UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS

Ecuaciones de Optimalidad Para el Criterio del Costo Promedio Sensible al Riesgo en Procesos de Decisión Markovianos Sobre un Espacio Finito

Tesis

Presentada por

ALFREDO ALANÍS DURÁN

Como Requisito Parcial
Para Obtener el Grado de

DOCTOR EN CIENCAS

CON

ORIENTACIÓN EN MATEMATICAS

San Nicolás de los Garza, Nuevo León, MÉXICO
Noviembre de 2013

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS

An Optimality System for the Risk-Sensitive Average Cost Criterion in Markov Decision Chains on a Finite State Space

A Dissertation Presented

by

ALFREDO ALANÍS DURÁN

In Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF SCIENCE

IN

MATHEMATICS

San Nicolás de los Garza, Nuevo León, MÉXICO
November, 2013

Declaration

I hereby confirm that this is my own work, and that the use of all material from other sources has been properly and fully acknowledged.

Alfredo Alanís Durán

An Optimality System for the Risk-Sensitive Average Cost Criterion in Markov Decision Chains on a Finite State Space

A Dissertation Presented

by

ALFREDO ALANÍS DURÁN

Approved by:

Dr. Rolando Cavazos-Cadena, Advisor
Universidad Autónoma Agraria Antonio Narro

Dra. María Aracelia Alcorta García, Advisor
Universidad Autónoma de Nuevo León

Dr. José Paz Pérez Padrón
Universidad Autónoma de Nuevo León

Dr. Francisco Javier Almaguer Martínez
Universidad Autónoma de Nuevo León

Dr. José Raúl Montes-de-Oca Machorro
Universidad Autónoma Metropolitana–Iztapalapa

An Optimality System for the Risk-Sensitive Average Cost Criterion in Markov Decision Chains on a Finite State Space

A Dissertation Presented

by

ALFREDO ALANÍS DURÁN

Abstract

This work concerns Markov decision chains endowed with the risk-sensitive average cost criterion, and the main goals are to characterize the optimal value function and to determine an optimal stationary policy. The exposition begins in Chapter 1 where the notion of Markov decision chain is introduced, and the ideas of risk-aversion and risk–sensitivity coefficient are briefly discussed. After this point, the risk-sensitive average cost criterion is formulated, and the main objectives are formally stated. Next, in Chapter 2 a fundamental theorem by Howard and Matheson (1972), as well as a recent extension on the characterization of the optimal average cost in terms of a single optimality equation are analyzed; such results require that, under the action of any stationary policy, every state can be visited with positive probability regardless of the initial sate, and the arguments used in this work emphasize the central role of that communication property. The presentation continues in Chapter 3 studying a recent theorem on the existence of solutions to the optimality equation for 'small'values of the risk-sensitivity coefficient, which was derived under the assumption that there exists a state that can be always reached with positive probability under the action of any stationary policy; the derivation presented in this work highlights the fundamental role of such an accessibility condition. The conclusions in Chapters 2 and 3 provide the motivation to pursue the main objective of this thesis, namely, *to establish a characterization of the optimal risk-sensitive average cost without imposing any condition on the structure of the transition law of the model.* This goal is achieved in Chapter 4, where the optimal risk-sensitive average cost function is characterized for general controlled Markov chains with finite state space and compact action sets, a result that is the *main contribution* of this thesis. It is supposed that the decision maker is risk-averse with constant risk-sensitivity coefficient and, under standard continuity–compactness conditions, it is proved that the (possibly non-constant) optimal value function is characterized by *a nested system of equations*, generalizing the conclusions s presented in the previous chapters, which require communication conditions on the transition law; moreover, it is shown that an optimal stationary policy can be derived form a solution of that system, and that the optimal superior and inferior limit average cost functions coincide. The approach used to obtain the main conclusions relies on the *discounted method* which, roughly, consists in using a family of contractive operators whose fixed points are used to approximate the optimal average index, to partition the state space in a family of equivalence classes, to determine a class of admissible actions at each state, and to construct a solution of a 'reduced' optimality equation on each equivalence class; the presentation of these results is based on the recent paper Alanís Durán and Cavazos-Cadena (2012). Finally, the exposition concludes in Chapter 5 with a retrospective view of the material presented in this work, and with the statement of two open problems concerning the extension of some of the conclusions in this work to models with denumerable state space.

Dedication

El esfuerzo detrás de esta tesis ha estado dedicado a mi familia, por su apoyo constante y siempre generoso. De manera especial, *a Lilia*, mi esposa, que no se cansa de empujar a todos en el sentido positivo para que se preparen y obtengan un grado superior al que tienen, *a mis hijos*, Lily, Alfredo (Sarahí) y Cecy, *a mis dos nietos*, el inquieto Elías y la tranquila de Ángela, y *a los que están por venir*.

También dedico este trabajo a mis compañeros de la Facultad de Ciencias Físico-Matemáticas, así como a todos los que hicieron posible llevar a buen término esta aventura que inició hace como tres años; bueno, comenzó en el CINVESTAV hace ya algún tiempo, y culmina ahora en el Postgrado de nuestra Facultad.

Alfredo Alanís Durán

Acknowledgement

Deseo reconocer al Doctor Rolando Cavazos Cadena, por su apoyo y guía para iniciar y concluir este trabajo basado en su experiencia y conocimiento en el campo de los Procesos Markovianos, así como a la Doctora Aracelia Alcorta, por su apoyo incondicional durante el programa de Doctorado, particularmente en las unidades de aprendizaje en el Postgrado.

Agradezco a la Facultad de Ciencias Físico-Matemáticas, por el decidido apoyo que recibí a través de la beca interna y la descarga académica durante la realización de mis estudios, en particular, a la MA Patricia Martínez Moreno, directora de nuestra Facultad.

Alfredo Alanís Durán

Contents

Chapter 1

General Perspective

This chapter presents an overview of this work. The notion of Markov decision chain is introduced, and the ideas of risk-aversion and risk–sensitivity coefficient are briefly discussed. After this point, the risk-sensitive average cost criterion is formulated and the main problem studied in this thesis is stated. The presentation concludes with an outline of the material in the following chapters.

1.1. Introduction

This work concerns discrete-time Markov decision chains, which are mathematical models for dynamical systems whose state X_t is observed at times $t = 0, 1, 2, 3, \ldots.$ by a decision maker (controller). After observing the state X_t at time t, the controller attempts to influence the evolution of the system by applying an action A_t, and such an intervention has two consequences: (i) A cost $C(X_t, A_t)$ is incurred, and (ii) regardless of the previous states and actions, the pair (X_t, A_t) determines the *probability distribution* of the state X_{t+1} to be observed at time $t + 1$. The rule (policy) π used by the controller to choose the actions determines the distribution of the cost process $\{C(X_t, A_t)\}$, and the performance of π is measured by an index (criterion) $J(x, \pi)$, which depends on the initial state $X_0 = x$ and involves the cost stream. The goal of the controller is to determine and apply a policy π^* satisfying

$$J(x, \pi^*) = \inf_\pi J(x, \pi) =: J^*(x)$$

for every initial state x; $J^*(\cdot)$ is the optimal value function and π^* is an optimal policy. In this work it is supposed that the system evolves on a finite state space, and the performance index is based on the assumption that the controller is *risk-averse*, that is, when facing a random cost Y, the decision maker is willing to pay a constant amount larger than the expectation $E[Y]$ in order to avoid the uncertain cost Y. Under a technical assumption on the risk-aversion of the controller, the performance criterion $J(x, \pi)$ considered in the subsequent development is given by *the risk-sensitive average cost per unit of time*, and the *main problem* studied in this thesis can be stated as follows:

> *To establish a characterization of the optimal risk-sensitive average cost function allowing to determine an optimal policy.*

The diverse ideas used to formulate this goal, together with relevant facts already available in the literature, will be analyzed in the remainder of the chapter, which is organized as follows: In Section 2 the notion of Markov decision chain is briefly discussed, whereas Section 3 is concerned with the ideas of risk-aversion and risk-sensitivity coefficient. Next, in Section

4 the risk-sensitive average cost criterion is specified and the main problem of the thesis is formally stated. The presentation concludes in Section 5 with an outline of the content of the following chapters.

1.2. Decision Model

A *discrete* Markov decision chain is specified as

$$\mathcal{M} = (S, A, \{A(x)\}, P, C),$$

where

(i) The *state space* S is a (nonempty) denumerable set endowed with the discrete topology;

(ii) The *action set* A is a metric space;

(iii) For each $x \in S$, $A(x) \subset A$ is the nonempty class of admissible actions (controls) at x;

(iv) The cost function C is a real-valued mapping defined on the set $\mathbb{K}$ of admissible pairs, which is given by

$$\mathbb{K} := \{(x, a) \,|\, a \in A(x),\ x \in S\}, \tag{1.2.1}$$

and

(v) $P = [p_{x\,y}(a) \,|(x, a) \in \mathbb{K},\ y \in S\,]$ is the controlled transition law, where

$$\sum_{y \in S} p_{x\,y}(a) = 1, \quad (x, a) \in \mathbb{K}.$$

The interpretation of this model $\mathcal{M}$ is as follows: At each time $t = 0, 1, 2, 3, \ldots$, a decision maker observes the state of a dynamical system evolving on S, say $X_t = x \in S$, and selects an admissible action $A_t = a \in A(x)$ to be applied to the system. Then, a cost $C(X_t, A_t) = C(x, a)$ is incurred and, regardless of the states observed and the actions applied before t, the state of the system at time $t+1$ will be $X_{t+1} = y$ with probability $p_{x\,y}(a) = p_{X_t\,y}(A_t)$; this is the Markov property of the descision process.

For each $t \geq 0$, $\mathbb{H}_t$ stands for the space of possible histories of the decision process up to time t, where $\mathbb{H}_0 := S$ and $\mathbb{H}_t := \mathbb{K}^t \times S$ when $t \geq 1$, and a generic element of $\mathbb{H}_t$ is denoted by $\mathbf{h}_t$, so that

$$\mathbf{h}_0 = x_0, \quad \text{and} \quad \mathbf{h}_t = (x_0, a_0, x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t), \quad t \geq 1, \tag{1.2.2}$$

where $x_t \in S$ and $(x_i, a_i) \in \mathbb{K}$. A *policy* π is a (possibly randomized) rule for choosing actions, and at each time t the action applied may depend on the whole observed history $\mathbf{h}_t$ up to time t. Formally, a policy is a sequence $\pi = \{\pi_t\}$, where π_t is a stochastic kernel on S given $\mathbb{H}_t$, that is,

(a) For each $\mathbf{h}_t \in \mathbb{H}_t$, $\pi_t(\cdot|\mathbf{h}_t)$ is a probability measure on the Borel σ-field of the action set set A, and is concentrated on the set of admissible actions at x_t, *i.e.*,

$$\pi_t(A(x_t)|\mathbf{h}_t) = 1,$$

and

(b) For each Borel subset B of the action space A, the mapping $\mathbf{h}_t \mapsto \pi_t(B|\mathbf{h}_t)$ is (Borel) measurable on the set $\mathbb{H}_t$.

When the controller chooses actions according to π and $\mathbf{h}_t$ is the observed history up to time t, the probability of applying a control A_t belonging to B is given by $\pi_t(B|\mathbf{h}_t)$; *the class of all policies is denoted by* $\mathcal{P}$. Define

$$\mathbb{F} := \prod_{x \in S} A(x), \tag{1.2.3}$$

a set that is naturally identified with the class of all functions $f\colon S \to A$ satisfying that $f(x) \in A(x)$ for every $x \in S$. A policy π is stationary if there exists $f \in \mathbb{F}$ such that, for every nonnegative integer t and $\mathbf{h}_t \in \mathbb{H}_t$, the probability measure $\pi_t(\cdot|\mathbf{h}_t)$ is concentrated at the point $f(x_t)$, *i.e.*, $\pi_t(\{f(x_t)\}|\mathbf{h}_t) = 1$; in this case π and f are naturally identified and, with this convention, $\mathbb{F} \subset \mathcal{P}$.

Given an initial state $X_0 = x \in S$ and the policy $\pi \in \mathcal{P}$ being used by the controller, the distribution of the state-action process $\{(X_t, A_t)\}$ is uniquely determined by the Ionescu-Tulcea theorem (Hinderer 1970, Hernández-Lerma 1989, Puterman 2005); such a distribution is denoted by P_x^π, whereas E_x^π stands for the corresponding expectation operator. The distribution P_x^π has the following *Markov property*:

For each $B \subset S$ and $t = 0, 1, 2, 3, \ldots$,

$$P_x^\pi[X_{t+1} \in B|X_i, A_i,\, i < t, X_t = x, A_t = a] = \sum_{y \in B} p_{x\,y}(a),$$

and then

$$P_x^\pi[X_{t+1} \in B|X_i, A_i,\, i < t, X_t] = \int_{A(X_t)} \left[\sum_{y \in B} p_{X_t,y}(a)\right] \pi_t(da|X_i, A_i, i < t, X_t); \quad (1.2.4)$$

applying this relation to a stationary policy $f \in \mathbb{F}$, it follows that

$$P_x^f[X_{t+1} \in B|X_i, A_i,\, i < t, X_t] = \sum_{y \in B} p_{X_t,y}(f(X_t)),$$

so that, under f, the state process $\{X_t\}$ is a Markov chain with time-invariant transition matrix $[p_{x\,y}(f(x)]_{x,y \in S}$.

The selection of an 'appropriate' policy π requires a criterion $V(x, \pi)$ measuring the performance of π when the initial state is $x \in S$; such a performance index can be thought of as 'a one-number summary' of the cost process $\{C(X_t, A_t)\}$. Once $V(x, \pi)$ has been specified, the objective of the controller is to use an *optimal policy* π^*, that is, a policy π^* satisfying

$$V(x, \pi^*) = V^*(x), \quad x \in S,$$

where the optimal value function $V^*(\cdot)$ is given by

$$V^*(x) = \inf_{\pi \in \mathcal{P}} V(x, \pi), \quad x \in S.$$

Examples of performance criteria are

(i) The total expected cost over a (possibly random) decision horizon T, which is given by

$$V(x, \pi) = E_x^\pi\left[\sum_{t=0}^{T-1} C(X_t, A_t)\right]; \quad (1.2.5)$$

(ii) The total discounted cost, defined by

$$V(x, \pi) = E_x^\pi\left[\sum_{t=0}^{\infty} \alpha^t C(X_t, A_t)\right], \quad (1.2.6)$$

where $\alpha \in (0, 1)$ is the discount factor, and

(iii) The expected average criterion specified as follows:

$$V(x,\pi) = \limsup_{m\to\infty} \frac{1}{m} E_x^\pi \left[\sum_{t=0}^{m-1} C(X_t, A_t) \right]. \tag{1.2.7}$$

Applications of Markov decision chains endowed with these (and other) criteria include a wide variety of areas, as machinary replacement, inventory management, economic growth, control of queues, fisheries and water reservoirs management, selling problems and transport networks; see, for instance, Ross (1992), Sennott (1996), Tijms (2003), Puterman (2005), or Bertsekas (2007, 2007a). In particular, the discounted criterion is widely used in economic models since, setting $\alpha = 1/(1+\rho)$ where ρ is the interest rate payed by a risk-less asset in a unit of time, the discounted criterion represents the value at time $t = 0$ of the accumulated cost that will be incurred in the future (Stokey and Lucas, 1989). Among the three criteria formulated above, the average cost index is mathematically the most challenging, since it involves the ergodic behavior of the state-action process $\{(X_t, A_t)\}$, and its analysis is based on recurrence properties of Markov processes as presented, for instance, in Loève (1977) or Billinglsley (1995) for the case of a discrete state space, and in Nummelin (2004) or Meyn (2009) for systems evolving on Borel spaces. A profound treatment of Markov decision chains endowed with diverse criteria, including the three performance indexes considered above, is contained in Hernández-Lerma (1988), Hernández-Lerma and Lasserre (1996, 1999) and Bertsekas and Shreve (1996).

The performance criterion analyzed in this work is a variant of the average index in (1.2.7), which is constructed by considering the *risk-sensitivity* of the controller before a random cost, a notion that is discussed below.

1.3. Risk-Sensitivity

The idea behind the performance criteria in (1.2.5)–(1.2.7) is that the controller values a random cost Y as much as the expectation $E[Y]$, so that the decision maker will be indifferent between paying the fixed amount $E[Y]$, or paying the uncertain cost Y. However, it is not difficult to visualize situations where the attitude before a random cost is different. For instance, consider the owner of an expensive new car paying \$500 for an insurance policy guaranteeing that, in case of a crash in the next year, he/she will receive an identical brand new vehicle. The cost of the car is \$300,000 and the owner feels that there is a small probability equal to 0.0001 of participating in a crash. What the owner foresees for the next year, is a random cost Y that can take the values \$0 and \$300,000 with probabilities 0 and 0.0001, respectively, so that $E[Y] = \$30$; however, \$500 were gladly paid to avoid facing the random cost Y, indicating that Y is assessed higher than its expectation $E[Y]$. Under certain assumptions on the behavior of a decision maker—the Von Neumann and Morgestern's rationality axioms—it can be proved that a random cost Y will be assessed using a *utility function* U, in such a way that Y will be valued as $E[U(Y)]$ (Berger, 2010). In this case, the real number

$$\mathcal{E} \equiv \mathcal{E}(Y) \tag{1.3.1}$$

satisfying

$$U(\mathcal{E}) = E[U(Y)] \tag{1.3.2}$$

is referred to as the *certain equivalent* of Y, and the controller will be indifferent between paying the certain amount $\mathcal{E}$ or incurring the random cost Y; also, when an offer of paying a fixed amount c to avoid the random cost Y is presented to the decision maker, the offer will be accepted if $c \leq \mathcal{E}(Y)$, and will be refused when $c > \mathcal{E}(Y)$. In the situation considered above, the random cost Y was avoided by the owner of the car paying \$500, so that $\mathcal{E}(Y) \geq$ \$500. Notice that the certain equivalent $\mathcal{E}(Y)$ is well-defined when Y is bounded and U is continuous and strictly increasing, properties that are supposed in the following discussion without explicit reference. On the other hand, since the relation (comparison) between two expected utilities $E[U(Y)]$ and $E[U(Y_1)]$ does not change when U is replaced by $\tilde{U} = aU + b$ where $a > 0$, it follows that *the utility function of the controller is determined up to an affine transformation* with positive slope.

A decision maker is referred to as

(i) *risk-neutral* if $\mathcal{E}(Y) = E[Y]$ always holds,

(ii) *risk-averse* if $\mathcal{E}(Y) > E[Y]$ when Y is a non-constant random variable, and

(iii) *risk-seeking* if $\mathcal{E}(Y) < E[Y]$ occurs for any non-constant random cost Y.

The attitude of the agent in the preceding paragraph is consistent with idea of risk-averse controller, since $\mathcal{E}(Y) \geq \$500 > \$30 = E[Y]$ for the cost Y under consideration. From (1.3.1) and (1.3.2) it follows that the notion of risk-neutrality corresponds to the case in which the utility function is the identity function, whereas the controller is risk-averse when

$$E[U(Y)] > U(E[Y])$$

if the random cost Y is non-constant; from Jensen's inequality (Rudin, 1982, Royden and Fitzpatrick, 2010), this property is equivalent to the *strict convexity* of the utility function U. Similarly, a controller is risk-seeking when its utility function is (strictly) concave. The quantity

$$\Delta(Y) := \mathcal{E}(Y) - E[Y], \tag{1.3.3}$$

is referred to as the *risk-premium* corresponding to Y, and was used by Pratt(1964) to determine a single number measuring the degree of risk-aversion of the controller when facing a random cost which is 'close' to any specific value y. Consider a bounded random variable Z with mean 0 and variance 1 and, for each $\sigma > 0$, define

$$Y(\sigma) = y + \sigma Z \tag{1.3.4}$$

In this case, $Y(\sigma)$ converges in probability to y as σ goes to zero, since $E[Y(\sigma)] = y$ and $\text{Var}\,[Y(\sigma)] = \sigma^2$; notice that

$$\Delta(Y(0)) = 0. \tag{1.3.5}$$

Proposition 1.3.1. Suppose that the utility function U has a continuous derivative of order 2 in the real line, and that $U' > 0$. In this context,

$$\frac{\Delta(Y(\sigma))}{\sigma^2} \to \frac{1}{2}\frac{U''(y)}{U'(y)} \quad \text{as } \sigma \to 0.$$

Proof. Notice that

$$\begin{aligned}\frac{U(\mathcal{E}(Y(\sigma))) - U(\mathcal{E}(Y(0)))}{\sigma} &= \frac{E[U(Y(\sigma)) - U(Y(0))]}{\sigma}\\ &= \frac{E[U(y + \sigma Z) - U(y)]}{\sigma};\end{aligned}$$

since U' is continuous and Z is a bounded random variable with null expectation, the bounded convergence theorem implies that

$$\begin{aligned}\lim_{\sigma \to 0} \frac{U(\mathcal{E}(Y(\sigma)) - U(\mathcal{E}(Y(0))}{\sigma} &= E\left[\lim_{\sigma \to 0} \frac{U(y + \sigma Z) - U(y)}{\sigma}\right]\\ &= E\left[U'(y)Z\right]\\ &= U'(y)E\left[Z\right] = 0.\end{aligned}$$

It follows that the mapping $\sigma \mapsto U(\mathcal{E}(Y(\sigma))$ has null derivative at $\sigma = 0$ and, since $U' > 0$, this implies that $\mathcal{E}(Y(\sigma))$ is also differentiable at $\sigma = 0$, and that its derivative is null at that point. Consequently (see (1.3.5)),

$$\begin{aligned}\lim_{\sigma\to 0} \frac{\Delta(Y(\sigma))}{\sigma} &= \lim_{\sigma\to 0} \frac{\Delta(Y(\sigma)) - \Delta(Y(0))}{\sigma}\\ &= \frac{d}{d\sigma}\Delta(Y(\sigma))\bigg|_{\sigma=0}\\ &= \frac{d}{d\sigma}[\mathcal{E}(Y(\sigma)) - E[Y(\sigma)]]\bigg|_{\sigma=0}\\ &= \frac{d}{d\sigma}[\mathcal{E}(Y(\sigma)) - y]\bigg|_{\sigma=0}\\ &= 0,\end{aligned}$$

so that

$$\Delta(Y(\sigma)) = o(\sigma).$$

From this point, Taylor's theorem yields that

$$\begin{aligned}U(\mathcal{E}(Y(\sigma)) &= U(y + \Delta(Y(\sigma)))\\ &= U(y) + U'(y)\Delta(Y(\sigma)) + O([\Delta(Y(\sigma))]^2) \qquad (1.3.6)\\ &= U(y) + U'(y)\Delta(Y(\sigma)) + o(\sigma^2).\end{aligned}$$

Next, recalling that Z is bounded, a second order Taylor expansion yields that

$$U(y + \sigma Z) = U(y) + U'(y)\sigma Z + \frac{1}{2}U''(y)\sigma^2 Z^2 + R(y, \sigma Z),$$

where the residual satsifies $|R(y,\sigma)| \le k(y,\sigma)\sigma^2 Z^2$ with $k(y,\sigma) \to 0$ as $\sigma \to 0$: it follows that

$$E[|R(y,\sigma Z)|] \le k(y,\sigma)\sigma^2 E[Z^2] = k(y,\sigma)\sigma^2 = o(\sigma^2),$$

and then

$$\begin{aligned}E[U(Y(\sigma)] &= E[U(y + \sigma Z)]\\ &= E\left[U(y) + U'(y)\sigma Z + \frac{1}{2}U''(y)\sigma^2 Z^2 + R(y,\sigma Z)\right]\\ &= U(y) + \sigma U'(y)E[Z] + \frac{1}{2}U''(y)\sigma^2 E[Z^2] + E[R(y,\sigma)]\\ &= U(y) + \frac{1}{2}U''(y)\sigma^2 + o(\sigma^2).\end{aligned}$$

Combining this relation with (1.3.6), *via* (1.3.1) and (1.3.2), it follows that

$$U'(y)\Delta(Y(\sigma)) = \frac{1}{2}U''(y)\sigma^2 + o(\sigma^2),$$

a relation that is equivalent to the desired conclusion. □

The above result shows that, for a random cost Y taking values in a 'small' neighborhood of y, twice the risk-premium $\Delta(Y)$ is proportional to $\operatorname{Var}[Y]$, and that the proportionality constant is given by $U''(y/U'(y)$; in this work it is supposed that this quantity is *a positive constant*, that is,

$$\frac{U''(y)}{U'(y)} \equiv \lambda > 0, \quad y \in \mathbb{R}, \qquad (1.3.7),$$

and λ will be referred to as the *risk-sensitivity coefficient* of the controller. Notice that the above display immediately yields that $U(y) = ae^{\lambda y} + b$ for all $y \in \mathbb{R}$, where a and b are constants and $a > 0$. Since a utility function is determined up to a location-scale transformation, hereafter it is supposed that the decision maker values a random cost Y using the *exponential utility function*

$$U(y) = e^{\lambda y}, \quad y \in \mathbb{R}, \tag{1.3.8}$$

and in this case the certain equivalent $\mathcal{E}(Y)$ is given by

$$\mathcal{E}(Y) = \frac{1}{\lambda} \log \left(E \left[e^{\lambda Y} \right] \right). \tag{1.3.9}$$

1.4. The Main Problem

As already mentioned, in this work the performance of a control policy will be measured by an average criterion under the condition that the decision maker is risk-averse with constant risk sensitivity $\lambda > 0$. The construction of such a criterion is as follows: Given a positive integer n, consider the total cost $\sum_{t=0}^{n-1} C(X_t, A_t)$ incurred by the controller after applying the first n actions $A_0, A_1, \ldots, A_{n-1}$. When the system is driven by π and x is the initial sate, the certain equivalent of that random cost is

$$J_n(x; \pi) = \frac{1}{\lambda} \log \left(E_x^\pi \left[e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t)} \right] \right); \tag{1.4.1}$$

see (1.3.9). With this notation, the (superior limit λ-sensitive) *average cost at the initial sate x corresponding to the policy* $\pi \in \mathcal{P}$ is given by

$$J(x; \pi) = \limsup_{n \to \infty} \frac{1}{n} J_n(x; \pi), \tag{1.4.2}$$

and the corresponding (λ-sensitive) optimal average cost function is given by

$$J^*(x) = \inf_{\pi \in \mathcal{P}} J(x; \pi). \tag{1.4.3}$$

The above criterion $J(x, \pi)$ measures the performance of π in terms of the largest limit point of the sequence of average costs over a finite horizon. Focusing on the smallest of such limit points, the *inferior limit average criterion* is obtained:

$$J_-(x; \pi) = \liminf_{n \to \infty} \frac{1}{n} J_n(x; \pi), \tag{1.4.4}$$

with corresponding optimal value function

$$J_-^*(x) = \inf_{\pi \in \mathcal{P}} J_-(x; \pi). \tag{1.4.5}$$

It follows from these specifications that the relation $J_-(x, \pi) \leq J(x, \pi)$ is always valid, so that $J_-^*(\cdot) \leq J^*(\cdot)$; it will be shown that, in the context of this work, the superior and inferior limit optimal value functions coincide.

Applications of risk-sensitive criteria to diverse areas are available, as decision theory (Lin, 2005), analysis of inventories (Caravani, 1986, Bouakiz and Sobel, 1992), productive maintenance (Gosavi, 2007), learning theory (Mihatsch and Neuneier, 2002), and mathematical finance (Bielecki *et al.* , 1999, Bäuerle and Rieder, 2013, Stettner 2004). On the other hand, the study of Markov decision chains endowed with the risk-sensitive average criterion can be traced back, at least, to Howard and Matheson (1972), Jacobson (1973), and Jaquette (1973, 1976). The case of finite models was considered in Howard and Matheson (1972), where the Perron- Frobenious theory of positive matrices was used to obtain an *optimality equation* characterizing the optimal risk-sensitive average cost and allowing to obtain an optimal stationary policy, conclusions that were obtained under the following communication assumption: regardless of the policy used to drive the system, for every pair of states x and y it is possible to visit y when the initial state is x. The theory sparkled again around 1990, with the works of Whittle (1990), Runolfsson (1994), James *et al.* (1994), Flemming and McEneany (1995), where models with continous time-parameter or Borel state space were considered. Discrete models were studied by Marcus *et al.* (1996), Flemming and Hernández-Hernández (1997), Hernández-Hernández and Marcus (1997), using a game theoretical to study the criterion (1.4.2). On the other hand, in Cavazos-Cadena and Fernández-Gaucherand (1998), it was shown the aforementioned results by Howard and Matheson can not be directly conveyed to models satisfying strong recurrence requirements: it was shown in that paper that the simultaneous Doeblin condition— under which there is sate that is accessible from any other state regardless of the policy employed—is not sufficient to ensure that the optimal risk-sensitive average cost is characterized by the optimality equation, establishing an interesting contrast with the risk-neutral average index in (1.2.7). On the positive side, in Cavazos-Cadena (2003) it was shown that, under the aforementioned simultaneous Doeblin condition, the optimality equation for the criterion (1.4.2) admits a solution whenever the risk-sensitivity coefficient is small enough; the results in these two last papers, together show that obtaining a general characterization of the optimal risk-sensitive average cost is an interesting problem. The average cost criterion in (1.4.2) has been studied for models with general sate space or unbounded cost function; see, for instance Di Masi and Stettner (1999, 2000, 2007), where the risk-sensitive average cost index is studied using contractive mappings under the condition that the transition law satisfies a strong mixing condition, Hernández-Hernández and Marcus (1999) and Jaśkiewicz (2007), where it was supposed that the cost function is strictly unbounded and, *via* game theoretical arguments, an optimality inequality was obtained at a class of states where the optimal average cost is minimized, whereas a similar conclusion was established in Cavazos-Cadena and Salem-Silva (2009) using standard dynamic programming arguments and Hölder's inequality.

As already noted, the characterization of the optimal value average cost function in terms of an optimality equation is not generally valid, even under strong recurrence conditions. This fact motivates *the main problem* studied in this thesis:

- To establish a characterization of the optimal risk-sensitive average cost function, in such a way that (a) no restriction on the transition mechanism of the system is required, and (b) an optimal stationary policy can be obtained.

The *main results* obtained in this thesis can be briefly described as follows:

- For Markov decision chains with finite state space, under mild continuity conditions it is proved that (a) The superior and inferior optimal value functions in (1.4.3) and (1.4.4) coincide, and (b) The optimal risk-sensitive average cost function can be characterized using a nested system of *local* optimality equations, from which an optimal stationary policy can be derived.

These conclusions generalize the available characterizations of the optimal risk-sensitive average cost in terms of a single optimality equation, which are only applicable in models satisfying strong conditions on the transition mechanism.

1.5. The Organization

The content of the following chapters reflects the learning experience of the recent years and, roughly, the presentation is divided into two parts: The first one analyzes the results motivating the main goal pursued in this work, whereas in the second part the main contribution of the thesis is established and some open problems for future research are posed. A special effort has been made to produce self-contained chapters, and a reader interested solely in the main contribution of the thesis can go directly to Chapter 4.

