. It follows from Theorem 5.1 that f is AC-optimal for the original CMP and the optimal average cost is $\alpha\widetilde{J}^*_{1-\alpha}(0)$. □

Remark 5.5. Note that if the ACOE has a bounded solution (ρ, h) then ρ is the optimal average cost for any initial condition. Hence, the existence of a bounded solution to the ACOE suggests that some kind of 'unichainedness' is in effect, since, for the multi-chain case the average cost would, in general, depend on the initial condition. The multi-chain version of the ACOE is

$$\min_{a \in U(i)} \sum_{j \in S} P(j \mid i,a)\rho(j) = \rho(i) \tag{5.12a}$$

$$\rho(i) + h(i) = \min_{a \in U_1(i)} \left\{ c(i,a) + \sum_{j \in S} P(j \mid i,a)h(j) \right\} \tag{5.12b}$$

where

$$U_1(i) = \left\{ a \in U(i) : \min_{a \in U(i)} \sum_{j \in S} P(j \mid i,a)\rho(j) = \rho(i) \right\}. \tag{5.12c}$$

This equation has been studied by Zijm [204] for countable state space. For more general state spaces it was extensively studied much earlier by Yushkevich [200] (see also [50]); this work will be discussed in the next section.

If (5.12) has a bounded solution $\rho(i)$, $h(i)$, where both ρ and h are bounded functions, then one can show as before that $\rho(i)$ is the optimal average cost starting from state $i \in S$ and a minimizing selector in (5.12) yields an average optimal stationary deterministic

policy. Under a certain 'geometric convergence condition' Zijm [204] has established the existence of a bounded solution to (5.12). Under the additional assumptions that under any stationary deterministic policy the corresponding process has at most a finite number of ergodic classes, he has shown that the geometric convergence condition is equivalent to a number of recurrence conditions of the type (A5.1)–(A5.5).

Hordijk [89] has established the existence of an average optimal $f \in \Pi_{SD}$ without utilizing the ACOE. Let Π_{SD} be endowed with the product topology. Then Π_{SD} is compact and metrizable. Let us consider the following assumptions.

(A5.7) For each $f \in \Pi_{SD}$ and $i \in S$, there exists a measure $\eta_i(f) \in \mathcal{P}(S)$ such that $\eta_i(f)(j) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n(f)(i,j)$.

(A5.8) $f \longmapsto \eta_i(f)$ is continuous for any $i \in S$.

(A5.9) For each $i \in S$, $\{\eta_i(f) : f \in \Pi_{SD}\}$ is tight (for a definition of tightness, see [130, Def. 3.1, p. 28]).

(A5.10) For each $f \in \Pi_{SD}$, the corresponding process is recurrent.

(A5.11) For each $f \in \Pi_{SD}$, the corresponding process does not have disjoint invariant sets.

(A5.12) $\{P(f)(i,\cdot) : i \in S, f \in \Pi_{SD}\}$ is tight.

It is easy to see that (A5.7)–(A5.8) imply that for each $i \in S$, $\{\eta_i(f) : f \in \Pi_{SD}\}$ is compact. Hence, in particular, (A5.7)–(A5.8) $\Rightarrow$ (A5.9). By definition, (A5.9) $\Rightarrow$ (A5.7). Also, it can easily be shown that (A5.9) and (A5.11) $\Rightarrow$ (A5.8), and (A5.12) $\Rightarrow$ (A5.9). However, (A5.12) may be easier to verify.

Theorem 5.7. *Each of the following five combinations of assumptions is sufficient for the existence of an average optimal $f \in \Pi_{SD}$: (A5.7, A5.8), (A5.9, A5.10), (A5.9, A5.11), (A5.10, A5.12), (A5.11, A5.12).*

Remark 5.6. The main idea behind the proof of this theorem can be traced back to the proof of Theorem 4.3. We give the main points and skip the details. Let $\beta_n \in (0,1)$ be a sequence such that $\beta_n \uparrow 1$, let $f_{\beta_n} \in \Pi_{SD}$ be β_n-discount optimal, and f_∞ be a limit point of $\{f_{\beta_n}\}$ in Π_{SD}. Suppose that $\rho^*(i)$ is a scalar satisfying $(1-\beta_n)J^*_{\beta_n}(i) \to \rho^*(i)$, for each $i \in S$ (along a suitable subsequence). Then by using Tauberian and ergodic theorems, one deduces that $J^*(i) = \rho^*(i)$ and f_∞ is average optimal under (A5.7, A5.8). Under (A5.9, A5.10), f_∞ is average optimal for initial states $i \in \widetilde{S} := \bigcup_i \operatorname{supp}(\eta_i(f_\infty))$, where 'supp' denotes the support. Then by (A5.10) there exists an $\bar{f}$ such that the corresponding process starting from any $i \in S \setminus \widetilde{S}$ reaches $\widetilde{S}$. Set

$$\widetilde{f}(i) = \begin{cases} \bar{f}(i) & \text{if } i \notin \widetilde{S} \\ f_\infty(i) & \text{if } i \in \widetilde{S}. \end{cases}$$

It follows that $\widetilde{f}$ is average optimal. The other cases can be dealt with in a similar manner.

5.2. Unbounded costs. We have thus far considered bounded costs only. There are practical situations (e.g. in queueing systems) where the cost is typically unbounded. We assume that $c \geq 0$ (cf. (A2.1)). Let us now consider the ACOE for unbounded c. Note that

the boundedness of c did not play any role in the proof of Theorem 5.1. For unbounded c the ACOE is unlikely to admit a bounded solution.

Lippman [112], [113] has studied controlled semi-Markov processes with unbounded costs. He has placed polynomial bounds on the movement of the process in one transition. He has made a further assumption that there exists an $f \in \Pi_{SD}$ such that both the mean first passage times and mean first passage costs from any state i to state 0 under the policy are finite. Moreover, if $f \in \Pi_{SD}$ is close to β-discount optimal for a sequence of discount factors, then it is AC-optimal. Lippman has employed the vanishing discount approach of Taylor and Ross to establish the existence of a solution (ρ, h) to the ACOE with h satisfying (5.2), thereby establishing the existence of an average optimal $f \in \Pi_{SD}$. He has also given some examples from queueing systems where his conditions are satisfied. However, his condition on the β-discounted value function appears to be very difficult to verify.

Hordijk [89] has used a Lyapunov stability condition to establish the existence of an average optimal $f \in \Pi_{SD}$.

(A5.13) (Lyapunov Condition) Let

$$\widetilde{P}(f)(i,j) = \begin{cases} P(f)(i,j), & j \neq 0 \\ 0, & j = 0. \end{cases}$$

There exists a function $w : S \to \mathbb{R}_+$ such that, for all $i \in S$,

(i) $c\big(i, f(i)\big) + 1 + \sum_j \widetilde{P}(f)(i,j)w(j) \leq w(i)$, for all $f \in \Pi_{SD}$.

(ii) $\sum_j P(f)(i,j)w(j)$ is continuous in f.

(iii) $\lim_{n\to\infty} \sum_j \widetilde{P}^n(f)(i,j)w(j) = 0$.

Theorem 5.8. *Under the above Lyapunov condition there exists an AC-optimal $f \in \Pi_{SD}$.*

Proof (Sketch). Let $f \in \Pi_{SD}$. For $i \in S$, we define

$$\tau_i = \min\big\{t \geq 1 \,:\, X_t = i\big\},$$

where X_t is governed by f. Then under $(A5.13)$, using the standard techniques of stochastic Lyapunov function method [105], [89], the following results can be proved:

$$E_i^f\big[\tau_0\big] \leq w(i) \tag{5.13}$$

$$E_i^f\left[\sum_{t=0}^{\tau_0 - 1} c\big(X_t, f(X_t)\big)\right] \leq w(i)\,. \tag{5.14}$$

Indeed, with $n \in \mathbb{N}$ and $n > 1$,

$$\begin{aligned} E_i^f\big[w(X_{n\wedge\tau_0}) \mid \mathfrak{F}_{n\wedge\tau_0}\big] - w(i) &= -E_i^f\left[\sum_{t=0}^{n\wedge\tau_0 - 1} E_i^f\big[w(X_{t+1}) \mid X_t\big] - w(X_t)\right] \\ &\leq -E_i^f\big[n \wedge \tau_0\big], \end{aligned}$$

where the last inequality is due to (A5.13). Hence,

$$E_i^f[n \wedge \tau_0] \leq w(i),$$

and letting $n \uparrow \infty$, (5.13) follows. Also, (5.14) can be proved along the same lines. By an ergodic theorem [129]

$$\lim_{N\to\infty} \frac{1}{N} E_0^f \left[\sum_{t=0}^{N-1} c(X_t, f(X_t))\right] = (E_0^f \tau_0)^{-1} E_0^f \left[\sum_{t=0}^{\tau_0 - 1} c(X_t, f(X_t))\right]$$
$$=: \rho(f)$$

Let

$$\rho^* := \inf_{f \in \Pi_{SD}} \rho(f).$$

Then $\rho^* \leq w(0)$. Define

$$h(i) = \inf_{f \in \Pi_{SD}} E_i^f \left[\sum_{t=0}^{\tau_0 - 1} \Big(c(X_t, f(X_t)) - \rho^*\Big)\right].$$

Then $h(0) = 0$. Using (5.13), (5.14) and (A5.13, iii), it can be shown that (ρ^*, h) is a solution of the ACOE with h satisfying (5.2), and the desired result follows. □

Remark 5.7.

(a) Note that by (A5.13, i) the cost function c does not grow faster than the Lyapunov function w. Thus, there is a restriction on the growth of c imposed by w. In CMP $w(i) = i$, $w(i) = i^2$ are typical examples of Lyapunov functions [89]. In the latter case, for example, we can treat only those unbounded cost functions which do not grow faster than quadratic functions.

(b) The condition (A5.13, iii) is crucial in showing that the cost potential h satisfies $\lim_{t\to\infty} \frac{1}{t} E_i^f h(X_t) = 0$, for all $f \in \Pi_{SD}$, and $i \in S$.

Federgruen, Hordijk and Tijms [52] have extended Hordijk's results by replacing the single attracting point $\{0\}$ by a finite set $K \subset S$. Their main assumption is: There exists a finite set $K \subset S$ such that for each initial state $i \in S$ the suprema over the mean hitting time of K and mean hitting costs are finite. This in turn is equivalent to the existence of a Lyapunov function $w : S \to \mathbb{R}_+$ satisfying (A5.13, i) where now $\widetilde{P}$ is defined as:

$$\widetilde{P}(f)(i,j) = \begin{cases} P(f)(i,j), & j \notin K,\ f \in \Pi_{SD} \\ 0, & j \in K. \end{cases}$$

Under the additional assumptions that (A5.13, ii) and (iii) hold, and the 'communication condition' that for any $f \in \Pi_{SD}$ the corresponding process has no two disjoint invariant sets, they have established the existence of a solution (ρ, h) to the ACOE by employing the vanishing discount approach and have shown that h satisfies (5.2). This work has been further extended by Federgruen, Schweitzer and Tijms [53]. They have dropped the

unichainedness assumption in [52]. Instead, they assume that any state can be reached from any other state via some policy. Under this and other conditions in [52] they have established the existence of a solution (ρ, h) to the ACOE, with h satisfying (5.2). They have deviated from the vanishing discount approach and have, instead, utilized Tychonoff's fixed point theorem in their analysis. We again note that in all these investigations, a restrictive growth condition on the cost function is imposed, as noted in Remark 5.7.

The Lyapunov stability condition necessarily imposes a blanket stability (i.e., positive recurrence) of certain states (cf. (5.13)) which may be very restrictive. On the other hand (5.2) is not easy to verify in general and, indeed, may not hold in the case of many queueing models [137]. Another generalization of the boundedness of the solution of the ACOE could be boundedness from below. This will be the case if the cost function has some 'monotone' properties, which naturally arise in various queueing models. This line of thought has been pursued in various ways in [24], [28], [72], [74], [75], [137], [138], [151], [152], [168], [186].

Sennott [151], [152] has prescribed very general conditions in this direction. We will now briefly describe them. Consider the following assumptions:

(A5.14) For every $i \in S$ and every $\beta \in (0,1)$, $J^*_\beta(i) < \infty$.

(A5.15) There exists a nonnegative integer L such that

$$h_\beta(i) := J^*_\beta(i) - J^*_\beta(0) \geq -L\,.$$

(A5.16) There exists a function $M : S \to \mathbb{R}_+$ such that $h_\beta(i) \leq M(i)$ for all $i \in S$ and any $\beta \in (0,1)$. For every $i \in S$ there exists an $a(i) \in U(i)$ such that

$$\sum_j P(j \mid i, a(i))M(j) < \infty\,.$$

Theorem 5.9. *Under (A5.14)–(A5.16), there exists an AC-optimal $f \in \Pi_{SD}$.*

Proof. Let $\beta_n \in (0,1)$ be such that $\beta_n \uparrow 1$. Let f_{β_n} be β_n-discount optimal. Let f be a limit point of f_{β_n} as $n \to \infty$. In order to simplify the notation all subsequences of β_n will also be denoted by β_n. By (A5.16), and a diagonal argument, there exists a function $h : S \to \mathbb{R}$ such that $\lim_{n\to\infty} h_{\beta_n}(\cdot) = h(\cdot)$. By (A5.15), $h(\cdot) \geq -L$. Let $\rho : S \to \mathbb{R}_+$ be a function such that $\lim_{n\to\infty}(1-\beta_n)J^*_{\beta_n}(i) = \rho(i)$. Using (A5.16) it is easy to see that $\rho(i) = \rho^*$, a constant. Now for $i \in S$,

$$(1-\beta_n)J^*_{\beta_n}(0) + h_{\beta_n}(i) = c\big(i, f_{\beta_n}(i)\big) + \beta_n \sum_{j\in S} P\big(j \mid i, f_{\beta_n}(i)\big)h_{\beta_n}(j)\,. \tag{5.15}$$

Fix an $i \in S$. Adding L to both sides to make $(h_{\beta_n}(i) + L) \geq 0$ and taking 'lim inf' on both sides of (5.15). Then by Fatou's lemma and the assumption of continuity of $P(j \mid i, \cdot)$ we conclude that

$$\rho^* + h(i) \geq c\big(i, f(i)\big) + \sum_j P\big(j \mid i, f(i)\big)h(j)\,.$$

Since $h(\cdot)$ is bounded below the proof of Theorem 5.1 can be modified to show that $J(i,f) \leq \rho^*$. By Theorem A.2, $J(i,\pi) \geq \rho^*$ for any $\pi \in \Pi$. Hence, $J(i,f) = J^*(i) = \rho^*$ and f is AC-optimal. □

Remark 5.8.

(a) From the above proof it is clear that if ρ is a scalar and $h : \boldsymbol{S} \to \mathbf{R}$ is bounded below and

$$\rho + h(i) \geq \min_{a \in U(i)} \left\{ c(i,a) + \sum_j P(j \mid i,a) h(j) \right\}, \tag{5.16}$$

then ρ is the optimal average cost, and any $f \in \Pi_{SD}$ selecting the minimum on the right-hand side of (5.16) is AC-optimal. In this case, we may replace the ACOE by an *average cost optimality inequality* (ACOI) viz. (5.16).

(b) If, for each $i \in \boldsymbol{S}$, $U(i)$ is finite, then in the above proof $f_{\beta_n}(i) = f(i)$ for large n. Then we can write for large n

$$\rho + h(i) = c\big(i, f(i)\big) + \beta_n \sum_j P\big(j \mid i, f(i)\big) h_{\beta_n}(j).$$

By Fatou's lemma

$$\rho + h(i) \geq c\big(i, f(i)\big) + \sum_j P\big(j \mid i, f(i)\big) h(j).$$

Consider the stronger assumption

(A5.17) Condition (A5.16) holds and $\sum_j P(j \mid i,a) M(j) < \infty$, for all $a \in \boldsymbol{A}$ and $i \in \boldsymbol{S}$.

Under (A5.17), using dominated convergence, it is easy to see that

$$\rho + h(i) = \min_{a \in U(i)} \left\{ c(i,a) + \sum_j P(j \mid i,a) h(j) \right\}$$

and we obtain the ACOE. If for each $i \in \boldsymbol{S}$, there is a finite set $R_i \subset \boldsymbol{S}$ such that $P(j \mid i, \cdot) = 0$ for $j \notin R_i$, then (A5.17) will obviously hold. Such a condition is satisfied for systems whose dynamics have a nearest-neighbour motion property [28].

(c) If there exists an $f \in \Pi_{SD}$, under which the process is ergodic, irreducible with an invariant measure $\eta(f) \in \mathcal{P}(\boldsymbol{S})$, and $\sum_i c\big(i, f(i)\big) \eta(f)(i) < \infty$, then (A5.14) and (A5.16) hold. (A5.15) holds if $J^*_\beta(i)$ is increasing in i. Direct conditions implying (A5.14)–(A5.17) can be found in [28], [32], [34], [74], [75], [151], [152], [168], [186]. See also [137], [138].

(d) Let $f \in \Pi_{SD}$ be a policy which attains the minimum on the right-hand side of (5.16). Fix an $i \in \boldsymbol{S}$. If the chain under f is positive recurrent at i, then one can show that equality holds at i in (5.16). However, the lack of positive recurrence at i may lead to strict inequality in (5.16). Cavazos-Cadena [33] had exhibited an example to demonstrate this. He has further shown in his example [33] that (A5.14)–(A5.16) are satisfied, but the ACOE does not admit any solution.

5.3. The convex analytic approach. We will now describe Borkar's convex analytic approach for the average cost case [20]–[26]. The convex analytic approach to the AC-problem is a natural extension of the linear programming approach when the state/action spaces are no longer finite. In this approach one views the control problem as the problem of minimizing a linear functional on the convex set of 'ergodic occupation measures' to be defined shortly [20]–[26]. This approach can also be used to treat other standard cost criteria, but it may be more involved for treating cases such as the DC criterion. On the other hand, it is more flexible and powerful for certain other purposes, e.g. pathwise average cost, constrained optimization problem, etc. Since the techniques involved here are entirely different from what we have thus far followed, we will embark on a more detailed discussion.

