

Alexey Piunovskiy  
Yi Zhang *Editors*

# Modern Trends in Controlled Stochastic Processes

Theory and Applications, V.III

# **Emergence, Complexity and Computation**

## **Volume 41**

### **Series Editors**

Ivan Zelinka, Technical University of Ostrava, Ostrava, Czech Republic

Andrew Adamatzky, University of the West of England, Bristol, UK

Guanrong Chen, City University of Hong Kong, Hong Kong, China

### **Editorial Board Members**

Ajith Abraham, MirLabs, USA

Ana Lucia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Juan C. Burguillo, University of Vigo, Spain

Sergej Čelikovský, Academy of Sciences of the Czech Republic, Czech Republic

Mohammed Chadli, University of Jules Verne, France

Emilio Corchado, University of Salamanca, Spain

Donald Davendra, Technical University of Ostrava, Czech Republic

Andrew Ilachinski, Center for Naval Analyses, USA

Jouni Lampinen, University of Vaasa, Finland

Martin Middendorf, University of Leipzig, Germany

Edward Ott, University of Maryland, USA

Linqiang Pan, Huazhong University of Science and Technology, Wuhan, China

Gheorghe Păun, Romanian Academy, Bucharest, Romania

Hendrik Richter, HTWK Leipzig University of Applied Sciences, Germany

Juan A. Rodríguez-Aguilar, IIIA-CSIC, Spain

Otto Rössler, Institute of Physical and Theoretical Chemistry, Tübingen, Germany

Vaclav Snasel, Technical University of Ostrava, Czech Republic

Ivo Vondrák, Technical University of Ostrava, Czech Republic

Hector Zenil, Karolinska Institute, Sweden

The Emergence, Complexity and Computation (ECC) series publishes new developments, advancements and selected topics in the fields of complexity, computation and emergence. The series focuses on all aspects of reality-based computation approaches from an interdisciplinary point of view especially from applied sciences, biology, physics, or chemistry. It presents new ideas and interdisciplinary insight on the mutual intersection of subareas of computation, complexity and emergence and its impact and limits to any computing based on physical limits (thermodynamic and quantum limits, Bremermann's limit, Seth Lloyd limits...) as well as algorithmic limits (Gödel's proof and its impact on calculation, algorithmic complexity, the Chaitin's Omega number and Kolmogorov complexity, non-traditional calculations like Turing machine process and its consequences,...) and limitations arising in artificial intelligence. The topics are (but not limited to) membrane computing, DNA computing, immune computing, quantum computing, swarm computing, analogic computing, chaos computing and computing on the edge of chaos, computational aspects of dynamics of complex systems (systems with self-organization, multiagent systems, cellular automata, artificial life,...), emergence of complex systems and its computational aspects, and agent based computation. The main aim of this series is to discuss the above mentioned topics from an interdisciplinary point of view and present new ideas coming from mutual intersection of classical as well as modern methods of computation. Within the scope of the series are monographs, lecture notes, selected contributions from specialized conferences and workshops, special contribution from international experts.

Indexed by zbMATH.

More information about this series at <http://www.palgrave.com/gp/series/10624>

Alexey Piunovskiy · Yi Zhang  
Editors

# Modern Trends in Controlled Stochastic Processes

Theory and Applications, V.III



Springer

*Editors*

Alexey Piunovskiy  
Department of Mathematical Sciences  
University of Liverpool  
Liverpool, UK

Yi Zhang  
Department of Mathematical Sciences  
University of Liverpool  
Liverpool, UK

ISSN 2194-7287

ISSN 2194-7295 (electronic)

Emergence, Complexity and Computation

ISBN 978-3-030-76927-7

ISBN 978-3-030-76928-4 (eBook)

<https://doi.org/10.1007/978-3-030-76928-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

This book contains extended versions of selected reports presented at the traditional Liverpool workshop on controlled stochastic processes in July 2021. These are independent research papers on Markov decision processes, optimal stopping problems, stochastic games, reinforcement learning, optimization algorithms, system control theory, queueing networks, scheduling, etc. Along with new theoretical results and open problems, many chapters contain case studies and applications to real-life problems. This book can be useful for active researchers in the aforementioned fields and also to practitioners interested in applying mathematical methods to the problems arising in finance, economics, queueing systems, telecommunication, and so on.

# Introduction

Alexey B. Piunovskiy<sup>✉</sup> and Yi Zhang

University of Liverpool, Department of Mathematical Sciences,  
Liverpool L69 7ZL, UK  
piunov@liv.ac.uk, yi.zhang@liverpool.ac.uk

The traditional workshop in Liverpool was initially scheduled for the summer 2020. Because of the COVID-19 pandemic, it was postponed till July 2021. Like in 2010 and 2015, we expect that world-class and active experts will be able to meet in Liverpool or at least to participate in a series of Zoom meetings to discuss interesting and challenging problems of stochastic optimal control. This book contains several extended reports from the mentioned forthcoming workshop. We hope, it will enable researchers, academics, and research students to get a sense of novel and interesting results, concepts, models, methods, and applications of controlled stochastic processes. Below, we briefly describe the topics touched in the further chapters. Roughly speaking, chapters [3–6, 8, 10–12, 15, 18, 19] are mainly theoretical, although include a lot of meaningful examples. Chapters [1, 2, 7, 9, 13, 14, 20] are more problem-oriented and contain case studies.

**Models and Methods.** Classical discrete-time Markov decision processes (MDPs) are considered in [3, 4, 6, 8, 9, 12, 15]; continuous-time Markov, semi-Markov, and more general processes are considered in [2, 5, 10, 11, 19]. Chapters [4, 14, 18] are about various types of stochastic games, including the game against the nature [4]. Let us underline that many authors investigate the models with partial information [3–5, 9, 12, 18, 19] which are deservedly considered to be more challenging.

As for the methods, dynamic programming is useful on many occasions [3, 4, 6, 8–10, 15]. When some probabilities (e.g., describing the dynamics of the process) are not precisely known, the Bayesian approach [9, 12, 14],  $Q$ -learning [3, 4], optimal filtering [19], robust control [1, 4, 12], and  $H_2$  control [5] can be useful. Let us also mention variational inequalities [11] and self-organizing algorithms [7]. Many authors suggested new effective numerical methods for tackling optimal control problems [3, 4, 6, 7, 10, 12, 13], especially arising from real-life case studies. Results of essential computer calculations and simulations are presented in [1, 3–5, 7, 9, 10, 13, 14, 18, 20].

Compared with the workshops in 2010 and 2015 [16, 17], we decided to give more attention to applications of the optimal control theory to real-life problems. As a result, the following case studies and meaningful examples are presented:

- regulation of the adaptive immune response [1];
- efficiency of allocating the same job(s) to several servers in queueing systems (survey) [2];
- forest management [3];
- control of moving objects [3,7];
- control of water resources [4];
- control of an unmanned aircraft subject to actuator faults [5];
- optimal economic growth [6];
- screening program for women breast cancer [9];
- portfolio optimization [10];
- scheduling theory [13];
- optimization of the strategies of a defender and an attacker (terrorist) in a generalized Blotto game [14];
- optimization of advertising efforts [18];
- Jackson networks [19];
- optimization of the targeted drug delivery system [20].

**Acknowledgements.** All the authors are thankful to the Engineering and Physical Sciences Research Council (EPSRC, UK, grant EP/T018216/1) and to the Research Centre in Mathematics and Modelling (RCMM, Uni. of Liverpool) for the financial support of the workshop “Modern Trends in Controlled Stochastic Processes: Theory and Applications” to be held at the Dept. of Mathematical Sciences of the University of Liverpool in July 2021.

## References

1. Almudevar, A.: A regulatory principle for robust reciprocal-time decay of the adaptive immune response. In: Piunovskiy, A., Zhang, Y. (eds.) Modern Trends in Controlled Stochastic Processes. ECC, vol. 41, pp. 298–312. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_15](https://doi.org/10.1007/978-3-030-76928-4_15)
2. Anton, E., Ayesta, U., Jonckheere, M., Verloop, I.M.: A survey of stability results for redundancy systems. In: Piunovskiy, A., Zhang, Y. (eds.) Modern Trends in Controlled Stochastic Processes. ECC, vol. 41, pp. 266–283. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_13](https://doi.org/10.1007/978-3-030-76928-4_13)
3. Avrachenkov, K.E., Borkar, V.S., Dolhare, H.P., Patil, K.: Full gradient DQN reinforcement learning: a provably convergent scheme. In: Piunovskiy, A., Zhang, Y. (eds.) Modern Trends in Controlled Stochastic Processes. ECC, vol. 41, pp. 192–220. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_10](https://doi.org/10.1007/978-3-030-76928-4_10)
4. Bäuerle, N., Glauner, A.: Q-learning for distributionally robust Markov decision processes. In: Piunovskiy, A., Zhang, Y. (eds.) Modern Trends in Controlled Stochastic Processes. ECC, vol. 41, pp. 108–128. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_6](https://doi.org/10.1007/978-3-030-76928-4_6)
5. de Oliveira, A.M., Costa, O.L.V.: Control of continuous-time Markov jump linear systems with partial information. In: Piunovskiy, A., Zhang, Y. (eds.) Modern Trends in Controlled Stochastic Processes. ECC, vol. 41, pp. 87–107. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_5](https://doi.org/10.1007/978-3-030-76928-4_5)
6. Deng, F., Guo, X., Zhang, Y.: On finite approximations to Markov decision processes with recursive and nonlinear discounting. In: Piunovskiy, A., Zhang, Y. (eds.) Modern Trends in Controlled Stochastic Processes. ECC, vol. 41, pp. 221–247. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_11](https://doi.org/10.1007/978-3-030-76928-4_11)

7. Diep, Q.B., Truong, T.C., Zelinka, I.: Swarm intelligence and swarm robotics in the path planning problem. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 313–327. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_16](https://doi.org/10.1007/978-3-030-76928-4_16)
8. Feinberg, E.A., Kasyanov, P.O., Zgurovsky, M.Z.: Average cost Markov decision processes with semi-uniform Feller transition probabilities. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 1–18. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_1](https://doi.org/10.1007/978-3-030-76928-4_1)
9. Horiguchi, M.: On an approach to evaluation of health care programme by Markov decision model. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 341–354. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_18](https://doi.org/10.1007/978-3-030-76928-4_18)
10. Huo, H., Wen, X.: First passage exponential optimality problem for semi-Markov decision processes. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 19–37. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_2](https://doi.org/10.1007/978-3-030-76928-4_2)
11. Jasso-Fuentes, H., Menaldi, J.-L., Vásquez-Rojas, F.: Optimal stopping problems for a family of continuous-time Markov processes. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 57–86. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_4](https://doi.org/10.1007/978-3-030-76928-4_4)
12. Kara, A.D., Yüksel, S.: Robustness to approximations and model learning in MDPs and POMDPs. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 166–191. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_9](https://doi.org/10.1007/978-3-030-76928-4_9)
13. Lipets, V., Zadorojnyi, A.: IBM crew pairing and rostering optimization (C-PRO) technology with MDP for optimization flow orchestration. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 284–297. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_14](https://doi.org/10.1007/978-3-030-76928-4_14)
14. Liu, L., Sonin, I.M.: Locks, bombs and testing: the case of independent locks. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 248–265. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_12](https://doi.org/10.1007/978-3-030-76928-4_12)
15. Piunovskiy, A.B.: Controlled random walk: conjecture and counter-example. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 38–56. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_3](https://doi.org/10.1007/978-3-030-76928-4_3)
16. Piunovskiy, A.B. (ed.): *Modern Trends in Controlled Stochastic Processes: Theory and Applications*. Luniver Press, Frome (2010)
17. Piunovskiy, A.B. (ed.): *Modern Trends in Controlled Stochastic Processes: Theory and Applications*, V.II. Luniver Press, Frome (2015)
18. Robles-Aguilar, A.D., González-Sánchez, D., Minjárez-Sosa, J.A.: Estimation of equilibria in an advertising game with unknown distribution of the response to advertising efforts. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 148–165. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_8](https://doi.org/10.1007/978-3-030-76928-4_8)
19. Semenikhin, K.V.: State estimation in partially observed stochastic networks with queueing applications. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 129–147. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_7](https://doi.org/10.1007/978-3-030-76928-4_7)
20. Tsompanas, M.-A., Bull, L., Adamatzky, A., Balazs, I.: Utilizing differential evolution into optimizing targeted cancer treatments. In: Piunovskiy, A., Zhang, Y. (eds.) *Modern Trends in Controlled Stochastic Processes*. ECC, vol. 41, pp. 328–340. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-76928-4\\_17](https://doi.org/10.1007/978-3-030-76928-4_17)

# Contents

<b>Average Cost Markov Decision Processes with Semi-Uniform Feller Transition Probabilities . . . . .</b>	1
Eugene A. Feinberg, Pavlo O. Kasyanov, and Michael Z. Zgurovsky	
<b>First Passage Exponential Optimality Problem for Semi-Markov Decision Processes . . . . .</b>	19
Haifeng Huo and Xian Wen	
<b>Controlled Random Walk: Conjecture and Counter-Example . . . . .</b>	38
Alexey B. Piunovskiy	
<b>Optimal Stopping Problems for a Family of Continuous-Time Markov Processes . . . . .</b>	57
Héctor Jasso-Fuentes, Jose-Luis Menaldi, and Fidel Vásquez-Rojas	
<b>Control of Continuous-Time Markov Jump Linear Systems with Partial Information . . . . .</b>	87
André Marcorin de Oliveira and Oswaldo Luiz do Valle Costa	
<b>Q-Learning for Distributionally Robust Markov Decision Processes . . . . .</b>	108
Nicole Bäuerle and Alexander Glauner	
<b>State Estimation in Partially Observed Stochastic Networks with Queueing Applications . . . . .</b>	129
Konstantin V. Semenikhin	
<b>Estimation of Equilibria in an Advertising Game with Unknown Distribution of the Response to Advertising Efforts . . . . .</b>	148
Alan D. Robles-Aguilar, David González-Sánchez, and J. Adolfo Minjárez-Sosa	
<b>Robustness to Approximations and Model Learning in MDPs and POMDPs . . . . .</b>	166
Ali Devran Kara and Serdar Yüksel	

<b>Full Gradient DQN Reinforcement Learning: A Provably Convergent Scheme . . . . .</b>	192
Konstantin E. Avrachenkov, Vivek S. Borkar, Hars P. Dolhare, and Kishor Patil	
<b>On Finite Approximations to Markov Decision Processes with Recursive and Nonlinear Discounting . . . . .</b>	221
Fan Deng, Xin Guo, and Yi Zhang	
<b>Locks, Bombs and Testing: The Case of Independent Locks . . . . .</b>	248
Li Liu and Isaac M. Sonin	
<b>A Survey of Stability Results for Redundancy Systems . . . . .</b>	266
Elene Anton, Urtzi Ayesta, Matthieu Jonckheere, and Ina Maria Verloop	
<b>IBM Crew Pairing and Rostering Optimization (C-PRO) Technology with MDP for Optimization Flow Orchestration . . . . .</b>	284
Vladimir Lipets and Alexander Zadorojniy	
<b>A Regulatory Principle for Robust Reciprocal-Time Decay of the Adaptive Immune Response . . . . .</b>	298
Anthony Almudevar	
<b>Swarm Intelligence and Swarm Robotics in the Path Planning Problem . . . . .</b>	313
Quoc Bao Diep, Thanh Cong Truong, and Ivan Zelinka	
<b>Utilizing Differential Evolution into Optimizing Targeted Cancer Treatments . . . . .</b>	328
Michail-Antisthenis Tsompanas, Larry Bull, Andrew Adamatzky, and Igor Balaz	
<b>On an Approach to Evaluation of Health Care Programme by Markov Decision Model . . . . .</b>	341
Masayuki Horiguchi	
<b>Author Index . . . . .</b>	355



# Average Cost Markov Decision Processes with Semi-Uniform Feller Transition Probabilities

Eugene A. Feinberg<sup>1</sup>(✉), Pavlo O. Kasyanov<sup>2</sup>, and Michael Z. Zgurovsky<sup>2</sup>

<sup>1</sup> Department of Applied Mathematics and Statistics, Stony Brook University,  
Stony Brook, NY 11794-3600, USA  
[eugene.feinberg@sunysb.edu](mailto:eugene.feinberg@sunysb.edu)

<sup>2</sup> Institute for Applied System Analysis, National Technical University  
of Ukraine “Kyiv Polytechnic Institute”, Kyiv, Ukraine  
[kasyanov@i.ua](mailto:kasyanov@i.ua), [mzz@kpi.ua](mailto:mzz@kpi.ua)  
<http://www.ams.sunysb.edu/~feinberg/>

**Abstract.** This paper studies average-cost Markov decision processes with semi-uniform Feller transition probabilities. This class of MDPs was recently introduced by the authors to study MDPs with incomplete information. This paper studies the validity of optimality inequalities, the existence of optimal policies, and the approximations of optimal policies by policies optimizing total discounted costs.

**Keywords:** MDP · Average-cost · Semi-uniform Feller transition probabilities

**AMS(2020) subject classification:** Primary 90C40 · Secondary 90C39

## 1 Introduction

This paper establishes the validity of the optimality inequality and the existence of stationary optimal policies for Markov Decision Processes (MDPs) with semi-uniform Feller transition probabilities. It also investigates approximations of optimal policies by policies minimizing discounted costs when the discount factor tends to 1. This class of MDPs with semi-uniform Feller transition probabilities was introduced in [12] because significant classes of problems with incomplete information can be reduced to belief MDPs with semi-uniform Feller transition probabilities.

The paper deals with MDPs with possibly unbounded cost functions and noncompact action sets. Such problems were studied in [11] for MDPs with weakly continuous transition probabilities and in [6, 17] for MDPs with setwise continuous transition probabilities. For MDPs with compact action sets, the models with weakly and setwise continuous probabilities were studied in [21].

## 2 Model Description

For a metric space  $\mathbb{S} = (\mathbb{S}, \rho_{\mathbb{S}})$ , where  $\rho_{\mathbb{S}}$  is a metric, let  $\tau(\mathbb{S})$  be the topology of  $\mathbb{S}$  (the family of all open subsets of  $\mathbb{S}$ ), and let  $\mathcal{B}(\mathbb{S})$  be its Borel  $\sigma$ -field, that is, the  $\sigma$ -field generated by all open subsets of the metric space  $\mathbb{S}$ . For  $s \in \mathbb{S}$  and  $\delta > 0$  denote by  $B_{\delta}(s)$  and  $\bar{B}_{\delta}(s)$  respectively the open and closed balls in the metric space  $\mathbb{S}$  of radius  $\delta$  with center  $s$  and by  $S_{\delta}(s)$  the sphere in  $\mathbb{S}$  of radius  $\delta$  with center  $s$ . Note that  $S_{\delta}(s) = \bar{B}_{\delta}(s) \setminus B_{\delta}(s)$ . For a subset  $S$  of  $\mathbb{S}$  let  $\bar{S}$  denote the *closure* of  $S$  and  $S^o$  the *interior* of  $S$ . Then  $S^o \subset S \subset \bar{S}$ .  $S^o$  is open and  $\bar{S}$  is closed.  $\partial S := \bar{S} \setminus S^o$  denotes the *boundary* of  $S$ . In particular,  $\partial B_{\delta}(s) = S_{\delta}(s)$ . We denote by  $\mathbb{P}(\mathbb{S})$  the *set of probability measures* on  $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$ . A sequence of probability measures  $\{\mu^{(n)}\}_{n=1,2,\dots}$  from  $\mathbb{P}(\mathbb{S})$  converges weakly to  $\mu \in \mathbb{P}(\mathbb{S})$  if for any bounded continuous function  $f$  on  $\mathbb{S}$

$$\int_{\mathbb{S}} f(s) \mu^{(n)}(ds) \rightarrow \int_{\mathbb{S}} f(s) \mu(ds) \quad \text{as } n \rightarrow \infty.$$

A sequence of probability measures  $\{\mu^{(n)}\}_{n=1,2,\dots}$  from  $\mathbb{P}(\mathbb{S})$  converges in total variation to  $\mu \in \mathbb{P}(\mathbb{S})$  if

$$\sup_{C \in \mathcal{B}(\mathbb{S})} |\mu^{(n)}(C) - \mu(C)| \rightarrow 0 \text{ as } n \rightarrow \infty; \quad (1)$$

see [3, 10, 13] for properties of these types of convergence of probability measures. Note that  $\mathbb{P}(\mathbb{S})$  is a separable metric space with respect to the topology of weak convergence for probability measures, when  $\mathbb{S}$  is a separable metric space; [20, Chapter II]. Moreover, according to [4, Theorem 8.3.2], if the metric space  $\mathbb{S}$  is separable, then the topology of weak convergence of probability measures on  $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$  coincides with the topology generated by the *Kantorovich-Rubinshtein metric*

$$\begin{aligned} \rho_{\mathbb{P}(\mathbb{S})}(\mu, \nu) := \\ \sup \left\{ \int_{\mathbb{S}} f(s) \mu(ds) - \int_{\mathbb{S}} f(s) \nu(ds) \mid f \in \text{Lip}_1(\mathbb{S}), \sup_{s \in \mathbb{S}} |f(s)| \leq 1 \right\}, \end{aligned} \quad (2)$$

$\mu, \nu \in \mathbb{P}(\mathbb{S})$ , where

$$\text{Lip}_1(\mathbb{S}) := \{f : \mathbb{S} \mapsto \mathbb{R}, |f(s_1) - f(s_2)| \leq \rho_{\mathbb{S}}(s_1, s_2), \forall s_1, s_2 \in \mathbb{S}\}.$$

For a Borel subset  $S$  of a metric space  $(\mathbb{S}, \rho_{\mathbb{S}})$ , where  $\rho_{\mathbb{S}}$  is a metric, we always consider the metric space  $(S, \rho_S)$ , where  $\rho_S := \rho_{\mathbb{S}}|_{S \times S}$ . A subset  $B$  of  $S$  is called open (closed) in  $S$  if  $B$  is open (closed respectively) in  $(S, \rho)$ . Of course, if  $S = \mathbb{S}$ , we omit “in  $\mathbb{S}$ ”. Observe that, in general, an open (closed) set in  $S$  may not be open (closed respectively). For  $S \in \mathcal{B}(\mathbb{S})$  we denote by  $\mathcal{B}(S)$  the Borel  $\sigma$ -field on  $(S, \rho_S)$ . Observe that  $\mathcal{B}(S) = \{S \cap B : B \in \mathcal{B}(\mathbb{S})\}$ .

For metric spaces  $\mathbb{S}_1$  and  $\mathbb{S}_2$ , a (Borel-measurable) *stochastic kernel*  $\Psi(ds_1|s_2)$  on  $\mathbb{S}_1$  given  $\mathbb{S}_2$  is a mapping  $\Psi(\cdot | \cdot) : \mathcal{B}(\mathbb{S}_1) \times \mathbb{S}_2 \mapsto [0, 1]$ , such that  $\Psi(\cdot | s_2)$  is

a probability measure on  $\mathbb{S}_1$  for any  $s_2 \in \mathbb{S}_2$ , and  $\Psi(B|\cdot)$  is a Borel-measurable function on  $\mathbb{S}_2$  for any Borel set  $B \in \mathcal{B}(\mathbb{S}_1)$ . A stochastic kernel  $\Psi(ds_1|s_2)$  on  $\mathbb{S}_1$  given  $\mathbb{S}_2$  defines a Borel measurable mapping  $s_2 \mapsto \Psi(\cdot|s_2)$  of  $\mathbb{S}_2$  to the metric space  $\mathbb{P}(\mathbb{S}_1)$  endowed with the topology of weak convergence. A stochastic kernel  $\Psi(ds_1|s_2)$  on  $\mathbb{S}_1$  given  $\mathbb{S}_2$  is called *weakly continuous (continuous in total variation)*, if  $\Psi(\cdot|s^{(n)})$  converges weakly (in total variation) to  $\Psi(\cdot|s)$  whenever  $s^{(n)}$  converges to  $s$  in  $\mathbb{S}_2$ . For one-point sets  $\{s_1\} \subset \mathbb{S}_1$ , we sometimes write  $\Psi(s_1|s_2)$  instead of  $\Psi(\{s_1\}|s_2)$ . Sometimes a weakly continuous stochastic kernel is called Feller, and a stochastic kernel continuous in total variation is called uniformly Feller [19].

Let  $\mathbb{S}_1, \mathbb{S}_2$ , and  $\mathbb{S}_3$  be Borel subsets of Polish spaces (a Polish space is a complete separable metric space), and  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$  be a stochastic kernel. For each  $A \in \mathcal{B}(\mathbb{S}_1)$ ,  $B \in \mathcal{B}(\mathbb{S}_2)$ , and  $s_3 \in \mathbb{S}_3$ , let:

$$\Psi(A, B|s_3) := \Psi(A \times B|s_3). \quad (3)$$

In particular, we consider *marginal* stochastic kernels  $\Psi(\mathbb{S}_1, \cdot|\cdot)$  on  $\mathbb{S}_2$  given  $\mathbb{S}_3$  and  $\Psi(\cdot, \mathbb{S}_2|\cdot)$  on  $\mathbb{S}_1$  given  $\mathbb{S}_3$ .

In this paper we consider a discrete-time *Markov decision process*, which is specified by a tuple  $(\mathbb{X}, \mathbb{A}, P, c)$ , where

- (i) the *state space*  $\mathbb{X}$  equals to  $\mathbb{X}_W \times \mathbb{X}_Y$ , where  $\mathbb{X}_W$  and  $\mathbb{X}_Y$  are Borel subsets of Polish spaces;
- (ii)  $\mathbb{A}$  is the *action space*, which is assumed to be a Borel subset of a Polish space;
- (iii)  $P$  is a stochastic kernel on  $\mathbb{X}_W \times \mathbb{X}_Y$  given  $\mathbb{X}_W \times \mathbb{X}_Y \times \mathbb{A}$ , which determines the distribution of the new state  $P(\cdot|w, y, a)$  on  $\mathbb{X}_W \times \mathbb{X}_Y$ , if  $(w, y) \in \mathbb{X}_W \times \mathbb{X}_Y$  is the current state and  $a \in \mathbb{A}$  is the current action, and it is assumed that the stochastic kernel  $P$  on  $\mathbb{X}$  given  $\mathbb{X}_W \times \mathbb{X}_Y \times \mathbb{A}$  is weakly continuous in  $(w, y, a) \in \mathbb{X}_W \times \mathbb{X}_Y \times \mathbb{A}$ ;
- (iv)  $x_0 = (w_0, y_0)$  is the *initial state*;
- (v)  $c : \mathbb{X}_W \times \mathbb{X}_Y \times \mathbb{A} \mapsto \mathbb{R}_+ = [0, +\infty]$  is a *one-step cost function*.

The Markov decision process *evolves* as follows. At time  $t = 0$ , the initial state  $x_0 = (w_0, y_0)$  is given. At each time epoch  $t = 0, 1, \dots$ , if the state of the system is  $(w_t, y_t) \in \mathbb{X}_W \times \mathbb{X}_Y$  and the decision-maker chooses an action  $a_t \in \mathbb{A}$ , then the cost  $c(w_t, y_t, a_t)$  is incurred and the system moves to state  $(w_{t+1}, y_{t+1})$  according to the transition law  $P(\cdot|w_t, y_t, a_t)$ .

Define the *histories*:  $h_0 := (w_0, y_0) \in \mathbb{H}_0$  and  $h_t := (w_0, y_0, a_0, w_1, y_1, a_1, \dots, w_{t-1}, y_{t-1}, a_{t-1}, w_t, y_t) \in \mathbb{H}_t$  for all  $t = 1, 2, \dots$ , where  $\mathbb{H}_0 := \mathbb{X}$  and  $\mathbb{H}_t := \mathbb{H}_{t-1} \times \mathbb{A} \times \mathbb{X}$  if  $t = 1, 2, \dots$ . Then a *policy* is defined as a sequence  $\pi = \{\pi_t\}$  such that, for each  $t = 0, 1, \dots$ ,  $\pi_t$  is a transition kernel on  $\mathbb{A}$  given  $\mathbb{H}_t$ . Moreover,  $\pi$  is called *nonrandomized* if each probability measure  $\pi_t(\cdot|h_t)$  is concentrated at one point. A nonrandomized policy is called *Markov* if all of the decisions depend only on the current state and time. A Markov policy is called *stationary* if all the decisions depend only on the current state. The *set of all policies* is denoted by  $\Pi$ . The Ionescu Tulcea theorem ([2, pp. 140–141] or [18, p.

178]) implies that a policy  $\pi \in \Pi$ , and initial state  $x_0 = (w_0, y_0)$  together with the transition kernel  $P$  determine a unique probability measure  $P_{x_0}^\pi$  on the set of all trajectories  $\mathbb{H}_\infty = (\mathbb{X}_W \times \mathbb{X}_Y \times \mathbb{A})^\infty$  endowed with the product of  $\sigma$ -field defined by Borel  $\sigma$ -fields of  $\mathbb{X}_W$ ,  $\mathbb{X}_Y$ , and  $\mathbb{A}$  respectively. The expectation with respect to this probability measure is denoted by  $\mathbb{E}_{x_0}^\pi = \mathbb{E}_{w_0, y_0}^\pi$ .

Let us specify the performance criterion. For a finite horizon  $T = 0, 1, \dots$ , and for a policy  $\pi \in \Pi$ , let the *expected total discounted costs* be

$$v_{T,\alpha}^\pi(x_0) := \mathbb{E}_{x_0}^\pi \sum_{t=0}^{T-1} \alpha^t c(w_t, y_t, a_t), \quad x_0 \in \mathbb{X}, \quad (4)$$

where  $\alpha \geq 0$  is the discount factor,  $v_{0,\alpha}^\pi(x_0) = 0$ . When  $T = \infty$ , (4) defines an *infinite horizon expected total discounted cost*, and we denote it by  $v_\alpha^\pi(x_0)$ . The *average cost per unit time* is defined as

$$w^\pi(x_0) := \limsup_{T \rightarrow \infty} \frac{1}{T} v_{T,1}^\pi(x_0), \quad x_0 \in \mathbb{X}. \quad (5)$$

For any function  $g^\pi(x_0)$ , including  $g^\pi(x_0) = v_{T,\alpha}^\pi(x_0)$ ,  $g^\pi(x_0) = v_\alpha^\pi(x_0)$ , and  $g^\pi(x_0) = w^\pi(x_0)$  define the *optimal cost*  $g(x_0) := \inf_{\pi \in \Pi} g^\pi(x_0)$ ,  $x_0 \in \mathbb{X}$ . A policy  $\pi$  is called *optimal* for the respective criterion, if  $g^\pi(x_0) = g(x_0)$  for all  $x_0 \in \mathbb{X}$ . For  $g^\pi = v_{t,\alpha}^\pi$ , the optimal policy is called *t-horizon discount-optimal*; for  $g^\pi = v_\alpha^\pi$ , it is called *discount-optimal*; and for  $g^\pi = w^\pi$ , it is called *average-cost optimal*.

It is well known (see, e.g., [2, Proposition 8.2]) that the functions  $v_{t,\alpha}(x)$  recursively satisfy the following *optimality equations* with  $v_{0,\alpha}(x) = 0$  for all  $x \in \mathbb{X}$ ,

$$v_{t+1,\alpha}(x) = \inf_a \left\{ c(x, a) + \alpha \int_{\mathbb{X}} v_{t,\alpha}(z) q(dz|x, a) \right\}, \quad x \in \mathbb{X}, \quad t = 0, 1, \dots. \quad (6)$$

In addition, a Markov policy  $\phi$ , defined at the first  $T$  steps by the mappings  $\phi_0, \dots, \phi_{T-1}$ , that satisfy for all  $t = 1, \dots, T$  the equations

$$v_{t,\alpha}(x) = c(x, \phi_{T-t}(x)) + \alpha \int_{\mathbb{X}} v_{t-1,\alpha}(z) q(dz|x, \phi_{T-t}(x)), \quad x \in \mathbb{X}, \quad (7)$$

is optimal for the horizon  $T$ ; see, e.g., [2, Lemma 8.7].

It is also well known ([2, Propositions 9.8 and 9.12] or [1, 5]) that  $v_\alpha$ , where  $\alpha \in (0, 1]$ , satisfies the following discounted cost optimality equation (DCOE):

$$v_\alpha(x) = \inf_a \left\{ c(x, a) + \alpha \int_{\mathbb{X}} v_\alpha(z) q(dz|x, a) \right\}, \quad x \in \mathbb{X}, \quad (8)$$

and a stationary policy  $\phi_\alpha$  is discount-optimal if and only if

$$v_\alpha(x) = c(x, \phi_\alpha(x)) + \alpha \int_{\mathbb{X}} v_\alpha(z) q(dz|x, \phi_\alpha(x)), \quad x \in \mathbb{X}. \quad (9)$$

### 3 Properties of Semi-Uniform Feller Stochastic Kernels

Let us consider some basic definitions.

**Definition 1.** Let  $\mathbb{S}$  be a metric space. A function  $f : \mathbb{S} \mapsto \mathbb{R}$  is called

- (i) lower semi-continuous (l.s.c.) at a point  $s \in \mathbb{S}$  if  $\liminf_{s' \rightarrow s} f(s') \geq f(s)$ ;
- (ii) upper semi-continuous at  $s \in \mathbb{S}$  if  $-f$  is lower semi-continuous at  $s$ ;
- (iii) continuous at  $s \in \mathbb{S}$  if  $f$  is both lower and upper semi-continuous at  $s$ ;
- (iv) lower/upper semi-continuous (continuous respectively) (on  $\mathbb{S}$ ) if  $f$  is lower/upper semi-continuous (continuous respectively) at each  $s \in \mathbb{S}$ .

For a metric space  $\mathbb{S}$ , let  $\mathbb{F}(\mathbb{S})$ ,  $\mathbb{L}(\mathbb{S})$ , and  $\mathbb{C}(\mathbb{S})$  be the spaces of all real-valued functions, all real-valued lower semi-continuous functions, and all real-valued continuous functions respectively defined on the metric space  $\mathbb{S}$ . The following definitions are taken from [7].

**Definition 2.** A set  $\mathbb{F} \subset \mathbb{F}(\mathbb{S})$  of real-valued functions on a metric space  $\mathbb{S}$  is called

- (i) lower semi-equicontinuous at a point  $s \in \mathbb{S}$  if  $\liminf_{s' \rightarrow s} \inf_{f \in \mathbb{F}} (f(s') - f(s)) \geq 0$ ;
- (ii) upper semi-equicontinuous at a point  $s \in \mathbb{S}$  if the set  $\{-f : f \in \mathbb{F}\}$  is lower semi-equicontinuous at  $s \in \mathbb{S}$ ;
- (iii) equicontinuous at a point  $s \in \mathbb{S}$ , if  $\mathbb{F}$  is both lower and upper semi-equicontinuous at  $s \in \mathbb{S}$ , that is,  $\limsup_{s' \rightarrow s} \inf_{f \in \mathbb{F}} |f(s') - f(s)| = 0$ ;
- (iv) lower/upper semi-equicontinuous (equicontinuous respectively) (on  $\mathbb{S}$ ) if it is lower/upper semi-equicontinuous (equicontinuous respectively) at all  $s \in \mathbb{S}$ ;
- (v) uniformly bounded (on  $\mathbb{S}$ ), if there exists a constant  $M < +\infty$  such that  $|f(s)| \leq M$  for all  $s \in \mathbb{S}$  and for all  $f \in \mathbb{F}$ .

Obviously, if a set  $\mathbb{F} \subset \mathbb{F}(\mathbb{S})$  is lower semi-equicontinuous, then  $\mathbb{F} \subset \mathbb{L}(\mathbb{S})$ . Moreover, if  $\mathbb{F}$  is equicontinuous, then  $\mathbb{F} \subset \mathbb{C}(\mathbb{S})$ .

Let  $\mathbb{S}_1$ ,  $\mathbb{S}_2$ , and  $\mathbb{S}_3$  be Borel subsets of Polish spaces, and  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$  be a stochastic kernel.