The organization of the subsequent material is as follows: In Chapter 2 the characterization of the optimal risk-sensitive average cost *via* a single optimality equation is analyzed, including the fundamental result by Howard and Matheson (1972) concerning models with finite action set, as well as an extension obtained in Cavazos-Cadena and Fernández-Gaucherand (2002) for the case of compact action sets. Next, in Chapter 3 models satisfying a weak form of the communication property—the existence of an accesible state—are studied, and an alternative proof is presented for the result in Cavazos-Cadena (2003) establishing the existence of a solution to the optimality equation when the risk-sensitivity coefficient is 'small enough'; the analysis in these two chapters shows clearly the essential role of the communication assumption to ensure the existence of solutions to the optimality equation. The presentation continues in Chapter 4 where the main contribution of the thesis is established, namely, under mild continuity-compactness assumptions, it is shown that (i) the optimal average cost function is characterized by a system of 'nested' optimality equations, (ii) that the superior an inferior risk-sensitive average criteria have the same optimal value function, and (iii) that a solution of the system of optimality equations renders an optimal stationary policy. Finally, the presentation concludes in Chapter 5 stating two problems for future research.

Chapter 2

The Risk-Sensitive Average Optimality Equation in Communicating Models

This chapter presents the results that motivate the main problem studied in this thesis. The exposition analyzes a fundamental theorem on the existence of solutions of the optimality equation originally established in the seminal paper by Howard and Matheson (1972), whose approach is based on matrix analysis, as well as an extension of that result formulated in Cavazos-Cadena and Fernández-Gaucherand (2002), where the conclusions were obtained using a probabilistic analysis of a total cost problem. The conclusions in these two works can be described as follows: If the Markov chain induced by any stationary policy is *communicating*, then the optimal risk-sensitive average cost is constant and is characterized by a single equation, which also renders an optimal stationary policy. The analysis performed below highlights the role of the communication assumption in the derivation of this result.

2.1. Introduction

This chapter analyzes two available results on the characterization of the optimal risk-sensitive average cost for *communicating* Markov decision chains evolving on a finite state space; for this class of models, under the action of an arbitrary stationary policy any state y can be reached with positive probability regardless of the initial state x. In that context, the main conclusions are that the optimal average cost does not depend on the initial state, and that its common value is characterized by a single optimality equation. Under the condition that the action set is also finite, this result was firstly established by Howard and Matheson (1972) using an algebraic approach based on the Perron-Frobenious theorem, and an extension to models with general compact action sets was given in Cavazos-Cadena and Fernández-Gaucherand (2002) *via* dynamic programming techniques applied to an auxiliary total cost problem up to the first return time to a given state. In the following sections these results are analyzed and alternative proofs are provided, highlighting the role of the communication assumption in both approaches. The discussion shows that, characterizing the optimal risk-sensitive average cost and finding an optimal stationary policy in a framework were the transition mechanism of the model is arbitrary, is certainly a very interesting problem.

The organization of the chapter is as follows: The next three sections concern Markov decision chains with finite state and action sets, and the exposition begins stating a version of the fundamental conclusions by Howard and Matheson as Theorems 2.2.1 and 2.2.2 in Section 2; the first theorem establishes the existence of solutions to the risk-sensitive average optimality equation, whereas the second one shows that the optimal average cost and an optimal stationary policy can be obtained when a solution to the optimality equation is available; in contrast with the original formulation, no restriction on the period of the transition matrix corresponding to any stationary policy is imposed. The proof of those conclusions relies on algebraic properties of nonnegative matrices which are presented in Section 3, and Howard and Matheson's results are established in Section 4. Next, form that point onwards, models with finite state space and *compact* action sets are considered, and a result on the existence of solutions to the optimality equation in that context is stated as Theorem 2.5.1 in Section 5. The proof of that result is approached combining probabilistic and dynamic programming ideas to analyze the total cost incurred up to the first return time to a given state; such a problem is studied in Sections 6 and 7, and the proof of Theorem 2.5.1 is finally presented in Section 8. The exposition concludes in Section 9 with some brief comments on the essential role played by the communication assumption in the derivation of the main results.

Notation. A generic vector in the Euclidean space $\mathbb{R}^k$ is denoted by a boldface letter, as in $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$, and is always considered as a *column* vector, whereas $\mathbb{1} = (1, 1, \ldots, 1)' \in \mathbb{R}^k$ stands for the vector with all of its components equal to 1. If A is square matrix and t is a nonnegative integer, then A^t is the t-fold product of A with itself, that is, $A^0 = I$(the identity matrix) and $A^t = A \times A^{t-1}$ for $t \geq 1$. Finally, inequalities and operations on vectors are interpreted componentwise; for instance, for $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$ and $\mathbf{y} = (y_1, y_2, \ldots, y_k)'$,

$$\mathbf{x} \leq \mathbf{y} \iff x_i \leq y_i, \quad i = 1, 2, \ldots, k,$$

and

$$\mathbf{x}^\alpha = (x_1^\alpha, x_2^\alpha, \ldots, x_k^\alpha)'$$

whenever the right-hand side is meaningful. On the other hand, if $\mathbb{K}$ is a topological space, $\mathcal{B}(\mathbb{K})$ stands for the class of all bounded and real-valued functions defined on $\mathbb{K}$, whereas $\|C\|{:}= \sup_{x\in\mathbb{K}} |C(x)| < \infty$ denotes the supremum norm of $C \in \mathcal{B}(\mathbb{K})$. Finally, for an event W, the corresponding indicator function is denoted by $I[W]$, and all the relations involving conditional expectations are valid with probability 1 with respect to the underlying probability measure.

2.2. The Optimality Equation

The fundamental results about the risk-sensitive average cost criterion were established in Howard and Matheson (1972) where, under mild conditions on the decision model, it was proved that (i) the optimal average cost does not depend on the initial state and is characterized by a single equation, and (ii) that an optimal stationary policy can be derived from a solution of such an equation. These conclusions are formally stated in the following two theorems.

Theorem 2.2.1. [Optimality Equation.] Let $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, P, C)$ be a Markov decision chain satisfying the following conditions:

(i) The state space S and the action set A are finite, and

(ii) *Under the action of each stationary policy $f \in \mathbb{F}$, the state space is communicating*, that is,

$$\text{For every } x, y \in S, \text{ there exists an integer } n \equiv n(x, y) > 0 \text{ such that} \quad P_x^f[X_n = y] > 0. \tag{2.2.1}$$

In this case, given a risk-sensitivity coefficient $\lambda > 0$, there exist a real number g and a function $h{:}\, S \to \mathbb{R}$ such that the following *optimality equation* holds:

$$e^{\lambda g + \lambda h(x)} = \min_{a\in A(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S} p_{x\,y}(a) e^{\lambda h(y)} \right]. \tag{2.2.2}$$

This theorem is slightly more general than the original result established in Howard and Matheson (1972) where, additionally, it was supposed that the Markov chain associated with any stationary policy is *aperiodic*. A proof of Theorem 2.2.1 will be given in Section 4 after the algebraic preliminaries presented in the following section. Next, the theorem below shows that a solution $(g, h(\cdot)) \in \mathbb{R} \times \mathcal{B}(S)$ of (2.2.2) determines the (λ-sensitive) optimal average cost, as well as an optimal stationary policy, showing clearly the importance of the above optimality equation in the analysis of the risk-sensitive average criterion.

Theorem 2.2.2. [Verification theorem.] Let $\mathcal{M} = (S, A, \{A(x)\}, P, C)$ be a Markov decision process satisfying the conditions (i) and (ii) in the statement of Theorem 2.2.1, and let $(g, h(\cdot)) \in \mathbb{R} \times \mathcal{B}(S)$ be a solution of the optimality equation (2.2.2). In this case the following assertions (i)–(iii) hold:

(i) For every policy $\pi \in \mathcal{P}$

$$J_-(x; \pi) \geq g,$$

where $J_-(x, \pi)$ is the inferior limit λ-sensitive average criterion as specified in (1.4.4).

(ii) If $f \in \mathbb{F}$ is such that

$$e^{\lambda g + \lambda h(x)} = e^{\lambda C(x, f(x))} \sum_{y \in S} p_{x\,y}(f(x)) e^{\lambda h(y)} \tag{2.2.3}$$

is satisfied for every $x \in S$, then

$$g = \lim_{n \to \infty} \frac{1}{n+1} J_n(x; f) = J(x, f), \quad x \in S.$$

(iii) At each state $x \in S$

$$J_-^*(x) = g = J^*(x)$$

so that the policy $f \in \mathbb{F}$ in (2.2.3) is λ-average optimal, and g is the λ-optimal average cost.

Proof. Let $(g, h(\cdot))$ be a solution of (2.2.2) and observe that, for each $x \in S$ and $\pi \in \mathcal{P}$, the following relations always hold with probability 1 with respect to P_x^π:

$$\begin{aligned} e^{\lambda g + \lambda h(X_t)} &\leq e^{\lambda C(X_t, A_t)} \sum_{y \in S} p_{X_t\,y}(A_t) e^{\lambda h(y)} \\ &= E_x^\pi \left[e^{\lambda C(X_t, A_t) + \lambda h(X_{t+1})} \middle| H_t \right], \end{aligned} \tag{2.2.4}$$

where

$$H_t = (X_0, A_0, X_1, A_1, \ldots, X_{t-1}, A_{t-1}, X_t) \tag{2.2.5}$$

is the history of the process up to time t, and the equality is due to the Markov property; see (1.2.4).

(i) Let $x \in S$ and $\pi \in \mathcal{P}$ be arbitrary. Observing that $e^{\sum_{t=0}^{n-1} \lambda C(X_t, A_t)}$ is $\sigma(H_n)$-measurable for every positive integer n, it follows that

$$\begin{aligned} E_x^\pi \left[e^{\lambda \sum_{t=0}^{n} C(X_t, A_t) + \lambda h(X_{n+1})} \middle| H_n \right] &= e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t)} E_x^\pi \left[e^{\lambda C(X_n, A_n) + \lambda h(X_{n+1})} \middle| H_n \right] \\ &\geq e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t)} e^{\lambda g + \lambda h(X_n)} \\ &= e^{\lambda g} e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t) + \lambda h(X_n)}, \end{aligned}$$

where (2.2.4) was used to set the inequality. It follows that

$$E_x^\pi\left[e^{\lambda\sum_{t=0}^{n} C(X_t,A_t)+\lambda h(X_{n+1})}\right] \geq e^{\lambda g} E_x^\pi\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t,A_t)+\lambda h(X_n)}\right]. \tag{2.2.6}$$

Notice now that (2.2.4) with $t=0$ implies that

$$E_x^\pi\left[e^{\lambda C(X_0,A_0)+\lambda h(X_1)}\right] \geq e^{\lambda g+\lambda h(x)} \tag{2.2.7}$$

Combining the two last displayed relations, a simple induction argument yields that for every $x \in S$ and $\pi \in \mathcal{P}$,

$$E_x^\pi\left[e^{\lambda\sum_{t=0}^{n} C(X_t,A_t)+\lambda h(X_{n+1})}\right] \geq e^{\lambda(n+1)g+\lambda h(x)}, \quad n=1,2,3,\ldots,$$

and then

$$\begin{aligned} e^{\lambda\|h\|+\lambda J_n(x;\pi)} &= e^{\lambda\|h\|}E_x^\pi\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t,A_t)}\right] \\ &\geq E_x^\pi\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t,A_t)+\lambda h(X_n)}\right] \\ &\geq e^{\lambda ng+\lambda h(x)} \\ &\geq e^{\lambda ng-\lambda\|h\|}; \end{aligned}$$

it follows that

$$J_n(x;\pi) \geq ng - 2\|h\|,$$

a relation that together with (1.4.4) immediately leads to

$$J_-(x;\pi) = \liminf_{n\to\infty}\frac{1}{n}J_n(x;\pi) \geq g.$$

(ii) Let $f \in \mathbb{F}$ be as in (2.2.3). Since $P_x^f[A_t = f(X_t)] = 1$, it follows that for each nonnegative integer t the equality holds in (2.2.4) P_x^f- almost surely, and then a conditional argument similar to the one used to establish part(i) yields that relations (2.2.6) and (2.2.7) also occur with equality when f is used instead of π, so that, by induction,

$$E_x^f\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t,A_t)+\lambda h(X_n)}\right] = e^{\lambda ng+\lambda h(x)}, \quad x\in S, \quad n=1,2,3,\ldots \tag{2.2.8}$$

Notice now that for each positive integer n

$$\begin{aligned} e^{J_n(x;f)} &= E_x^f\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t,A_t)}\right] \\ &= E_x^f\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t,A_t)+h(X_n)}e^{-\lambda h(X_n)}\right] \\ &\leq E_x^f\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t,A_t)+h(X_n)}\right]e^{\lambda\|h\|} \\ &= e^{\lambda ng+\lambda h(x)}e^{\lambda\|h\|} \\ &\leq e^{\lambda ng+2\lambda\|h\|} \end{aligned}$$

where (2.2.8) was used to set the third equality; thus, $J_n(x;f) \leq ng+2\|h\|$; similarly, (2.2.8) yields that the relation, $J_n(x;f) \geq ng-2\|h\|$ is always valid, and then

$$ng - 2\|h\|| \leq J_n(x;f) \leq ng+2\|h\| \quad x\in S, \quad n=1,2,3,\ldots,$$

so that $g = \lim_{n\to\infty}[n^{-1}J_n(x;f)]$.

(iii) A glance to (1.4.1)–(1.4.4) shows that parts (i) and (ii) together imply that, for every $x \in S$ and $\pi \in \mathcal{P}$,

$$J_-(x;\pi) \geq g \geq J(x;f) \geq J^*(x)$$

where f is as in (2.2.3) and the third inequality is due to the specification of the optimal value function $J^*(\cdot)$. Taking the infimum over $\pi \in \mathcal{P}$, the above display yields that $J_-^*(x) \geq g \geq J(x;f) \geq J^*(x)$ for every $x \in S$; hence, since $J_-^*(\cdot) \leq J^*(\cdot)$, it follows that $J_-^*(\cdot) = g = J^*(\cdot) = J(\cdot;f)$, concluding the argument. □

2.3. Algebraic Preliminaries

This section presents the technical tools that will be used to prove Theorem 2.2.1. In Howard and Matheson (1972) such a result was established using that, under appropriate conditons, the positive eigenvalue of a nonnegative matrix is larger than the module of any other eigenvalue, a fact that is part of the classical Perron-Frobenious theorem (Meyer, 1995). In the following section, Theorem 2.2.1 will be derived using an alternative approach; the argument is also based on (part of) the Perron-Frobenious theorem, and emphasizes the importance of the concept of *communicating matrix* in the study of the risk–sensitive average criterion, an idea that is introduced below.

Definition 2.3.1. Let A be a given matrix of order $k \times k$ such that

$$A_{i\,j} \geq 0, \quad i,j = 1,2,3,\ldots,k.$$

In this case, A is communicating if for every pair of integers $i,j \in \{1,2,\ldots,k\}$ there exists a sequence $i_0, i_1, \ldots, i_r$ contained in $\{1,2,3,\ldots,k\}$ such that

$$i_0 = i, \quad i_r = j, \quad \text{and} \quad A_{i_{t-1},i_t} > 0, \quad t = 1,2,\ldots,r. \tag{2.3.1}$$

Remark 2.3.1. The sequence $i_0 = i, i_1, \ldots, i_r = j$ in (2.3.1) is referred to as a *path* from i to j with length r. It is not difficult to see that if $i \neq j$ and a path from i to j exists, then a path with length less than k can be found.

(ii) The previous point and the specification of matrix multiplication yield that a nonnegative matrix A is communicating if, and only if, $\sum_{t=0}^{k-1} A^t > 0$. □

The main instrument that will be used to establish the existence of a solution to the optimality equation (2.2.2) is the following result, which is part of the conclusions of the Perron-Frobenious theorem (Meyer, 1995).

Theorem 2.3.1. Let A be a matrix of order $k \times k$ with *nonnegative components*, and suppose that A is communicating. In this case, the assertions (i)–(iii) below occur:

(i) A has a positive eigenvalue which admits a positive eigenvector. More precisely, there exists $\mu > 0$ and $\mathbf{m} \in (0,\infty)^k$ such that

$$A\mathbf{m} = \mu\mathbf{m};$$

the pair $(\mu, \mathbf{m})$ is referred to as a *positive eigenpair* of A.

(ii) The positive eigenvalue μ in part (i) is unique and is equal to the grow-rate of the multiplicative iterates of A, that is, for each no-null vector $\mathbf{x} \in [0,\infty)^k$,

$$\lim_{n\to\infty} [A^n\mathbf{x}]^{1/n} = \mu\mathbb{1}.$$

(iii) A nonnegative sub-eigenvector or super-eigenvector of A corresponding to μ is, necessarily, an eigenvector. More precisely, If $\mathbf{x} \in [0,\infty)^k$ satisfies that $A\mathbf{x} \geq \mu\mathbf{x}$ or $A\mathbf{x} \leq \mu\mathbf{x}$, then $A\mathbf{x} = \mu\mathbf{x}$.

As it will be apparent in the following argument, the proof of this result is simpler when all of the components of the matrix A are positive, a case that will analyzed separately. Also, the argument below shows that the second part follows immediately form the existence of a positive eigenvalue μ and the corresponding eigenvector $\mathbf{m}$ with positive components, whereas the third part depends heavily on the communication property. Before going any further, it is convenient to introduce the following notation.

Definition 2.3.2. Let A be a fixed matrix of order $k \times k$ with positive components.

(i) For each $\mathbf{x} = (x_1, \ldots, x_k)' \in (0,\infty)^k$ define

$$\mu(\mathbf{x}) = \min\left\{ \frac{[A\mathbf{x}]_i}{x_i} \,\middle|\, i = 1, 2, \ldots, k \right\},$$

and set

$$\mu^* = \sup_{\mathbf{x}\in(0,\infty)^k} \mu(\mathbf{x}) \tag{2.3.2}$$

(ii) The number a is given by

$$a = \max\left\{ \frac{A_{i\,j}}{A_{r\,j}} \,\middle|\, i, r, j = 1, 2, \ldots k \right\},$$

whereas the set $\mathcal{K}$ is specified by

$$\mathcal{K} = \left\{ \mathbf{x} \in (0,\infty)^k \,\middle|\, \frac{1}{ak} \leq x_i \leq \frac{a}{k},\ i = 1, 2, \ldots, k \right\}. \tag{2.3.3}$$

Observe that, $\mu(\cdot) > 0$, since all of the components of A are positive, a property that also yields that the number a is well-defined and belongs to $(0,\infty)$, so that the set $\mathcal{K}$ is a compact subset of the positive cone $(0,\infty)^k$. The following properties follow directly form Definition 2.3.2:

$$\mu(c\mathbf{x}) = \mu(\mathbf{x}), \quad \text{and} \quad A\mathbf{x} \geq \mu(\mathbf{x})\mathbf{x}, \quad \mathbf{x} \in (0,\infty)^k, \quad c > 0. \tag{2.3.4}$$

It will be shown below that μ^* in (2.3.2) is the positive eigenvalue of the matrix A; such a characterization is known as the Collatz-Wielandt relation (Meyer, 1995). The following auxiliary result is the starting point to establish Theorem 2.3.1.

Lemma 2.3.1. Let A be a matrix of order $k \times k$ be such that

$$A_{i\,j} > 0, \quad i, j = 1, 2, 3, \ldots, k.$$

Using the notation of Definition 2.3.2, the assertions (i)–(iv) below are valid.

(i) $\mu(A\mathbf{x}) \geq \mu(\mathbf{x})$ for every $\mathbf{x} \in (0,\infty)^k$;

(ii) The inclusion

$$\frac{A\mathbf{x}}{\mathbb{1}'A\mathbf{x}} \in \mathcal{K}$$

holds for every $\mathbf{x} \in (0,\infty)^k$.

Consequently,

(iii) μ^* satisfies that

$$\mu^* = \sup_{\mathbf{x}\in\mathcal{K}} \mu(\mathbf{x}),$$

and then
(iv) There exists $\mathbf{y}^* \in \mathcal{K} \subset (0,\infty)^k$ such that $\mu^* = \mu(\mathbf{y}^*)$.

Proof. (i) Let $\mathbf{x}$ be a given vector with positive components, and write $\mathbf{z} = A\mathbf{x}$. Using the inequality in (2.3.4) it follows that $A\mathbf{z} = A(A\mathbf{x}) \geq A(\mu(\mathbf{x})\mathbf{x}) = \mu(\mathbf{x})A\mathbf{x}$, that is, $A\mathbf{z} \geq \mu(\mathbf{x})\mathbf{z}$, a relation that leads to $\mu(A\mathbf{x}) = \mu(\mathbf{z}) \geq \mu(\mathbf{x})$, by Definition 2.3.2(i).
(ii) Notice that the specification of the number a in Definition 2.3.2(ii) yields that

$$A_{i\,j} \leq aA_{r\,j}, \quad r, i, j = 1, 2, 3, \ldots, k. \tag{2.3.5}$$

Given a fixed possible index r this implies that, for every $\mathbf{x} \in (0,\infty)^k$,

$$\mathbb{1}'A\mathbf{x} = \sum_{i=1}^k \sum_{j=1}^k A_{i\,j}x_j \leq \sum_{i=1}^k \sum_{j=1}^k aA_{r\,j}x_j = a\sum_{i=1}^k [A\mathbf{x}]_r = ak[A\mathbf{x}]_r$$

so that

$$\frac{1}{ak} \leq \frac{[A\mathbf{x}]_r}{\mathbb{1}'A\mathbf{x}}, \quad r = 1, 2, 3, \ldots, k. \qquad (2.3.6).$$

Similarly, for a fixed index i and $\mathbf{x} \in (0,\infty)^k$, (2.3.5) implies that

$$k[A\mathbf{x}]_i = \sum_{r=1}^k \sum_{j=1}^k A_{i\,j}x_j \leq \sum_{r=1}^k \sum_{j=1}^k aA_{r\,j}x_j = a\mathbb{1}'A\mathbf{x},$$

and then

$$\frac{[A\mathbf{x}]_i}{\mathbb{1}'A\mathbf{x}} \leq \frac{a}{k}, \quad i = 1, 2, 3, \ldots, k,$$

a relation that together with (2.3.6) leads to the desired conclusion; see (2.3.3).
(iii) Notice that (2.3.2) and part (i) together yield that $\mu^* = \sup_{\mathbf{x}\in(0,\infty)^k} \mu(A\mathbf{x})$, a fact that combined with the equality in (2.3.4) implies that

$$\mu^* = \sup_{\mathbf{x}\in(0,\infty)^k} \mu\left(\frac{A\mathbf{x}}{\mathbb{1}'A\mathbf{x}}\right) = \sup_{\mathbf{y}\in\mathcal{K}} \mu(\mathbf{y}).$$

where part (ii) was used to set the second equality.
(iv) Since the set $\mathcal{K}$ is compact and $\mu(\cdot)$ is a continuous mapping, there exists $\mathbf{y}^* \in \mathcal{K}$ such that $\mu(\mathbf{y}^*) = \sup_{\mathbf{y}\in\mathcal{K}} \mu(\mathbf{y})$, and then $\mu^* = \mu(\mathbf{y}^*)$, by part (iii). □

Next, the previous lemma will be used to establish the main conclusion of this section.

Proof of Theorem 2.3.1. The argument has been divided into two steps:

Case 1: All of the components of the matrix A are positive.
In this context, let μ^* be as in (2.3.2) and, using Lemma 2.3.1(iv), select a vector $\mathbf{y}^* \in (0,\infty)^k$ such that $\mu^* = \mu(\mathbf{y}^*)$, so that

$$A\mathbf{y}^* \geq \mu^*\mathbf{y}^*,$$

by (2.3.4).

(i) It will be verified that

$$A\mathbf{y}^* = \mu^*\mathbf{y}^*. \tag{2.3.7}$$

To achieve this goal, set

$$\mathbf{z} = A\mathbf{y}^* - \mu^*\mathbf{y}^*$$

and notice that $\mathbf{z} \geq 0$. Now, *suppose* that $\mathbf{z}$ is no-null. In this case, recalling that $A_{i\,j} > 0$ for every i, j, it follows that $A\mathbf{z} > 0$, that is, $A(A\mathbf{y}^*) - \mu^* A\mathbf{y}^* > 0$, so that there exists $\varepsilon > 0$ satisfying $A(A\mathbf{y}^*) - (1+\varepsilon)\mu^* A\mathbf{y}^* > 0$, *i.e.*,

$$A(A\mathbf{y}^*) > (1+\varepsilon)\mu^* A\mathbf{y}^*;$$

by Definition 2.3.2, this relation yields that $\mu(A\mathbf{y}^*) > (1+\varepsilon)\mu^*$, contradicting the fact that $\mu^*(> 0)$ is the supremum of the function $\mu(\cdot)$. This contradiction stems from the assumption that the vector $\mathbf{z}$ is no-null, and then $\mathbf{z} = 0$, which is equivalent to (2.3.7). Therefore, the pair $(\mu, \mathbf{m}) \equiv (\mu^*, \mathbf{y}^*)$ satisfies the first conclusion of Theorem 2.3.1.

(ii) Notice that $A\mathbf{m} = \mu\mathbf{m}$ leads to $A^n\mathbf{m} = \mu^n\mathbf{m}$ for every positive integer n, and then $[A^n\mathbf{m}]^{1/n} = \mu[\mathbf{m}]^{1/n}$, so that

$$\lim_{n\to\infty} [A^n\mathbf{m}]^{1/n} = \mu \lim_{n\to\infty} [\mathbf{m}]^{1/n} = \mu\mathbb{1}. \tag{2.3.8}$$

Now let $\mathbf{x} \in [0, \infty)^k$ be an arbitrary but fixed *no-null* vector, and notice that $\tilde{\mathbf{x}} = A\mathbf{x} > 0$. Since $\mathbf{m}$ has positive components, it follows that there exist positive constants c_0 and c_1 such that $c_0\mathbf{m} \leq \tilde{\mathbf{x}} \leq c_1\mathbf{m}$ so that

$$c_0 A^{n-1}\mathbf{m} \leq A^{n-1}\tilde{\mathbf{x}} = A^n\mathbf{x} \leq c_1 A^{n-1}\mathbf{m};;$$

taking the nth root and then the limit as n goes to ∞ in this relation, (2.3.8) yields that $[A^n\mathbf{x}]^{1/n} \to \mu\mathbb{1}$, a fact that implies the uniqueness of the positive eigenvalue μ.

(iii) Suppose that the vector $\mathbf{x} \in [0, \infty)^k$ satisfies $A\mathbf{x} \geq \mu\mathbf{x}$, so that $\mathbf{z} = A\mathbf{x} - \mu\mathbf{x} \geq 0$, and assume that $\mathbf{z}$ is no-null. In this case $\mathbf{x} \neq \mathbf{0}$ and, using that $A_{i\,j} > 0$ is always valid, it follows that $A\mathbf{z} > 0$, that is,

$$A\mathbf{z} = A(A\mathbf{x} - \mu\mathbf{x}) = A(A\mathbf{x}) - \mu A(\mathbf{x}) > 0.$$

Consequently, there exists $\varepsilon > 0$ such that $A\mathbf{y} - (1+\varepsilon)\mu\mathbf{y} \geq 0$, where $\mathbf{y} = A\mathbf{x}(> 0)$, and then $A^n\mathbf{y} \geq (1+\varepsilon)^n\mu^n\mathbf{y}$ for $n = 1, 2, 3, \ldots$; thus, $\lim_{n\to\infty}[A^n\mathbf{y}]^{1/n} \geq (1+\varepsilon)\mu\mathbb{1}$, in contradiction with part (ii). Therefore , if $\mathbf{x} \in [0, \infty)^k$ satisfies that $A\mathbf{x} \geq \mu\mathbf{x}$, then $A\mathbf{x} = \mu\mathbf{x}$. Similarly it can be established that if $A\mathbf{x} \leq \mu\mathbf{x}$ for some vector $\mathbf{x} \in [0, \infty)^k$, then $A\mathbf{x} = \mu\mathbf{x}$. This completes the proof of Theorem 2.3.1 when all of the components of A are positive.

Case 2: The nonnegative and communicating matrix A is arbitrary.

Define the matrices $\hat{A}$ and $\tilde{A}$ by

$$\hat{A} = I + A \quad \text{and} \quad \tilde{A} = \hat{A}^k. \tag{2.3.9}$$

and notice that

$$\tilde{A} = \sum_{r=0}^{k} \binom{k}{r} A^r > 0,$$

where the inequality is due to Remark 2.3.1. Applying the preceding Case 1 to this matrix $\tilde{A}$, there exists a pair $(\tilde{\mu}, \tilde{\mathbf{m}})$ satsifying

$$\tilde{A}\tilde{\mathbf{m}} = \tilde{\mu}\tilde{\mathbf{m}}, \tag{2.3.10}$$

where $\tilde{\mu} > 0$ and $\tilde{\mathbf{m}} \in (0, \infty)^k$ and, moreover,

$$\tilde{\mu}\mathbb{1} = \lim_{n\to\infty} [\tilde{A}^n \mathbf{m}]^{1/n} \geq \lim_{n\to\infty} [\mathbf{m}]^{1/n} = \mathbb{1},$$

where the inequality is due to the relation $\tilde{A} \geq I$; thus,

$$\tilde{\mu} \geq 1.$$

Define

$$\hat{\mu} = [\tilde{\mu}]^{1/k}, \quad \mathbf{m} = \left[\sum_{r=0}^{k-1} \hat{\mu}^{k-1-r} \hat{A}^r\right] \tilde{\mathbf{m}} \quad \text{and} \quad \mu = \hat{\mu} - 1, \tag{2.3.11}$$

and observe that the following factorizations hold:

$$\tilde{A} - \tilde{\mu}I = \hat{A}^k - \hat{\mu}^k I = (\hat{A} - \hat{\mu}I)\left[\sum_{r=0}^{k-1} \hat{\mu}^{k-1-r} \hat{A}^r\right] = \left[\sum_{r=0}^{k-1} \hat{\mu}^{k-1-r} \hat{A}^r\right](\hat{A} - \hat{\mu}I). \tag{2.3.12}$$

Using that $\hat{A} \geq A$ and $\tilde{\mu} \geq 1$, Remark 2.3.1 yields that the matrix within brackets in the above expressions has positive components, and then, since $\tilde{\mathbf{m}} > 0$, it follows that $\mathbf{m} > 0$. Notice that (2.3.10)–(2.3.12) together lead to

$$0 = (\tilde{A} - \tilde{\mu}I)\tilde{\mathbf{m}} = (\hat{A} - \hat{\mu}I)\left[\sum_{r=0}^{k-1} \hat{\mu}^{k-1-r} \hat{A}^r\right] \tilde{\mathbf{m}} = (\hat{A} - \hat{\mu}I)\mathbf{m}. \tag{2.3.13}$$

so that $\hat{A}\mathbf{m} = \hat{\mu}\hat{\mathbf{m}}$, an equality that is equivalent to

$$A\mathbf{m} = \mu\mathbf{m}; \tag{2.3.14}$$

see (2.3.9) and (2.3.11).