By replacing each $U(i)$ with $\prod_k U(k)$ and $P(j \mid i, \cdot)$ by its composition with the projection $\prod_k U(k) \to U(i)$, we may and will assume that the $U(i)$'s are replicas of a fixed compact metric space $\boldsymbol{A}$. We say that an $f \in \Pi_{SR}$ is *stable* if the corresponding process is positive recurrent. We will assume that under an $f \in \Pi_{SR}$ the process has $\boldsymbol{S}$ as its single communicating class. (This can be relaxed in some cases; see [26] for a discussion on this.) Therefore, f will have a unique invariant measure $\eta(f) \in \mathcal{P}(\boldsymbol{S})$ satisfying

$$\eta(f)P(f) = \eta(f)\,.$$

Let Π_{SSR} denote the space of stable stationary policies. Π_{SSD} is defined analogously. For an $f \in \Pi_{SSR}$ denote by $\hat{\eta}(f) \in \mathcal{P}(\boldsymbol{S} \times \boldsymbol{A})$ the 'ergodic occupation measure' defined by

$$\int_{\boldsymbol{S}\times\boldsymbol{A}} g\, d\hat{\eta}(f) = \sum_{i\in\boldsymbol{S}} \eta(f)(i) \int_{\boldsymbol{A}} g(i,a) f(i)(da)$$

for $g \in C_b(\boldsymbol{S} \times \boldsymbol{A})$. We will consider the sample path average cost optimality which is stronger than the usual AC-optimality. Let

$$I_R = \{\hat{\eta}(f) \,:\, f \in \Pi_{SSR}\}, \qquad I_D = \{\hat{\eta}(f) \,:\, f \in \Pi_{SSD}\}.$$

Note that $\hat{\eta}(f)$ can only be defined for an $f \in \Pi_{SSR}$. To consider optimality in Π we will need to consider the following empirical processes. Let $\pi \in \Pi$, and let (X_t, A_t) be the corresponding processes with initial law $\mu \in \mathcal{P}(\boldsymbol{S})$. Define the $\mathcal{P}(\boldsymbol{S} \times \boldsymbol{A})$-valued empirical process $\{\nu_t\}_{t\geq 1}$ by

$$\nu_t(C \times D) = \frac{1}{t} \sum_{s=0}^{t-1} I\{X_s \in C, A_s \in D\}, \qquad t \geq 1\,, \tag{5.17}$$

for C, D Borel in $\boldsymbol{S}$, $\boldsymbol{A}$, respectively. Let $\overline{\boldsymbol{S}} = \boldsymbol{S} \cup \{\infty\}$ be the one point compactification of $\boldsymbol{S}$. By abuse of notation, we may identify ν_t with the element of $\mathcal{P}(\overline{\boldsymbol{S}} \times \boldsymbol{A})$ that restricts to it on $\boldsymbol{S} \times \boldsymbol{A}$. Since $\mathcal{P}(\overline{\boldsymbol{S}} \times \boldsymbol{A})$ is compact, $\{\nu_t\}$, viewed as a sequence of $\mathcal{P}(\overline{\boldsymbol{S}} \times \boldsymbol{A})$-valued random variables, converges to a sample path dependent compact limit set in $\mathcal{P}(\overline{\boldsymbol{S}} \times \boldsymbol{A})$. We characterize this set in Lemma 5.1 below, the statement of which calls for some new notation. Note that any element $\nu \in \mathcal{P}(\overline{\boldsymbol{S}} \times \boldsymbol{A})$ can be decomposed as

$$\nu(B) = \delta_\nu \nu'\big(B \cap (\boldsymbol{S} \times \boldsymbol{A})\big) + (1 - \delta_\nu)\nu''\big(B \cap (\{\infty\} \times \boldsymbol{A})\big) \tag{5.18}$$

for B Borel in $\overline{S} \times A$, $\delta_\nu \in [0,1]$ is uniquely specified and $\nu' \in \mathcal{P}(S \times A)$ (respectively, $\nu'' \in \mathcal{P}(\{\infty\} \times A)$) is uniquely specified if $\delta_\nu > 0$ (respectively, $\delta_\nu < 1$). We may render ν', ν'' unique at all times by imposing an arbitrary fixed choice thereof when $\delta_\nu = 0$, respectively, 1.

Lemma 5.1. *Outside a set of zero probability (with respect to $\mathcal{P}_\mu^\pi$), the following holds: For any limit point ν of $\{\nu_t\}$ in $\mathcal{P}(\overline{S} \times A)$ for which $\delta_\nu > 0$,*

$$\nu' = \hat{\eta}(f)$$

for some $f \in \Pi_{SSR}$.

Proof. By the martingale stability theorem [114, p. 53]

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \Big[I\{X_s = i\} - E_\mu^\pi \big[I\{X_s = i\} \mid \mathfrak{F}_{s-1} \big] \Big] & \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \Big[I\{X_s = i\} - \sum_{j \in S} P(i \mid j, A_{s-1}) I\{X_{s-1} = j\} \Big] \\
&= \lim_{t\to\infty} \Big[\nu_t(\{i\} \times A) - \int P(i \mid \cdot, \cdot)\, d\nu_t \Big] \\
&= 0, \quad a.s.,
\end{aligned}$$

for each $i \in S$. Consider a sample path outside the set of zero probability on which the above fails for any $i \in S$. Then for any ν as in the statement of the lemma, we must have

$$\nu'(\{i\} \times A) \geq \int P(i \mid \cdot, \cdot)\, d\nu', \qquad i \in S.$$

Note that an inequality is obtained here, since the second term on the right hand side of (5.18) is obviously nonnegative. Summing over $i \in S$ on both sides, it follows that equality must hold. Decomposing ν' as $\nu'(i, da) = \overline{\nu}(i) f(i)(da)$, where $\overline{\nu} \in \mathcal{P}(S)$ is the marginal on S and $i \longmapsto f(i) \in \mathcal{P}(A)$ is a version of the regular conditional law which defines an element of Π_{SR}, we obtain

$$\overline{\nu}(i) = \sum_{j \in S} \overline{\nu}(j) P(f)(i, j).$$

Hence, $\overline{\nu} = \eta(f)$ and the conclusion follows. □

Lemma 5.2. *The sets I_R and I_D are closed; also, I_R is convex and has its extreme points in I_D.*

Proof. Let $\hat{\eta}(f_n) \in I_R$ and $\hat{\eta}(f_n) \to \nu$ for some ν in $\mathcal{P}(S \times A)$. Then for all $i \in S$

$$\hat{\eta}(f_n)(\{i\} \times A) = \int P(i \mid \cdot, \cdot)\, d\hat{\eta}(f_n), \qquad n \geq 1.$$

Letting $n \to \infty$

$$\nu(\{i\} \times A) = \int P(i \mid \cdot, \cdot)\, d\nu.$$

Now argue as in the proof of the preceding lemma to conclude that $\nu = \hat{\eta}(f)$ for some $f \in \Pi_{SSR}$. This proves that I_R is closed. The proof that I_D is closed is similar. Let $f_1, f_2 \in \Pi_{SSR}$ and $0 \leq \lambda \leq 1$. Define $f \in \Pi_{SSR}$ as follows:

$$f(i) = \frac{\lambda\eta(f_1)(i)f_1(i) + (1-\lambda)\eta(f_2)(i)f_2(i)}{\lambda\eta(f_1)(i) + (1-\lambda)\eta(f_2)(i)}.$$

Then using the properties of invariant measures, it is not difficult to see that

$$\eta(f) = \lambda\eta(f_1) + (1-\lambda)\eta(f_2)$$
$$\hat{\eta}(f) = \lambda\hat{\eta}(f_1) + (1-\lambda)\hat{\eta}(f_2),$$

showing that I_R is convex. Now let $f \in \Pi_{SSR}$ be such that for some $i_0 \in \boldsymbol{S}$ and $0 < \lambda < 1$, there exist $\phi_1, \phi_2 \in \boldsymbol{\mathcal{P}}(\boldsymbol{A})$ such that

$$\int P(\cdot \mid i_0, a)f(i_0)(da) = \lambda \int P(\cdot \mid i_0, a)\phi_1(da) + (1-\lambda)\int P(\cdot \mid i_0, a)\phi_2(da)$$

$$\int P(\cdot \mid i_0, a)\phi_1(da) \neq \int P(\cdot \mid i_0, a)\phi_2(da)$$

Define $f_1, f_2 \in \Pi_{SR}$ as

$$f_i(j) = \begin{cases} f(j), & j \neq i_0 \\ \phi_i, & j = i_0 \end{cases}$$

Then it can be shown [24] that $f_1, f_2 \in \Pi_{SSR}$, and any two of $\eta(f)$, $\eta(f_1)$, $\eta(f_2)$ are distinct from each other. Let $b \in (0,1)$ be such that

$$\lambda = b\eta(f_1)(i_0) \,/\, \big(b\eta(f_1)(i_0) + (1-b)\eta(f_2)(i_0)\big).$$

Then we can argue as before to conclude that

$$\hat{\eta}(f) = b\hat{\eta}(f_1) + (1-b)\hat{\eta}(f_2).$$

Therefore $\hat{\eta}(f)$ is not an extreme point of I_R. This implies that, for $\hat{\eta}(f')$ to be an extreme point of I_R, $P(\cdot \mid i, a)$ must be constant over $a \in \operatorname{supp}\big(f'(i)\big)$, for each $i \in \boldsymbol{S}$. Hence, $P(f'') = P(f')$, for all $f'' \in \Pi_{SSR}$ such that $\operatorname{supp}\big(f''(i)\big) \subset \operatorname{supp}\big(f'(i)\big)$, for each $i \in \boldsymbol{S}$. In this case, $\eta(f'') = \eta(f')$. Suppose that for some i, say $i = 1$, there exist $\alpha \in (0,1)$ and $\phi_1', \phi_2' \in \boldsymbol{\mathcal{P}}(\boldsymbol{A})$, $\phi_1' \neq \phi_2'$, such that $f'(1) = \alpha\phi_1' + (1-\alpha)\phi_2'$. Define $f_1', f_2' \in \Pi_{SSR}$ by

$$f_k' = \begin{cases} \phi_k' & \text{if } i = 1 \\ f'(i) & \text{if } i \neq 1 \end{cases} \qquad k = 1, 2.$$

It follows that $\eta(f') = \eta(f_1') = \eta(f_2')$. It is also easy to check that

$$\hat{\eta}(f') = \alpha\hat{\eta}(f_1') + (1-\alpha)\hat{\eta}(f_2'),$$
$$\hat{\eta}(f_1') \neq \hat{\eta}(f_2'),$$

which contradicts the extremality of $\hat{\eta}(f')$. Hence, $f'(1)$ must be a Dirac measure. Applying this argument to each $i \in S$, we deduce that $f \in \Pi_{SD}$. From this it follows that the extreme points of I_R lie in I_D. $\square$

We now proceed to show the existence of a sample path average cost optimal $f \in \Pi_{SSD}$. It is clear that a blanket stability condition or some condition on the cost that penalizes unstable behavior is required to give the desired existence. For example, consider the case $c(i,a) = \exp(-i)$ which rewards unstable behavior. Then the cost for any $f \in \Pi_{SSD}$ is a.s. positive while the cost for an unstable $f \in \Pi_{SD}$ is a.s. zero, making the latter optimal. We want to rule out this possibility, as stability is a very desirable property of a policy. We seek to find conditions under which our goal will be achieved. Let $f \in \Pi_{SSR}$. Define

$$\rho(f) := \int c \, d\hat{\eta}(f), \qquad \rho^* := \inf_{f \in \Pi_{SSR}} \rho(f).$$

Note that under $f \in \Pi_{SSR}$, $J(i,f) = \rho(f)$ for each $i \in S$. We consider two sets of hypotheses:

(A5.18) The Near-Monotonicity Condition:

$$\liminf_{i \to \infty} \min_{a \in A} c(i,a) > \rho^*.$$

Intuitively (A5.18) penalizes the drift of the process away from some finite set, requiring the optimal policy to exert some kind of a 'centripetal force' pushing the process back towards this finite set. Thus, the optimal policy gains the desired stability property. If $c(i,a) = k(i)$ for some $k : S \to \mathbb{R}_+$ and $k(i)$ is increasing, then this condition will automatically be satisfied. Such penalizing conditions quite often occur in queueing applications (see [20], [151], [152], [168], [186]).

(A5.19) Stability Condition (cf. (A5.7)–(A5.12)): $\Pi_{SR} = \Pi_{SSR}$ and I_R is compact.

(A5.19′) Equivalent conditions to (A5.19) are:

(i) $\Pi_{SD} = \Pi_{SSD}$ and I_D is compact.

(ii) The mean return times to a prescribed state (say 0) are uniformly integrable over all $f \in \Pi_{SR}$.

(iii) Same as (ii) but with Π_{SD} replacing Π_{SR}.

Theorem 5.10. *Under (A5.18) or (A5.19) there exists an $f \in \Pi_{SSD}$ which is sample path average cost optimal in Π_{SR}.*

Proof. From Lemma 5.2 it can be shown by an application of Choquet's theorem [25], [26], that if $\nu \longmapsto \int c \, d\nu$ attains its minimum on I_R, it will do so for an $f \in \Pi_{SD}$. Under (A5.19) it can be shown that $f \longmapsto \hat{\eta}(f)$ is continuous. Therefore, the desired result follows under (A5.19). We next consider the case under (A5.18). Let $f_n \in \Pi_{SR}$ be such that $\rho(f_n) \downarrow \rho^*$. By identifying $\hat{\eta}(f_n)$ with the element of $\mathcal{P}(\overline{S} \times A)$ that restricts to it on $S \times A$ for each n and then dropping to a subsequence if necessary, we may assume that $\hat{\eta}(f_n) \to \nu$ in $\mathcal{P}(\overline{S} \times A)$ for some ν. Let $n \to \infty$ in the equation

$$\hat{\eta}(f_n)(\{j\} \times A) = \int P(j \mid \cdot, \cdot) \, d\hat{\eta}(f_n), \qquad j \in S$$

and argue as in Lemma 5.1 to conclude that for ν' as in (5.18), $\delta_\nu > 0$ implies

$$\nu'(\{j\} \times A) = \int P(j \mid \cdot, \cdot)\, d\nu', \qquad j \in S.$$

Decomposing ν' as $\nu'(i, da) = \overline{\nu}(i) f(i)(da)$, $i \in S$ we have $\overline{\nu} = \eta(f)$ and therefore, $\nu' = \hat{\eta}(f)$. Let $c_m = c \wedge m$ for $m \geq 1$ and pick $\varepsilon > 0$ such that (A5.18) continues to hold with $\rho^* + \varepsilon$ in place of ρ^*. Then

$$\begin{aligned} \rho^* &= \lim_{n \to \infty} \int c\, d\hat{\eta}(f_n) \\ &\geq \lim_{n \to \infty} \int c_m\, d\hat{\eta}(f_n) \\ &\geq \delta_\nu \int c_m\, d\hat{\eta}(f) + (1 - \delta_\nu)\big((\rho^* + \varepsilon) \wedge m\big). \end{aligned}$$

Letting $m \to \infty$,

$$\rho^* \geq \delta_\nu \rho^* + (1 - \delta_\nu)(\rho^* + \varepsilon).$$

This is possible only if $\delta_\nu = 1$ and $\int c\, d\hat{\eta}(f) = \rho^*$. □

The above theorem, however, does not ensure optimality of the cost-minimizing policy in I_R with respect to arbitrary policies. For the near-monotone case this can be resolved without any further assumptions, but for the stable case, we need the following.

(A5.20) If $\tau = \min\{t \geq 1 : X_t = 0\}$, then

$$\sup_{\pi \in \Pi} E_0^\pi [\tau^2] < \infty.$$

Remark 5.9. (A5.20) clearly implies (A5.19). The converse need not be true, as can be shown by an explicit example [24]. Some sufficient conditions for (A5.20) are:

(i) A Lyapunov condition [28], we will describe shortly (cf. Theorem 5.11),
(ii) the strong uniform recurrence condition of Doeblin and its variants [174], and
(iii) the condition that there exist an $N < \infty$ for which

$$\sup_{\pi \in \Pi} \sup_i \mathcal{P}_i^\pi (\tau \geq N) < 1$$

where τ is as above.

Theorem 5.11. *Under (A5.18) or (A5.20) there exists an $f \in \Pi_{SD}$ which is sample path average cost optimal.*

Proof. Under (A5.20) it can be shown [26] that the processes ν_t as defined in (5.17) are tight over Π. Therefore, δ_ν as in the statement of Lemma 5.1 may be taken to be one. This resolves the case under (A5.20). Under (A5.18), let ν be a limit point of $\{\nu_t\}$ in $\mathcal{P}(\overline{S} \times A)$ along some subsequence. Then as in the proof of Theorem 5.9 it can be shown that

$$\liminf_{t \to \infty} \int c\, d\nu_t \geq \rho^*. \tag{5.19}$$

Since this is true for any limit point ν of $\{\nu_t\}$ in $\mathcal{P}(\overline{S} \times A)$ and for all sample points outside a set of probability zero, the desired result follows in this case also. □

Remark 5.10. Some open problems arising in this context are:

(i) Can (A5.20) be replaced by (A5.19) while retaining the desired optimality?

(ii) If $\Pi_{SR} = \Pi_{SSR}$, will (A5.19) hold automatically?

Remark 5.11. The condition in (5.19) implies a much stronger optimality which will be discussed in Section 6.

Now after the existence result of Theorem 5.11, an alternative treatment of the ACOE is possible. We will present a brief description without proofs. For details, see [24], [26], [28]. Define $h : S \to \mathbb{R}$ by

$$h(i) = E_i^{f_0}\left[\sum_{t=0}^{\tau-1}\Big(c\big(X_t, f_0(X_t)\big) - \rho^*\Big)\right], \qquad i \in S, \tag{5.20}$$

where $\tau = \min\{t \geq 1 : X_t = 0\}$ and $f_0 \in \Pi_{SD}$ is any sample path average cost optimal policy. In [22], [24], it is shown that $\big(h(\cdot), \rho^*\big)$ satisfies the ACOE under the following additional hypothesis called stability under local perturbations.

(A5.21) Given an $f \in \Pi_{SSD}$ with $\rho(f) < \infty$, any $f' \in \Pi_{SD}$ obtained from f by changing the actions at most finitely many states is also stable and $\rho(f') < \infty$.

A sufficient, though not necessary, condition for (A5.21) to hold is that every state has at most finitely many neighbors, i.e., for each $i \in S$, there is a finite set $R_i \subset S$ such that $P(j \mid i, \cdot) = 0$ for $j \notin R_i$.

In many cases, the solution (ρ^*, h) of the ACOE can be characterized (Theorem 5.12 below). The usual characterization of AC-optimal $f \in \Pi_{SD}$ in terms of the ACOE can also be proved for the foregoing.

Theorem 5.12. *Assume (A5.18) and let f_0, h be defined as above (cf. (5.20)). Let*

$$H = \big\{(\rho, w) : (\rho, w) \text{ satisfies the ACOE, } w(0) = 0, \inf w(\cdot) > -\infty\big\}.$$

Then (ρ^*, h) is the unique element of H corresponding to the minimum value of ρ (i.e., if (ρ', w') is another element of H, then $\rho' \geq \rho^*$ with equality if and only if $w' = h$).
Now, instead of (A5.18), suppose c is bounded and the following Lyapunov condition holds: There exists an $w : S \to \mathbb{R}_+$, a finite $A \subset S$ and an $\varepsilon > 0$ such that:

(a) $0 \in A$ *and the set* $\big\{i \in A^c : P(j \mid i, a) > 0, \text{ for some } j \in A,\ a \in A\big\}$ *is finite.*

(b) $\lim_{i \to \infty} w(i) = \infty$.

(c) *Under any* $\pi \in \Pi$, $\mu \in \mathcal{P}(S)$

$$E_\mu^\pi\big[(w(X_{t+1}) - w(X_t) + \varepsilon) I\{X_t \notin A\} \mid \mathfrak{F}_t\big] \leq 0, \quad a.s.$$

(d) *There exists a random variable Z and a scalar $\lambda > 0$ such that $E[\exp(\lambda Z)] < \infty$ and for all $b \geq 0$*

$$\mathcal{P}_\mu^\pi\big(|w(X_{t+1}) - w(X_t)| > b \mid \mathfrak{F}_t\big) \leq P(Z > b).$$

Then (ρ^*, h) *is the unique solution of the ACOE in the class* $\{(\rho, w) : w(0) = 0, \limsup_{i\to\infty} \frac{h(i)}{w(i)} < \infty\}$.