**Definition 3.** ([12]) A stochastic kernel  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$  is semi-uniform Feller if, for each sequence  $\{s_3^{(n)}\}_{n=1,2,\dots} \subset \mathbb{S}_3$  that converges to  $s_3$  in  $\mathbb{S}_3$  and for each bounded continuous function  $f$  on  $\mathbb{S}_1$ ,

$$\lim_{n \rightarrow \infty} \sup_{B \in \mathcal{B}(\mathbb{S}_2)} \left| \int_{\mathbb{S}_1} f(s_1) \Psi(ds_1, B | s_3^{(n)}) - \int_{\mathbb{S}_1} f(s_1) \Psi(ds_1, B | s_3) \right| = 0. \quad (10)$$

We recall that the marginal measure  $\Psi(ds_1, B | s_3)$ ,  $s_3 \in \mathbb{S}_3$ , is defined in (3). The term “semi-uniform” is used in Definition 3 because the uniform property holds in (10) only with respect to the first coordinate. If the uniform property holds with respect to both coordinates, then the stochastic kernel  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$  is continuous in total variation. Stochastic kernels continuous in total

variation are sometimes called uniformly Feller [19]. According to Corollary 1, a semi-uniform Feller stochastic kernel is weakly continuous.

By [2, Proposition 7.27], there exists a stochastic kernel  $\Phi$  on  $\mathbb{S}_1$  given  $\mathbb{S}_2 \times \mathbb{S}_3$  such that

$$\Psi(A \times B|s_3) = \int_B \Phi(A|s_2, s_3) \Psi(\mathbb{S}_1, ds_2|s_3), \quad (11)$$

$A \in \mathcal{B}(\mathbb{S}_1)$ ,  $B \in \mathcal{B}(\mathbb{S}_2)$ ,  $s_3 \in \mathbb{S}_3$ . The stochastic kernel  $\Phi(\cdot|s_2, s_3)$  on  $\mathbb{S}_1$  given  $\mathbb{S}_2 \times \mathbb{S}_3$  defines a measurable mapping  $\Phi : \mathbb{S}_2 \times \mathbb{S}_3 \rightarrow \mathbb{P}(\mathbb{S}_1)$ , where  $\Phi(s_2, s_3)(\cdot) = \Phi(\cdot|s_2, s_3)$ . According to [2, Corollary 7.27.1], for each  $s_3 \in \mathbb{S}_3$  the mapping  $\Phi(\cdot, s_3) : \mathbb{S}_2 \rightarrow \mathbb{P}(\mathbb{S}_1)$  is defined  $\Psi(\mathbb{S}_1, \cdot|s_3)$ -almost surely uniquely in  $s_2 \in \mathbb{S}_2$ . Consider the stochastic kernel

$$\phi(D \times B|s_3) := \int_B \mathbf{I}\{\Phi(s_2, s_3) \in D\} \Psi(\mathbb{S}_1, ds_2|s_3), \quad (12)$$

$D \in \mathcal{B}(\mathbb{P}(\mathbb{S}_1))$ ,  $B \in \mathcal{B}(\mathbb{S}_2)$ ,  $s_3 \in \mathbb{S}_3$ . In models for decision making with incomplete information,  $\phi$  is the transition probability between belief states, which are posterior distributions of states. Continuity properties of  $\phi$  play the fundamental role in the studies of models with incomplete information. Theorem 1 characterizes such properties, and this is the reason for the title of this section.

According to [2, Corollary 7.27.1], the particular choice of a stochastic kernel  $\Phi$  satisfying (11) does not effect the definition of  $\phi$  in (12) because for each  $s_3 \in \mathbb{S}_3$  the mapping  $\Phi(\cdot, s_3) : \mathbb{S}_2 \rightarrow \mathbb{P}(\mathbb{S}_1)$  is defined  $\Psi(\mathbb{S}_1, \cdot|s_3)$ -almost surely uniquely in  $s_2 \in \mathbb{S}_2$ .

Consider the following assumption.

**Assumption 1** *There exists a stochastic kernel  $\Phi$  on  $\mathbb{S}_1$  given  $\mathbb{S}_2 \times \mathbb{S}_3$  satisfying (11) such that, if a sequence  $\{s_3^{(n)}\}_{n=1,2,\dots} \subset \mathbb{S}_3$  converges to  $s_3 \in \mathbb{S}_3$  as  $n \rightarrow \infty$ , then there exists a subsequence  $\{s_3^{(n_k)}\}_{k=1,2,\dots} \subset \{s_3^{(n)}\}_{n=1,2,\dots}$  and a measurable subset  $B$  of  $\mathbb{S}_2$  such that  $\Psi(\mathbb{S}_1 \times B|s_3) = 1$  and*

$$\Phi(s_2, s_3^{(n_k)}) \text{ converges weakly to } \Phi(s_2, s_3), \quad \text{for all } s_2 \in B. \quad (13)$$

In other words, the convergence in (13) holds  $\Psi(\mathbb{S}_1, ds_2|s_3)$ -almost surely.

The following theorem provides necessary and sufficient conditions for semi-uniform Fellerness of a stochastic kernel  $\phi$  in terms of the properties of a given stochastic kernel  $\Psi$ . This theorem describes the necessary and sufficient conditions for the semi-uniform Feller property of the belief-MDPs in terms of the conditions on the transition kernel in the initial model for decision making with incomplete information.

**Theorem 1.** ([12, Theorem 5.14]) *For a given stochastic kernel  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$ , let the marginal kernel  $\Psi(\mathbb{S}_1, \cdot|\cdot)$  on  $\mathbb{S}_2$  given  $\mathbb{S}_3$  is continuous in total variation. Then the following conditions are equivalent:*

- (a) *the stochastic kernel  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$  is semi-uniform Feller;*

- (b) Assumption 1 holds;
- (c) if a sequence  $\{s_3^{(n)}\}_{n=1,2,\dots} \subset \mathbb{S}_3$  converges to  $s_3 \in \mathbb{S}_3$  as  $n \rightarrow \infty$ , then

$$\rho_{\mathbb{P}(\mathbb{S}_1)}(\Phi(s_2, s_3^{(n)}), \Phi(s_2, s_3)) \rightarrow 0 \text{ in probability } \Psi(\mathbb{S}_1, ds_2 | s_3), \quad (14)$$

- where  $\rho_{\mathbb{P}(\mathbb{S}_1)}$  is the Kantorovich-Rubinshtein metric defined in (2);
- (d) the stochastic kernel  $\phi$  on  $\mathbb{P}(\mathbb{S}_1) \times \mathbb{S}_2$  given  $\mathbb{S}_3$  is semi-uniform Feller;

and each of these statements implies that the stochastic kernels  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$  and  $\phi$  on  $\mathbb{P}(\mathbb{S}_1) \times \mathbb{S}_2$  given  $\mathbb{S}_3$  are weakly continuous.

**Corollary 1.** ([12, Corollary 5.15]) A semi-uniform Feller stochastic kernel  $\Psi$  on  $\mathbb{S}_1 \times \mathbb{S}_2$  given  $\mathbb{S}_3$  is weakly continuous.

For other properties of semi-uniform Feller stochastic kernels we refer to [12, Section 5].

## 4 Expected Discounted Costs

For a metric space  $\mathbb{U}$ , we denote by  $\mathbb{K}(\mathbb{U})$  the family of all nonempty compact subsets of  $\mathbb{U}$ .

For an  $\overline{\mathbb{R}}$ -valued function  $f$ , defined on a nonempty subset  $U$  of a metric space  $\mathbb{U}$ , consider the level sets

$$\mathcal{D}_f(\lambda; U) = \{y \in U : f(y) \leq \lambda\}, \quad \lambda \in \mathbb{R}. \quad (15)$$

We recall that a function  $f$  is *inf-compact* on  $U$  if all the level sets  $\mathcal{D}_f(\lambda; U)$  are compact.

Let  $\mathbb{S}_1, \mathbb{S}_2$ , and  $\mathbb{S}_3$  be Borel subsets of Polish spaces. Let  $LW(\mathbb{S}_1; \mathbb{S}_2)$  be the class of all nonnegative Borel-measurable functions  $\varphi : \mathbb{S}_1 \times \mathbb{S}_2 \mapsto \overline{\mathbb{R}}$  such that  $s_1 \mapsto \varphi(s_1, s_2)$  is lower semi-continuous on  $\mathbb{S}_1$  for each  $s_2 \in \mathbb{S}_2$ .

**Definition 4.** ([12]) A function  $u : \mathbb{S}_1 \times \mathbb{S}_2 \times \mathbb{S}_3 \mapsto \overline{\mathbb{R}}$  is called measurable  $\mathbb{K}$ -inf-compact if it is Borel-measurable and for each  $s_2 \in \mathbb{S}_2$  the function  $(s_1, s_3) \mapsto u(s_1, s_2; s_3)$  is  $\mathbb{K}$ -inf-compact on  $\mathbb{S}_1 \times \mathbb{S}_3$ , that is, for each  $s_2 \in \mathbb{S}_2$  the function  $(s_1, s_3) \mapsto u(s_1, s_2; s_3)$  is inf-compact on  $K \times \mathbb{S}_3$  for each  $K \in \mathbb{K}(\mathbb{S}_1)$ .

Consider a discrete-time MDP  $(\mathbb{X}, \mathbb{A}, q, c)$  with the state space  $\mathbb{X} = \mathbb{X}_W \times \mathbb{X}_Y$ , an action space  $\mathbb{A}$ , one-step costs  $c$ , and transition probabilities  $q$ . Assume that  $\mathbb{X}_W, \mathbb{X}_Y$ , and  $\mathbb{A}$  are Borel subsets of Polish spaces. For any  $\alpha \geq 0$  and  $u \in LW(\mathbb{X}_W; \mathbb{X}_Y)$ , we consider:

$$\eta_u^\alpha(x, a) = c(x, a) + \alpha \int_{\mathbb{X}} u(\tilde{x}) q(d\tilde{x} | x, a), \quad (x, a) \in \mathbb{X} \times \mathbb{A}. \quad (16)$$

The following assumption is used in this paper to prove the existence of optimal policies.

**Assumption 2** Let the following two conditions hold:

- (i) the function  $c : \mathbb{X} \times \mathbb{A} \mapsto \overline{\mathbb{R}}$  is nonnegative and measurable  $\mathbb{K}$ -inf-compact with  $\mathbb{S}_1 := \mathbb{X}_W$ ,  $\mathbb{S}_2 := \mathbb{X}_Y$ ,  $\mathbb{S}_3 := \mathbb{A}$ , and  $u = c$ ;
- (ii) the stochastic kernel  $q$  on  $\mathbb{X}_W \times \mathbb{X}_Y$  given  $\mathbb{X}_W \times \mathbb{X}_Y \times \mathbb{A}$  is semi-uniform Feller.

The following theorem, which is stronger theorem than Theorem 6.2 in [12], is the main result of this section.

**Theorem 2.** Let Assumption 2 hold. Then

- (i) the functions  $v_\alpha(w, y)$  and  $v_{t,\alpha}(w, y)$ ,  $t = 0, 1, \dots$ , belongs to  $LW(\mathbb{X}_W \times [0, 1]; \mathbb{X}_Y)$ , and  $v_{t,\alpha}(x) \uparrow v_\alpha(x)$  as  $t \rightarrow \infty$  for all  $(x, \alpha) \in \mathbb{X} \times [0, 1]$ ;
- (ii) for each  $x \in \mathbb{X}$  the functions  $\alpha \mapsto v_\alpha(x)$  and  $\alpha \mapsto v_{t,\alpha}(x)$ ,  $t = 0, 1, \dots$ , where  $\alpha \in [0, 1]$ , are nondecreasing, and they are continuous on the interiors of their domains;
- (iii) if  $t = 0, 1, \dots$ ,  $\alpha \in [0, 1]$ , and  $x \in \mathbb{X}$ , then  $v_{t+1,\alpha}(x) = \min_{a \in \mathbb{A}} \eta_{v_{t,\alpha}}^\alpha(x, a)$ , and the nonempty sets  $A_{t,\alpha}(x) := \{a \in \mathbb{A} : v_{t+1,\alpha}(x) = \eta_{v_{t,\alpha}}^\alpha(x, a)\}$  satisfy the properties: (a)  $\text{Gr}(A_{t,\alpha}) \in \mathcal{B}(\mathbb{X} \times [0, 1] \times \mathbb{A})$ , and (b)  $A_{t,\alpha}(x) = \mathbb{A}$ , if  $v_{t+1,\alpha}(x) = +\infty$ , and  $A_{t,\alpha}(x)$  is compact if  $v_{t+1,\alpha}(x) < +\infty$ ;
- (iv) for  $T = 1, 2, \dots$  and  $\alpha \in [0, 1]$ , if for a  $T$ -horizon Markov policy  $(\phi_0, \dots, \phi_{T-1})$  the inclusions  $\phi_{T-1-t}(x) \in A_{t,\alpha}(x)$  hold for all  $x \in \mathbb{X}$  and for all  $t = 0, \dots, T-1$ , then this policy is  $T$ -horizon optimal for the discount factor  $\alpha$ , and, in addition, there exist Markov optimal  $T$ -horizon policies  $(\phi_0^\alpha, \dots, \phi_{T-1}^\alpha)$  for the discount factor  $\alpha$  such  $\phi_t^\alpha(x) : \mathbb{X} \times [0, 1] \mapsto \mathbb{A}$  is Borel measurable for each  $t = 0, \dots, T-1$ ;
- (v) if  $\alpha \in [0, 1]$  and  $x \in \mathbb{X}$ , then  $v_\alpha(x) = \min_{a \in \mathbb{A}} \eta_{v_\alpha}^\alpha(x, a)$ , and the nonempty sets  $A_\alpha(x) := \{a \in \mathbb{A} : v_\alpha(x) = \eta_{v_\alpha}^\alpha(x, a)\}$  satisfy the properties: (a)  $\text{Gr}(A_\alpha) \in \mathcal{B}(\mathbb{X} \times [0, 1] \times \mathbb{A})$ , and (b)  $A_\alpha(x) = \mathbb{A}$ , if  $v_\alpha(x) = +\infty$ , and  $A_\alpha(x)$  is compact if  $v_\alpha(x) < +\infty$ ;
- (vi) for a discount factor  $\alpha \in [0, 1]$ , a stationary policy  $\phi$  is optimal for an infinite-horizon problem with this discount factor if and only if  $\phi(x) \in A_\alpha(x)$  for all  $x \in \mathbb{X}$ , and there exists a Borel measurable mapping  $\phi_\alpha : \mathbb{X} \mapsto \mathbb{A}$ , such that for each  $\alpha \in [0, 1]$  the stationary policy  $\phi_\alpha$  is optimal for the infinite-horizon problem with the discount factor  $\alpha$ .

Before the proof of Theorem 2, we provide Lemma 1, which is useful for establishing continuity properties of the value functions  $v_{t,\alpha}$  and  $v_\alpha(x)$ . The proof of this lemma uses Theorem 2.2 from [6]. For each  $(w, y, \alpha) \mapsto w_\alpha(w, y)$  from  $LW(\mathbb{X}_W \times \mathbb{R}_+; \mathbb{X}_Y)$ , where  $\mathbb{R}_+ := [0, +\infty)$ , we consider the function  $(w, y, \alpha) \mapsto w_\alpha^*(w, y) := \inf_{a \in \mathbb{A}} \eta_{w_\alpha}^\alpha(w, y, a)$  on  $\mathbb{X}_W \times \mathbb{X}_Y \times \mathbb{R}_+$ . We observe that, if for some  $x \in \mathbb{X}$  the function  $\alpha \mapsto w_\alpha(x)$  is nondecreasing, then the interior of its domain is the open interval  $(0, \alpha(x))$ .

**Lemma 1.** Let Assumption 2 hold, and let  $(w, y, \alpha) \mapsto w_\alpha(x)$  be a function from  $LW(\mathbb{X}_W \times \mathbb{R}_+; \mathbb{X}_Y)$  such that for each  $x \in \mathbb{X}$  the function  $\alpha \mapsto w_\alpha(x)$  is nondecreasing, and it is continuous on the interior of its domain. Then:

- (i) the function  $(x, a, \alpha) \mapsto \eta_{w_\alpha}^\alpha(x, a)$  is Borel measurable on  $\mathbb{X} \times \mathbb{A} \times \mathbb{R}_+$ , and for each  $y \in \mathbb{X}_Y$  the function  $(w, \alpha; a) \mapsto \eta_{w_\alpha}^\alpha(w, y, a)$  is  $\mathbb{K}$ -inf-compact on  $(\mathbb{X}_W \times \mathbb{R}_+) \times \mathbb{A}$ ;
- (ii) for each  $(x, a) \in \mathbb{X} \times \mathbb{A}$  the function  $\alpha \mapsto \eta_{w_\alpha}^\alpha(x, a)$  is nondecreasing and continuous in  $\alpha$  on the interior of its domain;
- (iii) the function  $(w, y, \alpha) \mapsto w_\alpha^*(w, y)$  belongs to  $LW(\mathbb{X}_W \times \mathbb{R}_+; \mathbb{X}_Y)$ ;
- (iv) for each  $x \in \mathbb{X}$  the function  $\alpha \mapsto w_\alpha^*(x)$  is nondecreasing and continuous on the interior of its domain;
- (v) there exists a Borel mapping  $(x, \alpha) \mapsto f_\alpha(x)$  of  $\mathbb{X} \times \mathbb{R}_+$  into  $\mathbb{A}$  such that  $f_\alpha(x) \in \mathbb{A}$  and  $w_\alpha^*(x) = \eta_{w_\alpha}^\alpha(x, f_\alpha(x))$  for all  $x \in \mathbb{X}$  and  $\alpha \geq 0$ ;
- (vi) the nonempty sets  $A_\alpha^*(x) = \{a \in \mathbb{A} : w_\alpha^*(x) = \eta_{w_\alpha}^\alpha(x, a)\}$ ,  $(x, \alpha) \in \mathbb{X} \times \mathbb{R}_+$ , satisfy the following properties: (a)  $\text{Gr}(A_\alpha^*) \in \mathcal{B}(\mathbb{X} \times \mathbb{R}_+ \times \mathbb{A})$ ; (b)  $A_\alpha^*(x) = \mathbb{A}$ , if  $w_\alpha^*(x) = +\infty$ , and  $A_\alpha^*(x)$  is compact if  $w_\alpha^*(x) < +\infty$ .

*Proof.* (i). The function  $(x, a, \alpha) \mapsto \eta_{w_\alpha}^\alpha(x, a)$  is nonnegative and nondecreasing in  $\alpha$  because  $(x, a) \mapsto c(x, a)$  and  $(x, \alpha) \mapsto w_\alpha(x)$  are nonnegative and nondecreasing in  $\alpha$ . Borel-measurability and continuity properties of  $(x, \alpha) \mapsto w_\alpha(x)$  and regularity of the transition kernel  $q$  imply that the function  $(x, a, \alpha) \mapsto \int_{\mathbb{X}} w_\alpha(z)q(dz|x, a)$  is Borel measurable on  $\mathbb{X} \times \mathbb{A} \times \mathbb{R}_+$ , which implies that the function  $(x, a, \alpha) \mapsto \eta_{w_\alpha}^\alpha(x, a)$  is Borel measurable on  $\mathbb{X} \times \mathbb{A} \times \mathbb{R}_+$ .

Fix an arbitrary  $y \in \mathbb{X}_Y$  and prove that the function  $(w, \alpha; a) \mapsto \eta_{w_\alpha}^\alpha(w, y, a)$  is  $\mathbb{K}$ -inf-compact on  $(\mathbb{X}_W \times \mathbb{R}_+) \times \mathbb{A}$ . According to Assumption 2(i), the function  $(w, a) \mapsto c(x, y, a)$  is  $\mathbb{K}$ -inf-compact on  $\mathbb{X}_W \times \mathbb{A}$ . If

$$\int_{\mathbb{X}} w_\alpha(\tilde{x})q(d\tilde{x}|w, y, a) \leq \liminf_{n \rightarrow \infty} \int_{\mathbb{X}} w_{\alpha^{(n)}}(\tilde{x})q(d\tilde{x}|w^{(n)}, y, a^{(n)}), \quad (17)$$

for all  $(w, a, \alpha) \in \mathbb{X}_W \times \mathbb{A} \times \mathbb{R}_+$  and  $\{w^{(n)}, a^{(n)}, \alpha^{(n)}\}_{n=1,2,\dots}$  converging to  $(w, a, \alpha)$ , then the function  $(w, \alpha; a) \mapsto \eta_{w_\alpha}^\alpha(w, y, a)$  is  $\mathbb{K}$ -inf-compact on  $(\mathbb{X}_W \times \mathbb{R}_+) \times \mathbb{A}$  since it is a sum of a  $\mathbb{K}$ -inf-compact function and a nonnegative lower semi-continuous function. Let us prove that (17) holds. On the contrary, there exist a sequence  $\{(w^{(n)}, a^{(n)}, \alpha^{(n)})\}_{n=1,2,\dots} \subset \mathbb{X}_W \times \mathbb{A} \times \mathbb{R}_+$  that converges to some  $(w, a, \alpha) \in \mathbb{X}_W \times \mathbb{A} \times \mathbb{R}_+$  and a constant  $\lambda$  such that

$$\int_{\mathbb{X}} w_{\alpha^{(n)}}(\tilde{x})q(d\tilde{x}|w^{(n)}, y, a^{(n)}) \leq \lambda < \int_{\mathbb{X}} w_\alpha(\tilde{x})q(d\tilde{x}|w, y, a), \quad (18)$$

for each  $n = 1, 2, \dots$ . Since the function  $\alpha \mapsto w_\alpha(x)$  is nondecreasing, without loss of generality, assume that  $\alpha^{(n)} \uparrow \alpha$  as  $n \rightarrow \infty$ . According to Theorem 1(a, b) applied to  $\Psi := q$ ,  $\mathbb{S}_1 := \mathbb{X}_W$ ,  $\mathbb{S}_2 := \mathbb{X}_Y$ ,  $\mathbb{S}_3 := \mathbb{X}_W \times \{y\} \times \mathbb{A}$ , there exists a stochastic kernel  $\Phi$  on  $\mathbb{X}_W$  given  $\mathbb{X}_Y \times \mathbb{X}_W \times \{y\} \times \mathbb{A}$  such that (11) and Assumption 1 hold. In particular, (18) implies that

$$\int_{\mathbb{X}_Y} \left[ \int_{\mathbb{X}_W} w_{\alpha^{(n)}}(\tilde{w}, \tilde{y})\Phi(d\tilde{w}|\tilde{y}, w^{(n)}, y, a^{(n)}) \right] q(\mathbb{X}_W, d\tilde{y}|w^{(n)}, y, a^{(n)}) \leq \lambda, \quad (19)$$

for each  $n = 1, 2, \dots$ , and there exist a subsequence  $\{(w^{(n_k)}, a^{(n_k)})\}_{k=1,2,\dots} \subset \{(w^{(n)}, a^{(n)})\}_{n=1,2,\dots}$  and a Borel set  $Y \in \mathcal{B}(\mathbb{X}_Y)$  such that  $q(\mathbb{X}_W \times Y|w, y, a) = 1$

and  $\Phi(\tilde{y}, w^{(n)}, y, a^{(n)})$  converges weakly to  $\Phi(\tilde{y}, w, y, a)$  in  $\mathbb{P}(\mathbb{X}_W)$  as  $k \rightarrow \infty$ , for all  $\tilde{y} \in Y$ . Therefore, since the function  $\tilde{w} \mapsto w_{\alpha^{(p)}}(\tilde{w}, \tilde{y})$  is nonnegative and lower semi-continuous for each  $\tilde{y} \in Y$  and  $p = 1, 2, \dots$ , Fatou's lemma for weakly converging probabilities [14, Theorem 1.1] implies that

$$\begin{aligned} \int_{\mathbb{X}_W} w_{\alpha^{(m)}}(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w, y, a) &\leq \\ \liminf_{k \rightarrow \infty} \int_{\mathbb{X}_W} w_{\alpha^{(m)}}(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w^{(n_k)}, y, a^{(n_k)}) &\leq \\ \liminf_{k \rightarrow \infty} \int_{\mathbb{X}_W} w_{\alpha^{(n_k)}}(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w^{(n_k)}, y, a^{(n_k)}), \end{aligned}$$

for each  $m = 1, 2, \dots$  and  $\tilde{y} \in Y$ , where the second inequality holds, since  $\alpha^{(n_k)} \uparrow \alpha$  as  $k \rightarrow \infty$ , and the function  $\alpha \mapsto w_\alpha(x)$  is nondecreasing. Therefore, the monotone convergence theorem implies

$$\int_{\mathbb{X}_W} w_\alpha(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w, y, a) \leq \liminf_{k \rightarrow \infty} \int_{\mathbb{X}_W} w_{\alpha^{(n_k)}}(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w^{(n_k)}, y, a^{(n_k)}),$$

for each  $\tilde{y} \in Y$ . For a fixed  $N = 1, 2, \dots$  we set

$$\begin{aligned} \varphi_k^N(\tilde{y}) &:= \min\left\{\int_{\mathbb{X}_W} w_{\alpha^{(n_k)}}(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w^{(n_k)}, y, a^{(n_k)}), N\right\}, \\ \varphi^N(\tilde{y}) &:= \min\left\{\int_{\mathbb{X}_W} w_\alpha(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w, y, a), N\right\}, \end{aligned}$$

$\tilde{y} \in Y$ ,  $k = 1, 2, \dots$ . Note that  $\varphi^N(\tilde{y}) \leq \liminf_{k \rightarrow \infty} \varphi_k^N(\tilde{y})$ , for each  $\tilde{y} \in Y$ . Therefore, uniform Fatou's lemma [13, Corollary 2.3] implies that

$$\int_{\mathbb{X}_Y} \varphi^N(\tilde{y}) q(\mathbb{X}_W, d\tilde{y}|w, y, a) \leq \liminf_{k \rightarrow \infty} \int_{\mathbb{X}_Y} \varphi_k^N(\tilde{y}) q(\mathbb{X}_W, d\tilde{y}|w^{(n_k)}, y, a^{(n_k)}) \leq \lambda,$$

for each  $N = 1, 2, \dots$ , where the second inequality follows from (19) since  $\varphi_k^N(\tilde{y}) \leq \int_{\mathbb{X}_W} w_{\alpha^{(n_k)}}(\tilde{w}, \tilde{y}) \Phi(d\tilde{w}|\tilde{y}, w^{(n_k)}, y, a^{(n_k)})$  for each  $\tilde{y} \in Y$ , and  $k = 1, 2, \dots$ . Thus, the monotone convergence theorem implies that

$$\int_{\mathbb{X}} w_\alpha(\tilde{x}) q(d\tilde{x}|w, y, a) = \lim_{N \rightarrow \infty} \int_{\mathbb{X}_Y} \varphi^N(\tilde{y}) q(\mathbb{X}_W, d\tilde{y}|w, y, a) \leq \lambda.$$

This is a contradiction to (18). Therefore, the function  $(w, \alpha; a) \mapsto \eta_{w_\alpha}^\alpha(w, y, a)$  is  $\mathbb{K}$ -inf-compact on  $(\mathbb{X}_W \times \mathbb{R}_+) \times \mathbb{A}$ .

(iii, v, vi). Statement (i) and Berge's theorem for noncompact action sets [9, Theorem 1.2] imply that the function  $(w, \alpha) \mapsto w_\alpha^*(w, y)$  is lower semi-continuous for each  $y \in \mathbb{X}_Y$ . Moreover, [6, Theorem 2.2, and Corollary 2.3 (i)] directly imply that the function  $(w, y, a) \mapsto w_\alpha^*(w, y)$  is Borel measurable and statements (v) hold. Property (vi)(a) follows from Borel measurability of  $(x, a, \alpha) \mapsto \eta_{w_\alpha}^\alpha(x, a)$  on  $\mathbb{X} \times \mathbb{A} \times \mathbb{R}_+$  and  $(x, \alpha) \mapsto w_\alpha^*(x)$  on  $\mathbb{X} \times \mathbb{R}_+$ ; and property (vi)(b) follows from inf-compactness of  $a \mapsto \eta_{w_\alpha}^\alpha(x, a)$  on  $\mathbb{A}$  for each  $(x, \alpha) \in \mathbb{X} \times \mathbb{R}_+$ .

(ii). The function  $\alpha \mapsto \alpha \int_{\mathbb{X}} w_\alpha(z) q(dz|x, a)$  is continuous on the interior of its domain for each  $(x, a) \in \mathbb{X} \times \mathbb{A}$ . This follows from Assumption 2 (ii) and [7, Theorem 6.1] because, according to Corollary 1, the stochastic kernel  $q$  is weakly continuous. So, the function  $\alpha \mapsto \eta_{w_\alpha}^\alpha(x, a)$  is continuous in  $\alpha$  on the interior of its domain.

(iv). Fix an arbitrary  $x \in \mathbb{X}$ . Statement (ii) implies that the function  $\alpha \mapsto w_\alpha^*(x)$  is nondecreasing. The continuity statement is nontrivial only if the interior of the domain of this function is not empty. Let  $(0, \alpha(x))$  be the interior domain of  $\alpha \mapsto w_\alpha^*(x)$ . We shall prove that the function  $w_\alpha^*(x)$  is continuous on  $(0, \alpha(x))$ . Let us fix an arbitrary  $\alpha' \in (0, \alpha(x))$ . We choose an arbitrary  $\beta \in (\alpha', \alpha(x))$ . Then  $w_\beta^*(x) < +\infty$ , and therefore  $\eta_{w_\beta}^\beta(x, a_\beta) < +\infty$  for some  $a_\beta \in \mathbb{A}$ . Then  $\eta_{w_\alpha}^\alpha(x, a_\beta) \leq \eta_{w_\beta}^\beta(x, a_\beta) < +\infty$  for all  $\alpha \in (0, \beta]$ . For each  $a \in \mathbb{A}$  the function  $g(\alpha, a) = \min\{\eta_{w_\alpha}^\alpha(x, a), \eta_{w_\alpha}^\alpha(x, a_\beta)\}$  is continuous in  $\alpha \in (0, \beta]$  as a minimum of two continuous functions, and  $w_\alpha^*(x) = \inf_{a \in \mathbb{A}} g(\alpha, a)$ . Since the infimum of upper semi-continuous functions is an upper semi-continuous function, the function  $\alpha \mapsto w_\alpha^*(x)$  is upper semi-continuous on  $(0, \beta]$ , and therefore it is upper semi-continuous on  $(0, \alpha(x))$ . According to statement (iii), the function  $\alpha \mapsto w_\alpha^*(x)$  is lower semi-continuous on  $\mathbb{R}_+$ . So, statement (iv) holds.  $\square$

*Proof of Theorem 2.* According to (6), the functions  $v_{t,\alpha}(x)$ ,  $t = 0, 1, \dots$ , recursively satisfy the optimality equations  $v_{t+1,\alpha}(x) = \inf_{a \in \mathbb{A}} \eta_{v_{t,\alpha}}^\alpha(x, a)$  with  $v_{0,\alpha}(x) = 0$ , for all  $(x, \alpha) \in \mathbb{X} \times [0, 1]$ . So, Lemma 1 (i, ii) sequentially applied to the functions  $v_{0,\alpha}(x)$ ,  $v_{1,\alpha}(x), \dots$ , imply statements (i, ii) of the theorem. In particular, statement (ii) of the theorem implies that these functions are lower semi-continuous in  $\alpha$  on the interiors of their domains. According to [2, Proposition 9.17],  $v_{t,\alpha}(x) \uparrow v_\alpha(x)$  as  $t \rightarrow +\infty$  for each  $(x, \alpha) \in \mathbb{X} \times [0, 1]$ . Therefore,  $v_\alpha(x) \in LW(\mathbb{X}_W \times [0, 1]; \mathbb{X}_Y)$ , and  $v_\alpha(x)$  is nondecreasing and lower semi-continuous in  $\alpha$  on the interior of its domain. Thus, statement (i) is proved.

In addition, (7) imply that a Markov policy defined at the first  $T$  steps by the mappings  $\phi_0^\alpha, \dots, \phi_{T-1}^\alpha$ , that satisfy for all  $t = 1, \dots, T$  the equations  $v_{t,\alpha}(x) = \eta_{v_{t-1,\alpha}}^\alpha(x, \phi_{T-t}^\alpha(x))$ , for each  $(x, \alpha) \in \mathbb{X} \times [0, 1]$ , is optimal for the horizon  $T$ . According to (8) and (9),  $v_\alpha(x)$  satisfies the discounted cost optimality equation  $v_\alpha(x) = \inf_{a \in \mathbb{A}} \eta_{v_\alpha}^\alpha(x, a)$  for each  $(x, \alpha) \in \mathbb{X} \times [0, 1]$ ; and a stationary policy  $\phi_\alpha(x)$  is discount-optimal if and only if  $v_\alpha(x) = \eta_{v_\alpha}^\alpha(x, \phi_\alpha(x))$  for each  $x \in \mathbb{X}$ . Statements (iii–vi) follow from these facts and Lemma 1 (v, vi).

To complete the proof of statement (ii), we need to show that for each fixed  $x \in \mathbb{X}$  the function  $\alpha \mapsto v_\alpha(x)$  is upper semi-continuous in the interior of its domain. Since  $v_\alpha(x)$  is nondecreasing and lower semi-continuous in  $\alpha$  on the interior of its domain, this means that we need to show that  $v_\alpha(x)$  is right-continuous in  $\alpha \in (0, 1)$  if  $v_\alpha(x) < +\infty$ . Indeed, if  $v_\alpha(x) < +\infty$ , let us consider a stationary optimal stationary policy  $\phi^\alpha$  whose existence is claimed in statement (vi). Then the function  $v_{\alpha+\Delta}^{\phi^\alpha}(x)$  is continuous in  $\Delta$  as a value of a converging power series. Therefore,

$$0 \leq v_{\alpha+\Delta}(x) - v_\alpha(x) = v_{\alpha+\Delta}(x) - v_\alpha^{\phi^\alpha}(x) \leq v_{\alpha+\Delta}^{\phi^\alpha}(x) - v_\alpha^{\phi^\alpha}(x) \downarrow 0$$

as  $\Delta \downarrow 0$ .  $\square$

## 5 Average Costs per Unit Time

Following [21], we assume that  $w^* := \inf_{x \in \mathbb{X}} w(x) < +\infty$ , that is, there exist  $x \in \mathbb{X}$  and  $\pi \in \Pi$  with  $w^\pi(x) < +\infty$ . Otherwise, if this assumption does not hold, then the problem is trivial, because  $w(x) = +\infty$  for all  $x \in \mathbb{X}$  and any policy  $\pi$  is average-cost optimal.

Define the following quantities for  $\alpha \in [0, 1)$ :

$$m_\alpha = \inf_{x \in \mathbb{X}} v_\alpha(x), \quad u_\alpha(x) = v_\alpha(x) - m_\alpha,$$

$$\underline{w} = \liminf_{\alpha \uparrow 1} (1 - \alpha)m_\alpha, \quad \bar{w} = \limsup_{\alpha \uparrow 1} (1 - \alpha)m_\alpha.$$

According to [21, Lemma 1.2],

$$0 \leq \underline{w} \leq \bar{w} \leq w^* < +\infty. \quad (20)$$

In this section we show that Assumption 2 and boundedness assumption Assumption B on the function  $u_\alpha$  introduced in [8], which is weaker than boundedness Assumption B introduced in [21], lead to the validity of stationary average-cost optimal inequalities and the existence of stationary policies. Stronger results hold under Assumption B.

**Assumption B.**  $\liminf_{\alpha \uparrow 1} u_\alpha(x) < +\infty$  for all  $x \in \mathbb{X}$ .

The above is weaker than the following assumption.

