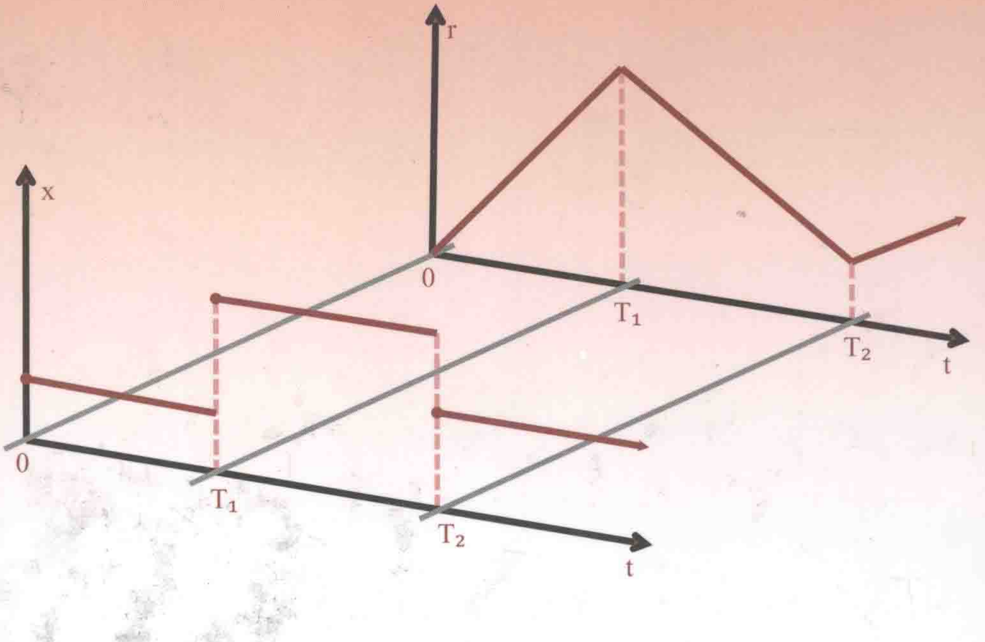
ICP Advanced Texts in Mathematics – Vol. 5

Tomás Prieto-Rumeau
Onésimo Hernández-Lerma

Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games



Imperial College Press

ICP Advanced Texts in Mathematics – Vol. 5

Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games

常州大学图书馆
藏书章

Tomás Prieto-Rumeau
Universidad Nacional de Educación a Distancia, Spain

Onésimo Hernández-Lerma
CINVESTAV-IPN, Mexico

ICP
Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ICP Advanced Texts in Mathematics — Vol. 5
SELECTED TOPICS ON CONTINUOUS-TIME CONTROLLED MARKOV CHAINS AND MARKOV GAMES

Copyright © 2012 by Imperial College Press

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-1-84816-848-0
ISBN-10 1-84816-848-9

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore by Mainland Press Pte Ltd.

Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games

ICP Advanced Texts in Mathematics ISSN 1753-657X

Series Editor: Dennis Barden *(Univ. of Cambridge, UK)*

Published

Vol. 1 Recent Progress in Conformal Geometry
by Abbas Bahri & Yong Zhong Xu

Vol. 2 The Geometry of Curvature Homogeneous Pseudo-Riemannian Manifolds
by Peter B. Gilkey

Vol. 3 Dynamics and Symmetry
by Michael J. Field

Vol. 4 The Solution of the $k(GV)$ Problem
by Peter Schmid

Vol. 5 Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games
by Tomás Prieto-Rumeau & Onésimo Hernández-Lerma

此为试读，需要完整PDF请访问：www.ertongbook.com

To Guadalupe and my parents, Emilio and Michèle.

To Marina, Gerardo, Adrián, Claudia, and Lucas.

Preface

This book concerns continuous-time controlled Markov chains and Markov games. The former, which are also known as continuous-time Markov decision processes, form a class of stochastic control problems in which a single decision-maker wishes to optimize a given objective function. In contrast, in a Markov game there are two or more decision-makers (or players, or controllers) each one trying to optimize his/her own objective function.

The main features of the control and game models studied in the book are that the time variable is *continuous*, the state space is *denumerable*, and the control (or action) sets are *Borel spaces*. Moreover, the transition and reward rates of the dynamical system may be *unbounded*. Controlled Markov chains and Markov games have many important applications in areas such as telecommunication networks, population and epidemic models, engineering, operations research, etc. Some of these applications are illustrated in this book.

We note that most of the material presented here is quite recent: it has been published in the last six years, and it appears in book form for the first time.