Remark 5.12. An alternative 'intrinsic' formulation of the ACOE is also possible. For any $f \in \Pi_{SSD}$, define $h_f : \boldsymbol{S} \to \mathbf{R}$ by

$$h_f(i) = E_i^f \left[\sum_{t=0}^{\tau-1} \Big(c\big(X_t, f(X_t)\big) - \rho(f) \Big) \right], \qquad i \in \boldsymbol{S}.$$

We say that f is *locally AC-optimal* if it yields a lower cost than any other element of Π_{SD} obtainable from f by changing f in at most finitely many states. In addition to the foregoing hypotheses, assume that every locally AC-optimal f is AC-optimal (for bounded c, a sufficient condition for this is that $\Pi_{SD} = \Pi_{SSD}$ and $\{\eta(f) : f \in \Pi_{SSD}\}$ is tight). One then has that f is sample path average cost optimal if and only if, for $i \in \boldsymbol{S}$

$$h_f(i) = \inf_a \Big\{ \sum_j P(j \mid i, a) h_f(j) + c(i, a) - \rho(f) \Big\}.$$

This statement is 'intrinsic' in the sense that all quantities (i.e., $h_f, \rho(f)$) are computable in terms of f. An interesting open problem is to characterize the most general conditions under which local AC-optimality implies AC-optimality.

Remark 5.13. The Lyapunov condition in Theorem 5.12(ii) implies (A5.20) and has many other implications [26], but the condition (ii)(d) there is rather strong and due to this it may be difficult to construct such a function in a given situation. A partial answer to this question is given in [72]. It would be interesting to investigate if the Lyapunov conditions studied by [89], [53] (cf. (A5.13)), which do not involve (ii)(d) above, imply (A5.20).

§6. Borel State and Action Spaces

We consider in this section the case in which $\boldsymbol{S}$ and $\boldsymbol{A}$ are general Borel spaces. This is a natural setting for many problems, e.g. control of stock in water reservoirs, allocation of a resource between production and consumption, control of biological populations, harvesting a natural resource; see [17], [50], [80], for several examples, and references therein. Also, the equivalent formulation of POCMP in terms of the conditional distribution of the (unobservable) state leads to a problem with an uncountable Borel state space, as we are going to see in Section 7.

In this more general context, the ACOE is written as

$$\begin{aligned} \rho(x) + h(x) &= \inf_{a \in U(x)} \Big\{ c(x, a) + \int_{\boldsymbol{S}} h(y) P(dy \mid x, a) \Big\} \\ &= T(h)(x), \qquad x \in \boldsymbol{S}, \end{aligned} \tag{6.1}$$

where $\rho, h \in \mathcal{M}(\boldsymbol{S})$. As in Section 5, a pair of functions (ρ, h) as above is called a solution to the ACOE, and if ρ and h are bounded, we will say that the solution is bounded. Also, as in Theorem 5.1, our aim is to relate the AC problem to the existence of solutions to the ACOE. We have the following.

Theorem 6.1. *Suppose that (ρ, h) is a solution to the ACOE, and that for each policy $\pi \in \Pi_M$, the following holds*

$$\lim_{t\to\infty} E_x^\pi \left[\frac{h(X_t)}{t}\right] = 0\,, \qquad \forall\, x \in \boldsymbol{S}\,. \tag{6.2}$$

Then we have that

(i) *there holds*

$$\limsup_{n\to\infty} \frac{1}{n+1} E_x^\pi \left[\sum_{t=0}^{n} \rho(X_t)\right] \le J(x,\pi)\,, \tag{6.3}$$

and if $\pi \in \Pi_{SD}$ is such that $\pi(x)$ attains the infimum in (6.1), then equality is attained in (6.3);

(ii) *if $\rho(x) = \rho^* \in \mathbb{R}$, for all $x \in \boldsymbol{S}$, then $J^*(x) = \rho^*$, for all $x \in \boldsymbol{S}$, and any $\pi^* \in \Pi_{SD}$ such that $\pi^*(x)$ attains the infimum in (6.1) is average optimal.*

The proof of Theorem 6.1 follows that of Theorem 5.1 and is essentially contained in [173], and more explicitly in [78], [144]; see also [76, pp. 66–68], [80, pp. 53–55], [146, pp. 93–94]. Note that (i) above says that if $\rho(\cdot)$ is taken as the cost function to define another CMP $(\boldsymbol{S}, \boldsymbol{A}, U, P, \rho)$ then, for any $\pi \in \Pi_M$, the average cost incurred under the cost function $\rho(\cdot)$ does not exceed that under cost function $c(\cdot,\cdot)$.

Given the results above, it is of interest to find conditions under which there exists a solution (ρ, h) to the ACOE, satisfying (6.2). If h is bounded, then (6.2) is satisfied trivially. Also, if the random variables $\{h(X_t)\}$ are uniformly integrable under $\mathcal{P}_x^\pi$, for $\pi \in \Pi_M$ and $x \in \boldsymbol{S}$, then there exists a constant $0 < K_x^\pi < \infty$ such that $E_x^\pi\big[|h(X_t)|\big] \le K_x^\pi$. Hence, if such a uniform integrability condition holds under *every* $\pi \in \Pi_M$ and $x \in \boldsymbol{S}$, then (6.2) is also satisfied trivially. The latter approach has been used by Shwartz and Makowski, for some queueing problems [162], [163], [164].

6.1. Bounded costs. We first assume that $c(\cdot,\cdot)$ is bounded. When there are bounded solutions (ρ, h) to the ACOE, then much stronger results than those in Theorem 6.1 (i) can be obtained. To state these, some definitions are needed.

Let R and H be bounded, measurable real-valued functions on $\boldsymbol{S}$, i.e., $R, H \in \mathcal{M}_b(\boldsymbol{S})$, and let $\pi^* \in \Pi$. Following the terminology of Yushkevich and Dynkin [50], the triplet (R, H, π^*) is said to be *canonical* if

$$J_N(x,\pi^*,H) = J_N^*(x,H) = H(x) + NR(x)\,, \qquad \forall\, N \in \mathbb{N}_0\,, \quad x \in \boldsymbol{S}\,, \tag{6.4}$$

and $\pi^* \in \Pi$ is said to be a *canonical policy* if it is an element of some canonical triplet. Note that if (R, H, π^*) is a canonical triplet, then π^* is N-stage optimal, *for all* $N \in \mathbb{N}_0$, when H is taken as the terminal cost. This concept was introduced by Yushkevich [200]. For finite models Denardo and Fox [36] used a similar approach.

A policy $\pi^* \in \Pi$ is said to be *strong average optimal* if

$$\limsup_{N\to\infty} \frac{1}{N} J_N(x,\pi^*) \le \liminf_{N\to\infty} \frac{1}{N} J_N(x,\pi)\,, \qquad \forall\, x \in \boldsymbol{S}\,, \quad \forall\, \pi \in \Pi\,. \tag{6.5}$$

Alternate definitions of strong average optimality are given in [67], [68]. Clearly, a strong average optimal policy π^* is also average optimal, and the limit of the sequence $\{\frac{1}{N}J_N(x,\pi^*)\}$, as $N \to \infty$, exists. An interpretation of (6.5) is that the 'most pessimistic' average performance under π^*, is no worse than the most 'optimistic' performance under any other policy. We have the following.

Theorem 6.2. *Let $\pi^* \in \Pi_{SD}$, let $\rho, h \in \mathcal{M}_b(S)$ and $c \in \mathcal{M}_b(K)$. Then (ρ, h, π^*) is a canonical triplet if and only if*

$$\rho(x) = \inf_{a \in U(x)} \left\{ \int_S \rho(y) P(dy \mid x, a) \right\} \tag{6.6}$$

and

$$\rho(x) + h(x) = \inf_{a \in U(x)} \left\{ c(x,a) + \int_S h(y) P(dy \mid x, a) \right\} \tag{6.7}$$

and $\pi^*(x)$ attains the infimum in both (6.6) and (6.7), for all $x \in S$.

Proof. (Necessity) Let (ρ, h, π^*) be a canonical triplet. Then, by (6.4)

$$\begin{aligned} h(x) + \rho(x) + N\rho(x) &= J^*_{N+1}(x,h) \\ &= T\big(J^*_N\big)(x) \\ &= c\big(x, \pi^*(x)\big) + \int_S J^*_N(y,h) P\big(dy \mid x, \pi^*(x)\big)\,. \end{aligned} \tag{6.8}$$

Since $J_0(x, \pi^*, h) = J^*_0(x,h) = h(x)$, then (6.7) follows from (6.8), by letting $N = 0$. Furthermore, since $\rho(\cdot)$, $h(\cdot)$, and $c(\cdot,\cdot)$ are bounded, then dividing both sides of (6.8) by N and letting $N \to \infty$, yields (6.6).
(Sufficiency) Let (ρ, h) satisfy (6.6) and (6.7), and let $\pi^*(x)$ attain the infimum in these expressions. We use induction to show that (ρ, h, π^*) is a canonical triplet. For $N = 0$, this is trivially satisfied. Suppose $N \in \mathbb{N}_0$ is the first integer for which (6.4) fails, then

$$\begin{aligned} J^*_N(x,h) &= T(J^*_{N-1})(x) \\ &= T(h + (N-1)\rho)(x) \\ &= \inf_{a \in U(x)} \left\{ c(x,a) + \int_S h(y) P(dy \mid x, a) + (N-1) \int_S \rho(y) P(dy \mid x, a) \right\} \\ &\geq T(h)(x) + (N-1) \inf_{a \in U(x)} \left\{ \int_S \rho(y) P(dy \mid x, a) \right\} \\ &= T(h)(x) + (N-1)\rho(x) = h(x) + N\rho(x). \end{aligned}$$

On the other hand,

$$\begin{aligned} J^*_N(x,h) &\leq J_N(x, \pi^*, h) \\ &= c\big(x, \pi^*(x)\big) + \int_S J^*_{N-1}(y, \pi^*, h) P\big(dy \mid x, \pi^*(x)\big) \\ &= c\big(x, \pi^*(x)\big) + \int_S \big[h(y) + (N-1)\rho(y)\big] P\big(dy \mid x, \pi^*(x)\big) \\ &= T(h)(x) + (N-1)\rho(x) = h(x) + N\rho(x) \end{aligned}$$

contradicting our hypothesis. Therefore, (ρ, h, π^*) is a canonical triplet. □

The results in Theorem 6.2 were obtained by Yushkevich [200]; see also [50]. Note that (6.7) is the ACOE, and (6.6) allows $\rho(\cdot)$ to be treated as a constant, with respect to the optimization problem. Of course, if $\rho(x) = \rho^*$, for all $x \in S$, then (6.6) is satisfied trivially. The coupled equations (6.6) and (6.7) were apparently introduced by Howard [92, pp. 61–62], in the context of finite state CMP for which, under some policies, $\{X_t\}$ has several ergodic classes, i.e., the so-called multi-chain case. In this case, different ergodic classes may have different optimal average cost, and $\rho(\cdot)$ gives this cost, as will be shown.

From Theorem 6.2 we see that the canonical policy π^* is a measurable selector for both (6.6) and (6.7). However, condition (A2.2) in Section 2 is not enough to guarantee the existence of selectors in either (6.6) or (6.7), since ρ and h are assumed to be bounded and measurable functions, but not necessarily lower semicontinuous. For this situation, the following condition is needed.

(A6.1) The transition kernel $P(\cdot \mid x, a)$ is *strongly continuous* in (x, a); that is, $u \in \mathcal{M}_b(S)$ implies $\int_S u(y)P(dy \mid \cdot, \cdot) \in C_b(K)$.

It follows that under conditions (A2.1), (A2.3) and (A6.1), measurable selectors exist for each of (6.6) and (6.7), and $\pi^* \in \Pi_{SD}$ will be a canonical policy if and only if it is a selector for both (6.6) and (6.7). If (ρ, h, π^*) is a canonical triplet, then (ρ, h) solves the ACOE, and (6.2) is satisfied, since h is bounded. Consequently, the results of Theorem 6.1 follow. The next result presents other important implications.

Theorem 6.3. *Let (ρ, h, π^*) be a canonical triplet, and let $c \in \mathcal{M}_b(K)$. Then, for each $x \in S$,*

(i) $J_N(x, \pi^*) \leq J_N(x, \pi) + \operatorname{span}(h)$, *for every* $\pi \in \Pi$;

(ii) π^* *is strong average optimal;*

(iii) $J(x, \pi^*) = J^*(x) = \rho(x)$;

(iv) $h^-(x) + \frac{\rho(x)}{1-\beta} \leq J^*_\beta(x) \leq h^+(x) + \frac{\rho(x)}{1-\beta}$;

(v) *If* $\rho(x) = \rho^* \in \mathbb{R}$, *for all* $x \in S$, *then for every* $\pi \in \Pi$

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t) \geq \rho^*, \qquad \mathcal{P}^\pi_x\text{–a.s.},$$

when $X_0 = x$, *and* $\{A_t\}$ *is generated using the policy* π. *Furthermore*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t) = \rho^*, \qquad \mathcal{P}^\pi_x\text{–a.s.},$$

if and only if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} \Phi(X_t, A_t) = 0\,, \qquad \mathcal{P}^\pi_x\text{–a.s.},$$

where $\Phi : K \to \mathbb{R}$ *is given by*

$$\Phi(x, a) := c(x, a) + \int h(y)P(dy \mid x, a) - \rho^* - h(x)\,.$$

(vi) π^* *is sample path average cost optimal.*

Proof. To prove (i), note that for all $\pi \in \Pi$

$$J_N(x,\pi^*,h) = E_x^{\pi^*}\left[\sum_{t=0}^{N-1} c(X_t,A_t) + h(X_N)\right]$$
$$\leq E_x^{\pi}\left[\sum_{t=0}^{N-1} c(X_t,A_t) + h(X_N)\right] = J_N(x,\pi,h)$$

Hence,

$$J_N(x,\pi^*) \leq J_N(x,\pi) + E_x^{\pi}[h(X_N)] - E_x^{\pi^*}[h(X_N)]$$
$$\leq J_N(x,\pi) + \operatorname{span}(h), \qquad \forall\, \pi \in \Pi.$$

By the boundedness of $h(\cdot)$, we have that

$$\lim_{N\to\infty} \frac{1}{N} J_N(x,\pi^*,h) = \lim_{N\to\infty}\left[\frac{h(x)+N\rho(x)}{N}\right] = \rho(x)\,.$$

Furthermore, since $J_N(x,\pi^*,h) = J_N(x,\pi^*) + E_x^{\pi^*}[h(X_N)]$, then

$$\rho(x) = \lim_{N\to\infty} \frac{1}{N} J_N(x,\pi^*)$$

and (ii)–(iii) follows from (i).

Next, since (ρ,h) solve the ACOE, then (ρ,h^-) and (ρ,h^+) are also solutions to the ACOE. Since $h^-(\cdot) \leq 0 \leq h^+(\cdot)$, then by Lemma 2.1 we have that $T(h^-) \leq T(\beta h^-) = T_\beta(h^-)$, and $T(h^+) \geq T(\beta h^+) = T_\beta(h^+)$. Then, (iv) follows by induction, using Theorem 2.1 (iv); see [62].

Turning our attention to (v) and (vi), observe that, due to (6.7), $\Phi(x,a) \geq 0$, for all $(x,a) \in \boldsymbol{K}$. Also, by the (Markov) property (2.3) in Section 2, we have that, for any $\pi \in \Pi$,

$$\Phi(X_t,A_t) = E_x^{\pi}\Big[c(X_t,A_t) + h(X_{t+1}) - \rho^* - h(X_t) \;\Big|\; H_t, A_t\Big], \qquad \mathcal{P}_x^{\pi}\text{-a.s.}$$

Let

$$Z_t := c(X_t,A_t) + h(X_{t+1}) - h(X_t) - \rho^* - \Phi(X_t,A_t)\,,$$

and

$$M_N := \sum_{t=0}^{N-1} Z_t = \sum_{t=0}^{N-1} c(X_t,A_t) - N\rho^* + h(X_N) - h(X_0) - \sum_{t=0}^{N-1} \Phi(X_t,A_t)\,.$$

Note that $\{Z_t\}$ is a $(\mathfrak{S}_t, \mathcal{P}_x^{\pi})$ martingale difference, where $\mathfrak{S}_t := \sigma(H_{t+1}, A_{t+1})$. Since $\{Z_t\}$ is bounded uniformly in t, by the martingale stability theorem

$$\lim_{N\to\infty} \frac{M_N}{N} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} Z_t = 0\,, \qquad \mathcal{P}_x^{\pi}\text{-a.s.}$$

Therefore, by the boundedness of $h(\cdot)$,

$$\lim_{N\to\infty}\left[\frac{1}{N}\sum_{t=0}^{N-1}c(X_t,A_t)-\rho^*-\frac{1}{N}\sum_{t=0}^{N-1}\Phi(X_t,A_t)\right]=0\,,\qquad \mathcal{P}_x^\pi\text{-a.s.}$$

Finally, (v) and (vi) follow since $\Phi(x,a)\geq 0$, for all $(x,a)\in \boldsymbol{K}$, and since for a canonical policy π^*, $\Phi(X_t,A_t)=0$, $\mathcal{P}_x^\pi$-a.s. $\square$

The results in (i)–(iii) of Theorem 6.3 are essentially contained in [50, Chap. 7]; that in (iv) is motivated by similar results in [132] and [62]; (v) and (vi) are due to Georgin [70], see also [80, pp. 52–55]. Also, the function Φ defined in (v) was introduced by Mandl [120] and is often referred to as *Mandl's discrepancy function.*

In view of Theorem 6.3, it follows that a canonical triplet yields the desired results. We therefore look for conditions on the primary objects like the cost function c and transition kernel P which imply the existence of a canonical triplet, so that the theory can be used in a given practical situation. To this end, a standard procedure is to assume some ergodicity conditions which will ensure the existence of a bounded solution to the ACOE. We have already discussed several such conditions for the countable state case (cf. (A5.1)–(A5.5)). Analogues of such assumptions are also available in the literature, an extensive survey of which appears in [84]. We will focus on a particular ergodicity condition which not only subsumes many such conditions but also facilitates easily implementable numerical schemes. Our presentation here follows essentially that in [80, Chap. 3].

(A6.2) There exists a number $\alpha<1$ such that

$$\sup_{k,k'\in\boldsymbol{K}}\left\|P(\cdot\mid k)-P(\cdot\mid k')\right\|_{TV}\leq 2\alpha\,,$$

where $\|\cdot\|_{TV}$ denotes the total variation norm.

Example 6.1. Let $\boldsymbol{S}=\mathbb{R}$, $\boldsymbol{A}\subset\mathbb{R}$, a compact set. Consider the system

$$X_{t+1}=F(X_t,A_t)+G(X_t)W_t\,,\qquad X_0=x\,,$$

where $F:\mathbb{R}\times\boldsymbol{A}\to\mathbb{R}$, $G:\mathbb{R}\to\mathbb{R}$ are bounded, continuous and $G(\cdot)>0$, and $\{W_t\}$ is a sequence of independent $N(0,1)$ random variables ($N(a,b)$ stands for the Gaussian distribution with mean a and variance b). In this case the transition kernel is given by

$$P(\cdot\mid x,a)=N\big(F(x,a),G^2(x)\big)\,.$$

Using the assumed conditions on F,G one can show that (A6.2) holds. We omit the details. An important consequence of (A6.2) is given below; for a proof and further discussion see [80, Chap. 3].