**Assumption B.**  $\sup_{\alpha \in [0, 1)} u_\alpha(x) < +\infty$  for all  $x \in \mathbb{X}$ .

In the rest of this paper we assume that Assumption B holds. In view of Theorem 2 (i), if  $v_\alpha(x) = +\infty$  for some  $(x, \alpha) \in \mathbb{X} \times [0, 1)$ , then  $u_\beta(x) = v_\beta(x) = +\infty$  for all  $\beta \in [\alpha, 1)$ , and  $u(x) = +\infty$ , where  $m_\beta$  is finite in view of (20). Thus Assumption B implies that  $v_\alpha(x) < +\infty$ , and therefore  $u_\alpha(x) < +\infty$  for all  $(x, \alpha) \in \mathbb{X} \times [0, 1)$ . Under Assumption 2, in view of (20) and Theorem 2(i,ii),  $m_\alpha : [0, 1) \mapsto \mathbb{R}_+$  is a nondecreasing upper semi-continuous function as an infimum of the family of the continuous functions, and therefore  $u_\alpha(w, y) = v_\alpha(w, y) - m_\alpha \in LW(\mathbb{X}_W \times [0, 1); \mathbb{X}_Y)$ .

Let us define the following nonnegative functions on  $\mathbb{X}_W \times \mathbb{X}_Y$ :

$$\begin{aligned} U_\beta(w, y) &:= \inf_{\alpha \in [\beta, 1)} u_\alpha(w, y), \\ \underset{\sim}{U}_\beta(w, y) &:= \liminf_{w' \rightarrow w} U_\beta(w', y), \\ u(w, y) &:= \sup_{\beta \in [0, 1)} \underset{\sim}{U}_\beta(w, y), \end{aligned} \quad (21)$$

$\beta \in [0, 1)$ ,  $x \in \mathbb{X}$ . To establish the Borel measurable properties for these functions we need to make the following assumption.

**Assumption 3** *The space  $\mathbb{X}_W$  is  $\sigma$ -compact.*

**Lemma 2.** *Let  $\beta \in [0, 1)$ . Under Assumptions 2 and 3, the functions  $U_\beta, \tilde{U}_\beta, u : \mathbb{X} \mapsto \mathbb{R}_+$  defined in (21) are Borel measurable on  $\mathbb{X}$ . Moreover, the functions  $U_\beta(w, y)$  and  $u(w, y)$  are lower semi-continuous in  $w$  for each  $y \in \mathbb{X}_Y$ .*

*Proof.* Fix  $\beta \in [0, 1)$ . Borel measurability of  $(w, y) \mapsto U_\beta(w, y)$  follows from (21) and [6, Theorem 2.1] applied to the Borel spaces  $\mathbb{X}$  and  $[0, 1)$ , set-valued map  $B(x) = [\beta, 1)$  for all  $x \in \mathbb{X}$ , and the function  $u(x, \alpha) := u_\alpha(x) \in LW([0, 1); \mathbb{X})$ . Let us prove the Borel measurability of  $(w, y) \mapsto \tilde{U}_\beta(w, y)$ . Indeed, consider the function

$$u(w', \alpha, w, y, \delta) := u_\alpha(w', y) \chi\{w' \in \bar{B}_\delta(w)\},$$

$w', w \in \mathbb{X}_W, y \in \mathbb{X}_Y, \alpha \in [\beta, 1), \delta > 0$ , where  $\chi\{\text{"True"}\} := 0$ , and  $\chi\{\text{"False"}\} := +\infty$ . Since the nonnegative functions  $(w', \alpha, y) \mapsto u_\alpha(w', y)$  and  $(w', w, \delta) \mapsto \chi\{w' \in \bar{B}_\delta(w)\}$  belong to  $LW(\mathbb{X}_W \times [0, 1); \mathbb{X}_Y)$  and  $LW(\mathbb{X}_W; \mathbb{X}_W \times (0, +\infty))$  respectively, then the function  $u$  belongs to  $LW(\mathbb{X}_W \times [0, 1); \mathbb{X} \times (0, +\infty))$ . Therefore, according to Feinberg and Kasyanov [6, Theorem 2.1] applied to the Borel space  $\mathbb{X} := \mathbb{X} \times (0, +\infty)$ ,  $\sigma$ -compact space  $\mathbb{A} := \mathbb{X}_W \times [0, 1)$ , set-valued map  $B(w, \delta) = \mathbb{X}_W \times [\beta, 1)$  for all  $w \in \mathbb{X}_W$ , and the function  $u \in LW(\mathbb{A}; \mathbb{X})$ , we have that the function

$$(w, y) \mapsto U(w, y) := \inf_{w' \in \bar{B}_\delta(w)} \inf_{\alpha \in [\beta, 1)} u_\alpha(w', y)$$

is Borel measurable. Therefore, the function  $(w, y) \mapsto \tilde{U}_\beta(w, y)$  is Borel measurable because

$$\tilde{U}_\beta(w, y) = \sup_{n=1,2,\dots} \inf_{w' \in \bar{B}_{\frac{1}{n}}(w)} \inf_{\alpha \in [\beta, 1)} u_\alpha(w', y),$$

$w \in \mathbb{X}_W$  and  $y \in \mathbb{X}_Y$ , and a supremum of countable family of Borel measurable functions is Borel measurable. Note that lower semi-continuity of  $\tilde{U}_\beta(w, y)$  in  $w$  directly follows from its definition (21). Therefore, according to (21), the function  $(w, y) \mapsto u(w, y)$  is Borel measurable and it is lower semi-continuous in  $w$  for each  $y \in \mathbb{X}_Y$  as a supremum of countable family of Borel measurable functions  $\{\tilde{U}_{1-\frac{1}{n}}(w, y)\}_{n=1,2,\dots}$  which are lower semi-continuous in  $w$ .  $\square$

In view of the definition of  $u$  in Assumption B,

$$u(w, y) = \lim_{\beta \uparrow \sim} U_\beta(w, y), \quad w \in \mathbb{X}_W, y \in \mathbb{X}_Y. \quad (22)$$

Under Assumptions 2 and 3 the following sets can be defined for  $u$  introduced in (21):

$$\begin{aligned} A^u(x) &:= \{a \in \mathbb{A} : \bar{w} + u(x) \geq \eta_u^1(x, a)\}, \\ A_u(x) &:= \left\{a \in \mathbb{A} : \min_{a^* \in \mathbb{A}} \eta_u^1(x, a^*) = \eta_u^1(x, a)\right\}, \quad x \in \mathbb{X}. \end{aligned}$$

In view of Lemma 1, the sets  $A_u(x)$  are nonempty and compact for all  $x \in \mathbb{X}$ . In the following theorem we show that Assumption 2 and boundedness assumption Assumptions B on the functions  $\{u_\alpha\}_{\alpha \in [0,1]}$  lead to the validity of stationary average-cost optimal inequalities and the existence of stationary policies. [8, Theorems 3 and 4] are respectively counterparts to Theorem 3.3 and the main result in [16] for MDPs with weakly continuous transition probabilities. Assumption B and some additional conditions lead to the validity of optimality equations for average-costs MDPs. In [15] such sufficient conditions are provided for MDPs with weakly continuous transition probabilities and applied to inventory control. More general sufficient conditions for validity of optimality equations are provided in [7, Section 7] for MDPs with weakly and setwise continuous transition probabilities.

**Theorem 3.** *Let Assumptions 2, 3, and B hold. Then for infinite-horizon average costs per unit time there exists a stationary optimal policy  $\phi$  satisfying*

$$\bar{w} + u(x) \geq \eta_u^1(x, \phi(x)), \quad x \in \mathbb{X}, \quad (23)$$

with  $u$  defined in (21), and for this policy

$$w(x) = w^\phi(x) = \limsup_{\alpha \uparrow 1} (1 - \alpha)v_\alpha(x) = \bar{w} = w^*, \quad x \in \mathbb{X}. \quad (24)$$

Moreover, the following statements hold:

- (a) the function  $u : \mathbb{X} \mapsto \mathbb{R}_+$  defined in (21) is Borel measurable;
- (b) the nonempty sets  $A^u(x)$ ,  $x \in \mathbb{X}$ , satisfy the following properties: (b<sub>1</sub>)  $\text{Gr}(A^u) \in \mathcal{B}(\mathbb{X} \times \mathbb{A})$ ; (b<sub>2</sub>) for each  $x \in \mathbb{X}$  the set  $A^u(x)$  is compact;
- (c) if  $\varphi(x) \in A^u(x)$  for all  $x \in \mathbb{X}$  for a stationary policy  $\varphi$ , then  $\varphi$  satisfies (23) and (24), with  $u$  defined in (21) and with  $\phi = \varphi$ , and  $\varphi$  is optimal for average costs per unit time;
- (d) the sets  $A_u(x)$  are compact and  $A_u(x) \subset A^u(x)$  for all  $x \in X$ , and there exists a stationary policy  $\varphi$  with  $\varphi(x) \in A_u(x) \subset A^u(x)$  for all  $x \in \mathbb{X}$ .

The proof of Theorem 3 uses the following statement.

**Lemma 3.** *Under Assumptions 2, 3, and B,*

$$\bar{w} + u(x) \geq \min_{a \in \mathbb{A}} \eta_u^1(x, a), \quad x \in \mathbb{X}. \quad (25)$$

*Proof.* Fix an arbitrary  $\varepsilon^* > 0$ . Due to the definition of  $\bar{w}$ , there exists  $\alpha_0 \in (0, 1)$  such that

$$\bar{w} + \varepsilon^* > (1 - \alpha)m_\alpha, \quad \alpha \in [\alpha_0, 1]. \quad (26)$$

According to Lemma 2, the  $\mathbb{R}_+$ -valued function  $(w, y) \mapsto \tilde{U}_\alpha(w, y)$  is Borel measurable for all  $\alpha \in (0, 1)$ . Therefore, the function  $\eta_{\tilde{U}_\alpha}^\alpha(x, a)$  is well-defined.

Let us prove that

$$\bar{w} + \varepsilon^* + u(x) \geq \min_{a \in \mathbb{A}} \eta_{\tilde{U}_\alpha}^\alpha(x, a), \quad x \in \mathbb{X}, \alpha \in [\alpha_0, 1]. \quad (27)$$

Indeed, Theorem 2 (v) and (26) imply that

$$\begin{aligned} \bar{w} + \varepsilon^* + u_\beta(w, y) &> (1 - \beta)m_\beta + u_\beta(w, y) = v_\beta(w, y) - \beta m_\beta \\ &= \min_{a \in \mathbb{A}} \eta_{u_\beta}^\beta(w, y, a) \geq \min_{a \in \mathbb{A}} \eta_{\tilde{U}_\alpha}^\alpha(w, y, a), \end{aligned}$$

for each  $w \in \mathbb{X}_W$ ,  $y \in \mathbb{X}_Y$ , and  $\alpha, \beta \in [\alpha_0, 1)$  such that  $\beta \geq \alpha$ . Since the right-hand side of the above inequality does not depend on  $\beta \in [\alpha, 1)$ , by taking the infimum in  $\beta \in [\alpha, 1)$ , we obtain that

$$\bar{w} + \varepsilon^* + U_\alpha(w, y) \geq \min_{a \in \mathbb{A}} \eta_{U_\alpha}^\alpha(w, y, a) \geq \min_{a \in \mathbb{A}} \eta_{\tilde{U}_\alpha}^\alpha(w, y, a), \quad (28)$$

for all  $w \in \mathbb{X}_W$ ,  $y \in \mathbb{X}_Y$ , and  $\alpha \in [\alpha_0, 1)$ . Since the function  $c$  is measurable  $\mathbb{K}$ -inf-compact and, due to Lemma 2,  $\tilde{U}_\alpha \in LW(\mathbb{X}_W; \mathbb{X}_Y)$ , and the function  $w \mapsto \min_{a \in \mathbb{A}} \eta_{\tilde{U}_\alpha}^\alpha(w, y, a)$  is nonnegative lower semi-continuous function for each  $y \in \mathbb{X}_Y$ . Therefore, (28) implies that

$$\bar{w} + \varepsilon^* + U_\alpha(w, y) \geq \min_{a \in \mathbb{A}} \eta_{\tilde{U}_\alpha}^\alpha(w, y, a), \quad (29)$$

for all  $w \in \mathbb{X}_W$ ,  $y \in \mathbb{X}_Y$ , and  $\alpha \in [\alpha_0, 1)$ . Thus, since the function  $\tilde{U}_\alpha(w, y)$  is nonincreasing in  $\alpha \in [0, 1)$ , inequalities (27) hold in view of (22).

Let us fix an arbitrary  $x \in \mathbb{X}$ . By Lemma 1 (v, vi), for every  $\alpha \in [0, 1)$  there exists  $a_\alpha \in \mathbb{A}$  such that  $\min_{a \in \mathbb{A}} \eta_{\tilde{U}_\alpha}^\alpha(x, a) = \eta_{\tilde{U}_\alpha}^\alpha(x, a_\alpha)$ . Since  $\tilde{U}_\alpha \geq 0$ , for  $\alpha \in [\alpha_0, 1)$ , inequality (27) can be continued as

$$\bar{w} + \varepsilon^* + u(x) \geq \eta_{\tilde{U}_\alpha}^\alpha(x, a_\alpha) \geq c(x, a_\alpha). \quad (30)$$

Thus, for all  $\alpha \in [\alpha_0, 1)$

$$a_\alpha \in \mathcal{D}_{\eta_{\tilde{U}_\alpha}^\alpha(x, \cdot)}(\bar{w} + \varepsilon^* + u(x)) \subset \mathcal{D}_{c(x, \cdot)}(\bar{w} + \varepsilon^* + u(x)) \subset \mathbb{A}.$$

Since the function  $c(x, \cdot)$  is inf-compact, the nonempty set  $\mathcal{D}_{c(x, \cdot)}(\bar{w} + \varepsilon^* + u(x))$  is compact. Therefore, for every sequence  $\beta^{(n)} \uparrow 1$  of numbers from  $[\alpha_0, 1)$  there is a subsequence  $\{\alpha^{(n)}\}_{n \geq 1}$  such that the sequence  $\{a_{\alpha^{(n)}}\}_{n \geq 1}$  converges and  $a_* := \lim_{n \rightarrow \infty} a_{\alpha^{(n)}} \in \mathbb{A}$ . Consider a sequence  $\alpha^{(n)} \uparrow 1$  such that  $a_{\alpha^{(n)}} \rightarrow a_*$  for some  $a_* \in \mathbb{A}$ . Due to (22) and Lemma 2, similarly to the proof of (17), we obtain that

$$\liminf_{n \rightarrow \infty} \alpha^{(n)} \int_{\mathbb{X}} U_{\alpha^{(n)}}(z) q(dz|x, a^{(n)}) \geq \int_{\mathbb{X}} u(z) q(dz|x, a_*). \quad (31)$$

Therefore, since the function  $c$  is lower semi-continuous in  $a$ , (30) imply

$$\begin{aligned} \bar{w} + \varepsilon^* + u(x) &\geq \liminf_{n \rightarrow \infty} \eta_{\tilde{U}_{\alpha^{(n)}}}^{\alpha^{(n)}}(x, a_{\alpha^{(n)}}) \\ &\geq c(x, a_*) + \int_{\mathbb{X}} u(z) q(dz|x, a_*) \geq \min_{a \in \mathbb{A}} \eta_u^1(x, a^*), \end{aligned}$$

which implies (25) because  $\varepsilon^* > 0$  is arbitrary.  $\square$

*Proof of Theorem 3.* For statement (a) see (22) and the following sentence. Since  $\text{Gr}(A^u) = \{(x, a) \in \text{Gr}(A) : g(x, a) \geq 0\}$ , where  $g(x, a) = \bar{w} + u(x) - c(x, a) - \int_{\mathbb{X}} u(y)q(dy|x, a)$  is a Borel function, the set  $\text{Gr}(A^u)$  is Borel. The sets  $A^u(x)$ ,  $x \in \mathbb{X}$ , are compact because for each  $x \in \mathbb{X}$  the function  $a \mapsto \eta_u^1(x, a)$  is inf-compact on  $\mathbb{A}$  as a sum of inf-compact and nonnegative lower semi-continuous functions. Thus, statement (b) is proved. The Arsenin-Kunugui theorem implies the existence of a stationary policy  $\phi$  such that  $\phi(x) \in A^u(x)$  for all  $x \in \mathbb{X}$ . Statement (d) follows from and Lemma 1(v) because each  $a_* \in A_u(x)$  satisfies  $\eta_u^1(x, a_*) = \min_{a^* \in \mathbb{A}} \eta_u^1(x, a^*) \leq \bar{w} + u(x)$ , where the inequality holds since  $A^u(x) \neq \emptyset$ . The remaining conclusions of Theorem 3 follow from Lemma 3 and [21, Proposition 1.3] stating that inequalities (23) imply optimality of the policy  $\phi$  and (24).  $\square$

Under Assumptions 2, 3, and B, consider the sequence  $\alpha^{(n)} \uparrow 1$  such that  $(1 - \alpha^{(n)})m_{\alpha^{(n)}} \rightarrow \underline{w}$  as  $n \rightarrow \infty$ . Let us define the following nonnegative functions on  $\mathbb{X}_W \times \mathbb{X}_Y$ :

$$\begin{aligned} U_m(w, y) &:= \inf_{n \geq m} u_{\alpha^{(n)}}(w, y), \\ \underset{\sim}{U_m}(w, y) &:= \liminf_{w' \rightarrow w} U_m(w', y), \\ u(w, y) &:= \sup_{m \rightarrow \infty} \underset{\sim}{U_m}(w, y), \end{aligned} \quad (32)$$

$m = 1, 2, \dots, x \in \mathbb{X}$ .

**Theorem 4.** Suppose Assumptions 2, 3, and B hold. Then all the conclusions of Theorem 3 hold and, in addition, for a stationary policy  $\phi$  satisfying (23) with  $u$  defined in (32),

$$w^\phi(x) = \underline{w} = \lim_{\alpha \uparrow 1} (1 - \alpha)v_\alpha(x) = \lim_{N \rightarrow \infty} \frac{1}{N} v_{N,1}^\phi(x), \quad x \in \mathbb{X}. \quad (33)$$

*Proof* repeats the proof of Theorem 3 if we replace  $[\alpha, 1)$  with  $\{\alpha^{(n)}\}_{n \geq m}$ ; cf. [11, Theorem 4].  $\square$

## 6 Approximation of Average Cost Optimal Policies by $\alpha$ -discount Optimal Policies

Under Assumptions 2, 3, and B, consider a nondecreasing sequence  $\alpha^{(n)} \uparrow 1$  such that  $(1 - \alpha^{(n)})m_{\alpha^{(n)}} \rightarrow \underline{w}$  as  $n \rightarrow \infty$ . Consider the nonnegative functions defined in (32). For a family of sets  $\{\text{Gr}(A_{\alpha^{(n)}})\}_{n=1,2,\dots}, x \in \mathbb{X}$ , from Theorem 2, let us set:

$$A^{app}(x) := \left\{ a \in A^u(x) : (x, a) \in \tilde{A} \right\}, \quad x \in \mathbb{X},$$

where  $(w, y, a) \in \tilde{A}$  if and only if there exist a subsequence  $\{\gamma^{(n)}\}_{n=1,2,\dots} \subset \{\alpha^{(n)}\}_{n=1,2,\dots}$  and a sequence  $\{w^{(n)}, a^{(n)}\}_{n=1,2,\dots} \subset \text{Gr}(A_{\alpha^{(n)}})$  that converges to  $(w, a)$  as  $n \rightarrow \infty$ .

**Theorem 5.** Under Assumptions 2, 3, and B, the graph  $\text{Gr}(A^{app})$  is a Borel subset of  $\text{Gr}(A^*)$ , and for each  $x \in \mathbb{X}$  the set  $A^{app}(x)$  is nonempty and compact. Furthermore, there exists a stationary policy  $\phi^{app}$  such that  $\phi^{app}(x) \in A^{app}(x)$  for all  $x \in \mathbb{X}$ , and any such policy is average-cost optimal.

*Proof* is similar to the proof of [8, Theorem 5] with minor changes; cf. the proof of Theorem 3.  $\square$

**Corollary 2.** (cf. [8, Corollary 3]) Under Assumptions 2, 3, and B, for any stationary average-cost optimal policy  $\phi^{app}$ , such that  $\phi^{app}(x) \in A^{app}(x)$  for all  $x \in \mathbb{X}$ , for every  $(w, y) \in \mathbb{X}$  there exist  $\alpha_n \uparrow 1$  and  $w_n \rightarrow w$  as  $n \rightarrow +\infty$  such that for some  $a_n \in A_{\alpha_n}(w_n, y)$ ,  $n \geq 1$ , the equality  $\phi^{app}(w, y) = \lim_{n \rightarrow +\infty} a_n$  holds.

## References

1. Bäuerle, N., Rieder, U.: Markov Decision Processes with Applications to Finance. Springer, Berlin (2011)
2. Bertsekas, D.P., Shreve, S.E.: Stochastic Optimal Control: The Discrete-Time Case. Academic Press, New York (1978)
3. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
4. Bogachev, V.I.: Measure Theory, vol. II. Springer, Berlin (2007)
5. Dynkin, E.B., Yushkevich, A.A.: Controlled Markov Processes. Springer, New York (1979)
6. Feinberg, E.A., Kasyanov, P.O.: MDPs with setwise continuous transition probabilities. [arXiv:2011.01325](https://arxiv.org/abs/2011.01325) (2020)
7. Feinberg, E.A., Kasyanov, P.O., Liang, Y.: Fatou's lemma in its classical form and Lebesgue's convergence theorems for varying measures with applications to Markov decision processes. *Theory Probab. Appl.* **65**(2), 270–291 (2020)
8. Feinberg, E.A., Kasyanov, P.O., Zadoianchuk, N.V.: Average-cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res.* **37**(4), 591–607 (2012)
9. Feinberg, E.A., Kasyanov, P.O., Zadoianchuk, N.V.: Berge's theorem for noncompact image sets. *J. Math. Anal. Appl.* **397**(1), 255–259 (2013)
10. Feinberg, E.A., Kasyanov, P.O., Zgurovsky, M.Z.: Convergence of probability measures and Markov decision models with incomplete information. *Proc. Steklov Inst. Math.* **287**(1), 96–117 (2014)
11. Feinberg, E.A., Kasyanov, P.O., Zgurovsky, M.Z.: Partially observable total-cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res.* **41**(2), 656–681 (2016)
12. Feinberg, E.A., Kasyanov, P.O., Zgurovsky, M.Z.: Markov decision processes with incomplete information and semi-uniform Feller transition probabilities. In preparation (2021)
13. Feinberg, E.A., Kasyanov, P.O., Zgurovsky, M.Z.: Uniform Fatou's lemma. *J. Math. Anal. Appl.* **444**(1), 550–567 (2016)
14. Feinberg, E.A., Kasyanov, P.O., Zadoianchuk, N.V.: Fatou's lemma for weakly converging probabilities. *Theory Probab. Appl.* **58**(4), 683–689 (2014)
15. Feinberg, E.A., Liang, Y.: On the optimality equation for average cost Markov decision processes and its validity for inventory control. *Ann. Oper. Res.* (2017). <https://doi.org/10.1007/s10479-017-2561-9>

16. Hernández-Lerma, O.: Adaptive Markov Control Processes. Springer, New York (1989)
17. Hernández-Lerma, O.: Average optimality in dynamic programming on Borel spaces - Unbounded costs and controls. *Syst. Control Lett.* **17**(3), 237–242 (1991)
18. Hernández-Lerma, O., Lassere, J.B.: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer, New York (1996)
19. Papanicolaou, G.C.: Asymptotic analysis of stochastic equations. In: Rosenblatt, M. (ed.) Studies in Probability Theory, pp. 111–179. Mathematical Association of America, Washington DC (1978)
20. Parthasarathy, K.R.: Probability Measures on Metric Spaces. Academic Press, New York (1967)
21. Schäl, M.: Average optimality in dynamic programming with general state space. *Math. Oper. Res.* **18**(1), 163–172 (1993)



# First Passage Exponential Optimality Problem for Semi-Markov Decision Processes

Haifeng Huo<sup>(✉)</sup> and Xian Wen

Department of School of Science, Guangxi University of Science and Technology,  
Liuzhou 5451006, China  
[xiaohuo08ok@163.com](mailto:xiaohuo08ok@163.com), [wenxian879@163.com](mailto:wenxian879@163.com)

**Abstract.** This paper deals with the exponential utility maximization problem for semi-Markov decision process with Borel state and action spaces, and nonnegative reward rates. The criterion to be optimized is the expected exponential utility of the total rewards before the system state enters the target set. Under the regular and compactness-continuity conditions, we establish the corresponding optimality equation, and prove the existence of an exponential utility optimal stationary policy by an invariant embedding technique. Moreover, we provide an iterative algorithm for calculating the value function as well as the optimal policies. Finally, we illustrate the computational aspects of an optimal policy with an example.

**Keywords:** Semi-Markov decision processes · Exponential utility · First passage time · Value iterative approach · Optimality equation · Optimal policy

**AMS(2020) subject classification:** Primary 90C40 · Secondary 90C39

## 1 Introduction

Semi-Markov decision processes (SMDPs), as an important class of stochastic control problems, have been widely studied [1, 10, 11, 15, 20, 28, 31]. The commonly used criteria for SMDPs are the finite horizon expected criterion [8, 14, 26, 28], the expected discounted criterion [1, 3, 10, 13, 25, 27], and the average criterion [10, 23, 31–33]. These criteria are linear utility functions of the total rewards (i.e. are risk-neutral), which only focus on the expected total rewards of a system during a fixed or a random horizon, and therefore cannot reflect the decision maker's attitude toward risk.

To exhibit the attitude of a decision maker in the face of risk (i.e. risk-seeking or risk-averse), the risk sensitive criteria, which include the exponential utility criterion, have been considered for discrete-time MDPs (DTMDPs) [2, 4–6, 21, 22], and continuous-time MDPs (CTMDPs) [7, 9, 30, 34]. Specifically, Jaquette [21] first introduced the exponential utility to DTMDPs. For the resulting

optimization problem, Chung and Sobel [6] established the corresponding optimality equation by means of the Banach fixed point theorem. Cavazos-Cadena and Montes-De-Oca [4, 5] gave conditions ensuring the existence of optimal policies for the positive dynamic programming, where the state space is considered to be finite in [4], and denumerable in [5]. Jaśkiewicz [22] considered the Borel state and action spaces, and establish the convergence of the  $n$ -stage optimal expected total reward and the existence of an optimal stationary policy. Baüerle and Rieder [2] considered a more general problem than the classic risk sensitive optimization problem, which is called minimizing a certainty equivalent. They solved the optimization problem by an ordinary MDP with extended state space, and proved the existence of an optimal policy under some suitable conditions. For the case of CTMDPs, Ghosh and Saha [7] studied the risk sensitive control in discrete state space. They obtain the value function as a solution to the Hamilton Jacobi Bellman equation, and proved the existence of an optimal Markov control for finite horizon problem, and the existence of an optimal stationary control for infinite horizon problem. Wei [30] dealt with risk sensitive cost criterion for finite horizon CTMDPs with denumerable state space and Borel action space. Under suitable conditions, he proved the existence of the Feynman-Kac formula and an optimal deterministic Markov policy. For the same problem as in [30], Guo, Liu and Zhang [9] investigated the case when the transition and cost rates may be unbounded. They proved that the value function is the unique solution to the optimality equation, and showed the existence of an optimal policy via the Feynman-Kac formula. Few literature [34] applied the uniformization technique to reducing the CTMDPs problem with exponential utility to an equivalent DTMDPs. Recently, Huang, Lian and Guo [17] considered the risk sensitive unconstrained and constrained problems for SMDPs with Borel state space, unbounded cost rates and general utility functions, and proved the existence of the Bellman equation and the optimal policies under some continuity-compactness conditions by using the occupation measure approach.

All of this existing literature shows that all the aforementioned MDPs for the risk-sensitive criterion have two common features: the horizon is finite or infinite, the control model is DTMDPs or CTMDPs. However, such as those encountered in many real world situations, many models in ruin problems [20, 29], reliability [20, 24], and maintenance [20] are considered with a random horizon, and described as SMDPs. Moreover, compared to DTMDPs and CTMDPs (under stationary policies), SMDPs are more general stochastic optimal models, in which the holding time of the system state can be allowed to follow any arbitrary probability distribution. This is the main reason for considering a random horizon for SMDPs in this paper.

Compared with the existing research work for risk-sensitive SMDPs in [17], this paper has some new features as follows: First, in order to make the conclusion more closely fit the actual situation, we pay more attention to the time horizon is the random first passage time, which is more general than those in [17]. Second, since the random first passage time is considered in our control model, by Remark

4.2 in [17], we know that the occupation measure approach is not suitable for our model, because the definition of the occupation measure is based on the discount factor. Instead, we use a so-called minimum nonnegative solution approach to establish the optimality equation and prove the existence of optimal policies. Third, we are mainly concerned with the calculation and existence of the optimal policies, while the purpose of the works in [17] is to establish the existence condition of the optimal policies. Due to these, we develop a value iteration algorithm to calculate the value function and the optimal policy, which is new and the key feature in our paper.

To the best of our knowledge, the risk-sensitive optimality problem for SMDPs in first passage has not been studied yet.

Motivated by the above discussion, we investigate in this paper the first passage risk-sensitive optimality problems for SMDPs. We focus on both the existence conditions and the computational algorithms of an optimal policy, thus we limit the choice of risk-sensitive criteria to the exponential utility criterion (e.g. [2, 6, 21, 34]), which maximizes the expected exponential utility of the total rewards before the state of system enters the target set. More precisely, in order to ensure the existence of an optimal stationary policy, we impose the standard regular condition to ensure that the state process is non-explosive, which is similar to those given in [13–15, 18] for SMDPs (see Lemma 1). Second, compared with [13–15, 18], which are mainly limited to denumerable state space and finite action set, we consider more general Borel state and action spaces. Then, we need to introduce a new continuity-compactness condition (see Assumption 2). Under the regular and continuity-compactness conditions, we establish the corresponding optimality equation, and prove that the value function is a solution to this optimality equation. Moreover, we show the existence of an exponential utility optimal stationary policy by using an invariant embedding technique (see Assumption 1). Furthermore, a value iteration algorithm for computing the value function as well as the optimal policies, in a finite number of iterations, is provided. Finally, an example illustrating the computational methodology of an optimal stationary policy and the value function is given.

The rest of this paper is organized as follows. In Sect. 2, we introduce the semi-Markov decision model and state the first passage exponential utility optimality problem. The main optimality results are stated and proved in Sect. 3. In Sect. 4, an example is provided to illustrate the computational aspects of an optimal policy.

## 2 Model Description

Models of first passage exponential utility SMDPs are defined by

$$\{S, A, (A(x), x \in S), Q(u, y|x, a), B, r(x, a)\} \quad (1)$$

with the following components:

- (a)  $S$  denotes a Borel state space, endowed with the Borel  $\sigma$ -algebras  $\mathcal{B}(S)$ .

- (b)  $A$  denotes a Borel action space, endowed with the Borel  $\sigma$ -algebras  $\mathcal{B}(A)$ .
- (c)  $A(x) \in \mathcal{B}(A)$  represents the set of allowable actions when the system is at state  $x \in S$ .  $K := \{(x, a) | x \in S, a \in A(x)\}$  represents the set of all feasible pairs of states and actions.
- (d)  $Q(\cdot, \cdot | x, a)$  is a semi-Markov kernel on  $R^+ \times S$  given  $K$ , where  $R^+ := [0, \infty)$ . For any  $u \in R^+, D \in \mathcal{B}(S)$ , when the action  $a \in A(x)$  is taken in state  $x$ ,  $Q(u, D | x, a)$  denotes the joint probability that the holding time of the system is no more than  $u \in R^+$  and the state  $x$  changes into the set  $D$ . The semi-Markov kernel  $Q(\cdot, \cdot | x, a), (x, a) \in K$  has the following features:
  - (i) For any  $D \in \mathcal{B}(S)$ ,  $Q(\cdot, D | x, a)$  is a non-decreasing, right continuous function from  $R^+$  to  $[0, 1]$  with  $Q(0, D | x, a) = 0$ .
  - (ii) For any  $u \in R^+$ ,  $Q(u, \cdot | x, a)$  is a sub-stochastic kernel on the state space  $S$ .
- (e)  $B$  is target set, which is a measurable subset of  $S$ , and usually represents the set of failure (or ruin) states of a system.
- (f)  $r(x, a)$  denotes the reward rate, which is assumed to be nonnegative measurable function on  $K$  such that  $r(x, \cdot) \equiv 0$  for all  $x \in B$ .

The first passage SMDP with exponential utility evolves as follows: When the system state is  $x_0 \in B^c$  at time  $t_0 = 0$ , the decision maker selects an admissible action  $a_0$  from the action set  $A(x_0)$ , where  $B^c$  denotes the complement of  $B$ . Consequently, the system stays in the state  $x_0$  up to time  $t_1$ . At this point the system jumps to state  $x_1$  with probability  $p(x_1 | x_0, a_0)$ , and earns a reward  $r(x_0, a_0)(t_1 - t_0)$ . If the state  $x_1 \in B$ , the system will stay at the target set  $B$  forever. If the state  $x_1 \in B^c$ , a new decision epoch  $t_1$  comes along. Then, based on the present state  $x_1$  and the previous state  $x_0$ , the decision maker chooses an action  $a_1 \in A(x_1)$  and the process is repeated. Thus, during its evolution, the system receives a series of rewards. The decision maker aims at maximizing the exponential utility of the total rewards before the state of the system first reaches the target set  $B$ .

Let

$$h_k := (x_0, a_0, t_1, x_1, a_1, \dots, t_k, x_k), \quad (2)$$

be an admissible history up to the  $k$ -th decision epoch, where  $t_{m+1} \geq t_m \geq 0$ ,  $x_m \in S, a_m \in A(x_m)$  for  $m = 0, 1, \dots, k-1$ ,  $x_k \in S$ . From the evolution of SMDPs, we know that  $t_{k+1}$  ( $k \geq 0$ ) denotes the  $(k+1)$ -th decision epoch,  $x_k$  denotes the state of the system on  $[t_k, t_{k+1})$ ,  $a_k$  denotes an action, which is chosen by the decision maker at time  $t_k$ .  $\theta_{k+1} := t_{k+1} - t_k$  denotes the sojourn time at state  $x_k$ , which may follow any given probability distribution.

The set of all admissible histories  $h_k$  is denoted by  $H_k$ , that is  $H_0 := S$  and  $H_k := (S \times A \times (0, +\infty))^k \times S$ .

For the sake of the optimality problem, we shall pay close attention to some classes of policies that we introduce below.

**Definition 1.** A sequence  $\pi = \{\pi_k, k \geq 0\}$  is called stochastic history-dependent policy if, for any  $k = 0, 1, 2, \dots$ , the stochastic kernel  $\pi_k$  on  $A(x_k)$  given  $H_k$  satisfies

$$\pi_k(A(x_k)|h_k) = 1 \text{ for any } h_k \in H_k.$$

Denote by  $\Pi$  the set of all stochastic history-dependent policies,  $\phi$  the set of all stochastic kernels  $\varphi$  on  $A(x)$  given  $S$  such that  $\varphi(A(x)|x) = 1$ , and  $F$  the family of all Borel measurable functions  $f$  from  $S$  to  $A(x)$  for all  $x \in S$ .

**Definition 2.** A policy  $\pi = \{\pi_k\} \in \Pi$  is called stochastic Markov if there exists a sequence of stochastic kernels  $\{\varphi_k\}$  such that  $\pi_k(\cdot|h_k) = \varphi_k(\cdot|x_k)$  for  $k \geq 0, h_k \in H_k$ , and  $\varphi_k \in \phi$ . For simplicity, we denote such a policy by  $\pi = \{\varphi_k\}$ .

A stochastic Markov policy  $\pi = \{\varphi_k\}$  is called stochastic stationary if all the  $\varphi_k$  are independent of  $k$ . Such a policy is denoted by  $\varphi$ , for simplicity.

A stochastic Markov policy  $\pi = \{\varphi_k\}$  is called deterministic Markov if each  $\varphi_k(\cdot|x_k)$  is concentrated at  $f_k(x_k) \in A(x_k)$  for some measurable functions  $\{f_k\}$  with  $k \geq 0, x_k \in S$ , and  $f_k \in F$ .

A deterministic Markov policy  $\pi = \{f_k\}$  is called deterministic stationary if all the measurable functions  $f_k$  are independent of  $k$ . For simplicity, such a policy is denoted by  $f$ .