(i) Since $\mathbf{m}$ has positive coordinates, from (2.3.14) it is sufficient to show that $\mu > 0$. To achieve this goal, first notice that $\mu \neq 0$; indeed, if $\mu = 0$ the above display yields that $A\mathbf{m} = 0$, and then, since A has nonnegative components and $\mathbf{m} \in (0, \infty)^k$, it follows that A is null, contradicting that A is a communicating matrix. On the other hand, as already noted, the relation $\tilde{\mu} \geq 1$ holds, so that $\mu \geq 0$ (see (2.3.11)), and then μ is positive.

(ii) This part follows using the same argument as in Case 1 above.

(iii) Let $\mathbf{x} \in [0, \infty)^k$ be such that $A\mathbf{x} \geq \mu\mathbf{x}$. In this case, (2.3.9) and (2.3.11) yield that

$$\hat{A}\mathbf{x} - \hat{\mu}\mathbf{x} \geq 0,$$

as well as $\tilde{A}\mathbf{x} \geq \tilde{\mu}\mathbf{x}$; from this last inequality, an application of Case 1 to the matrix $\tilde{A}$ yields that $(\tilde{A} - \tilde{\mu}I)\mathbf{x} = 0$, and combining this fact with the factorization in (2.3.12) it follows that

$$\left[\sum_{r=0}^{k-1} \hat{\mu}^{k-1-r} \hat{A}^r\right](\hat{A} - \hat{\mu}I)\mathbf{x} = 0;$$

As already noted, all of the components of the matrix within brackets are positive, whereas the vector $\hat{A}\mathbf{x} - \hat{\mu}\mathbf{x}$ has nonnegative coordinates, so that that the above display yields that $(\hat{A} - \hat{\mu}I)\mathbf{x} = 0$, which is equivalent to $A\mathbf{x} = \mu\mathbf{x}$, by (2.3.9) and (2.3.11). A similar argument can be used to show that, if $A\mathbf{x} \leq \mu\mathbf{x}$ for some $\mathbf{x} \in [0, \infty)^k$, then $A\mathbf{x} = \mu\mathbf{x}$. □

2.4. Proof of Theorem 2.2.1

In this section Theorem 2.3.1 will be used to establish the existence of solutions of the optimality equation (2.2.2). Let $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, P, C)$ be a Markov decision process with finite state and action spaces, so that the set $\mathbb{F} = \prod_{x\in S} A(x)$ is also finite. For each $f \in \mathbb{F}$ define the matrix $A(f)$ whose rows and columns are indexed by the elements of S as follows:

$$A(f)_{x\,y} = e^{\lambda C(x,f(x))} p_{x\,y}(f(x)), \quad x, y \in S. \tag{2.4.1}$$

Next, suppose that $x, y \in S$ and the positive integer n are such that $P_x^f[X_n = y] > 0$. In this case, there exist states $x_1, x_2, \ldots x_{n-1} \in S$ such that

$$P_x^f[X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}, X_n = y] > 0$$

a relation that can be more explicitly written as

$$p_{x\,x_1}(f(x)) p_{x_1\,x_2}(f(x)) p_{x_2\,x_3}(f(x)) \cdots p_{x_{n-2}\,x_{n-1}}(f(x)) p_{x_{n-1}\,y}(f(x)) > 0,$$

and *via* (2.4.1) this inequality is equivalent to

$$A(f)_{x\,x_1} A(f)_{x_1\,x_2} A(f)_{x_2\,x_3} \cdots A(f)_{x_{n-2}\,x_{n-1}} A(f)_{x_{n-1}\,y} > 0,$$

so that, under the condition (2.2.1) in Theorem 2.2.1, each matrix $A(f)$ is communicating; see Definition 2.3.1. Consequently, an application of Theorem 2.3.1 yields that, for each $f \in \mathbb{F}$, there exists $\mu(f) > 0$ and $\mathbf{m}(f) \in (0, \infty)^k$ such that

$$\mu(f)\mathbf{m}(f) = A(f)\mathbf{m}(f). \tag{2.4.2}$$

Define

$$\mu^* = \min_{f\in\mathbb{F}} \mu(f) \tag{2.4.3}$$

and, recalling that $\mathbb{F}$ is finite, select $f^* \in \mathbb{F}$ such that

$$\mu(f^*) = \mu^*; \tag{2.4.4}$$

finally, set

$$\mathbf{m}^* = \mathbf{m}(f^*). \tag{2.4.5}$$

With this notation, it follows that $\mu^*\mathbf{m}^* = A(f^*)\mathbf{m}^*$, an equation that can be more explicitly written as

$$\mu^*\mathbf{m}^*_x = e^{\lambda C(x,f^*(x))} \sum_{y\in S} p_{x\,y}(f^*(x))\mathbf{m}^*_y, \quad x \in S; \tag{2.4.6}$$

Proof of Theorem 2.2.1. It will be proved that

$$\mu^*\mathbf{m}^*_x = \min_{a\in A(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S} p_{x\,y}(a)\mathbf{m}^*_y \right], \quad x \in S. \tag{2.4.7}$$

To achieve this goal, for each $x \in S$ select a minimizer $\tilde f(x) \in A(x)$ of the right hand side of this equation, so that

$$\begin{aligned} e^{\lambda C(x,\tilde f(x))} \sum_{y\in S} p_{x\,y}(\tilde f(x))\mathbf{m}^*_y &= \min_{a\in A(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S} p_{x\,y}(a)\mathbf{m}^*_y \right] \\ &\le e^{\lambda C(x,f^*(x))} \sum_{y\in S} p_{x\,y}(f^*(x))\mathbf{m}^*_y \\ &= \mu^*\mathbf{m}^*_x, \end{aligned} \tag{2.4.8}$$

where (2.4.6) was used to set the last equality. Since $x \in S$ is arbitrary, using (2.4.1) this relation is equivalent to $A(\tilde{f})\mathbf{m}^* \leq \mu^*\mathbf{m}^*$, and then, since $\mu^* \leq \mu(\tilde{f})$ (see (2.4.3)), it follows that

$$A(\tilde{f})\mathbf{m}^* \leq \mu^*\mathbf{m}^* \leq \mu(\tilde{f})\mathbf{m}^*.$$

Thus, starting form $A(\tilde{f})\mathbf{m}^* \leq \mu(\tilde{f})\mathbf{m}^*$, an application of Theorem 2.3.1(iii) to the matrix $A(\tilde{f})$ yields that $A(\tilde{f})\mathbf{m}^* = \mu(\tilde{f})\mathbf{m}^*$, an equality that combined with the previous display yields that $A(\tilde{f})\mathbf{m}^* = \mu^*\mathbf{m}^*$, that is, for every $x \in S$,

$$e^{\lambda C(x,\tilde{f}(x))} \sum_{y \in S} p_{x\,y}(\tilde{f}(x))\mathbf{m}^*_y = \mu\mathbf{m}^*_x;$$

see (2.4.1). This relation and (2.4.8) together imply that (2.4.7) holds. To conclude, set

$$g = \frac{1}{\lambda}\log(\mu^*) \quad \text{and} \quad h(x) = \frac{1}{\lambda}\log(\mathbf{m}^*_x), \quad x \in S;$$

with this notation (2.4.7) is equivalent to the optimality equation (2.2.2). □

2.5. Models with Compact Action Sets

In this section the fundamental result in Theorem 2.2.1 is extended to the case of Markov decision chains with finite state space and *compact action sets*. Besides mild continuity conditions to be stated below, the basic structural assumption on the model is, again, that each stationary policy induces a communicating Markov chain. Within this modified framework, the existence of solutions to the optimality equation will be established using a probabilistic analysis of a total cost problem, and the argument emphasizes the central role of the communication condition. Throughout the remainder $\mathcal{M} = (S, A, \{A(x)\}_{x \in S}, P, C)$ is a Markov decision process, where the state space S is finite and the action set A is a metric space. Additionally, the following conditions are enforced.

Assumption 2.5.1. *(i) For each $x \in S$ the action set $A(x)$ is a compact subspace of A;*

(ii) For each $x, y \in S$ the mappings $a \mapsto C(x,a)$ and $a \mapsto p_{x\,y}(a)$ are continuous functions of $a \in A(x)$;

(iii) For each stationary policy $f \in \mathbb{F}$ the Markov chain induced by f is communicating; see (2.2.1).

Theorem 2.5.1. For an arbitrary risk-sensitivity coefficient $\lambda > 0$, under Assumption 2.5.1 there exist $g \in \mathbb{R}$ and $h\colon S \to \mathbb{R}$ such that the optimality equation holds, that is,

$$e^{\lambda g + \lambda h(x)} = \min_{a \in A(x)} \left[e^{\lambda C(x,a)} \sum_{y \in S} p_{x\,y}(a) e^{\lambda h(y)} \right].$$

This theorem was originally established in Cavazos-Cadena and Fernández-Gaucherand (2002) using probabilistic ideas to analyze a risk-sensitive *total cost problem* for controlled Markov chains (Cavazos-Cadena *et al.* 2000, Cavazos-Cadena and Montes-de-Oca 2000, 2000a). In the following sections an alternative (simpler) derivation will be presented, which is also based on the total cost criterion, but applied to *uncontrolled* models.

Remark 2.5.1. Notice that, under Assumption 2.5.1, for each $x \in S$ the right-hand side of the optimality equation has a minimizer $f(x)$, so that the policy $f \in \mathbb{F}$ satisfies (2.2.3), and it is not difficult to see that the conclusions of Theorem 2.2.2 hold with the same proof. □

A verification of the above result will be presented in Section 8, after the preliminaries established in the following two sections. The argument relies heavily on the concept of *first return time* which is used to define an auxiliary problem with the risk-sensitive total cost criterion; the relevant notions are introduced below.

2.6. Stopping Times and Total Costs

The main idea used below to establish Theorem 2.5.1 is the notion of *first return time*, which is now introduced.

Definition 2.6.1. Let $z \in S$ be a fixed state. The first return time to state z is defined by

$$T_z := \min\{n > 0 \,|\, X_n = z\},$$

where the minimum of the empty set is ∞.

Notice that for each positive integer t,

$$[T_z = t] = [X_t = z, X_r \neq z \text{ if } 1 \leq r < t] \in \sigma(H_t) \tag{2.6.1}$$

so that T_z is a *stopping time* with respect to the filtration $\{\sigma(H_t)\}$; see (2.2.5). The following consequence of the *communication property* in Assumption 2.5.1(iii) will play an important role in the subsequent development.

Lemma 2.6.1. Let f be a stationary policy. Under Assumption 2.5.1(iii), for different states $x, z \in S$,

$$P_z^f[T_x < T_z] > 0.$$

Proof. Given different states x and z, Assumption 2.5.1(iii) yields that there exists a positive integer n such that

$$P_z^f[X_n = x] > 0.$$

Now let m be the minimum positive integer n satisfying this relation, so that

$$P_z^f[X_m = x] > 0 \quad \text{and} \quad P_z^f[X_t = x] = 0 \text{ when } 1 \leq t < m.$$

Notice now that the above equality and the Markov property together imply that, for each positive integer r less than m, $P_z^f[T_z = r, X_m = x | H_r] = I[T_z = r] P_z^f[X_{m-r} = x] = 0$, so that $P_z^f[T_z = r, X_m = x] = 0$. It follows that

$$P_z^f[X_m = x] = P_z[X_m = x, T_z \geq m] = P_z[X_m = x, T_z > m],$$

where the second equality used that x and z are different. The two last displays yield that $P_z^f[X_m = x, T_z > m] > 0$ and, form this point, the conclusion follows observing that $[X_m = x, T_z > m] \subset [T_x < T_z]$. □

Now suppose that the system runs until a given state z is reached in a positive time, and that the system is halted at that moment. Also, assume that the system is driven by a stationary policy f and that a cost $D(x)$ is incurred every time that the state is x before the first return time to z. In this case, $\sum_{t=0}^{T_z - 1} D(X_t)$ is the total cost incurred while the system is running, and a special notation for such a quantity is now introduced.

Definition 2.6.2. Let $f \in \mathbb{F}$ and the function $D\colon S \to \mathbb{R}$ be arbitrary, Given a state $z \in S$, the function $h_{z,f,D}\colon S \to \mathbb{R}$ is defined by

$$h_{z,f,D}(x) {:=}\, \frac{1}{\lambda} \log \left(E_x^f \left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} \right] \right).$$

Notice that $h_{z,f,D}(\cdot)$ is well-defined—since it is expressed in terms of the expectation of a nonnegative quantity—but may attain the value $+\infty$. The following lemma establishes some elementary properties of the functions $h_{z,f,D}$.

Lemma 2.6.2. Suppose that the communication property in Assumption 2.5.1(iii) holds and, given $f \in \mathbb{F}$ and $D : S \to \mathbb{R}$, assume that $h_{z,f,D}(z) < \infty$ for some state $z \in S$. In this case,
(i) $h_{z,f,D}(x) < \infty$ for every $x \in S$;
(ii) The function $h_{z,f,D}(\cdot)$ satisfies *the dynamic programming equation*

$$e^{\lambda h_{z,f,D}(x)} = e^{\lambda D(x)} \left[p_{x\,z}(f(x)) + \sum_{y \in S \setminus \{z\}} p_{x\,y}(f(x)) e^{\lambda h_{z,f,D}(y)} \right], \quad x \in S. \tag{2.6.2}$$

In particular,
(iii) If $h_{z,f,D}(z) = 0$ then

$$e^{\lambda h_{z,f,D}(x)} = e^{\lambda D(x)} \sum_{y \in S} p_{x\,y}(f(x)) e^{\lambda h_{z,f,D}(y)}, \quad x \in S. \tag{2.6.3}$$

Proof. (i) Let $x \in S \setminus \{z\}$ be arbitrary, and notice that there exists a positive integer r such that $P_z[T_x = r < T_z] > 0$, by Lemma 2.6.1; using that the event $[T_x = r < T_z]$ is $\sigma(H_r)$-measurable (see (2.2.5)), it follows that

$$\begin{aligned}
&E_z^f \left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} \middle| H_r \right] \\
&\qquad \ge E_z^f \left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} I[T_x = r < T_z] \middle| H_r \right] \\
&\qquad \ge e^{\lambda \sum_{t=0}^{r-1} D(X_t)} I[T_x = r < T_z] E_z^f \left[e^{\lambda \sum_{t=r}^{T_z-1} D(X_t)} \middle| H_r \right] \\
&\qquad \ge e^{-\lambda r \|D\|} I[T_x = r < T_z] E_z^f \left[e^{\lambda \sum_{t=r}^{T_z-1} D(X_t)} \middle| H_r \right] \\
&\qquad = e^{-\lambda r \|D\|} I[T_x = r < T_z] E_x^f \left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} \right] \\
&\qquad = e^{-\lambda r \|D\|} I[T_x = r < T_z] e^{\lambda h_{z,f,D}(x)}
\end{aligned}$$

where, using that $X_r = x$ on the event $[T_x = r]$, the first equality is due to the Markov property, and the second one follows from Definition 2.6.2. After taking the expected value with respect to P_z^f, it follows that

$$\begin{aligned}
e^{h_{z,f,D}(z)} &= E_z^f \left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} \right] \\
&\ge P_z^f [T_x = r < T_z] e^{-\lambda r \|D\|} e^{h_{z,f,D}(x)}
\end{aligned}$$

and then, since the probability is the above expression is positive, $h_{z,f,D}(z) < \infty$ implies that $h_{z,f,D}(x)$ is also finite.

(ii) Given $x \in S$, Definition 2.6.2 yields that

$$\begin{aligned} & e^{\lambda h_{z,f,D}(x)} \\ & \quad = E_x^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)}\right] \\ & \quad = E_x^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} I[X_1 = z]\right] + \sum_{y\in S\setminus\{z\}} E_x^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} I[X_1 = y]\right]. \end{aligned} \qquad (2.6.4).$$

Notice now that $T_z > 1$ on the event $[X_1 = y]$ when $y \neq z$, and then, by the Markov property,

$$\begin{aligned} E_x^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} I[X_1 = y] \middle| X_1\right] &= e^{\lambda D(x)} I[X_1 = y] E_y^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)}\right] \\ &= e^{\lambda D(x)} I[X_1 = y] e^{\lambda h_{z,f,D}(y)}, \end{aligned} \qquad (2.6.5)$$

and taking the expectation with respect to P_x^f, it follows that

$$E_x^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} I[X_1 = y]\right] = e^{\lambda D(x)} p_{x\,y}(f(x)) e^{\lambda h_{z,f,D}(y)}.$$

On the other hand, the equality $[T_z = 1] = [X_1 = z]$ leads to

$$E_x^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)} I[X_1 = z]\right] = e^{\lambda D(x)} E_x^f\left[I[X_1 = z]\right] = e^{\lambda D(x)} p_{x\,z}(f(x)),$$

and the conclusion follows combining this equality with (2.6.4) and (2.6.5).

(iii) When $h_{z,f,D}(z) = 1$, equations (2.6.3) and (2.6.2) are equivalent. □

Throughout the remainder, for each $x \in S$ the indicator function of the singleton $\{x\}$ is denoted by δ_x, that is,

$$\delta_x(y) = \begin{cases} 0, & \text{if } y \neq x, \\ 1, & \text{when } y = x. \end{cases} \qquad (2.6.6)$$

The following result is the essential instrument that will be used to prove Theorem 2.5.1. Among the assertions in the next theorem, the fourth one is the most important and can be roughly described as follows: The class of all real valued functions D satisfying that $h_{z,f,D}(z) < 0$, is an open set in $\mathcal{B}(S)$.

Theorem 2.6.1. For a given policy $f \in \mathbb{F}$ and $D\colon S \to \mathbb{R}$, the following assertions (i)–(iii) hold:

(i) If $D_1\colon S \to \mathbb{R}$ is such that $D_1(\cdot) \geq D(\cdot)$ with $D_1(y) > D(y)$ for some state $y \in S$, then $h_{z,f,D_1}(z) > h_{z,f,D}(z)$ when $h_{z,f,D}(z)$ is finite.

(ii) If $h_{z,f,D}(z) \leq 0$ for some state z, then

$$h_{y,f,D+a\delta_z}(y) \leq 0 \text{ for } \textit{every state } y,$$

where

$$a = -h_{z,f,D}(z). \qquad (2.6.7)$$

Consequently,

(iii) If $h_{z,f,D}(z) < 0$ at some state z, then there exists a positive constant b such that $h_{x,f,D+b\delta_z}(x) < 0$ for all $x \in S$ and, moreover,

(iv) If $h_{z,f,D}(z) < 0$ at some state z, then the relation

$$h_{x,f,D+b}(x) < 0 \text{ for all } x \in S$$

holds for some $b > 0$.

Proof. (i) Let D and D_1 be two real valued functions such that $D_1(x) \geq D(x)$ for every x, with strict inequality at $y \in S$, and suppose that $h_{z,f,D}(z) < \infty$; since the inequality $h_{z,f,D_1}(z) > h_{z,f,D}(z)$ is clear when $y = z$, assume that $y \neq z$. In this case

$$e^{\lambda \sum_{t=0}^{T_z-1} D_1(X_t)} \geq e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)}$$

with strict inequality on the event $T_y < T_z$, which has positive probability with respect to P_z^f, by Lemma 2.6.1. It follows that

$$e^{\lambda h_{z,f,D_1}(z)} = E_z^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D_1(X_t)}\right] > E_z^f\left[e^{\lambda \sum_{t=0}^{T_z-1} D(X_t)}\right] = e^{\lambda h_{z,f,D}(z)},$$

and then $h_{z,f,D_1}(z) > h_{z,f,D}(z)$.
(ii) Let $z \in S$ and notice that $a \geq 0$. From (2.6.2) it follows that

$$1 = e^{\lambda h_{z,f,D}(z)+\lambda a} = e^{\lambda(D(z)+a)}\left[p_{z\,z}(f(x)) + \sum_{y \in S\setminus\{z\}} p_{z\,y}(f(x)) e^{\lambda h_{z,f,D}(y)}\right] \tag{2.6.8}$$

Now, define $D_1, h_1\colon S \to \mathbb{R}$ as follows:

$$\begin{gathered} D_1(x) = D(x) \text{ and } h_1(x) = h_{z,f,D}(x) \text{ if } x \neq z, \\ D_1(z) = D(z) + a \text{ and } h_1(x) = 0 \text{ if } x = z; \end{gathered} \tag{2.6.9}$$

notice that

$$D_1 = D + a\delta_z, \tag{2.6.10}$$

by (2.6.6). From the specifications of D_1 and h_1, it follows *via* (2.6.2) and (2.6.8), that for every state x

$$e^{\lambda h_1(x)} = e^{\lambda D_1(x)} \sum_{y \in S\setminus\{z\}} p_{x\,y}(f(x)) e^{\lambda h_1(y)},$$

that is,

$$e^{\lambda h_1(x)} = E_x^f\left[e^{\lambda D_1(X_0)+\lambda h_1(X_1)}\right], \quad x \in S. \qquad (2.6.11),$$

Next, let $y \in S$ be arbitrary but fixed. It will be proved, by induction, that

$$\begin{aligned} e^{\lambda h_1(y)} = \sum_{r=1}^{n} & E_y^f\left[e^{\sum_{t=0}^{T_y-1} \lambda D_1(X_t)+\lambda h_1(y)} I[T_y = r]\right] \\ & + E_y^f\left[e^{\sum_{t=0}^{n-1} \lambda D_1(X_t)+\lambda h_1(X_n)} I[T_y > n]\right]. \end{aligned} \tag{2.6.12}$$

for every positive integer n. To establish this assertion, notice that (2.6.11) yields that

$$\begin{aligned} e^{\lambda h_1(y)} &= E_y^f\left[e^{\lambda D_1(X_0)+\lambda h_1(X_1)}\right] \\ &= E_y^f\left[e^{\lambda D_1(X_0)+\lambda h_1(X_1)} I[T_y = 1]\right] + E_y^f\left[e^{\lambda D_1(X_0)+\lambda h_1(X_1)} I[T_y > 1]\right] \\ &= E_y^f\left[e^{\lambda D_1(X_0)+\lambda h_1(y)} I[T_y = 1]\right] + E_y^f\left[e^{\lambda D_1(X_0)+\lambda h_1(X_1)} I[T_y > 1]\right] \end{aligned}$$

where the fact that $X_1 = y$ on the event $[T_y = 1]$ was used to set the third equality; this verifies assertion (2.6.12) when $n = 1$. Suppose now that (2.6.12) holds for certain positive integer n, and notice that (2.6.11) and the Markov property yield that, with probability 1 with respect to P_y^f,

$$\begin{aligned} e^{\lambda h_1(X_n)} &= E_{X_n}^f\left[e^{\lambda D_1(X_n)+\lambda h_1(X_{n+1})}\right] \\ &= E_y^f\left[e^{\lambda D_1(X_n)+\lambda h_1(X_{n+1})}\middle| H_n\right], \end{aligned}$$

and then, since $e^{\sum_{t=0}^{n-1}\lambda D_1(X_t)}$ and $I[T_y > n]$ are $\sigma(H_n)$-measurable, it follows that

$$\begin{aligned} e^{\sum_{t=0}^{n-1}\lambda D_1(X_t)+\lambda h_1(X_n)}I[T_y > n] &= e^{\sum_{t=0}^{n-1}\lambda D_1(X_t)}I[T_y > n]e^{\lambda h_1(X_n)} \\ &= e^{\sum_{t=0}^{n-1}\lambda D_1(X_t)}I[T_y > n]E_y^f\left[e^{\lambda D_1(X_n)+\lambda h_1(X_{n+1})}\middle| H_n\right] \\ &= E_y^f\left[e^{\sum_{t=0}^{n-1}\lambda D_1(X_t)}I[T_y > n]e^{\lambda D_1(X_n)+\lambda h_1(X_{n+1})}\middle| H_n\right] \\ &= E_y^f\left[e^{\sum_{t=0}^{n}\lambda D_1(X_t)+\lambda h_1(X_{n+1})}I[T_y > n]\middle| H_n\right]; \end{aligned}$$

taking expectation with respect to P_y^f, this leads to

$$\begin{aligned} E_y^f\left[e^{\sum_{t=0}^{n-1}\lambda D_1(X_t)+\lambda h_1(X_n)}I[T_y > n]\right] &= E_y^f\left[e^{\sum_{t=0}^{n}\lambda D_1(X_t)+\lambda h_1(X_{n+1})}I[T_y > n]\right] \\ &= E_y^f\left[e^{\sum_{t=0}^{n}\lambda D_1(X_t)+\lambda h_1(X_{n+1})}I[T_y = n+1]\right] \\ &\quad + E_y^f\left[e^{\sum_{t=0}^{n}\lambda D_1(X_t)+\lambda h_1(X_{n+1})}I[T_y > n+1]\right] \\ &= E_y^f\left[e^{\sum_{t=0}^{T_y-1}\lambda D_1(X_t)+\lambda h_1(y)}I[T_y = n+1]\right] \\ &\quad + E_y^f\left[e^{\sum_{t=0}^{n}\lambda D_1(X_t)+\lambda h_1(X_{n+1})}I[T_y > n+1]\right]. \end{aligned}$$

Combining this expression with the induction hypothesis, it follows that (2.6.12) is also valid with $n + 1$ instead of n. Using (2.6.12) it follows that

$$e^{\lambda h_1(y)} \geq \sum_{r=1}^{n} E_y^f\left[e^{\sum_{t=0}^{T_y-1}\lambda D_1(X_t)+\lambda h_1(y)}I[T_y = r]\right],$$

and then

$$\begin{aligned} e^{\lambda h_1(y)} &\geq \sum_{r=1}^{\infty} E_y^f\left[e^{\sum_{t=0}^{T_y-1}\lambda D_1(X_t)+\lambda h_1(y)}I[T_y = r]\right] \\ &= E_y^f\left[e^{\sum_{t=0}^{T_y-1}\lambda D_1(X_t)+\lambda h_1(y)}I[T_y < \infty]\right] \\ &= E_y^f\left[e^{\sum_{t=0}^{T_y-1}\lambda D_1(X_t)+\lambda h_1(y)}\right] \\ &= e^{\lambda h_1(y)}E_y^f\left[e^{\sum_{t=0}^{T_y-1}\lambda D_1(X_t)}\right] \end{aligned}$$

where the second equality is due to the fact that, for a communicating Markov chain over a finite state space, the return time to a given state is finite with probability 1 (Loève, 1980, Billingsley, 2010). The above relation yields that $1 \geq E_y^f\left[e^{\sum_{t=0}^{T_y-1}\lambda D_1(X_t)}\right] = e^{\lambda h_{y,f,D_1}(y)}$, which in turn leads to

$$h_{y,f,D+a\delta_z}(y) = h_{y,f,D_1}(y) \leq 0, \tag{2.6.13}$$

where the equality is due to (2.6.10); the conclusion follows since $y \in S$ is arbitrary.

(iii) Suppose that $h_{z,f,D}(z) < 0$, so that the number a in (2.6.7) is positive. Setting $b = a/2$ it follows that $D(x)+b\delta_z(x) \le D(x)+a\delta_z(x)$ for every $x \in S$, with strict inequality for $x = z$. In this case, part (i) yields that for every $y \in S$ the inequality $h_{y,f,D+b\delta_z}(y) < h_{y,f,D+a\delta_z}(y)$ holds, and the conclusion follows form (2.6.13).

(iv) Suppose that $h_{z,f,D}(z) < 0$ and write $S = \{x_1, x_1, \ldots, x_k\}$ where $z = x_1$, so that $h_{x_1,f,D}(x_1) < 0$. In this case, an application of part (iii) yields that there exists $b_1 > 0$ such that

$$h_{y,f,D+b_1\delta_{x_1}}(y) < 0, \quad y \in S.$$

In particular, $h_{x_2,f,D+b_1\delta_{x_1}}(x_2) < 0$, and part (iii) with x_2 and $D + b_1\delta_{x_1}$ instead of z and D, respectively, yields that there exists $b_2 > 0$ such that

$$h_{y,f,D+b_1\delta_{x_1}+b_2\delta_{x_2}}(y) < 0, \quad y \in S.$$

Repeating this argument it follows that there exist positive constants $b_1, b_2, \ldots b_k$ such that

$$h_{y,f,D+b_1\delta_{x_1}+b_2\delta_{x_2}+\cdots+b_k\delta_{x_k}}(y) < 0 \text{ for all } y \in S. \tag{2.6.14}$$

Setting $b = \min\{b_1, b_2, \ldots, b_k\}/2$, it follows that b is positive and that

$$D + b < D + b_1\delta_{x_1} + b_2\delta_{x_2} + \cdots + b_k\delta_{x_k};$$

from this point, part (i) yields that $h_{y,f,D+b}(y) < h_{y,f,D+b_1\delta_{x_1}+b_2\delta_{x_2}+\cdots+b_k\delta_{x_k}}(y)$ for every state y, and then (2.6.14) implies that $h_{y,f,D+b}(y) < 0$. □

2.7. Relative Costs up to a Return Time

This section contains the results about relative costs that will be used to establish Theorem 2.5.1. To begin with, for each stationary policy $f \in \mathbb{F}$ define the section $C_f \colon S \to \mathbb{R}$ of the cost function C by

$$C_f(x) = C(x, f(x)), \quad x \in S, \tag{2.7.1}$$

and let $z \in S$ be an arbitrary state that will fixed throughout the remainder. Given $\gamma \in \mathbb{R}$, the total cost relative to γ incurred up to the first return to state z is $\sum_{t=0}^{T_z-1}[C_f(X_t) - \gamma]$, and $R_{f,\gamma}(x)$ stands for the certain equivalent of this quantity when x is the initial state; more explicitly,

$$R_{f,\gamma}(x) = \frac{1}{\lambda} \log\left(E_x^f\left[e^{\lambda \sum_{t=0}^{T_z-1}[C_f(X_t)-\gamma]} \right]\right), \quad x \in S. \tag{2.7.2}$$

From this specification it is not difficult to see that $R_{f,\gamma}$ is a monotone decreasing function of γ, that is,

$$R_{f,\gamma} \ge R_{f,\gamma_1}, \quad \gamma \le \gamma_1, \tag{2.7.3}$$

as well as

$$R_{f,\|C_f\|} \le 0, \quad \text{and} \quad R_{f,-\|C_f\|-1} > 0, \tag{2.7.4}$$

since $C_f - \|C_f\| \le 0$ and $C_f + \|C_f\| + 1 > 0$. Also, notice that with the notation in the previous section

$$R_{f,\gamma}(\cdot) = h_{z,f,C_f-\gamma}(\cdot). \tag{2.7.5}$$

The following result, which a a consequence of the communication property in Assumption 2.5.1(iii), will play an essential role in the proof of Theorem 2.5.1, Define the set $G(f)$ by

$$G(f) := \{\gamma \in \mathbb{R} \,|\, R_{f,\gamma}(z) \leq 0\}. \tag{2.7.6}$$

and

$$\gamma(f) = \min G(f) < \infty. \tag{2.7.7}$$

Theorem 2.7.1. Let $f \in \mathbb{F}$ and $z \in S$ be fixed. With the notation in (2.7.1)–(2.7.7), the following assertions (i)–(ii) hold:

(i) $\gamma(f)$ is finite, satisfies that the inequality $\gamma(f) \geq -\|C_f\| - 1$, and

$$G(f) = [\gamma(f), \infty). \tag{2.7.8}$$

(ii) $R_{f,\gamma(f)}(z) = 0$, and then

(iii) The following (Poisson) equation holds:

$$e^{\lambda R_{f,\gamma(f)}(x)} = e^{\lambda[C_f(x)-\gamma(f)]} \sum_{y\in S} p_{x\,y}(f(x)) e^{\lambda R_{f,\gamma(f)}(y)}, \quad x \in S. \tag{2.7.9}$$

Proof. (i) Notice that (2.7.4) and (2.7.6) together yield that $G(f)$ is a nonempty and proper subset of $\mathbb{R}$ contained in $(-\|C_f\| - 1, \infty)$; moreover, using (2.7.3) it follows that $G(f)$ is an interval extending to ∞, so that $G(f) = (\gamma(f), \infty)$ or $G(f) = [\gamma(f), \infty)$, where $\gamma(f) > -\|C_f\| - 1$. Therefore, to establish (2.7.8) it is sufficient to show that $\gamma(f) \in G(f)$. To achieve this goal select a sequence $\{\gamma_n\} \subset G(f)$ such that $\gamma_n \searrow \gamma(f)$. In this case $R_{f,\gamma_n}(z) \leq 0$ and $e^{\lambda \sum_{t=0}^{T_z-1}[C_f(X_t)-\gamma_n]} \nearrow e^{\lambda \sum_{t=0}^{T_z-1}[C_f(X_t)-\gamma(f)]}$, so that the monotone convergence theorem leads to

$$\begin{aligned} 1 &\geq e^{\lambda R_{f,\gamma_n}(z)} \\ &= E_z^f\left[e^{\lambda \sum_{t=0}^{T_z-1}[C_f(X_t)-\gamma_n]}\right] \nearrow E_z^f\left[e^{\lambda \sum_{t=0}^{T_z-1}[C_f(X_t)-\gamma(f)]}\right] = e^{\lambda R_{f,\gamma(f)}(z)}, \end{aligned}$$

and then $R_{f,\gamma(f)}(z) \leq 0$, so that $\gamma(f) \in G(f)$.