Lemma 6.1. *Suppose (A6.2) holds. Then, for any $f \in \Pi_{SD}$, the corresponding process $\{X_t\}$ has a unique invariant measure $\eta(f) \in \mathcal{P}(S)$ satisfying*

$$\left\| P^t\big(\cdot \mid x, f(x)\big) - \eta(f)(\cdot) \right\|_{TV} \leq 2\alpha^t, \qquad t = 0, 1, \ldots, \tag{6.9}$$

where $P^t\big(\cdot \mid x, f(x)\big)$ denotes the t-step transition probability measure under f with $X_0 = x$.

Remark 6.1. (a) Lemma 6.1 also holds for any $f \in \Pi_{SR}$.

(b) It follows from (6.9) that for any $f \in \Pi_{SD}$, $P^t\big(\cdot \mid x, f(x)\big)$ converges to $\eta(f)$ in total variation norm, uniformly in x and at a geometric rate.

(c) It is clear that for any $f \in \Pi_{SD}$,

$$J(\mu, f) = \int_S c(x, f(x))\eta(f)(dx)$$

for any initial law μ.

(d) Compare (6.9) with (A5.5). In view of Theorem 5.4, (A6.2) may be viewed as a representative counterpart of assumptions (A5.1)–(A5.5) for the general state space case.

We now introduce the concept of span-contraction.

Definition 6.1. Let $T : \mathcal{M}_b(S) \to \mathcal{M}_b(S)$. T is said to be a *span-contraction* if, for some $\gamma \in [0, 1)$,

$$\operatorname{span}(Tu - Tv) \leq \gamma \operatorname{span}(u - v), \qquad \text{for all } u, v \in \mathcal{M}_b(S).$$

Let $\sim$ be the equivalence relation on $\mathcal{M}_b(S)$ defined by $u \sim v$ if and only if there exists some constant C such that $u(x) - v(x) = C$ for all $x \in S$. Let $\widetilde{\mathcal{M}}_b(S) = \mathcal{M}_b(S)/\sim$, the quotient space, endowed with the quotient norm induced by the span seminorm. For $v \in \mathcal{M}_b(S)$, let $\tilde{v}$ denote the corresponding element of $\widetilde{\mathcal{M}}_b(S)$ and $\widetilde{T} : \widetilde{\mathcal{M}}_b(S) \to \widetilde{\mathcal{M}}_b(S)$ be the canonically induced map, i.e., $\widetilde{T}\tilde{v} = \widetilde{Tv}$, $v \in \mathcal{M}_b(S)$. It is easily seen that if T is a span-contraction on $\mathcal{M}_b(S)$ then $\widetilde{T}$ is a contraction on $\widetilde{\mathcal{M}}_b(S)$ and, therefore, has a unique fixed point. In turn, it follows that the map T has a span-fixed point, i.e., there exists a $v^* \in \mathcal{M}_b(S)$ such that $\operatorname{span}(Tv^* - v^*) = 0$ or equivalently, $Tv^* - v^*$ is a constant. It also follows that any two span-fixed points of T must differ by a constant.

We now replace (A2.3) with the following:

(A6.3) (i) The multifunction $U(x)$ is continuous;
(ii) $c(\cdot, \cdot) \in C_b(K)$.

We have the following result; for a proof see [80, Lemma 3.5].

Lemma 6.2. *Under (A2.2), (A6.2) and (A6.3), the operator T defined in (2.5) maps $C_b(S)$ to $C_b(S)$ and is a span-contraction.*

Corollary 6.1. *Under (A2.2), (A6.2) and (A6.3) the ACOE has a bounded solution* $(\rho^*, h^*) \in \mathbf{R} \times C_b(S)$.

Proof. This follows from the fact that there exists a $h^* \in C_b(S)$ such that $\text{span}(Th^* - h^*) = 0$. Hence, $Th^* = h^* + \rho^*$ for some constant ρ^*. □

Remark 6.2. (a) Assume (A6.2) and (A6.3). Let $(\rho^*, h^*) \in \mathbf{R} \times C_b(S)$ be a solution to the ACOE and fix $x_0 \in S$. Define $h(\cdot) = h^*(\cdot) - h^*(x_0)$. Then (ρ^*, h) is also a solution to the ACOE. By the span-contraction property of T, it is the unique solution in $\mathbf{R} \times C_b(S)$ satisfying $h(x_0) = 0$, i.e., if $(\rho', h') \in \mathbf{R} \times C_b(S)$ is any other solution of the ACOE in $\mathbf{R} \times C_b(S)$ such that $h'(x_0) = 0$, then $\rho' = \rho$ and $h' = h$. (b) In view of the span-contraction property of the operator T, the value iteration scheme described in Section 4 can be extended to this case; for details we refer to [80, Chap. 3].

Remark 6.3. In Section 4, we have identified the duality between the linear programming formulation and the ACOE under the irreducibility assumption. This has been extended by Yamada [199] to the case when the state space S is a compact subset of $\mathbf{R}^n$ and the transition law has a density which satisfies a certain 'positivity' condition. Hernández-Lerma et.al. [82] have further extended this result to the Borel state space setting under condition (A6.2).

Kurano [102]–[104] has studied the problem for compact state and action spaces, under the hypothesis of Doeblin. Doeblin's condition for the general state space can be described as follows.

(A6.4) There exists a nontrivial finite measure μ on $\big(S, \mathcal{B}(S)\big)$, a positive integer ℓ and an $\varepsilon > 0$ such that

$$P^\ell\big(A \mid x, f(x)\big) \geq 1 - \varepsilon\,, \qquad \text{if } \mu(A) \geq \varepsilon\,,$$

for all $f \in \Pi_{SD}$ and $x \in S$.

Theorem 6.4. *Let the state and action spaces be compact and (A6.3), (A6.4) hold. Then there exist an* $f \in \Pi_{SD}$ *and a set* $A \in \mathcal{B}(S)$ *with* $\mu(A) > \varepsilon$ *such that* $P\big(A \mid x, f(x)\big) = 1$ *for all* $x \in S$, *and* f *is optimal provided that the initial law is supported on the set* A.

Further assume the following.

(A6.5) (Reachability). For any $x \in S$ and $D \in \mathcal{B}(S)$ with $\mu(D) > \varepsilon$ (μ and ε as in (A6.4)) there exists a $\pi \in \Pi$ such that

$$P_x^\pi\Big(\bigcup_{t=0}^{\infty}\{X_t \in D\}\Big) = 1\,.$$

(A6.6) One of the following two conditions is satisfied:

(i) $\mu(\partial D) = 0$ if $\mu(D) > 0$, where ∂D denotes the boundary of D.

(ii) For each $D \in \mathcal{B}(S)$ with $\mu(D) > \varepsilon$, $P(D \mid x, a)$ is continuous in (x, a).

Theorem 6.5. *Under (A6.3)–(A6.6) there exists an $f \in \Pi_{SD}$ which is optimal.*

Remark 6.4. (a) The proof of Theorem 6.3 exploits the idea involved in Lemma 5.1 of extracting a stationary randomized policy from a limit point of empirical processes. A novel idea in [102] is to remove the randomization by using the ergodic decomposition of Markov processes under (A6.4). The compactness is used to ensure the tightness of the empirical processes under any policy. This can be dropped if the cost function has a penalizing condition or if there is a blanket stability of Lyapunov type. The details closely mimic the development at the end of Section 5. (b) Wijngaard [197] has also obtained the existence of an optimal $f \in \Pi_{SD}$ under Doeblin's condition using an operator theoretic method.

We will now discuss the vanishing discount approach to obtain a bounded solution to the ACOE. For a fixed $x_0 \in S$, let $h_\beta(\cdot) = J^*_\beta(\cdot) - J^*_\beta(x_0)$ denote the differential discounted value function. For a general state space, the usual diagonalization procedure used on a countable state space is not amenable. Nevertheless, if $h_\beta(\cdot)$ is uniformly bounded and equicontinuous, then one can use a more subtle diagonalization involving the Arzela-Ascoli theorem to take the required limits and obtain a bounded solution to the ACOE. This was studied by Ross [144]. Following [17], [70], [71], we will discuss some sufficient conditions to obtain the required uniform boundedness and equicontinuity of $h_\beta(\cdot)$.

(A6.7) For each $\beta \in (\beta', 1)$, for some $0 < \beta' < 1$, and $f_\beta \in \Pi_{SD}$, the corresponding state process has a unique invariant probability measure $\eta(f_\beta)$ such that

$$\sup_{\substack{x \in S \\ \beta \in (\beta', 1)}} \sum_{t=1}^{\infty} \| P^t(\cdot \mid x, f_\beta(x)) - \eta(f_\beta)(\cdot) \|_{TV} < \infty . \tag{6.10}$$

The following result is now easy to establish.

Lemma 6.3. *Under (A6.1), (A6.3) and (A6.7), $h_\beta(\cdot) := J^*_\beta(\cdot) - J_\beta(x_0)$, $x_0 \in S$ fixed, is uniformly bounded and equicontinuous for $\beta \in (\beta', 1)$.*

Corollary 6.2. *Under (A6.1), (A6.3) and (A6.7), the ACOE has a solution (ρ^*, h) such that $h \in C_b(S)$.*

Remark 6.5. If (A6.4) is satisfied and we further impose the condition that for every $f \in \Pi_{SD}$, the corresponding state process has a single ergodic class, then (6.10) holds. In particular, if $P(dy \mid x, a)$ has a density $p(y, x, a)$, with respect to some σ-finite measure μ, and there exists a nonnegative measurable function p_0 satisfying $\int p_0(y)\mu(dy) > 0$ and $p(y, x, a) \geq p_0(y)$, for all (x, a), then (A6.4) holds and (6.10) can be easily verified. If $(x, a) \to p(y, x, a)$ is continuous, then by Scheffe's theorem, $p(\cdot \mid x, a)$ is strongly continuous in (x, a).

6.2. Unbounded costs. We now drop the boundedness condition on the cost function and discuss some recent developments involving refinements and extensions of the vanishing discount approach. Since for unbounded costs the uniform boundedness of the differential

discounted value function $h_\beta(\cdot)$ is rather unnatural, one attempts to extend the procedure of [151], [152] to the present case. To this end, we make the following analogues of (A5.14)–(A5.16):

(A6.8) There exists a nonnegative function $b \in \mathcal{M}(S)$, a constant $M \geq 0$, and a sequence $\{\beta_n\} \subset (0,1)$, $\beta_n \uparrow 1$, such that for all $x \in S$,
  (i) $-M \leq h_{\beta_n}(x) \leq b(x)$
  (ii) $\int_S b(y)P(dy \mid x,a) < \infty$, for all $a \in U(x)$.

(A6.9) There exists a policy π and an initial state $\hat{x}$ such that $J(\hat{x},\pi) < \infty$.

(A6.10) There exists $\beta' \in (0,1)$ such that $\sup_{\beta \in (\beta',1)} \tilde{h}_\beta(x) < \infty$ where $\tilde{h}_\beta(x) = J^*_\beta(x) - \inf_{x \in S} J^*_\beta(x)$.

(A6.11) The transition kernel $P(\cdot \mid x,a)$ is strongly continuous in a, for each $x \in S$.

Under (A6.8) and (A6.11), defining $h(x) = \liminf_{n \to \infty} h_{\beta_n}(x)$, $x \in S$, and using Fatou's lemma, one can show that there exists a constant ρ^* such that

$$\lim_{n' \to \infty} (1 - \beta_{n'}) J^*_{\beta_{n'}}(x) = \rho^*, \qquad \text{for all } x \in S, \tag{6.11}$$

where $\beta_{n'} \uparrow 1$ is a subsequence of $\{\beta_n\}$, and

$$\rho^* + h(x) \geq \min_{a \in U(x)} \left\{ c(x,a) + \int_S h(y) P(dy \mid x,a) \right\}, \qquad x \in S, \tag{6.12}$$

which is the ACOI (see (5.16)) for this case. Similarly, under (A6.9)–(A6.11), one can find a constant ρ^* such that along a suitable sequence $\beta_n \in (\beta',1)$, $\beta_n \uparrow 1$, $\lim_{n\to\infty}(1-\beta_n)\inf_{x\in S} J^*_\beta(x) = \rho^*$. Then defining $h(x) = \liminf_{n\to\infty} \tilde{h}_{\beta_n}(x)$ one can deduce (6.12). Thus, we have the following result.

Theorem 6.6. *Under (A6.8) and (A6.11) or under (A6.9)–(A6.11), there exists a constant ρ^* and a function h which is bounded below and satisfies (6.12). Any policy $\pi \in \Pi_{SD}$ realizing the minimum on the right-hand side of (6.12) is average optimal and ρ^* is the minimum average cost.*

Remark 6.6. For details, we refer to [81], [83], [136]. In the case of a countable state space a number of sufficient conditions on the transition kernel and the cost function which enable us to verify (A5.14)–(A5.16) are available, as mentioned in Section 5. This does not seem to be the case for a general Borel state space model, although several interesting examples have been studied in [81], [83] and [136]. Also, (A6.11) is a very strong condition and will not, in general, be satisfied for the transition kernel of the equivalent problem for a partially observable model. Thus, this case needs further investigation. Finally, note that (A6.10) may in principle be easier to verify than (A6.8).

Remark 6.7. We note that Theorem 6.6 provides only an ACOI and not the ACOE. In many situations, the discounted value function is convex (e.g. in linear systems with quadratic cost [14]), or concave (e.g. the separated problem in partially observable models). This class of problems has been used in [59] to obtain the ACOE under (A6.8), (A6.11) and some additional assumptions.

§7. Partially Observable Controlled Markov Processes

Thus far, we have assumed that the complete history of the process H_t is available to the decision-maker, at each stage $t \in T$. However, in many situations some components of the state process may not be directly available to the controller since, e.g. it may be impossible or too costly to measure these. Furthermore, due to imprecisions in the measuring devices, only noisy observations of the state may be available. When these situations arise, the problem is said to be a partially observable controlled Markov process. We study here POCMP with finite or countably infinite state and observation spaces, and finite or compact action set. A major portion of our exposition concentrates on the vanishing discount method, where we see that the particular structure of the POCMP can be employed to yield stronger results than those available for general Borel spaces. We also review Borkar's convex analytic approach, specialized to the partially observable case [26].

7.1. Models with partial state information.

The model for this problem is essentially that in [50, Chap. 8] and is as follows. The state process is described by a pair $\{X_t, Y_t\}_{t\in T}$ taking values in a product of Borel spaces $\boldsymbol{X} \times \boldsymbol{Y}$. Only the second component $\{Y_t\}_{t\in T}$ of the state process is available for decision-making and, reflecting this, $\boldsymbol{Y}$ is called the *observation* or *message space* and Y_t the *observation process.* With $\boldsymbol{A}$ denoting the action space, the evolution of the system is governed by a measurable stochastic kernel P on $\boldsymbol{X} \times \boldsymbol{Y}$ given $\boldsymbol{X} \times \boldsymbol{Y} \times \boldsymbol{A}$.

Let $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$ be an initial distribution of the state. Decomposing (disintegrating) the measure μ, we have

$$\mu(dx, dy) = \overline{Q}_0(dy)\, \psi_0(dx \mid y)\,,$$

where $\overline{Q}_0$ is the marginal of μ on $\boldsymbol{Y}$ and ψ_0 is a version of the regular conditional law, defined $\overline{Q}_0$-a.s.; we pick any version from this equivalence class and keep it fixed thereafter. Note that knowledge of μ, since the value of Y_0 is available to the controller, implies that an *a posteriori* distribution ψ_0 (given $Y_0 = y$) for the unobserved initial state is introduced. We include ψ_0 into the *observed history* by letting

$$\overline{\boldsymbol{H}}_0 := \mathcal{P}(\boldsymbol{X}) \times \boldsymbol{Y}\,, \qquad \overline{\boldsymbol{H}}_t := \overline{\boldsymbol{H}}_{t-1} \times \boldsymbol{Y} \times \boldsymbol{A}\,, \quad t \in \mathbb{N}_0\,.$$

The set of admissible actions is specified by a strict, measurable, compact-valued multifunction $U : \boldsymbol{Y} \to \mathcal{B}(\boldsymbol{A})$. Hence, in this context, an admissible policy is a sequence $\pi = \{\pi_t\}_{t\in T}$ of Borel measurable stochastic kernels π_t on $\boldsymbol{A}$ given $\overline{\boldsymbol{H}}_t$ satisfying, for all $t \in T$, the constraint

$$\pi_t\big(U(y_t) \,\big|\, \overline{h}_t\big) = 1\,, \qquad \forall\, \overline{h}_t \in \overline{\boldsymbol{H}}_t\,.$$

The set of all admissible policies is again denoted by Π.

Remark 7.1. In general, decisions take into account past and present information, and not just the last observation. Notice that the constraints on the actions cannot depend on the unobservable component X_t of the state. If this type of constraint must be included in the model, then it must be provided to the controller as an additional observation. Similarly, if

the cost process $\{c(X_t, Y_t, A_t)\}$ is available to the controller, then it should also be regarded as an additional component in the observation process [50, p. 201].

Remark 7.2. Quite often μ is specified as

$$\mu(dx, dy) = Q_0(dy \mid x)\mu_0(dx),$$

where $\mu_0 \in \mathcal{P}(\boldsymbol{X})$ is an initial distribution for X_0, and Q_0 is a stochastic kernel on $\boldsymbol{Y}$ given $\boldsymbol{X}$ [15, Chap. 10], [80, Chap. 4].

With $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$ and an admissible policy π specified, there exists a unique probability measure $\mathcal{P}_\mu^\pi$ on $\big(\boldsymbol{\Omega}, \mathcal{B}(\boldsymbol{\Omega})\big)$, where $\boldsymbol{\Omega} := (\boldsymbol{X} \times \boldsymbol{Y} \times \boldsymbol{A})^\infty$, defined by

$$\begin{aligned}\mathcal{P}_\mu^\pi(dx_0, dy_0, da_0, \ldots, da_{t-1}, dx_t, dy_t) = \\ \mu(dx_0, dy_0)\, \pi_0(da_0 \mid \psi_0, y_0)\, P(dx_1, dy_1 \mid x_0, y_0, a_0) \\ \cdots\ \pi_{t-1}(da_{t-1} \mid \psi_0, y_0, a_0, \ldots, y_{t-1})\, P(dx_t, dy_t \mid x_{t-1}, y_{t-1}, a_{t-1}).\end{aligned}$$

7.2. Transformation into a completely observable model. A common approach in the analysis of a partially observable (PO) model is to construct a completely observable (CO) model, equivalent to the original one in the sense that corresponding policies have equal costs. The advantages in doing this are obvious, since the theory of CO problems is much better developed. However, the price usually paid is that the dimensionality of the new state space is substantially larger than that of the original one.

Such an equivalent CO problem can be obtained in many ways. The main idea is to specify an *information state process* that summarizes, at each time, all relevant information for decision-making. Clearly, $\overline{H}_t = (\psi_0, Y_0, A_0, \ldots, A_{t-1}, Y_t)$ can be used as an information state process, but this leads to a nonstationary CO model, in which "growing memory" difficulties arise; see [15, Chap. 10]. We present here the more standard approach where the inferential knowledge of X_t is summarized using its conditional probability distribution, given the entire observed history up to time t. We first present the construction of the equivalent CO model for general Borel state spaces and then specialize to models with countable state space. Also, the following assumption will be in effect throughout this Section:

(A7.1) The transition kernel $P(\cdot \mid x, y, a)$ and the cost function $c(x, y, a)$ do not depend on y, and $U(y) = \boldsymbol{A}$, for all $y \in \boldsymbol{Y}$.

