Systems and Control E_3

G. George Yin
Hanqin Zhang
Qing Zhang

Applications of Two-time-scale Markovian Systems

(双时间尺度的马尔可夫系统的应用)

Science Press
Beijing

Applications of Two-time-scale Markovian Systems

（双时间尺度的马尔可夫系统的应用）

G. George Yin

Hanqin Zhang

Qing Zhang

Science Press

Beijing

内 容 简 介

本书主要包含两部分。第一部分是马尔可夫系统的渐近性质。首先回顾了已有的关于双时间尺度、有限状态马尔可夫系统的结果，然后集中讨论了在双时间尺度框架下，具有可数状态空间的马尔可夫系统和切换扩散系统的渐近性质。同时考虑了具有广泛应用背景、其生成算子也依赖于系统状态本身的此两类系统的渐近性质。书中第二部分集中讨论了这两类系统在随机制造业、排队网络、金融工程、保险与风险管理等领域的应用。 为了便于阅读，书中每个章节相对独立。本书致力于对以随机制造业、排队网络、金融工程、保险与风险管理、过程控制的 Wonham 滤波等不同领域的实际应用为背景、具有双时间尺度这一共同特性、大规模复杂随机系统的优化与控制问题的理论研究。希望本书对马尔可夫系统的建模、分析、优化、仿真和控制有一定的参考价值。

本书可作为应用数学、应用概率和运筹与控制领域专家、学者和研究生的参考书。

图书在版编目(CIP)数据

双时间尺度的马尔可夫系统的应用= Applications of Two-time-scale Markovian Systems: 英文/(美)殷刚等著. —北京：科学出版社, 2013
(系统与控制丛书)

ISBN 978-7-03-037349-6

Ⅰ. 双… Ⅱ. 殷… Ⅲ. 马尔可夫过程-研究-英文 Ⅳ. O211.62

中国版本图书馆 CIP 数据核字 (2013) 第 081552 号

责任编辑：姚庆爽 / 责任校对：邹慧卿
责任印制：张 倩 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新科印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2013 年 5 月第 一 版 开本：B5(720 × 1000)
2013 年 5 月第一次印刷 印张：14
字数：343 000

定价：80.00 元

(如有印装质量问题，我社负责调换)

To Meimei for her continuing support

George Yin

To my mentor Guang-Hui Hsu

Hanqin Zhang

To my family for their understanding and support

Qing Zhang

Editorial Board

G. George Yin
Department of Mathematics
Wayne State University
Detroit, MI 48202, U.S.A.
gyin@math.wayne.edu

Qing Zhang
Department of Mathematics
University of Georgia
Athens, GA 30602, U.S.A.
qingz@math.uga.edu

Hanqin Zhang
Academy of Mathematics and
Systems Science, Academia Sinica
Beijing, 100190, China
hanqin@amt.cn.ac
and
NUS Business School
National University of Singapore
Singapore, 117592
bizzhq@nus.edu.sg

Responsible Editor: Qingshuang Yao

Published by Science Press
16 Donghuangchenggen North Street
Beijing 100717, P. R. China

Printed in Beijing

ISBN 978-7-03-037349-6

Preface

Many real-life systems are of large scale and complex. An effective way of reducing computational complexity is to formulate the underlying problems by noting the inherited hierarchical structure. In the early 1960s, nearly decomposable system models came into being. Mathematically, such an approach leads to the formulation of two-time-scale systems. In fact, the idea may be traced back to much earlier work. In the turn of the twentieth century, L. Prandtl published his seminal paper "On fluid motion with small friction," which was one of the essential building blocks of the boundary layer methodology, and set up the foundation of the singular perturbation theory. Loosely, the idea of the singular perturbation theory is: One may take advantage of the structure of the underlying systems to obtain reduced or limit systems. The original systems may have a large number of states, but the limit systems contain fewer states. Thus, using the two-time-scale formulation enables us to substantially reduce the computational effort. To rigorously justify these, we need to put the analysis in a solid foundation.

Devoted to large scale and complex systems, this book encompasses such applications as insurance and risk management, financial engineering, queueing networks, and the Wonham filtering among others. With seemingly diverse range of applications, the different problems are closely connected through the central theme of two-time-scale formulation. We hope that the book can serve as a user guide for modeling, analysis, optimization, and computation for a wide variety of Markovian systems. Emphasizing on applications, this book is an outgrowth of a draft of a survey paper. Our original aims were to take an updated account on applications of two-time-scale Markovian systems in production planning, queueing networks, financial engineering, and related fields since late 1990s, with emphasis on the use of the associated probability distributions and transition probabilities obtained from forward equations of Markov chains. Nevertheless, soon enough, we realized that the manuscript became too long to fit into a single paper. In addition, our continued work on the subject leads to further applications. Not only can two-time-scale

Markov chains be treated, but also singularly perturbed diffusions and singularly perturbed switching jump diffusions can be dealt with in many applications. It appears to be a good idea to document the progress in a book-type archive so that the results can be easily retrieved.