(ii) Assume that $R_{f,\gamma(f)}(z) < 0$, In this case, using that $R_{f,\gamma(f)} = h_{z,f,C_f-\gamma(f)}$, an application of Theorem 2.6.1(iv) with the function $C_f - \gamma(f)$ instead of D yields that there exists a positive number b such that

$$0 > h_{z,f,C_f-\gamma(f)+b}(z) = h_{z,f,C_f-(\gamma(f)-b)}(z) = R_{f,\gamma(f)-b}(z),$$

so that $\gamma(f) - b \in G(f)$, which is a contradiction; see (2.7.7). Thus, $R_{f,\gamma(f)}(z) \geq 0$, and the conclusion follows since $R_{f,\gamma(f)}(z) \leq 0$, by part (i).

(iii) Using (2.7.5), the desired conclusion (2.7.9) is equivalent to the dynamic programming equation (2.6.3) applied with the function $C_f - \gamma(f)$ instead of D. □

In the above theorem the policy $f \in \mathbb{F}$ is fixed, and the conclusions depend only on the communication property of the Markov chain induced by f. The following result uses the full strength of Assumption 2.5.1. Set

$$g^* := \inf_{f\in\mathbb{F}} \gamma(f). \tag{2.7.10}$$

which is a finite number. Next, select a sequence of policies $\{f_n\} \subset \mathbb{F}$ such that

$$\gamma_n \equiv \gamma(f_n) \to g^* \text{ as } n \to \infty. \tag{2.7.11}$$

Since $\mathbb{F} = \prod_{x \in S} A(x)$ is a compact metric space, by Assumption 2.5.1(i), taking a subsequence if necessary, without loss of generality it can be assumed that for a policy $f^* \in \mathbb{F}$,

$$\lim_{n \to \infty} f_n(x) = f^*(x), \quad x \in S. \tag{2.7.12}$$

Theorem 2.7.2. Suppose that Assumption 2.5.1 holds. In this case assertions (i) and (ii) below are valid.
(i) For each $x \in S$ and $r = 1, 2, 3, \ldots$,

$$\lim_{n\to\infty} E_x^{f_n}\left[e^{\lambda \sum_{t=0}^{r-1}[C_{f_n}(X_t)-\gamma(f_n)]} I[T_z = r]\right] = E_x^{f^*}\left[e^{\lambda \sum_{t=0}^{r-1}[C_{f^*}(X_t)-g^*]} I[T_z = r]\right]$$

Consequently,
(ii) $1 \geq E_z^{f^*}\left[e^{\lambda \sum_{t=0}^{T_z-1}[C_{f^*}(X_t)-g^*]}\right]$, and then
(iii) $g^* = \gamma(f^*)$.

Proof. (i) Let $x \in S$ be fixed and, for a positive integer r, let S_r stand for the class of all trajectories $(x_0, x_1, \ldots, x_r) \in S \times S \times \cdots \times S$ satisfying that $x_0 = x$, $x_r = z$ and $x_t \neq z$ for $1 \leq t < r$. Notice now that, for every positive integer n,

$$\begin{aligned}
&E_x^{f_n}\left[e^{\lambda \sum_{t=0}^{r-1}[C_{f_n}(X_t)-\gamma(f_n)]} I[T_z = r]\right] \\
&\qquad = \sum_{(x_0,x_1,x_2,\ldots,x_r)\in S_r} e^{\lambda \sum_{t=0}^{r-1}[C_{f_n}(x_t)-\gamma(f_n)]} \prod_{i=1}^{r} p_{x_{i-1}\,x_i}(f_n(x_{i-1})) \\
&\qquad = \sum_{(x_0,x_1,x_2,\ldots,x_r)\in S_r} e^{\lambda \sum_{t=0}^{r-1}[C(x_t f_n(x_t))-\gamma(f_n)]} \prod_{i=1}^{r} p_{x_{i-1}\,x_i}(f_n(x_{i-1}));
\end{aligned}$$

where (2.7.1) was used to set the second equality. Observe now that the continuity properties in Assumption 2.5.1 together with the convergences (2.7.10) and (2.7.12) yield that

$$\begin{aligned}
&\lim_{n\to\infty} E_x^{f_n}\left[e^{\lambda \sum_{t=0}^{r-1}[C_{f_n}(X_t)-\gamma(f_n)]} I[T_z = r]\right] \\
&\qquad = \lim_{n\to\infty} \sum_{(x_0,x_1,x_2,\ldots,x_n)\in S_r} e^{\lambda \sum_{t=0}^{r-1}[C(x_t,f_n(x_t))-\gamma(f_n)]} \prod_{i=1}^{r} p_{x_{i-1}\,x_i}(f_n(x_{i-1})) \\
&\qquad = \sum_{(x_0,x_1,x_2,\ldots,x_r)\in S_r} e^{\lambda \sum_{t=0}^{r-1}[C(x_t,f^*(x_t))-g^*]} \prod_{i=1}^{r} p_{x_{i-1}\,x_i}(f^*(x_{i-1})) \\
&\qquad = E_x^{f^*}\left[e^{\lambda \sum_{t=0}^{r-1}[C_{f^*}(X_t)-g^*]} I[T_z = r]\right]
\end{aligned}$$

(ii) Recall that $R_{f_n,\gamma(f_n)}(z) = 0$, by Theorem 2.7.1(ii), so that (2.7.2) yields that

$$\begin{aligned}
1 &= e^{R_{f_n,\gamma(f_n)}(z)} \\
&= E_z^{f_n}\left[e^{\lambda \sum_{t=0}^{T_z-1}[C_{f_n}(X_t)-\gamma(f_n)]}\right] \\
&= \sum_{r=1}^{\infty} E_z^{f_n}\left[e^{\lambda \sum_{t=0}^{r-1}[C_{f_n}(X_t)-\gamma(f_n)]} I[T_z = r]\right]
\end{aligned}$$

It follows that $1 \geq \sum_{r=1}^{m} E_z^{f_n}\left[e^{\lambda\sum_{t=0}^{r-1}[C_{f_n}(X_t)-\gamma(f_n)]}I[T_z=r]\right]$ for every positive m; taking the limit as n goes to ∞ in this last inequality, part (i) yields that

$$1 \geq \sum_{r=1}^{m} E_x^{f^*}\left[e^{\lambda\sum_{t=0}^{r-1}[C_{f^*}(X_t)-g^*]}I[T_z=r]\right]$$

and then, since m is arbitrary,

$$1 \geq \sum_{r=1}^{\infty} E_x^{f^*}\left[e^{\lambda\sum_{t=0}^{n-1}[C_{f^*}(X_t)-g^*]}I[T_z=r]\right] = E_x^{f^*}\left[e^{\lambda\sum_{t=0}^{n-1}[C_{f^*}(X_t)-g^*]}\right].$$

(iii) Notice that part (ii) yields that $R_{f^*,g^*}(z) \leq 0$, a relation that combined with (2.7.6) and (2.7.8) implies that $g^* \geq \gamma(f^*)$, and it follows that $g^* = \gamma(f^*)$, since g^* is the minimum value of the mapping $f \mapsto \gamma(f)$ over $f \in \mathbb{F}$. □

2.8. Proof of Theorem 2.5.1

In this section the existence of a solution to the optimality equation is established uder Assumption 2.5.1. To begin with, let $g^* \in \mathbb{R}$ and $f^* \in \mathbb{F}$ be as in (2.7.10)–(2.7.12). Since $g^* = \gamma(f^*)$, Theorem 2.7.1(iii) and (2.7.5) yield that, for every $x \in S$,

$$\begin{aligned} e^{\lambda g^*+\lambda R_{f^*,g^*}(x)} &= e^{\lambda C_{f^*}(x)}\sum_{y\in S} p_{x\,y}(f^*(x))e^{\lambda R_{f^*,g^*}(y)} \\ &= e^{\lambda C(x,f^*(x))}\sum_{y\in S} p_{x\,y}(f^*(x))e^{\lambda R_{f^*,g^*}(y)}, \quad x \in S \end{aligned} \tag{2.8.1}$$

where the second equality is due to the specification of the function C_{f^*}; see (2.7.1).

Proof of Theorem 2.5.1. Suppose that Assumption 2.5.1 holds. It will be shown that the pair $(g^*, R_{f^*,g^*}(\cdot))$ satisfies the optimality equation. To begin with observe that, for each $x \in S$, the mapping $a \mapsto e^{\lambda C(x,a)}\sum_{y\in S} p_{x\,y}(a)e^{\lambda R_{f^*,g^*}(y)}$ is continuous in $a \in A(x)$, and then such a function attains its minimum at some action $\tilde{f}(x) \in A(x)$, since the $A(x)$ is compact. Thus, the stationary policy $\tilde{f}$ satisfies

$$\begin{aligned} &e^{\lambda C(x,\tilde{f}(x))}\sum_{y\in S} p_{x\,y}(\tilde{f}(x))e^{\lambda R_{f^*,g^*}(y)} \\ &\qquad = \min_{a\in A(x)}\left[e^{\lambda C(x,a)}\sum_{y\in S} p_{x\,y}(a)e^{\lambda R_{f^*,g^*}(y)}\right], \quad x \in S, \end{aligned} \tag{2.8.2}$$

an equality that together with (2.8.1) yields that $e^{\lambda C(x,\tilde{f}(x))}\sum_{y\in S} p_{x\,y}(\tilde{f}(x))e^{\lambda R_{f^*,g^*}(y)} \leq e^{\lambda g^*+\lambda R_{f^*,g^*}(x)}$ for every state x, and then, there exists a function $\Delta\colon S \to [0,\infty)$ such that

$$e^{\lambda g^*+\lambda R_{f^*,g^*}(x)} = e^{\lambda[C(x,\tilde{f}(x))+\Delta(x)]}\sum_{y\in S} p_{x\,y}(\tilde{f}(x))e^{\lambda R_{f^*,g^*}(y)}, \quad x \in S, \tag{2.8.3}$$

a relation that is equivalent to

$$e^{\lambda R_{f^*,g^*}(x)} = E_x^{\tilde{f}}\left[e^{\lambda[C(X_0,\tilde{f}(X_0))+\Delta(X_0)-g^*]+\lambda R_{f^*,g^*}(X_1)}\right],$$

and then

$$\begin{aligned} e^{\lambda R_{f^*,g^*}(x)} &= E_x^{\tilde f}\left[e^{\lambda[C_{\tilde f}(X_0)+\Delta(X_0)-g^*]+\lambda R_{f^*,g^*}(X_1)}\right] \\ &= E_x^{\tilde f}\left[e^{\lambda[C_{\tilde f}(X_0)+\Delta(X_0)-g^*]+\lambda R_{f^*,g^*}(X_1)} I[T_z=1]\right] \\ &\qquad\qquad + E_x^{\tilde f}\left[e^{\lambda[C_{\tilde f}(X_0)+\Delta(X_0)-g^*]+\lambda R_{f^*,g^*}(X_1)} I[T_z>1]\right], \quad x\in S. \end{aligned}$$

Combining the two last displays, an induction argument using the Markov property yields that for every positive integer n and $x \in S$

$$\begin{aligned} e^{\lambda R_{f^*,g^*}(x)} &\geq \sum_{r=1}^{n} E_x^{\tilde f}\left[e^{\lambda\sum_{t=0}^{r-1}[C_{\tilde f}(X_t)+\Delta(X_t)-g^*]+\lambda R_{f^*,g^*}(X_r)} I[T_z=r]\right] \\ &\qquad\qquad + E_x^{\tilde f}\left[e^{\lambda[C_{\tilde f}(X_n)+\Delta(X_n)-g^*]+\lambda R_{f^*,g^*}(X_n)} I[T_z>n]\right], \end{aligned}$$

so that

$$\begin{aligned} e^{\lambda R_{f^*,g^*}(x)} &\geq \lim_{n\to\infty}\sum_{r=1}^{n} E_x^{\tilde f}\left[e^{\lambda\sum_{t=0}^{r-1}[C_{\tilde f}(X_t)+\Delta(X_t)-g^*]+\lambda R_{f^*,g^*}(X_r)} I[T_z=r]\right] \\ &= \sum_{r=1}^{\infty} E_x^{\tilde f}\left[e^{\lambda\sum_{t=0}^{r-1}[C_{\tilde f}(X_t)+\Delta(X_t)-g^*]+\lambda R_{f^*,g^*}(X_r)} I[T_z=r]\right] \\ &= E_x^{\tilde f}\left[e^{\lambda\sum_{t=0}^{T_z-1}[C_{\tilde f}(X_t)+\Delta(X_t)-g^*]+\lambda R_{f^*,g^*}(z)}\right]. \end{aligned}$$

Setting $x = z$, the above dispay leads to $1 \geq E_x^{\tilde f}\left[e^{\lambda\sum_{t=0}^{T_z-1}[C_{\tilde f}(X_t)+\Delta(X_t)-g^*]}\right]$, that is, $1 \geq e^{\lambda h_{z,\tilde f,[C_{\tilde f}+\Delta-g^*]}(z)}$, by Definition 2.6.2, and then

$$h_{z,\tilde f,[C_{\tilde f}+\Delta-g^*]} \leq 0.$$

Suppose now that the nonnegative function Δ is positive at some state. In this case, Theorem 2.6.1(i) yields that $h_{z,\tilde f,[C_{\tilde f}+\Delta-g^*]}(z) > h_{z,\tilde f,[C_{\tilde f}-g^*]}(z)$, an inequality that together with the above display implies that

$$R_{\tilde f,g^*}(z) = h_{z,\tilde f,[C_{\tilde f}-g^*]}(z) < 0,$$

where the inequality is due to (2.7.5). It follows that

$$g^* \in G(\tilde f) = [\gamma(\tilde f), \infty),$$

by (2.7.6) and Theorem 2.7.1(i); since$R_{\tilde f,\gamma(\tilde f)}(z) = 1$, by Theorem (ii), the two last displays together imply that $\gamma(\tilde f) < g^*$, which a contradiction, since g^* is the minimum value of the mapping $f \mapsto \gamma(f)$ over $f \in \mathbb{F}$. Consequently, $\Delta(x) = 0$ for every state x, so that

$$e^{\lambda g^*+\lambda R_{f^*,g^*}(x)} = e^{\lambda C(x,\tilde f(x))}\sum_{y\in S} p_{x\,y}(\tilde f(x))e^{\lambda R_{f^*,g^*}(y)} \quad \text{for all } x \in S;$$

see (2.8.3). Combining this property with (2.8.2) it follows that

$$e^{\lambda g^*+\lambda R_{f^*,g^*}(x)} = \min_{a\in A(x)}\left[e^{\lambda C(x,a)}\sum_{y\in S} p_{x\,y}(a)e^{\lambda R_{f^*,g^*}(y)}\right], \quad x \in S,$$

that is, the pair (g^*, R_{f^*,g^*}) satisfies the optimality equation. □

2.9. The Communication Property and the Optimality Equation

In this chapter two theorems on the existence of solutions of the optimality equation were studied, and the proof of both results relied on the communication assumption of the Markov chains induced by any stationary policy. For the algebraic approach by Howard and Matheson, the essential instrument of analysis was (the Perron-Forbenious) Theorem 2.3.1 asserting that, for a nonnegative and communicating matrix A, the following properties (a) and (b) hold:

(a) There exists a positive eigenpair $(\mu, \mathbf{m})$ for A, and

(b) That a nonnegative sub-(or super-) eigenvector $\mathbf{x}$ corresponding to μ—that is, a vector $\mathbf{x} \in [0, \infty)$ satsifying $A\mathbf{x} \leq \mu\mathbf{x}$ or $A\mathbf{X} \geq \mathbf{x}$—necessarily belongs to the eigenspace of μ.

Although a positive eigenpair for a nonnegative matrix may exist even if A is noncommunicating, in general the above property (b) depends on the communication property of the matrix A.

On the other hand, the argument used to establish Theorem 2.5.1 relied on the basic property in Lemma 2.6.1, which is a consequence of the communication condition (2.2.1). In the following chapter, a result on the characterization of the optimal average cost for systems satisfying a weak form of communication will be analyzed.

Chapter 3

The Optimality Equation for Markov Decision Chains with an Accessible State

In this chapter a class of models satisfying a weak form of communication is studied. Such a requirement is formulated in terms of the existence of a fixed state, say $z \in S$, such that under the action of each stationary policy and regardless of the initial state, z is visited with positive probability. This assumption is a form of the simultaneous Doeblin condition, which has been widely used in the analysis of Markov decision chains with risk-neutral average index. The main objective of the presentation is to show that a solution to the optimality equation can be generally ensured only when the risk-aversion coefficient is 'small enough', and that, even if the optimal average cost function is constant, it may not be characterized by a single optimality equation. Besides standard dynamic programming ideas, the approach used below relies on elementary continuity arguments, providing an alternative approach to the one used in Cavazos-Cadena (2003) where, under the conditions in this chapter. the optimality equation was studied *via* contractive operators.

3.1. Introduction

This chapter concerns Markov decision models satisfying a weak form of the communication requirement in (2.2.1). The describe the version used below, let $z \in S$ be fixed, and consider the following conditions (a) and (b):

(a) Under the action of each stationary policy, the state z is *accessible* regardless of the initial state, that is, given $x \in S$ and $f \in \mathbb{F}$, there exists a positive integer n such that $P_x^f[X_n = z] > 0$, and

(b) For each $x \in S$ and $f \in \mathbb{F}$, there exists a positive integer m such that $P_z^f[X_m = x] > 0$.

It is not difficult to verify that the communication property (2.2.1) is equivalent to the occurrence of *both* conditions (a) and (b). On the other hand, under condition (a) each stationary policy has a single recurrent class which always contains z, properties that for the risk-neutral average criterion ensure that the corresponding optimality equation has a solution. However, in the risk-averse context of this work, it will be shown that condition (a) alone renders a solution to the optimality equation only if the risk-aversion coefficient λ is 'small enough', and that such an existence result generally fails for arbitrary $\lambda > 0$. Moreover, a simple example will be used to show that, even is the (optimal) average cost function is constant, in general it is not characterized by a single optimality equation.

The organization of the subsequent material is as follows: In Section 2 the accessibility condition is formally introduced, the result on the existence of solutions to the optimality equation is stated as Theorem 3.2.1, and an example is given to illustrate the lack of solutions when $\lambda > 0$ is arbitrary. Next, in Section 3 it is shown that when the state z is accessible, the right-tails of the distribution of the return time T_z decay at a geometric rate, a conclusion that is used in Section 4 to study the total relative utility of the cost incurred before the first return to z. Finally, Theorem 3.2.1 is proved in Section 5 and the presentation concludes in Section 6 with some brief comments.

3.2. Models with an Accessible State

In this section the risk-sensitive average optimality equation will be studied under the following condition on the transition law: There exists a (fixed) state $z \in S$ such that, under the action of any stationary policy, the state z can be reached with positive probability regardless of the initial state, a requirement that, together with mild continuity-compactness conditions on the decision model, is formally stated below.

Assumption 3.2.1. *(i) The state space S is finite and $A(x)$ is a compact set for each $x \in S$; (ii) For each $x, y \in S$ the mappings $a \mapsto C(x, a)$ and $a \mapsto p_{x\,y}(a)$ are continuous in $a \in A(x)$; (iii) For some state $z \in S$, the following property occurs::*

$$\begin{gathered}\text{Given } x \in S \text{ and } f \in \mathbb{F}, \text{ there exists a positive integer } n(x, f) \equiv n \\ \text{such that } P_x^f[X_n = z] > 0.\end{gathered} \tag{3.2.1}$$

Throughout the remainder, z is a (fixed) state such that the above condition (3.2.1) holds. The following theorem establishes that Assumption 3.2.1 guarantees the existence of solutions to the optimality equation whenever the risk-sensitivity coefficient λ is sufficiently close to zero.

Theorem 3.2.1. Let $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, P, C)$ be a Markov decision chain satisfying Assumption 3.2.1. In this case, there exists a positive number β such that

If $\lambda \in (0, \beta)$, then the optimality equation

$$e^{\lambda g + \lambda h(x)} = \min_{a\in A(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S} p_{x\,y}(a) e^{\lambda h(y)} \right], \quad x \in S,$$

is satisfied by some pair $(g, h(\cdot)) \in \mathbb{R} \times \mathcal{B}(S)$.

This result was originally obtained in Cavazos-Cadena (2003), where the conclusions were derived using the 'discounted approach', a technique that will play an important role in Chapter 4, were it will be used to establish the main contribution of this thesis. On the other hand, in the subsequent development a different method, highlighting the role of the accessible state z, will be used to establish Theorem 3.2.1 in Section 5; the argument relies on a simple probabilistic analysis of condition (3.2.1) presented in Section 3, which will be used in Section 4 to study an auxiliary total cost problem. Notice that the accessibility condition in Assumption 3.2.1(iii) is weaker that the communication requirement (2.2.1) and, accordingly, the above theorem ensures the existence of a solution to the optimality equation only when the risk sensitivity coefficient λ is 'small enough', that is, less than a certain positive number β, which will be specified during the proof of Theorem 3.2.1. Before going any further, it is interesting to see whether or not, within the framework of Assumption 3.2.1, a solution to the optimality equation can be obtained for *arbitrary* $\lambda \in (0, \infty)$ and not only for λ 'close' to zero. The following example shows that, in general, the answer is negative.

Example 3.2.1. Consider a Markov chain over the state space $S = \{0, 1\}$, with transition probabilities determined by

$$p_{1\,0} = 0.5 = p_{1\,1} \quad \text{and} \quad p_{0\,0} = 1,$$

whereas the cost function $C\colon S \to \mathbb{R}$ be given by

$$C(0) = 0, \quad C(1) = 1.$$

Notice that these specifications determine a Markov decision chain for which the action set consists of a unique action, and that Assumption 3.2.1 holds in this example. □

In this example the state 0 is absorbing (*i.e.*, $p_{0\,0} = 1$) and $C(0) = 0$, so that

$$J(0) = 0 \tag{3.2.2}$$

regardless of λ.

Proposition 3.2.1. Let $\lambda > 0$ be arbitrary. For the model in Example 3.2.1, the (Poisson) equation

$$e^{\lambda g + \lambda h(x)} = e^{\lambda C(x)} \sum_{y \in S} p_{x\,y} e^{\lambda h(y)}, \quad x \in S, \tag{3.2.3}$$

is satisfied by a pair $(g, h(\cdot)) \in \mathbb{R} \times \mathcal{B}(S)$ if, and only if,

$$\lambda < \log(2).$$

Proof. The Poisson equation (3.2.3) can be explicitly written as

$$\begin{aligned} e^{\lambda g + \lambda h(1)} &= e^{\lambda} \left[.0.5 e^{\lambda h(1)} + 0.5 e^{\lambda h(0)} \right] \\ e^{\lambda g + \lambda h(0)} &= e^{\lambda h(0)} \end{aligned} \tag{3.2.4}$$

The second equality yields that $g = 0$, and then the above system of two scalar equations is equivalent to the single equation

$$\left(1 - \frac{e^{\lambda}}{2}\right) e^{\lambda[h(1) - h(0)]} = \frac{e^{\lambda}}{2}$$

If $\lambda \geq \log(2)$, then the left-hand side of this equality is less than or equal to zero, whereas the right-hand side is positive, so that no solution exists in this case.
When $\lambda < \log(2)$, the above equation holds with

$$h(0) = 0 \quad \text{and} \quad h(1) = \frac{1}{\lambda} \log\left(\frac{e^{\lambda}/2}{1 - e^{\lambda}/2}\right),$$

and the equalities (3.2.4) are satisfied by $(g, h(\cdot))$, where $g = 0$ and $h(\cdot)$ is given above. □

Next, in the following proposition the average cost will be determined when (3.2.3) does not have a solution.

Proposition 3.2.2. In the context of Example 3.2.1, let the risk sensitivity coefficient λ be such that

$$\lambda \geq \log(2).$$

In this context,

$$J(1) = \frac{1}{\lambda}\log\left(\frac{e^\lambda}{2}\right);$$

particularly,

$$J(1) = 0 \text{ if } \lambda = \log(2).$$

Proof. When the initial state is 1, notice that that $C(X_t) = 1$ before the system arrives to state 0, that is, when $t < T_0$, whereas $C(X_t) = 0$ when $t \geq T_0$, so that $\sum_{t=0}^{n-1} C(X_t) = \min\{T_z, n\} \equiv T_z \wedge n$, and then

$$\begin{aligned}
e^{\lambda J_n(1)} &= E_1\left[e^{\lambda \sum_{t=0}^{n-1} C(X_t)}\right] \\
&= E_1\left[e^{\lambda T_z \wedge n}\right] \\
&= \sum_{r=1}^{n} E_1\left[e^{\lambda[r\wedge n]} I[T_0 = r]\right] + E_1\left[e^{\lambda n} I[T_0 > n]\right] \\
&= \sum_{r=1}^{n} e^{\lambda r} P_1[T_0 = r] + e^{\lambda n} P_{[}T_0 > n]
\end{aligned}$$

Notice now that the specification of the transition law in Example 3.2.1 yields that

$$P[T_0 = r] = 1/2^r = P[T_0 > r] = 1/2^r, \quad r = 1, 2, 3, \ldots$$

so that

$$e^{\lambda J_n(1)} = \sum_{r=1}^{n} \frac{e^{r\lambda}}{2^r} + \frac{e^{n\lambda}}{2^n}.$$

and then, since $e^\lambda/2 \geq 1$,

$$\left(\frac{e^\lambda}{2}\right)^n \leq e^{\lambda J_n(1)} \leq (n+1)\left(\frac{e^\lambda}{2}\right)^n$$

so that

$$\frac{1}{\lambda}\log\left(\frac{e^\lambda}{2}\right) \leq \frac{1}{n}J_n(1) \leq \frac{1}{n\lambda}\log(n+1) + \frac{1}{\lambda}\log\left(\frac{e^\lambda}{2}\right)$$

and taking the limit as n goes to ∞, this relation yields that

$$J(1) = \lim_{n\to\infty}\frac{1}{n}J_n(1) = \frac{1}{\lambda}\log\left(\frac{e^\lambda}{2}\right);$$

thus, $J(1) = 0$ if $e^\lambda = 2$, and $J(1) > 0$ when $e^\lambda > 2$. □

For the model in Example 3.2.1, Proposition 3.2.2 and the equality (3.2.2) together yield that the average cost is not constant when $\lambda > \log(2)$ and then, in accordance with Proposition 3.2.1, the Poisson equation (3.2.3) does not have a solution (otherwise, the average cost must be constant). On the other hand, when $\lambda = \log(2)$ *the average cost function $J(\cdot)$ is constant* and equal to 0, by Proposition 3.2.2 and (3.2.2), *but the Poisson equation does not have a solution*, by Proposition 3.2.1, showing that, *even if the average cost is constant, its value is not characterized by a single optimality equation* (which in the context of uncontrolled models is referred to as the *Poisson equation*). This is fact provides a strong motivation to pursue the main objective of this thesis, namely, to provide a general characterization of the (λ-sensitive) optimal average reward for arbitrary Markov decision chains over a finite state space.

3.3. The Simultaneous Doeblin Condition

In this section an equivalent formulation of the accessibility property in (3.2.1) is formulated; in the form presented below such a condition is usually referred to as the simultaneous Doeblin condition. First, it is convenient to introduce the following notation.