The class of all stochastic Markov, stochastic stationary, deterministic Markov, and deterministic stationary policies are, respectively, denoted by  $\Pi_{RM}, \Pi_{RS}, \Pi_{DM}$  and  $\Pi_{DS}$ . Clearly,  $\phi = \Pi_{RS} \subset \Pi_{RM} \subset \Pi$  and  $F = \Pi_{DS} \subset \Pi_{DM} \subset \Pi$ .

For the sake of mathematical rigor, we need to construct a well-suited probability space. Define a sample space  $\Omega := \{(x_0, a_0, t_1, x_1, a_1, \dots, t_k, x_k, a_k, \dots) | x_0 \in S, a_0 \in A(x_0), t_l \in (0, \infty], x_l \in S, a_l \in A(x_l) \text{ for each } 1 \leq l \leq k, k \geq 1\}$ . Let  $\mathcal{F}$  be the Borel  $\sigma$ -algebra of the sample space  $\Omega$ . For any  $\omega := (x_0, a_0, t_1, x_1, a_1, \dots, t_k, x_k, a_k, \dots) \in \Omega$ , we define the random variables  $T_k, X_k, A_k$  on  $(\Omega, \mathcal{F})$  as follows:

$$T_k(\omega) := t_k, X_k(\omega) := x_k, A_k(\omega) := a_k, T_\infty(\omega) := \lim_{k \rightarrow \infty} T_k(\omega). \quad (3)$$

In what follows, for the purpose of simplicity, we omit the argument  $\omega$ .

Moreover, we define the state process  $\{x_t, t \geq 0\}$  and the action process  $\{A_t, t \geq 0\}$  on  $(\Omega, \mathcal{F})$  by

$$x_t := \sum_{k \geq 0} I_{\{T_k \leq t < T_{k+1}\}} X_k + \Delta I_{\{t \geq T_\infty\}},$$

$$A_t := \sum_{k \geq 0} I_{\{T_k \leq t < T_{k+1}\}} A_k + a_\Delta I_{\{t \geq T_\infty\}},$$

where  $I_D(\cdot)$  denotes the indicator function on the set  $D$ ,  $\Delta \notin E$  is a cemetery state, and  $a_\Delta$  is an isolated point.

For any policy  $\pi \in \Pi$  and initial state  $x \in S$ , in the light of the Ionescu Tulcea theorem (e.g., Proposition C.10 in [11]), there exist a unique probability measure  $P_x^\pi$  on the measurable space  $(\Omega, \mathcal{F})$  such that,

$$P_x^\pi(A_k \in \Gamma | T_0, X_0, A_0, \dots, T_k, X_k) = \pi_k(\Gamma | T_0, X_0, A_0, \dots, T_k, X_k), \quad (4)$$

$$P_x^\pi(T_{k+1} - T_k \leq u, X_{k+1} \in D | T_0, X_0, A_0, \dots, T_k, X_k, A_k) = Q(u, D | X_k, A_k),$$

for each  $u \in R^+, \Gamma \in \mathcal{B}(A), D \in \mathcal{B}(S), k \geq 0$ . We shall use  $\mathbb{E}_x^\pi$  to represent the expectation operator with respect to  $P_x^\pi$ .

To avoid the possibility that the system generates an infinite number of jumps within a fixed finite horizon, we need to impose the following condition.

**Assumption 1** For any  $\pi \in \Pi, x \in S$ ,  $P_x^\pi(T_\infty = \infty) = 1$ .

To ease the verification of Assumption 1, we state the following sufficient condition for its validity.

**Lemma 1.** If  $Q(\delta, S|x, a) \leq 1 - \varepsilon$  with some constants  $\delta, \varepsilon > 0$  and  $(x, a) \in K$ , then Assumption 1 holds.

*Proof.* The proof follows directly from Proposition 2.1 in [14]. □

*Remark 1.(a)* A key feature of Lemma 1 is that the condition is imposed on the semi-Markov kernel, and can be directly verified.

(b) Lemma 1 is the standard regular condition, which is similar to the classic expected criteria for SMDPs, see, for instance [13–15, 18].

The random variable  $\tau_B$  is given by

$$\tau_B = \begin{cases} \inf\{t \geq 0 : x_t \in B\}, & \text{if } \{t \geq 0 : x_t \in B\} \neq \emptyset; \\ +\infty, & \text{otherwise.} \end{cases} \quad (5)$$

represents the first passage time for which the state process  $\{x_t, t \geq 0\}$  first enters the target set  $B$ .

For any  $x \in S$  and  $\pi \in \Pi$ , we define the first passage exponential utility criterion by

$$V^\pi(x) := E_x^\pi \left( e^{-\gamma \int_0^{\tau_B} r(x_t, A_t) dt} \right), \quad (6)$$

where  $\gamma > 0$  represents the risk aversion coefficient, which expresses the degree of risk aversion that the decision makers face to the level of the total rewards before the state of the system first enters the target set.

**Definition 3.** A policy  $\pi^* \in \Pi$  is called an optimal policy, if

$$V^{\pi^*}(x) = \sup_{\pi \in \Pi} V^\pi(x), x \in S. \quad (7)$$

The corresponding value function is given by

$$V^*(x) := \sup_{\pi \in \Pi} V^\pi(x), x \in S. \quad (8)$$

*Remark 2.* Note that for any  $\pi \in \Pi$  and initial state  $x \in B$ , in view of (5), (6) and (8), we have  $\tau_B = 0$  and  $V^*(x) = V^\pi(x) = 1$ . In order to avoid this trivial case, our arguments consider only the case  $x \in B^c$ .

### 3 Main Results

In this section, we will state the main results concerning the first passage exponential utility optimality problem for SMDPs.

**Notation:** Let  $\mathcal{V}_m$  denotes the set of all Borel measurable functions from  $S$  to  $[0, 1]$ . For any  $x \in B^c, V \in \mathcal{V}_m, \varphi \in \phi, a \in A(x)$ , we define the operators  $M^a V, M^\varphi V$  and  $MV$  as follows:

$$\begin{aligned} M^a V(x) &:= \int_B \int_0^{+\infty} e^{-\gamma r(x,a)u} Q(du, dy|x, a) \\ &\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,a)u} V(y) Q(dy|x, a), \\ M^\varphi V(x) &:= \int_{A(x)} \varphi(da|x) M^a V(x), \\ MV(x) &:= \sup_{a \in A(x)} M^a V(x). \end{aligned}$$

For any  $\varphi \in \phi$ , we also define the operators  $(M^n V, n \geq 1), ((M^\varphi)^n V, n \geq 1)$  as follows:

$$M^{n+1} V = M(M^n V), (M^\varphi)^{n+1} V = M^\varphi((M^\varphi)^n V), n \geq 1.$$

Since the state and action space are Borel space, in order to ensure the existence of optimal policies, it follows from [28, 31, 32], we need establish the following continuity-compactness condition, and which is trivially satisfied for the case of denumerable state space and finite action set  $A(x)$  with  $x \in S$ .

- Assumption 2.** (a) For any  $x \in B^c, A(x)$  is compact;  
(b) For each fixed  $V \in \mathcal{V}_m, \int_{y \in S} \int_0^{+\infty} e^{-\gamma r(x,a)u} V(y) Q(dy|x, a)$  is upper semicontinuous and inf-compact on  $K$ .

**Lemma 2.** Suppose that Assumptions 1 and 2 hold. Then the operators  $M^a$  and  $M$  have the following properties:

- (a) For any  $U, V \in \mathcal{V}_m$ , if  $U \geq V$ , then  $M^a U(x) \geq M^a V(x)$  and  $MU(x) \geq MV(x)$  for any  $x \in S$  and  $a \in A(x)$ .
- (b) For any  $V \in \mathcal{V}_m$ , there exists a policy  $f \in \Pi_{DS}$  such that  $MV(x) = M^f V(x)$  for any  $x \in S$ .

*Proof.* (a) This statement follows from the definitions of operators  $M^a$  and  $M$ .  
(b) Assuming the validity of Assumption 1 and 2, and invoking the measurable selection theorem (Theorem B.6 in [28]), we conclude that, for each  $x \in S$ , there is a stationary policy  $f \in F$  with  $M^f V(x) = MV(x) = \sup_{a \in A(x)} M^a V(x)$ .

□

Since state process  $\{x_t, t \geq 0\}$  is non-explosive and the reward rate is non-negative, in view of the monotone convergence theorem, we can rewrite  $V^\pi(x)$  as follows:

$$\begin{aligned}
V^\pi(x) &= E_x^\pi \left( e^{-\gamma \int_0^{\tau_B} r(x_t, A_t) dt} \right) \\
&= E_x^\pi \left( e^{-\gamma \sum_{m=0}^{\infty} \int_{T_m}^{T_{m+1}} I_{\{\tau_B > t\}} r(x_t, A_t) dt} \right) \\
&= E_x^\pi \left( e^{-\gamma \sum_{m=0}^{\infty} \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} \right) \\
&= \lim_{n \rightarrow \infty} E_x^\pi \left( e^{-\gamma \sum_{m=0}^n \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} \right). \tag{9}
\end{aligned}$$

We shall find it essential to define the sequence  $\{V_n^\pi(x), n = -1, 0, 1, \dots\}$  by

$$\begin{aligned}
V_{-1}^\pi(x) &:= 1, \\
V_n^\pi(x) &:= E_x^\pi \left( e^{-\gamma \sum_{m=0}^n \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} \right).
\end{aligned}$$

Obviously,  $V_n^\pi(x) \geq V_{n+1}^\pi(x)$  for any  $n \geq -1$  and  $\lim_{n \rightarrow \infty} V_n^\pi(x) = V^\pi(x)$  for all  $x \in B^c$ .

**Proposition 1.** *For each  $\pi = \{\pi_0, \pi_1, \dots\} \in \Pi$  and  $x \in S$ . Then, there exists a policy  $\pi' = \{\varphi_0, \varphi_1, \dots\} \in \Pi_{RM}$ , satisfying  $V^\pi(x) = V^{\pi'}(x)$ .*

*Proof.* Since  $V^\pi(x) = E_x^\pi \left( e^{-\gamma \sum_{m=0}^{\infty} \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} \right)$  in (9), to prove this proposition we need to prove that, for each  $x \in S$ , there exists a randomized Markov policy  $\pi' = \{\varphi_0, \varphi_1, \dots\} \in \Pi_{RM}$  such that

$$\begin{aligned}
P_x^{\pi'}(X_k \in D, T_{n+1} - T_n > u, A_k \in \Gamma) \\
= P_x^\pi(X_k \in D, T_{n+1} - T_n > u, A_k \in \Gamma)
\end{aligned}$$

with  $k = 0, 1, \dots, u \in R^+, D \in \mathcal{B}(S), \Gamma \in \mathcal{B}(A)$ .

Thus, in view of property (4), it suffices to show that

$$P_x^{\pi'}(X_k \in D, A_k \in \Gamma) = P_x^\pi(X_k \in D, A_k \in \Gamma). \tag{10}$$

Along the same arguments as in the proof of Theorem 5.5.1 in [28], one can prove (10) by induction on the integer  $k$ .  $\square$

Proposition 1 states, in particular, that in seeking optimal policies for (7), it is sufficient to limit the search to the set of randomized Markov policies. Thus, from now on, we will limit our attention to  $\Pi_{RM}$ .

The following lemma is required to establish the optimality equation.

**Lemma 3.** *Under Assumption 1 and 2, for any  $x \in S$ ,  $n \geq -1$ , and  $\pi = \{\varphi_0, \varphi_1, \dots\} \in \Pi_{RM}$ , the following statements hold.*

- (a)  $V_n^\pi \in \mathcal{V}_m$  and  $V^\pi \in \mathcal{V}_m$ .  
(b)  $V_{n+1}^\pi(x) = M^{\varphi_0} V_n^{1\pi}(x)$  and  $V^\pi(x) = M^{\varphi_0} V^{1\pi}(x)$ , with  ${}^1\pi := \{\varphi_1, \varphi_2, \dots\}$  being the 1-shift policy of  $\pi$ .  
In particular, for any  $f \in F$ ,  $V_{n+1}^f(x) = M^f V_n^f(x)$  and  $V^f(x) = M^f V^f(x)$ .

*Proof.* (a) We shall prove the first statement of (a) by induction on the integer  $n \geq -1$ . The statement is trivial for  $n = -1$  since  $V_{-1}^\pi(x) = 1 \in \mathcal{V}_m$  for any  $x \in S$  and  $\pi \in \Pi_{RM}$ . Assume the statement holds for any  $n < k$ . Then, by (4) and the property of conditional expectation, we have

$$\begin{aligned}
& V_{k+1}^\pi(x) \\
&= E_x^\pi \left( e^{-\gamma \sum_{m=0}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} \right) \\
&= E_x^\pi [E_x^\pi [e^{-\gamma \sum_{m=0}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} | T_0, x_{T_0}, A_0, T_1, x_{T_1}]] \\
&= \int_{A(x)} \varphi_0(da|x) \\
&\quad \times \int_S \int_0^{+\infty} E_x^\pi \left( e^{-\gamma (\int_0^{T_1} r(x_t, A_t) dt + \sum_{m=1}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=1}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt)} \right. \\
&\quad \left. | T_0 = 0, x_{T_0} = x, A_0 = a, T_1 = u, x_{T_1} = y \right) Q(du, dy|x, a) \\
&= \int_{A(x)} \varphi_0(da|x) \int_B \int_0^{+\infty} e^{-\gamma r(x, a)u} Q(du, dy|x, a) + \int_{A(x)} \varphi_0(da|x) \\
&\quad \times \int_{B^c} \int_0^{+\infty} E_x^\pi \left( e^{-\gamma (\int_0^{T_1} r(x_t, A_t) dt + \sum_{m=1}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=1}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt)} \right. \\
&\quad \left. | T_0 = 0, x_{T_0} = x, A_0 = a, T_1 = u, x_{T_1} = y \right) Q(du, dy|x, a) \\
&= \int_{A(x)} \varphi_0(da|x) [\int_B \int_0^{+\infty} e^{-\gamma r(x, a)u} Q(du, j|x, a) \\
&\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x, a)u} E_y^{{}^1\pi} \left( e^{-\gamma \sum_{m=0}^k \int_{T_m}^{T_{m+1}} I_{\{\cap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} \right) \\
&\quad \times Q(du, dy|x, a)] \\
&= \int_{A(x)} \varphi_0(da|x) [\int_B \int_0^{+\infty} e^{-\gamma r(x, a)u} Q(du, dy|x, a) \\
&\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x, a)u} V_k^{1\pi}(y) Q(du, dy|x, a)] \\
&:= M^{\varphi_0} V_k^{1\pi}(x)
\end{aligned}$$

which together with induction hypothesis implies that  $V_{k+1}^\pi(x)$  is a measurable function and  $V_{k+1}^\pi(x) \leq 1$ . Thus,  $V_n^\pi \in \mathcal{V}_m$  for all  $n \geq -1$ . Since the limit of a convergent sequence of measurable functions is itself a measurable function, we obtain  $\lim_{n \rightarrow \infty} V_n^\pi = V^\pi \in \mathcal{V}_m$ . This concludes the proof of (a).

(b) From the proof of part (a), we can deduce that, for any  $x \in B^c$  and  $n \geq -1$ ,

$$V_{n+1}^\pi(x) = M^{\varphi_0} V_n^{1\pi}(x). \quad (11)$$

Letting  $n \rightarrow \infty$  in (11) and using the monotone convergence theorem, we obtain

$$V^\pi(x) = M^{\varphi_0} V^{1\pi}(x).$$

In particular, for  $\pi = f \in F$ , we have  $V^f(x) = M^f V^f(x)$ .  $\square$

*Remark 3.* For any  $x \in B^c$  and  $f \in F$ , one can use Lemma 3 to develop an efficient iteration algorithm for the computation of the function  $V^f(x)$  based on the following:  $V^f(x) = \lim_{n \rightarrow \infty} V_n^f(x)$  where  $V_{-1}^f(x) := 1$  and  $V_{n+1}^f(x) = M^f V_n^f(x)$  for  $n \geq 0$ .

The following theorem states the existence of an optimality equation.

**Theorem 1.** *Under Assumption 1 and 2, the following hold.*

- (a) *For each  $n \geq -1$ , let  $V_{n+1}^* := MV_n^*$  with  $V_{-1}^* := 1$ . Then,  $\lim_{n \rightarrow \infty} V_n^* = V^* \in \mathcal{V}_m$ .*
- (b) *The value function  $V^*$  is a solution to the optimality equation  $V^* = MV^*$ .*
- (c) *There is a policy  $f^* \in F$  such that  $V^*(x) = M^f V^*(x)$ ,  $x \in B^c$ .*

*Proof.* (a) Using Lemma 2(a) and the definition of the operator  $M$ , we obtain  $0 \leq V_{n+1}^*(x) \leq V_n^*(x) \leq 1$  and  $V_n^* \in \mathcal{V}_m$ ,  $n \geq -1$ , for any  $x \in B^c$ . Thus,  $\tilde{V} := \lim_{n \rightarrow \infty} V_n^* \in \mathcal{V}_m$ , since the limit of a convergent sequence of measurable function is also measurable. To complete the proof of part (a), we need to prove that  $\tilde{V} = V^*$ .

We first show by induction on  $n \geq -1$  that for any  $x \in B^c$  and  $\pi = \{\varphi_0, \varphi_1, \dots\} \in \Pi_{RM}$

$$V_n^*(x) \geq V_n^\pi(x). \quad (12)$$

It is clear that  $V_{-1}^* = V_{-1}^\pi = 1$  for any  $\pi \in \Pi_{RM}$ . Suppose that (12) holds for any  $n \leq k$ . By the induction hypothesis, the definition of the operator  $M$  and Lemma 3(b), we have

$$V_{k+1}^*(x) = MV_k^*(x) \geq MV_k^{1\pi}(x) \geq M^{\varphi_0} V_k^{1\pi}(x) = V_{k+1}^\pi(x).$$

Letting  $n \rightarrow \infty$  in (12), we obtain  $\tilde{V}(x) = \lim_{n \rightarrow \infty} V_n^*(x) \geq V^\pi(x)$  with  $\pi \in \Pi_{RM}$ . Since  $\pi$  is arbitrary, we conclude that  $\tilde{V}(x) \geq V^*(x)$ .

We need, now, to prove the reverse inequality  $\tilde{V}(x) \leq V^*(x)$ . For any  $x \in B^c$ ,  $n \geq -1$ , let  $A_n := \{a \in A(x) | M^a V_n^*(x) \geq M\tilde{V}(x)\}$  and  $A^* := \{a \in A(x) | M^a \tilde{V}(x) = M\tilde{V}(x)\}$ . By the compactness-continuity condition in Assumption 2 and the convergence  $V_n^* \downarrow \tilde{V}$ , we conclude that  $A_n$  and  $A^*$  are nonempty and compact, and that  $A_n \downarrow A^*$ . It follows from the measurable selection theorem (Theorem B.6 in [28]) that, for each  $n \geq 1$ , there exist  $a_n \in A_n$

such that  $M^{a_n} V_n^*(x) = MV_n^*(x)$ . Hence, using compactness and the convergence  $A_n \downarrow A^*$ , we deduce that there exist an  $a^* \in A^*$  and a subsequence  $\{a_{n_k}\}$  of  $\{a_n\}$  such that  $a_{n_k} \rightarrow a^*$ . Since  $V_n^* \downarrow \tilde{V}$ , by Lemma 3(a), for any given  $n \geq 1$ , we have

$$M^{a_{n_k}} V_{n_k}^*(x) \leq M^{a_{n_k}} V_n^*(x) \quad \forall n_k \geq n.$$

Letting  $k \rightarrow \infty$  and using the upper semicontinuity condition in Assumption 2 give

$$\tilde{V}^*(x) \leq M^{a^*} V_n^*(x),$$

which together with the convergence  $V_n^* \downarrow \tilde{V}$  imply

$$\tilde{V}^*(x) \leq M^{a^*} \tilde{V}(x) \leq M\tilde{V}(x),$$

By Lemma 2(b), there exists a stationary policy  $f \in F$  such that

$$\tilde{V}(x) \leq M\tilde{V}(x) = M^f \tilde{V}(x).$$

Moreover, using Lemma 2(a), Lemma 3(b) and Remark 3, we obtain

$$\tilde{V}(x) \leq (M^f)^n \tilde{V}(x) \leq (M^f)^n V_{-1}^f(x) = V_{n-1}^f(x).$$

Letting  $n \rightarrow \infty$ , and invoking Remark 3, we obtain  $\tilde{V}(x) \leq V^f(x) \leq V^*(x)$ , which proves the part (a) of the theorem.

(b) By virtue of Lemma 3(b), we know that for any  $x \in B^c$  and  $\pi \in \Pi_{RM}$ , we have

$$V^\pi(x) = M^{\varphi_0} V^1 \pi(x) \leq M^{\varphi_0} V^*(x) \leq MV^*(x).$$

Taking the supremum over all policies  $\pi \in \Pi_{RM}$  implies  $V^*(x) \leq MV^*(x)$ .

The reverse inequality is proved as follows: From the definition of  $V_n^*$ , for any  $x \in B^c$  and  $a \in A(x)$ ,

$$V_{n+1}^*(x) = MV_n^*(x) \geq M^a V_n^*(x).$$

Letting  $n \rightarrow \infty$  and using the monotone convergence theorem, we obtain

$$V^*(x) \geq M^a V^*(x),$$

which implies that  $V^*(x) \geq MV^*(x)$  since  $a \in A(x)$  is arbitrary. This proves  $V^* = MV^*$ .

(c) The statement in (c) follows from Lemma 2.  $\square$

To guarantee the uniqueness of solution of the optimality equation and the existence of the optimal policies, we require the following additional condition (i.e., Assumption 3).

**Assumption 3** For any  $x \in B^c$ ,  $f \in \Pi_s$ ,  $P_x^f(\tau_B < +\infty) = 1$ .

*Remark 4.*(a) Assumption 3 means that, when the initial state of such system is  $X_0 = x \in S$ , the controlled state process  $\{x_t, t \geq 0\}$  will eventually enter the target set  $B$  under the policy  $f \in F$ .

- (b) Letting  $X_n := x_{T_n}, n = 0, 1, \dots$ ,  $T_n$  denotes the jump epoch. Then, we obtain a discrete-time embedded chain  $\{X_n, n \geq 0\}$ . For every  $x \in B^c$ , using Theorem 3.3 in [16], we know that Assumption 3 can be rewritten as follows:

$$P_x^f(\tau_B < +\infty) = P_x^f\left(\bigcup_{n=1}^{\infty} \{X_n \in B\}\right) = 1,$$

which is equivalent to

$$P_x^f\left(\bigcap_{n=1}^{\infty} \{X_n \in B^c\}\right) = 0. \quad (13)$$

- (c) Using Proposition 3.3 in [19], we also obtain a sufficient condition to verify Assumption 3. There exist a constant  $\alpha > 0$  such that  $\int_B P(dy|x, a) \geq \alpha$  for  $(x, a) \in B^c \times A(x)$ , then Assumption 3 holds.

**Lemma 4.** Suppose that Assumptions 1 and 3 hold.

- (a) If  $U, V \in \mathcal{V}_m$  are such that  $U(x) - V(x) \leq M^f(U - V)(x)$  with  $x \in B^c, f \in \Pi_s$ , then  $U(x) \leq V(x)$ .
- (b) For any  $f \in \Pi_s$ ,  $V^f \in \mathcal{V}_m$  is the unique solution to the equation  $V = M^f V$ .

*Proof.* (a) For any  $U, V \in \mathcal{V}_m$ ,  $x \in B^c, f \in \Pi_s$ , we will show the following conclusion by induction,

$$(M^f)^n(U - V)(x) \leq P_x^f\left(\bigcap_{k=1}^n \{X_k \in B^c\}\right), n \geq 1. \quad (14)$$

For  $n = 1$ , it follows from  $U, V \in \mathcal{V}_m$  that

$$\begin{aligned} M^f(U - V)(x) &= M^f U(x) - M^f V(x) \\ &= \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,f)u} (U - V)(y) Q(du, dy | x, a) \\ &\leq \int_{B^c} \int_0^{+\infty} Q(du, dy | x, a) \\ &= P_x^f(X_1 \in B^c). \end{aligned}$$

Suppose that (14) holds for  $n = k$ . Then, by using the induction hypothesis and the nonnegativity of the reward rate, we have

$$\begin{aligned}
(M^f)^{k+1}(U - V)(x) &= M^f(M^f)^k(U - V)(x) \\
&= \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,f)u} (M^f)^k(U - V)(y) \\
&\quad \times Q(du, dy|x, a) \\
&= \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,f)u} P_y^f(\bigcap_{l=1}^k \{X_l \in B^c\}) \\
&\quad \times Q(du, dy|x, a) \\
&\leq \int_{B^c} \int_0^{+\infty} P_y^f(\bigcap_{l=1}^k \{X_l \in B^c\}) Q(du, dy|x, a). \quad (15)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&P_x^f(\bigcap_{l=1}^{k+1} \{X_l \in B^c\}) \\
&= E_x^f[I_{\{\bigcap_{l=1}^{k+1} \{X_l \in B^c\}\}}] \\
&= E_x^f[E_x^f[I_{\bigcap_{l=1}^{k+1} \{X_l \in B^c\}}|X_0, X_1]] \\
&= \int_{B^c} \int_0^{+\infty} P_x^f\left(\bigcap_{l=1}^{k+1} \{X_l \in B^c\}|X_0 = x, X_1 = y\right) Q(du, dy|x, a) \\
&= \int_{B^c} \int_0^{+\infty} P_y^f\left(\bigcap_{l=1}^k \{X_l \in B^c\}\right) Q(du, dy|x, a),
\end{aligned}$$

from which together with (15) and the induction, we have for all  $n \geq 1$ ,

$$U(x) - V(x) \leq (M^f)^n(U(x) - V(x)) \leq P_x^f(\bigcap_{k=1}^n \{X_k \in B^c\}). \quad (16)$$

Letting  $n \rightarrow \infty$ , using (13), we obtain

$$U(x) - V(x) \leq P_x^f(\bigcap_{k=1}^{\infty} \{X_k \in B^c\}) = 0.$$

Then,  $U(x) \leq V(x)$ , for  $x \in S$ .

(b) For any  $x \in S, f \in F$ , it follows from Lemma 2(b) that  $V^f(x) \in \mathcal{V}_m$  satisfies the equation  $V(x) = M^f V(x)$ . If  $U(x)$  is another solution to the equation  $U(x) = M^f U(x)$  on  $S$ , and thus  $U(x) - V^f(x) = M^f(U(x) - V^f(x))$ , which together with the statement in part (a), we know  $U(x) = V^f(x)$  and the uniqueness of solution to the equation is proved.  $\square$

**Theorem 2.** Suppose that Assumption 1, 2 and 3 hold. Then, the following statements hold.

- (a) The value function  $V^*$  is the unique solution to the optimality equation  $V^* = MV^*$ .
- (b) There is a policy  $f^* \in F$  which satisfies  $V^* = M^{f^*}V^*$ ,  $V^* = V^{f^*}$  and such a policy  $f^* \in F$  is optimal.

*Proof.* (a) It follows from Lemma 3 (b) that  $V^*$  satisfies the equation  $V^* = MV^*$ . Then, by Lemma 2(b), there exists a stationary policy  $f^* \in F$  such that  $V^* = M^{f^*}V^*$ . Moreover,  $U$  is another solution of the equation  $U = MU$ . Similarly, the existence of a policy  $f' \in F$  satisfying  $U = M^{f'}U$  is ensured by Lemma 2(b). Then, we have  $V^* - U \leq M^{f^*}(V^* - U)$ . Combining this inequality and Lemma 4 yields that  $V^* \leq U$ . Similarly, we obtain  $U - V^* \leq M^{f'}(U - V^*)$  and  $U \leq V^*$ , which implies  $U = V^*$  and the uniqueness of  $V^*$  is achieved.

(b) Since  $V^* \in \mathcal{V}_m$ , for any  $x \in B^c$ , Lemma 2 guarantees the existence of a stationary policy  $f^* \in F$  such that

$$V^*(x) = M^{f^*}V^*(x),$$

which together with Lemma 3 and Remark 11 yield

$$V^* = \lim_{n \rightarrow \infty} (M^{f^*})^n V^* \leq \lim_{n \rightarrow \infty} (M^{f^*})^n V_{-1}^{f^*} = \lim_{n \rightarrow \infty} V_{n-1}^{f^*} = V^{f^*}.$$

This implies the optimality of  $f^*$ .  $\square$

Theorem 1 leads to the following iterative algorithm for computing the value function and the corresponding optimal policies.

#### The value iteration algorithm procedure:

**Step 1:** For any  $x \in B^c$ , set  $V_{-1}^*(x) := 1$ .

**Step 2:** According to Theorem 1, the value  $V_{n+1}^*(x)$ ,  $n \geq 1$ , is iteratively computed as:

$$\begin{aligned} M^a V_n^*(x) &= \int_B \int_0^{+\infty} e^{-\gamma r(x,f)u} Q(du, dy|x, a) \\ &\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,f)u} V_n^*(y) Q(du, dy|x, a), \\ V_{n+1}^*(x) &= \sup_{a \in A(x)} \{M^a V_n^*(x)\}. \end{aligned}$$

**Step 3:** When  $|V_{n+1}^* - V_n^*| < 10^{-12}$ , the iteration stops. Since  $V_n^*$  is very close to  $V_{n+1}^*$ , one can view  $V_{n+1}^*$  as a good approximation of the value function  $V^*$ . In addition, Lemma 2 and Theorem 2 ensure the existence of a policy  $f^* \in F$  such that  $MV^* = M^{f^*}V^*$ , and this policy  $f^*$  is optimal. Or else, go back to step 2 and replace  $n$  with  $n + 1$ .

## 4 Example

In this section, an example is given to illustrate our main results, and to demonstrate the computation of an optimal stationary policy and the corresponding value function using the above described iterative algorithm.

*Example 1.* Consider a company using idle funds for financial management. When the company has some idle funds (which is denoted by state 1), the decision maker gets the reward at the rate of return  $r(1, a_{11}) \geq 0$  through deposit method  $a_{11}$  or the reward at the rate of return  $r(1, a_{12}) \geq 0$  through another deposit method  $a_{12}$ . When the company has plenty of idle funds (which is denoted by state 2), the decision maker can choose a financial management  $a_{21}$  earning in a reward rate  $r(2, a_{21}) \geq 0$  or another financing way  $a_{22}$  earning in a reward rate  $r(2, a_{22}) \geq 0$ . When the company goes bankrupt (which is denoted by state 0), the decision-maker does not need to choose any way of financing  $a_{01}$  and cannot get any reward  $r(0, a_{01}) = 0$ .

Suppose that the evolution mechanism of this system is described as a SMDP. When the system state is 1, the decision maker selects an admissible action  $a_{1n}, n = 1, 2$ . Then, the system stays at the state 1 with a random time satisfying the uniform distribution in the region  $[0, u(1, a_{1n})], n = 1, 2$ . After the system state lingers for a period of time, it will move to a new state  $j \in \{0, 2\}$  with the probability  $p(j|1, a_{1n}), n = 1, 2$ . When the action  $a_{2n}$  is selected  $n = 1, 2$ , the system stays at 2 with a random time satisfying the exponential distribution with the parameter  $\lambda(2, a_{2n})$ . Consequently, the system jumps to state  $j \in \{0, 1\}$  with the probability  $p(j|2, a_{2n}), n = 1, 2$ .

The corresponding parameters of this SMDPs are given as follows: The state space  $S = \{0, 1, 2\}$ , the target set  $B = \{0\}$  and the admissible action sets  $A(0) = \{a_{01}\}, A(1) = \{a_{11}, a_{12}\}, A(2) = \{a_{21}, a_{22}\}$ , the risk-sensitivity coefficient  $\gamma = 1$ . The transition probabilities are assumed to be given

$$\begin{aligned} p(0|0, a_{01}) &= 1, & p(0|1, a_{11}) &= \frac{1}{2}, & p(2|1, a_{11}) &= \frac{1}{2}, \\ p(0|1, a_{12}) &= \frac{2}{3}, & p(2|1, a_{12}) &= \frac{1}{3}, & p(0|2, a_{21}) &= \frac{3}{10}, \\ p(1|2, a_{21}) &= \frac{7}{10}, & p(0|2, a_{22}) &= \frac{2}{5}, & p(1|2, a_{22}) &= \frac{3}{5}. \end{aligned} \quad (17)$$

In addition, the corresponding distribution parameters are given by

$$\begin{aligned} u(1, a_{11}) &= 30, & u(1, a_{12}) &= 40, \\ \lambda(2, a_{21}) &= 0.11, & \lambda(2, a_{22}) &= 0.13. \end{aligned} \quad (18)$$

and the reward rates are given by

$$\begin{aligned} r(1, a_{11}) &= 0.0035, & r(1, a_{12}) &= 0.011, \\ r(2, a_{21}) &= 0.013, & r(2, a_{22}) &= 0.015. \end{aligned}$$

In this model, we mainly focus on the existence and calculation parts of an optimal policy and the value function for first passage exponential utility criterion. As can be seen from the discussion in Sect. 3 above, we first need to verify Assumption 1, 2 and 3. Indeed, by (17) and (18), we know that Assumption 1 and 3 are satisfied. Moreover, since the state space is denumerable and the action space  $A$  is finite, Assumption 2 is trivially satisfied. Thus, by Theorem 1 and 2,

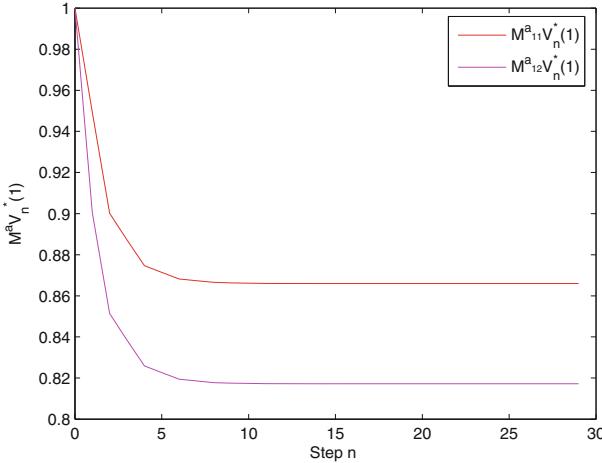
the value iteration technique can be used for evaluating the value function and the exponential optimal policies as follows:

**Step 1:** Let  $V_{-1}^*(x) := 1, x = 1, 2$ .

**Step 2:** For  $x = 1, 2, n \geq 1$ , using Theorem 1 (a), we obtain

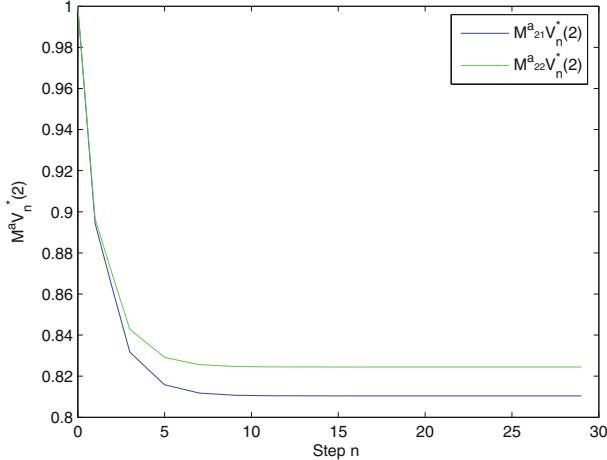
$$\begin{aligned}
V_n^*(1) &= MV_{n-1}^*(1), \\
&= \max \left\{ \frac{1}{2} \times \frac{1}{30} \times \int_0^{30} e^{-0.0035u} du \right. \\
&\quad + \frac{1}{2} \times \frac{1}{30} \times \int_0^{30} e^{-0.0035u} du \times V_{n-1}^*(2), \\
&\quad \left. \frac{2}{3} \times \frac{1}{40} \times \int_0^{40} e^{-0.011u} du + \frac{1}{3} \times \frac{1}{40} \times \int_0^{40} e^{-0.011u} du \times V_{n-1}^*(2) \right\} \\
V_n^*(2) &= MV_{n-1}^*(2), \\
&= \max \left\{ \frac{3}{10} \times 0.11 \times \int_0^{+\infty} e^{-0.123u} du \right. \\
&\quad + \frac{7}{10} \times 0.11 \times \int_0^{+\infty} e^{-0.123u} du \times V_{n-1}^*(1), \\
&\quad \left. \frac{2}{5} \times 0.13 \times \int_0^{+\infty} e^{-0.145u} du + \frac{3}{5} \times 0.13 \times \int_0^{+\infty} e^{-0.145u} du \times V_{n-1}^*(1) \right\}
\end{aligned}$$

**Step 3:** When  $|V_n^* - V_{n-1}^*| < 10^{-12}$ , go to step 4, the value  $V_n^*$  is usually approximated as  $V^*$ ; otherwise, go to step  $n + 1$  and go back to step 2.