The book contains two parts. The first part is an account on the asymptotic properties of Markovian systems. After reviewing previous results on two-time-scale Markov chains with a finite state space, systems having countable state spaces are treated. In addition, switching diffusion limits with continuous-state-dependent generators are treated, which largely extends the applicability of the results. The second part of the book is devoted to a number of applications arising from manufacturing, queueing systems, financial engineering, insurance risk management. Hopefully, these results will be of interests to many people working in the related areas and an even wider range of applications. In addition, similar to the format of lecture notes, we made effort such that each chapter is fairly independent of the others. Thus, each chapter can be read independently without referring much to other chapters. The book can be adopted as a text for a graduate level special topic course.

We began our work on two-time-scale Markovian systems in the mid-1990s. Our book [212] provides a comprehensive treatment of two-time-scale Markov chains with finite state spaces, whereas this book focuses on applications that were obtained after the publication of the aforementioned book and that were not contained in the updated and revised edition [213] either. In addition, countable state space cases and switching diffusion limits are also treated in the current book. We hope that this book will serve as a modest spur to induce others to come forward with valuable contributions.

This book project could not have been completed without the help and encouragement of many people. Since the early 1990s, we have been working with Rafail Khasminskii on a number of research projects. From him, we have learned much about diffusions and singular perturbations. Our thanks also go to our colleagues, Grazyna Badowski, Qi He, Zhuo Jin, Ruihua Liu, Xuerong Mao, John Moore, Dung Tien Nguyen, Son Luu Nguyen, Suresh Sethi, Qingshuo Song, Jianwu Wang, Fubao Xi, Hailiang Yang, Chenggui Yuan, Xunyu Zhou, and Chao Zhu, who have worked with us on a number of related projects. During the years, our work has been supported by (not at the same time) the National Science Foundation, the Naval Research Office, the Air Force Office of Scientific Research, the Army Research Office, and the National Natural Science Foundation of China. Their support and encouragement are gratefully acknowledged. We thank the book series editor Lei Guo for his consideration, encouragement, and support. Our thanks also go to Xiangping Yang and Qingshuang Yao of Science Press for their support and assistance

to finalize the book.

Detroit — George Yin
Beijing and Singapore — Hanqin Zhang
Athens — Qing Zhang

Contents

1

Introduction

Numerous applications in physical sciences, biological sciences, social sciences, and engineering are based on decision making for systems involving two-time scales. In the past two decades, there have been increasing demands for modeling, analysis, and optimization of such systems with uncertainty. Resurgent research efforts have been devoted to studying the behavior of random dynamic systems. With the motivation from the existing and emerging applications, blending stochastic averaging and singular perturbation methods together, many new results have been obtained. The demands and rapid progress necessitate the dissemination of these results in a collective work, which provides an easy access for researchers from different fields. In response to the needs, we put together this book to substantially update the recent progress on applications in insurance risk management, financial engineering, queueing networks, and the Wonham filtering of two-time-scale Markovian systems. It presents a review of important mathematical tools and demonstrates their utilities.

1.1 Two-time-scale Markovian Systems

Markov chains embrace a wide variety of applications in modeling complex systems such as population dynamics, queueing networks, and manufacturing systems. In recent years, applications of Markovian models have emerged from wireless communications, queueing systems, internet traffic modeling, and financial engineering. These models are general enough to capture the system uncertainty and are mathematically trackable on the other hand. Most dynamic systems in the real world are inevitably large and complex, due to their interactions with numerous subsystems. For the aforementioned applications, the rapid progress in technology and the increasing complexity in modeling have made the control and optimization tasks more challenging.