Definition 3.3.1. Suppose that Assumption 3.2.1 holds, and let $z \in S$ be a fixed state such that (3.2.1) is valid. In this context, for each nonnegative integer n the function $M_n: S \to [0, 1]$ is defined by

$$M_n(x) = \sup_{\pi\in\mathcal{P}} P_x^\pi[T_z > n], \quad x \in S. \tag{3.3.1}$$

Theorem 3.3.1. Under Assumption 3.2.1, the following assertions (i)–(ii) hold:
(i) For each state $x \in S$,

$$\lim_{n\to\infty} M_n(x) = 0.$$

Consequently,
(ii) There exist a positive constant B and $\rho \in (0, 1)$ such that, for every $x \in S$ and $\pi \in \mathcal{P}$,

$$P_x^\pi[T_z > n] \le B\rho^n, \quad x \in S, \quad \pi \in \mathcal{P}, \quad n = 1, 2, 3, \ldots \tag{3.3.2}$$

and then

$$P_x^\pi[T_z = n] \le P_x^\pi[T_z > n - 1] \le \frac{B}{\rho}\rho^n, \quad x \in S, \quad \pi \in \mathcal{P}, \quad n = 1, 2, 3, \ldots \tag{3.3.3}$$

Remark 3.3.1. The relation (3.3.2) implies that, for every initial state x and $\pi \in \mathcal{P}$,

$$E_x^\pi[T_z] = \sum_{n=0}^\infty P_x^\pi[T_z > n] \le \sum_{n=0}^\infty B\rho^n \le B/(1 - \rho) < \infty.$$

Conversely, suppose that the inequality $E_x^\pi[T_z] \le \tilde{B}$ for some constant $\tilde{B}$ regardless of x and π, and let n_0 be a positive integer larger than $\tilde{B}$. In this case, Markov's inequality yields that $P_x^\pi[T_z > n_0] \le \tilde{B}/n_0 = \tilde{\rho} < 1$, and an induction argument using the Markov property of the process allows to conclude that $P_x^\pi[T_z > qn_0] \le \tilde{\rho}^q$ for every positive integer q. Given a positive integer n, write $n = qn_0 + r$ where $0 \le r < n_0$ and notice that

$$P_x^\pi[T_z > n] \le P_x^\pi[T_z > qn_0] \le \tilde{\rho}^q = \tilde{\rho}^{-r}[\tilde{\rho}^{1/n_0}]^n \le \tilde{\rho}^{-n_0}[\tilde{\rho}^{1/n_0}]^n,$$

and it follows that (3.3.2) holds with $B = \tilde{\rho}^{-n}$ and $\rho = \tilde{\rho}^{1/n_0}$. In short, (3.3.2) is equivalent to

$$\sup_{x\in S, \pi\in\mathcal{P}} E_x^\pi[T_z] < \infty,$$

which is the *simultaneous Doeblin condition*. □

The proof of Theorem 3.3.1 relies on the following result.

Lemma 3.3.1. Under Assumption 3.2.1, there exists an integer n_0 such that

$$\sup_{x\in S,\, f\in \mathbb{F}} P_x[T_z > n_0] =: \rho_0 < 1. \tag{3.3.4}$$

Proof. Let $x \in S$ be arbitrary, and denote by S_n the class of trajectories $(x_0, x_1, \ldots, x_n)$ in S^{n+1} such that $x_0 = x$ and $x_t \neq z$ when $1 \leq t \leq n$. Notice now that

$$P_x^f[T_z > n] = \sum_{(x_0,x_1,\ldots,x_n)\in S_n} p_{x\,x_1}(f(x))p_{x_1\,x_2}(f(x))\cdots p_{x_{n-2}\,x_{n-1}}(f(x))p_{x_{n-1}\,y}(f(x)),$$

and then Assumption 3.2.1(i) yields that
(a) The mapping $f \mapsto P_x^f[T_z > n]$ is continuous in $f \in \mathbb{F}$.
Now, let $\tilde f \in \mathbb{F}$ be fixed and, using Assumption 3.2.1(iii), select an integer $n(x, \tilde f)$ such that $P_x^{\tilde f}[T_z > n_0(x, \tilde f)] < 1$, so that, using the above continuity property, there exists an open set $V(x, \tilde f) \subset \mathbb{F}$ such that

$$\tilde f \in V(x, \tilde f) \quad \text{and} \quad P_x^f[T_z > n_0(x, \tilde f)] < 1 \quad \text{for all } f \in V(x, \tilde f).$$

The family $\{V(x, \tilde f)\}_{\tilde f \in \mathbb{F}}$ is an open covering of the compact set $\mathbb{F}$, and then there exist a finite set of policies $\tilde f_1, \tilde f_2, \ldots, \tilde f_r$ such that $\mathbb{F} = \bigcup_{i=1}^r V(x, \tilde f_i)$; setting

$$n_0(x) = \max\{n(x, \tilde f_i),\, i = 1, 2, \ldots, r\},$$

it follows that $P_x^f[T_z > n_0(x)] < 1$ for every $f \in \mathbb{F}$, and then the above continuity property (a) yields that

$$\sup_{f\in\mathbb{F}} P_x^f[T_z > n_0(x)] =: \rho(x) < 1,$$

so that $\sup_{x\in S,\, f\in\mathbb{F}} P_x[T_z > n_0] = \max_{x\in S} \rho(x) < 1$, where $n_0 = \max_{x\in S} n_0(x)$. □

Proof of Theorem 3.3.1. Given a nonnegative integer n, $x \in S$ and $\pi \in \mathcal{P}$, an application of the Markov property yields that

$$\begin{aligned}
P_x^\pi[T_z > n+1|A_0, X_1] &= P_x^\pi[X_1 \neq z, X_2 \neq z, \ldots X_n \neq z, X_{n+1} \neq z|A_0, X_1] \\
&= I[X_1 \neq z]P_x^\pi[X_2 \neq z, \ldots X_n \neq z, X_{n+1} \neq z|A_0, X_1] \\
&= I[X_1 \neq z]P_{X_1}^{\tilde\pi}[X_1 \neq z, X_2 \neq z, \ldots X_n \neq z] \\
&\leq I[X_1 \neq z]M_n(X_1),
\end{aligned}$$

where the inequality is due to the specification of $M_n(\cdot)$ in Definition 3.3.1, and the 'shifted' policy $\tilde\pi$ is obtained from π by prefixing a history with (x, A_0), that is,

$$\tilde\pi_t(\cdot|x_0, a_0, \ldots a_{r-1}, x_{r-1}, x_r) = \pi_{t+1}(\cdot|x, A_0, x_0, a_0, \ldots a_{r-1}, x_{r-1}, x_r).$$

It follows that

$$\begin{aligned}
P_x^\pi[T_z > n+1] &= E_x^\pi[P_x^\pi[T_z > n+1|A_0, X_1]] \\
&\leq E_x^\pi[I[X_1 \neq z]M_n(X_1)] \\
&= \int_{A(x)} \left[\sum_{y\neq z} p_{x\,y}(a)M_n(y)\right] \pi_0(da|x) \\
&\leq \max_{a\in A(x)} \left[\sum_{y\neq z} p_{x\,y}(a)M_n(y)\right],
\end{aligned}$$

and then, since π is an arbitrary policy,

$$M_{n+1}(x) \leq \max_{a\in A(x)} \left[\sum_{y\neq z} p_{x\,y}(a) M_n(y)\right], \quad x \in S. \tag{3.3.5}$$

On the other hand, from Definition 3.3.1 it follows that, for each $x \in S$, $M_n(x)$ is a decreasing function of n, so that

$$\lim_{n\to\infty} M_n(x) =: M(x) \in [0, 1]. \tag{3.3.6}$$

exists for every x; taking the limit as n goes to ∞ in both sides of (3.3.5), and recalling that S is finite, it follows that

$$M(x) \leq \max_{a\in A(x)} \left[\sum_{y\neq z} p_{x\,y}(a) M(y)\right], \quad x \in S.$$

For each $x \in S$, the term within brackets in this last display is a continuous function of $a \in A(x)$; since the set $A(x)$ is compact, the maximum is attained at an action $f^*(x) \in A(x)$, so that

$$M(x) \leq \sum_{y\neq z} p_{x\,y}(f^*(x)) M(y);$$

this relation implies that

$$M(x) \leq E_x^{f^*}[I[X_1 \neq z]M(X_1)] = E_x^{f^*}[I[T_z > 1]M(X_1)] \tag{3.3.7}$$

as well as $M(X_n) \leq E_{X_n}^{f^*}[I[X_{n+1} \neq z]M(X_{n+1})|H_n]]$, by the Markov property, and then

$$\begin{aligned} M(X_n)I[T_z > n] &\leq I[T_z > n]E_{X_n}^{f^*}[I[X_{n+1} \neq z]M(X_{n+1})|H_n] \\ &= E_{X_n}^{f^*}[I[T_z > n]I[X_{n+1} \neq z]M(X_{n+1})|H_n] \\ &= E_{X_n}^{f^*}[I[T_z > n+1]M(X_{n+1})|H_n] \end{aligned}$$

so that $E_x^{f^*}[M(X_n)I[T_z > n]] \leq E_x^{f^*}[I[T_z > n+1]M(X_{n+1})]$. This last inequality and (3.3.7) together imply that for every $x \in S$ and every nonnegative integer n,

$$M(x) \leq E_x^{f^*}[M(X_n)I[T_z > n]] \leq \max_{y\in S} M(y) P_x^{f^*}[T_z > n],$$

and then, since $x \in S$ is arbitrary,

$$\max_{x\in S} M(x) \leq \max_{y\in S} M(y) \max_{x\in S} P_x^{f^*}[T_z > n].$$

Now, select the positive integer n_0 as in Lemma 3.3.1 to conclude that

$$\max_{x\in S} M(x) \leq \rho_0 \max_{y\in S} M(y),$$

an inequality that, since $\rho_0 < 1$, leads to $\max_{y\in S} M(y) = 0$. Thus, $M(\cdot) = 0$, and then $\lim_{n\to\infty} M_n(x) = 0$, by (3.3.6) .

(ii) Using part (i), select an integer $n_1 > 0$ such that

$$P_x^\pi[X_1 \neq z, X_2 \neq z, \ldots, X_{n_1} \neq z] = P_x^\pi[T_z > n_1] \leq \frac{1}{2}, \quad x \in S, \quad \pi \in \mathcal{P}; \tag{3.3.8}$$

via the Markov property, it follows that, for every state $x \in S$ and $\pi \in \mathcal{P}$

$$P_x^\pi[X_{t+1} \neq z, X_{t+2} \neq z, \ldots, X_{t+n_1} \neq z | H_t] \leq \frac{1}{2},$$

and then

$$\begin{aligned} P_x^\pi[T_z > n_1 + t | H_t] &= P_x^\pi[T_z > t, X_{t+1} \neq z, X_{t+2} \neq z, \ldots, X_{t+n_1} \neq z | H_t] \\ &= I[T_z > t] P_x^\pi[X_{t+1} \neq z, X_{t+2} \neq z, \ldots, X_{t+n_1} \neq z | H_t] \\ &\leq \frac{1}{2}[T_z > t] \end{aligned}$$

so that the inequality

$$P_x^\pi[T_z > n_1 + t] \leq \frac{1}{2} P_x^\pi[T_z > t]$$

is always valid. From this relation and (3.3.8) an induction argument yields that

$$P_x^\pi[T_z > qn_1] \leq \left(\frac{1}{2}\right)^q, \quad x \in S, \quad \pi \in \mathcal{P}, \quad q = 1, 2, 3, \ldots$$

To conclude, given a nonnegative integer n, write $n = qn_1 + r$, where $0 \leq r < n_1$, and notice that, for every $x \in S$ and $\pi \in \mathcal{P}$,

$$P_x^\pi[T_z > n] \leq P_x^\pi[T_z > qn_1] \leq \left(\frac{1}{2}\right)^q = \left(\frac{1}{2^{1/n_1}}\right)^{n_1 q} = 2^{r/n_1} \left(\frac{1}{2^{1/n_1}}\right)^{n_1 q + r},$$

and then

$$P_x^\pi[T_z > n] \leq 2^{1-1/n_1} \left(\frac{1}{2^{1/n_1}}\right)^n,$$

so that the desired conclusion is satisfied with $B = 2^{1-1/n_1}$ and $\rho = 1/2^{1/n_1}$. □

3.4. Total Relative Cost

In this section an auxiliary total cost problem is introduced and some basic continuity properties of the corresponding optimal value function are established. To begin with, notice that the conclusion of Theorem 3.2.1 is clear when the cost function C is null, so that, from this point onwards, it will be supposed that $\|C\| > 0$. With this in mind, define the number $\beta > 0$ by

$$\beta = \frac{-\log(\rho)}{2\|C\|}, \tag{3.4.1}$$

where $\rho \in (0, 1)$ is as in Theorem 3.3.1(ii), and suppose throughout the remainder that the risk-sensitivity coefficient λ satisfies that

$$\lambda \in (0, \beta). \tag{3.4.2}$$

In this case $2\lambda\|C\| < -\log(\rho)$, and there exists δ such that

$$\delta > 0 \quad \text{and} \quad 2\lambda[\|C\| + \delta] < -\log(\rho) \tag{3.4.3}$$

Notice that Theorem 3.3.1(ii) implies that, for every $x \in S$ and $\pi \in \mathcal{P}$,

$$\begin{aligned} E_x^\pi\left[e^{2\lambda[\|C\|+\delta]T_z}\right] &= \sum_{n=1}^{\infty} e^{2\lambda[\|C\|+\delta]n} P_x^\pi[T_z = n] \\ &\leq \frac{B}{\rho} \sum_{n=1}^{\infty} e^{2\lambda[\|C\|+\delta]n} \rho^n \end{aligned}$$

and then

$$E_x^\pi\left[e^{2\lambda[\|C\|+\delta]T_z}\right] = \frac{B}{\rho}\sum_{n=1}^\infty\left(e^{2\lambda[\|C\|+\delta]}\rho\right)^n =: \hat{B} < \infty, \quad x \in S, \quad \pi \in \mathcal{P}. \tag{3.4.4}$$

where the inequality is due the relation $e^{2\lambda[\|C\|+\delta]}\rho < 1$, by (3.4.3). Consider now the total relative cost with respect to $g \in \mathbb{R}$ incurred before the first return time to state z, which is given by $\sum_{t=0}^{T_z-1}[C(X_t, A_t - g]$; in the sequel, the expected utility of this quantity will play an important role, and a special notation is now introduced.

Definition 3.4.1. For each $g \in (-\|C\| - \delta,\, \|C\| + \delta)$, $x \in S$ and $\pi \in \mathcal{P}$, the expected utility corresponding to $\sum_{t=0}^{T_z-1}[C(X_t, A_t - g]$ when x is the initial state and π is the policy employed is denoted by

$$u(x, \pi; g) := E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}\right],$$

whereas

$$u^*(x; g) := \inf_{\pi\in\mathcal{P}} u(x, \pi; g)$$

stands for the optimal utility at x.

Lemma 3.4.1. With the notation in (3.4.1)–(3.4.4), the following assertions (i)—(iii) hold for every $x \in S$ and $\pi \in \mathcal{P}$:

(i) For each $g \in (-\|C\| - \delta,\, \|C\| + \delta)$

$$u(x, \pi; g) < \hat{B}.$$

Moreover,

(ii) If $|g| \le \|C\|$ and $|h| < \delta$, then

$$|u(x, \pi; g + h) - u(x, \pi; g)| \le |h|\frac{\hat{B}}{\lambda\delta}.$$

Consequently,

(iii) The function $u^*(x; \cdot)$ is continuous in $[-\|C\|, \|C\|]$ and satisfies that $u^*(x; \|C\|) \le 1$ and $u^*(x; -\|C\|) \ge 1$.

Proof. (i) Suppose that $|g| < \|C\| + \delta$ and notice that

$$\begin{aligned}\left|\lambda\sum_{t=0}^{T_z-1}[C(X_t, A_t) - g]\right| &\le \lambda\sum_{t=0}^{T_z-1}[\|C\| + |g|]\\ &= \lambda T_z[\|C\| + |g|]\\ &\le \lambda T_z[2\|C\| + |\delta]\end{aligned}$$

so that (3.4.4) and Definition 3.4.1 together yield that $u(x, \pi; g) \le \hat{B}$.

(ii) let $h \in (-\delta, \delta)$ and $g \in [-\|C\|, \|C\|]$ by arbitrary. Using the inequalities $|e^x - 1| \leq |x|e^{|x|}$ and $|x| \leq e^{a|x|}/a$, which are valid for every $x \in \mathbb{R}$ and $a > 0$, it follows that

$$\begin{aligned}
|u(x,\pi;g+h) - u(x,\pi;g)| &= \left| E_x^\pi \left[e^{\lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g-h]} \right] - E_x^\pi \left[e^{\lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]} \right] \right| \\
&= \left| E_x^\pi \left[(e^{-\lambda h T_z} - 1) e^{\lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]} \right] \right| \\
&\leq E_x^\pi \left[|e^{-\lambda h T_z} - 1| e^{\lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]} \right] \\
&\leq E_x^\pi \left[|hT_z||e^{\lambda|h|T_z} e^{\lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]} \right] \\
&= |h| E_x^\pi \left[T_z e^{\lambda|h|T_z + \lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]} \right] \\
&\leq \frac{|h|}{\lambda\delta} E_x^\pi \left[e^{\lambda\delta T_z} e^{\lambda|h|T_z + \lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]} \right] \\
&= \frac{|h|}{\lambda\delta} E_x^\pi \left[e^{\lambda\delta T_z + \lambda|h|T_z + \lambda \sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]} \right].
\end{aligned}$$

Since $|h| < \delta$ and $|g| \leq \|C\|$, it follows that

$$\left| \delta T_z + \lambda|h|T_z + \lambda \sum_{t=0}^{T_z-1} [C(X_t, A_t) - g] \right| \leq 2[\|C\| + \delta]T_z$$

and combining these two last displays with (3.4.4) it follows that

$$|u(x,\pi;g+h) - u(x,\pi;g)| \leq |h| \frac{\hat{B}}{\lambda\delta}.$$

(iii) Let $g, g_1 \in [-\|C\|, \|C\|]$ be arbitrary numbers satisfying $|g_1 - g| < \delta$, and notice that part (ii) yields that, for every $x \in S$ and $\pi \in \mathcal{P}$, $|u(x,\pi;g_1) - u(x,\pi;g)| \leq |g_1 - g|\hat{B}/[\lambda\delta]$, a relation that leads to $u(x,\pi;g_1) \leq u(x,\pi;g) + |g_1 - g|\hat{B}/[\lambda\delta]$, and then, taking the infimum with respect to $\pi \in \mathcal{P}$, Definition 3.4.1 yields that

$$u^*(x;g_1) \leq u^*(x;g) + |g_1 - g|\hat{B}/[\lambda\delta];$$

similarly, interchanging the roles of g and g_1,

$$u^*(x;g) \leq u^*(x;g_1) + |g_1 - g|\hat{B}/[\lambda\delta].$$

and these two last dsiplays together yield that

$$|u^*(x;g_1) - u^*(x;g)| \leq |g_1 - g| \frac{\hat{B}}{\lambda\delta}, \quad g, g_1 \in [-\|C\|, \|C\|\,],$$

so that $u^*(x;\cdot)$ is a (Lipchitz-)continuous function on the interval $[-\|C\|, \|C\|\,]$. To conclude, notice that $C(X_t, A_t) - g \geq 0$ when $g = -\|C\|$, and $C(X_t, A_t) - g \leq 0$ if $g = \|C\|$, so that $e^{\lambda \sum_{t=0}^{T_z-1}[C(x_t,A_t)-g]} \geq 1$ if $g = -\|C\|$ and $e^{\lambda \sum_{t=0}^{T_z-1}[C(x_t,A_t)-g]} \leq 1$ when $g = \|C\|$; hence, the inequalities

$$u(x,\pi;-\|C\|) \geq 1 \quad \text{and} \quad u(x,\pi;\|C\|) \leq 1$$

are always valid, by Definition 3.4.1, and taking the infimum with respect to π, these relations yield that $u^*(x;\|C\|) \leq 1$ and $u^*(x;-\|C\|) \geq 1$. □

3.5. Proof of Theorem 3.2.1

The conclusion in Theorem 3.2.1 will be verified by combining the continuity properties of the optimal utility u^* in Lemma 3.4.1(iii), with the dynamic programming equation stated below.

Lemma 3.5.1. For each $g \in [-\|C\|, \|C\|\,]$ the optimal utility function $u^*(\cdot; g)$ satisfies the following dynamic programming equation:

$$u^*(x; g) = \inf_{a \in A(x)} \left[e^{\lambda[C(x,a)-g]} \left(p_{x\,z}(a) + \sum_{y \in S \setminus \{z\}} p_{x\,y}(a) u^*(y; g) \right) \right]. \tag{3.5.1}$$

Proof. Let $g \in [-\|C\|, \|C\|\,]$ and $\varepsilon > 0$ be arbitrary but fixed, and for each $y \in S$, select a policy $\pi^y = \{\pi_t^y\} \in \mathcal{P}$ such that

$$u^*(y; g) \leq u(y, \pi^y; g) + \varepsilon. \tag{3.5.2}$$

Now, take an arbitrary policy $f \in \mathbb{F}$ and define the new policy $\pi = \{\pi_t\}$ as follows:

- $\pi_0(\{f(x)\}|x) = 1$ for every $x \in S$ and,
- For $t \geq 1$, $\pi_t(\cdot|x_0, a_0, x_1, a_1, \ldots, x_t) = \pi_{t-1}^{x_1}(\cdot|x_1, a_1, \ldots, x_t)$.

The behavior of a controller using this policy π can be described as follows: At time $t = 0$ actions are chosen according to f whereas, from time 1 onwards, if $x_1 = y$ is observed, then the decisions are taken using π^y as if the process had started again at time 1. With this specification, an application of the Markov property yields that, for each $y \in S$ and regardless of the initial state $x \in S$,

$$\begin{aligned}
&E_x^\pi \left[e^{\lambda \sum_{t=0}^{T_z - 1} [C(X_t, A_t) - g]} I[T_z > 1] \middle| X_1 \right] \\
&\quad = E_x^\pi \left[e^{\lambda[C(x,f(x))-g]} e^{\lambda \sum_{t=1}^{T_z - 1} [C(X_t, A_t) - g]} I[X_1 \neq z] \middle| X_1 \right] \\
&\quad = e^{\lambda[C(x,f(x))-g]} I[X_1 \neq z] E_{X_1}^{\pi^{X_1}} \left[e^{\lambda \sum_{t=0}^{T_z - 1} [C(X_t, A_t) - g]} \right] \\
&\quad = e^{\lambda[C(x,f(x))-g]} I[X_1 \neq z] u(X_1, \pi^{X_1}; g) \\
&\quad \leq e^{\lambda[C(x,f(x))-g]} I[X_1 \neq z] [u^*(X_1) + \varepsilon]
\end{aligned}$$

and, taking the expectation with respect to P_x^π, this yields that

$$\begin{aligned}
E_x^\pi \left[e^{\lambda \sum_{t=0}^{T_z - 1} [C(X_t, A_t) - g]} I[T_z > 1] \right] &\leq e^{\lambda[C(x,f(x))-g]} \sum_{y \in S \setminus \{z\}} p_{x\,y}(f(x)[u^*(y) + \varepsilon] \\
&\leq e^{\lambda[C(x,f(x))-g]} \sum_{y \in S \setminus \{z\}} p_{x\,y}(f(x)u^*(y) + e^{2\lambda\|C\|}\varepsilon.
\end{aligned}$$

Combining this relation with the equality

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{T_z - 1} [C(X_t, A_t) - g]} I[T_z = 1] \right] = e^{\lambda[C(x,f(x))-g]} p_{x\,z}(f(x)),$$

it follows that

$$\begin{aligned} u^*(x;g) &\leq u(x,\pi;g) \\ &= E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}\right] \\ &= E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z=1]\right] + E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z>1]\right] \\ &\leq e^{\lambda[C(x,f(x))-g]}p_{x\,z}(f(x)) + e^{\lambda[C(x,f(x))-g]}\sum_{y\in S\setminus\{z\}} p_{x\,y}(f(x))u^*(y) + e^{\lambda\|C\|}\varepsilon, \end{aligned}$$

where the first inequality is due to the specification of the optimal utility u^* in Definition 3.4.1; since $\varepsilon > 0$ is arbitrary, it follows that

$$u^*(x;g) \leq e^{\lambda[C(x,f(x))-g]}\left(p_{x\,z}(f(x)) + \sum_{y\in S\setminus\{z\}} p_{x\,y}(f(x))u^*(y;g)\right).$$

This inequality holds for every $f \in \mathbb{F}$, so that the action $f(x)$ in the above expression is an arbitrary element of the action set $A(x)$, and then

$$u^*(x;g) \leq \inf_{a\in A(x)}\left[e^{\lambda[C(x,a))-g]}\left(p_{x\,z}(a) + \sum_{y\in S\setminus\{z\}} p_{x\,y}(a))u^*(y;g)\right)\right]. \tag{3.5.3}$$

To establish the reverse inequality, let $\pi \in \mathcal{P}$ be an arbitrary policy, and notice that, by the Markov property, for every initial state $x \in S$ the following relations hold:

$$\begin{aligned} &E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z>1]\middle| A_0, X_1\right] \\ &\qquad = E_x^\pi\left[e^{\lambda[C(x,A_0))-g]}e^{\lambda\sum_{t=1}^{T_z-1}[C(X_t,A_t)-g]}I[X_1\neq z]\middle| A_0, X_1\right] \\ &\qquad = e^{\lambda[C(x,A_0)-g]}I[X_1\neq z]E_{X_1}^{\pi^{(x,A_0)}}\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}\right] \end{aligned}$$

where the shifted policy $\pi^{(x,A_0)}$ is defined as follows: For each $t = 0, 1, 2, \ldots$,

$$\pi_t^{(x,A_0)}(\cdot|x_0,a_0,x_1,a_1,\ldots,x_{t-1},a_{t-1},x_t) = \pi_{t+1}(\cdot|x,A_0,x_0,a_0,x_1,a_1,\ldots,x_{t-1},a_{t-1},x_t).$$

Since $E_{X_1}^{\pi^{(x,A_0)}}\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}\right] \geq u^*(X_1;g)$, it follows that

$$E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z>1]\middle| A_0, X_1\right] \geq e^{\lambda[C(x,A_0)-g]}I[X_1\neq z]u^*(X_1;g),$$

and taking the expectation with respect to P_x^π, this relation leads to

$$\begin{aligned} E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z>1]\right] &\geq E_x^\pi\left[e^{\lambda[C(x,A_0)-g]}I[X_1\neq z]u^*(X_1;g)\right] \\ &= \int_{a\in A(x)}\left(e^{\lambda[C(x,a)-g]}\sum_{y\in S\setminus\{z\}} u^*(y;g)\right)\pi_0(da|x). \end{aligned}$$

On the other hand,

$$E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z=1]\right]=\int_{a\in A(x)}e^{\lambda[C(x,a)-g]}p_{x\,z}(a)\pi_0(da|x),$$

a relation that together with the previous display leads to

$$\begin{aligned}
&E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}\right]\\
&\quad=E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z=1]\right]+E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}I[T_z>1]\right]\\
&\quad\geq\int_{a\in A(x)}\left(e^{\lambda[C(x,a)-g]}\sum_{y\in S\setminus\{z\}}p_{x\,y}(a)u^*(y;g)\right)\pi_0(da|x)\\
&\qquad\qquad\qquad+\int_{a\in A(x)}e^{\lambda[C(x,a)-g]}p_{x\,z}(a)\pi_0(da|x)\\
&\quad=\int_{a\in A(x)}e^{\lambda[C(x,a)-g]}\left(p_{x\,z}(a)+\sum_{y\in S\setminus\{z\}})p_{x\,y}(a)u^*(y;g)\right)\pi_0(da|x),
\end{aligned}$$

and then

$$E_x^\pi\left[e^{\lambda\sum_{t=0}^{T_z-1}[C(X_t,A_t)-g]}\right]\geq\inf_{a\in A(x)}e^{\lambda[C(x,a)-g]}\left(p_{x\,z}(a)+\sum_{y\in S\setminus\{z\}}p_{x\,y}(au^*(y;g)\right),$$

After taking the infimum with respect to $\pi\in\mathcal{P}$, it follows that

$$u^*(x;g)\geq\inf_{a\in A(x)}\left[e^{\lambda[C(x,a)-g]}\left(p_{x\,z}(a)+\sum_{y\in S\setminus\{z\}}p_{x\,y}(a)u^*(y;g)\right)\right]$$

and then, recalling that $x\in S$ is arbitrary, the conclusion follows combining this last relation with (3.5.3). □

The existence of solutions to the optimality equation under Assumption 3.2.1 can be now established as follows:

Proof of Theorem 3.2.1. By Lemma 3.4.1(ii), the function $u^*(z;\cdot)$ is continuous in the interval $[-\|C\|,\,\|C\|\,]$ and satisfies the relations

$$u^*(z;-\|C\|)\geq 1\quad\text{and}\quad u^*(z;\|C\|)\leq 1.$$

Thus, by the intermediate value property, there exists $g^*\in[-\|C\|,\,\|C\|\,]$ such that

$$u^*(z,g^*)=1,$$

and then Lemma 3.5.1 yields that

$$u^*(x;g^*)=\inf_{a\in A(x)}\left[e^{\lambda[C(x,a)-g^*]}\sum_{y\in S}p_{x\,y}(a)u^*(y;g)\right],\quad x\in S.$$

To conclude, define $h^*: S \to \mathrm{I\!R}$ by

$$h^*(x) := \frac{1}{\lambda}\log(u^*(x; g^*)), \quad x \in S,$$

and notice that the above dynamic programming equation is equivalent to

$$e^{\lambda h^*(x)} = \inf_{a\in A(x)} \left[e^{\lambda[C(x,a)-g^*]} \sum_{y\in S} p_{x\,y}(a) e^{\lambda h^*(y)} \right], \quad x \in S,$$

showing that the optimality equation is satisfied by the pair $(g^*, h^*(\cdot)) \in \mathrm{I\!R} \times \mathcal{B}(S)$. □

3.6. Conclusion

In this chapter the existence of solutions to the (λ-sensitive) optimality equation was studied within the framework determined by Assumption 3.2.1. Besides standard continuity-compactness properties, such an assumption postulates the accessibility condition (3.2.1), which ensures that, under the action of any stationary policy, some state z can be reached with positive probability regardless of the initial state, but does not guarantee that, starting from z, any other state can be eventually visited. Thus, Assumption 3.2.1 is weaker that Assumption 2.5.1 and, accordingly, the main result of this chapter, stated as Theorem 3.2.1, only ensures that the optimality equation has a solution whenever the risk-aversion coefficient is small enough; moreover, it was shown in Example 3.2.1 that such a conclusion can not be extended to include every positive risk-sensitivity coefficient. The proof of Theorem 3.2.1 was approached using a total utility cost criterion, providing an alternative method to the one used in Cavazos-Cadena (2003), where Theorem 3.2.1 was obtained *via* the discounted technique, a method that will be used in the following chapter.

On the other hand, the analysis of Example 3.2.1 presented in Propositions 3.2.1 and 3.2.2 pointed out a remarkable fact:

> Even if (i) all the stationary policies have a single recurrent class and share a recurrent state, *and* (ii) the (optimal) average cost function $J^*(\cdot)$ is constant, in general $J^*(\cdot)$ is not characterized by a single equation

This fact shows that the main goal of this thesis, namely, to provide a general characterization of the optimal risk-sensitive average index in Markov decision chains over a finite state space, is a very interesting problem; the proposed solution will be presented in the following chapter.

Chapter 4

Optimality Systems for Average Markov Decision Chains Under Risk-Aversion

In this chapter the optimal risk-sensitive average cost function is characterized for general controlled Markov chains with finite state space and compact action sets, a result that is the *main contribution* of this thesis. It is supposed that the decision maker is risk-averse with constant risk-sensitivity coefficient and, under standard continuity–compactness conditions, it is proved that the (possibly non-constant) optimal value function is characterized by *a nested system of equations*, generalizing the characterizations presented in the previous chapters which require communication conditions on the transition law; moreover, it is shown that an optimal stationary policy can be derived form a solution of that system, and that the optimal superior and inferior limit average cost functions coincide. The approach in this chapter relies on the *discounted method* which, roughly, consists in using a family of contractive operators whose fixed points are used to approximate the optimal average index. The presentation of the subsequent material is self-contained and is based on Alanís-Durán and Cavazos-Cadena (2012).