Given a PO model $(\boldsymbol{X} \times \boldsymbol{Y}, \boldsymbol{A}, U, P, c)$, satisfying (A7.1), we construct a CO model $\big(\mathcal{P}(\boldsymbol{X}), \boldsymbol{A}, \widetilde{U}, \mathcal{K}, \tilde{c}\big)$ as follows. Let $\{\Psi_t, Y_t\}_{t \in T}$ and $\{\widetilde{H}_t\}_{t \in T}$ denote the state process and the history spaces respectively. The set of admissible actions is selected by letting $\widetilde{U}(\psi) = \boldsymbol{A}$, for all $\psi \in \mathcal{P}(\boldsymbol{X})$. We define the cost function $\tilde{c}$ by

$$\tilde{c}(\psi, a) := \int_{\boldsymbol{X}} c(x, a)\, \psi(dx), \qquad \psi \in \mathcal{P}(\boldsymbol{X}). \tag{7.1}$$

It remains to construct the transition kernel $\mathcal{K}$. Working on the canonical sample space $\widetilde{\boldsymbol{\Omega}} = (\mathcal{P}(\boldsymbol{X}) \times \boldsymbol{A})^\infty$ we first define a stochastic kernel q on $\boldsymbol{X} \times \boldsymbol{Y}$ given $\mathcal{P}(\boldsymbol{X}) \times \boldsymbol{A}$, by

$$q(dx, dy \mid \psi, a) := \int_{\boldsymbol{X}} P(dx, dy \mid x', a)\, \psi(dx'), \qquad \psi \in \mathcal{P}(\boldsymbol{X}) \tag{7.2}$$

and decomposing q, we obtain

$$q(dx, dy \mid \psi, a) = Q(dy \mid \psi, a)\, \Psi(dx \mid \psi, a, y)\,. \tag{7.3}$$

Equation (7.3) is the *filtering equation.* For fixed (ψ, a), the map $y \mapsto \Psi$, as defined implicitly in (7.3), is a measurable mapping from $\boldsymbol{Y}$ to $\boldsymbol{\mathcal{P}}(\boldsymbol{X})$. Consequently, along with the distribution Q on $\boldsymbol{Y}$, it induces a distribution $\mathcal{K}$ on $\boldsymbol{\mathcal{B}}\big(\boldsymbol{\mathcal{P}}(\boldsymbol{X})\big)$ which is a measurable function of (ψ, a), or, in other words, a stochastic kernel on $\boldsymbol{\mathcal{P}}(\boldsymbol{X})$ given $\boldsymbol{\mathcal{P}}(\boldsymbol{X}) \times \boldsymbol{A}$. It follows that the model $\big(\boldsymbol{\mathcal{P}}(\boldsymbol{X}), \boldsymbol{A}, \mathcal{K}, \tilde{c}\big)$, with state process $\{\Psi_t\}_{t \in T}$, forms a completely observable controlled Markov process, with transition kernel given by

$$\mathcal{K}(B \mid \psi, a) := \int_{\boldsymbol{Y}} I\big\{\Psi(\cdot \mid \psi, a, y) \in B\big\}\, Q(dy \mid \psi, a)\,, \qquad B \in \boldsymbol{\mathcal{B}}\big(\boldsymbol{\mathcal{P}}(\boldsymbol{X})\big)\,. \tag{7.4}$$

The distribution $\tilde{\mu}_0$ of Ψ_0, corresponding to an initial distribution μ of the PO model, is taken to be

$$\tilde{\mu}_0(B) := \int_{\boldsymbol{Y}} \mu(B, dy)\,, \qquad B \in \boldsymbol{\mathcal{B}}\big(\boldsymbol{\mathcal{P}}(\boldsymbol{X})\big)\,. \tag{7.5}$$

Given a history $\overline{h}_t = (\psi_0, y_0, \ldots, a_{t-1}, y_t) \in \overline{\boldsymbol{H}}_t$ in the PO model we can construct $\psi_1, \psi_2, \ldots$ in a recursive manner by starting from ψ_0 and, having obtained ψ_{t-1}, solving for Ψ in (7.3), with $(\psi, a, y) = (\psi_{t-1}, a_{t-1}, y_t)$, and letting $\psi_t = \Psi$. In this manner we obtain a corresponding history $\tilde{h}_t = (\psi_0, a_0, \ldots, a_{t-1}, \psi_t) \in \widetilde{\boldsymbol{H}}_t$ for the CO model; we denote this correspondence by the map $g_t : \overline{\boldsymbol{H}}_t \to \widetilde{\boldsymbol{H}}_t$. We can then assign to each admissible policy $\tilde{\pi} \in \widetilde{\Pi}$ in the CO model a corresponding policy $\pi = g^*(\tilde{\pi})$ in the PO model, defined by

$$\pi_t(\cdot \mid \overline{h}_t) := \tilde{\pi}_t\big(\cdot \mid g_t(\overline{h}_t)\big)\,, \qquad \overline{h}_t \in \overline{\boldsymbol{H}}_t\,. \tag{7.6}$$

Clearly every policy $\pi \in \Pi$ can also be regarded as a policy in $\widetilde{\Pi}$; in other words, the map g^* is onto. If $\mathcal{P}_{\tilde{\mu}}^{\tilde{\pi}}$ is the probability measure induced by the policy $\tilde{\pi}$ and the initial distribution $\tilde{\mu}$ (corresponding to μ) on the canonical sample space $\widetilde{\Omega}$, then for each $C \in \boldsymbol{\mathcal{B}}(\boldsymbol{X})$

$$\mathcal{P}_{\mu}^{g^*(\tilde{\pi})}(X_t \in C \mid \overline{H}_t = \overline{h}_t) = \Psi_t(C)\,, \qquad \mathcal{P}_{\tilde{\mu}}^{\tilde{\pi}}\text{-a.s.} \tag{7.7}$$

Utilizing (7.1), (7.4) and (7.5), it can be verified that

$$E_{\mu}^{g^*(\tilde{\pi})}\big[c(X_t, A_t)\big] = E_{\tilde{\mu}}^{\tilde{\pi}}\big[\tilde{c}(\Psi_t, A_t)\big]\,, \qquad \forall\, t \in T\,, \tag{7.8}$$

thus establishing that the two models are indeed equivalent as claimed. It follows that the process Ψ_t summarizes all information, relevant for control purposes, and is called for this purpose a *sufficient statistic* (see [157], [158] and [49]). We define the set of *separated policies* Π_S as those policies $\pi \in \Pi$ for which there a Markov policy $\tilde{\pi}$ on the equivalent CO problem such that $\pi = g^*(\tilde{\pi})$, as defined in (7.6). In other words, with $\tilde{\pi} = \{f_t\}_{t \in T} \in \widetilde{\Pi}_M$, $f_t : \boldsymbol{\mathcal{P}}(\boldsymbol{X}) \to \boldsymbol{\mathcal{P}}(\boldsymbol{A})$ and for each initial distribution $\mu \in \boldsymbol{\mathcal{P}}(\boldsymbol{S})$,

$$\pi_t(\cdot \mid \overline{h}_t) = f_t\big(\Psi_t\big)(\cdot)\,, \qquad \mathcal{P}_{\tilde{\mu}_0}^{\tilde{\pi}}\text{-a.s.}$$

Thus, the actions taken using a separated policy only depend on $\overline{H}_t$ through the conditional distribution of X_t. In other words, the following *separation principle* holds: if an optimal policy exists in Π, one exists in Π_S. Hence, the process can be controlled optimally by first estimating the state via the conditional distribution, and choosing control actions based solely on the latter. These and other results, in various degrees of generality, were independently obtained by various authors, e.g. [3], [5], [87], [134], [159], [147], [170], [171], [195], [201].

Example 7.1. A partially observable version of the stochastic nonlinear system in Example 2.1 is described by the equations

$$\begin{aligned} X_{t+1} &= F(X_t, A_t, W_t)\,, \\ Y_t &= G(X_t, A_{t-1}, V_t)\,, \\ Y_0 &= G_0(X_0, V_0)\,, \end{aligned}$$

where G and G_0 are Borel measurable, and the disturbance $\{V_t\}_{t\in T}$ is an i.i.d. sequence of random variables taking values in a Borel space $\boldsymbol{V}$, with a common distribution $\mathcal{P}_V$; furthermore, it is assumed that X_0, $\{W_t\}$ and $\{V_t\}$ are mutually independent.

We now specialize to the case where the state space $\boldsymbol{X} \times \boldsymbol{Y}$ is a finite or countably infinite set, the action space $\boldsymbol{A}$ is a finite or compact set and with assumption (A7.1) in effect. Thus, $U(y) = \boldsymbol{A}$, for all $y \in \boldsymbol{Y}$, and the kernel of the process takes the form $P(x', y' \mid x, a)$. We also assume that the cost c and the kernel P are continuous with respect to $a \in \boldsymbol{A}$. The space $\mathcal{P}(\boldsymbol{X})$ is identified with the set Δ of probability vectors, i.e.,

$$\Delta := \Big\{\psi \in [0,1]^{\boldsymbol{X}} \;:\; \sum_{x\in\boldsymbol{X}} \psi(x) = 1\Big\} \tag{7.9}$$

endowed with the topology given by the metric

$$d(\psi_1, \psi_2) := \sum_{x\in\boldsymbol{X}} \big|\psi_1(x) - \psi_2(x)\big| = \|\psi_1 - \psi_2\|_1\,,$$

where $\|\cdot\|_1$ stands for the standard ℓ_1-norm on $\mathbb{R}^{\boldsymbol{X}}$.

In general, the recursive (filtering) equation (7.3) used to compute ψ_{t+1}, is obtained via a decomposition of measures technique, see [15, Chap. 10], [50, Chap. 8], [80, Chap. 4], [201]. This is particularly simple to accomplish (using Bayes rule) when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are countable, or when the system is described by a linear system function and the disturbances are Gaussian, see [5], [14], [100], [170], [171]. For this purpose, we need the following definitions (compare with (7.2)–(7.3)).

$$q(x, y \mid \psi, a) := \sum_{x'\in\boldsymbol{X}} P(x, y \mid x', a)\,\psi(x') \tag{7.10}$$

$$V(y, \psi, a) := \sum_{x\in\boldsymbol{X}} q(x, y \mid \psi, a) \tag{7.11}$$

$$T(y, \psi, a)(\cdot) := \begin{cases} \frac{q(\cdot\,, y\mid\psi, a)}{V(y,\psi,a)}\,, & \text{if } V(y,\psi,a) \neq 0 \\ 0\,, & \text{otherwise} \end{cases} \tag{7.12}$$

Note that the map $\psi \to T(y,\psi,a)$ maps Δ into itself. In the countable case ψ_t can be computed by letting $\psi_t = T(y_t, \psi_{t-1}, a_{t-1})$. Here, $V(y,\psi,a)$ is interpreted as the (one-step ahead) conditional probability of the observation being y given an *a priori* distribution ψ for the core state, under decision a. Likewise, $T(y,\psi,a)$ is interpreted as the *a posteriori* conditional probability distribution of the core state given decision a was made, observation y obtained, and an *a priori* distribution ψ. Also, the kernel in (7.4) takes the form

$$\mathcal{K}(B \mid \psi, a) := \sum_{y \in \boldsymbol{Y}} V(y,\psi,a)\, I\{T(y,\psi,a) \in B\}\,, \qquad B \in \mathcal{B}(\Delta)\,, \tag{7.13}$$

while the cost $\tilde{c}$ is computed by

$$\tilde{c}(\psi, a) := \sum_{x \in \boldsymbol{X}} c(x,a)\psi(x)\,. \tag{7.14}$$

Remark 7.3. It is common to specify, instead of the kernel P, a transition kernel $\overline{P}$ on $\boldsymbol{X}$ given $\boldsymbol{X} \times \boldsymbol{A}$ and an *observation kernel* $\overline{Q}$ on $\boldsymbol{Y}$ given $\boldsymbol{X} \times \boldsymbol{A}$ [14], [61], [80], [124], [166]. Note that this is only a special case of our presentation, which happens when the kernel P admits the decomposition

$$P(x, y \mid x', a) = \overline{Q}(y \mid x, a)\overline{P}(x \mid x', a)\,.$$

In this case, we can express (7.10)–(7.12) in a convenient vector form by viewing ψ as an element of $\mathbb{R}^{\boldsymbol{X}}$ and defining the transition matrix $\big[\overline{P}(a)\big]_{x,x'} := \overline{P}(x \mid x', a)$ and the observation matrix $Q_y(a) := \operatorname{diag}\big\{Q(y \mid x, a) : x \in \boldsymbol{X}\big\}$. Then with $\overline{q}$ denoting the vector in $\mathbb{R}^{\boldsymbol{X}}$ defined by $\overline{q}_x(y \mid \psi, a) := q(x, y \mid \psi, a)$ and $1' = (1, \ldots, 1)$, we have

$$\overline{q}(y \mid \psi, a) = \overline{Q}_y(a)\overline{P}(a)\psi \tag{7.10$'$}$$

$$V(y,\psi,a) = 1'\overline{Q}_y(a)\overline{P}(a)\psi \tag{7.11$'$}$$

and analogously for (7.12).

Note that a *nonrandomized* separated admissible policy can be viewed as a sequence of maps $\pi_t : \Delta \to \boldsymbol{A}$. Then an equivalent, *completely observable*, discounted cost problem (DC$'$) can be formulated as finding a separated admissible policy which minimizes

$$J_\beta(\psi, \pi) := E^{\pi}_{\psi_0}\left[\sum_{t=0}^{\infty} \beta^t \tilde{c}(\Psi_t, A_t)\right].$$

The average cost problem (AC$'$) is analogously defined.

Note that the one-stage cost function $\tilde{c}(\psi, a)$ is linear in $\psi \in \Delta$. It is easy to show that the expectation operator corresponding to the kernel $\mathcal{K}$ preserves concavity (convexity) [49], [6]. The following results complement those in Theorem 2.1.

Theorem 7.1. *For a (DC') decision problem, $J^*_\beta(\cdot)$ is a concave function, for all $0 < \beta < 1$. The DCOE is given by*

$$J^*_\beta(\psi) = \min_{a \in A}\left\{ \tilde{c}(\psi, a) + \beta \sum_{y \in Y} V(y, \psi, a) J^*_\beta\big(T(y, \psi, a)\big) \right\}, \tag{7.15}$$

and any (nonrandomized) separated stationary policy which attains the minimum above is optimal.

Remark 7.4. The optimality equation (7.15) is obtained from the general theory of CMP [15], [80]. For other results, see [5]–[7], [14], [49], [124], [157], [166], [167], [165].

In this context, a pair (ρ, h) is said to be a solution to the ACOE if, for all $\psi \in \Delta$,

$$\rho + h(\psi) = \min_{a \in A}\left\{ \tilde{c}(\psi, a) + \sum_{y \in Y} V(y, \psi, a) h\big(T(y, \psi, a)\big) \right\}. \tag{7.16}$$

7.3. The vanishing discount approach. As shown in Section 5, for a countable state space CMP, boundedness conditions on the differential discounted value function were sufficient for solutions to the corresponding ACOE to exist. We consider here the following hypothesis:

(A7.2) There exists a sequence $\beta_n \uparrow 1$, such that h_{β_n} is bounded.

Despite the fact that the model $(\Delta, A, \mathcal{K}, \tilde{c})$ has a general Borel state space, it has two special features which simplify the analysis via the vanishing discount method. The first of these features is the concavity of the discounted value function while the second is the fact that the kernel $\mathcal{K}(\cdot \mid \psi, a)$ vanishes on the complement of a countable set (for fixed ψ and a) and, thus, the integrals with respect to $\mathcal{K}$ reduce to infinite sums.

For the finite state and action space case, the concavity of the discounted value function has been exploited by Platzman [132] and by Ohnishi et al. [128]. These authors utilize the fact that a collection of concave functions defined on some relatively open convex set C which are finite and pointwise bounded, is uniformly bounded and equi-Lipschitzian relative to any closed subset of C [139, Th. 10.6]. Thus, under condition (A7.2), the finite dimensionality of Δ and the concavity of $h_\beta(\cdot)$ are used in [128], [132] to obtain a bounded solution (ρ^*, h) to the ACOE, via the vanishing discount approach. In particular, they partition Δ into its interior, its vertices, and its edges, i.e.,

$$\Delta = \bigcup_{j \in \mathcal{J}} \Delta_j .$$

Note that $|\mathcal{J}| = 2^{|X|+1} - 1$ and that each set Δ_j is a relatively open convex set. Given a sequence $\beta_n \uparrow 1$, then the concavity of $h_\beta(\cdot)$ and (A7.2) are used to obtain subsequences $\beta_n(j)$ such that $\{h_{\beta_n(j)}(\cdot)\}$ converges on Δ_j. Platzman [132] defines a metric on Δ which accomplishes this partition. Let

$$\mathcal{I}(\psi) := \big\{ i \in X \,:\, \psi(i) > 0 \big\}, \qquad \psi \in \Delta,$$

$$d(\psi_1, \psi_2) := 1 - \min\left\{ \frac{\psi_1(i)}{\psi_2(i)} \,:\, i \in \mathcal{I}(\psi_2) \right\}, \qquad \psi_1, \psi_2 \in \Delta,$$

$$D(\psi_1, \psi_2) := \max\big\{ d(\psi_1, \psi_2); d(\psi_2, \psi_1) \big\}.$$

In [131, pp. 88-89], Platzman shows that $D(\cdot,\cdot)$ is a metric which leaves Δ disconnected and with components identical to the elements of the partition $\{\Delta_j\}_{j\in\mathcal{J}}$. The following is shown in [132, Lemma A.1].

Lemma 7.1. *Let $f : \Delta \to \mathbf{R}$ be concave and bounded below; then*

$$\left|f(\psi_1) - f(\psi_2)\right| \leq \operatorname{span}(f)\, D(\psi_1, \psi_2)\,.$$

Hence, under (A7.2), $\{h_\beta(\cdot)\}_{\beta\in(0,1)}$ is an equi-Lipschitzian family, with common Lipschitz constant given by the (smallest) uniform bound, and the Arzela-Ascoli Theorem can be used as in [144] to obtain a bounded solution to the ACOE.

If the state space is infinite the above method does not work, simply because the partition induced by the Platzman metric results in a nonseparable space. In this situation the particular structure of the kernel has been employed in [61] to develop a theoretical framework based on the notion of *invariant* subsets (sub-processes) of a CMP and sufficient conditions are given for the existence of solutions to the ACOE, in the case of a finite action space. The key point is to note that if we let $B(\psi,a) := \{T(y,\psi,a) : y \in \boldsymbol{Y}\}$, which is a countable set since $\boldsymbol{Y}$ is countable, then $\mathcal{K}(B(\psi,a) \mid \psi, a) = 1$. Thus, at any time $t \in \mathbf{N}_0$, the set of possible *next states* for Ψ_t is the set $\bigcup_{a\in\boldsymbol{A}} B(\Psi_t, a)$, which is countable, provided $\boldsymbol{A}$ is finite. This special structure has also been identified by other authors, e.g. [5, p. 187], [132, p. 369], [166, pp. 19-20].