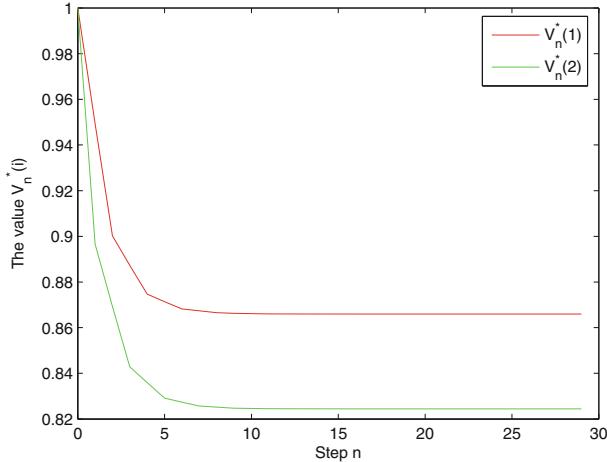


**Fig. 1.** The function  $M^a V_n^*(1)$

**Step 4:** Plot out the graphs of the value functions  $M^{a_{ij}} V_n^*(i)$  and  $V_n^*(i)$ ,  $i = 1, 2; j = 1, 2$ , see Figs. 1, 2 and 3.



**Fig. 2.** The function  $M^a V_n^*(2)$



**Fig. 3.** The value function  $V_n^*(i)$

Moreover, for  $x = 1$ , using Theorem 1, 2, Fig. 1 and Fig. 2, we know that

$$MV^*(1) = V^*(1) = M^{a_{11}} V^*(1).$$

For  $x = 2$ , we also obtain

$$MV^*(2) = V^*(2) = M^{a_{22}} V^*(2).$$

According to the above analysis and Theorem 2, we obtain the optimal stationary policy  $f^*(1) = a_{12}, f^*(2) = a_{21}$  and the value function  $V^*(1) = 0.8660, V^*(2) = 0.8245$ .

**Acknowledgement.** This work was supported by National Natural Science Foundation of China (Grant No. 11961005, 11801590); Foundation of Guangxi Educational Committee (Grant No. KY2019YB0369); Ph.D. research startup foundation of Guangxi University of Science and Technology (Grant No. 18Z06); Guangxi Natural Science Foundation Program (Grant No. 2020GXNSFAA297196).

## References

1. Baüerle, N., Rieder, U.: *Markov Decision Processes with Applications to Finance*. Springer, Heidelberg (2011)
2. Baüerle, N., Rieder, U.: More risk-sensitive Markov decision processes. *Math. Oper. Res.* **39**, 105–120 (2014)
3. Cao, X.R.: Semi-Markov decision problems and performance sensitivity analysis. *IEEE Trans. Autom. Control* **48**, 758–769 (2003)
4. Cavazos-Cadena, R., Montes-De-Oca, R.: Optimal stationary policies in risk-sensitive dynamic programs with finite state space and nonnegative rewards. *Appl. Math. (Warsaw)* **27**, 167–185 (2000)
5. Cavazos-Cadena, R., Montes-De-Oca, R.: Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces. *Math. Meth. Oper. Res.* **52**, 133–167 (2000)
6. Chung, K.J., Sobel, M.J.: Discounted MDP's: distribution functions and exponential utility maximization. *SIAM J. Control Optim.* **25**, 49–62 (1987)
7. Ghosh, M.K., Saha, S.: Risk-sensitive control of continuous time Markov chains. *Stochastics* **86**, 655–675 (2014)
8. Ghosh, M.K., Saha, S.: Non-stationary semi-Markov decision processes on a finite horizon. *Stoch. Anal. Appl.* **31**, 183–190 (2013)
9. Guo, X., Liu, Q.L., Zhang, Y.: Finite horizon risk-sensitive continuous-time Markov decision processes with unbounded transition and cost rates. *4OR* **17**, 427–442 (2019)
10. Guo, X.P., Hernández-Lerma, O.: *Continuous-Time Markov Decision Processes: Theory and Applications*. Springer, Berlin (2009)
11. Hernández-Lerma, O., Lasserre, J.B.: *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, New York (1996)
12. Howard, R.A., Matheson, J.E.: Risk-sensitive Markov decision processes. *Manage. Sci.* **18**, 356–369 (1972)
13. Huang, Y.H., Guo, X.P.: Discounted semi-Markov decision processes with nonnegative costs. *Acta Math. Sin. (Chinese Ser.)* **53**, 503–514 (2010)
14. Huang, Y.H., Guo, X.P.: Finite horizon semi-Markov decision processes with application to maintenance systems. *Eur. J. Oper. Res.* **212**, 131–140 (2011)
15. Huang, Y.H., Guo, X.P.: Mean-variance problems for finite horizon semi-Markov decision processes. *Appl. Math. Optim.* **72**, 233–259 (2015)
16. Huang, Y.H., Guo, X.P., Song, X.Y.: Performance analysis for controlled semi-Markov process. *J. Optim. Theory Appl.* **150**, 395–415 (2011)
17. Huang, Y.H., Lian, Z.T., Guo, X.P.: Risk-sensitive semi-Markov decision processes with general utilities and multiple criteria. *Adv. Appl. Probab.* **50**, 783–804 (2018)
18. Huang, X.X., Zou, X.L., Guo, X.P.: A minimization problem of the risk probability in first passage semi-Markov decision processes with loss rates. *Sci. China Math.* **58**, 1923–1938 (2015)

19. Huo, H.F., Zou, X.L., Guo, X.P.: The risk probability criterion for discounted continuous-time Markov decision processes. *Discrete Event Dyn. Syst.* **27**, 675–699 (2017)
20. Janssen, J., Manca, R.: *Semi-Markov Risk Models for Finance, Insurance, and Reliability*. Springer, New York (2006)
21. Jaquette, S.C.: A utility criterion for Markov decision processes. *Manage. Sci.* **23**, 43–49 (1976)
22. Jaśkiewicz, A.: A note on negative dynamic programming for risk-sensitive control. *Oper. Res. Lett.* **36**, 531–534 (2008)
23. Jaśkiewicz, A.: On the equivalence of two expected average cost criteria for semi Markov control processes. *Math. Oper. Res.* **29**, 326–338 (2013)
24. Limnios, N., Oprisan, G.: *Semi-Markov Processes and Reliability*. Birkhäuser, Boston (2001)
25. Luque-Vásquez, F., Minjárez-Sosa, J.A.: Semi-Markov control processes with unknown holding times distribution under a discounted criterion. *Math. Meth. Oper. Res.* **61**, 455–468 (2005)
26. Mamer, J.W.: Successive approximations for finite horizon semi-Markov decision processes with application to asset liquidation. *Oper. Res.* **34**, 638–644 (1986)
27. Nollau, V.: Solution of a discounted semi-Markovian decision problem by successive overrelaxation. *Optimization* **39**, 85–97 (1997)
28. Puterman, M.L.: *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York (1994)
29. Schäl, M.: Control of ruin probabilities by discrete-time investments. *Math. Meth. Oper. Res.* **70**, 141–158 (2005)
30. Wei, Q.D.: Continuous-time Markov decision processes with risk-sensitive finite-horizon cost criterion. *Math. Meth. Oper. Res.* **84**, 1–27 (2016)
31. Wei, Q.D., Guo, X.P.: New average optimality conditions for semi-Markov decision processes in Borel spaces. *J. Optim. Theory Appl.* **153**, 709–732 (2012)
32. Wei, Q.D., Guo, X.P.: Constrained semi-Markov decision processes with ratio and time expected average criteria in Polish spaces. *Optimization* **64**, 1593–1623 (2015)
33. Yushkevich, A.A.: On semi-Markov controlled models with average reward criterion. *Theory Probab. Appl.* **26**, 808–815 (1982)
34. Zhang, Y.: Continuous-time Markov decision processes with exponential utility. *SIAM J. Control Optim.* **55**, 1–24 (2017)



# Controlled Random Walk: Conjecture and Counter-Example

Alexey B. Piunovskiy<sup>(✉)</sup>

Department of Mathematical Sciences, University of Liverpool,  
Liverpool L69 7ZL, UK  
[piunov@liv.ac.uk](mailto:piunov@liv.ac.uk)

**Abstract.** In this paper we investigate the following conjecture about the random walk on the positive integer lattice, starting from a large point  $i > 0$  and up to the absorption at negative points: *on the first steps, one has to maximize the expected reward coming from passing through one point on the lattice*. Under appropriate conditions, this conjecture is true. The counter-example shows that sometimes it is not valid.

**Keywords:** Random walk · Markov decision process · Turnpike · Total expected cost

**AMS(2020) Subject Classification:** Primary 90C40 · Secondary 90C39

## 1 Introduction

The current article is an attempt to study the conjecture formulated by Prof. I.Sonin in a private communication.

The random walk on the positive integer lattice, starting from  $X_0 = i$ , is defined by equation

$$X_t = X_{t-1} - Z_t(a_t)$$

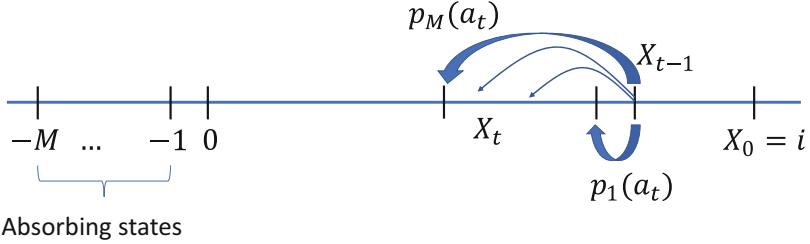
and is terminated as soon as  $X_t < 0$ . Here  $\{Z_t(a)\}_{t=1}^\infty$  are mutually independent positive integer-valued random variables depending on the action  $a \in \mathbf{A} = \{a_1, a_2, \dots, a_N\}$ , with the given probability distribution

$$P(Z(a) = m) = p_m(a), \quad m = 1, 2, \dots, M.$$

See Fig. 1.

On each step  $t$ , the associated expected reward equals  $R^{a_t}$ . For example, if  $R_{Z(a)}(a)$  is the reward associated with the action  $a \in \mathbf{A}$  and the value  $Z(a)$ , then

$$R^a = \sum_{m=1}^M R_m(a)p_m(a).$$



**Fig. 1.** Random walk.

For a fixed  $a \in \mathbf{A}$ , for a large initial state  $i$ , the total expected number of steps up to the absorption equals  $\approx \frac{i}{L^a}$ , where  $L^a := E[Z(a)] = \sum_{m=1}^M m p_m(a)$ . Thus, the total expected reward is  $\approx i \frac{R^a}{L^a}$ . To put it slightly different, the expected reward, coming from passing through one point on the lattice, equals  $\approx \frac{R^a}{L^a}$ . The **conjecture** to be investigated reads as follows:

There is such  $I < \infty$  that, if  $X_{t-1} \geq I$ , then the optimal action  $a_t \in \mathbf{A}_*$ ,  
where  $\mathbf{A}_* := \{a \in \mathbf{A} : \frac{R^a}{L^a} = c_* := \max_{a \in \mathbf{A}} \frac{R^a}{L^a}\}$

In the current article, we provide sufficient conditions for this conjecture to be valid. The numerical example in Sect. 4 shows that in general it is not the case. In Sect. 5, we formulate the similar statement for the discounted version of the described model. All the proofs are presented in the Appendix.

In what follows, if  $P$  is a matrix, then  $P_{s,\cdot}$  denotes its  $s$ -th row. We say that a stochastic matrix  $P$  is ergodic or aperiodic if the corresponding Markov chain is so. The maximum (minimum) over the empty set equals  $-\infty (+\infty)$ .

## 2 MDP Formulation and Preliminaries

Obviously, we deal with the Markov decision process (MDP) with the state space

$$\mathbf{X} := \{-M, -M + 1, \dots, -1, 0, 1, 2, \dots\},$$

action space

$$\mathbf{A} := \{a_1, a_2, \dots, a_N\},$$

the transition probability

$$\tilde{P}_{i,j}(a) = \begin{cases} p_m(a), & \text{if } i \geq 0, j = i - m, m = 1, 2, \dots, M; \\ 0 & \text{otherwise,} \end{cases}$$

and the reward function

$$r_i(a) := \begin{cases} 0, & \text{if } i < 0; \\ R^a, & \text{if } i \geq 0. \end{cases}$$

The initial state is  $i \in \mathbf{X}$ , and we consider MDP  $\langle \mathbf{X}, \mathbf{A}, \tilde{P}, r \rangle$  with the total expected reward, with (random) states and actions

$$X_0 = i, A_1, X_1, A_2, \dots .$$

The definition of a strategy  $\pi$  (past-dependent, randomized) is conventional [1, 2, 4, 6];  $E_i^\pi$  is the corresponding mathematical expectation;

$$V(i) := \sup_{\pi} E_i^\pi \left[ \sum_{t=1}^{\infty} r_{X_{t-1}}(A_t) \right] \quad (2)$$

is the Bellman function for this MDP;  $i \in \mathbf{X}$ . Since the reward  $r$  is bounded and the process  $X_t$  is ultimately absorbed at  $\{-M, -M+1, \dots, -1\}$  after (maximum)  $i + 1$  time steps, the function  $V$  is finite-valued. It is well known (see, e.g., [2, Ch.4] or [1, §9.5]) that the function  $V$  is the unique solution to the optimality (Bellman) equation

$$\begin{aligned} V(i) &= \max_{a \in \mathbf{A}} \left\{ R^a + \sum_{m=1}^M V(i-m)p_m(a) \right\} \quad \text{for } i \geq 0; \\ V(i) &= 0 \quad \text{for } i = -M, -M+1, \dots, -1, \end{aligned} \quad (3)$$

which can be solved successively for  $i = 0, 1, \dots$ . Now the conjecture (1) is reformulated as follows:

$$\begin{aligned} \text{There is such } I < \infty \text{ that, for all } i \geq I, \\ \text{the maximum in (3) is only provided by } a \in \mathbf{A}_*. \end{aligned} \quad (4)$$

It is natural to call the interval  $\{I, I+1, \dots\}$  ‘Turnpike’.

**Lemma 1.** *Function*

$$\tilde{W}(i) := V(i) - c_* i, \quad i \in \mathbf{X} \quad (5)$$

is the (unique) uniformly bounded function satisfying equation

$$\begin{aligned} \tilde{W}(i) &= -c_* i \quad \text{for } i = -M, -M+1, \dots, -1; \\ \tilde{W}(i) &= \max_{a \in \mathbf{A}} \left\{ L^a \left( \frac{R^a}{L^a} - c_* \right) + \sum_{m=1}^M \tilde{W}(i-m)p_m(a) \right\} \quad \text{for } i \geq 0, \end{aligned} \quad (6)$$

which can be solved successively for  $i = 0, 1, \dots$ . Hence

$$V(i) = c_* i + O(1) \quad \text{when } i \rightarrow \infty.$$

Moreover, for each  $i \in \mathbf{X}$ , the maxima in (5) and in (6) are provided by the same values of  $a \in \mathbf{A}$ .

Every value of  $i \in \mathbf{X}$  can be uniquely represented as

$$i = (k - 1)M + s, \quad \text{where } s \in \mathbf{S} := \{0, 1, \dots, M - 1\}, \quad k = 0, 1, 2, \dots \quad (7)$$

For each  $i \in \mathbf{X}$  with the corresponding values of  $k$  and  $s$ , we denote  $\tilde{W}(i)$ , introduced in (5), as  $W^k(s)$ . Now Eq. (6) takes the following form:

$$\begin{aligned} W^0(s) &= -c_*(-M + s) \quad \text{for } s \in \mathbf{S}; \\ W^{k+1}(0) &= \max_{a \in \mathbf{A}} \left\{ L^a \left( \frac{R^a}{L^a} - c_* \right) + \sum_{j=0}^{M-1} W^k(j)p_{M-j}(a) \right\}, \\ W^{k+1}(1) &= \max_{a \in \mathbf{A}} \left\{ L^a \left( \frac{R^a}{L^a} - c_* \right) + W^{k+1}(0)p_1(a) + \sum_{j=1}^{M-1} W^k(j)p_{M-j+1}(a) \right\}, \\ &\dots \\ W^{k+1}(M-1) &= \max_{a \in \mathbf{A}} \left\{ L^a \left( \frac{R^a}{L^a} - c_* \right) + W^{k+1}(M-2)p_1(a) + W^{k+1}(M-3)p_2(a) \right. \\ &\quad \left. + \dots + W^{k+1}(0)p_{M-1}(a) + W^k(M-1)p_M(a) \right\}, \quad k = 0, 1, \dots \end{aligned} \quad (8)$$

After we introduce the stochastic matrix

$$P(a) := \begin{pmatrix} P_{0,0}(a) = p_M(a) & P_{0,1}(a) = p_{M-1}(a) & \dots & P_{0,M-1}(a) = p_1(a) \\ P_{1,0}(a) = p_1(a) & P_{1,1}(a) = p_M(a) & \dots & P_{1,M-1}(a) = p_2(a) \\ \dots & & & \dots \\ P_{M-1,0}(a) = p_{M-1}(a) & P_{M-1,1}(a) = p_{M-2}(a) & \dots & P_{M-1,M-1}(a) = p_M(a) \end{pmatrix}, \quad (9)$$

the obtained equations for  $W^k(s)$  can be rewritten as

$$\begin{aligned} W^0(s) &= -c_*(-M + s) \quad \text{for } s \in \mathbf{S}; \\ W^{k+1}(s) &= \max_{a \in \mathbf{A}} \left\{ L^a \left( \frac{R^a}{L^a} - c_* \right) + \sum_{j=0}^{s-1} W^{k+1}(j)P_{s,j}(a) + \sum_{j=s}^{M-1} W^k(j)P_{s,j}(a) \right\} \\ &\quad \text{for } s \in \mathbf{S}, \quad k \geq 0. \end{aligned} \quad (10)$$

Iterations (10) are similar to the Gauss-Seidel version of the value iteration algorithm for the average reward MDP (see [6, §6.3.3]). For a fixed  $k \geq 0$ , we substitute the expression for  $W^{k+1}(0)$  in the formula for  $W^{k+1}(1)$ , the expression for  $W^{k+1}(0)$  and the modified expression for  $W^{k+1}(1)$  in the formula for  $W^{k+1}(2)$  and so on. After we denote  $\mathbf{D}$  the finite set of all mappings from  $\mathbf{S}$  to  $\mathbf{A}$ , called below ‘decisions’, iterations (10) can be represented in the form

$$\begin{aligned} W^0(s) &= -c_*(-M + s) \quad \text{for } s \in \mathbf{S}; \\ W^{k+1}(s) &= \max_{d \in \mathbf{D}} U_d \circ W^k(s) \quad \text{for } s \in \mathbf{S}, \quad k \geq 0, \end{aligned}$$

where

$$\begin{aligned}
U_d \circ W^k(0) &:= L^{d(0)} \left( \frac{R^{d(0)}}{L^{d(0)}} - c_* \right) + \sum_{j=0}^{M-1} W^k(j) P_{0,j}(d(0)) \\
U_d \circ W^k(1) &:= L^{d(1)} \left( \frac{R^{d(1)}}{L^{d(1)}} - c_* \right) + P_{1,0}(d(1)) L^{d(0)} \left( \frac{R^{d(0)}}{L^{d(0)}} - c_* \right) \\
&\quad + P_{1,0}(d(1)) \sum_{j=0}^{M-1} W^k(j) P_{0,j}(d(0)) + \sum_{j=1}^{M-1} W^k(j) P_{1,j}(d(1)) \\
&\text{and so on, up to } U_d \circ W^k(M-1).
\end{aligned}$$

In what follows, each function  $W : \mathbf{S} \rightarrow \mathbb{R}$  is identified with the column vector  $W \in \mathbb{R}^M$ , and the both notations  $W(s) = W_s$  are in use. Now one can rewrite iterations (10) in the matrix form:

$$\begin{aligned}
W^0(s) &= -c_*(-M+s), \quad s \in \mathbf{S}; \\
W^{k+1} &= \max_{d \in \mathbf{D}} U_d \circ W^k = \max_{d \in \mathbf{D}} \{\mathcal{R}(d) + \mathcal{Q}(d)W^k\}, \quad k = 0, 1, \dots,
\end{aligned} \tag{11}$$

where, for fixed  $d \in \mathbf{D}$ , the column vector  $\mathcal{R}(d) \in \mathbb{R}^M$  and the rows of the square  $M \times M$  matrix  $\mathcal{Q}(d)$  are defined recursively:

$$\begin{aligned}
\mathcal{R}_0(d) &= L^{d(0)} \left( \frac{R^{d(0)}}{L^{d(0)}} - c_* \right); \\
\mathcal{R}_{l+1}(d) &= L^{d(l+1)} \left( \frac{R^{d(l+1)}}{L^{d(l+1)}} - c_* \right) + \sum_{j=0}^l P_{l+1,j}(d(l+1)) \mathcal{R}_j(d), \\
l &= 0, \dots, M-2;
\end{aligned} \tag{12}$$

$$\begin{aligned}
\mathcal{Q}_{0,j}(d) &= P_{0,j}(d(0)), \quad j = 0, 1, \dots, M-1; \\
\mathcal{Q}_{l+1,j}(d) &= \begin{cases} \sum_{i=0}^l [P_{l+1,i}(d(l+1)) \mathcal{Q}_{i,j}(d)] & \text{for } j < l+1; \\ \sum_{i=0}^l [P_{l+1,i}(d(l+1)) \mathcal{Q}_{i,j}(d)] + P_{l+1,j}(d(l+1)) & \text{for } j \geq l+1. \end{cases} \\
l &= 0, \dots, M-2;
\end{aligned} \tag{13}$$

Clearly,  $\mathcal{R}_s(d) \leq 0$  for all  $s \in \mathbf{S}$  and  $d \in \mathbf{D}$ . We underline that the rows  $\mathcal{Q}_{l,.}(d)$  of the matrix  $\mathcal{Q}(d)$  with  $l \leq s$  do not depend on the values of  $d(s+1), d(s+2), \dots, d(M-1)$ . Note also that, for every function  $W$  on  $\mathbf{S}$ , there is  $\hat{d} \in \mathbf{D}$  providing the component-wise maximum to  $U_d \circ W$ . Indeed, the values  $\hat{d}(s)$  can be calculated successively for  $s = 0, 1, \dots, M-1$ , and any other mapping  $d$  is such that  $U_d \circ W \leq T_{\hat{d}} \circ W$  component-wise. In other words,  $\hat{d}$  solves the vector optimization problem  $U_d \circ W \rightarrow \max_{d \in \mathbf{D}}$ , i.e., this problem is well defined: the Pareto set contains the unique point  $T_{\hat{d}} \circ W$ .

- Lemma 2.** (a) The matrix  $\mathcal{Q}(d)$  is stochastic for all  $d \in \mathbf{D}$ , provided the original matrix  $P(a)$  is stochastic for all  $a \in \mathbf{A}$ .  
 (b) Suppose  $d \in \mathbf{D}$  is such that  $P_{s,s}(d(s)) > 0$  for all  $s \in \mathbf{S}$ . Then, for all  $s, l \in \mathbf{S}$ ,  $\mathcal{Q}_{s,l}(d) > 0$  provided  $P_{s,l}(d(s)) > 0$ .

**Definition 1.** Decisions  $d \in \mathbf{D}$  satisfying the property  $d(s) \in \mathbf{A}_*$  for all  $s \in \mathbf{S}$  will be called trivial. Equivalently, a decision  $d \in \mathbf{D}$  is trivial if and only if  $\mathcal{R}(d) = \mathbf{0}$ . Here and below,  $\mathbf{0} \in \mathbb{R}^M$  is the zero vector. The set of all trivial decisions is denoted as  $\mathbf{D}_*$ .

The conjecture (4), and also (1) is now reformulated as follows:

There exists  $K$  such that, for all  $k \geq K$ , the maximum in (11) (14)  
 is only provided by the trivial decisions  $d \in \mathbf{D}_*$ .

Note that all the vectors  $W^0, W^1, \dots$  are uniformly bounded by Lemma 1, and the maxima in (4) and (10) are provided by the same values of  $a \in \mathbf{A}$ .

### 3 Main Results

**Condition 1.** There exists  $J$  such that, for every sequence of mappings  $\hat{d}_1, \hat{d}_2, \dots, \hat{d}_J \in \mathbf{D}_*$ , the matrix  $\mathcal{Q}(\hat{d}_1)\mathcal{Q}(\hat{d}_2)\dots\mathcal{Q}(\hat{d}_J)$  contains no zeroes.

**Theorem 1.** If Condition 1 is satisfied then the conjecture (14) (and also (4) and (1)) is valid.

In the following statements, the sufficient conditions for the conjecture (14) to be valid are given in terms of the original matrix  $P(a)$ .

**Corollary 1.** Suppose  $\mathbf{A}_* = \{a_*\}$  is a singleton (consequently  $\mathbf{D}_* = \{d_*\}$  is a singleton with  $d_*(s) \equiv a_*$ ). Let the matrix  $P(a_*)$  be ergodic. Assume additionally that  $p_M(a_*) > 0$ . Then Condition 1 is satisfied, and hence the conjecture (14) (and also (4) and (1)) is valid.

When using a different method of attack, one can prove the following statement. (See [5, Cor.2].)

**Proposition 1.** Suppose  $p_M(a) > 0$  for all  $a \in \mathbf{A}$  and, for any two states  $i, j \in \mathbf{S}$ , there exists a path  $i_0 = i \rightarrow i_1 \rightarrow \dots \rightarrow i_N = j$  in  $\mathbf{S}$  such that, for any  $a_0, a_1, \dots, a_{N-1} \in \mathbf{A}$ ,

$$P_{i_0, i_1}(a_0)P_{i_1, i_2}(a_1)\dots P_{i_{N-1}, i_N}(a_{N-1}) > 0.$$

Then the conjecture (14) (and also (4) and (1)) is valid.

The matrix  $P(a)$  has a cyclic structure. Thus, the conditions of Proposition 1 are satisfied if there is  $m < M$  having no common divisors with  $M$  such that  $p_M(a) > 0$  and  $p_m(a) > 0$  for all  $a \in \mathbf{A}$ .

Let us briefly discuss the connection of the conjecture (14) and the Turnpike Theorem for the average reward MDP established in [5]. During the proof of Theorem 1, it was shown that

$$\lim_{k \rightarrow \infty} sp(W^{k+1} - W^k) = 0, \quad (15)$$

where  $sp(W) := \max_{s \in \mathbf{S}} W(s) - \min_{s \in \mathbf{S}} W(s)$  is the ‘span-seminorm’ and the vectors  $W^k$  come from the iterations (11). Condition (15) is sufficient for the Turnpike Theorem [5, Thm.1] which is strictly connected with the conjecture (14). Namely, under mild additional requirements that Turnpike Theorem implies the validity of the conjecture (14): see [5, Thm.3]. By the way, Proposition 1 also follows from the Turnpike Theorem [5, Thm.1]. One can show that in the example from Sect. 4  $\lim_{k \rightarrow \infty} sp(W^{k+1} - W^k) = 2 \left[ 1 - \frac{h-2}{\varepsilon} \right] > 0$ : see [5, §5.3].

Suppose Condition 1 is satisfied,  $K$  is as in the proof of Theorem 1,  $k \geq K$  is arbitrarily fixed and  $d \notin \mathbf{D}_*$ . Then, according to the proof of Theorem 1 (see (24) and (25)), for each  $s \in \mathbf{S}$  such that  $d(s) \notin \mathbf{A}_*$ ,

$$\mathcal{R}_s(d) + \mathcal{Q}_{s,.}(d)W^k < \max_{d \in \mathbf{D}_*} \{\mathcal{R}_s(d) + \mathcal{Q}_{s,.}(d)W^k\} = W^{k+1}(s).$$

Therefore, going back to (10), we again have the strict inequality

$$W^{k+1}(s) > \max_{a \in \mathbf{A} \setminus \mathbf{A}_*} \left\{ L^a \left( \frac{R^a}{L^a} - c_* \right) + \sum_{j=0}^{s-1} W^{k+1}(j)P_{s,j}(a) + \sum_{j=s}^{M-1} W^k(j)P_{s,j}(a) \right\}$$

and finally, going back to (3):

$$\text{for all } i = (k-1)M + s, \quad V(i) > \max_{a \in \mathbf{A} \setminus \mathbf{A}_*} \left\{ R^a + \sum_{m=1}^M V(i-m)p_m(a) \right\},$$

i.e., for the valid conjecture (4) we have the following:

$$\max_{a \in \mathbf{A}} \left\{ R^a + \sum_{m=1}^M V(i-m)p_m(a) \right\} \max_{a \in \mathbf{A} \setminus \mathbf{A}_*} \left\{ R^a + \sum_{m=1}^M V(i-m)p_m(a) \right\} \quad (16)$$

for all  $i \geq I := (K-1)M$ .

## 4 Counter-Example

In this subsection, we show that the conjecture (4) (and also (1) and (14)) may be not valid if the conditions formulated in Sect. 3 are not satisfied.

Put

$$\mathbf{A} := \{a_1, a_2\}, \quad M := 3, \quad \varepsilon \in (0, 1), \quad p_2(a_1) = 1, \quad p_2(a_2) = 1 - \varepsilon, \quad p_3(a_2) = \varepsilon,$$

where  $\varepsilon \in (0, 1)$ ; other probabilities being zero. Finally, let  $R^{a_1} := 2$  and  $R^{a_2} := h \in (2, 2 + \varepsilon)$ . Now

$$L^{a_1} = 2, \quad L^{a_2} = 2 + \varepsilon, \quad \frac{R^{a_1}}{L^{a_1}} = 1, \quad \frac{R^{a_2}}{L^{a_2}} = \frac{h}{2 + \varepsilon} < 1, \quad c_* = 1, \quad \mathbf{A}_* = \{a_1\}.$$

Below, we study the iterations (3).

Since  $h > 2$ , obvious calculations lead to the following expressions:

$$V(0) = V(1) = \max\{2, h\} = h;$$

$V(2) = \max\{2 + V(0) = 2 + h; \quad h + (1 - \varepsilon)V(0) + \varepsilon V(-1) = h + (1 - \varepsilon)h\} = 2 + h$  because

$$\frac{2}{1 - \varepsilon} = 2[1 + \varepsilon + \varepsilon^2 + \dots] > 2 + \varepsilon > h \implies 2 > (1 - \varepsilon)h.$$

$V(3) = \max\{2 + V(1) = 2 + h; \quad h + (1 - \varepsilon)V(1) + \varepsilon V(0) = 2h\} = 2h$ . Further properties of the function  $V$  are given in the following lemma.

**Lemma 3.** *For all  $j \geq 1$ , the following statements hold.*

(i) *For even steps  $i = 2j$ ,*

$$V(2j) = 2j + h,$$

*and maximum in (3) is provided by  $a_1$  only.*

(ii) *For odd steps  $i = 2j - 1$ ,*

$$V(2j - 1) < \frac{2\varepsilon(j - 1) + (1 + \varepsilon)h - 2}{\varepsilon}.$$

(iii) *For odd steps  $i = 2j + 1$ ,*

$$V(2j + 1) = (1 - \varepsilon)V(2j - 1) + (1 + \varepsilon)h + 2\varepsilon(j - 1),$$

*and maximum in (3) is provided by  $a_2$  only.*

Therefore, for all odd values of  $i$ , the maximum in (3) is provided only by  $a_2 \notin \mathbf{A}_*$ . The conjecture (4) (and also (1) and (14)) is not valid.

In this example,  $\mathbf{D}_* = \{d_*\}$  with  $d_*(s) \equiv a_* = a_1$ . The matrices  $P(a_*)$  and  $\mathcal{Q}(d_*)$  look as follows:

$$P(a_*) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}; \quad \mathcal{Q}(d_*) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

and are periodic;  $p_M(a_*) = 0$ . Thus, Theorem 1, Corollary 1 and Proposition 1 are not applicable.

## 5 Discounted Model

In this section,  $\beta \in (0, 1)$  is the discount factor, expression (2) is replaced with

$$V^\beta(i) := \sup_{\pi} E_i^\pi \left[ \sum_{t=1}^{\infty} \beta^{t-1} r_{X_{t-1}}(A_t) \right], \quad (17)$$

and the optimality equation looks like

$$\begin{aligned} V^\beta(i) &= \max_{a \in \mathbf{A}} \left\{ R^a + \beta \sum_{m=1}^M V^\beta(i-m)p_m(a) \right\} \quad \text{for } i \geq 0; \\ V^\beta(i) &= 0 \quad \text{for } i = -M, -M+1, \dots, -1. \end{aligned} \quad (18)$$

Like previously, it can be solved successively for  $i = 0, 1, \dots$ . We put  $R^* := \max_{a \in \mathbf{A}} R^a$  and

$$\mathbf{A}^* := \{a \in \mathbf{A} : R^a = R^*\}. \quad (19)$$

The so-called turnpike theory in discounted models (see [6, §6.8]), [7]) leads to the following statement (cf (4)):

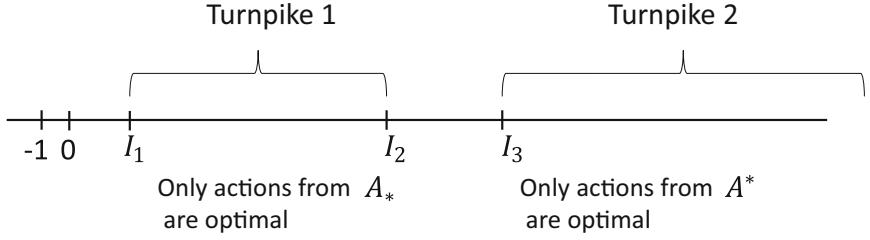
**Theorem 2.** *There is such  $I < \infty$  that, for all  $i \geq I$ , the maximum in (18) is only provided by  $a \in \mathbf{A}^*$ .*

Some recent developments of the turnpike theory for discounted MDPs can be found in [3].

It is interesting to look at what happens if the discount factor  $\beta$  is close to 1, assuming that the Condition 1 is satisfied (more generally, assuming the conjecture (4) to be valid for  $\beta = 1$ ). Denote the corresponding  $I$  as  $I_1$ , i.e.,  $I_1 = (K-1)M$  with  $K$  as in the proof of Theorem 1 (see the end of Sect. 3), and fix an arbitrary  $I_2 > I_1$ . Obviously, for each  $i \in \mathbf{X}$ ,  $\lim_{\beta \rightarrow 1^-} V^\beta(i) = V(i)$  with  $V$  as in (3). According to (16), there is  $\beta_0 \in (0, 1)$  such that, for all  $i = I_1, I_1+1, \dots, I_2$ , for all  $\beta \in [\beta_0, 1]$

$$\max_{a \in \mathbf{A}} \left\{ R^a + \sum_{m=1}^M V^\beta(i-m)p_m(a) \right\} > \max_{a \in \mathbf{A} \setminus \mathbf{A}^*} \left\{ R^a + \sum_{m=1}^M V^\beta(i-m)p_m(a) \right\}.$$

Thus, for a fixed  $\beta \in [\beta_0, 1]$ , for all  $i = I_1, I_1+1, \dots, I_2$ , the maximum in (18) is only provided by  $a \in \mathbf{A}_*$ . Of course, by Theorem 2, there is a finite  $I_3 > I_2$  such that, for all  $i \geq I_3$ , the maximum in (18) is only provided by  $a \in \mathbf{A}^*$ . Recall that  $\mathbf{A}_*$  and  $\mathbf{A}^*$  are given by (1) and (19). One can say that, for  $\beta$  close to 1, there are two turnpikes, where only actions from  $\mathbf{A}_*$  and  $\mathbf{A}^*$  are optimal in the model (17); see Fig. 2. When  $\beta$  approaches 1,  $I_2$  and  $I_3$  go to infinity, and in the limiting case  $\beta = 1$  we have just the conjecture (4).