4.1. Introduction

This chapter is concerned with discrete-time Markov decision processes (MDPs) evolving on a finite state space. The system is driven by a risk-averse decision maker with constant risk sensitivity coefficient $\lambda > 0$, and the performance of a control policy is measured by the (superior limit) risk-sensitive average cost criterion. It is supposed that the action set is a compact metric space, and that the cost function and the transition law depend continuously on the action applied, but otherwise they are arbitrary; in particular, no communication conditions are imposed on the transition law, so that the optimal value function may not be constant. Within that framework, the following problem is addressed:

> *To characterize the optimal value function using a system of equations form which an optimal stationary policy can be determined.*

The study of stochastic systems endowed with the risk-sensitive average criterion can be traced back, at least, to the seminal papers by Howard and Matheson (1972), Jacobson (1973) and Jaquette(1973; 1976). Recently, there has been an intensive work on (controlled) stochastic system endowed with the risk-sensitive average criterion; see, for instance, Flemming and McEneany (1995), Di Masi and Stettner (1999, 2000, 2007), Jaśkiewicz (2007), Sladký and Montes-de-Oca (2008), Sladký (2008) and the references there in. A fundamental result on the existence of solutions of the risk-sensitive optimality equation was obtained by Howard and Matheson (1972), where controlled Markov chains with finite state and action spaces were studied, and it was shown that the optimal average cost is determined by a *single* equation whenever each stationary policy determines a communicating Markov chain. In such a case, the optimal average cost function is constant, say g, and the existence of a solution to the optimality equation was established using the Perron-Frobenius theory of nonnegative matrices (Gantmakher, 1959). Other approaches have been used to obtain a solution to the optimality equation: the main result in Hernández-Hernández and Marcus (1996) is based on game theoretical ideas, the approach in Cavazos-Cadena and Fernández-Gaucherand (2002) relies on the risk-sensitive total cost criterion, and the discounted technique—involving contractive mappings—was employed in Di Masi and Stettner (1999) and Cavazos-Cadena (2003). On the other hand, there is an interesting contrast between the risk-neutral and the risk-sensitive average cost criteria: Under strong recurrence conditions, like the simultaneous Doeblin condition—under which the Markov chain determined by each stationary policy has a single recurrent class—the risk-neutral optimality equation has a solution, but a similar conclusion is not valid in the risk-sensitive context, even if the optimal average cost is constant (Cavazos-Cadena and Fernández-Gaucherand, 1999, Cavazos-Cadena and Hernández-Hernandez, 2005). Thus, the characterization of the optimal risk-sensitive average cost can not be based, in general, on a single equation, and the problem posed above is an interesting and natural one.

The characterization of a general (risk-sensitive) optimal average cost function was recently studied in Sladký (2008) for models with finite state *and* action sets; in that paper, the analysis is based on Perron-Frobenius decompositions of a family of nonnegative matrices (Rothblum and Whittle, 1982, Sladký, 1979, 1980, Whittle, 1953, Zijm, 1983). On the other hand, the discounted approach has also been employed to study the case of a non necessarily constant optimal average index; see, for instance, Hernández-Hernández and Marcus (1999) for models with denumerable state space, and Jaśkiewicz (2007) and Cavazos-Cadena and Salem-Silva (2009), which concern MDPs with Borel state space. Roughly, in those papers a characterization of the optimal average cost is obtained at *some* states where the optimal performance index attains its minimum. In the context of this work, the discounted technique will play a central role to obtain a *complete* characterization of the optimal average cost function.

The main results of this work involve the idea of *optimality system* introduced in Section 3, and can be briefly described as follows: An optimality system is determined by

(i) a partition $S_1, \ldots, S_k$ of a state space,

(ii) a sequence of pairs $\{(g_i, h_i(\cdot))\}_{i=1,\ldots,k}$, where g_i is a real number and h_i is a function defined on S_i,

(iii) the specification of a (generally proper) subset of $B(x)$ of the original set admissible actions $A(x)$ at each state x.

In terms of these objects, an equation—similar to the usual optimality equation—is stipulated for every $i = 1, 2, \ldots, k$, and the following conclusions, extending those in Cavazos-Cadena and Hernández-Hernández (2006) for *uncontrolled* models, are obtained:

(1) An optimality system characterizes the optimal average cost function and renders an optimal stationary policy (the verification theorem);

(2) There exists and optimality system (the existence theorem).

The approach used below to establish these conclusions relies on basic probabilistic and dynamic programming ideas, which are used to establish the verification theorem, whereas the discounted method is employed to derive the existence result.

The organization of the exposition is as follows: In Section 2 a brief description of the decision model is presented and, after introducing the notion of optimality system in Section

3, the verification and existence results are stated as Theorems 3.1 and 3.2, respectively. Then, in Section 4 a technical result on the inferior limit average criterion is presented, and it is used to establish the verification theorem in Section 5. Next in Section 6 the discounted approach is used to specify the components of an optimality system, and the presentation concludes in Section 7 with a proof of the existence theorem.

Notation. The set of all nonnegative integers is denoted by $\mathbb{N}$ and, for a given topological space $\mathbb{K}$, $\mathcal{B}(\mathbb{K})$ stands for the Banach space of all bounded functions $C\colon \mathbb{K} \to \mathbb{R}$ equipped with the supremum norm:

$$\|C\| := \sup_{x\in\mathbb{K}} |C(x)|.$$

On the other hand, for $x \in \mathbb{K}$, $\delta_x(\cdot)$ is the Dirac's measure concentrated at x, that is, for every Borel subset $D \subset \mathbb{K}$, $\delta_x(D) = 1$ if $x \in D$, and $\delta_x(D) = 0$ when $x \notin D$. If A is an event, the corresponding indicator function is denoted by $I[A]$ and, as usual, all relations involving conditional expectations are supposed to hold almost surely with respect to the underlying probability measure.

4.2. Decision Model

Throughout the remainder $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, C, P)$ is an MDP, where the state space S is a finite set endowed with the discrete topology, and the action set A is a metric space. For each $x \in S$, $A(x) \subset A$ is the nonempty set of admissible actions at x, whereas $\mathbb{K} := \{(x, a) \,|\, a \in A(x), x \in S\}$ is the class of admissible pairs. On the other hand, $C \in B(\mathbb{K})$ is the cost function and $P = [p_{x\,y}(\cdot)]$ is the controlled transition law on S given $\mathbb{K}$, that is, for each $(x, a) \in \mathbb{K}$ and $z \in S$, $p_{x\,z}(a) \geq 0$ and $\sum_{y\in S} p_{x\,y}(a) = 1$. This model $\mathcal{M}$ is interpreted as follows: At each time $t \in \mathbb{N}$ the decision maker observes the state of a dynamical system, say $X_t = x \in S$, and selects the action (control) $A_t = a \in A(x)$. Then, a cost $C(x, a)$ is incurred and, regardless of the previous states and actions, the state of the system at time $t + 1$ will be $X_{t+1} = y \in S$ with probability $p_{x\,y}(a)$; this is the Markov property of the decision process.

Assumption 4.2.1. *(i) For each $x \in S$, $A(x)$ is a compact subset of A.*

(ii) For every $x, y \in S$, the mappings $a \mapsto C(x, a)$ and $a \mapsto p_{x\,y}(a)$ are continuous in $a \in A(x)$.

Policies. The space $\mathbb{H}_t$ of possible histories up to time $t \in \mathbb{N}$ is defined by $\mathbb{H}_0 := S$ and $\mathbb{H}_t := \mathbb{K}^{t-1} \times S$, $t \geq 1$. A generic element of $\mathbb{H}_t$ is denoted by $\mathbf{h}_t = (x_0, a_0, \ldots, x_i, a_i, \ldots, x_t)$, where $a_i \in A(x_i)$. A policy $\pi = \{\pi_t\}$ is a special sequence of stochastic kernels: For each $t \in \mathbb{N}$ and $\mathbf{h}_t \in \mathbb{H}_t$, $\pi_t(\cdot|\mathbf{h}_t)$ is a probability measure on A concentrated on $A(x_t)$, and for each Borel subset $B \subset A$, the mapping $\mathbf{h}_t \mapsto \pi_t(B|\mathbf{h}_t)$, $\mathbf{h}_t \in \mathbb{H}_t$, is Borel measurable; when the controller chooses actions according to π the control A_t applied at time t belongs to $B \subset A$ with probability $\pi_t(B|\mathbf{h}_t)$, where $\mathbf{h}_t$ is the observed history of the process up to time t. The class of all policies is denoted by $\mathcal{P}$. Given the policy π being used for choosing actions and the initial state $X_0 = x$, the distribution of the state-action process $\{(X_t, A_t)\}$ is uniquely determined (Araposthatis *et al.* 1993, Puterman 2005), and such a distribution and the corresponding expectation operator are denoted by P_x^π and E_x^π, respectively. Next, define $\mathbb{F} := \prod_{x\in S} A(x)$ and notice that $\mathbb{F}$ is a compact metric space, which consists of all functions $f\colon S \to A$ such that $f(x) \in A(x)$ for each $x \in S$. A policy π is *stationary* if there exists a sequence $f \in \mathbb{F}$ such that the probability measure $\pi_t(\cdot|\mathbf{h}_t)$ is always concentrated at $f(x_t)$, and in this case π and f are naturally identified; with this convention, $\mathbb{F} \subset \mathcal{P}$.

Performance Criterion. As already mentioned, the decision maker is supposed to be *risk-averse* with constant risk-sensitivity coefficient $\lambda > 0$, that is, the controller assesses a random cost Y using the expectation of $e^{\lambda Y}$; the certain equivalent of Y is the real number

$\mathcal{E}[Y]$ determined by $e^{\lambda\mathcal{E}[Y]} = E[e^{\lambda Y}]$, so that the controller is indifferent between paying the certain equivalent $\mathcal{E}[Y]$ for sure, or incurring the random cost Y. It follows that

$$\mathcal{E}[Y] = \frac{1}{\lambda}\log\left(E[e^{\lambda Y}]\right),$$

whereas Jensen's inequality yields that if Y has finite expectation, then $\mathcal{E}[Y] \geq E[Y]$ and the strict inequality holds if Y is non constant. Suppose now that the controller is driving the system using policy $\pi \in \mathcal{P}$ starting at $x \in S$, and let $J_n(\lambda, \pi, x)$ be the certain equivalent of the total cost $\sum_{t=0}^{n-1} C(X_t, A_t)$ incurred before time n, that is,

$$J_n(\lambda, \pi, x) = \frac{1}{\lambda}\log\left(E_x^\pi\left[e^{\lambda\sum_{t=0}^{n-1} C(X_t, A_t)}\right]\right). \tag{4.2.1}$$

With this notation, the (long-run superior limit) λ-sensitive average cost at state x under policy π is given by

$$J(\lambda, \pi, x) := \limsup_{n\to\infty} \frac{1}{n} J_n(\lambda, \pi, x), \tag{4.2.2}$$

and

$$J^*(\lambda, x) := \inf_{\pi\in\mathcal{P}} J(\lambda, \pi, x), \quad x \in S, \tag{4.2.3}$$

is the optimal λ-sensitive average cost function; a policy $\pi^* \in \mathcal{P}$ is λ-optimal if $J(\lambda, \pi^*, x) = J^*(\lambda, x)$ for each $x \in S$.

Remark 4.2.1. When $X_0 = x$, the inferior limit λ-sensitive average criterion $J_-(\lambda, \pi, x)$ corresponding to $\pi \in \mathcal{P}$ is defined by

$$J_-(\lambda, \pi, x) := \liminf_{n\to\infty} \frac{1}{n} J_n(\lambda, \pi, x), \tag{4.2.4}$$

and the corresponding (inferior limit) λ-optimal value function is given by

$$J_*(\lambda, x) := \inf_{\pi\in\mathcal{P}} J_-(\lambda, \pi, x), \quad x \in S \tag{4.2.5}$$

so that $J_*(\lambda, \cdot) \leq J^*(\lambda, \cdot)$; as it will be shown below, under Assumption 4.2.1 the optimal value functions $J_*(\lambda, \cdot)$ and $J^*(\lambda, \cdot)$ coincide. □

The Problem. The optimality equation corresponding to the average criterion in (4.2.2) is given by

$$e^{\lambda(g+h(x))} = \inf_{a\in A(x)}\left[e^{\lambda C(x,a)}\sum_{y\in S} p_{x\,y}(a)e^{\lambda h(y)}\right], \quad x \in S, \tag{4.2.6}$$

where g is a real number and $h: S \to \mathbb{R}$ is a given function. When this equation is satisfied by the pair $(g, h(\cdot)) \in \mathbb{R} \times \mathcal{B}(S)$, the optimal average cost function $J^*(\lambda, \cdot)$ is constant and equal to g; moreover, Assumption 4.2.1 yields that there exists a policy $f^* \in \mathbb{F}$ such that

$$e^{\lambda(g+h(x))} = e^{\lambda C(x,f^*(x))}\sum_{y\in S} p_{x\,y}(f^*(x))e^{\lambda h(y)}, \quad x \in S,$$

and such a policy f^* is λ-optimal. As already noted, a pair $(g, h(\cdot))$ satisfying (4.2.6) exists when the whole state space is a communicating class under the action of each stationary policy; however, it was shown un Cavazos-Cadena and Hernández-Hernández (2006) that, if the Markov chain associated with some $f \in \mathbb{F}$ has two or more recurrent classes, or if the set of transient states is nonempty, then (4.2.6) may not have a solution, even if the optimal average cost function is constant. On the other hand, for *uncontrolled* Markov chains it was recently shown in Cavazos-Cadena and Hernández-Hernández (2007) that, in general, the average cost function is determined by a *system* of local Poisson equations, and *the main problem* considered in this thesis consists in extending such a conclusion to the present context of controlled models. The results in this direction involve the idea of *optimality system*, which is introduced in the following section.

4.3. Optimality Systems and Main Results

In this section the main conclusions of this note are stated as Theorems 3.1 and 3.2 below. These results involve the idea of *optimality system*, which extends the notion of optimality equation and allows to characterize the optimal value function in terms of a system of equations, as well as to obtain a λ-optimal stationary policy.

Definition 4.3.1. Let $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, C, P)$ be the MDP described in Section 4.2. An *optimality system* for $\mathcal{M}$ is a vector of triplets

$$\mathcal{O} = ((S_1, g_1, h_1), (S_2, g_2, h_2), \ldots, (S_k, g_k, h_k)) \tag{4.3.1}$$

satisfying the following conditions:

(i) $S_1, S_2, \ldots, S_k$ is a partition of S.

(ii) For each $i = 1, 2, \ldots, k$, $(g_i, h_i(\cdot)) \in \mathbb{R} \times \mathcal{B}(S_i)$ and

$$g_1 \le g_2 \le \cdots \le g_k. \tag{4.3.2}$$

(iii) For each $i = 1, 2, \ldots, k$,

$$B(x) := \{a \in A(x) \mid \textstyle\sum_{y\in S_1\cup S_2\cup\cdots\cup S_i} p_{x\,y}(a) = 1\}, \quad x \in S_i \quad \text{is nonempty.} \tag{4.3.3}$$

(iv) For each $i = 1, 2, \ldots, k$,

$$e^{\lambda(g_i + h_i(x))} = \inf_{a\in B(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S_i} p_{x\,y}(a) e^{\lambda h_i(y)} \right], \quad x \in S_i. \tag{4.3.4}$$

Remark 4.3.1. Notice that (4.3.4) implies that, for every $x \in S_i$, $\sum_{y\in S_i} p_{x\,y}(a) > 0$ for all $a \in B(x)$, since $e^{\lambda(g_i + h_i(x))} > 0$. □

The number k of triplets in $\mathcal{O}$ will be referred to as *the order* of $\mathcal{O}$. The above idea is an extension of the notion of J-system used in Cavazos-Cadena and Hernández-Hernández (2006) to characterize the average cost function for an uncontrolled Markov chain. In the present controlled context, the following result shows that an optimality system renders (i) the optimal value function, (ii) the equality of the superior and inferior limit optimal value functions, as well as (iii) a λ-optimal stationary policy.

Theorem 4.3.1. [Verification.] Let $\mathcal{M}$ be the model described in Section 2 and suppose that Assumption 4.2.1 holds. If $\mathcal{O} = ((S_1, g_1, h_1), (S_2, g_2, h_2), \ldots, (S_k, g_k, h_k))$ is an optimality system for $\mathcal{M}$, then the following assertions (i)–(iii) hold:

(i) For each $i = 1, 2, \ldots, k$, the optimal average cost at each state $x \in S_i$ is given by g_i:

$$J^*(\lambda, x) = g_i, \quad x \in S_i.$$

Moreover,

(ii) $J_*(\lambda, x) = J^*(\lambda, x)$ for all $x \in S$; see (4.2.3) and (4.2.5).

(iii) Suppose that the stationary policy $f \in \mathbb{F}$ satisfies that

$$f(x) \in B(x), \quad x \in S \tag{4.3.5}$$

and

$$e^{\lambda(g_i+h_i(x))} = \left[e^{\lambda C(x,f(x))} \sum_{y \in S_i} p_{x\,y}(f(x)) e^{\lambda h_i(y)} \right], \quad x \in S_i, \quad i = 1, 2, \ldots, k. \tag{4.3.6}$$

In this case f is λ-optimal and

$$\lim_{n\to\infty} \frac{1}{n} J_n(\lambda, f, x) = J^*(\lambda, x), \quad x \in S.$$

Notice that Assumption 4.2.1 yields that the set $B(x)$ in (4.3.3) is always compact, a fact that using (4.3.4) implies the existence of a stationary policy f satisfying (4.3.5) and (4.3.6). The following result establishes the existence of an optimality system.

Theorem 4.3.2. [Existence.] Under Assumption 4.2.1, there exists an optimality system $\mathcal{O}$ for model $\mathcal{M}$.

The proof of Theorems 4.3.1 and 4.3.2 will be presented in Sections 5 and 7, respectively, after establishing the necessary preliminary results. The argument used to establish the verification result relies on standard probabilistic and dynamic programming arguments, whereas the existence of an optimality system will be obtained *via* the risk-sensitive discounted criterion.

4.4. A Lower Bound for the Inferior Limit Average Criterion

In this section a basic technical tool that will be used to prove Theorem 4.3.1 is established. The main objective is to show that if $\mathcal{O}$ is as optimality system for model $\mathcal{M}$, then a lower bound for the optimal inferior limit average cost function can be obtained, a result that is precisely stated in the following theorem.

Theorem 4.4.1. Let $\mathcal{O}$ in (4.3.1) be an optimality system for model $\mathcal{M}$. In this case, g_i is a lower bound for the inferior limit λ-sensitive average cost criterion at each state $x \in S_i$:

$$J_*(\lambda, x) \geq g_i, \quad x \in S_i, \quad i = 1, 2, \ldots, k; \tag{4.4.1}$$

see Remark 4.2.1.

This result will be proved below by induction. Since the argument is rather technical, to ease the presentation the simple auxiliary facts involved in the argument are established in the following three lemmas.

Lemma 4.4.1. If $\mathcal{O} = ((S_1, g_1, h_1), (S_2, g_2, h_2), \ldots, (S_k, g_k, h_k))$ is an optimality system for model $\mathcal{M}$, then the following assertions (i) and (ii) hold:

(i) For each positive integer n,

$$\frac{1}{n} J_n(\lambda, \pi, x) \geq g_k - \frac{2\|h_k\|}{n}, \quad x \in S_k, \quad \pi \in \mathcal{P}.$$

Consequently,

(ii) At each state $x \in S_k$, the constant g_k is a lower bound for the optimal inferior limit average cost function:

$$J_*(\lambda, x) \geq g_k, \quad x \in S_k;$$

see (4.2.1), (4.2.4) and (4.2.5).

Proof. Since $S_1 \cup \cdots \cup S_k = S$, from (4.3.3) it follows that $B(x) = A(x)$ when $x \in S_k$, and then the fourth part in Definition 4.3.1 yields that

$$e^{\lambda(g_k + h_k(x))} \leq e^{\lambda C(x,a)} \sum_{y \in S_k} p_{x\,y}(a) e^{\lambda h_k(y)}, \quad a \in A(x), \quad x \in S_k. \tag{4.4.2}$$

Now let $\pi \in \mathcal{P}$ be arbitrary. After integrating both sides of the above inequality with respect to $\pi_0(\cdot|x)$, it follows that

$$e^{\lambda(g_k + h_k(x))} \leq E_x^\pi \left[e^{\lambda C(X_0, A_0) + \lambda h_k(X_1)} I[X_1 \in S_k] \right], \quad x \in S_k, \quad \pi \in \mathcal{P}. \tag{4.4.3}$$

On the other hand, for every positive integer n, the Markov property yields that

$$\begin{aligned} E_x^\pi &\left[e^{\lambda \sum_{t=0}^{n} C(X_t, A_t) + \lambda h_k(X_{n+1})} I[X_r \in S_k,\, 1 \leq r \leq n+1] \middle| (X_m, A_m), m = 1, \ldots n \right] \\ &= e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t)} I[X_r \in S_k,\, 1 \leq r \leq n] e^{\lambda C(X_n, A_n)} \sum_{y \in S_k} p_{X_n\,y}(A_n) e^{\lambda h_k(y)} \\ &\geq e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t)} I[X_r \in S_k,\, 1 \leq r \leq n] e^{\lambda g_k + \lambda h_k(X_n)} \end{aligned}$$

where (4.4.2) was used to set the inequality. Therefore,

$$\begin{aligned} E_x^\pi &\left[e^{\lambda \sum_{t=0}^{n} C(X_t, A_t) + \lambda h_k(X_{n+1})} I[X_r \in S_k,\, 1 \leq r \leq n+1] \right] \\ &\geq e^{\lambda g_k} E_x^\pi \left[e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t) + \lambda h_k(X_n)} I[X_r \in S_k,\, 1 \leq r \leq n] \right] \end{aligned}$$

Combining this last relation and (4.4.3), a simple induction argument yields that, for every positive integer n, $x \in S_k$ and $\pi \in \mathcal{P}$,

$$E_x^\pi \left[e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t) + \lambda h_k(X_n)} I[X_r \in S_k,\, 1 \leq r \leq n] \right] \geq e^{\lambda(n g_k + h_k(x))},$$

and then

$$\begin{aligned} e^{\lambda(J_n(\lambda, \pi, x) + \|h_k\|)} &\geq E_x^\pi \left[e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t) + \lambda h_k(X_n)} \right] \\ &\geq E_x^\pi \left[e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t) + \lambda h_k(X_n)} I[X_r \in S_k,\, 1 \leq r \leq n] \right] \\ &\geq e^{\lambda(n g_k + h_k(x))} \geq e^{\lambda(n g_k - \|h_k\|)} \end{aligned}$$

so that, for every positive integer n,

$$\frac{1}{n} J_n(\lambda, \pi, x) \geq g_k - \frac{2\|h_k\|}{n}, \quad x \in S_k, \quad \pi \in \mathcal{P},$$

establishing part (i), and then the second assertion follows from (4.2.4) and (4.2.5). □

Now let the optimality system $\mathcal{O}$ be as in (4.3.1), suppose that $k > 1$ and set

$$\hat{S} = S_1 \cup \cdots \cup S_{k-1}. \tag{4.4.4}$$

Next, let $x \in \hat{S}$ be arbitrary, so that there exists $i < k$ such that $x \in S_i$ for some $i < k$; since

$$a \in B(x) \Rightarrow \sum_{y \in S_1 \cup \cdots \cup S_i} p_{x\,y}(a) = 1 \Rightarrow \sum_{y \in \hat{S}} p_{x\,y}(a) = 1,$$

it follows that

$$\hat{A}(x) := \{a \in A(x) \mid \textstyle\sum_{y \in \hat{S}} p_{x\,y}(a) = 1\}, \quad x \in \hat{S}, \tag{4.4.5}$$

is always nonempty. Set $\hat{\mathbb{K}} := \{(x, a) | x \in \hat{S}, a \in \hat{A}(x)\}$ and define the transition $\hat{P} = [\hat{p}_{x\,y}]$ and $\hat{C}: \hat{\mathbb{K}} \to \mathbb{R}$ by

$$\hat{p}_{x\,y}(a) := p_{x\,y}(a), \quad \hat{C}(x, a) := C(x, a), \quad (x, a) \in \hat{\mathbb{K}}, \quad y \in \hat{S}. \tag{4.4.6}$$

Definition 4.4.1. Let $\mathcal{O}$ be an optimality system for model $\mathcal{M}$ as in Definition 4.3.1, and suppose that the order k of $\mathcal{O}$ is larger than 1. With the notation in (4.4.4)–(4.4.6), the reduced model $\hat{\mathcal{M}}$ is specified by

$$\hat{\mathcal{M}} = (\hat{S}, A, \{\hat{A}(x)\}_{x \in \hat{S}}, \hat{C}, \hat{P}) \tag{4.4.7}$$

Combining Definitions 4.3.1 and 4.4.1 the following lemma follows immediately.

Lemma 4.4.2. If $\mathcal{O} = ((S_1, g_1, h_1), (S_2, g_2, h_2), \ldots, (S_k, g_k, h_k))$ is an optimality system for model model $\mathcal{M}$, where $k > 1$, then

$$\hat{\mathcal{O}} = ((S_1, g_1, h_1), (S_2, g_2, h_2), \ldots, (S_{k-1}, g_{k-1}, h_{k-1})), \tag{4.4.8}$$

is an optimality system for the reduced model $\hat{\mathcal{M}}$. Moreover, setting

$$\hat{B}(x) := \{x \in \hat{A}(x) \mid \sum_{y \in S_1 \cup \cdots \cup S_i} \hat{p}_{x\,y} = 1\}, \quad x \in S_i, \quad i = 1, 2, \ldots, k-1,$$

the equality $\hat{B}(x) = B(x)$ holds for every $x \in \hat{S}$; see (4.3.3).

Remark 4.4.1. The class of policies for model $\hat{\mathcal{M}}$ will be denoted by $\hat{\mathcal{P}}$. For $\Delta \in \hat{\mathcal{P}}$, $\hat{J}_-(\lambda, \Delta, \cdot)$ denotes the inferior limit λ-sensitive average cost criterion associated with Δ, and $\hat{J}_*(\lambda, \cdot) = \inf_{\Delta \in \hat{\mathcal{P}}} \hat{J}_-(\lambda, \Delta, \cdot)$ stands for the optimal inferior limit average cost function for model $\hat{\mathcal{M}}$. □

The following lemma is the final step before the proof of Theorem 4.4.1. Write

$$H_n := (X_0, A_0, \ldots, X_{n-1}, A_{n-1}, X_n) \tag{4.4.9}$$

Lemma 4.4.3. Let $\mathcal{O}$ in (4.3.1) be an optimality system for model $\mathcal{M}$, where $k > 1$. Suppose that for some $r \in \{1, 2, \ldots, k-1\}$, the state $x \in S_r$ and $\pi \in \mathcal{P}$ satisfy

$$J_-(\lambda, \pi, x) < g_r. \tag{4.4.10}$$

In this case, the following assertions (i)–(iii) hold:

(i) With probability 1 with respect to P_x^π, the actions chosen by π after observing H_n always belong to $\hat{A}(X_n)$. More precisely,

$$1 = P_x^\pi[\pi_n(\hat{A}(X_n)|H_n) = 1], \quad n \in \mathbb{N}.$$

Now let $w\colon \hat{S} \to A$ be a stationary policy for model $\hat{\mathcal{M}}$, that is, $w(x) \in \hat{A}(x)$ for each $x \in \hat{S}$, and define the policy $\Delta \in \hat{\mathcal{P}}$ as follows: For each $n \in \mathbb{N}$ and $\mathbf{h}_n \in \hat{\mathbb{H}}_n$,

$$\Delta_n(D|\mathbf{h}_n) := \pi_n(D \cap \hat{A}(x_n)|\mathbf{h}_n) + (1 - \pi_n(\hat{A}(x_n)|\mathbf{h}_n))\delta_{w(x_n)}(D), \quad D \in \mathcal{B}(A). \quad (4.4.11)$$

With this notation,
(ii) For every $n \in \mathbb{N}$,

$$P_x^{\Delta}[H_n \in D] = P_x^{\pi}[H_n \in D], \quad D \in \mathcal{B}(\hat{\mathbb{H}}_n), \quad (4.4.12)$$

and then,
(iii) $\hat{J}_-(\lambda, \Delta, x) = J_-(\lambda, \pi, x) < g_r$.

Proof. (i) The argument is by contradiction. Suppose that, for some $n \in \mathbb{N}$,

$$0 < P_x^{\pi}[\pi_n(A(X_n) \setminus \hat{A}(X_n)|H_n) > 0]. \quad (4.4.13)$$

Notice now that

$$\begin{aligned} P_x^{\pi}[X_{n+1} \in S_k|H_n] &= \int_{A(X_n)} \sum_{y \in S_k} p_{X_n\, y}(a)\pi_n(da|H_n) \\ &\geq \int_{A(X_n)\setminus\hat{A}(X_n)} \sum_{y \in S_k} p_{X_n\, y}(a)\pi_n(da|H_n) \\ &= \int_{A(X_n)\setminus\hat{A}(X_n)} \sum_{y \in S\setminus\hat{S}} p_{X_n\, y}(a)\pi_n(da|H_n). \end{aligned}$$

For $a \in A(X_n) \setminus \hat{A}(X_n)$ the summation inside the integral is positive, by (4.4.5), and then the integral is larger that zero on the event $[\pi_n(A(X_n) \setminus \hat{A}(X_n)|H_n) > 0]$. It follows from (4.4.13) that $P_x^{\pi}[X_{n+1} \in S_k|H_n] > 0$ with positive P_x^{π}-probability, so that

$$P_x^{\pi}[X_{n+1} \in S_k] > 0. \quad (4.4.14)$$

Next, given $\tilde{\mathbf{h}}_n \in \mathbb{H}_n$ and $\tilde{a} \in A(x_n)$, define the (shifted) policy $\pi^{\tilde{\mathbf{h}}_n, \tilde{a}}$ as follows:

$$\pi_t^{\tilde{\mathbf{h}}_n, \tilde{a}}(\cdot|\mathbf{h}_t) := \pi_{n+1+t}(\cdot|\tilde{h}_n, \tilde{a}, \mathbf{h}_t), \quad \mathbf{h}_t \in \mathbb{H}_t, \quad t \in \mathbb{N}.$$

With this specification, the Markov property yields that for every $m > n + 1$

$$\begin{aligned} &E_x^{\pi}[e^{\lambda \sum_{t=0}^{m-1} C(X_t, A_t)} I[X_{n+1} \in S_k]|H_n, A_n, X_{n+1}] \\ &\quad = e^{\lambda \sum_{t=0}^{n} C(X_t, A_t)} I[X_{n+1} \in S_k] E_{X_{n+1}}^{\pi^{H_n, A_n}}[e^{\lambda \sum_{t=0}^{m-n-2} C(X_t, A_t)}] \\ &\quad \geq e^{-\lambda(n+1)\|C\|} I[X_{n+1} \in S_k] E_{X_{n+1}}^{\pi^{H_n, A_n}}[e^{\lambda \sum_{t=0}^{m-n-2} C(X_t, A_t)}] \\ &\quad \geq e^{-\lambda(n+1)\|C\|} I[X_{n+1} \in S_k] e^{\lambda J_{m-n-1}(\lambda, \pi^{H_n, A_n}, X_{n+1})}; \end{aligned}$$

see (4.2.1). From this point, Lemma 4.4.1(i) yields that

$$\begin{aligned} &E_x^{\pi}[e^{\lambda \sum_{t=0}^{m-1} C(X_t, A_t)} I[X_{n+1} \in S_k]|H_n, A_n, X_{n+1}] \\ &\quad \geq e^{-\lambda(n+1)\|C\|} I[X_{n+1} \in S_k] e^{\lambda(m-n-1)g_k - 2\lambda\|h_k\|} \end{aligned}$$

and then

$$\begin{aligned}e^{\lambda J_m(\lambda,\pi,x)} &= E_x^\pi[e^{\lambda\sum_{t=0}^{m-1} C(X_t,A_t)}]\\ &\geq E_x^\pi[e^{\lambda\sum_{t=0}^{m-1} C(X_t,A_t)} I[X_{n+1}\in S_k]]\\ &\geq e^{-\lambda(n+1)\|C\|}P_x^\pi[X_{n+1}\in S_k]e^{\lambda(m-n-1)g_k-2\lambda\|h_k\|}\end{aligned}$$

Using (4.4.14), this inequality immediately yields that

$$J_-(\lambda,\pi,x)=\liminf_{m\to\infty}\frac{1}{m}J_m(\lambda,\pi,x)\geq g_k$$

and then (4.4.10) implies that $g_k < g_r$, an inequality that, recalling that $r < k$, contradicts (4.3.2). Therefore, (4.4.13) does not hold and it follows that $0 = P_x^\pi[\pi_n(A(X_n)\setminus \hat{A}(X_n)|H_n) > 0]$, that is, $1 = P_x^\pi[\pi_n(\hat{A}(X_n)|H_n) = 1]$.