One of the main goals of this book is to study the so-called advanced optimality criteria for controlled Markov chains (e.g., bias, overtaking, sensitive discount, and Blackwell optimality), which are refinements of the basic criteria, namely, discounted and average reward optimality. To make this a self-contained book, we also give the main results on the existence of controlled Markov chains and the basic optimality criteria. For the corresponding technical details — some of which have been skipped here — the reader can consult Guo and Hernández-Lerma's *Continuous-Time Markov Decision Processes: Theory and Applications* [52].

A particular emphasis is made regarding the application of the results presented in the book. One of our main concerns is to propose assumptions on the control and game models that are easily verifiable (and verified) in practice. Furthermore, we study an algorithm to solve a certain class of control models, and establish some approximation results that allow us to give precise numerical approximations of the solutions to some problems of practical interest.

Hence, the book has an adequate balance between, on the one hand, theoretical results and, on the other hand, applications and computational issues. It is worth mentioning that the latter were, somehow, missing in the literature on continuous-time controlled Markov chains.

Finally, the topic of zero-sum two-person continuous-time Markov games, for both the basic — discounted and average payoff — and some "advanced" optimality criteria — bias and overtaking equilibria — appears for the first time in book form.

This book is mainly addressed to researchers in the fields of stochastic control and stochastic games. Indeed, it provides an extensive, rigorous, and up-to-date analysis of continuous-time controlled Markov chains and Markov games. It is also addressed to advanced undergraduate and beginning graduate students because the reader is not supposed to have a high mathematical background. In fact, a working knowledge of calculus, linear algebra, probability, and continuous-time Markov chains (at the level of, say, Chapter 4 of R. Durrett's book *Essentials of Stochastic Processes* [31]) should suffice to understand the material herein. As already mentioned, the reader interested in the theoretical foundations of controlled Markov chains can consult [52].

We have carefully written this book, with great dedication and commitment. We apologize, however, for any errors and omissions it might contain.

The authors, April 2011

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

Chapter 1

Introduction

1.1. Preliminary examples

Before giving a formal definition of control and game models, we propose two motivating examples.

1.1.1. *A controlled population system*

We describe next a *controlled population system* inspired by the models in [52, Example 7.2] and [136, Sec. IV]. (In Sec. 9.4 below we will consider a generalization of this controlled population system.) We call it a controlled system because there is a *controller* (also known as a *decision-maker*) who observes a *random dynamical system*, and takes *actions* so as to optimize the system's behavior according to a given *optimality criterion.*

The state space The state variable, denoted by i, is the population size, which takes values in the *state space*

$$S = \{0, 1, 2, \ldots\}.$$

We suppose that the population system is observed *continuously in time*, at times labeled $t \geq 0$. The time horizon may be finite (that is, we observe the state variable on a time horizon $0 \leq t \leq T$, for some finite time $T > 0$) or infinite (which means that the population system is observed at all times $t \geq 0$). We will denote by $x(t) \in S$ the random state of the system at time $t \geq 0$. We will refer to $\{x(t)\}_{t \geq 0}$ as the *state process.*

The sources of variation of the population are described next.

The birth rate The population is subject to a natural birth rate, denoted by $\lambda > 0$. This means that each individual of the population can give birth to a new individual with a transition probability rate which equals λ. More precisely, suppose that the population size is $i \in S$ at time $t \geq 0$, and let $P_i^+[t, t+\delta]$ denote the probability that a new individual is born on the time interval $[t, t+\delta]$. Then we have that

$$\lim_{\delta \downarrow 0} \frac{P_i^+[t, t+\delta]}{\delta} = \lambda i \quad \forall\, t \geq 0. \tag{1.1}$$

Note that, in (1.1), the "individual" birth rate λ is multiplied by the population size i.

The death rate The population is also subject to a natural death rate $\mu > 0$. So, if the population size is i at time $t \geq 0$, and $P_i^-[t, t+\delta]$ denotes the probability that an individual dies on the time interval $[t, t+\delta]$, then we have

$$\lim_{\delta \downarrow 0} \frac{P_i^-[t, t+\delta]}{\delta} = \mu i \quad \forall\, t \geq 0. \tag{1.2}$$

The immigration rate At this point, note that the birth and death rates described above are *not controlled*, meaning that the decision-maker cannot modify them. That is why we called them the *natural* birth and death rates. On the other hand, the immigration rate, defined next, may be controlled by the decision-maker.

Put $A = [a_1, a_2] \subset \mathbb{R}^+$, and let $a \in A$ be the controlled immigration rate. The interpretation is that the decision-maker can encourage or discourage immigration by following suitable immigration policies. Hence, when the decision-maker *chooses* the control $a \in A$, the probability $P_i^a[t, t+\delta]$ that an immigrant individual arrives, on the time interval $[t, t+\delta]$, at the population under study when its size is i at time $t \geq 0$ verifies that

$$\lim_{\delta \downarrow 0} \frac{P_i^a[t, t+\delta]}{\delta} = a \quad \forall\, t \geq 0. \tag{1.3}$$

We assume that the controller takes actions continuously in time. So, let $a(t)$ in A, for $t \geq 0$, denote the controller's action at time t.