We briefly summarize the work in [61]. The notions of *descendents*, *ancestors* and *relatives* of a point $\psi \in \Delta$ are first introduced. The descendents of ψ are defined as the smallest subset of Δ containing ψ which is invariant under the action of the maps in the collection $\{T(y,\cdot,a) : y \in \boldsymbol{Y}, a \in \boldsymbol{A}\}$, while the ancestors of ψ are defined as all the points in Δ which reach ψ under the application a finite sequence of these maps. Finally, the relatives of a point ψ, denoted by $\mathcal{R}_\psi^{(1)}$, is the set formed by the union of its descendents and ancestors. Note that the definition of the descendents is an extension, to the present context, of Doob's concept of *consequent sets* [44, p. 206]. Subsequently, the *genealogical tree* GT_ψ of ψ is defined by

$$GT_\psi := \bigcup_{n\in\mathbb{N}} \mathcal{R}_\psi^{(n)},$$

where the sets $\mathcal{R}_\psi^{(n)}$ are defined recursively as

$$\mathcal{R}_\psi^{(n+1)} := \bigcup_{s\in\mathcal{R}_\psi^{(n)}} \mathcal{R}_s^{(1)}, \qquad n \in \mathbb{N}\,.$$

The descendents of a point form a countable set, but the ancestors can in general be uncountably many. To guarantee that the relatives and hence the genealogical tree of a point is a countable set the following condition is introduced.

(A7.3) For all $y \in \boldsymbol{Y}$, $a \in \boldsymbol{A}$, and $\psi \in \Delta$, $T^{-1}(y,\psi,a)$ is a countable set.

Introduce the relation $\psi \sim \psi'$ if $GT_\psi = GT_{\psi'}$. It follows that "$\sim$" defines an *equivalence relation* on Δ resulting in a partition of Δ into equivalence classes which are precisely

the sets GT_ψ. Under conditions (A7.2)–(A7.3) the standard diagonalization argument can be employed on each equivalence class GT_ψ to construct a pair (ρ^*, h_{GT_ψ}) which solves the ACOE on GT_ψ (the boundedness hypothesis (A7.2) can be weakened by letting the constant M depend on the equivalence class). Then, by defining $h(\psi) := h_{GT_\psi}(\psi)$ for all $\psi \in \mathbf{\Delta}$, (ρ^*, h) clearly solves the ACOE on $\mathbf{\Delta}$. One peculiarity of this approach is that the resulting function h is not guaranteed to be measurable. This is not a major problem though, since an important consequence of the particular structure (with finite action space) is that the "measurability of various objects is of no essential concern" for the equivalent problem [15, p. xi]. The approach in [61] fails when the action space $\boldsymbol{A}$ is not finite.

Since the vanishing discount method relies heavily on the boundedness of the differential discounted value function, the problem of finding sufficient conditions on the cost and the kernel of the process for this to hold becomes an important one. Platzman [132] has given (reachability and detectability) conditions for (A7.2) to hold; however, these conditions are difficult to verify. On the other hand, many models of interest possess special properties, which allow the verification of condition (A7.2) very easily.

Suppose that a partial order "$\prec_{\mathbf{\Delta}}$" has been defined on $\mathbf{\Delta}$, and let "$\prec_{\boldsymbol{A}}$" denote a linear order on $\boldsymbol{A}$; we assume that $\boldsymbol{A}$ is finite. We also identify $\boldsymbol{X}$ with $\mathbb{N}_0$ and endow it with its natural ordering.

Definition 7.1. Consider $\big((\mathbf{\Delta}, \prec_{\mathbf{\Delta}}), (\boldsymbol{A}, \prec_{\boldsymbol{A}}), \mathcal{K}, \tilde{c}\big)$, and let $\psi_1, \psi_2 \in \mathbf{\Delta}$. We say that:

(i) the value functions are *monotone* if

$$\psi_1 \prec_{\mathbf{\Delta}} \psi_2 \implies J^*_\beta(\psi_1) \leq J^*_\beta(\psi_2), \qquad \text{for all } 0 < \beta < 1,$$

(ii) a (nonrandomized) stationary separated policy π is *monotone* if

$$\psi_1 \prec_{\mathbf{\Delta}} \psi_2 \implies \pi(\psi_1) \prec_{\boldsymbol{A}} \pi(\psi_2).$$

Two frequently used partial orders on $\mathbf{\Delta}$ are the *stochastic dominance* $\prec_{st}$ and the *monotone likelihood ratio* $\prec_{lr}$, defined below.

Definition 7.2. Let $\psi_1, \psi_2 \in \mathbf{\Delta}$; we say that:

(i) $\psi_1 \prec_{st} \psi_2$ if $\displaystyle\sum_{i \geq q} \psi_1(i) \leq \sum_{i \geq q} \psi_2(i)$, for all $q \in \boldsymbol{X}$, and

(ii) $\psi_1 \prec_{lr} \psi_2$ if $\psi_1(j)\psi_2(i) \leq \psi_1(i)\psi_2(j)$, for all $i, j \in \boldsymbol{X}$ such that $i \leq j$.

Let e^j denote the element of $\mathbf{\Delta}$ with the j^{th} component equal to 1, $j \in \boldsymbol{X}$; thus, e.g. $e^0 = (1, 0, 0, \dots)$. The following is easily shown.

Lemma 7.2. *If $\psi_1, \psi_2 \in \mathbf{\Delta}$ and $\psi_1 \prec_{lr} \psi_2$, then $\psi_1 \prec_{st} \psi_2$. Also, for all $\psi \in \mathbf{\Delta}$, $e^0 \prec_{lr} \psi$.*

Definition 7.3. An action $a_j \in \boldsymbol{A}$ is called a reset action if, for some $j \in \boldsymbol{X}$, $T(y, \psi, a_j) = e^j$, for all $y \in \boldsymbol{Y}$ and $\psi \in \mathbf{\Delta}$.

A reset action a_j corresponds to the core state of the system being j, with probability one, at the next time epoch after action a_j has been taken. This type of action arises naturally in manufacturing systems subject to inspection, maintenance, and replacement. The following results derive from the work of Sondik [166]; see also [61].

Lemma 7.3. *If there exists a reset action $a_j \in \boldsymbol{A}$, then*

$$J^*_\beta(\psi) - J^*_\beta(e^j) \le \tilde{c}(\psi, a_j), \qquad \forall\, \psi \in \boldsymbol{\Delta}.$$

If $\boldsymbol{X}$ is finite, and for each $j \in \boldsymbol{X}$ there is a corresponding reset action, then for each $\beta \in (0,1)$ there exists $J \in \boldsymbol{X}$ such that

$$0 \le J^*_\beta(\psi) - J^*_\beta(e^J) \le M, \qquad \forall\, \psi \in \boldsymbol{\Delta},$$

where $M := \max\{c(i,a) \mid i \in \boldsymbol{X}, a \in \boldsymbol{A}\}$.

Remark 7.5. Note that if $J^*_\beta(\cdot)$ is monotone with respect to $\prec_{lr}$, and if there is an action $a_0 \in \boldsymbol{A}$ that resets the state to e^0, then $0 \le J^*_\beta(\psi) - J^*_\beta(e^0) \le \tilde{c}(\psi, a_0)$ uniformly in $\beta \in (0,1)$. Furthermore, note that when $\boldsymbol{X}$ is finite, a constant $M > 0$ exists such that $\tilde{c}(\psi, a_0) \le M$, for all $\psi \in \boldsymbol{\Delta}$, and thus condition $(A7.2)$ holds.

Models with a *replacement* action that resets the system to an "as new" state e^0 have been considered in [2], [115], [127], [128], [145], [187]–[191], [184], [185]. Related problems are those considered in [64], where a reset action to a most desirable state is available, and in [88], where (maintenance) reset actions a_j are available for all $j \neq 0$, with $\boldsymbol{X}$ a finite set.

7.4. The convex analytic method. We will now briefly describe Borkar's convex analytic approach. The action set $\boldsymbol{A}$ is assumed to be any compact metric space. We also assume that c and P are continuous in a. We will consider the pathwise average cost. This cannot in general be written as an equivalent cost in terms of $\{\Psi_t\}$, but it is natural to propose

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{c}(\Psi_t, A_t) \tag{7.17}$$

as a substitute. Any $\mu \in \mathcal{P}(\boldsymbol{\Delta} \times \boldsymbol{A})$ can be decomposed as

$$\mu(d\psi, da) = \overline{\mu}(d\psi)\Phi(\psi)(da)\,. \tag{7.18}$$

where $\overline{\mu}$ is the marginal of μ on $\boldsymbol{\Delta}$ and Φ is the regular conditional law defined $\overline{\mu}$–a.s. We shall always work with one arbitrary representative of this equivalence class. Define $\Gamma \subset \mathcal{P}(\boldsymbol{\Delta} \times \boldsymbol{A})$ by

$$\begin{aligned} \Gamma &= \left\{ \mu \in \mathcal{P}(\boldsymbol{\Delta} \times \boldsymbol{A}) \;\middle|\; \begin{array}{l}\text{For } \overline{\mu}, \Phi \text{ as in (7.18), } \overline{\mu} \text{ is invariant under} \\ \text{the stationary randomized policy } \Phi\end{array} \right\} \\ &= \left\{ \mu \in \mathcal{P}(\boldsymbol{\Delta} \times \boldsymbol{A}) \;\middle|\; \iiint f(\psi)\mathcal{K}(d\psi \mid \psi', a)\Phi(\psi')(da)\overline{\mu}(d\psi') \right. \\ &\qquad\qquad \left. = \int f\, d\overline{\mu}, \quad \text{for all } f \in C_b(\boldsymbol{\Delta}) \right\}. \end{aligned} \tag{7.19}$$

From (7.19) one can easily check that Γ is closed. Note that the set of invariant probability measures for the process $\{\Psi_t\}$ controlled by a stationary randomized policy Φ, when

nonempty, need not be a singleton. In general, it will form a closed convex set in $\mathcal{P}(\Delta)$, the extreme points of which correspond to ergodic measures. That is, the above process with one of these extreme measures (say, μ) as the initial condition will be ergodic. Then (7.17) will a.s. equal $\int \tilde{c}\, d\mu$. In view of the ergodic decomposition of a stationary Markov process, this will also be the case for other invariant measures (which will be a convex combination of the ergodic ones). Define

$$\rho^* = \inf_{\mu \in \Gamma} \int \tilde{c}\, d\mu\,.$$

We assume that $\rho^* < \infty$. We consider two alternative conditions under which the above infimum will be a minimum.

(A7.4) Near-monotone case: c satisfies

$$\lim_{i \to \infty} \inf_{a} c(i,a) = \infty$$

(A7.5) Stable case. Condition (A5.19′) (ii) holds.

Observe that the 'near-monotonicity' condition here is more restrictive than the one used in Section 5. We now state the following result; the proof is analogous to that of Theorem 5.10.

Lemma 7.4. *Under either (A7.4) or (A7.5), the map $\mu \mapsto \int \tilde{c}\, d\mu$ attains its minimum on Γ.*

Define the $\mathcal{P}(\Delta \times A)$-valued process $\{\eta_t\}$ by

$$\int f\, d\eta_t = \frac{1}{t} \sum_{m=0}^{t-1} f(\Psi_t, A_t), \qquad t \geq 1\,, \qquad f \in C_b(\Delta \times A)\,,$$

where $\{\Psi_t\}$ is governed by some policy. Again we can prove the following analogue of Lemma 5.1.

Lemma 7.5. *With probability 1, any limit point of $\{\eta_t\}$ in $\mathcal{P}(\Delta \times A)$ lies in Γ.*

Consider the near-monotone case. Suppose that for a given sample path, a subsequence of $\{\eta_t\}$ has no limit point in $\mathcal{P}(\Delta \times A)$. Arguments similar to those in the proof of Theorem 5.1 can be used to show that the cost must go to $+\infty$ along this subsequence. In view of Lemma 7.5, this leads to

$$\liminf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{c}(\Psi_t, A_t) \geq \rho^*\,, \qquad \text{a.s.} \tag{7.20}$$

Along with Lemma 7.4, this would seem to lead to the existence of an optimal stationary randomized policy. There is, however, one catch. It is not a priori clear that any initial law for $\{\Psi_t\}$ would be in the domain of attraction of the element(s) of Γ that minimize the cost (or, for that matter, whether this domain of attraction can be reached in a finite random time from any initial law under some policy). Similar 'reachability' problems surface when

one tries to extend the dynamic programming equations. These are circumvented under somewhat stringent conditions in [132] as we have already discussed.

Finally, one can prove the convexity of Γ. Again it is unclear how (and whether) one can characterize the extreme points of Γ as those corresponding to stationary (non-randomized) policies. As the ACOE is not available in this approach, the existence of an optimal stationary policy remains an open issue in general. In the stable case, it is not clear if (7.20) holds and thus this case remains open to investigation. To sum up, the convex analytic approach to POCMP needs to be further studied.

§8. Multiobjective and Constrained Models

An important success of the convex analytic approach discussed in Section 5 is in the domain of multiobjective problems, in which there is more than one cost (objective) function. We will first consider a multiobjective CMP with average cost criterion recast as a CMP with several constraints. CMP with one or multiple constraints have been studied in [1], [16], [26], [39], [40], [43], [90], [91], [94], [116], [125], [141], [142], [162]. Our presentation follows [25], [26].

We consider the case when $\boldsymbol{S} = \{0, 1, 2, \dots\}$, $\boldsymbol{A}$, the action space, is a prescribed compact metric space and $P(j \mid i, a)$ is continuous in a for fixed i, j. Also, $U(i) = \boldsymbol{A}$ for all $i \in \boldsymbol{S}$. In the constrained CMP problem we have in addition to the cost function $c \in C_b(\boldsymbol{S} \times \boldsymbol{A})$, m additional 'costs' $c_i \in C_b(\boldsymbol{S} \times \boldsymbol{A})$, $1 \le i \le m$, and are required to satisfy

$$a_i \le \int c_i \, d\hat{\eta}(f) \le b_i \,, \qquad 1 \le i \le m \tag{8.1}$$

for prescribed numbers $b_i > a_i$, $f \in \Pi_{SD}$ and $\hat{\eta}(f) \in \boldsymbol{\mathcal{P}}(\boldsymbol{S} \times \boldsymbol{A})$ is as in Section 5. (We are assuming all costs are bounded for the sake of simplicity. Also, we are confining our attention to Π_{SSR}; this suffices under reasonable hypotheses, as we saw in Section 5.) We will assume condition (A5.20) in Section 5.

Recall that $I_R = \{\hat{\eta}(f) : f \in \Pi_{SSR}\}$. Let $\widetilde{I}_R$ be the subset of I_R where the constraints (8.1) are satisfied. Then $\widetilde{I}_R$ is closed and convex. We assume also that it is compact (this will be true under (A5.20) in Section 5). Under this assumption one can show as in Section 5 that there exists an $f^* \in \Pi_{SR}$ which is optimal for this problem. We will now proceed to show that f^* requires randomization in at most m states.

Let $g \in C_b(\boldsymbol{S} \times \boldsymbol{A})$. For some $a \in \mathbb{R}$, let $H = I_R \cap \{\psi : \int g \, d\psi \le a\}$, assumed to be nonempty. Clearly H is closed and convex. Let $\hat{\eta}(f)$ be an extreme point of H. Suppose it is not an extreme point of I_R itself. Then there exist distinct measures $\hat{\eta}(f_{11})$, $\hat{\eta}(f_{12})$ such that at least one of them (say $\hat{\eta}(f_{11})$) is not in H and $\hat{\eta}(f)$ is a convex combination of the two. Suppose $\hat{\eta}(f_{21}) \in I_R \setminus H$, $\hat{\eta}(f_{22})$ is another such pair. Then it can be shown that $\hat{\eta}(f_{ij})$, $1 \le i, j \le 2$, are collinear (I_R, $\widetilde{I}_R$, H, etc. are viewed as subsets of $\mathfrak{M}(\boldsymbol{S} \times \boldsymbol{A})$, the Banach space of finite signed measures on $\boldsymbol{S} \times \boldsymbol{A}$). Therefore, all pairs of points in I_R satisfying: (a) at least one of them is not in H, and (b) $\hat{\eta}(f)$ is a convex combination thereof, lie on a single straight line in $\mathfrak{M}(\boldsymbol{S} \times \boldsymbol{A})$. Let Z denote the intersection of this line with I_R. Under our hypotheses on I_R, Z is a closed finite line segment. Let $\eta(f_1)$, $\eta(f_2)$

denote its end points. Then it can be shown that $\eta(f_i)$, $i = 1,2$ are extreme points of I_R. By Lemma 5.2, $f_i \in \Pi_{SSD}$; also, f_1 and f_2 are distinct since $\hat{\eta}(f)$ is not an extreme point of I_R. Therefore, there exists an $a' \in (0,1)$ such that

$$\hat{\eta}(f) = a'\hat{\eta}(f_1) + (1-a')\hat{\eta}(f_2)\,.$$

Arguing as in the proof of Lemma 5.2, it is clear that for each $i \in S$ we may take $f(i)$ to be a convex combination of $f_1(i)$ and $f_2(i)$. Let $\widetilde{f} \in \Pi_{SD}$ be such that for each $i \in S$, $\widetilde{f}(i) =$ either $f_1(i)$ or $f_2(i)$. Then under our hypotheses ((A5.20) of Section 5) one can show that $\hat{\eta}(\widetilde{f}) \in Z$. Now consider Z as a union of two closed line segments Z_1 and Z_2, Z_1 being the line segment between $\hat{\eta}(f_1)$ and $\hat{\eta}(f)$, and Z_2 that between $\hat{\eta}(f_2)$ and $\hat{\eta}(f)$. Let $\{f_n'\}$ be a sequence in Π_{SD}, defined as follows: $f_0' = f_1$, and

$$f_n'(i) = \begin{cases} f_2(i), & i \le n \\ f_1(i), & i > n. \end{cases}$$

Then by the above considerations $\hat{\eta}(f_n') \in Z$. Since $f_n' \to f_2$ as $n \to \infty$, we conclude that $\hat{\eta}(f_n') \to \hat{\eta}(f_2)$ (the map $f \mapsto \hat{\eta}(f)$ is continuous under (A5.19)). Thus, the sequence $\hat{\eta}(f_n)$, $n \ge 0$, starts in Z_1 and eventually moves into Z_2. Let n denote the first time this happens. Then either $\hat{\eta}(f_n') = \hat{\eta}(f)$ or $\hat{\eta}(f)$ is a convex combination of $\hat{\eta}(f_n)$ and $\hat{\eta}(f_{n-1})$. Since $f_n'(i) = f_{n-1}'(i)$ for $i \ne n$, the arguments employed in Lemma 5.2 show that we may take $f(i) =$ the Dirac measure at $f_n'(i)$ for $i \ne n$ and $f(n) =$ a suitable convex combination of Dirac measures at $f_1(n)$ and $f_2(n)$. We have established the following result.

Theorem 8.1. *Each extreme point of H corresponds to an $\hat{\eta}(f)$ such that $f \in \Pi_{SR}$ satisfies: for all but at most one i, $f(i)$ is a Dirac measure at some point of A. For the single remaining i, if any, $f(i)$ is a convex combination of two such Dirac measures.*

A variant of the above theorem leads to the following result [27].

Theorem 8.2. *The minimum of $\nu \mapsto \int c\,d\nu$ on $\widetilde{I}_R$ is attained at an $\hat{\eta}(f) \in \widetilde{I}_R$ where f is either deterministic or satisfies: There are states $i_1, \ldots, i_k \in S$ and positive integers $n_1, \ldots, n_k > 1$ such that f requires randomization among n_j values at state i_j, $1 \le j \le k$, requires no randomization for the remaining states, and $\sum_{i=1}^{k} n_i \le m$.*

Once this existence result is available, necessary conditions for optimality can be obtained from the standard Lagrange multiplier theory.