(ii) The argument is by induction. For $n = 0$, both sides of (4.4.12) are equal to $\delta_x(D)$. Assume now that (4.4.12) holds for certain nonnegative integer n, and let $D \in \mathcal{B}(\hat{\mathbb{H}}_n)$, $D_1 \in \mathcal{B}(\hat{A})$ and $D_2 \subset \hat{S}$ be arbitrary. Next observe that

$$P_x^\pi[H_n\in D, A_n\in D_1, X_{n+1}\in D_2|H_n] = I[H_n\in D]\int_{a\in D_1}\sum_{y\in D_2}p_{X_n\,y}(a)\pi_n(da|H_n);$$

since the equality $\Delta_n(\cdot|H_n) = \pi_n(\cdot|H_n)$ holds P_x^π-a.s., by part (i) and (4.4.11), it follows that

$$\begin{aligned}&P_x^\pi[H_n\in D, A_n\in D_1, X_{n+1}\in D_2|H_n]\\ &\qquad= I[H_n\in D]\int_{a\in D_1}\sum_{y\in D_2}p_{X_n\,y}(a)\Delta_n(da|H_n)\quad P_x^\pi\text{-a.s.},\end{aligned}$$

and then

$$P_x^\pi[H_n\in D, A_n\in D_1, X_{n+1}\in D_2] = \int_{\mathbf{h}_n\in D}\left[\int_{a\in D_1}\sum_{y\in D_2}p_{x_n\,y}(a)\Delta_n(da|\mathbf{h}_n)\right]P_x^\pi[d\mathbf{h}_n].$$

By the induction hypothesis the distribution of H_n is the same under P_x^π and P_x^Δ, so that

$$P_x^\pi[H_n\in D, A_n\in D_1, X_{n+1}\in D_2] = \int_{\mathbf{h}_n\in D}\left[\int_{a\in D_1}\sum_{y\in D_2}p_{x_n\,y}(a)\Delta_n(da|\mathbf{h}_n)\right]P_x^\Delta[d\mathbf{h}_n],$$

that is,

$$P_x^\pi[H_n\in D, A_n\in D_1, X_{n+1}\in D_2] = P_x^\Delta[H_n\in D, A_n\in D_1, X_{n+1}\in D_2].$$

Since $D\in\mathcal{B}(\hat{H}_n), D_1\in\mathcal{B}(\hat{A})$ and $D_2\subset\hat{S}$ are arbitrary, Theorem 10.4 in Billingsley (1995) yields that (4.4.12) holds with $n+1$ instead of n.

(iii) The previous part yields that $E_x^\Delta[e^{\lambda\sum_{t=0}^{n-1}C(X_t,A_t)}] = E_x^\pi[e^{\lambda\sum_{t=0}^{n-1}C(X_t,A_t)}]$ for every positive integer n. Therefore, $J_n(\lambda,\Delta,x)/n = J_n(\lambda,\pi,x)/n$, and the conclusion follows taking the inferior limit as n goes to ∞ in both sides of this equality. □

After the above preliminaries, the proof of the main result of this section is presented below.

Proof of Theorem 4.4.1. The argument is by induction in the order k of the optimality system $\mathcal{O}$. If $k = 1$ then (4.4.1) follows from Lemma 4.4.1(ii). Suppose now that (4.4.1) holds when $k = m - 1$ for certain integer $m \geq 2$, and let $\mathcal{O}$ be an optimality system for $\mathcal{M}$ with order m. The reduced optimality system $\hat{\mathcal{O}}$ in Definition 4.4.1 has order $m - 1$, and then the optimal inferior limit average cost corresponding to $\hat{\mathcal{M}}$ satisfies

$$\hat{J}_*(\lambda, x) \geq g_i, \quad x \in S_i, \quad i = 1, 2, \ldots, m-1. \tag{4.4.15}$$

by the induction hypothesis, a fact that will be used to verify that

$$J_*(\lambda, x) \geq g_i, \quad x \in S_i, \quad i = 1, 2, \ldots, m-1. \tag{4.4.16}$$

Indeed, if this relation fails, there exist $r < m$ and a state $x \in S_r$ such that $J_*(\lambda, x) < g_r$, and then $J_-(\lambda, \pi, x) < g_r$ for some policy $\pi \in \mathcal{P}$. Using Lemma 4.4.3(iii), there exists a policy $\Delta \in \hat{\mathcal{P}}$ such that $\hat{J}_-(\lambda, \Delta, x) = J_-(\lambda, \pi, x) < g_r$, and then $\hat{J}_*(\lambda, x) < g_r$, contradicting (4.4.15). Thus, (4.4.16) holds, whereas an application of Lemma 4.4.1(i) to the present optimality system of order m yields that $J^*(\lambda, x) \geq g_m$ for all $x \in S_m$, a fact that together with (4.4.16) yields that (4.4.1) holds when $k = m$, concluding the argument. □

4.5. Proof of the Verification Theorem

In this section Theorem 4.3.1 will be established. The argument combines Theorem 4.4.1 with the following result, which provides an upper bound for the (superior limit) average cost function associated with a stationary policy f satisfying (4.3.6). Although such a result can be obtained from Cavazos-Cadena and Hernández-Hernández (2006), for the sake of completeness a different proof is presented, which uses simple probabilistic arguments. The following notation is involved in the argument: For each set $W \subset S$, the corresponding hitting time is given by

$$T_W := \min\{n > 0 \,|\, X_n \in W\}, \tag{4.5.1}$$

where the minimum of the empty set is ∞.

Theorem 4.5.1. (i) Let f be a stationary policy as in the statement of Theorem 4.3.1(ii). In this case,

$$J(\lambda, f, x) \leq g_i, \quad x \in S_i, \quad i = 1, 2, \ldots, k. \tag{4.5.2}$$

Consequently,

(ii) For each $i \in \{1, 2, \ldots, k\}$ and $x \in S_i$, $J^*(\lambda, x) \leq g_i$.

Proof. To begin with, notice that (4.3.3) and (4.3.5) together imply that the set $S_1 \cup \cdots S_i$ is closed under the action of policy f, that is, for each $i = 1, 2, \ldots, k$,

$$x \in S_1 \cup \cdots \cup S_i \quad \text{and} \quad p_{x\,y}(f(x)) > 0 \Rightarrow y \in S_1 \cup \cdots \cup S_i. \tag{4.5.3}$$

On the other hand, starting from (4.3.6), a standard induction argument using the Markov property yields that, for every positive integer n

$$e^{\lambda(ng_i + h_i(x))} = E_x^f[e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t) + \lambda h_i(X_n)} I[X_t \in S_i,\, t < n]], \quad x \in S_i, \quad i = 1, 2, \ldots, k. \tag{4.5.4}$$

Since (4.5.3) implies that $1 = P_x^f[X_t \in S_1]$ for every $x \in S_1$ and $t \in \mathbb{N}$, it follows that if the initial state x belongs to S_1, the equality $e^{\lambda(ng_1 + h_1(x))} = E_x^f \left[e^{\lambda \sum_{t=0}^{n-1} C(X_t, A_t)} e^{\lambda h_1(X_n)} \right]$ holds for each $n > 0$, and in this case $E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} \right] \leq e^{\lambda(ng_1 + 2\|h_1\|)}$, that is,

$$J_n(\lambda, x, f) \leq ng_1 + 2\|h_1\|, \quad x \in S_1, \quad n = 1, 2, 3, \ldots \tag{4.5.5}$$

see (4.2.1). Next, for $i \in \{1, 2, \ldots, k\}$ consider the following claim.

$\mathcal{C}_i$: $J(\lambda, f, x) \leq g_i$ for every $x \in S_i$.

It will be proved, by induction, that $\mathcal{C}_i$ is valid for every $i = 1, 2, \ldots, k$. To achieve this goal, observe that (4.5.5) implies that

$$J(\lambda, f, x) = \limsup_{n\to\infty} \frac{1}{n} J_n(\lambda, x, f) \leq g_1, \quad x \in S_1,$$

so that $\mathcal{C}_1$ is valid. Now, suppose that $\mathcal{C}_j$ holds for $j = 1, 2, \ldots, i-1$, where $i \in \{2, 3, \ldots, k\}$. In this case, given $\varepsilon > 0$, for each $x \in S_j$ with $1 \leq j \leq i-1$, there exists a positive integer $N(x)$ such that $J_n(\lambda, f, x)/n \leq g_j + \varepsilon$ for $n \geq N(x)$, a relation that *via* (4.2.1) is equivalent to

$$E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} \right] \leq e^{\lambda n (g_j + \varepsilon)}, \quad n \geq N(x);$$

since $g_j \leq g_i$ for $j < i$, by (4.3.2), it follows that

$$E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} \right] \leq D(\varepsilon) e^{\lambda n (g_i + \varepsilon)}, \quad x \in S_1 \cup \cdots \cup S_{i-1}, \quad n = 1, 2, 3, \ldots, \tag{4.5.6}$$

where, setting

$$\tilde{D}(\varepsilon) := \max \left\{ e^{-\lambda n (g_i + \varepsilon)} E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} \right] \,\middle|\, 1 \leq n < N(x),\ x \in S_1 \cup \cdots \cup S_{i-1} \right\},$$

$D(\varepsilon)$ is given by

$$D(\varepsilon) = \max\{\tilde{D}(\varepsilon), 1\}.$$

Next, let $x \in S_i$ be arbitrary but fixed, and observe that (4.5.3) yields that

$$P_x^f [X_t \in S_1 \cup \cdots \cup S_i,\ t = 1, 2, 3, \ldots] = 1. \tag{4.5.7}$$

Combining this relation with the specification of the hitting time T_W in (4.5.1), it follows that for every positive integers n and r the following equalities occur with probability 1 with respect to P_x^f:

$$I[T_{S_1 \cup \cdots \cup S_{i-1}} = r] = I[X_m \in S_i, 1 \leq m < r] I[X_r \in S_1 \cup \cdots \cup S_{i-1}]$$
$$I[T_{S_1 \cup \cdots \cup S_{i-1}} > n-1] = I[X_m \in S_i, 1 \leq m \leq n-1]$$

Therefore, for a positive integer n,

$$\begin{aligned} E_x^f &\left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} \right] \\ &= \sum_{r=1}^{n-1} E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} I[T_{S_1 \cup \cdots \cup S_{i-1}} = r] \right] \\ &\quad + E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} I[T_{S_1 \cup \cdots \cup S_{i-1}} > n-1] \right] \\ &= \sum_{r=1}^{n-1} E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} I[X_m \in S_i, 1 \leq m < r] I[X_r \in \cup_{j=1}^{i-1} S_j] \right] \\ &\quad + E_x^f \left[e^{\sum_{t=0}^{n-1} C(X_t, A_t)} I[X_t \in S_i,\ t = 1, 2, \ldots, n-1] \right]. \end{aligned} \tag{4.5.8}$$

To continue, each one of the terms in this last equality will be analyzed. First, recalling that $x \in S_i$, notice that (4.5.4) immediately implies that

$$E_x^f \left[e^{\sum_{t=0}^{r-1} C(X_t, A_t)} I[X_t \in S_i,\ t = 1, 2, \ldots, r-1] \right] \leq e^{\lambda (r g_i + 2\|h_i\|)}, \quad r = 1, 2, 3, \ldots. \tag{4.5.9}$$

Next, for $r \in \{1, 2, \ldots, n-1\}$, the Markov property yields

$$\begin{aligned}
&E_x^f\left[e^{\sum_{t=0}^{n-1} C(X_t,A_t)} I[X_m \in S_i, 1 \le m < r] I[X_r \in \cup_{j=1}^{i-1} S_j] \,\middle|\, H_r\right] \\
&\quad = I[X_m \in S_i, 1 \le m < r] I[X_r \in \cup_{j=1}^{i-1} S_j] e^{\sum_{t=0}^{r-1} C(X_t,A_t)} E_{X_r}^f\left[e^{\sum_{t=0}^{n-r-1} C(X_t,A_t)} \,\middle|\, H_r\right] \\
&\quad \le I[X_m \in S_i, 1 \le m < r] I[X_r \in \cup_{j=1}^{i-1} S_j] e^{\sum_{t=0}^{r-1} C(X_t,A_t)} D(\varepsilon) e^{\lambda(n-r)(g_i+\varepsilon)} \\
&\quad \le I[X_m \in S_i, 1 \le m < r] e^{\sum_{t=0}^{r-1} C(X_t,A_t)} D(\varepsilon) e^{\lambda(n-r)(g_i+\varepsilon)}
\end{aligned}$$

where (4.5.6) was used to set the first inequality. Thus,

$$\begin{aligned}
&E_x^f\left[e^{\sum_{t=0}^{n-1} C(X_t,A_t)} I[X_m \in S_i, 1 \le m < r] I[X_r \in \cup_{j=1}^{i-1} S_j]\right] \\
&\quad \le D(\varepsilon) e^{\lambda(n-r)(g_i+\varepsilon)} E_x^f\left[e^{\sum_{t=0}^{r-1} C(X_t,A_t)} I[X_m \in S_i, 1 \le m < r]\right] \\
&\quad \le D(\varepsilon) e^{\lambda(n-r)(g_i+\varepsilon)} e^{\lambda(r g_i + 2\|h_i\|)} \\
&\quad \le e^{2\lambda\|h_i\|} D(\varepsilon) e^{\lambda n(g_i+\varepsilon)}
\end{aligned}$$

where (4.5.9) was used to set the second inequality. Combining this last display and (4.5.9), from (4.5.8) it follows that

$$\begin{aligned}
e^{\lambda J_n(\lambda,f,x)} &= E_x^f\left[e^{\sum_{t=0}^{n-1} C(X_t,A_t)}\right] \\
&\le \sum_{r=1}^{n-1} e^{2\lambda\|h_i\|} D(\varepsilon) e^{\lambda n(g_i+\varepsilon)} + e^{2\lambda\|h_i\|} e^{\lambda n g_i} \\
&\le n e^{2\lambda\|h_i\|} \max\{D(\varepsilon), 1\} e^{\lambda n(g_i+\varepsilon)}
\end{aligned}$$

that is,

$$J_n(\lambda, f, x) \le \frac{\log(n) + 2\lambda\|h_i\| + \log(\max\{D(\varepsilon), 1\})}{\lambda} + n(g_i + \varepsilon),$$

a relation that leads to

$$J(\lambda, f, x) = \limsup_{n\to\infty} \frac{1}{n} J_n(\lambda, f, x) \le g_i + \varepsilon;$$

since $x \in S_i$ and $\varepsilon > 0$ are arbitrary, it follows that $\mathcal{C}_i$ holds, concluding the induction argument. Therefore $\mathcal{C}_j$ occurs for every $j = 1, 2, \ldots, k$, a fact that is equivalent to (4.5.2). □

Proof of Theorem 4.3.1. Since $J_*(\lambda, \cdot) \le J^*(\lambda, \cdot)$, Theorems 4.4.1 and 4.5.1(ii) together yield that

$$g_i \le J_*(\lambda, x) \le J^*(\lambda, x) \le g_i, \quad x \in S_i, \quad i = 1, 2, \ldots, k.$$

a relation that immediately implies parts (i) and (ii). Now, let $f \in \mathbb{F}$ be as in (4.3.5) and (4.3.6). Using that $J(\lambda, f, \cdot) \ge J_-(\lambda, f, \cdot) \ge J_*(\lambda, \cdot)$, by (4.2.2), (4.2.4) and (4.2.5), the above displayed relation and Theorem 4.5.1(i) lead to

$$J(\lambda, f, x) = J_-(\lambda, f, x) = g_i = J^*(x), \quad x \in S_i, \quad i = 1, 2, \ldots, k,$$

where part (i) was used to set the last equality. Therefore, f is λ-optimal and, *via* (4.2.2) and (4.2.4), $\lim_{n\to\infty} J_n(\lambda, f, x)/n = J^*(\lambda, x)$ for all $x \in S$, completing the proof. □

4.6. Discounted Approach

This section presents the necessary technical tools that will be used to establish the existence of an optimality system for model $\mathcal{M}$. The approach relies on the discounted operators introduced below which, when λ is small enough and appropriate communication conditions are satisfied by the transition law, have been used to construct solutions of the optimality equation (4.2.6) (Di Masi and Stettner 1999, Cavazos-Cadena 2003).

Definition 4.6.1. Given $\alpha \in (0,1)$ define the operator $T_\alpha\colon \mathcal{B}(S) \to \mathcal{B}(S)$ as follows: For each $V \in \mathcal{B}(S)$ and $x \in S$, $T_\alpha[V](x)$ is determined by

$$e^{\lambda T_\alpha[V](x)} = \inf_{a\in A(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S} p_{x\,y}(a) e^{\lambda \alpha V(y)} \right], \quad x \in S. \tag{4.6.1}$$

According to this specification, $T_\alpha[V](x)$ is the minimum certain equivalent of the random cost $C(X_0, A_0) + \alpha V(X_1)$ that can be achieved when the initial state is $X_0 = x$. On the other hand, it is not difficult to see that T is a monotone and α-homogeneous operator, that is, for $V, W \in \mathcal{B}(S)$ (i) $V \ge W$ implies that $T[V] \ge T[W]$, and (ii) $T[V + r] = T[V] + \alpha r$ for every $r \in \mathbb{R}$. Combining these properties with the relation $W - \|W - V\| \le V \le W + \|W - V\|$, it follows that

$$T[W] - \alpha\|W - V\| \le T[V] \le T[W] + \alpha\|W - V\|, \quad V, W \in \mathcal{B}(S), \tag{4.6.2}$$

so that $\|T[W] - T[V]\| \le \alpha\|V - W\|$, showing that T_α is a contractive operator on the space $\mathcal{B}(S)$ endowed with the maximum norm. Consequently, by Banach's fixed point theorem, there exists a unique function $V_\alpha \in \mathcal{B}(S)$ satisfying $T_\alpha[V_\alpha] = V_\alpha$, that is,

$$e^{\lambda V_\alpha(x)} = \inf_{a\in A(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S} p_{x\,y}(a) e^{\lambda \alpha V_\alpha(y)} \right], \quad x \in S, \quad \alpha \in (0,1). \tag{4.6.3}$$

Notice now that (4.6.1) yields that $T_\alpha[0](x) = \inf C(x,a)$, so that $\|T_\alpha[0]\| \le \|C\|$. Using (4.6.2) with V_α and 0 instead of W and V, respectively, it follows that

$$(1 - \alpha)\|V_\alpha\| \le \|C\|. \tag{4.6.4}$$

In the remainder of the section, the family $\{V_\alpha\}_{\alpha\in(0,1)}$ of fixed points will be used to construct the components of an optimality system, and the idea in the following definition is the essential step in that direction. Throughout the remainder, $\{\alpha_m\} \subset (0,1)$ is a fixed sequence satisfying the following requirements:

$$\alpha_m \nearrow 1 \quad \text{as} \quad m \nearrow \infty, \tag{4.6.5}$$

and

$$\begin{aligned}&\text{For evey } x, y \in S, \text{ the following limits exist:}\\ &\lim_{m\to\infty} [V_{\alpha_m}(x) - V_{\alpha_m}(y)] \in [-\infty, \infty]\\ &\lim_{m\to\infty} (1 - \alpha_m) V_{\alpha_m}(x) \in [-\|C\|,\ \|C\|]\end{aligned} \tag{4.6.6}$$

where the last inclusion follows from (4.6.4).

Definition 4.6.2. The relation '$\sim$' in the state space S is specified as follows:

$$x \sim y \iff \lim_{m\to\infty} [V_{\alpha_m}(x) - V_{\alpha_m}(y)] \in (-\infty, \infty). \tag{4.6.7}$$

From this definition it is not difficult to see that '$\sim$' is an equivalence relation, and then it induces a partition of S into equivalence classes. Notice that for $x, y \in S$, (4.6.6) and Definition 4.6.2 yield that

$$x \not\sim y \iff \lim_{m\to\infty}[V_{\alpha_m}(x) - V_{\alpha_m}(y)] = \infty \quad \text{or} \quad \lim_{m\to\infty}[V_{\alpha_m}(x) - V_{\alpha_m}(y)] = -\infty; \qquad (4.6.8)$$

moreover,

$$\begin{gathered}\text{if } x \sim x_1 \text{ and } y \sim y_1 \text{ and } \lim_{m\to\infty}[V_{\alpha_m}(x) - V_{\alpha_m}(y)] = \infty,\\ \text{then} \lim_{m\to\infty}[V_{\alpha_m}(x_1) - V_{\alpha_m}(y_1)] = \infty.\end{gathered} \qquad (4.6.9)$$

Definition 4.6.3. The relation '$\prec$' in the family of equivalence classes determined by the the equivalence relation in (4.6.7) is defined as follows: If $\mathcal{E}$ and $\mathcal{E}'$ are two different equivalence classes, then

$$\mathcal{E} \prec \mathcal{E}' \iff \lim_{m\to\infty}[V_{\alpha_m}(x) - V_{\alpha_m}(y)] = \infty \text{ for some } x \in \mathcal{E}' \text{ and some } y \in \mathcal{E}.$$

By (4.6.9) this relation is well-defined, whereas (4.6.8) implies that $\prec$ is a (strict) total order, that is, if $\mathcal{E}$ and $\mathcal{E}'$ are two different equivalences classes, then either $\mathcal{E} \prec \mathcal{E}'$ or $\mathcal{E}' \prec \mathcal{E}$. Moreover, combining the above definition and (4.6.9), it follows that

$$\mathcal{E} \prec \mathcal{E}' \iff \lim_{m\to\infty}[V_{\alpha_m}(x) - V_{\alpha_m}(y)] = \infty \quad \text{for all } x \in \mathcal{E}' \text{ and all } y \in \mathcal{E}. \qquad (4.6.10)$$

Throughout the remainder,

$$S_1^*, \ldots, S_k^* \text{ are the different equivalence clasess of } S \text{ with respect to '}\sim\text{'} \qquad (4.6.11)$$

where, without loss of generality, the labeling of the equivalence classes is such that

$$S_i^* \prec S_{i+1}^* \quad 1 \leq i < k; \qquad (4.6.12)$$

also, the states $x_1, \ldots, x_k$ are fixed and satisfy

$$x_i \in S_i^*, \quad i = 1, 2, \ldots, k. \qquad (4.6.13)$$

Now, for $i \in \{1, 2, \ldots, k\}$, define

$$g_i^* := \lim_{m\to\infty}(1 - \alpha_m)V_{\alpha_m}(x_i), \qquad (4.6.14)$$

and

$$h_i^*(x) = \lim_{m\to\infty}[V_{\alpha_m}(x) - V_{\alpha_m}(x_i)], \quad x \in S_i^*. \qquad (4.6.15)$$

Notice that $g_i^* \in [-\|C\|, \|C\|]$, by (4.6.4) whereas, observing that $x_i \sim x$ for every $x \in S_i$, from Definition 4.6.2 it follows that $h_i(x)$ is finite for every $x \in S_i^*$; the above objects S_i^*, g_i^* and $h_i^*(\cdot)$ will be used to build an optimality system for model $\mathcal{M}$.

4.7. Proof of the Existence Result

In this section it will be verified that an optimality system for model $\mathcal{M}$ exists. With the notation in (4.6.11)–(4.6.15), define the sequence of triplets $\mathcal{O}^*$ as follows:

$$\mathcal{O}^* := ((S_1^*, g_1^*, h_1^*), \ldots, (S_k^*, g_k^*, h_k^*)). \qquad (4.7.1)$$

Proof of Theorem 4.3.2. It will be shown that $\mathcal{O}^*$ specified above is an optimality system for model $\mathcal{M}$. To achieve this goal, the four conditions in Definition 4.3.1 will be verified.

(i) Since $S_1^*, \ldots, S_k^*$ are the different equivalence classes of S with respect to the equivalence relation in Definition 4.6.2, those S_i^* sets form a partition of S.

(ii) As already noted, g_i^* is a finite number and $h_i^* \in \mathcal{B}(S_i^*)$. Now let $i < j$ be arbitrary in $\{1, 2, \ldots, k\}$. Recall now that $x_i \in S_i$ and $x_j \in S_j$, by (4.6.13), and combine Definition 4.6.3 with (4.6.10) and (4.6.12) to obtain that $\lim_{m\to\infty}[V_{\alpha_m}(x_j) - V_{\alpha_m}(x_i)] = \infty$, so that $V_{\alpha_m}(x_j) > V_{\alpha_m}(x_i)$ for m large enough, a fact that leads to

$$g_j^* = \lim_{m\to\infty}(1-\alpha_m)V_{\alpha_m}(x_j) \geq \lim_{m\to\infty}(1-\alpha_m)V_{\alpha_m}(x_i) = g_i^*,$$

and then $g_1^* \leq \cdots \leq g_k^*$.

(iii) Setting

$$B^*(x) = \{a \in A(x) \,|\, \sum_{y \in S_1^* \cup \cdots \cup S_i^*} p_{x\,y}(a) = 1\}, \quad x \in S_i^*, \quad i = 1, 2, \ldots, k, \tag{4.7.2}$$

it will be shown below that $B^*(x)$ is always a nonempty set. To achieve this goal, notice that Assumption 4.2.1 yields that, for each $\alpha \in (0, 1)$, there exists a policy $f_\alpha \in \mathbb{F}$ such that, for every $x \in S$,

$$e^{\lambda V_\alpha(x)} = e^{\lambda C(x, f_\alpha(x))} \sum_{y \in S} p_{x\,y}(f_\alpha(x)) e^{\lambda \alpha V_\alpha(y)}. \tag{4.7.3}$$

Now, let the sequence $\{\alpha_m\}$ be as in (4.6.5) and (4.6.6), and consider the sequence $\{f_{\alpha_m}\} \subset \mathbb{F}$. Recalling that $\mathbb{F}$ is a compact metric space, taking a subsequence (if necessary), without loss of generality it can be assumed that there exists $f^* \in \mathbb{F}$ such that

$$\lim_{m\to\infty} f_{\alpha_m}(x) = f^*(x). \tag{4.7.4}$$

Next, it will be shown that $f^*(x)$ always belongs to $B^*(x)$, an assertion that will be verified by contradiction. Let $i \in \{1, 2, \ldots, k\}$ and $x \in S_i^*$ be arbitrary but fixed, and *suppose* that

$$p_{x\,z}(f^*(x)) > 0 \quad \text{for some } z \in S_j^* \text{ where } j > i. \tag{4.7.5}$$

Replacing α by α_m in (4.7.3) and multiplying both sides of the resulting equality by $e^{-\lambda V_{\alpha_m}(x_i)}$, where x_i is the fixed state in (4.6.13), direct calculations yield that

$$e^{\lambda(1-\alpha_m)V_{\alpha_m}(x_i) + \lambda[V_{\alpha_m}(x) - V_{\alpha_m}(x_i)]} = e^{\lambda C(x, f_{\alpha_m}(x))} \sum_{y \in S} p_{x\,y}(f_{\alpha_m}(x)) e^{\lambda \alpha_m [V_{\alpha_m}(y) - V_{\alpha_m}(x_i)]}, \tag{4.7.6}$$

and then

$$e^{\lambda(1-\alpha_m)V_{\alpha_m}(x_i) + \lambda[V_{\alpha_m}(x) - V_{\alpha_m}(x_i)]} \geq e^{\lambda C(x, f_{\alpha_m}(x))} p_{x\,z}(f_{\alpha_m}(x)) e^{\lambda \alpha_m [V_{\alpha_m}(z) - V_{\alpha_m}(x_i)]}. \tag{4.7.7}$$

Since $x, x_i \in S_i^*$, taking the limit as m goes to ∞ in both sides of this inequality, the continuity of the transition law and the cost function together with (4.6.6), (4.6.14), (4.6.15) and (4.7.4), lead to

$$e^{\lambda g_i^* + \lambda h_i^*(x)} \geq e^{\lambda C(x, f^*(x))} p_{x\,z}(f^*(x)) e^{\lambda \lim_{m\to\infty}[V_{\alpha_m}(z) - V_{\alpha_m}(x_i)]};$$

since $z \in S_j^*$ and $x_i \in S_i^*$ with $j > i$, *via* (4.6.10) and (4.6.12) it follows that

$$\lim_{m\to\infty}[V_{\alpha_m}(z) - V_{\alpha_m}(x_i)] = \infty,$$

so that, recalling that λ and $p_{x\,z}(f^*(x))$ are positive, the above display yields that

$$e^{\lambda g_i^* + \lambda h_i^*(x)} \geq \infty,$$

a contradiction that stems from (4.7.5). Therefore, $p_{x\,z}(f^*(x)) = 0$ when $z \in S_j^*$ with $j > i$, and it follows that

$$\sum_{y \in S_1^* \cup \cdots \cup S_i^*} p_{x,y}(f^*(x)) = 1,$$

that is,

$$f^*(x) \in B^*(x); \tag{4.7.8}$$

since $x \in S_i^*$ and $i \in \{1, 2, \ldots, k\}$ were arbitrary in this argument, it follows that $B^*(x)$ is always a nonempty set.