**Fig. 2.** Turnpikes for  $\beta \approx 1$ .

## 6 Summary

In this article, we studied the conjecture (1) (equivalent to (4) and (14)) and showed that in general it is not valid. In the discounted case, Turnpike Theorem 2 always holds. Under appropriate conditions, when the discount factor is close to 1, there are two turnpikes.

## Appendix

*Proof of Lemma 1.* The case of  $i = -M, -M + 1, \dots, -1$  is obvious.

For  $i \geq 0$ ,

$$\begin{aligned}\tilde{W}(i) &= \max_{a \in \mathbf{A}} \left\{ R^a + \sum_{m=1}^M [\tilde{W}(i-m) + c_*(i-m)] p_m(a) \right\} - c_* i \\ &= \max_{a \in \mathbf{A}} \left\{ R^a - c_* L^a + \sum_{m=1}^M \tilde{W}(i-m) p_m(a) \right\}.\end{aligned}$$

Equalities (6) are proved, and the maxima in (4) and in (6) are provided by the same values of  $a \in \mathbf{A}$ .

Finally, keeping in mind that

- $|\tilde{W}(i)| \leq |c_*| M$  for  $i < 0$ ,
- $\frac{R^a}{L^a} - c_* \leq 0$  for all  $a \in \mathbf{A}$ , and
- $\frac{R^a}{L^a} - c_* = 0$  for  $a \in \mathbf{A}_* \neq \emptyset$ ,

it is easy to prove by induction that  $|\tilde{W}(i)| \leq |c_*| M$  for all  $i = 0, 1, 2, \dots$   $\square$

*Proof of Lemma 2.* (a) All the elements of the matrix  $\mathcal{Q}(d)$  are obviously non-negative.

Clearly,

$$\sum_{j=0}^{M-1} \mathcal{Q}_{0,j}(d) = \sum_{j=0}^{M-1} P_{0,j}(d(0)) = 1$$

Suppose  $\sum_{j=0}^{M-1} \mathcal{Q}_{i,j}(d) = 1$  for all  $i \leq l$  for some  $l \in \{0, 1, \dots, M-2\}$  and consider  $l+1$ :

$$\begin{aligned} \sum_{j=0}^{M-1} \mathcal{Q}_{l+1,j}(d) &= \sum_{j=0}^l \sum_{i=0}^l [P_{l+1,i}(d(l+1)) \mathcal{Q}_{i,j}(d)] \\ &\quad + \sum_{j=l+1}^{M-1} \left( \sum_{i=0}^l [P_{l+1,i}(d(l+1)) \mathcal{Q}_{i,j}(d)] + P_{l+1,j}(d(l+1)) \right) \\ &= \sum_{i=0}^l \left( \sum_{j=0}^{M-1} \mathcal{Q}_{i,j}(d) \right) P_{l+1,i}(d(l+1)) + \sum_{j=l+1}^{M-1} P_{l+1,j}(d(l+1)) \\ &= \sum_{i=0}^{M-1} P_{l+1,i}(d(l+1)) = 1. \end{aligned}$$

The last equality is by the induction supposition.

(b) If  $l \geq s$  then this statement follows directly from the definition (13):  $\mathcal{Q}_{s,l}(d) \geq P_{s,l}(d(s))$ .

Suppose  $l < s$ . Then, again using (13), we have  $\mathcal{Q}_{s,l}(d) \geq P_{s,l}(d(s)) \mathcal{Q}_{l,l}(d)$ . Since  $\mathcal{Q}_{l,l}(d) \geq P_{l,l}(d(l)) > 0$ , we finally obtain that

$$\mathcal{Q}_{s,l}(d) > 0, \quad \text{if } P_{s,l}(d(s)) > 0.$$

□

For the proof of Theorem 1 we need the following lemma.

**Lemma 4.** Suppose  $\vec{\alpha} \in \mathbb{R}^M$  is a substochastic row vector and  $\sum_{i=0}^{M-1} \alpha_i \mathcal{R}_i(d) = 0$  for some  $d \in \mathbf{D}$ . Then the row vector  $\vec{\alpha} \mathcal{Q}(d)$  coincides with the row vector  $\vec{\alpha} \mathcal{Q}(\hat{d})$  for some  $\hat{d} \in \mathbf{D}_*$  with  $\hat{d}(i) = d(i)$  if  $\alpha_i > 0$ .

*Proof.* Suppose  $\alpha_0 \in [0, 1]$  and consider the substochastic row vector  $\vec{\alpha} := (\alpha_0, 0, \dots, 0) \in \mathbb{R}^M$  such that  $\sum_{i=0}^{M-1} \alpha_i \mathcal{R}_i(d) = 0$ , where  $d \in \mathbf{D}$  is fixed. Then the row vector  $\vec{\alpha} \mathcal{Q}(d)$  coincides with the row vector  $\vec{\alpha} \mathcal{Q}(\hat{d})$  for some  $\hat{d} \in \mathbf{D}_*$  with  $\hat{d}(0) = d(0)$  if  $\alpha_0 > 0$ . Indeed, for  $\alpha_0 > 0$ ,  $\mathcal{R}_0(d) = 0$ , and we put  $\hat{d}(0) := d(0) \in \mathbf{A}_*$ . The other values  $\hat{d}(i) \in \mathbf{A}_*$  for  $i = 1, 2, \dots, M-1$  can be taken arbitrarily leading to equalities

$$\vec{\alpha} \mathcal{Q}(\hat{d}) = \left( \alpha_0 P_{0,j}(\hat{d}(0)) \right)_{j=0}^{M-1} = (\alpha_0 P_{0,j}(d(0)))_{j=0}^{M-1} = \vec{\alpha} \mathcal{Q}(d).$$

If  $\alpha_0 = 0$  then one can take an arbitrary  $\hat{d} \in \mathbf{D}_*$ :

$$\vec{\alpha} \mathcal{Q}(\hat{d}) = \vec{\alpha} \mathcal{Q}(d) = \mathbf{0}.$$

We proceed further by induction. Suppose the statement of the lemma is valid for all  $\vec{\alpha}$  satisfying condition  $\alpha_l = \alpha_{l+1} = \dots = \alpha_{M-1} = 0$  for some

$1 \leq l \leq M - 1$  and  $\hat{d}(l), \hat{d}(l+1), \dots, \hat{d}(M-1)$  can be arbitrary in  $\mathbf{A}_*$ . Consider a vector  $\vec{\alpha}$  with  $\alpha_{l+1} = \alpha_{l+2} = \dots = \alpha_{M-1} = 0$  and let  $d \in \mathbf{D}$  be such that  $\sum_{i=0}^{M-1} \alpha_i \mathcal{R}_i(d) = 0$ .

Since, for each fixed  $\tilde{d} \in \mathbf{D}$ , the product  $\vec{\alpha} \cdot \vec{\mathcal{Q}}(\tilde{d})$  does not depend on the rows  $\mathcal{Q}_{j,\cdot}(\tilde{d})$  with  $j \geq l + 1$ , the values  $\tilde{d}(j)$  with such  $j$  do not affect the value of  $\vec{\alpha} \cdot \vec{\mathcal{Q}}(\tilde{d})$ . (See (13)) Hence one can put  $\hat{d}(l+1), \hat{d}(l+2), \dots, \hat{d}(M-1) \in \mathbf{A}_*$  arbitrarily.

In case  $\alpha_l = 0$  the statement of the lemma holds by the induction supposition. Below,  $\alpha_l > 0$  and hence  $\mathcal{R}_l(d) = 0$  and  $d(l) \in \mathbf{A}_*$ . Therefore, we put  $\hat{d}(l) := d(l)$ . Moreover, in the current situation  $\mathcal{R}_j(d) = 0$  for all  $j \in \{0, 1, \dots, l-1\}$  with positive values of  $P_{l,j}(d(l))$ : see (12). Thus, for the row vector

$$\vec{\alpha}' := (\alpha_0 + \alpha_l P_{l,0}(d(l)), \dots, \alpha_{l-1} + \alpha_l P_{l,l-1}(d(l)), 0, \dots, 0) \in \mathbb{R}^M$$

we have  $\sum_{i=0}^{M-1} \alpha'_i \mathcal{R}_i(d) = 0$ , and we will use the induction supposition for  $\vec{\alpha}'$  to complete the proof. Note also that the vector  $\vec{\alpha}'$  is substochastic.

For any  $\tilde{d} \in \mathbf{D}$ , according to (13), the elements of the row vector  $\vec{\alpha}' \cdot \vec{\mathcal{Q}}(\tilde{d})$  are as follows:

$$\begin{aligned} \sum_{j=0}^l \alpha_j \mathcal{Q}_{j,0}(\tilde{d}) &= \sum_{j=0}^{l-1} \alpha_j \mathcal{Q}_{j,0}(\tilde{d}) + \alpha_l \sum_{i=0}^{l-1} [P_{l,i}(\tilde{d}(l)) \mathcal{Q}_{i,0}(\tilde{d})] \\ &= \sum_{j=0}^{l-1} \{\alpha_j + \alpha_l P_{l,j}(\tilde{d}(l))\} \mathcal{Q}_{j,0}(\tilde{d}); \\ \sum_{j=0}^l \alpha_l \mathcal{Q}_{j,1}(\tilde{d}) &= \sum_{j=0}^{l-1} \{\alpha_j + \alpha_l P_{l,j}(\tilde{d}(l))\} \mathcal{Q}_{j,1}(\tilde{d}); \\ &\quad \dots \quad \dots \quad \dots \\ \sum_{j=0}^l \alpha_l \mathcal{Q}_{j,l-1}(\tilde{d}) &= \sum_{j=0}^{l-1} \{\alpha_j + \alpha_l P_{l,j}(\tilde{d}(l))\} \mathcal{Q}_{j,l-1}(\tilde{d}); \\ \sum_{j=0}^l \alpha_l \mathcal{Q}_{j,l}(\tilde{d}) &= \sum_{j=0}^{l-1} \{\alpha_j + \alpha_l P_{l,j}(\tilde{d}(l))\} \mathcal{Q}_{j,l}(\tilde{d}) + \alpha_l P_{l,l}(\tilde{d}(l)); \\ &\quad \dots \quad \dots \quad \dots \\ \sum_{j=0}^l \alpha_l \mathcal{Q}_{j,M-1}(\tilde{d}) &= \sum_{j=0}^{l-1} \{\alpha_j + \alpha_l P_{l,j}(\tilde{d}(l))\} \mathcal{Q}_{j,M-1}(\tilde{d}) + \alpha_l P_{l,M-1}(\tilde{d}(l)). \end{aligned}$$

To put it differently, for the mapping  $d$  we have

$$\vec{\alpha} \cdot \vec{\mathcal{Q}}(d) = \vec{\alpha}' \cdot \vec{\mathcal{Q}}(d) + (0, \dots, 0, \alpha_l P_{l,l}(d(l)), \dots, \alpha_l P_{l,M-1}(d(l))).$$

According to the induction supposition, there is  $\hat{d} \in \mathbf{D}_*$  (with fixed  $\hat{d}(l) = d(l) \in \mathbf{A}_*$  and arbitrary values  $\hat{d}(l+1), \dots, \hat{d}(M-1) \in \mathbf{A}_*$  which do not appear in the provided expressions) such that

$$\overset{\rightarrow}{\alpha'} \mathcal{Q}(d) = \overset{\rightarrow}{\alpha'} \mathcal{Q}(\hat{d}).$$

Therefore,

$$\overset{\rightarrow}{\alpha} \mathcal{Q}(d) = \overset{\rightarrow}{\alpha} \mathcal{Q}(\hat{d}).$$

Besides, for  $i = 0, 1, \dots, l - 1$ , if  $\alpha'_i > 0$  then  $\hat{d}(i) = d(i)$ ; hence, if  $\alpha_i > 0$  then  $\hat{d}(i) = d(i)$  for all  $i = 0, 1, \dots, l - 1, l$ .

The proof is completed.  $\square$

*Proof of Theorem 1.* Let  $T_k := \max_{s \in \mathbf{S}} W^k(s)$  and  $t_k := \min_{s \in \mathbf{S}} W^k(s)$  and let us show that the sequence  $\{T_k\}_{k=0}^\infty$  decreases and the sequence  $\{t_k\}_{k=0}^\infty$  increases.

Since, for every  $d \in \mathbf{D}$ ,  $\mathcal{R}(d) \leq 0$  and the matrix  $\mathcal{Q}(d)$  is stochastic (see Lemma 2(a)),

$$\{\mathcal{R}(d) + \mathcal{Q}(d)W^k\}_s \leq T_k \quad \text{for all } s \in \mathbf{S}.$$

Hence,  $T_{k+1} \leq T_k$ .

For every  $d \in \mathbf{D}_*$ ,  $\mathcal{R}(d) = 0$ , so

$$\{\mathcal{R}(d) + \mathcal{Q}(d)W^k\}_s \geq t_k \quad \text{for all } s \in \mathbf{S}.$$

Hence,  $W^{k+1}(s) \geq t_k$  and  $t_{k+1} \geq t_k$ .

Therefore, there exist limits  $T^\infty := \lim_{k \rightarrow \infty} T_k \geq t^\infty := \lim_{k \rightarrow \infty} t_k$  which are finite because of Lemma 1. Later, under the imposed condition, it will be clear that these limits coincide.

Let

$$\tilde{\Delta} := - \max_{d \in \mathbf{D}, s \in \mathbf{S}: \mathcal{R}_s(d) < 0} \mathcal{R}_s(d) \quad \text{and} \quad q := \min_{d \in \mathbf{D}, s, l \in \mathbf{S}: \mathcal{Q}_{s,l}(d) > 0} \mathcal{Q}_{s,l}(d).$$

Since the sets  $\mathbf{D}$  and  $\mathbf{S}$  are finite,  $\tilde{\Delta} > 0$ ,  $q \in (0, 1]$ , and we introduce an arbitrary  $\Delta \in (0, \tilde{\Delta})$  and

$$\varepsilon := \frac{\Delta}{2} \left( \frac{q}{2 - q} \right)^J > 0.$$

Let  $K$  be such that

$$T_K < T^\infty + \varepsilon \quad (\text{hence } T_K < T^\infty + \frac{\Delta}{2}).$$

Since the sequence  $\{T_k\}_{k=0}^\infty$  decreases to  $T^\infty$ , we conclude that

$$T_k < T^\infty + \varepsilon < T^\infty + \frac{\Delta}{2} \quad \text{for all } k \geq K. \quad (20)$$

We intend to show that

$$\forall s \in \mathbf{S} \quad W^K(s) > T^\infty - \frac{\Delta}{2}. \quad (21)$$

To do this, fix  $\tilde{s} \in \mathbf{S}$  such that  $W^{K+J}(\tilde{s}) = T_{K+J}$  and let us prove by induction the following statement

**A.** For each  $j = 0, 1, \dots, J$ , there exist mappings  $\hat{d}_1, \hat{d}_2, \dots, \hat{d}_j \in \mathbf{D}_*$  such that, for the  $\tilde{s}$ -th row of the stochastic matrix  $\mathcal{Q}(\hat{d}_1)\mathcal{Q}(\hat{d}_2) \dots \mathcal{Q}(\hat{d}_j)$  denoted below as  $\vec{\gamma}^j$ , if  $\gamma_s^j > 0$ , then  $W^{K+J-j}(s) > T^\infty - \left(\frac{2-q}{q}\right)^j \varepsilon$ .

Let  $j = 0$ . Then no mappings  $\hat{d} \in \mathbf{D}_*$  are considered,  $\vec{\gamma}^0 = (\delta_{\tilde{s}, s})_{s=0}^{M-1}$  is the basic row vector with element 1 on the  $\tilde{s}$ -th place, and  $\vec{W}^{K+J}(\tilde{s}) > T^\infty - \varepsilon$  because

- $W^{K+J}(\tilde{s}) = T_{K+J}$  by the definition of  $\tilde{s}$ ;
- and  $T_{K+J} \geq T^\infty$  because the sequence  $\{T_k\}_{k=0}^\infty$  decreases to  $T^\infty$ .

Suppose the statement **A** is valid for some  $j \in \{0, 1, \dots, J-1\}$  and the mappings  $\hat{d}_1, \hat{d}_2, \dots, \hat{d}_j \in \mathbf{D}_*$  are fixed;

$$\vec{\gamma}^j = \vec{\gamma}^0 \mathcal{Q}(\hat{d}_1)\mathcal{Q}(\hat{d}_2) \dots \mathcal{Q}(\hat{d}_j),$$

where  $\vec{\gamma}^0 = (\delta_{\tilde{s}, s})_{s=0}^{M-1}$  as before. Let  $d_{j+1} \in \mathbf{D}$  be such that

$$W^{K+J-j} = \mathcal{R}(d_{j+1}) + \mathcal{Q}(d_{j+1})W^{K+J-(j+1)}.$$

For  $s \in \mathbf{S}$  such that  $\gamma_s^j > 0$  we have inequality

$$W^{K+J-j}(s) > T^\infty - \left(\frac{2-q}{q}\right)^j \varepsilon \quad (22)$$

according to the inductive supposition. For such value of  $s$ , suppose  $\mathcal{R}_s(d_{j+1}) < 0$ . Then

$$W^{K+J-j}(s) \leq -\Delta + T_{K+J-(j+1)} < -\Delta + T^\infty + \frac{\Delta}{2} = T^\infty - \frac{\Delta}{2}$$

because of (20): remember,  $J - (j + 1) \geq 0$ . Further,

$$W^{K+J-j}(s) < T^\infty - \left(\frac{2-q}{q}\right)^j \varepsilon \leq T^\infty - \left(\frac{2-q}{q}\right)^j \varepsilon$$

which contradicts (22). Therefore, if  $\gamma_s^j > 0$  then  $\mathcal{R}_s(d_{j+1}) = 0$  and  $\sum_{s=0}^{M-1} \gamma_s^j \mathcal{R}_s(d_{j+1}) = 0$ .

Using Lemma 4 with the (sub)stochastic vector  $\vec{\gamma}^j$ , we conclude that the row vector  $\vec{\gamma}^j \mathcal{Q}(d_{j+1})$  coincides with the row vector  $\vec{\gamma}^j \mathcal{Q}(\hat{d}_{j+1})$  for some  $\hat{d}_{j+1} \in \mathbf{D}_*$ . Now the  $\tilde{s}$ -th row of the matrix  $\mathcal{Q}(\hat{d}_1)\mathcal{Q}(\hat{d}_2) \dots \mathcal{Q}(\hat{d}_{j+1})$  equals

$$\vec{\gamma}^{j+1} = \vec{\gamma}^j \mathcal{Q}(d_{j+1}) = \vec{\gamma}^j \mathcal{Q}(d_{j+1}).$$

Suppose  $\gamma_l^{j+1} > 0$ . Then there is at least one index  $s$  such that  $\gamma_s^j > 0$  and  $\mathcal{Q}_{s,l}(d_{j+1}) > 0$ ; hence  $\mathcal{Q}_{s,l}(d_{j+1}) \geq q$ . As was proved above,  $\mathcal{R}_s(d_{j+1}) = 0$ . Consider equality

$$W^{K+J-j}(s) = \mathcal{R}_s(d_{j+1}) + \mathcal{Q}_{s,.}(d_{j+1})W^{K+J-(j+1)} = \mathcal{Q}_{s,.}(d_{j+1})W^{K+J-(j+1)}.$$

Since  $T_{K+J-(j+1)} < T^\infty + \varepsilon \leq T^\infty + \left(\frac{2-q}{q}\right)^j \varepsilon$  (see (20)),

$$W^{K+J-j}(s) \leq \mathcal{Q}_{s,l}(d_{j+1}) W^{K+J-(j+1)}(l) + (1 - \mathcal{Q}_{s,l}(d_{j+1})) \left[ T^\infty + \left(\frac{2-q}{q}\right)^j \varepsilon \right].$$

In case

$$W^{K+J-(j+1)}(l) \leq T^\infty - \left(\frac{2-q}{q}\right)^{j+1} \varepsilon$$

we have

$$\begin{aligned} W^{K+J-j}(s) &\leq \mathcal{Q}_{s,l}(d_{j+1}) \left[ T^\infty - \left(\frac{2-q}{q}\right)^{j+1} \varepsilon \right] \\ &\quad + (1 - \mathcal{Q}_{s,l}(d_{j+1})) \left[ T^\infty + \left(\frac{2-q}{q}\right)^j \varepsilon \right] \\ &= T^\infty + \left(\frac{2-q}{q}\right)^j \varepsilon - \mathcal{Q}_{s,l}(d_{j+1}) \left[ \left(\frac{2-q}{q}\right)^j \varepsilon + \left(\frac{2-q}{q}\right)^{j+1} \varepsilon \right] \\ &\leq T^\infty + \left(\frac{2-q}{q}\right)^j \varepsilon - q \left(\frac{2-q}{q}\right)^j \varepsilon \left[ 1 + \frac{2-q}{q} \right] \\ &\quad (\text{because } \mathcal{Q}_{s,l}(d_{j+1}) \geq q) \\ &= T^\infty + \left(\frac{2-q}{q}\right)^j \varepsilon [1 - q - (2-q)] = T^\infty - \left(\frac{2-q}{q}\right)^j \varepsilon \end{aligned}$$

which contradicts (22). Therefore

$$W^{K+J-(j+1)}(l) > T^\infty - \left(\frac{2-q}{q}\right)^{j+1} \varepsilon,$$

and the statement **A** is proved for  $j+1$ .

When  $j = J$ , the vector  $\vec{\gamma}^J$  contains no zeroes; hence

$$W^K(s) > T^\infty - \left(\frac{2-q}{q}\right)^J \varepsilon = T^\infty - \frac{\Delta}{2} \quad \text{for all } s \in \mathbf{S}$$

by the definition of  $\varepsilon$ . Inequality (21) is proved.

Note also that inequality (21) implies that  $t_K > T^\infty - \frac{\Delta}{2}$  and, since the sequence  $\{t_k\}_{k=0}^\infty$  increases to  $t^\infty$ ,  $t^\infty \geq T^\infty - \frac{\Delta}{2}$ . Noting that  $\Delta \in (0, \tilde{\Delta})$  was arbitrary, we conclude that  $t^\infty = T^\infty$ .

Now, if  $d \notin \mathbf{D}_*$  then, for some  $s \in \mathbf{S}$ ,

$$\mathcal{R}_s(d) + \mathcal{Q}_{s,\cdot}(d) W^K < -\Delta + T^\infty + \frac{\Delta}{2} = T^\infty - \frac{\Delta}{2}$$

according to (20). On the other hand, for each  $d \in \mathbf{D}_*$ , for all  $s \in \mathbf{S}$ ,  $\mathcal{Q}_{s,\cdot}(d) W^K > T^\infty - \frac{\Delta}{2}$  because of (21), meaning that

$$W^{K+1} = \max_{d \in \mathbf{D}} \{\mathcal{R}(d) + \mathcal{Q}(d)W^K\} = \mathcal{R}(d_*^{K+1}) + \mathcal{Q}(d_*^{K+1})W^K = \mathcal{Q}(d_*^{K+1})W^K$$

with  $d_*^{K+1} \in \mathbf{D}_*$ .

The following statement can be easily proved by induction.

**B.** For each  $k \geq K$

$$W^k(s) > T^\infty - \frac{\Delta}{2} \quad \text{for all } s \in \mathbf{S}$$

and the maximum in equation

$$W^{k+1} = \max_{d \in \mathbf{D}} \{\mathcal{R}(d) + \mathcal{Q}(d)W^k\} \quad (23)$$

is provided necessarily by  $d_*^{k+1} \in \mathbf{D}_*$ .

As was shown above, this statement is valid for  $k = K$ .

Suppose it is valid for some  $k - 1 \geq K$  and consider the case of  $k$ . Firstly, the vector

$$W^k = \mathcal{Q}(d_*^k)W^{k-1}$$

is such that (for all  $s \in \mathbf{S}$ )  $W^k(s) > T^\infty - \frac{\Delta}{2}$  because of the inductive supposition:  $W^{k-1}(s) > T^\infty - \frac{\Delta}{2}$  for all  $s \in \mathbf{S}$ . Secondly, like previously, if  $d \notin \mathbf{D}_*$ , then, for some  $s \in \mathbf{S}$ ,

$$\mathcal{R}_s(d) + \mathcal{Q}_{s,\cdot}(d)W^k < -\Delta + T^\infty + \frac{\Delta}{2} = T^\infty - \frac{\Delta}{2}. \quad (24)$$

(This inequality holds for each  $s \in \mathbf{S}$  with  $d(s) \notin \mathbf{A}_* \implies \mathcal{R}_s(d) < 0 \implies \mathcal{R}_s(d) < -\Delta$ .) And, for each  $d \in \mathbf{D}_*$ , for all  $s \in \mathbf{S}$ ,

$$\mathcal{R}_s(d) + \mathcal{Q}_{s,\cdot}(d)W^k = \mathcal{Q}_{s,\cdot}(d)W^k > T^\infty - \frac{\Delta}{2}. \quad (25)$$

Thus the maximum in (23) at  $k$  is provided necessarily by  $d_*^{k+1} \in \mathbf{D}_*$ .

The proof is completed.  $\square$

*Proof of Corollary 1.* It is sufficient to show that the matrix  $\mathcal{Q}(d_*)$  is ergodic. The mapping  $d_*$  satisfies the condition of Lemma 2(b):  $P_{s,s}(d_*(s)) = P_{s,s}(a_*) = p_M(a_*) > 0$  for all  $s \in \mathbf{S}$ . Hence, for all  $s, l \in \mathbf{S}$ , if  $P_{s,l}(d_*(i)) = P_{s,l}(a_*) > 0$  then  $\mathcal{Q}_{s,l}(d_*) > 0$ . Therefore, the matrix  $\mathcal{Q}(d_*)$  is ergodic because the matrix  $P(a_*)$  is ergodic. The proof is completed.  $\square$

*Proof of Lemma 3.* When  $j = 1$ , Items (i) and (iii) are valid by the preliminary calculations, and Item (ii) comes from the following:

$$2\varepsilon(j-1) + (1+\varepsilon)h - 2 - \varepsilon V(2j-1) = (1+\varepsilon)h - 2 - \varepsilon h = h - 2 > 0.$$

Suppose statements (i), (ii) and (iii) hold for some  $j \geq 1$  and consider  $j+1$ .

(i) For  $i = 2(j+1)$ , using the induction supposition, we estimate the difference

$$\begin{aligned} & 2 + V(2j) - [h + (1 - \varepsilon)V(2j) + \varepsilon V(2j - 1)] \\ & > 2 + 2j + h - h - (1 - \varepsilon)(2j + h) - \varepsilon \frac{2\varepsilon(j - 1) + (1 + \varepsilon)h - 2}{\varepsilon} \\ & = 4 + 2\varepsilon - 2h = 2[2 + \varepsilon - h] > 0. \end{aligned}$$

The inequality is according to statement (ii) at  $j$ .

Thus,  $V(2(j+1)) = 2(j+1) + h$ , and the maximum in (3) is provided only by  $a_1$ .

(ii)

$$\begin{aligned} V(2j + 1) &= (1 - \varepsilon)V(2j - 1) + (1 + \varepsilon)h + 2\varepsilon(j - 1) \\ &< (1 - \varepsilon) \frac{2\varepsilon(j - 1) + (1 + \varepsilon)h - 2}{\varepsilon} + (1 + \varepsilon)h + 2\varepsilon(j - 1) \\ &= \frac{2\varepsilon j + (1 + \varepsilon)h - 2}{\varepsilon}, \end{aligned}$$

so that statement (ii) is valid for  $j + 1$ .

(iii) For  $i = 2(j+1) + 1$ , using the induction supposition, we estimate the difference

$$\begin{aligned} & h + (1 - \varepsilon)V(2j + 1) + \varepsilon V(2j) - [2 + V(2j + 1)] \\ & = h - \varepsilon V(2j + 1) + \varepsilon[2j + h] - 2 \\ & > h(1 + \varepsilon) - [2\varepsilon j + h - 2 + \varepsilon h] + 2\varepsilon j - 2 = 0, \end{aligned}$$

where the inequality is by the proved above Item (ii) for  $j + 1$ . Recall also that  $V(2j) = 2j + h$ . Therefore,

$$V(2(j+1) + 1) = h + (1 - \varepsilon)V(2j + 1) + \varepsilon[2j + h],$$

and we see that statement (iii) is valid for  $j + 1$  and the maximum in (3) is provided only by  $a_2$ .  $\square$

*Proof of Theorem 2.* Introduce function

$$\tilde{W}(i) := V^\beta(i) - \frac{R^*}{1 - \beta}, \quad i \in \mathbf{X},$$

which obviously satisfies equation

$$\begin{aligned} \tilde{W}(i) &= -\frac{R^*}{1 - \beta} \quad \text{for } i = -M, -M + 1, \dots, -1; \\ \tilde{W}(i) &= \max_{a \in \mathbf{A}} \left\{ (R^a - R^*) + \beta \sum_{m=1}^M \tilde{W}(i - m)p_m(a) \right\} \quad \text{for } i \geq 0. \end{aligned}$$

Like previously, (see (7)), we replace the argument  $i$  with  $k = 0, 1, 2, \dots$  and  $s \in \mathbf{S} = \{0, 1, \dots, M - 1\}$ , denote  $W^k(s) := \tilde{W}(i)$  and finish with equations

like (8). The only difference is that  $p_m(a)$  is replaced by  $\beta p_m(a)$ , and the initial condition is

$$W^0(s) = -\frac{R^*}{1-\beta} \quad \text{for } s \in \mathbf{S}.$$

We obtain iterations (cf (11))

$$\begin{aligned} W^0(s) &= -\frac{R^*}{1-\beta}, \quad s \in \mathbf{S}; \\ W^{k+1} &= \max_{d \in \mathbf{D}} \{\mathcal{R}^\beta(d) + \mathcal{Q}^\beta(d)W^k\}, \quad k = 0, 1, \dots, \end{aligned} \quad (26)$$

where

$$\begin{aligned} \mathcal{R}_0^\beta(d) &= R^{d(0)} - R^*; \\ \mathcal{R}_{l+1}^\beta(d) &= R^{d(l+1)} - R^* + \beta \sum_{j=0}^l P_{l+1,j}(d(l+1)) \mathcal{R}_j^\beta(d), \\ l &= 0, \dots, M-2, \end{aligned}$$

and the matrix  $\mathcal{Q}^\beta$  is given by (13) with  $P$  being replaced by  $\beta P$ . Similarly to (11), the maximum in the expression

$$U \circ W := \max_{d \in \mathbf{D}} \{\mathcal{R}^\beta(d) + \mathcal{Q}^\beta(d)W\}$$

is provided by some  $\hat{d} \in \mathbf{D}$  for each  $W \in \mathbb{R}^M$ : the values  $\hat{d}(s)$  can be calculated successively for  $s = 0, 1, \dots, M-1$ . Note also that  $\mathcal{R}^\beta(d) \leq 0$  for all  $d \in \mathbf{D}$ .

The matrix  $\mathcal{Q}^\beta(d)$  is (uniformly with respect to  $d$ ) strictly substochastic, i.e.,

$$0 < \sum_{j=0}^{M-1} \mathcal{Q}_{l,j}^\beta(d) \leq \beta < 1 \quad \text{for all } l \in \mathbf{S}:$$

the proof is identical to the proof of Lemma 2(a). Therefore, the mapping  $U$  is a contraction in the space  $\mathbb{R}^M$  with the uniform norm: see the proof of Proposition 6.2.4 in [6]. The maximum  $\max_{d \in \mathbf{D}} \mathcal{R}^\beta(d) = \mathbf{0}$  (the zero vector in  $\mathbb{R}^M$ ) is provided by those and only those  $d \in \mathbf{D}$ , for which  $d(s) \in \mathbf{A}^*$  for all  $s \in \mathbf{S}$ . Therefore, the unique fixed point of the operator  $U$  is  $W^\infty = \mathbf{0}$  and  $\lim_{k \rightarrow \infty} W^k = \mathbf{0}$ . Below,

$$\mathbf{D}^* := \{d \in \mathbf{D} : d(s) \in \mathbf{A}^* \text{ for all } s \in \mathbf{S}\},$$

and

$$U \circ W^\infty = \mathcal{R}^\beta(d) + \mathcal{Q}^\beta(d)W^\infty = \mathcal{R}^\beta(d) = W^\infty = \mathbf{0}$$

if and only if  $d \in \mathbf{D}^*$ . The theorem will be proved if we show that, for some  $K < \infty$ , the maximum in (26) at all  $k \geq K$  is only provided by  $d \in \mathbf{D}^*$ .

Denote

$$\Delta := \min_{d \in \mathbf{D} \setminus \mathbf{D}^*} \min_{s \in \mathbf{S}: \mathcal{R}_s^\beta(d) < 0} \{-\mathcal{R}_s^\beta(d)\}.$$

The spaces  $\mathbf{D}$  and  $\mathbf{S}$  are finite, and  $\Delta > 0$ .

If  $\Delta = +\infty$  then  $\mathbf{D}^* = \mathbf{D}$  and the proof is finished. (One can put  $K = 0$ .)

Suppose  $\Delta < +\infty$ . Then, for each  $d \in \mathbf{D} \setminus \mathbf{D}^* \neq \emptyset$ , for each  $s \in \mathbf{S}$  such that  $\mathcal{R}_s^\beta(d) < 0$ ,

$$\mathcal{R}_s^\beta(d) \leq -\Delta.$$

Let us choose  $0 < \varepsilon < \frac{\Delta}{2}$  and fix  $K \geq 0$  such that

$$\max_{j \in \mathbf{S}} |W^k(j)| < \varepsilon \quad \text{for all } k \geq K.$$

Now, for each  $k \geq K$ , if  $d \notin \mathbf{D}^*$  provides the maximum in (26), then there is  $s \in \mathbf{S}$  such that  $\mathcal{R}_s^\beta(d) < 0$ , and, for each such  $s$ ,

$$W^{k+1}(s) = \mathcal{R}_s^\beta(d) + \sum_{j=0}^{M-1} \mathcal{Q}_{s,j}^\beta(d) W^k(j) \leq \mathcal{R}_s^\beta(d) + \varepsilon \leq -\Delta + \varepsilon.$$

(Recall that  $\mathcal{Q}^\beta(d)$  is a substochastic matrix.) Since  $W^{k+1}(s) > -\varepsilon$ , we obtain the strict inequalities

$$W^{k+1}(s) < W^{k+1}(s) + \varepsilon - \Delta + \varepsilon < W^{k+1}(s).$$

The obtained contradiction shows that, for all  $k \geq K$ , only the decisions from  $\mathbf{D}^*$  provide the maximum in (26).  $\square$

## References

1. Hernandez-Lerma, O., Lasserre, J.B.: Further Topics on Discrete-Time Markov Control Processes. Springer, New York (1999)
2. Kallenberg, L.C.M.: Markov Decision Processes. Lecture Notes. University of Leiden, The Netherlands (2010)
3. Lewis, M.E., Paul, A.: Uniform turnpike theorems for finite Markov decision processes. *Math. Oper. Res.* **44**, 1145–1160 (2019)
4. Piunovskiy, A.: Optimal Control of Random Sequences in Problems with Constraints. Kluwer, Dordrecht (1997)
5. Piunovskiy, A.: Turnpikes and random walk. [arXiv:2102.09341](https://arxiv.org/abs/2102.09341) (2021)
6. Puterman, M.: Markov Decision Processes. Wiley, New York - Chichester - Brisbane - Toronto - Singapore (1994)
7. Shapiro, J.F.: Turnpike planning horizons for a Markovian decision model. *Manag. Sci.* **14**, 292–300 (1968)



# Optimal Stopping Problems for a Family of Continuous-Time Markov Processes

Héctor Jasso-Fuentes<sup>1(✉)</sup>, Jose-Luis Menaldi<sup>2</sup>, and Fidel Vásquez-Rojas<sup>1</sup>

<sup>1</sup> Department of Mathematics, CINVESTAV-IPN, Apartado Postal 14-740,  
07000 Mexico City, Mexico

[hjasso@math.cinvestav.mx](mailto:hjasso@math.cinvestav.mx), [fvasquez@math.cinvestav.mx](mailto:fvasquez@math.cinvestav.mx)

<sup>2</sup> Department of Mathematics, Wayne State University, Detroit, MI 48202, USA  
[menaldi@wayne.edu](mailto:menaldi@wayne.edu)

**Abstract.** In this chapter we study the well-known optimal stopping problem applied to a general family of continuous-time Markov processes. The approach to follow is merely analytic and it is based on the characterization of stopping problems through the study of a certain variational inequality; namely one solution of this inequality will coincide with the optimal value of the stopping problem. In addition, by means of this characterization, it is possible to find the so-named continuation region, and as a byproduct obtaining the optimal stopping time. The most of the material is based on the semigroup theory, infinitesimal generators and resolvents. The chapter is a complete version of the former presentation without detailed proofs in [25].