The catastrophe rate In addition, we suppose that the population is subject to "catastrophes". The rate $b \in B = [b_1, b_2] \subset \mathbb{R}^+$ at which catastrophes occur is controlled by the decision-maker (for instance, by using adequate medical policies or implementing fire prevention programs, the controller can decrease the catastrophe rate b).

Moreover, the catastrophe is supposed to have a random size. This means that, if a catastrophe occurs when the population size is $i \in S$, then the probability that $1 \leq k \leq i$ individuals die in the catastrophe is $\gamma_i(k)$. We suppose that $\gamma_i(k) > 0$ and that

$$\sum_{1 \leq k \leq i} \gamma_i(k) = 1.$$

Consequently, the transition rate from a state $i > 0$ to a state $0 \leq j < i$ corresponding to a catastrophe under the action $b \in B$ is

$$b \cdot \gamma_i(i-j). \tag{1.4}$$

More explicitly, the rate b corresponds to the catastrophe, and then, *conditional on the catastrophe*, $i-j$ individuals perish with probability $\gamma_i(i-j)$. The new state of the system is thus j.

We denote by $b(t) \in B$, for $t \geq 0$, the action chosen by the controller at time $t \geq 0$.

The action set As seen in the previous paragraphs, the controller chooses his/her actions in the set $A \times B$. We will refer to $A \times B$ as the *action set.*

The transition rate matrix Our previous discussion on the dynamics of the system can be summarized by the transition rate matrix $[q_{ij}(a,b)]_{i,j \in S}$. Here, $q_{ij}(a,b)$ denotes the transition rate from the state $i \in S$ (row) to the state $j \in S$ (column) when the controller chooses the actions $a \in A$ and $b \in B$. This transition rate matrix is

$$\begin{pmatrix}
-a & a & 0 & 0 & 0 & 0 & \cdots \\
\mu + b & -(\mu+\lambda) - a - b & \lambda + a & 0 & 0 & 0 & \cdots \\
b\gamma_2(2) & 2\mu + b\gamma_2(1) & -2(\mu+\lambda) - a - b & 2\lambda + a & 0 & 0 & \cdots \\
b\gamma_3(3) & b\gamma_3(2) & 3\mu + b\gamma_3(1) & -3(\mu+\lambda) - a - b & 3\lambda + a & 0 & \cdots \\
\vdots & \vdots & & & & \vdots & \ddots
\end{pmatrix}.$$

The matrix $[q_{ij}(a,b)]_{i,j \in S}$ above is constructed as follows. The transition rate from state i to $i+1$ is obtained by summing the corresponding transition rates in (1.1) and (1.3). For the transition rate from i to $i-1$ we proceed similarly, and we sum (1.2) and (1.4) for $j = i-1$. Finally, the transition rate from i to j, for $0 \leq j < i-1$, is given by (1.4). The diagonal terms $q_{ii}(a,b)$ are such that the rows of the transition rate matrix sum to zero. This technical requirement comes from the fact that the transition probabilities sum to one (by rows), and so the corresponding derivatives must sum to zero. Further details on this issue are given in Chapter 2.

Policies A *control policy* or simply a *policy* is a “rule” that prescribes the actions chosen by the controller. Typically, a policy is a function of the form

$$\varphi(t,i)=(a,b)\in A\times B,$$

which is given the following interpretation: the controller observes the state of the system $x(t)=i\in S$ at time $t\geq 0$, and then he/she takes the actions $a(t)=a\in A$ and $b(t)=b\in B$. The process $\{x(t),a(t),b(t)\}_{t\geq 0}$, which describes the evolution of the system and the controller's actions, is called the *state-action process.*

The above defined policies are called *Markovian* because they depend only on the current state of the system, say $x(t)$, and the time variable t. In general, however, although we will not consider them in this book, policies can be *history-dependent.* This means that the actions $a(t)$ and $b(t)$ may depend on the *history* $\{x(s),a(s),b(s)\}_{0\leq s\leq t}$ of the state-action process up to time t. More specific classes of policies (such as *randomized* or *stationary* policies) will be introduced in Chapter 2.

The reward rates We suppose that the controller earns rewards (or incurs costs) continuously in time. Typically, the reward rates depend on both the state of the system and the actions.

For this controlled population system, suppose that there is a reward rate function $R(i)$ depending on the population size $i\in S$. Usually, R will be an increasing function of i. In some particular cases, R will be a linear function.