(iv) It will be verified that

$$e^{\lambda(g_i^* + h_i^*(x))} = \inf_{a \in B^*(x)} \left[e^{\lambda C(x,a)} \sum_{y \in S_i^*} p_{x\,y}(a) e^{\lambda h_i^*(y)} \right], \quad x \in S_i^*. \tag{4.7.9}$$

Let $i \in \{1, 2, \ldots, k\}$ and $x \in S_i^*$ be arbitrary but fixed. Now take an arbitrary action $a \in B^*(x) \subset A(x)$ and notice that (4.7.2) yields that $p_{x\,y}(a) = 0$ when $y \notin S_1^* \cup \cdots \cup S_i^*$. Using this fact (4.6.3) implies that, for every positive integer m,

$$e^{\lambda V_{\alpha_m}(x)} \leq e^{\lambda C(x,a)} \sum_{y \in S_1^* \cup \cdots \cup S_i^*} p_{x\,y}(a) e^{\lambda \alpha_m V_{\alpha_m}(y)},$$

and multiplying both sides of this inequality by $e^{-\lambda V_{\alpha_m}(x_i)}$ it follows that

$$e^{\lambda(1-\alpha_m)V_{\alpha_m}(x_i) + \lambda[V_{\alpha_m}(x) - V_{\alpha_m}(x_i)]} \leq e^{\lambda C(x,a)} \sum_{y \in S_1^* \cup \cdots \cup S_i^*} p_{x\,y}(a) e^{\lambda \alpha_m [V_{\alpha_m}(y) - V_{\alpha_m}(x_i)]};$$

recalling that $x_i \in S_i^*$ and using (4.6.6), (4.6.14) and (4.6.15), taking the limit as m goes to ∞ in both sides of the above inequality the following relation is obtained:

$$\begin{aligned} e^{\lambda g_i^* + \lambda h_i^*(x)} \leq {} & e^{\lambda C(x,a)} \sum_{y \in S_i^*} p_{x\,y}(a) e^{\lambda h_i^*(y)} \\ & + e^{\lambda C(x,a)} \sum_{y \in \cup_{1 \leq j < i} S_j^*} p_{x\,y}(a) e^{\lambda \lim_{m\to\infty} [V_{\alpha_m}(y) - V_{\alpha_m}(x_i)]}. \end{aligned} \tag{4.7.10}$$

Since

$$\lim_{m\to\infty} [V_{\alpha_m}(y) - V_{\alpha_m}(x_i)] = -\infty \text{ when } y \in S_j^* \text{ with } j < i, \tag{4.7.11}$$

by (4.6.10) and (4.6.12), the positivity of λ yields that the second summation in the above display vanishes, so that

$$e^{\lambda g_i^* + \lambda h_i^*(x)} \leq e^{\lambda C(x,a)} \sum_{y \in S_i^*} p_{x\,y}(a) e^{\lambda h_i^*(y)}$$

and then, since $a \in B^*(x)$ was arbitrary in this argument,

$$e^{\lambda g_i^* + \lambda h_i^*(x)} \leq \inf_{a \in B^*(x)} \left[e^{\lambda C(x,a)} \sum_{y \in S_i^*} p_{x\,y}(a) e^{\lambda h_i^*(y)} \right]. \tag{4.7.12}$$

To establish the reverse inequality, notice that (4.7.6) yields that

$$\begin{aligned} &e^{\lambda(1-\alpha_m)V_{\alpha_m}(x_i)+\lambda[V_{\alpha_m}(x)-V_{\alpha_m}(x_i)]} \\ &\qquad \geq e^{\lambda C(x,f_{\alpha_m}(x))} \sum_{y\in S_i^*\cup\cdots\cup S_i^*} p_{x\,y}(f_{\alpha_m}(x))e^{\lambda\alpha_m[V_{\alpha_m}(y)-V_{\alpha_m}(x_i)]}. \end{aligned}$$

Taking the limit as m goes to ∞, the specifications of g_I^* and $h_i^*(\cdot)$ together with Assumption 4.2.1 and (4.7.4) lead to

$$\begin{aligned} e^{\lambda g_i^*+\lambda h_i^*(x)} &\geq e^{\lambda C(x,f^*(x))} \sum_{y\in S_i^*} p_{x\,y}(f^*(x))e^{\lambda h_i^*(y)} \\ &\qquad + e^{\lambda C(x,f_{\alpha_m}(x))} \sum_{y\in\cup_{1\leq j<i}S_j} p_{x\,y}(f^*(x))e^{\lambda \lim_{m\to\infty}[V_{\alpha_m}(y)-V_{\alpha_m}(x_i)]} \end{aligned}$$

and then (4.7.11) and the positivity of λ yield that

$$\begin{aligned} e^{\lambda g_i^*+\lambda h_i^*(x)} &\geq e^{\lambda C(x,f^*(x))} \sum_{y\in S_i^*} p_{x\,y}(f^*(x))e^{\lambda h_i^*(y)} \\ &\geq \inf_{a\in B^*(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S_i^*} p_{x\,y}(a)e^{\lambda h_i^*(y)} \right] \end{aligned}$$

where the second inequality follows from the inclusion in (4.7.8). This display and (4.7.12) together imply that

$$e^{\lambda g_i^*+\lambda h_i^*(x)} = \inf_{a\in B^*(x)} \left[e^{\lambda C(x,a)} \sum_{y\in S_i^*} p_{x\,y}(a)e^{\lambda h_i^*(y)} \right];$$

since $i\in\{1,2,\ldots,k\}$ and $x\in S_i^*$ are arbitrary, (4.7.9) follows.

In short, it has been verified that $\mathcal{O}^*$ in (4.7.1) is an optimality system for $\mathcal{M}$, establishing the conclusion of Theorem 4.3.2. □

Chapter 5

Conclusion and Open Problems

In this final part of the exposition, the results presented in the previous chapters are briefly discussed, emphasizing the main contribution of this thesis. Next, the essential tool used to derive the conclusions in Chapter 4, namely, the existence of a solution to the discounted dynamic programming equation, is briefly discussed in the context of Markov decision chains with denumerable sate space, and two open problems concerning extensions of the main conclusions of the thesis to models with infinite state space are posed.

5.1. A Retrospective View

In this work Markov decision chains evolving on a finite state space have been studied. Starting with a brief discussion of the idea of risk-aversion, the notion of risk-premium of a random cost Y was defined as the excess with respect to the expected value $E[Y]$ that the controller is willing to pay in order to avoid the uncertain cost Y. Following Pratt (1964), a single number measuring the controller's aversion to risk when Y takes values around a point $y \in \mathbb{R}$ was determined, and the analysis showed that

(i) Twice the risk-premium is proportional to the variance of the random cost, where

(ii) The proportionality constant is given by the quotient of the second and first derivatives of the underlying utility function evaluated a y; such a quotient is a natural measure of the *risk-sensitivity* of the decision maker when facing a random cost taking values around y.

Under the basic condition that the risk-sensitivity coefficient of the controller is a positive constant λ , the utility function is exponential and the certain equivalent of the random cost Y is easily determined by (1.3.9), an expression that was used to specify the risk-sensitive average performance criterion in (1.4.1) and (1.4.2). After this point, the main problem of the thesis was sated as follows:

> *To establish a characterization of the optimal risk-sensitive average cost function in such a way that an optimal stationar policy can be obtaied.*

The already available results on this problem, concerning models satisfying strong communication conditions ensuring that the optimal average cost function is a constant and characterized by a single optimality equation, were studied in Chapters 2, and the analysis in that context showed the central role of the communication assumption in the algebraic approach by Howard and Matheson (1972), and in the probabilistic method in Cavazos-Cadena and Fernández-Gaucherand (2002). Next, in Chapter 3 it was shown that when the system is not fully communicating, even a constant optimal average cost function is not, in general, characterized by a single optimality equation and, under the simultaneous Doeblin condition, under which a state z is accessible regardless of the starting point of the system, it was proved that the existence of a solution to the optimality equation is guaranteed only if the risk-sensitivity coefficient is small enough.

The discussion in Chapters 3 and 4 these two chapters provided a strong motivation to pursue the main goal of this work, and the main contributions were established in Chapter 4 as Theorems 4.3.1 and 4.3.2, were, regardless of the communication structure of the model, under mild continuty compactness conditions, it was proved that the optimal risk-sensitive average cost function is characterized by a system of nested equations, and that a solution of such a system renders an optimal stationary policy; moreover, the optimal value function is the same if the superior or inferior limit are used in the specification of the average performance index. .

The basic tool used to prove the existence of an optimality system—established in Thorem 4.3.2—was the discounted method, and a central assumption behind that result was the finiteness of the state space, so that, when the action sets are compact, any continuous cost function is bounded. Accordingly, the problems described below concern the application of the discounted technique in Markov decision chains with (infinite) *denumerable state space* and possibly *unbounded cost function*. To pave the route to the statement of the problems, the existence of solution to the discounted dynamic programming equation is briefly discussed for model with a general denumerable state space in the following section.

5.2. Discounted Dynamic Programming Equation

Consider a Markov decision chain $\mathcal{M} = (S, A, \{A(x)\}, P, C)$ where the spate space is *denumerable* and, to begin with, also assume that the cost function is bounded, that is, $\|C\| < \infty$. In this context, for each $\alpha \in (0,1)$ the same argument used in Section 4.6 yields that the operator T_α in (4.6.1) is contractive on the space $\mathcal{B}(S)$ of bounded functions defined on the state space S, and then there exists a $V_\alpha \in \mathcal{B}(S)$ satisfying the dynamic programming equation (4.6.3). Assume now that the cost function is just bounded below, that is, $C \geq b$ for some real number b; replacing C by $C - b$, without loss of generality suppose that C is nonnegative, and define $C_r\colon \mathbb{K} \to \mathbb{R}$ be $C_r =\colon C \wedge r$ for each $r \geq 0$. It follows that C_r is bounded, and then, for each $\alpha \in (0,1)$, there exists a unique bounded funtion $V_{\alpha,r}\colon S \to [0,\infty)$ such that

$$e^{\lambda V_{\alpha,r}(x)} = \inf_{a \in A(x)} \left[e^{\lambda C_r(x,a)} \sum_{y \in S} p_{x\,y}(a) e^{\lambda \alpha V_{\alpha,r}(y)} \right], \quad x \in S$$

The monotonicity of the operator T_α immediately yields that $V_{\alpha,r}(\cdot)$ increases with r, so that

$$\lim_{r\to\infty} V_{\alpha,r}(\cdot) =\colon V_\alpha(\cdot)$$

is well defined and, assuming that $V_\alpha(x) < \infty$, it can be seen that the equality in the above display is preserved after the passage to the limit:

$$e^{\lambda V_\alpha(x)} = \inf_{a \in A(x)} \left[e^{\lambda C(x,a)} \sum_{y \in S} p_{x\,y}(a) e^{\lambda \alpha V_\alpha(y)} \right],$$

so that the dynamic programming equation holds for V_α as soon as this function is finite. The fixed point V_α is characterized by the above equation toghether with the following property:

For each $x \in S$, $V_\alpha(x) = \inf W(x)$, where W is bounded below and satisfies

$$e^{\lambda W(x)} \geq \inf_{a \in A(x)} \left[e^{\lambda C(x,a)} \sum_{y \in S} p_{x\,y}(a) e^{\lambda \alpha W(y)} \right], \quad x \in S.$$

The problem posed in the following section involves the functions $\{V_\alpha\}_{\alpha\in(0,1)}$ and the optimal average reward function J^*.

5.3. Approximations in the Case of a Penalized Cost Function

As it was established in the proof of Theorem 4.3.2 , the optimal average reward $J(x)$ satisfies that, for every $x \in S$,

$$J^*(x) = \lim_{k\to\infty} (1-\alpha_k)V_{\alpha_k}(x), \tag{5.3.1}$$

so that, if the state space is finite, the normalized sequence $\{(1-\alpha_k)V_{\alpha_k}\}$ converges to the optimal average reward function J^*. For models with denumerable state space the following result is available.

Suppose that the cost function has a penalized structure in the following sense:

$$\begin{gathered}\text{For each real number } r \text{, the set}\\ \{x\in S \,|\, C(x,a)\le r \text{ for some } a\in A(x)\} \text{ is finite.}\end{gathered} \tag{5.3.2}$$

In this context, given a sequence $\{\alpha_n\}$ increasing to 1, there exists a subsequence $\{\alpha_{n_k}\}$ and a state $z \in S$ such that

(a) $V_{\alpha_{n_k}}(z) = \min_{x\in S} V_{\alpha_{n_k}}(x)$, and

(b) $\lim_{k\to\infty}(1-\alpha_{n_k})V_{\alpha_{n_k}}(x) = J^*(x)$ at every state x such that the sequence $\{V_{\alpha_{n_k}}(x) - V_{\alpha_{n_k}}(z)\}_{k=1,2,3,\ldots}$ is bounded.

This result was firstly obtained in Hernández-Hernández and Marcus (1999) using a game theoretical approach, and an extension to models with Borel state space was given in Jaśkiewicz (2007); a different approach based on Hölder's inequality was presented in Cavazos-Cadena and Salem-Silva (2009). Using he terminology in Section 4.6, the convergence (5.3.1) has been already established at states in the minimal class S_1^*, but not at states in other classes; notice that if the state space is irreducible, the above convergence holds at every state (Borkar and Meyn, 2002).

Problem 1: Let $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, P, C)$ be a Markov decision chain with denumerable sate space, and assume that the cost function has a penalized structure, in the sense that (5.3.2) holds. Is it true that the convergence

$$\lim_{\alpha\nearrow 1}(1-\alpha)V_\alpha(x) = J^*(x)$$

holds at every state $x \in S$?

Of course, the question makes sense for an arbitrary (nonnegative) cost function, but the problem is challenging even within the restricted framework determined by (5.3.2). In the risk-neutral context, the above convergence holds; see, for instance, Sennott (1999).

5.4. General Denumerable Models

Before stating the next problem, it is convenient to point out a consequence of the characterization of the (λ-sensitive) optimal average cost function in terms of an optimality system as in the previous chapter. Given a number $g \in \mathbb{R}$, define

$$S^g := \{x\in S \,|\, J^*(x) > g\}. \tag{5.4.1}$$

Now, by simplicity, consider *an uncontrolled Markov chain over a finite state space*, and notice that the main results in the previous chapter render the following conclusion about the (λ-sensitive) average cost function $J(\cdot)$.

> The average cost $J(\cdot)$ is determined by a system of equations of the form
>
> $$e^{\lambda(g_i+h_i(x))} = \left[e^{\lambda C(x)} \sum_{y\in S_i} p_{x\,y} e^{\lambda h_i(y)}\right], \quad x \in S_i, \quad i = 1, 2, \ldots, k,$$
>
> where $S_1, S_2, \ldots, S_k$ is a partition of the state space, $g_1 \le g_2 \le \cdots \le g_k$ and $p_{x\,y} = 0$ when $x \in S_i$ and $y \in S_j$ with $i < j$. In these circumstances, $J(x) = g_i$ if $x \in S_i$.

As it was shown in Chapter 4, the above displayed relation yields that,

$$J(x) = \lim_{n\to\infty} \frac{1}{n\lambda} \log\left(E_x\left[e^{\lambda \sum_{t=0}^{n-1} C(X_t)} I[X_t \in S_i,\, 0 \le t < n]\right]\right), \quad x \in S_i,$$

whereas the original specification of $J(\cdot)$ establishes that

$$J(x) = \lim_{n\to\infty} \frac{1}{n\lambda} \log\left(E_x\left[e^{\lambda \sum_{t=0}^{n-1} C(X_t)}\right]\right), \quad x \in S.$$

Consider now a nonempty set S^g as in (5.4.1). Let $x \in S^g$ be arbitary and let S_i be the class that contains x; in this case $S_i \subset S^g$ and then

$$1 \ge I[X_t \in S^g,\, 0 \le t < n-1] \ge I[X_t \in S_i\, 0 \le t < n-1].$$

The three last displays together imply that

$$J(x) = \lim_{n\to\infty} \frac{1}{n\lambda} \log\left(E_x\left[e^{\lambda \sum_{t=0}^{n-1} C(X_t)} I[X_t \in S^g,\, 0 \le t < n]\right]\right), \quad x \in S^g,$$

an equality that leads to the following conclusion:

For each state $x \in S^g$, the average cost $J(x)$ depends only on
the costs incurred while the system stays in S^g (5.4.2)

Next, an example will be used to analyze this property for a Markov chain on a denumerable state space.

Example 5.4.1. Consider a Markov chain with state space S specified as follows: S is the union of the set $\mathbb{N}$ of nonnegative integers, and other denumerable disjoint set whose elements are denoted by $\overline{1}, \overline{2}, \overline{3}, \ldots$:

$$S = \{0, 1, 2, 3, \ldots\} \cup \{\overline{1}, \overline{2}, \overline{3}, \ldots\}.$$

The transition law is determined by

$$p_{0\,0} = 1, \quad p_{x\,x-1} = 1, \quad x = 1, 2, 3, \ldots,$$

$$p_{\overline{x},\overline{x+1}} = p = 1 - p_{\overline{x},x}, \quad x = 1, 2, 3, \ldots,$$

where $p \in (0, 1)$ will be specified latter. Finally, let the cost function be given as follows:

$$C(x) = 1, \quad C(\overline{x}) = 0 = C(0), \quad x = 1, 2, 3, \ldots.$$

Considering that the action set is a singleton, these specifications determine a controlled Markov chain. □

In the following proposition the (risk-sensitive) average cost for the above Markov chain will be analyzed.

Proposition 5.4.1. In the context of Example 5.4.1 , select the parameter p as

$$p = e^{-\lambda/2}.$$

In this case, the (λ-sensitive) average cost function $J(\cdot)$ satisfies the following relations:

$$J(x) = 0, \quad x = 0, 1, 2, 3, \ldots,$$

and

$$J(\overline{x}) \geq \frac{1}{4}, \quad x = 1, 2, 3, \ldots.$$

Proof. Suppose that the initial state is a nonnegative integer x. In this case, the system visits the states $x, x_1, x-2, \ldots, 1$ at times $0, 1, 2, 3, \ldots, x-1$, incurring a cost 1 ate each step and, from time x onwards, stays at state 0 incurring a null cost. It follows that

$$J_n(x) = x \wedge (n+1), \quad x, n = 0, 1, 2, 3, \ldots,$$

so that

$$J(x) = \lim_{n\to\infty} \frac{1}{n+1} J_n(x) = 0, \quad x = 0, 1, 2, 3, \ldots.$$

Next, suppose that the initial state is $\overline{x}$, where x is a positive integer. In this case, for each even integer $m = 2k > 0$, the following trajectory of length m has probability $p^{k-1}(1-p)$:

$$\overline{x}, \overline{x+1}, \ldots, \overline{x+k-1}, x+k-1, x+k-2, \ldots, x.$$

Along this trajectory, the total cost incurred is equal to $k = m/2$, and then

$$e^{\lambda J_m(\overline{x})} \geq p^{k-1}(1-p)e^{\lambda k} = \frac{1-p}{p}(pe^{\lambda})^{m/2} = \frac{1-p}{p} e^{m\lambda/4}$$

a relation that leads to $J(\overline{x}) \geq \liminf_{m\to\infty} \frac{1}{m} J_{m-1}(\overline{x}) \geq 1/4$, concluding the argument. □

Now, in the context of Example 5.4.1, consider the set S^0 as in (5.4.1), and notice that

$$S^0 = \{\overline{1}, \overline{2}, \overline{3}, \ldots\}$$

and

$$J(x) \geq 1/4, \quad x \in S^0,$$

by Proposition 5.4.1. On the other hand, the specification of the cost function in Example 5.4.1 prescribes that the cost function is null at each state in S^0, so that the *the costs incurred while the system stays in S^0 are null* and, consequently, the property (5.4.2) fails in the contest of the present example.

Problem 2: Given a Markov decision chain $\mathcal{M} = (S, A, \{A(x)\}_{x\in S}, P, C)$ with denumerable state space, find a characterization of the optimal (risk-sensitive) average cost allowing to obtain an optimal stationary policy.

The above discussion shows that an answer to Problem 2 will *not* be a direct generalization of the results in Chapter 4.

Bibliography

[1]. A. Alanís-Durán and R. Cavazos-Cadena (2012), An optimality system for finite average Markov decision chains under risk-aversion, *Kybernetika*, **48**, 83–104.

[2]. A. Araposthathis, V. K. Borkar, E. Fernández-Gaucherand, M. K. Gosh and S. I. Marcus (1993), Discrete-time controlled Markov processes with average cost criteria: a survey, *SIAM Journal on Control and Optimization*, **31** (1993), 282–334.

[3]. J. O. Berger (2010), Statistical Decision Theory and Bayesian Analysis, 2nd. Edition, *Springer*, New York.

[4]. D. P. Bertsekas (2007), Dynamic Programming and Optimal Control, Vol. I, *Athena Scientific*, Belmont, Massachusetts.

[5]. D. P. Bertsekas (2007a), Dynamic Programming and Optimal Control, Vol. II: Approximate Dynamic Programming, *Athena Scientific*, Belmont, Massachusetts.

[6]. D. P. Bertsekas and S. E. Shreve (1996), Stohastic Optimal Control: The Discrete-Time Case, *Athena Scientific*, Belmont, Massachusetts.

[7]. T. R. Bielecki, D. Hernández-Hernández and S. R. Pliska (1999), Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management, *Mathematical Methods of Operations Research*, **50**, 167–188.

[8]. N. Bäuerle and U. Rieder (2013), More Risk-Sensitive Markov Decision Processes, *Mathematics of Operation Research*, To appear.

[9]. P. Billingsley (1995), Probability and Measure, 3rd. Edition, *Wiley*, New York.

[10]. V. S. Borkar and S. Meyn (2002), Risk-Sensitive optimal control for Markov decision processes with monotone cost, **27**, *Mathematics of Operations Research*, 192–209

[11]. M. Bouakiz and M. J. Sobel (1992), Inventory Control with an Exponential Utility Criterion, *Operations Research*, **40**, . 603–608.

[12]. P. Caravani (1986), On extending linear-quadratic control to non-symmetric risky objective, *Journal of Economic Dynamics and Control*, **10**, 83–88.

[13]. R. Cavazos–Cadena, E. Fernández-Gaucherand (1999), Controlled Markov chains with risk-sensitive criteria: average cost, optimality equations and optimal solutions, *Mathematical Methods of Operations Research*, **43**, 121–139.

[14]. R. Cavazos–Cadena and R. Montes-de-Oca (2000), Optima stationary policies in risk-sensitive dynamic program s with finite state space and nonnegative rewards, *Applicationes Mathematicae*, **27** , 167–185.

[15]. R. Cavazos–Cadena and R. Montes-de-Oca (2000a), Nearly Optimal Policies in Risk-Sensitive Positive Dynamic Programming on Discrete Spaces, *Mathematical Methods of Operations Research*, **52**, 133–167.

[16]. R. Cavazos-Cadena, E. Feinberg and R. Montes-de-Oca (2000) , A Note on the Existence of Optimal Policies in Total Reward Dynamic Programs with Compact Action Sets, *Mathematics of Operations Research.* **25**, 657—666.

[17]. R. Cavazos-Cadena and R. Montes-de-Oca, Optimal Stationary Policies in Risk- Sensitive Dynamic Programs with Finite State Space and Nonnegative

[18]. R. Cavazos–Cadena, E. Fernández–Gaucherand (2002), Risk-sensitive control in communicating average Markov decision chains, In: M. Dror, P. L'Ecuyer and F. Szidarovsky (Eds.): *Modelling Uncertainty: An examination of Stochastic Theory, Methods and Applications*, Kluwer, Boston, pp. 525-544.

[19]. R. Cavazos–Cadena (2003), Solution to the risk-sensitive average cost optimality equation in a class of Markov decision processes with finite state space, *Mathematical Methods of Operations Research*, **57**, 263–285.

[20]. R. Cavazos–Cadena, D. Hernández-Hernández (2006), A characterization of the optimal risk-sensitive average cost in finite controlled Markov chains, *Annals of Applied Probability*, **15**, 175–212.

[21]. R. Cavazos–Cadena, D. Hernández-Hernández (2006), A system of Poisson equations for a non-constant Varadhan functional on a finite state space *Applied Mathematics and Optimization*, **53**, 101–119.

[22]. R. Cavazos–Cadena and F. Salem-Silva (2009), The Discounted Method and Equivalence of Average Criteria for Risk-Sensitive Markov Decision Processes on Borel Spaces, *Applied Mathematics & Optimization*, **61**, 167–190.

[23]. A. Gosavi (2007), A risk-sensitive approach to total productive maintenance *Automatica*, **42**, 1321–1330

[24]. G. B. Di Masi, L. Stettner: Risk-sensitive (1999), control of discrete time Markov processes with infinite horizon, *SIAM Journal on Control and Optimization*, **38**, 61–78.

[25]. G. B. Di Masi, L. Stettner (2000), Infinite horizon risk sensitive control of discrete time Markov processes with small risk, *Systems & Control Letters*, **40**, 15–20.

[26]. G. B. Di Masi, L. Stettner (2006), Remarks on Risk Neutral and Risk Sensitive Portfolio Optimization, in: in: Y. Kabanov, R. Liptser and J.Stoyanov (Eds.), From Stochastic Calculus to Mathematical Finance, *The Shiryaev Festschrift*, Springer, New York.

[27]. G. B. Di Masi, L. Stettner (2007), Infinite horizon risk sensitive control of discrete time Markov processes under minorization property, *SIAM Journal on Control and Optimization*, 46, 231–252.

[28]. W. H. Flemming and W. M. McEneany (1995), Risk-sensitive control on an infinite horizon, *SIAM Journal on Control and Optimization*, **33**, , 1881–1915.

[29]. W. H. Flemming and D. Hernández-Hernández (1997), Risk-sensitive control of finite state machines on an infinite horizon I, *SIAM Journal on Control and Optimization*, **33**, , 1881–1915.

[30]. F. R. Gantmakher (1959), The Theory of Matrices, *Chelsea*, London.

[31]. D. Hernández-Hernández, S. I. Marcus (1996), Risk-sensitive control of Markov processes in countable state space, *Systems and Control Letters*, **29**, 147–155.

[32]. D. Hernández-Hernández, S. I. Marcus (1999), Existence of Risk Sensitive Optimal Stationary Policies for Controlled Markov Processes, *Applied Mathematics and Optimization*, **40**, 273–285.

[33]. O. Hernández-Lerma (1989) Adaptive Markov Control Processes, Springer, New York.

[34]. O. Hernández-Lerma and J. B. Lasserre (1996), Discrete-Time Markov Control Processes: Basic Optimality Criteria, Springer, New York.

[35]. O. Hernández-Lerma and J. B. Lasserre (1999) Further Topics on Discrete-Time Markov Control Processes, Springer, New York.

[36]. A. R. Howard, J. E. Matheson (1972), Risk-sensitive Markov decision processes, *Management Sciences*, **18**, 356–369.

[37]. K. Hinderer (1970), Foundations of Non-stationary Dynamic Programming with Discrete-Time Parameter, Springer, New York.

[38]. D. H. Jacobson (1973), Optimal stochastic linear systems with exponential performance criteria and their relation to stochastic differential games, *IEEE Transactions on Automatic Control*, **18**, 124–131.

[39]. M. R. James, J. S. Baras and R. J. Elliot (1994), Risk-sensitive optimal control and dynamic games for partially observed discrete-time nonlinear systems, *IEEE Transactons on Automatic Control*, **39**, 780-792.

[40]. S. C. Jaquette (1973), Markov decison processes with a new optimality criterion: discrete time, *The Annals of Statistics*, **1**, 496–505.

[41]. S. C. Jaquette (1976), A utility criterion for Markov decision processes, *Management Sciences*, **23**, 43–49.

[42]. A. Jaśkiewicz (2007), Average optimality for risk sensitive control with general state space, *Annals of Applied Probability*, **17**, 654–675.

[43]. Y. Lin (2005), Decision theretic planning under risk-sensitive objectives, Ph. D. dissertation, Georgia Institute of Technology.

[44]. M. Loève (1977), Probability Theory IVol. I, *Springer*, New York.

[45]. S. I . Marcus, E. Fernández-Gaucherand, D. Hernández-Hernández, S. Coralupi and P. Frad (1996), Risk-sensitive Markov Decision Processes, in : it Systems & COntrol in the Twenty-Fisrt Century, Progress in Systems and Control, Birkhäuser. Editors: CI. Byrnes, B.N.Datta, D.S. Gilliam, C. F. Martin, 263-279.

[46]. S. Meyn and R. L. Tweedie (2009), Markov Chains and Stochastic Stability, *Cambridge University Press*, London.

[47]. C. Meyer (2000). Matrix Analysis and Applied Linear Algebra, *SIAM*, Philadelphia.

[48]. O. MIhatsch and R. Neunner (2002), Risk-Sensitive Reinforcement Learning, *Machine Learning*, **49**, 267-290

[49]. E. Nummelin (2004), General Irreducible Markov Chains and Non-Negative Operators, *Cambridge University Press*, London.

[50]. U. G. Rothblum and P. Whittle (1982), Growth optimality for branching Markov decision chains. *Math. Oper. Res.* **7**, 582–601.

[51]. J. W. Pratt (1964), Risk aversion in the small and in the large, *Econometrica*, **32**, 122–136.

[52]. M. L. Puterman (2005), Markov Decision Processes: Discrete Stochastic Dynamic Programming, *Wiley*, New York.

[53]. S. M. Ross (1992) Aplied Probability Models with Optimization Applications, *Dover*, New York.

[54]. H. L. Royden and P. Fitzpatrick (2010), Real Analysis, *Prentice Hall*, New York.

[55]. W. Rudin (1986), Real and Complex Analysis, *McGraw-Hill*, New York.

[56]. T. Runolfsson (1994), The equivalence between infinite horizon control of stochastic systems with exponential-of-integral performance index and stochastic differential games, *IEEE Transactions on Automatic Control*, **39**, 1551–1563.

[57]. L. I. Sennott (1998), Stochastic Dynamic Programming and the Control of Queueing Systems, *Wiley*, New York.

[58]. K. Sladký (1979), Successive approximation methods for dynamic programming models. In: Proc. of the Third Formator Symposium on the Analysis of Large-Scale Systems (J. Beneš and L. Bakule, eds.). Academia, Prague 1979, pp. 171–189.

[59]. K. Sladký (1980), Bounds on discrete dynamic programming recursions I. *Kybernetika*, **16** (1980), 526-547.

[60]. L. Stettner (2004), Risk-Sensitive Portfolio Optimization With Completely and Partially Observed Factors, *IEEE Transactions on Automatic Control*, **49**, 457–464.

[61]. N. L. Stokey and R. E. Lucas (1989), Recursive Methods in Economic Dynamics, *Harvard University Press*, Cambridge, Massachusetts.

[62]. H. C. Tijms (2003), A First Course in Stochastic Models, *Wiley*, New York.

[63]. P. Whittle (1990), Risk-sensitive optimal control, *Wiley*, New York.