**Keywords:** Optimal stopping times · Continuous-time Markov processes · Variational inequalities

**AMS(2020) Subject Classification:** Primary 60G40 · Secondary 60J25 · 49J40

## 1 Introduction

Optimal stopping problems are perhaps one of the most interesting and studied problems in the theory of stochastic processes. Successful methods have been developed during decades to show the existence and several characterizations of optimal stopping times. The most studied methods to address these problems are definitely the theory of Snell envelopes and backward-reflected stochastic differential equations—see [7, 14, 15, 17, 18], but on the other hand, there is also another useful method that tackles stopping problems from a merely analytical viewpoint—see [3, 5, 22, 23, 29, 31, 32], among others.

One of the main differences of the second method with respect to the former is the assumption of a Markovian structure of the process, so in principle it could seem more restrictive. However, its analytical nature allows the use of sophisticated tools of functional analysis, set topology, or even more, the use of numerical approximations of the original (and theoretic) problem—see for instance [16].

In this work we shall apply the analytical approach we have already mentioned, and extend several works on this line. But before to specify the details, we can depart to mentioning some pioneer works on this analytical direction, such as [3–5]. All these works were focused on the study of both optimal stopping and impulsive control problems associated to non-degenerated diffusion processes. Based on these works, several authors followed the same line (with both/either theoretical and/or applied viewpoints) that have produced during decades a spread of knowledge on this field.

Other former but nor less important works were developed by Robin [31, 32] and later by Stettner [34] that also applied analytical tools for solving optimal stopping problems on general continuous-time Markov-Feller processes. Within the analysis of the aforementioned papers, we highlight the assumptions on the state space of either type: locally compact or compact.

Following with the description of the former literature, we can quote Menaldi's works [22–24] as well as the one by Menaldi and Sritharan [28], in which the authors analyzed two great families of Markov-Feller process: (1) degenerate stochastic differential equations with either jumps or without jumps, and (2) Navier-Stokes equations; in all these mentioned works the authors take advantage to the particularities of the model in order to explore the regularity of the optimal values. As for the discrete-time models there is a handful of works such as [6, 19, 20, 30],

In this work we use the same line as Robin's works [31, 32] but we drop the local-compactness assumption of the state space. It is important to say that our model is based on the existence of a Markov process that lives on a *fixed* probability space, whereas in the aforementioned references, this space is constructed through the canonical space. This implies that both works are not a special case of each other. Actually, we are somehow inspired from the ideas scattered in reference [25]. One difference of this reference with respect to this proposal, is the nature of the dynamical system and also the general details, since in this work we detail point by point all the arguments of the proofs.

The content of this paper is organized as follows: In Sect. 2, we describe the class of Markov processes we are interested in, and its associated semigroup. Due to a minimal set of assumptions imposed to this process, we will be forced to introduce a seminorm that measures the maximum value of functions along the trajectories (rather than over the whole space, that is the usual case of the supremum norm). This seminorm, produces some properties of the aforementioned semigroup such as a kind of Feller version that is measured through this seminorm. By the end of the section, we will define the corresponding infinitesimal generator and the resolvent operators that both together play a substantial role within the analysis of the optimal stopping problems. In Sect. 3, we will turn our attention to the study of the so-called penalized problem, whose main characteristic is the associated parametric family of functional equations that will be analyzed in this part; in particular, the existence and regularity of these functional equations are ensured. Later we will consider a certain variational inequality. This inequality satisfies the following two nice properties: (i) one of

its subsolutions becomes the limit of the (unique) solutions of the aforementioned family of parametric equations and (ii) the maximal sub-solution of this inequality is just the minimal cost of our stopping problem; this last property will be proved later in Sect. 4, in which we will also provide a characterization of the optimal stopping time as a hitting time associated to a given set so-called continuation region or contact set.

## 2 A Family of Markov Processes

In this section we introduce the dynamics of our stopping problem. This dynamics consists of a continuous-time Markov process that in turn defines a family of operators so-called the semigroup of the process. With these elements it is possible to introduce both an infinitesimal generator and a resolvent operator related to that semigroup. These latter operators will play a substantial role for the analysis of the optimal stopping problem. The way to construct the above mentioned mathematical objects is not straightforward due to the generality of the state space.

### 2.1 Preliminaries

Let  $\mathcal{E} := (\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$  be a fixed filtered probability space, satisfying the usual conditions (i.e., the filtration  $\{\mathcal{F}_t\}_{t \geq 0}$  is right-continuous and  $\mathcal{F}_0$  contains all subsets of the  $\mathbb{P}$ -null sets). Besides, let us consider an open subset  $\mathcal{O}$  of a Banach space with norm  $|\cdot|$ . Throughout this work we will be working with an abstract *homogeneous*  $\mathcal{O}$ -valued stochastic process  $\{y(t, x)\}_{t \geq 0}$ , with initial condition  $x \in \mathcal{O}$  (i.e.  $\mathbb{P}(y(0, x) = x) = 1$ ), defined on  $\mathcal{E}$ .

A first consequence of the above mathematical objects is the definition of the space  $B(\mathcal{O})$  consisting of all measurable functions  $h : \mathcal{O} \rightarrow \mathbb{R}$  such that

$$h(y(t, x)) \in L_1(\Omega, \mathbb{R}), \quad \forall t \geq 0, \quad x \in \mathcal{O}; \quad (1)$$

we note that every bounded measurable function belongs to this space.

With these preliminary elements, we can establish the following assumptions for  $\{y(t, x)\}_{t \geq 0}$ :

**Assumption 1.** *The mapping*

$$(t, x) \mapsto \mathbb{P}(y(t, x) \in B) \quad \text{is measurable } \forall B \in \mathfrak{B}(\mathcal{O}), \quad (2)$$

where  $\mathfrak{B}(\mathcal{O})$  denotes the  $\sigma$ -algebra generated by  $\mathcal{O}$ . In addition:

- (a) There exist constants  $\alpha_0 > 0$ , and  $k \geq 1$ , as well as a measurable function  $w : \mathcal{O} \rightarrow [1, +\infty)$  satisfying  $\lim_{|x| \rightarrow \infty} w(x) = \infty$ , such that all together satisfy the following:

(a.1)

$$\mathbb{E} \left[ \sup_{s \geq 0} \{e^{-\alpha_0 s} w(y(s, x))\} \right] \leq kw(x), \quad \forall x \in \mathcal{O}, \quad \text{and} \quad (3)$$

(a.2)

$$\mathbb{E} [e^{-\alpha_0 s} w(y(s, x))] \leq w(x), \quad \forall x \in \mathcal{O} \quad \text{and} \quad \forall s \geq 0, \quad (4)$$

where  $\mathbb{E}[\cdot]$  is the expectation associated to  $\mathbb{P}$ .

(b) *The Markov property:*

$$\mathbb{P}(y(t+s, x) \in B | \mathcal{F}_s) = \mathbb{P}(y(t, y(s, x)) \in B), \quad \text{a.s.} \quad \forall t, s \geq 0, \quad B \in \mathfrak{B}(\mathcal{O}). \quad (5)$$

The right-hand side of the above equality is understood as the evaluation of the mapping  $z \mapsto \mathbb{P}(y(t, z) \in B)$  at  $z = y(s, x)$ .

(c) *The following relation holds true for all  $s, t \geq 0, x \in \mathcal{O}$*

$$\mathbb{E}[h(y(t, y(s, x)))] = \mathbb{E}[h(y(s, y(t, x)))] , \quad \text{a.s.} \quad \forall h \in B(\mathcal{O}), \quad (6)$$

where the left-hand side means the evaluation of  $z \mapsto \mathbb{E}[h(y(t, z))]$  at  $z = y(s, x)$ , and the right-hand side is the evaluation of  $f \mapsto \mathbb{E}[f(y(t, x))]$  at  $f = h(y(s, \cdot))$ .

(d) *For each  $x \in \mathcal{O}$ ,  $t \mapsto y(t, x)$  has not discontinuities of second kind. Moreover, for all  $x \in \mathcal{O}$  and  $\varepsilon > 0$  there is  $\delta > 0$  such that if  $0 \leq t \leq \delta$  then*

$$\mathbb{P}\left(\sup_{0 \leq s \leq \frac{1}{\varepsilon}} |y(t+s, x) - y(s, x)| \geq \varepsilon\right) < \varepsilon. \quad (7)$$

*Remark 1.*(a) The measurability assumption (2), is a well known fact, as it is established in Dellacherie and Meyer [12], Ethier and Kurtz [13], Rogers and Williams [33]. A clear consequence of the above property is that  $(t, x) \mapsto \mathbb{E}[h(y(t, x))]$  is measurable for every simple function  $h : \mathcal{O} \rightarrow \mathbb{R}$ . Thus, a standard convergence procedure to each  $h \in B(\mathcal{O})$  from sequences of simple functions, yields that

$$(t, x) \mapsto \mathbb{E}[h(y(t, x))] \quad \text{is measurable} \quad \forall h \in B(\mathcal{O}). \quad (8)$$

In particular, equation (6) is well-defined.

- (b) It is worth to say that properties (3), (4), and (7) are common in special cases of Markov processes, such as those that come from solutions of both ordinary and partial stochastic differential equations—see Bensoussan and Lions [3, 5], Bensoussan [4], Menaldi [22–24], Menaldi and Sritharan [26–28].
- (c) It is not difficult to prove that the Markov property (5) is equivalent to this one:

$$\mathbb{E}[h(y(t, y(s, x)))] = \mathbb{E}[h(y(t+s, x)) | \mathcal{F}_s] \quad \forall t \geq s \geq 0, \quad x \in \mathcal{O}, \quad \forall h \in B(\mathcal{O}). \quad (9)$$

- (d) Condition (6) is a kind of uniqueness on the paths. This type of relation is satisfied for a big family of Markov processes  $\{y(t, x)\}_{t \geq 0}$ , for instance the well-known family of Ito's process (with or without jumps, of finite or infinite dimension)—see Bensoussan and Lions [3, 5], Bensoussan [4], Menaldi [22–24], Menaldi and Sritharan [26–28], Da Prato [10, 11], among others.

(e) By writing the set of right-discontinuities of  $\{y(t, x)\}_{t \geq 0}$  as

$$\cup_{\varepsilon > 0} \cap_{\delta > 0} \left\{ \sup_{0 \leq t \leq \delta} |y(t + s, x) - y(s, x)| \geq \varepsilon \right\},$$

and with the aid of (7), it is not difficult to show that  $\{y(t, x)\}_{t \geq 0}$  has right-continuous paths. Also, since the process has not second order discontinuities, we can conclude that it is *càdlàg*.

We will also think over the space of functions  $h \in B(\mathcal{O})$  with the property of

$$\sup_{x \in \mathcal{O}} \frac{|h(x)|}{w(x)} < \infty. \quad (10)$$

This space is denoted by  $B_w(\mathcal{O})$  that will be endowed with the norm

$$\|h\|_w := \sup_{x \in \mathcal{O}} \frac{|h(x)|}{w(x)}. \quad (11)$$

It is common to say that every function in  $B_w(\mathcal{O})$  satisfies a finite  $w$ -growth. In addition, it is not difficult to show that  $(B_w(\mathcal{O}), \|\cdot\|_w)$  is a Banach space.

Finally, using the (fixed) constant  $\alpha_0 > 0$  appearing in (3), we introduce the family of seminorms  $\{p(\cdot, x)\}_{x \in \mathcal{O}}$  on  $B(\mathcal{O})$  by

$$p(h, x) = \mathbb{E} \left[ \sup_{s \geq 0} \{ e^{-\alpha_0 s} |h(y(s, x))| \} \right], \quad \forall x \in \mathcal{O}. \quad (12)$$

Each element of the above family is in fact a seminorm because  $p(h, x) \geq 0$ ,  $p(ah, x) = |a|p(h, x)$  for all  $a \in \mathbb{R}$  and  $p(h + g, x) \leq p(h, x) + p(g, x)$ , but if  $p(h, x) = 0$  then  $\{h(y(s, x))\}_{s \geq 0}$  is indistinguishable of the constant process equal to zero. Using this seminorm, we shall denote by  $B_p(\mathcal{O})$  the subspace of  $B(\mathcal{O})$  consisting of functions  $h$  satisfying

$$p(h, x) < \infty, \quad \forall x \in \mathcal{O}. \quad (13)$$

Note that the definition of this later space, together with the definition of  $B_w(\mathcal{O})$  in (11), and the assumption in (3), all together yield that  $B_w(\mathcal{O}) \subseteq B_p(\mathcal{O}) \subseteq B(\mathcal{O})$ .

## 2.2 The Associated Semigroup

For  $\alpha \geq \alpha_0$ , with  $\alpha_0$  as in (3), we define the family of operators  $\{\Phi_\alpha(t)\}_{t \geq 0}$  on  $B_p(\mathcal{O})$  by

$$\Phi_\alpha(t)h(x) = \mathbb{E}[e^{-\alpha t} h(y(t, x))], \quad \forall x \in \mathcal{O}, h \in B_p(\mathcal{O}), t \geq 0. \quad (14)$$

In view of  $\Phi_\alpha(t)$  is essentially an integral (with respect to the probability measure  $\mathbb{P}$ ), we have that it is monotone, that is,  $h \geq 0$  implies  $\Phi_\alpha(t)h \geq 0$  for any  $t \geq 0$ . Besides, from the definition of  $\Phi_\alpha(t)$  in (14), it is clear that  $\Phi_\alpha(0)h = h$ . This

family of operators also satisfies the *semigroup property*  $\Phi_\alpha(t)\Phi_\alpha(s) = \Phi_\alpha(t+s)$  that follows directly from the Markov property, namely for  $h \in B_p(\mathcal{O})$ ,

$$\begin{aligned}\Phi_\alpha(t)\Phi_\alpha(s)h(x) &= \mathbb{E}[e^{-\alpha t}\Phi_\alpha(s)h(y(t, x))] = \mathbb{E}[e^{-\alpha t}\mathbb{E}[e^{-\alpha s}h(y(t, y(s, x)))]]] \\ &= \mathbb{E}[e^{-\alpha(t+s)}\mathbb{E}[h(y(t+s, x))|\mathcal{F}_s]] = \mathbb{E}[e^{-\alpha(t+s)}h(y(t+s, x))] \\ &= \Phi_\alpha(t+s)h(x).\end{aligned}$$

If  $h \in B_w(\mathcal{O})$  then, using the inequality (4) as well as the norm in (11), we get the following

$$\begin{aligned}|\Phi_\alpha(t)h(x)| &\leq \mathbb{E}[e^{-\alpha t}|h(y(t, x))|] = \mathbb{E}[e^{-\alpha t}\frac{|h(y(t, x))|}{w(y(t, x))}w(y(t, x))] \\ &\leq \|h\|_w \mathbb{E}[e^{-\alpha_0 t}w(y(t, x))] \leq \|h\|_w w(x), \quad \forall x \in \mathcal{O}.\end{aligned}$$

Hence,

$$\|\Phi_\alpha(t)h\|_w \leq \|h\|_w. \quad (15)$$

The semigroup property naturally arises when the operators  $\Phi_\alpha(t)$  are defined as an integral with respect to a given transition probability kernel  $q(x, t, \cdot) = \mathbb{P}[y(t, x) \in \cdot]$  that in turn satisfies the well-known Chapman-Kolmogorov equations. This last type of equations is very common in specific models, such as continuous-time Markov chains, Lévy Processes, partial stochastic differential equations, to mention a few. (See [1, 2, 10, 11], among others). The family of the operators  $\Phi_\alpha$  defined in (14) will be called throughout this work as the associated *semigroup* of the Markov process  $\{y(t, x)\}_{t \geq 0}$ .

As we will see in the following result, the semigroup  $\Phi_\alpha$  satisfies the contraction property with respect to the seminorm  $p(\cdot, x)$ . The details are as follows.

**Proposition 1.** *For each  $h \in B_p(\mathcal{O})$ ,  $t, s \geq 0$  and  $x \in \mathcal{O}$  we have that*

$$p(\Phi_\alpha(t)h, x) \leq p(h, x).$$

*Proof.* Fixed  $h \in B_p(\mathcal{O})$ ,  $t, s \geq 0$  and  $x \in \mathcal{O}$ , we have

$$\begin{aligned}p(\Phi(t)h, x) &= \mathbb{E}\left[\sup_{s \geq 0}\{e^{-\alpha_0 s}|\mathbb{E}[e^{-\alpha t}h(y(t, y(s, x)))]|\}\right] \\ &= \mathbb{E}\left[\sup_{s \geq 0}\{e^{-\alpha_0 s}|\mathbb{E}[e^{-\alpha t}h(y(s, y(t, x)))]|\}\right] \quad (\text{by (6)}) \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\sup_{s \geq 0}\{e^{-\alpha_0 s}e^{-\alpha t}|h(y(s, y(t, x)))|\}\right]\right].\end{aligned}$$

On the other hand, it is not difficult to prove that the Markov property in (1) implies the Markov property (see e.g. [35, Section 5.2.2.]) in the following sense

$$\mathbb{E}\left[\sup_{s \geq 0}\{e^{-\alpha_0 s}e^{-\alpha t}|h(y(s, y(t, x)))|\}\right] = \mathbb{E}\left[\sup_{s \geq 0}\{e^{-\alpha_0 s}e^{-\alpha t}|h(y(s+t, x))|\}\middle|\mathcal{F}_t\right].$$

Then we can conclude that

$$\begin{aligned}
p(\Phi(t)h, x) &\leq \mathbb{E}[\mathbb{E}[\sup_{s \geq 0}\{e^{-\alpha_0 s} e^{-\alpha t} |h(y(s + t, x))|\} | \mathcal{F}_t]] \\
&= \mathbb{E}[\sup_{s \geq 0}\{e^{-\alpha_0 s} e^{-\alpha t} |h(y(s + t, x))|\}] \\
&\leq \mathbb{E}[\sup_{s \geq 0}\{e^{-\alpha_0 s} |h(y(s, x))|\}] = p(h, x).
\end{aligned}$$

□

*Remark 2.*(a) The assumption in (2) together with (15), give us that  $\Phi_\alpha(t)$  leaves invariant the space  $B_w(\mathcal{O})$ ; actually, our family of operators  $t \mapsto \Phi_\alpha(t)$  satisfies the properties of the so-called *monotone semigroup of contractions* defined on  $B_w(\mathcal{O})$ .

(b) Even more, (2) and Proposition 1 also give the invariance of the semigroup  $\Phi_\alpha$  over the set  $B_p(\mathcal{O})$ .

**Continuity of the Semigroup.** In many situations, the above semigroup satisfies the so-called *strong* continuity (see [1, 2, 8, 10] among others)

$$\|\Phi_\alpha(t)h - h\| \rightarrow 0, \quad \text{as } t \downarrow 0, \tag{16}$$

applied to a suitable space of functions  $h$ —for example, the set of continuous functions that vanish at infinity. The above case is very common when the dimension of  $\mathcal{O}$  is either finite-dimensional or locally compact. However, there exist situations when  $\mathcal{O}$  does not hold the previous two properties—for example, assume that  $\mathcal{O}$  is a Hilbert space as in references [26–28]), so convergence (16) is no longer valid. However, it is possible to obtain a sort of continuity type in the next weaker sense (see, for instance Böttcher et al. [8], Menaldi [25], or Menaldi and Sritharan [28]).

$$\Phi_\alpha(t)h(x) - h(x) \rightarrow 0, \quad \text{as } t \downarrow 0 \quad \forall x \in \mathcal{O}, \tag{17}$$

where  $h$  is Borel measurable. One of the disadvantages of this later continuity is that it produces a lack of regularity of some sophisticated mathematical objects (i.e., infinitesimal generator, the resolvent operator, among others), whose definitions depend strongly from the convergence in (17).

Since our hypotheses of the state space  $\mathcal{O}$  are not restricted to the cases of finite dimension nor local compactness, it is expected to not obtain convergence of type (16), even when we could use the norm  $\|\cdot\|_w$ . To avoid this drawback, we shall seek an intermediate convergence, weaker than (16) but a little stronger than (17) so that we are in conditions to achieve regularity properties for the infinitesimal generator and on the resolvent operator. The key point is to define a suitable functions set whose elements are continuous in certain sense but at the same time, the semigroup applied to this set can be continuous in seminorm (see Definition 2 below).

Let us now define the concept of convergence in seminorm that is crucial to define continuity in seminorm sense.

**Definition 1.** We say that a sequence  $h_n$  in  $B_p(\mathcal{O})$  converges in seminorm to some  $h$  in  $B_p(\mathcal{O})$  as  $n \rightarrow \infty$ , denoted by  $s - \lim_{n \rightarrow \infty} h_n = h$ , if

$$\lim_{n \rightarrow \infty} p(h_n - h, x) = 0, \quad \forall x \in \mathcal{O}. \quad (18)$$

Moreover, if the elements of the above sequence are in  $B_w(\mathcal{O})$  then we say that  $h_n$  converges boundedly in seminorm to  $h$  as  $n \rightarrow \infty$ , denoted by  $bs - \lim_{n \rightarrow \infty} h_n = h$ , provided the following conditions are satisfied

$$\begin{cases} \sup_{n \in \mathbb{N}} \|h_n\|_w < \infty; \\ s - \lim_{n \rightarrow \infty} h_n = h. \end{cases} \quad (19)$$

Note that for each  $x \in \mathcal{O}$ ,  $t \geq 0$ , and  $h \in B_w(\mathcal{O})$ , a simple use of the bound (3) yields that

$$\begin{aligned} p(h, x) &= \mathbb{E}[\sup_{s \geq 0} e^{-\alpha_0 s} |h(y(s, x))|] \\ &\leq \mathbb{E}[\sup_{s \geq 0} e^{-\alpha_0 s} \|h\|_w w(y(s, x))] \leq k \|h\|_w w(x) < \infty. \end{aligned} \quad (20)$$

The above relation means that convergence in norm implies convergence in seminorm which, at the same time, implies pointwise convergence.

**Definition 2.** We define the subspace  $C_p(\mathcal{O})$  of  $B_p(\mathcal{O})$  that is conformed by the functions  $h$  such that:

- (a)  $s - \lim_{t \downarrow 0} \Phi_\alpha(t)h = h$ ,
- (b) for each  $x \in \mathcal{O}$  we have that  $\{h(y(s, x))\}_{s \geq 0}$  is a càdlàg process.

We also denote the intersection  $C_p(\mathcal{O}) \cap B_w(\mathcal{O})$  by  $C_p^w(\mathcal{O})$ .

The next proposition shows further properties of the sets  $C_p(\mathcal{O})$  and  $C_p^w(\mathcal{O})$ .

**Proposition 2.** Under Assumption 1, we have

- (a) The sets  $C_p(\mathcal{O})$  and  $C_p^w(\mathcal{O})$  are non-empty.
- (b) For every  $t \geq 0$ :
  - (b.1)  $\Phi_\alpha(t)h \in C_p(\mathcal{O})$  when  $h \in C_p(\mathcal{O})$ ,
  - (b.2)  $\Phi_\alpha(t)h \in C_p^w(\mathcal{O})$  when  $h \in C_p^w(\mathcal{O})$ .

*Proof.* (a) Let  $C_u(\mathcal{O})$  be the space of bounded uniformly continuous functions and take  $h \in C_u(\mathcal{O})$ . Note that

$$\begin{aligned} p(\Phi_\alpha(t)h - h, x) &\leq p(\Phi_\alpha(t)h - e^{-\alpha t}h, x) + p(e^{-\alpha t}h - h, x) \\ &\leq p(\Phi_\alpha(t)h - e^{-\alpha t}h, x) + (e^{-\alpha t} - 1)p(h, x), \end{aligned}$$

where  $(e^{-\alpha t} - 1) \rightarrow 0$  when  $t \downarrow 0$ . So, we aim to show

$$p(\Phi_\alpha(t)h - e^{-\alpha t}h, x) \rightarrow 0$$

when  $t \downarrow 0$ . Namely, for any  $t \geq 0$  and  $x \in \mathcal{O}$ , we have

$$\begin{aligned} p(\Phi_\alpha(t)h - e^{-\alpha t}h, x) &\leq \mathbb{E}\left[\sup_{s \geq 0} e^{-\alpha_0 s} |h(y(t+s, x)) - h(y(s, x))|\right] \\ &\leq \mathbb{E}\left[\sup_{0 \leq s \leq T} e^{-\alpha_0 s} |h(y(t+s, x)) - h(y(s, x))|\right] \\ &\quad + \mathbb{E}\left[\sup_{s \geq T} e^{-\alpha_0 s} |h(y(t+s, x)) - h(y(s, x))|\right] \end{aligned} \quad (21)$$

for any  $T > 0$ . In order to bound this expression, let us take  $\varepsilon > 0$  and choose  $0 < \delta_1 \leq \varepsilon$  such that  $|x - \bar{x}| < \delta_1$  implies  $|h(x) - h(\bar{x})| < \varepsilon$ . In turn, in virtue of (7), let us choose  $0 < \delta_0 \leq \delta_1$  such that for all  $0 \leq t \leq \delta_0$  we have

$$\mathbb{P}\left(\sup_{0 \leq s \leq \frac{1}{\delta_1}} |y(t+s, x) - y(s, x)| \geq \delta_1\right) < \delta_1.$$

Letting  $T = \frac{1}{\delta_0}$  we get

$$\begin{aligned} &\mathbb{E}\left[\sup_{0 \leq s \leq \frac{1}{\delta_0}} e^{-\alpha_0 s} |e^{-\alpha t}h(y(t+s, x)) - h(y(s, x))|\right] \\ &\leq \mathbb{E}\left[\sup_{0 \leq s \leq \frac{1}{\delta_0}} e^{-\alpha_0 s} |e^{-\alpha t}h(y(t+s, x)) - h(y(s, x))|\right. \\ &\quad \times \mathbf{1}_{\sup_{0 \leq s \leq \frac{1}{\delta_0}} |y(t+s, x) - y(s, x)| < \delta_0}] \\ &\quad + \mathbb{E}\left[\sup_{0 \leq s \leq \frac{1}{\delta_0}} e^{-\alpha_0 s} |e^{-\alpha t}h(y(t+s, x)) - h(y(s, x))|\right. \\ &\quad \times \mathbf{1}_{\sup_{0 \leq s \leq \frac{1}{\delta_0}} |y(t+s, x) - y(s, x)| \geq \delta_0}. \end{aligned}$$

The fact that  $h$  is bounded (uniformly), gives us

$$\begin{aligned} &\mathbb{E}\left[\sup_{0 \leq s \leq \frac{1}{\delta_0}} e^{-\alpha_0 s} |e^{-\alpha t}h(y(t+s, x)) - h(y(s, x))| \mathbf{1}_{\sup_{0 \leq s \leq \frac{1}{\delta_0}} |y(t+s, x) - y(s, x)| \geq \delta_0}\right] \\ &\leq 2 \|h\|_\infty \mathbb{P}\left(\sup_{0 \leq s \leq \frac{1}{\delta_0}} |y(t+s, x) - y(s, x)| \geq \delta_0\right) < 2 \|h\|_\infty \varepsilon, \end{aligned} \quad (22)$$

where we have denoted by  $\|\cdot\|_\infty$  the supremum norm. On the other hand, the uniform continuity of  $h$  gives us

$$\begin{aligned} &\mathbb{E}\left[\sup_{0 \leq s \leq \frac{1}{\delta_0}} e^{-\alpha_0 s} |e^{-\alpha t}h(y(t+s, x)) - h(y(s, x))|\right. \\ &\quad \times \mathbf{1}_{\sup_{0 \leq s \leq \frac{1}{\delta_0}} |y(t+s, x) - y(s, x)| < \delta_0}] < \varepsilon. \end{aligned} \quad (23)$$

We have for the second term in the right-hand side of (21)

$$\begin{aligned} &\mathbb{E}\left[\sup_{s \geq \frac{1}{\delta_0}} e^{-\alpha_0 s} e^{-\alpha t} |h(y(t+s, x)) - h(y(s, x))|\right] \leq \sup_{s \geq \frac{1}{\delta_0}} 2e^{-\alpha_0 s} \|h\|_\infty \\ &\leq 2e^{-\alpha_0 \frac{1}{\delta_0}} \|h\|_\infty. \end{aligned} \quad (24)$$

Using the estimations (22), (23) and (24) in (21) we get  $p(\Phi_\alpha(t)h - e^{-\alpha t}h, x) \rightarrow 0$  as  $t \downarrow 0$ . This proves that  $h$  satisfies part (a) of Definition 2. But also note that  $h$  trivially satisfies Definition 2(b) because  $\{y(t, x)\}_{t \geq 0}$  is càdlàg. Therefore, we can easily conclude that  $C_u(\mathcal{O}) \subset C_p^w(\mathcal{O}) \subset C_p(\mathcal{O})$ , which proves part (a) of this proposition.

(b.1) Let  $h \in C_p(\mathcal{O})$ . Now, in virtue of Proposition 1, we have that

$$p(\Phi_\alpha(t)h, x) \leq p(h, x),$$

for each  $x \in \mathcal{O}$  and  $t \geq 0$ . Hence

$$\begin{aligned} p(\Phi_\alpha(s)\Phi_\alpha(r)h - \Phi_\alpha(r)h, x) &= p(\Phi_\alpha(r)(\Phi_\alpha(s)h - h), x) \\ &\leq p(\Phi_\alpha(s)h - h, x) \rightarrow 0, \quad s \downarrow 0. \end{aligned}$$

This shows that  $\Phi_\alpha(r)h \in C_p(\mathcal{O})$  for all  $r \geq 0$ , for every element  $h \in C_p(\mathcal{O})$ . It remains to prove that the process  $\{\Phi_\alpha(t)h(y(s, x))\}_{s \geq 0}$  is càdlàg for each  $x \in \mathcal{O}$  and  $t \geq 0$ . To do this, let  $s_0 \geq 0$  and  $\{s_n\}_{n \in \mathbb{N}}$  be a decreasing sequence in  $[0, \infty)$  converging to  $s_0$ . Take  $t \geq 0$  and  $x \in \mathcal{O}$ . We have that  $\{h(y(s+t, x))\}_{s \geq 0}$  is a càdlàg process and  $\sup_{s \geq 0} e^{-\alpha_0 s} |h(y(s+t, x))| \in L_1(\Omega)$  because  $h \in C_p(\mathcal{O})$  and satisfies (13). Hence, applying Theorem 45 in [12], the right continuity of both the filtration and the process  $h(y(s, x))$ , as well as the Markov property, we deduce

$$\begin{aligned} \lim_{n \rightarrow \infty} e^{-\alpha_0 s_n} e^{\alpha t} \Phi_\alpha(t)h(y(s_n, x)) &= \lim_{n \rightarrow \infty} \mathbb{E}[e^{-\alpha_0 s_n} h(y(s_n + t, x)) | \mathcal{F}_{s_n}] \\ &= \mathbb{E}[e^{-\alpha_0 s_0} h(y(s_0 + t, x)) | \mathcal{F}_{s_0}] = e^{-\alpha_0 s_0} e^{\alpha t} \Phi_\alpha(t)h(y(s_0, x)), \quad \text{a.s.} \end{aligned}$$

Due to the continuity of the exponential function, from the above we deduce that  $\lim_{s \downarrow s_0} \Phi_\alpha(t)h(y(s, x)) = \Phi_\alpha(t)h(y(s_0, x))$ , a.s. On the other hand, using again Theorem 45 in [12] and the existence of left-limits of the process  $h(y(s, x))$  we get  $\lim_{s \uparrow s_0} \Phi_\alpha(t)h(y(s, x)) = \mathbb{E}[e^{-\alpha t} h(y(t + s_0^-, x)) | \mathcal{F}_{s_0^-}]$ , a.s. Therefore, the process  $\{\Phi_\alpha(t)h(y(s, x))\}_{s \geq 0}$  is càdlàg.

(b.2) If  $h \in C_p^w(\mathcal{O})$ , then we have that  $\|\Phi_\alpha(t)h\|_w \leq \|h\|_w < \infty$  due to (15), yielding that  $\Phi_\alpha(t)h \in C_p^w(\mathcal{O})$ .  $\square$

Our next target is to describe a closedness properties of both  $C_p(\mathcal{O})$  and  $C_p^w(\mathcal{O})$  under (boundedly) seminorm-convergence. For this end, we will prove the next ancillary result.

**Lemma 1.** *Consider a sequence of functions  $\{h_n\}_{n \in \mathbb{N}}$  together with a function  $h$  all contained in  $B(\mathcal{O})$ . For each  $x \in \mathcal{O}$ , suppose that  $\lim_{n \downarrow 0} p(h_n - h, x) = 0$ . Then, there exists a subsequence  $\{n_k\}_{k \in \mathbb{N}}$  (dependent of  $x$ ), such that*

$$\lim_{k \rightarrow \infty} \sup_{s \geq 0} \{e^{-\alpha_0 s} |h_{n_k}(y(s, x)) - h(y(s, x))|\} = 0, \quad \text{a.s.}$$

*Proof.* We note that convergence in seminorm implies that

$$\sup_{s \geq 0} \{e^{-\alpha_0 s} |h_n(y(s, x)) - h(y(s, x))|\} \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (25)$$

where the last convergence is of  $L_1(\Omega, \mathbb{R})$  type. Then, the above sequence converges also in measure and this yields the existence of a subsequence which converges a.s.  $\square$

**Theorem 1.** *Let  $h$  and  $\{h_n\}_{n \in \mathbb{N}}$  be functions all in  $B(\mathcal{O})$ . Then, under Assumption 1, the following assertions hold true.*

- (a) *If  $h_n \in B_p(\mathcal{O})$  and  $s - \lim_{n \rightarrow \infty} h_n = h$  then  $h \in B_p(\mathcal{O})$ .*
- (b) *If  $h_n \in C_p(\mathcal{O})$  and  $s - \lim_{n \rightarrow \infty} h_n = h$  then  $h \in C_p(\mathcal{O})$ .*
- (c) *If  $h_n \in C_p^w(\mathcal{O})$  and  $\text{bs} - \lim_{n \rightarrow \infty} h_n = h$  then  $h \in C_p^w(\mathcal{O})$ .*

*Proof.* (a) Given  $x \in \mathcal{O}$ , there exists  $n \in \mathbb{N}$  such that  $p(h - h_n, x) \leq 1$  and we have that  $|h| \leq |h - h_n| + |h_n|$ . Then due to the triangular inequality of the seminorm, we get  $p(h, x) \leq p(h - h_n, x) + p(h_n, x) < \infty$  and therefore  $h \in B_p(\mathcal{O})$ .