In addition, we assume that there is a cost rate $C_1(i,a)$ associated with the action $a\in A$ when the population size is $i\in S$. (The function C_1 captures, e.g., the cost of the immigration policy, but also the benefits of having a larger working population.) Similarly, the cost rate for controlling the catastrophe rate $b\in B$ when the population size is $i\in S$ is denoted by $C_2(i,b)$.

Hence, if the decision-maker selects the actions $(a(t),b(t))\in A\times B$ when the state of the system is $x(t)\in S$, at time $t\geq 0$, then he/she obtains an infinitesimal net reward

$$(R(x(t))-C_1(x(t),a(t))-C_2(x(t),b(t)))\cdot\delta$$

on the “small” time interval $[t,t+\delta]$.

The optimality criterion The optimality criterion is concerned with the performance evaluation of the policies. As an illustration, if the controller wants to maximize his/her total expected reward on the finite horizon $[0, T]$, then he/she will consider

$$\varphi \mapsto E^{\varphi}\left[\int_0^T [R(x(t)) - C_1(x(t), a(t)) - C_2(x(t), b(t))]dt\right] \tag{1.5}$$

for each policy φ. In (1.5), E^{φ} denotes expectation under the policy φ. Hence, the *finite horizon control problem* consists in finding a policy with the maximal total expected reward (1.5).

Suppose now that there is a depreciation rate $\alpha > 0$ (related to the inflation rate), and that the controller wants to maximize his/her total expected rewards brought to their present value. The *discounted optimality criterion* consists in finding a policy φ that maximizes

$$E^{\varphi}\left[\int_0^{\infty} e^{-\alpha t}[R(x(t)) - C_1(x(t), a(t)) - C_2(x(t), b(t))]dt\right].$$

Furthermore, we can assume that the controller has a given budget, say θ, for the (discounted) expenses on the immigration policy and the catastrophe prevention programs. In this case, the controller has to find the policy that maximizes the expected discounted reward

$$E^{\varphi}\left[\int_0^{\infty} e^{-\alpha t} R(x(t))dt\right]$$

within the class of policies that satisfy the constraint

$$E^{\varphi}\left[\int_0^{\infty} e^{-\alpha t}[C_1(x(t), a(t)) + C_2(x(t), b(t))]dt\right] \leq \theta.$$

This is a *constrained control model* similar to those that we will study in Chapter 8.

Conclusions Finally, we summarize the main elements of the controlled population system described above.

- The state space. (As in the example in this section, the control models studied in this book have denumerable state space.)
- The action set.
- The transition and reward rates.
- The optimality criterion.

1.1.2. *A prey-predator game model*

In the control model described in Sec. 1.1.1, there was a single controller handling the stochastic dynamical system. In a game model, we suppose that there are *several players*. Our next example is a simplified version of the Kolmogorov prey-predator model, which is based on the Lotka–Volterra equation; see, e.g., [19, 39].

The state space We assume that there are two interacting species in a given environment: species 1 is the prey, while species 2 is the predator. The bidimensional state variable (i, j) stands for the total population i and j of the prey and predator species, respectively. So, the state space is

$$S = \{0, 1, 2, \ldots\} \times \{0, 1, 2, \ldots\}.$$

The state process $x(t) = (i(t), j(t))$, for $t \geq 0$, gives the (random) size of the two populations at time $t \geq 0$. As in the control model in Sec. 1.1.1, the state variable is observed continuously at times $t \geq 0$.

The action set The prey species takes actions $a \in A \subset \mathbb{R}$. The action a models the struggle for survival of the prey (for instance, moving to safer areas under the hunting pressure of the predator). The predator species takes actions $b \in B \subset \mathbb{R}$, which model its hunting intensity (e.g., involving more individuals in hunting). Hence, the action set for species 1 is A, while B is the action set for species 2.

Therefore, this model is a *two-player* game in which the prey and predator species compete.

Strategies of the players The players use Markov policies (as defined in Sec. 1.1.1). More precisely, the players observe the state of the system $(i(t), j(t))$, and then they independently choose their actions $a(t) \in A$ and $b(t) \in B$. Hence, the policies of the players are given by functions

$$\phi : [0, \infty) \times S \to A \quad \text{and} \quad \psi : [0, \infty) \times S \to B,$$

for species 1 and 2, respectively, which depend on the state of the system and the time $t \geq 0$.

A usual convention in game models is that the players' policies are referred to as *strategies*, rather than policies.

The transition rates We assume that the two species have natural birth and death rates. Namely, the birth and death rates of the prey species are $\lambda_1 > 0$ and $\mu_1 > 0$, respectively, while the corresponding birth and death rates of the predator species are $\lambda_2 > 0$ and $\mu_2 > 0$.

此为试读，需要完整PDF请访问：www.ertongbook.com