Theorem 8.3. *There exist λ_i, $\beta_i \ge 0$, $1 \le i \le k$ such that $\hat{\eta}(f)$ as in Theorem 8.2 minimizes*

$$\eta \mapsto F(\eta, \{\lambda_i\}, \{\beta_i\}) := \int c\,d\eta - \sum_{i=1}^{k} \lambda_i\big(b_i - \textstyle\int c_i\,d\eta\big) - \sum_{i=1}^{k} \beta_i\big(\textstyle\int c_i\,d\eta - a_i\big)$$

on I_R. Furthermore, if $\widetilde{I}_R$ has nonempty interior, the following saddle point property holds: for all $\overline{\lambda}_i, \overline{\beta}_i \ge 0$, $1 \le i \le k$, $\eta \in I_R$

$$F\big(\hat{\eta}(f), \{\overline{\lambda}_i\}, \{\overline{\beta}_i\}\big) \le F\big(\hat{\eta}(f), \{\lambda_i\}, \{\beta_i\}\big) \le F\big(\eta, \{\lambda_i\}, \{\beta_i\}\big)\,.$$

Remark 8.1. The result in Theorem 8.1 cannot be improved in general. Indeed, in [26] there is a counterexample to show the non-existence of an optimal $f \in \Pi_{SD}$ for the CMP with one constraint.

Remark 8.2. We have discussed the stable case only. Analogous results can be obtained for the near-monotone case (conditions similar to (A5.18)). For details, we refer to [25].

Remark 8.3. When the action set $\boldsymbol{A}$ is countable, analogous results are obtained in [1].

We next consider another multiobjective CMP with AC criterion. We have m cost functions $c_i \in C_b(\boldsymbol{S} \times \boldsymbol{A})$, $1 \le i \le m$. All cost functions are of equal importance and as a result, the optimality problem cannot be recast as a constrained one. Therefore, we directly deal with the optimality problem with a vector cost criterion. This has been studied in [47], [73].

Let I_R be compact. Consider the vector cost criterion

$$\left(\int c_1 \, d\hat{\eta}(f), \ldots, \int c_m \, d\hat{\eta}(f)\right), \qquad \hat{\eta}(f) \in I_R \,.$$

In general, there need not exist an $f \in \Pi_{SSR}$ that minimizes all of $\int c_i \, d\hat{\eta}(f)$ over I_R. This motivates the concept of Pareto optimality. An $f \in \Pi_{SSR}$ is said to be *Pareto optimal* if there does not exist any $\overline{f} \in \Pi_{SSR}$ for which $\int c_i \, d\hat{\eta}(\overline{f}) \le \int c_i \, d\hat{\eta}(f)$, $1 \le i \le m$, with inequality being strict for at least one i. Pareto optimality is clearly the minimal requirement for any reasonable notion of an optimal solution for the multiobjective problem with no priority among objectives. The Pareto optimal solutions can be characterized as follows.

Theorem 8.4. *Any $f \in \Pi_{SSR}$ which minimizes $\sum_{i=1}^{m} \lambda_i \int c_i \, d\hat{\eta}(f)$ for some $\lambda_i > 0$, $1 \le i \le m$, is Pareto optimal. Conversely, any Pareto optimal $\overline{f} \in \Pi_{SSR}$ minimizes the above functional for some choice of $\lambda_i \ge 0$, $1 \le i \le m$.*

Remark 8.4. Note that the converse is only partial since we have $\lambda_i \ge 0$ in place of $\lambda_i > 0$. It becomes exact if $\boldsymbol{S}$ and $\boldsymbol{A}$ are finite.

One often reduces a vector cost criterion as above to a scalar one by introducing a 'utility function'. One such case is that of finding the 'shadow minimum' for the problem of minimizing the vector cost $\nu \mapsto \left[\int c_1 \, d\nu, \ldots, \int c_m \, d\nu\right] \in \mathbb{R}^m$ on I_R. Letting L denote the range of this map, L can be shown to be closed and convex. Suppose $y_i^* = \min\left\{\int c_i \, d\nu : \nu \in I_R\right\}$, $1 \le i \le m$. Let $y^* = (y_1^*, \ldots, y_m^*)$. The point y^* is called the ideal (or utopian) point. The point $x^* \in L$ which is closest to y^* is called the *shadow minimum.* This point is unique and is characterized by

$$\langle y^* - x^*, z - x^* \rangle \le 0 \,, \qquad z \in \boldsymbol{S} \,.$$

For finite $\boldsymbol{S}$ and $\boldsymbol{A}$, a combined linear-quadratic program can find x^* explicitly [73]. The point x^* is easily seen to be Pareto optimal.

§9. Conclusions

We hope this survey has provided a useful presentation of the problems and techniques in average cost control of Markov processes. As is amply clear, there is not a globally applicable approach. Instead, one expects to build a library of special tricks, a collection of simple verifiable sufficient conditions under which the problem is accessible, possibly with different techniques. Going one step further, there are the more difficult partially observable and multiobjective problems. Though these have seen some significant results of late, there remains much more that eludes satisfactory analysis. A similar comment applies to computational aspects and adaptive control, two topics we have not touched upon here. For computational aspects we refer to [79], [85], [133], [176], and for adaptive control, [26], [80], [99]. Also, we have not dealt with the vast literature on *sensitive* optimality [133], [178], nor with some some other criteria, such as overtaking [108], variance sensitive [194], and weighted cost [58], [63], [96]. Finally, the discrete-time models have interesting applications to continuous-time problems, for which we refer to [14, Sect. 6.7], [106], [155], [202].

§10. Acknowledgements

The authors would like to thank the anonymous referee and the Associate Editor Steven E. Shreve for their constructive criticism and comments which helped to improve this paper. Our thanks also to Prof. Linn I. Sennott for pointing out an error in the original statement of Theorem 5.1. We were blessed by having an excellent typist, Joan Van Cleave, who rose above mere patience when faced with numerous revisions of our "final draft." Finally, E. Fernández-Gaucherand wishes to thank Prof. O. Hernández-Lerma of CINVESTAV-IPN, México, for useful discussions.

References

[1] E. Altman and A. Shwartz, *Markov decision problems and state-action frequencies*, SIAM J. Control Optim. **29** (1991), 786–809.

[2] V. A. Andriyanov, I. A. Kogan and G. A. Umnov, *Optimal control of a partially observable discrete Markov process*, Automat. & Remote Control **4** (1980), 555–561.

[3] M. Aoki, *Optimal control of partially observable Markovian systems*, J. Franklin Institute **280** (1965), 367–386.

[4] K. J. Arrow, T. Harris and J. Marshak, *Optimal inventory policy*, Econometrica **19** (1951), 250–272.

[5] K. J. Åström, *Optimal control of Markov processes with incomplete state information*, J. Math. Anal. Appl. **10** (1965), 174–205.

[6] ———, *Optimal control of Markov processes with incomplete state information, II. The convexity of the loss function*, J. Math. Anal. Appl. **26** (1969), 403–406.

[7] ———, *Stochastic control problems*, Mathematical Control Theory (W. A. Coppel, ed.), Lecture Notes in Mathematics, vol. 680, SpringerVerlag, Berlin, 1978, pp. 1–69.

[8] J. A. Bather, *Optimal decision procedures for finite Markov chains, I: Examples*, Adv. Appl. Prob. **5** (1973), 328–339; *II: Communicating systems*, Adv. Appl. Prob. **5** (1973), 521–540; *III: General convex systems*, Adv. Appl. Prob. **5** (1973), 541–553.

[9] R. Bellman, *A Markovian decision problem*, J. Math. Mech. **6** (1957), 679–684.

[10] ———, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[11] ———, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NJ, 1961.

[12] R. Bellman and D. Blackwell, *On a Particular Non-Zero Sum Game*, RM-250, RAND Corp., Santa Monica, CA, 1949.

[13] R. Bellman and J. P. La Salle, *On Non-Zero Sum Games and Stochastic Processes*, RM-212, RAND Corp., Santa Monica, CA, 1949.

[14] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, PrenticeHall, Englewood Cliffs, NJ, 1987.

[15] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[16] F. J. Beutler and K. W. Ross, *Optimal policies for controlled Markov chain with a constraint*, J. Math. Anal. Appl. **112** (1985), 236–256.

[17] R. N. Bhattacharya and M. Majumdar, *Controlled semi-Markov model under long-run average rewards*, J. Stat. Planning and Inference **22** (1989), 223–242.

[18] D. Blackwell, *Discrete dynamic programming*, Ann. Math. Statist. **33** (1962), 719–726.

[19] ———, *Discounted dynamic programming*, Ann. Math. Statist. **36** (1965), 226–235.

[20] V. S. Borkar, *Controlled Markov chains and stochastic networks*, SIAM J. Control Optim. **21** (1983), 652–666.

[21] ———, *On minimum cost per unit time control of Markov chains*, SIAM J. Control Optim. **22** (1984), 965–984.

[22] ———, *Control of Markov chains with long-run average cost criterion*, Stochastic Differential Systems, Stochastic Control Theory and Applications (W. Fleming and P. L. Lions, eds.), The IMA Volumes in Mathematics and Its Applications, vol. 10, SpringerVerlag, Berlin, 1988, pp. 57–77.

[23] ———, *A convex analytic approach to Markov decision processes*, Probab. Th. Rel. Fields **78** (1988), 583–602.

[24] ———, *Control of Markov chains with long-run average cost criterion: the dynamic programming equations*, SIAM J. Control Optim. **27** (1989), 642–657.

[25] ———, *Controlled Markov chains with constraints*, Proceedings of the Workshop on Recent Advances in Modelling and Control of Stochastic Systems, Bangalore, January 1991, to be published by Indian Academy of Sciences.

[26] ———, *Topics in Controlled Markov Chains*, Pitman Research Notes in Math. No. 240, Longman Scientific and Technical, Harlow, 1991.

[27] ———, *Ergodic control of Markov chains with constraints — the general case*, preprint (1991).

[28] V. S. Borkar and M. K. Ghosh, *Ergodic and adaptive control of nearest neighbour motions*, Math. Control Signals Systems **4** (1991), 81–98.

[29] L. D. Brown and R. Purves, *Measurable selection of extrema*, Annals of Statistics **1** (1973), 902–912.

[30] R. Cavazos-Cadena, *Necessary and sufficient conditions for a bounded solution to the optimality equation in average reward Markov decision chains*, Systems & Control Letters **10** (1988), 71–78.

[31] ———, *Necessary conditions for the optimality equation in average reward Markov decision processes*, Appl. Math. Optim. **19** (1989), 97–112.

[32] ———, *Recent results on conditions for the existence of average optimal stationary policies*, Annals of Operations Research **28** (1991), 3–26.

[33] ———, *A counterexample on the optimality equation in Markov decision chains with the average cost criterion*, Systems & Control Letters **16** (1991), 387–392.

[34] R. Cavazos-Cadena and L. I. Sennott, *Comparing recent assumptions for the existence of average optimal stationary policies*, Operations Research Letters (to appear).

[35] R. Ya. Chitashvili, *A controlled finite Markov chain with an arbitrary set of decisions*, Theory Prob. Applications **20** (1975), 839–846.

[36] E. V. Denardo and B. L. Fox, *Multichain Markov renewal programs*, SIAM J. Appl. Math. **16** (1968), 468–487.

[37] C. Derman, *On sequential decisions and Markov chains*, Management Science **9** (1962), 16–24.

[38] ———, *Denumerable state Markov decision processes — average cost criterion*, Ann. Math. Statist. **37** (1966), 1545–1553.

[39] ———, *Finite State Markovian Decision Processes*, Academic Press, 1970.

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

[41] C. Derman and R. E. Strauch, *A note on memoryless rules for controlling sequential control processes*, Ann. Math. Statist. **37** (1966), 276–278.

[42] C. Derman and A. F. Veinott, Jr., *A solution to a countable system of equations arising in Markovian decision processes*, Ann. Math. Statist. **38** (1967), 582–584.

[43] ———, *Constrained Markov decision chains*, Management Science **19** (1972), 389–390.

[44] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[45] A. W. Drake, *Observation of a Markov Process Through a Noisy Channel*, D. Sc. Thesis, Dept. of Electrical Engineering, MIT, 1962.

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

[47] S. Durinovic, H. M. Lee, M. N. Katehakis and J. A. Filar, *Multiobjective Markov decision process with average reward criterion*, Large Scale Systems **10** (1986), 215–226.

[48] A. Dvoretzky, J. Keifer and J. Wolfowitz, *The inventory problem*, Econometrica **20** (1956), 187–222; 450–466.

[49] E. B. Dynkin, *Controlled random sequences*, Theory Prob. Applications **10** (1965), 1–14.

[50] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[51] A. Federgruen, A. Hordijk and H. C. Tijms, *Recurrent conditions in denumerable state Markov decision processes*, Dynamic Programming and its Applications (M. L. Puterman, ed.), Academic Press, New York, 1978, pp. 3–22.

[52] ———, *Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion*, Stoch. Proc. Appl. **9** (1979), 223–235.

[53] A. Federgruen, P. J. Schweitzer and H. C. Tijms, *Denumerable undiscounted semi-Markov decision processes with unbounded rewards*, Mathematics of Operations Research **8** (1983), 298–313.

[54] A. Federgruen and H. C. Tijms, *The optimality equation in average cost denumerable state semi-Markov decision problems, recurrence conditions and algorithms*, J. Appl. Prob. **15** (1978), 356–373.

[55] E. A. Feinberg, *On controlled finite state Markov processes with compact control sets*, Theory Prob. Applications **20** (1975), 856–862.

[56] ———, *The existence of a stationary ε-optimal policy for a finite Markov chain*, Theory Prob. Applications **23** (1978), 297–313.

[57] ———, *An ε-optimal control of a finite Markov chain with an average reward criterion*, Theory Prob. Applications **25** (1980), 70–81.

[58] ———, *Controlled Markov processes with arbitrary numerical criteria*, Theory Prob. Applications **27** (1982), 486–503.

[59] E. Fernández-Gaucherand, *Controlled Markov Processes on the Infinite Planning Horizon: Optimal and Adaptive Control*, Ph. D. dissertation, Electrical and Computer Engineering Dept., University of Texas at Austin, 1991.

[60] E. Fernández-Gaucherand, A. Arapostathis and S. I. Marcus, *On partially observable Markov decision processes with an average cost criterion*, Proc. 28th IEEE Conf. on Decision and Control, Tampa, FL (1989), 1267–1272.

[61] ———, *On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes*, Annals of Operations Research **29** (1991), 439–470.

[62] ———, *Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes*, Systems & Control Letters **15** (1990), 425–432.

[63] E. Fernández-Gaucherand, M. K. Ghosh and S. I. Marcus, *Controlled Markov processes on the infinite planning horizon with a weighted cost criterion*, Proceedings of the IV Latin American Congress in Probability and Mathematical Statistics, México City, México (1990) (to appear).

[64] C. H. Fine, *A quality control model with learning effects*, Operations Research **36** (1988), 437–444.

[65] L. Fisher and S. M. Ross, *An example in denumerable decision processes*, Ann. Math. Statist. **39** (1968), 674–675.

[66] J. Flynn, *Averaging versus discounting in dynamic programming: a counterexample*, Annals of Statistics **2** (1974), 411–413.

[67] ———, *Conditions for the equivalence of optimality criteria in dynamic programming*, Annals of Statistics **4** (1976), 936–953.

[68] ———, *On optimality criteria for dynamic programs with long finite horizons*, J. Math. Anal. Appl. **76** (1980), 202–208.

[69] N. Furukawa, *Markovian decision processes with compact action spaces*, Ann. Math. Statist. **43** (1972), 1612–1622.

[70] J.-P. Georgin, *Contrôle des chaines de Markov sur des espaces arbitraires*, Ann. Inst. H. Poincaré **14, Sect. B** (1978), 255–277.

[71] ———, *Estimation et contrôle des chaines de Markov sur des espaces arbitraires*, Lecture Notes Math., No. 636, Springer-Verlag, Berlin, 1978, pp. 71–113.

[72] M. K. Ghosh, *Ergodic and Adaptive Control of Markov Processes*, Ph. D. Thesis, Indian Institute of Science, Bangalore, India, 1988.

[73] ———, *Markov decision processes with multiple costs*, Operations Research Letters **9** (1990), 257–260.

[74] M. K. Ghosh and S. I. Marcus, *Ergodic control of Markov chains*, Proc. 29th IEEE Conf. on Decision and Control, Honolulu, Hawaii (1990), 258–263.

[75] ———, *On strong average optimality of Markov decision processes with unbounded costs*, Operations Research Letters **11** (1992) (to appear).

[76] I. I. Gihman and A. V. Skorohod, *Controlled Stochastic Processes*, Springer-Verlag, New York, 1979.

[77] D. Gillette, *Stochastic games with zero stop probabilities*, Contributions to the Theory of Games, III, Annals of Math. Studies, No. 39, Princeton University Press, Princeton, NJ, 1957, pp. 71–187.

[78] L. G. Gubenko and E. S. Statland, *On controlled, discrete-time Markov decision processes*, Theory Probab. Math. Statist. **7** (1975), 47–61.

[79] M. Haviv and M. L. Puterman, *An improved algorithm for solving communicating average reward Markov decision processes*, Annals of Operations Research **29** (1991), 229–242.

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

[81] ———, *Average optimality in dynamic programming on Borel spaces: Unbounded costs and controls*, preprint (1990).

[82] O. Hernández-Lerma, J. C. Hennet and J. B. Lasserre, *Average cost Markov decision processes: optimality conditions*, J. Math. Anal. Appl. (to appear).

[83] O. Hernández-Lerma and J. B. Lasserre, *Average cost optimal policies for Markov control processes with Borel state space and unbounded costs*, Systems & Control Letters **15** (1990), 349–356.

[84] 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 **29** (1991), 29–46.

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

[86] C. J. Himmelberg, T. Parthasarathy and F. S. Van Vleck, *Optimal plans for dynamic programming problems*, Mathematics of Operations Research **1** (1976), 390–394.

[87] K. Hinderer, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameters*, Lect. Notes Oper. Res. Math. Syst., vol. 33, Springer-Verlag, Berlin, 1970.

[88] W. J. Hopp and S. C. Wu, *Multiaction maintenance under Markovian deterioration and incomplete information*, Naval Res. Logist. Quart. **35** (1988), 447–462.

[89] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Math. Centre Tract, No. 51, Mathematisch Centrum, Amsterdam, 1974.

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

[91] ———, *Constrained undiscounted stochastic dynamic programming*, Mathematics of Operations Research **9** (1984), 276–289.

[92] R. Howard, *Dynamic Programming and Markov Decision Processes*, MIT Press, Cambridge, MA, 1960.

[93] G. Hübner, *On the fixed points of the optimal reward operator in stochastic dynamic programming with discount factor greater than one*, Zeit. Angew. Math. Mech. **57** (1977), 477–480.

[94] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*, Math. Centre Tract, No. 148, Mathematisch Centrum, Amsterdam, 1983.

[95] S. Karlin, *The structure of dynamic programming models*, Naval Res. Logist. Quart. **2** (1955), 285–294.

[96] D. Krass, J. A. Filar and S. Sinha, *A weighted Markov decision process*, Operations Research (to appear).

[97] N. V. Krylov, *Construction of an optimal strategy for a finite controlled chain*, Theory of Prob. **10** (1965), 45–54.