Take, for instance, the design of a manufacturing system, in which the rate of production depends on current inventory, the future demand, as well as marketing expenditure etc. One uses Markovian models to describe the uncertain machine capacity, to delineate the random demand, and to model random environment and

other stochastic factors. Taking into consideration of the various scenarios results in the system models becoming much more complex. The complex model is welcomed on one hand owing to its comprehensive nature, but it poses substantial difficulty for designing optimal controls. A direct consequence is that the amount computation to reach the optimality is often infeasible. Rather than using a brute force approach to tackle the problem, a more appropriate way to handle the systems is needed. Here two-time-scale method comes in a natural way. For example, a production floor level manager needs to pay attention to day-to-day fluctuation of production capacity, while at the production planning level, only aggregative information at the floor level is required for longer term decision making. There is an inherent hierarchical structure. If one can effectively use this structure, the amount of computation needed in the optimal control design could be substantially reduced. Intuitively, by suitable aggregation, one treats a much "smaller" or "reduced" system with a lot less number of states and variables to deal with. Like this example, two-time-scale formulations are also widely adopted in the queueing networks, as well as large-scale optimization tasks.

In these examples, incorporating all the important factors into the models often results in the state space of the underlying Markov chain being fairly large. As a consequence, the system is too complex to handle and the exact solutions are difficult to obtain. To overcome the difficulty, we look for approximations instead, in which a two-time scale may be introduced to highlight the different rates of variations. In fact, like the examples mentioned, different elements in a large system frequently evolve at different rates. Some of them vary rapidly and others change slowly. The dynamic system evolves as if different components used different clocks or time scales. We should keep in mind that "fast" vs. "slow" and "long time" vs. "short time" are all relative terms, and that time-scale separation is frequently inherent in the underlying problems. For instance, equity investors in a stock market can be classified as long-term investors and short-term investors. The long-term investors consider a relatively longtime horizon and make decisions based on weekly or monthly performance of the stock, whereas short-term investors focus on returns in short terms, namely, daily or an even shorter period. Their time scales are in sharp contrast presenting an example of inherently different time scales. The two-time scale approach relies on decomposing the states of the Markov chains into several recurrent classes or possibly several recurrent classes plus a group of transient states. The essence is that within each recurrent class the interactions are strong and among different recurrent classes the interactions are weak.

Since exact or closed-form solutions to large systems are difficult to obtain, one is often contented with approximate solutions. Because of the precise mathematical models being difficult to establish, a near-optimal control becomes a viable,

and sometimes, the only alternative. Such near optimality requires much less computational effort and often results in more robust policy to attenuate unwanted disturbances. One of the central themes of this book is to present approximate solutions for large-scale Markovian systems using their multiple-time-scale nature. Before treating these problems, results on asymptotic properties of two-time-scale Markov chains including limits of probability distribution vectors, transition probability matrices, and suitably scaled occupation measures will be presented. To integrate analytic and probabilistic methods enables us to have a comprehensive understanding of the structures of the Markovian models, leading to a systematic treatment for systems involving time-scale separation.

1.2 Literature Review

For a Markov chain with a finite but large state space, a decomposition approach is often attractive. For example, to treat large-scale time-inhomogeneous Markovian queueing systems, one of the commonly used methods is decomposition, which consists of breaking the underlying network into smaller pieces (e.g., one station in each piece); see Bitran and Tirupati [23], Reiman [171], and Whitt [191], among others. The decomposition divides the large state space of the Markov process, which completely characterizes the queueing system, into a number of subspaces. Frequently, the transitions within each subspace are much more intensive and frequent than that of among different subspaces.

Ideally, one would like to divide the underlying problem into subproblems that can be solved completely independently; we can then paste together the solutions of the subproblems to obtain the solution to the entire problem. If, for example, the transition matrix of the Markov chain is decomposable into several subtransition matrices (in a diagonal block form), the problem can be solved easily by the aforementioned decomposition methods. Unfortunately, the real world is not ideal. Rather than complete decomposability, one frequently encounters systems that are not completely decomposable but only close to completely decomposable. In the early 1960s, Ando and Fisher proposed the so-called nearly completely decomposable matrix models; see Simon and Ando [178]. Such a notion has subsequently been applied to queueing networks for evaluating certain performance measures such as a queue length (see Courtois [45]) and to economics for reduction of complexity of large-scale systems (see Simon and Ando [178]).

Recent advances in the study of large-scale systems, for example, in production planning, have posed new challenges and provided opportunities for an in-depth understanding of two-time-scale or singularly perturbed Markov chains; see Delebecque and Quadrat [51], Pan and Başar [160], Pervozvanskii and Gaitsgory [166],

Phillips and Kokotovic [168], Sethi and Zhang [174], Sethi, Zhang, and Zhang [175], and Yin and Zhang [212], among others. As alluded to previously, for real-world problems, one often faces large-scale systems with uncertainty. Using the idea of hierarchical decomposition and aggregation to deal with a Markovian system enables us to treat a much simpler system with less complexity; see Sethi and Zhang [174] and Sethi, Zhang, and Zhang [175] for stochastic manufacturing systems, and see also Avramovic, Chow, Kokotovic, Peponides, and Winkelman [13] and Chow, Winkelman, Pai and Sauer [44] for applications to power systems. From a modeling point of view, this amounts to setting up the problems involving different time scales to facilitate the use of singular perturbation methodology for solving the problem. Here, singular perturbation is interpreted in a broader sense, including both deterministic perturbation methods and stochastic averaging. For general references on singular perturbation methods, we refer the reader to Bogoliubov and Mitropolskii [29], Kevorkian and Cole [105], O'Malley [156], Vasil'eava and Butuzov [185], Wasow [187], and references therein.

To gain basic understanding of such systems, it is important to understand the structural properties of the Markov processes. Two-time-scale or singularly perturbed continuous-time Markov chains were treated in Khasminskii, Yin, and Zhang [114, 115], Massey and Whitt [147], Pan and Başar [160], Phillips and Kokotovic [168], and Yin and Zhang [212]. Two-time-scale discrete-time Markov processes were investigated by Avrachenkov, Filar, and Haviv [12], Tse, Gallager and Tsitsiklis [183], and Yin and Zhang [214, 215]. The two-time-scale Markov processes were used in various applications in telecommunications, Markov decision processes, control and optimization problems; see Abbad, Filar, and Bielecki [2], Blankenship [26], Hoppensteadt and Miranker [88], and Naidu [153]. Under a somewhat different setup, averaging of switching and diffusion approximations were analyzed in Anisimov [9].

From the stochastic aspect, two-time-scale stochastic systems have been studied by a host of researchers throughout the years. Khasminskii [106] established a stochastic version of the averaging principle, and brought forward the notion of fast and slow processes in [107]; see also related references in Skorohod [176]. Kushner [127, 128] treated two-time-scale systems in the form of a pair of diffusions as well as wideband noise disturbances. He treated control, optimization, and filtering problems, and introduced the notion of near optimality. Recently, Kabanov and Pergamenshchikov [102] considered asymptotic analysis and control of two-time-scale systems. Korolyuk and Limnios [120] treated systems with semi-Markov processes. Using analytic methods to tackle probabilistic problems was considered by Papanicolaou [163]. Friedlin and Wentzel [73] examined large deviations of stochastic systems from a random perturbation perspective.

1.3 Why Do We Need This Book?

As alluded to in the previous section, the underlying problems often require a combined approach of stochastic analysis and singular perturbation theory. There are many excellent expository articles and textbooks devoted to singular perturbation theory; there are as many excellent treaties for stochastic systems.

Why do we need this book? First, rapid progress in science and technology with applications in various fields has increasingly demanded stochastic modeling and characterization. Taking into account of new applications often requires dealing with hybrid systems involving both continuous dynamics and discrete events. Although stochastic control has become a mature field, optimal controls of hybrid randomly switching processes, switching diffusions, and applications are still somewhat different from the well-known results in the literature. It will be beneficial to bring out the salient features.

Being an effective machinery to handle reduction of complexity, because complex systems consist of many layers, components, and subsystems, to bring out the inherent hierarchical structure, two-time-scale methods provide an ideal approach leading to much reduction of computational effort. To replace the original systems by an averaged system enables us to ignore the meticulous variables, and concentrate on the more important characteristics of the systems. The two-time-scale hybrid Markovian systems will complement the results on two-time diffusions, and on two-time deterministic dynamic systems. Collecting these results is an effort to put the available results in a toolbox to be convenient to use. Our book (Yin and Zhang [212]) provides a comprehensive treatment of two-time-scale Markov chains with finite state spaces, whereas this book collects results on applications that were obtained after the publication of the aforementioned book and that were not contained in the updated version [213] either. In addition, countable state space cases and switching diffusion limits are also treated in the current book.

The needs in applications require the investigation of two-time-scale Markov systems. The applications on the other hand stimulated the advances in the theoretical study of such systems. Two-time-scale formulations have been considered by engineers, management scientists, financial practitioner, and applied mathematicians from very different angles. Although they have enjoyed many applications, new challenges arise in dealing with hybrid systems involving Markovian driving processes. Treating such hybrid systems, a combined approach of stochastic analysis and analytical perturbation methods have been proven to be an effective treatment. It will be useful to summarize some of the recent developments.

Furthermore, the results of two-time Markovian systems are somewhat scattered in the literature. It will be helpful to gather the applications of two-time Markov