[98] P. R. Kumar, *Simultaneous identification and adaptive control of unknown systems over finite parameter sets*, IEEE Trans. Automat. Control **AC-28** (1983), 68–76.

[99] ———, *A survey of some of results in stochastic adaptive control*, SIAM J. Control Optim. **23** (1985), 329–380.

[100] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliff, NJ, 1986.

[101] M. Kurano, *Markov decision processes with a Borel measurable cost function: the average case*, Mathematics of Operations Research **11** (1986), 309–320.

[102] ———, *The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin*, SIAM J. Control Optim. **27** (1989), 296–307.

[103] ———, *Average Cost Markov Decision Processes under the Hypothesis of Doeblin*, Report No. 9, Dept. Mathematics, Faculty of Education, Chiba, Japan, 1989.

[104] ———, *On Optimality Inequalities in Average Cost Markov Decision Processes with Doeblin's Conditions*, Report No. 1, Dept. Mathematics, Faculty of Education, Chiba, Japan, 1990.

[105] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[106] ———, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim. **28** (1990), 999–1048.

[107] B. L. Lamond and M. L. Puterman, *Generalized inverses in discrete time Markov decision processes*, SIAM J. Math. Anal. Appl. **10** (1989), 118–134.

[108] A. Leizarowitz, *Infinite horizon optimization for a finite state Markov chain*, SIAM J. Control Optim. **25** (1987), 1601–1618.

[109] G. de Leve, *Generalized Markov Decision Processes, Part I: Model and Method*, Math. Centre Tract No. 3, Mathematisch Centrum, Amsterdam, 1964.

[110] ———, *Generalized Markov Decision Processes, Part II: Probabilistic Background*, Math. Centre Tract No. 4, Mathematisch Centrum, Amsterdam, 1964.

[111] G. de Leve, A. Federgruen and H. C. Tijms, *A general Markov decision method*, Adv. Appl. Prob. **9** (1977), 296–335.

[112] S. A. Lippman, *Semi-Markov decision processes with unbounded rewards*, Management Science **19** (1973), 717–731.

[113] ———, *On dynamic programming with unbounded rewards*, Management Science **21** (1975), 1225–1233.

[114] M. Loéve, *Probability Theory II*, Springer-Verlag, Berlin, 1978.

[115] W. S. Lovejoy, *Some monotonicity results for partially observed Markov decision processes*, Operations Research **35** (1987), 736–743.

[116] D.-J. Ma, A. M. Makowski and A. Shwartz, *Estimation and optimal control for constrained Markov chains*, Proc. 25th IEEE Conf. on Decision and Control, Athens (1986), 994–999.

[117] A. Maitra, *Dynamic Programming for Countable State Systems*, Doctoral Thesis, University of California, Berkeley, 1964.

[118] ———, *Dynamic programming for countable state systems*, Sankhya, Ser. A **27** (1965), 241–248.

[119] ———, *Discounted dynamic programming on compact metric spaces*, Sankhya, Ser. A **30** (1968), 211–216.

[120] P. Mandl, *Estimation and control in Markov chains*, Adv. Appl. Prob. **6** (1974), 40–60.

[121] A. Manne, *Linear programming and sequential decisions*, Management Science **6** (1960), 259–267.

[122] A. Martin-Löf, *Existence of a stationary control for a Markov chain maximizing the average reward*, Operations Research **15** (1967), 866–871.

[123] B. L. Miller and A. F. Veinott, Jr., *Discrete dynamic programming with a small interest rate*, Ann. Math. Statist. **40** (1969), 366–370.

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

[125] P. Nain and K. W. Ross, *Optimal priority assignment with hard constraint*, IEEE Trans. Automat. Control **AC-31** (1986), 883–888.

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

[127] M. Ohnishi, H. Kawai and H. Mine, *An optimal inspection and replacement policy under incomplete state information*, European J. Oper. Res. **27** (1986), 117–128.

[128] M. Ohnishi, H. Mine and H. Kawai, *An optimal inspection and replacement policy under incomplete state information: average cost criterion*, Stochastic Models in Reliability Theory (S. Osaki and Y. Hatoyama, eds.), Lect. Notes Econ. Math. Syst., vol. 235, Springer-Verlag, Berlin, 1984, pp. 187–197.

[129] S. Orey, *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand, London, 1971.

[130] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[131] L. K. Platzman, *Finite Memory Estimation and Control of Finite Probabilistic Systems*, Ph. D. dissertation, Dept. of Electrical Engineering and Computer Science, MIT, 1977.

[132] ———, *Optimal infinite horizon undiscounted control of finite probabilistic systems*, SIAM J. Control Optim. **18** (1980), 362–380.

[133] M. L. Puterman, *Markov decision processes*, Handbooks in Operation Research and Management Science (D. P. Heyman and M. J. Sobel, eds.), vol. 2, North Holland, Amsterdam, 1990, pp. 331-434.

[134] D. Rhenius, *Incomplete information in Markovian decision models*, Annals of Statistics **2** (1974), 1327–1334.

[135] U. Rieder, *Measurable selection theorems for optimization problems*, Manuscripta Mathematica **24** (1978), 115–131.

[136] R. K. Ritt and L. I. Sennott, *Optimal stationary policies in general state Markov decision chains with finite action set*, Mathematics of Operations Research (to appear).

[137] D. R. Robinson, *Markov decision chains with unbounded costs and applications to the control of queues*, Adv. Appl. Prob. **8** (1976), 159–176.

[138] ———, *Optimality conditions for a Markov decision chain with unbounded costs*, J. Appl. Prob. **17** (1980), 996–1003.

[139] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[140] Z. Rosberg, P. Varaiya and J. Walrand, *Optimal control of service in tandem queues*, IEEE Trans. Automat. Control **AC-27** (1982), 600–610.

[141] K. W. Ross, *Randomized and past dependent policies for Markov decision processes with multiple constraints*, Operations Research **37** (1989), 474–477.

[142] K. W. Ross and R. Varadarajan, *Markov decision processes with sample path constraints: the communicating case*, Operations Research **37** (1989), 780–790.

[143] S. M. Ross, *Non-discounted denumerable Markovian decision models*, Ann. Math. Statist. **39** (1968), 412–423.

[144] ———, *Arbitrary state Markovian decision processes*, Ann. Math. Statist. **39** (1968), 2118–2122.

[145] ———, *Quality control under Markovian deterioriation*, Management Science **17** (1971), 587–596.

[146] ———, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.

[147] Y. Sawaragi and T. Yoshikawa, *Discrete time Markovian decision processes with incomplete state observation*, Ann. Math. Statist. **41** (1970), 78–86.

[148] M. Schäl, *On continuous dynamic programming with discrete time parameters*, Z. Wahrsch. Verw. Gebiete **21** (1972), 279–288.

[149] ———, *On dynamic programming: compactness of the space of policies*, Stoch. Proc. Appl. **3** (1975), 345–364.

[150] ———, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie verw. Gebiete **32** (1975), 179–196.

[151] L. I. Sennott, *A new condition for the existence of optimal stationary policies in average cost Markov decision processes*, Operations Research Letters **5** (1986), 17–23.

[152] ———, *Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs*, Operations Research **37** (1989), 626–633.

[153] ———, *Average cost semi-Markov decision processes and the control of queueing systems*, Probab. in Eng. & Info. Sci. **3** (1989), 247–272.

[154] ———, *The average cost optimality equation and critical number policies*, preprint (1991).

[155] R. F. Serfozo, *An equivalence between continuous and discrete time Markov decision processes*, Operations Research **27** (1979), 616–620.

[156] L. Shapley, *Stochastic games*, Proc. Nat. Acad. Sci. USA **39** (1953), 1095–1100.

[157] A. N. Shiryaev, *On the theory of decision functions and control by an observation process with incomplete data*, Selected Translations in Mathematical Statistics and Probability, vol. 6, American Mathematical Society, Providence, RI, 1966, pp. 162–188.

[158] ———, *Some new results in the theory of controlled random sequences*, Selected Translations in Mathematical Statistics and Probability, vol. 8, American Mathematical Society, Providence, RI, 1970, pp. 49–130.

[159] ———, *On Markov sufficient statistics in non-additive Bayes problems of sequential analysis*, Theory Prob. Applications **9** (1964), 604–618.

[160] S. E. Shreve and D. P. Bertsekas, *Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control*, SIAM J. Control Optim. **16** (1978), 953–978.

[161] ———, *Dynamic programming in Borel spaces*, Dynamic Programming and its Applications (M. L. Puterman, ed.), Academic Press, New York, 1978, pp. 115–130.

[162] A. Shwartz and A. M. Makowski, *An optimal adaptive scheme for two competing queues with constraints*, Analysis and Optimization of Systems, Lecture Notes on Control and Information Sciences (A. Bensoussan and J. L. Lions, eds.), Springer-Verlag, Berlin, 1986.

[163] ———, *On the Poisson Equation for Markov Chains*, Report No. EE-646, Faculty of Electrical Engineering, Technion: Israel Institute of Technology, 1987.

[164] ———, *Comparing policies in Markov decision processes: Mandl's lemma revisited*, Mathematics of Operations Research **15** (1990), 155–174.

[165] R. D. Smallwood and E. J. Sondik, *The optimal control of partially observable Markov process over a finite horizon*, Operations Research **21** (1973), 1071–1088.

[166] E. J. Sondik, *The Optimal Control of Partially Observable Markov Processes*, Ph. D. dissertation, Electrical Engineering Dept., Stanford University, 1971.

[167] ———, *The optimal control of partially observable Markov decision problems over the infinite horizon: discounted costs*, Operations Research **26** (1978), 282–304.

[168] S. S. Stidham Jr. and R. R. Weber, *Monotonic and insensitive optimal policies for control of queues with unbounded costs*, Operations Research **37** (1989), 611–625.

[169] R. E. Strauch, *Negative dynamic programming*, Ann. Math. Statist. **37** (1966), 871–890.

[170] C. Striebel, *Sufficient statistics in the optimum control of stochastic systems*, J. Math. Anal. Appl. **12** (1965), 576–593.

[171] ———, *Optimal Control of Discrete Time Stochastic Systems*, Lect. Notes Econ. Math. Syst., vol. 110, Springer-Verlag, Berlin, 1975.

[172] R. Sznjder and J. A. Filar, *Some comments on a theorem of Hardy and Littlewood*, J. Optimiz. Th. & Appl. **75** (1992) (to appear).

[173] H. M. Taylor, *Markovian sequential replacement processes*, Ann. Math. Statist. **38** (1965), 1677–1694.

[174] L. C. Thomas, *Connectedness conditions for denumerable state Markov decision processes*, Recent Developments in Markov Decision Processes (R. Hartley, L. C. Thomas and D. F. White, eds.), Academic Press, New York, 1980, pp. 181–204.

[175] H. C. Tijms, *On Dynamic Programming with Arbitrary State Space, Compact Action Space and the Average Reward as Criterion*, Report BW 55/75, Mathematisch Centrum, Amsterdam, 1975.

[176] ———, *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley, Chichester, 1986.

[177] J. Van der Wal and J. Wessels, *Markov decision processes*, Statist. Neerlandica **39** (1985), 219–233.

[178] A.F. Veinott, *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist. **40** (1969), 1635–1660.

[179] O. V. Viskov and A. N. Shiryaev, *On controls leading to optimal stationary models*, Trudy Mat. Inst. Steklov Akad. Nauk SSSR **71** (1964), 35–45, (Russian).

[180] D. H. Wagner, *Survey of measurable selection theorems*, SIAM J. Control Optim. **15** (1977), 859–903.

[181] H. M. Wagner, *On the optimality of pure strategies*, Management Science **6** (1960), 268–269.

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

[183] ———, *Statistical Decision Functions*, Wiley, New York, 1950.

[184] R. Wang, *Optimal replacement policy with unobservables states*, J. Appl. Prob. **14** (1977), 340–348.

[185] ———, *Computing optimal quality control policies — two actions*, J. Appl. Prob. **14** (1977), 826–832.

[186] R. R. Weber and S. S. Stidham Jr., *Optimal control of service rates in networks of queues*, Adv. Appl. Prob. **15** (1987), 202–218.

[187] C. C. White, *A Markov quality control process subject to partial observation*, Management Science **23** (1977), 843–852.

[188] ———, *Optimal inspection and repair of a production process subject to deterioration*, J. Oper. Res. Soc. **29** (1978), 235–243.

[189] ———, *Bounds on optimal cost for a replacement problem with partial observation*, Naval Res. Logist. Quart. **26** (1979), 415–422.

[190] ———, *Optimal control — limit strategies for a partially observed replacement problem*, Internat. J. Systems Science **10** (1979), 321–331.

[191] ———, *Monotone control laws for noisy, countable-state Markov chains*, European J. Oper. Res. **5** (1980), 124–132.

[192] C. C. White and D. J. White, *Markov decision processes*, European J. Oper. Res. **39** (1989), 1–16.

[193] D. J. White, *Dynamic programming of Markov chains and the method of successive approximations*, J. Math. Anal. Appl. **6** (1963), 373–376.

[194] ———, *Mean, variance, and probabilistic criteria in finite Markov decision processes: a review*, J. Optimiz. Th. & Appl. **56** (1988), 1–29.

[195] P. Whittle, *Sequential decision processes with essential unobservables*, Adv. Appl. Prob. **1** (1969), 271–287.

[196] ———, *Optimization over Time: Dynamic Programming and stochastic control, II*, John Wiley & Sons, Chichester, 1983.

[197] J. Wijngaard, *Stationary Markovian decision problems and perturbation theory of quasicompact linear operators*, Mathematics of Operations Research **2** (1977), 91–102.

[198] ———, *Existence of average optimal strategies in Markovian decision problems with strictly unbounded costs*, Dynamic Programming and its Applications (M. L. Puterman, ed.), Academic Press, New York, 1978, pp. 369–386.
[199] K. Yamada, *Duality theorem in Markovian decision problems*, J. Math. Anal. Appl. **50** (1975), 579–595.
[200] A. A. Yushkevich, *On a class of strategies in general Markov decision models*, Theory Prob. Applications **18** (1973), 777–779.
[201] ———, *Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces*, Theory Prob. Applications **21** (1976), 153–158.
[202] ———, *On reducing a jump controllable Markov model to a model with discrete time*, Theory Prob. Applications **25** (1980), 58–59.
[203] A. A. Yushkevich and R. Ya. Chitashvili, *Controlled random sequences and Markov chains*, Russian Math. Surveys **37** (1982), 239–274
[204] H. Zijm, *The optimality equations in multichain denumerable Markov decision processes with average cost criterion: The bounded cost case*, Statistics and Decisions **3** (1985), 143–165.

APPENDIX

Multifunctions and measurable selectors. Let $\boldsymbol{V}$ and $\boldsymbol{W}$ denote non-empty Borel spaces, and let $2^{\boldsymbol{W}}$ denote the collection of all *nonempty* subsets of $\boldsymbol{W}$. A *multifunction* (or set-valued function) Φ from $\boldsymbol{V}$ to $\boldsymbol{W}$ is a map $\Phi : \boldsymbol{V} \to 2^{\boldsymbol{W}}$. The subset $\mathrm{Dom}(\Phi) := \{v \in \boldsymbol{V} : \Phi(v) \neq \varnothing\}$ is called the *domain* of Φ. When $\mathrm{Dom}(\Phi) = \boldsymbol{V}$ we say that the map Φ is *strict*. In what follows we assume that Φ is a strict multifunction. If, for each $v \in \boldsymbol{V}$, $\Phi(v)$ is a compact (closed, measurable) subset of $\boldsymbol{W}$, then Φ is said to be *compact (closed, measurable)-valued*. A *selector* (or selection) of Φ is a function $\varphi : \boldsymbol{V} \to \boldsymbol{W}$ such that $\varphi(v) \in \Phi(v)$, for all $v \in \mathrm{Dom}(\Phi)$. The set of (Borel) measurable selectors of Φ will be denoted by $\mathcal{S}(\Phi)$. The *graph* of Φ, denoted by $\mathrm{Graph}(\Phi)$, is defined as

$$\mathrm{Graph}(\Phi) := \{(v, w) : v \in \boldsymbol{V},\ w \in \Phi(v)\}.$$

For a set $W \in 2^{\boldsymbol{W}}$, we define

$$\Phi^{-1}[W] := \{v \in \boldsymbol{V} : \Phi(v) \cap W \neq \varnothing\},$$

and we say that Φ is (Borel) *measurable* if $\Phi^{-1}[B] \in \boldsymbol{\mathcal{B}}(\boldsymbol{V})$, for each *closed* subset B of $\boldsymbol{W}$. If Φ is *closed-valued*, then measurability of Φ implies that $\mathrm{Graph}(\Phi) \in \boldsymbol{\mathcal{B}}(\boldsymbol{V} \times \boldsymbol{W})$; furthermore, if Φ is *compact-valued*, then the converse also holds [86], [180, Th. 4.2]. The multifunction Φ is called *upper semicontinuous* if for every $v \in \boldsymbol{V}$ and every open set $G \supset \Phi(v)$ there exists a neighborhood N of v such that $\Phi(v') \subset G$, for all $v' \in N$; it is called *lower semicontinuous* if for every $v \in \boldsymbol{V}$ and every open set G such that $G \cap \Phi(v) \neq \varnothing$, $\Phi^{-1}(v)$ contains an open neighborhood of v. Also, Φ is said to be *continuous* if it is both upper and lower semicontinuous.

The following result, in different variations, has been shown by several authors [15, Sect. 7.5], [46, Lemma 6, p. 38], [50, Chap. 2], [86], [150], and summarized also in [80], [180, Th. 9.1].

Theorem A.1. *Let Φ be a compact-valued, measurable, strict multifunction from $\boldsymbol{V}$ to $\boldsymbol{W}$. Let $f : \mathrm{Graph}(\Phi) \to \mathbb{R}$ be a measurable function, such that for each $v \in \boldsymbol{V}$, $f(v, \cdot)$*

is lower semicontinuous on $\Phi(v)$. Then there exists a measurable selector $\varphi^* \in \mathcal{S}(\Phi)$ such that

$$f(v,\varphi^*(v)) = \min_{w\in\Phi(v)}\{f(v,w)\}, \qquad \forall\, v \in \boldsymbol{V}.$$

Let $f^* : \boldsymbol{V} \to \mathbb{R}$, defined by $f^*(v) := f\big(v,\varphi^*(v)\big)$. If Φ is upper semicontinuous and f is bounded below, then $f^* \in \mathcal{L}(\boldsymbol{V})$. Also, if Φ is continuous and $f \in C_b(\boldsymbol{V}\times\boldsymbol{W})$, then $f^* \in C_b(\boldsymbol{V})$.

A Tauberian theorem. The following Tauberian theorem plays a very important role in the analysis of the average cost criterion. For its proof, which is very hard to locate in the literature in this particular format, we refer the reader to [172].

Theorem A.2. *Let $\{a_n\}$ be a sequence of nonnegative numbers and $\beta \in (0,1)$. Then*

$$\liminf_{N\to\infty}\frac{1}{N}\sum_{m=0}^{N-1} a_m \le \liminf_{\beta\uparrow 1}(1-\beta)\sum_{n=0}^{\infty}\beta^n a_n$$
$$\le \limsup_{\beta\uparrow 1}(1-\beta)\sum_{n=0}^{\infty}\beta^n a_n \le \limsup_{N\to\infty}\frac{1}{N}\sum_{m=0}^{N-1} a_m .$$