(b) Let us suppose  $h_n \in C_p(\mathcal{O})$  and  $s - \lim_{n \rightarrow \infty} h_n = h$ . Then we have that

$$\begin{aligned} p(\Phi_\alpha(t)h - h, x) &\leq p(\Phi_\alpha(t)h - \Phi_\alpha(t)h_n, x) + p(\Phi_\alpha(t)h_n - h_n, x) + p(h_n - h, x) \\ &\leq 2p(h_n - h, x) + p(\Phi_\alpha(t)h_n - h_n, x). \end{aligned}$$

Letting  $t \downarrow 0$  and hence  $n \rightarrow \infty$  to the last expression, we get  $\lim_{t \downarrow 0} p(\Phi_\alpha(t)h - h, x) = 0$ , for each  $x \in \mathcal{O}$ . On the other hand, a simple use of Lemma 1 ensures the existence of a subsequence  $\{n_k\}_{k \in \mathbb{N}}$  such that

$$\sup_{s \geq 0} \{e^{-\alpha_0 s} |h_{n_k}(y(s, x)) - h(y(s, x))|\} \rightarrow 0, \quad \text{a.s.} \quad (26)$$

when  $k \rightarrow \infty$ . Note that this subsequence  $x$ . Let  $t_0 \geq 0$ , we have that

$$\begin{aligned} &|h(y(t, x)) - h(y(t_0, x))| \\ &\leq |h(y(t, x)) - h_{n_k}(y(t, x))| + |h_{n_k}(y(t, x)) - h_{n_k}(y(t_0, x))| \\ &\quad + |h_{n_k}(y(t_0, x)) - h(y(t_0, x))| \\ &\leq (e^{\alpha_0 t} + e^{\alpha_0 t_0}) \sup_{s \geq 0} e^{-\alpha_0 s} |h_{n_k}(y(s, x)) - h(y(s, x))| \\ &\quad + |h_{n_k}(y(t, x)) - h_{n_k}(y(t_0, x))|. \end{aligned} \quad (27)$$

Since  $t \mapsto h_{n_k}(y(t, x))$  is right-continuous and considering the convergence (26), we then apply the limits  $t \downarrow t_0$  and hence  $k \rightarrow \infty$  on the last expression and obtain  $\lim_{t \downarrow t_0} |h(y(t, x)) - h(y(t_0, x))| = 0$  a.s. On the other hand, in the same way as in (27), we get

$$\begin{aligned} |h(y(t, x)) - h(y(t_0^-, x))| &\leq (e^{\alpha_0 t} + e^{\alpha_0 t_0}) \sup_{s \geq 0} e^{-\alpha_0 s} |h_{n_k}(y(s, x)) - h(y(s, x))| \\ &\quad + |h_{n_k}(y(t, x)) - h_{n_k}(y(t_0^-, x))|. \end{aligned}$$

We apply the limits  $t \uparrow t_0$  and hence  $k \rightarrow \infty$  on the last expression and obtain  $\lim_{t \uparrow t_0} |h(y(t, x)) - h(y(t_0^-, x))| = 0$  a.s. due to the left-limits existence.

(c) If  $h_n \in C_p(\mathcal{O})$  and  $\text{bs} - \lim_{n \rightarrow \infty} h_n = h$ , then we have that

$$s - \lim_{n \rightarrow \infty} h_n = h$$

and  $\sup_{n \in \mathbb{N}} \|h_n\|_w$ , then due to part (b), we have that  $h \in C_p(\mathcal{O})$  and we need to demonstrate  $\|h\|_w < \infty$ . Namely, we have that seminorm convergence implies pointwise convergence, hence  $\frac{|h(x)|}{w(x)} = \lim_{n \rightarrow \infty} \frac{|h_n(x)|}{w(x)} \leq \sup_{n \in \mathbb{N}} \|h_n\|_w < \infty$  implying  $\|h\|_w < \infty$  and therefore  $h \in C_p^w(\mathcal{O})$ .  $\square$

### 2.3 The Infinitesimal Generator and the Resolvent

We define the infinitesimal generator  $(D(\mathcal{A}_\alpha), \mathcal{A}_\alpha)$  associated to the semigroup  $\Phi_\alpha$  as follows

$$\begin{cases} D(\mathcal{A}_\alpha) := \{h \in C_p^w(\mathcal{O}) : \exists \text{ bs} - \lim_{t \downarrow 0} \frac{h - \Phi_\alpha(t)h}{t}\}; \\ \mathcal{A}_\alpha h := \text{bs} - \lim_{t \downarrow 0} \frac{h - \Phi_\alpha(t)h}{t}. \end{cases} \quad (28)$$

*Remark 3.* In virtue of Definition 2 and Theorem 1, every limit in (28) belongs to  $C_p^w(\mathcal{O})$ .

Recall from Assumption 1 that  $t \mapsto \Phi_\alpha(t)h(x)$  is measurable for every  $h \in B(\mathcal{O})$  and  $x \in \mathcal{O}$ , then we are in conditions to define the *resolvent operator*  $\{\mathcal{R}_\alpha\}_{\alpha > \alpha_0}$  by

$$\mathcal{R}_\alpha h(x) = \int_0^\infty \Phi_\alpha(t)h(x) dt, \quad \forall x \in \mathcal{O}, \quad h \in B(\mathcal{O}), \quad (29)$$

where the integral is taken in the Lebesgue sense for real valued functions. A direct consequence of this definition is that  $\mathcal{R}_\alpha h$  is Borel measurable, for each fixed  $\alpha$ . Also, if  $h \in B_p(\mathcal{O})$  then Fubini's Theorem along with Proposition 1 yield

$$\begin{aligned} p(\mathcal{R}_\alpha h, x) &\leq \mathbb{E} \left[ \sup_{s \geq 0} e^{-\alpha_0 s} \int_0^\infty |\Phi_\alpha(t)h(y(s, x))| dt \right] \\ &\leq \int_0^\infty \mathbb{E} \left[ \sup_{s \geq 0} e^{-\alpha_0 s} |\Phi_\alpha(t)h(y(s, x))| \right] dt = \int_0^\infty p(\Phi_\alpha(t)h, x) dt \\ &= \int_0^\infty e^{-(\alpha - \alpha_0)t} p(\Phi_{\alpha_0}(t)h, x) dt \leq \frac{1}{\alpha - \alpha_0} p(h, x), \end{aligned} \quad (30)$$

which implies  $\mathcal{R}_\alpha h \in B_p(\mathcal{O})$ . Moreover, if  $h \in B_w(\mathcal{O})$  then we have

$$\|\mathcal{R}_\alpha h\|_w \leq \int_0^\infty e^{-(\alpha - \alpha_0)t} \|\Phi_{\alpha_0}(t)h\|_w dt \leq \frac{1}{\alpha - \alpha_0} \|h\|_w < \infty, \quad (31)$$

and so  $\mathcal{R}_\alpha h \in B_w(\mathcal{O})$ .

Our next goal is to prove the stronger fact that  $\mathcal{R}h \in C_p^w(\mathcal{O})$  when  $h \in C_p^w(\mathcal{O})$ , that is, we will show that  $\mathcal{R}_\alpha$  maps  $C_p^w(\mathcal{O})$  into itself. Such results will be provided in Theorem 2 below. Before doing this, we will check some useful properties:

In the same way as in (30) it is easy to demonstrate that

$$p\left(\int_a^b \Phi_\alpha(t)h dt, x\right) \leq \int_a^b p(\Phi_\alpha(t)h, x) dt, \quad \text{for every } 0 \leq a \leq b \leq \infty. \quad (32)$$

Besides, we can interchange the semigroup and the resolvent; namely, for every  $\beta > \alpha_0$  and  $\alpha \geq \alpha_0$ , using Fubini's Theorem we get

$$\begin{aligned} \mathcal{R}_\beta \Phi_\alpha(t)h(x) &= \int_0^\infty \Phi_\alpha(t)\Phi_\beta(s)h(x) ds = \int_0^\infty \mathbb{E}[e^{-\alpha t}\Phi_\beta(s)h(y(t, x))] ds \\ &= \mathbb{E}\left[e^{-\alpha t} \int_0^\infty \Phi_\beta(s)h(y(t, x)) ds\right] = \Phi_\alpha(t)\mathcal{R}_\beta h(x). \end{aligned} \quad (33)$$

The use of Fubini's Theorem is justified since

$$\int_0^\infty \mathbb{E}[e^{-\alpha t}|\Phi_\beta(s)h(y(t, x))|] ds \leq \Phi_\alpha(t)\mathcal{R}_\beta|h|(x) \leq \frac{1}{\beta - \alpha_0} \|h\|_w w(x).$$

Our next result uses the following notation:

$$u(t) = \Phi_\alpha(t)h, \quad \text{for a given } h \in C_p(\mathcal{O}) \text{ and } \alpha > \alpha_0. \quad (34)$$

**Lemma 2.** *Fix  $x \in \mathcal{O}$ . Then:*

- (a) *For all  $t_0 \geq 0$  we have that  $\lim_{t \rightarrow t_0} p(u(t) - u(t_0), x) = 0$ .*
- (b) *We have  $\lim_{t \rightarrow \infty} p(u(t), x) = 0$ .*
- (c) *For all  $\varepsilon > 0$  there exists  $\delta = \delta(x, \varepsilon) > 0$  such that if  $|t - s| \leq \delta$  then  $p(u(t) - u(s), x) \leq \varepsilon$ .*

*Proof.* (a) Using Proposition 1, the semigroup property and the continuity in seminorm at  $t = 0$  of the semigroup, it is straightforward to show that

$$\lim_{t \rightarrow t_0^+} p(\Phi_\alpha(t)h - \Phi_\alpha(t_0)h, x) = \lim_{t \rightarrow t_0^-} p(\Phi_\alpha(t)h - \Phi_\alpha(t_0)h, x) = 0,$$

which proves (a).

(b) By Proposition 1, we have that  $p(u(t), x) = p(e^{-(\alpha-\alpha_0)t}\Phi_{\alpha_0}(t)h, x) \leq e^{-(\alpha-\alpha_0)t}p(h, x)$ . Due to  $\alpha - \alpha_0 > 0$ , we can take  $T > 0$  such that

$$e^{-(\alpha-\alpha_0)t}p(h, x) \leq \varepsilon$$

for all  $t \geq T$ .

(c) Using part (b) above, we can take  $T > 0$  large enough such that for all  $t \geq T$ ,  $p(u(t), x) \leq \frac{\varepsilon}{2}$ . Also, by the compactness of  $[0, T]$  and (a) above, we can find  $\delta > 0$  such that  $|t - s| < \delta$  and  $t, s \leq T$  imply  $p(u(t) - u(s), x) < \varepsilon$ . On the other hand, if  $s, t \geq T$  we get  $p(u(t) - u(s), x) \leq p(u(t), x) + p(u(s), x) \leq \varepsilon$ .  $\square$

**Lemma 3.** *For each  $u$  as in (34), there exists a sequence of functions  $u_n : [0, \infty) \rightarrow C_p(\mathcal{O})$  such that*

$$\lim_{n \rightarrow \infty} \sup_{t \geq 0} p(u(t) - u_n(t), x) = 0. \quad (35)$$

Moreover, if  $h \in C_p^w(\mathcal{O})$  then we can choose the above sequence such that  $u_n(t) \in C_p^w(\mathcal{O})$ , for all  $n \in \mathbb{N}$  and  $t \geq 0$ .

*Proof.* For fixed  $x \in \mathcal{O}$  and  $n \in \mathbb{N}$ , we define

$$\begin{aligned} E_{n,k} &:= [\frac{k-1}{n}, \frac{k}{n}), \quad k = 1, \dots, n^2, \\ F_n &:= [n, \infty). \end{aligned}$$

Define also the sequence of functions

$$u_n(t) := \sum_{k=1}^{n^2} u(t_k) \mathbf{1}_{E_{n,k}}(t) + u(n) \mathbf{1}_{F_n}(t), \quad (36)$$

with  $t_k = \frac{k-1}{n}$ . Note by Proposition 2, for all  $n \in \mathbb{N}$  and  $t \geq 0$ , each  $u_n(t)$  is in  $C_p(\mathcal{O})$  because they are linear combination of functions in  $C_p(\mathcal{O})$ . In the same way, if  $h \in C_p^w(\mathcal{O})$  in (34) then in virtue of this same proposition,  $u \in C_p^w(\mathcal{O})$ , yielding also  $u_n(t) \in C_p^w(\mathcal{O})$ . The limit (35) follows easily from estimations in Lemma 2.  $\square$

*Remark 4.* We know that  $u_n(t)$  belongs to  $C_p(\mathcal{O})$  (resp. to  $C_p^w(\mathcal{O})$ ) if  $h \in C_p(\mathcal{O})$  (resp.  $\in C_p^w(\mathcal{O})$ ). Also, because of the definition of  $u_n$  in (36) we have that for each  $x \in \mathbb{R}$  the function  $t \mapsto u_n(t)(x) = \sum_{k=1}^{n^2} u(t_k)(x) \mathbf{1}_{E_{n,k}}(t) + u(n)(x) \mathbf{1}_{F_n}(t)$  is simple and real valued. Hence, given  $\beta > 0$  the function  $t \mapsto e^{-\beta t} u_n(t)(x)$  is Lebesgue integrable with integral given by

$$\int_a^b e^{-\beta t} u_n(t)(x) dt = \sum_{k=1}^{n^2} u(t_k)(x) \int_{E_{n,k} \cap [a,b]} e^{-\beta t} dt + u(n)(x) \int_{F_n \cap [a,b]} e^{-\beta t} dt. \quad (37)$$

We note that the above integral, as a function of  $x$ , belongs to  $C_p(\mathcal{O})$  (resp. to  $C_p^w(\mathcal{O})$ ), because it is a sum of functions in  $C_p(\mathcal{O})$  (resp.  $C_p^w(\mathcal{O})$ ). Then, we simply denote this integral by  $\int_a^b e^{-\beta t} u_n(t) dt$ .

We have arrived to our first main result regarding the regularity of the resolvent  $\mathcal{R}_\alpha$ , when the integrand satisfies that regularity.

**Theorem 2.** *Assume that Assumption 1 is valid. Then, for all  $0 \leq a \leq b \leq \infty$ , and  $\beta > 0$ , the next relation holds true*

$$s - \lim_{n \rightarrow \infty} \int_a^b e^{-\beta t} u_n(t) dt = \int_a^b e^{-\beta t} u(t) dt, \quad (38)$$

for the functions  $u$  and  $\{u_n\}$  introduced in Lemma 3. In particular, we have that  $\mathcal{R}_\alpha h$  is in  $C_p(\mathcal{O})$ . Analogously, we obtain the same result with  $C_p^w(\mathcal{O})$  instead of  $C_p(\mathcal{O})$  if  $h \in C_p^w(\mathcal{O})$  with  $\text{bs-lim}$  instead of  $s-\lim$  in (38).

*Proof.* By the inequality in (32) as long with Lemma 3, we get

$$\begin{aligned} p\left(\int_a^b e^{-\beta t} u_n(t) dt - \int_a^b e^{-\beta t} u(t) dt, x\right) &\leq \int_a^b e^{-\beta t} p(u_n(t) - u(t), x) dt \\ &\leq \sup_{t \in [a,b]} p(u_n(t) - u(t), x) \int_a^b e^{-\beta t} dt \rightarrow 0, \end{aligned}$$

when  $n \rightarrow \infty$ . That is  $\text{s-lim}_{n \rightarrow \infty} \int_a^b e^{-\beta t} u_n(t) dt = \int_a^b e^{-\beta t} u(t) dt$ , that implies  $\int_a^b e^{-\beta t} u(t) dt \in C_p(\mathcal{O})$  due to Theorem 1. Moreover, in the case of  $h \in C_p^w(\mathcal{O})$  we have that  $u(t) \in C_p^w(\mathcal{O})$  and  $\|u(t)\|_w = \|\Phi_\alpha(t)h\|_w \leq \|h\|_w$  for all  $t \geq 0$ . Using this last inequality together with (37) we get

$$\begin{aligned} \left\| \int_a^b e^{-\beta t} u_n(t) dt \right\|_w &\leq \sum_{k=1}^{n^2} \|h\|_w \int_{E_{n,k} \cap [a,b]} e^{-\beta t} dt + \|h\|_w \int_{F_n \cap [a,b]} e^{-\beta t} dt \\ &= \|h\|_w \int_a^b e^{-\beta t} dt < \infty. \end{aligned} \quad (39)$$

Hence,  $\sup_{n \in \mathbb{N}} \left\| \int_a^b e^{-\beta t} u_n(t) dt \right\|_w < \infty$ , and therefore

$$\text{bs} - \lim_{n \rightarrow \infty} \int_a^b e^{-\beta t} u_n(t) dt = \int_a^b e^{-\beta t} u(t) dt,$$

that implies  $\int_a^b e^{-\beta t} u(t) dt \in C_p^w(\mathcal{O})$ , again due to Theorem 1. In particular, taking  $\beta = \frac{\alpha-\alpha_0}{2} > 0$ ,  $u(t) = \Phi_{\beta+\alpha_0}(t)h$ ,  $a = 0$ , and  $b = \infty$ , we obtain

$$\begin{aligned} \mathcal{R}_\alpha h(x) &= \int_0^\infty \Phi_\alpha(t)h(x) dt = \int_0^\infty e^{-(\alpha-\alpha_0)t} \Phi_{\alpha_0}(t)h(x) dt \\ &= \int_0^\infty e^{-\frac{\alpha-\alpha_0}{2}t} \Phi_{\frac{\alpha-\alpha_0}{2}+\alpha_0}(t)h(x) dt = \int_0^\infty e^{-\beta t} u(t)(x) dt. \end{aligned}$$

Thus,  $\mathcal{R}_\alpha h$  is in  $C_p(\mathcal{O})$  (resp. in  $C_p^w(\mathcal{O})$  when  $h \in C_p^w(\mathcal{O})$ ).  $\square$

The next result is a useful property of the integrals of semigroups that is very common in finite-dimensional spaces.

**Lemma 4.** *Let  $h \in C_p^w(\mathcal{O})$ . For any  $t_0 \geq 0$  we have*

$$\text{bs} - \lim_{t \downarrow 0} \frac{1}{t} \int_{t_0}^{t_0+t} \Phi_\alpha(s)h ds = \Phi_\alpha(t_0)h. \quad (40)$$

*Proof.* Let  $t_0 \geq 0$  and fix  $x \in \mathcal{O}$ . By Theorem 1 (c), we get that  $\frac{1}{t} \int_{t_0}^{t_0+t} \Phi_\alpha(s)h ds \in C_p^w(\mathcal{O})$ . Since  $t \mapsto \Phi_\alpha(t)h$  is continuous in seminorm, given  $\varepsilon > 0$  we consider  $\delta > 0$  such that  $|t_0 - s| < \delta$  implies  $p(\Phi_\alpha(s)h - \Phi_\alpha(t_0)h, x) < \varepsilon$ . Hence, if  $|t| \leq \delta$  then, by (32) we get

$$\begin{aligned} p\left(\frac{1}{t} \int_{t_0}^{t_0+t} \Phi_\alpha(s)h ds - \Phi_\alpha(t_0)h, x\right) &= p\left(\frac{1}{t} \int_{t_0}^{t_0+t} [\Phi_\alpha(s)h - \Phi_\alpha(t_0)h] ds, x\right) \\ &\leq \frac{1}{t} \int_{t_0}^{t_0+t} p(\Phi_\alpha(s)h - \Phi_\alpha(t_0)h, x) ds < \varepsilon. \end{aligned}$$

On the other hand, using (15) we get

$$\left\| \frac{1}{t} \int_{t_0}^{t_0+t} \Phi_\alpha(s)h ds \right\|_w \leq \frac{1}{t} \int_{t_0}^{t_0+t} \|\Phi_\alpha(s)h\|_w ds \leq \frac{1}{t} \int_{t_0}^{t_0+t} \|h\|_w ds = \|h\|_w. \quad (41)$$

Thus, we have proved  $\text{bs} - \lim_{t \downarrow 0} \frac{1}{t} \int_{t_0}^{t_0+t} \Phi_\alpha(s)h ds = \Phi_\alpha(t_0)h$ .  $\square$

Our next definition has to do with the differentiability of semigroups.

**Definition 3.** We say that  $t \mapsto \Phi_\alpha(t)h$  is boundedly differentiable in seminorm in a fixed point  $r \geq 0$  if the limit

$$\text{bs} - \lim_{t \rightarrow 0} \frac{\Phi_\alpha(t+r)h - \Phi_\alpha(r)h}{t}$$

exists in  $C_p^w(\mathcal{O})$ .

*Remark 5.(a)* If  $h \in C_p^w(\mathcal{O})$  and the above limit exists, then Theorem 1(c) ensures that this limit belongs to  $C_p^w(\mathcal{O})$ .

- (b) The boundedly differentiability in seminorm implies the pointwise differentiability; i.e., for each  $x \in \mathcal{O}$ ,  $\lim_{t \downarrow 0} \frac{\Phi_\alpha(t+r)h(x) - \Phi_\alpha(t)h(x)}{t}$ .

The next theorem shows a relation between the semigroup  $\Phi_\alpha$  and the infinitesimal generator  $\mathcal{A}_\alpha$ , among other important properties.

**Theorem 3.** Suppose that Assumption 1 is valid. Then, for each  $h \in D(\mathcal{A}_\alpha)$ , we have that  $\Phi_\alpha(t)h \in D(\mathcal{A}_\alpha)$  for all  $t > 0$ . Furthermore, the function  $t \mapsto \Phi_\alpha(t)h$  is boundedly differentiable in seminorm on  $(0, \infty)$ , and the following relation holds

$$-\frac{d}{dt}(\Phi_\alpha(t)h) = \mathcal{A}_\alpha \Phi_\alpha(t)h = \Phi_\alpha(t)\mathcal{A}_\alpha h, \quad \forall t > 0. \quad (42)$$

(The derivative on the left-hand side is understood in the sense of boundedly differentiability in seminorm.)

*Proof.* First note that

$$\frac{1}{s}(\Phi_\alpha(t)h - \Phi_\alpha(t+s)h) = \Phi_\alpha(t) \frac{1}{s}(h - \Phi_\alpha(s)h). \quad (43)$$

Next, by using the fact of  $s - \lim_{s \downarrow 0} \frac{1}{s}(h - \Phi_\alpha(s)h) = \mathcal{A}_\alpha h$  as long with Proposition 1, we have that

$$-\frac{d^+}{dt} \Phi_\alpha(t)h = s - \lim_{s \downarrow 0} \frac{1}{s}(\Phi_\alpha(t)h - \Phi_\alpha(t+s)h) = \Phi_\alpha(t)\mathcal{A}_\alpha h.$$

On the other hand, taking into account (43) we get

$$\begin{aligned} \left\| \frac{1}{s}(\Phi_\alpha(t)h - \Phi_\alpha(t+s)h) \right\|_w &\leq \frac{1}{s} \|\Phi_\alpha(t)(h - \Phi_\alpha(s)h)\|_w \\ &\leq \frac{1}{s} \|h - \Phi_\alpha(s)h\|_w \leq \sup_{s \geq 0} \frac{1}{s} \|h - \Phi_\alpha(s)h\|_w < \infty. \end{aligned}$$

The last inequality is due to the boundedly convergence in seminorm  $\text{bs} - \lim_{s \downarrow 0} \frac{1}{s}(h - \Phi_\alpha(s)h)$  in (19) applied to the definition of  $\mathcal{A}_\alpha$ . Hence  $\Phi_\alpha(t)h \in D(\mathcal{A}_\alpha)$  and  $\mathcal{A}_\alpha \Phi_\alpha(t)h = \Phi_\alpha(t)\mathcal{A}_\alpha h$ . In the same way it is possible to show that  $-\frac{d}{dt} \Phi_\alpha(t)h = \mathcal{A}_\alpha \Phi_\alpha(t)h$ , which proves (42).  $\square$

The next two results are crucial for our analysis: the first one shows the denseness of the domain  $D(\mathcal{A}_\alpha)$  into the space  $C_p^w(\mathcal{O})$ , whereas the second proves that the resolvent is the inverse operator of the generator; that is,  $\mathcal{A}_\alpha^{-1} = \mathcal{R}_\alpha$ .

**Theorem 4.** *Under the assumption of Theorem 3, the domain  $D(\mathcal{A}_\alpha)$  is dense in  $C_p^w(\mathcal{O})$  in the sense of the boundedly seminorm-convergence.*

*Proof.* Take  $h \in C_p^w(\mathcal{O})$  and define  $h_n := n \int_0^{\frac{1}{n}} \Phi_\alpha(s)h \, ds$ . By the proof of Lemma 4, we know that  $h_n \in C_p^w(\mathcal{O})$  and  $\text{bs-lim}_{n \rightarrow \infty} h_n = h$ , so it is sufficient to show that  $h_n \in D(\mathcal{A}_\alpha)$ . Indeed, using Fubini's Theorem, we have that

$$\begin{aligned} & \Phi_\alpha(t)h_n(x) \\ &= \mathbb{E}[e^{-\alpha t} n \int_0^{\frac{1}{n}} \Phi_\alpha(s)h(y(t, x)) \, ds] = n \int_0^{\frac{1}{n}} \mathbb{E}[e^{-\alpha t} \Phi_\alpha(s)h(y(t, x))] \, ds \\ &= n \int_0^{\frac{1}{n}} \Phi_\alpha(s+t)h(x) \, ds = n \int_t^{t+\frac{1}{n}} \Phi_\alpha(s)h(x) \, ds. \end{aligned}$$

Then, we obtain

$$\begin{aligned} \frac{1}{t}(h_n - \Phi_\alpha(t)h_n) &= n \left( \frac{1}{t} \int_0^{\frac{1}{n}} \Phi_\alpha(s)h \, ds - \frac{1}{t} \int_t^{t+\frac{1}{n}} \Phi_\alpha(s)h \, ds \right) \\ &= n \left( \frac{1}{t} \int_0^t \Phi_\alpha(s)h \, ds - \frac{1}{t} \int_{\frac{1}{n}}^{t+\frac{1}{n}} \Phi_\alpha(s)h \, ds \right). \end{aligned}$$

Using this last fact together with Lemma 4, we get  $s - \lim_{t \downarrow 0} \frac{1}{t}(h_n - \Phi_\alpha(t)h_n) = n(h - \Phi_\alpha(\frac{1}{n})h)$ . We have also the relation

$$\begin{aligned} \left\| \frac{1}{t}(h_n - \Phi_\alpha(t)h_n) \right\|_w &\leq \frac{n}{t} \left\| \int_0^t \Phi_\alpha(s)h \, ds \right\|_w + \frac{n}{t} \left\| \int_{\frac{1}{n}}^{t+\frac{1}{n}} \Phi_\alpha(s)h \, ds \right\|_w \\ &\leq \frac{n}{t} \int_0^t \|\Phi_\alpha(s)h\|_w \, ds + \frac{n}{t} \int_{\frac{1}{n}}^{t+\frac{1}{n}} \|\Phi_\alpha(s)h\|_w \, ds \leq 2n \|h\|_w. \end{aligned}$$

Hence,  $h_n \in D(\mathcal{A}_\alpha)$ .

**Theorem 5.** *Let Assumption 1 hold true. Then, for each  $\alpha > 0$ , the operator  $\mathcal{A}_\alpha$  from  $D(\mathcal{A}_\alpha)$  to  $C_p^w(\mathcal{O})$  is bijective. Besides, the following identity is satisfied*

$$\mathcal{A}_\alpha^{-1} = \mathcal{R}_\alpha.$$

*Proof.* Let us show first that  $\mathcal{A}_\alpha$  is surjective. Let  $h \in C_p^w(\mathcal{O})$  and  $s \geq 0$ . Using (33) we obtain

$$\Phi_\alpha(s)\mathcal{R}_\alpha h(x) = \mathcal{R}_\alpha\Phi_\alpha(s)h(x) = \int_0^\infty \Phi_\alpha(t+s)h(x) \, dt = \int_s^\infty \Phi_\alpha(t)h(x) \, dt.$$

Then,

$$\begin{aligned} \frac{1}{s}(\mathcal{R}_\alpha h - \Phi_\alpha(s)\mathcal{R}_\alpha h) &= \frac{1}{s} \int_0^\infty \Phi_\alpha(t)h(x) dt - \frac{1}{s} \int_s^\infty \Phi_\alpha(t)h(x) dt \\ &= \frac{1}{s} \int_0^s \Phi_\alpha(t)h(x) dt. \end{aligned}$$

By Lemma 4, we deduce that  $\text{bs-lim}_{s \downarrow 0} \frac{1}{s}(\mathcal{R}_\alpha h - \Phi_\alpha(s)\mathcal{R}_\alpha h) = h$  which implies  $\mathcal{R}_\alpha h \in D(\mathcal{A}_\alpha)$  and  $\mathcal{A}_\alpha \mathcal{R}_\alpha h = h$ , and therefore  $\mathcal{A}_\alpha$  is surjective. Now, let us show that  $\mathcal{A}_\alpha$  is injective. Take  $h \in D(\mathcal{A}_\alpha)$  such that  $\mathcal{A}_\alpha h = 0$ . By Theorem 3, we have that  $\frac{d}{dt}\Phi_\alpha(t)h(x) = -\Phi_\alpha(t)\mathcal{A}_\alpha h(x) = 0$  for all  $x \in \mathcal{O}$ , which implies that  $t \mapsto \Phi_\alpha(t)h(x)$  is a real constant. But,  $|\Phi_\alpha(t)h(x)| \leq e^{-(\alpha-\alpha_0)t} \|h\|_w w(x)$ , so,  $\lim_{t \rightarrow \infty} \Phi_\alpha(t)h(x) = 0$ . Moreover, we have  $\Phi_\alpha(0)h(x) = h(x)$  and then,  $h(x) = 0$  for all  $x \in \mathcal{O}$ . Thus, we have concluded that  $\mathcal{A}_\alpha$  is invertible with inverse given by  $\mathcal{R}_\alpha$ .  $\square$

As a direct consequence of both Theorems 5 and 3 we can get, for all  $h \in C_p^w(\mathcal{O})$ , the relation

$$\mathcal{R}_\alpha h - \Phi_\alpha(t)\mathcal{R}_\alpha h = \int_0^t \Phi_\alpha(s)h ds = \int_0^t \Phi_\alpha(s)\mathcal{A}_\alpha \mathcal{R}_\alpha h ds. \quad (44)$$

We conclude this section by providing some properties of the operators  $\mathcal{A}_\alpha$  and  $\mathcal{R}_\alpha$ .

**Proposition 3.** *For all  $h \in C_p^w(\mathcal{O})$  and  $\beta > 0$ , we have the next relation*

$$\text{bs-lim}_{\alpha \rightarrow \infty} \alpha \mathcal{R}_{\alpha+\beta} h = h. \quad (45)$$

*Proof.* By definition of the resolvent and (33) we get the resolvent equation:

$$\mathcal{R}_\alpha \mathcal{R}_\beta = \frac{1}{\alpha - \beta} (\mathcal{R}_\beta - \mathcal{R}_\alpha). \quad (46)$$

Next, we will prove  $\lim_{\alpha \rightarrow \infty} p(\alpha \mathcal{R}_\alpha h - h, x) = 0$  for all  $h \in C_p^w(\mathcal{O})$ . Let us assume first that  $h \in D(\mathcal{A}_\alpha)$  and let us take  $g \in C_p^w(\mathcal{O})$  such that  $h = \mathcal{R}_\beta g$ . We have

$$\alpha \mathcal{R}_\alpha h = \alpha \mathcal{R}_\alpha \mathcal{R}_\beta g = \frac{\alpha}{\alpha - \beta} (\mathcal{R}_\beta g - \mathcal{R}_\alpha g) = \frac{\alpha}{\alpha - \beta} h - \frac{\alpha}{\alpha - \beta} \mathcal{R}_\alpha g.$$

It is easy to see that  $\lim_{\alpha \rightarrow \infty} \left\| \frac{\alpha}{\alpha - \beta} h - h \right\|_w = 0$  and  $\lim_{\alpha \rightarrow \infty} \left\| \frac{\alpha}{\alpha - \beta} \mathcal{R}_\alpha g \right\|_w = 0$ , where the last limit is due to (31). Therefore,

$$\lim_{\alpha \rightarrow \infty} \|\alpha \mathcal{R}_\alpha h - h\|_w = 0.$$

By (20) we see that the above convergence in norm implies the convergence in seminorm:  $\text{s-lim}_{\alpha \rightarrow \infty} \alpha \mathcal{R}_\alpha h = h$ . Now, consider the general case  $h \in C_p^w(\mathcal{O})$ . Let  $h_n$  be a sequence in  $D(\mathcal{A}_\alpha)$  such that  $\text{bs-lim}_{n \rightarrow \infty} h_n = h$ . We have

$$|\alpha \mathcal{R}_\alpha h - h| \leq |\alpha \mathcal{R}_\alpha h - \alpha \mathcal{R}_\alpha h_n| + |\alpha \mathcal{R}_\alpha h_n - h_n| + |h_n - h|,$$

applying (30) to the above inequality we get

$$0 \leq p(\alpha \mathcal{R}_\alpha h - h, x) \leq \frac{\alpha}{\alpha - \alpha_0} p(h - h_n, x) + p(\alpha \mathcal{R}_\alpha h_n - h_n, x) + p(h_n - h, x).$$

Letting  $\alpha \rightarrow \infty$  and hence  $n \rightarrow \infty$  in the last inequality, we easily deduce that  $\lim_{\alpha \rightarrow \infty} p(\alpha \mathcal{R}_\alpha h - h, x) = 0$ ; in other words  $s - \lim_{\alpha \rightarrow \infty} \alpha \mathcal{R}_\alpha h = h$ . Moreover, by (31) we get  $\|\alpha \mathcal{R}_\alpha h\|_w \leq \alpha / (\alpha - \alpha_0) \|h\|_w$ , and so  $bs - \lim_{\alpha \rightarrow \infty} \alpha \mathcal{R}_\alpha h = h$ . It remains to show (45). For this purpose, let  $\beta > 0$  and note that  $\alpha \mathcal{R}_{\alpha+\beta} = (\alpha + \beta) \mathcal{R}_{\alpha+\beta} - \beta \mathcal{R}_{\alpha+\beta}$ , we know that  $bs - \lim_{\alpha \rightarrow \infty} (\alpha + \beta) \mathcal{R}_{\alpha+\beta} h = h$  and  $bs - \lim_{\alpha \rightarrow \infty} \beta \mathcal{R}_{\alpha+\beta} = 0$ , hence  $bs - \lim_{\alpha \rightarrow \infty} \alpha \mathcal{R}_{\alpha+\beta} h = h$ .  $\square$

**Proposition 4.** *Given  $\alpha > \alpha_0$  and  $\beta \geq 0$ , we have*

$$\mathcal{A}_{\alpha+\beta} = \mathcal{A}_\alpha + \beta I. \quad (47)$$

*Proof.* Let  $h \in D(\mathcal{A}_\alpha)$ . Then,

$$h - \Phi_{\alpha+\beta}(t)h = h - e^{-\beta t} \Phi_\alpha(t)h = h - \Phi_\alpha(t)h + (1 - e^{-\beta t}) \Phi_\alpha(t)h.$$

Multiplying by  $\frac{1}{t}$  the last expression, and hence letting  $t \downarrow 0$ , we get  $\mathcal{A}_{\alpha+\beta}h := bs - \lim_{t \downarrow 0} \frac{1}{t} (h - \Phi_{\alpha+\beta}(t)h) = \mathcal{A}_\alpha h + \beta h$ .  $\square$

### 3 The Optimal Stopping Problem

This section deals with an optimal stopping control problem whose dynamical system is of Markov type studied in Sect. 2. The total cost consists of both a running cost that is paid when the dynamic is still running and a stopping cost that must to be paid once the dynamic is stopped. The way to tackle this problem is through a characterization of the optimal cost (value function) regarded as the maximal subsolution of a variational inequality defined later. In addition, by means of this characterization, it is also possible to find the well-known continuation region that in turn provides the associated optimal stopping time viewed as the first hitting time of that region.

#### 3.1 The Statement of the Problem

In this subsection we start our analysis recalling some mathematical objects introduced in Sect. 2. Namely, we recall the underlying stochastic process, consisting of the homogeneous Markov process  $\{y(t, x)\}_{t \geq 0}$ ,  $x \in \mathcal{O}$  defined on the probability space  $\mathcal{E} := (\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ , with state space  $(\mathcal{O}, |\cdot|)$ , satisfying  $\mathbb{P}(y(0, x) = x) = 1$  as well as the properties established in Assumption 1.

We bring to mind that a stopping time is a random variable  $\tau$  with values in the no-negative real numbers set such that the event  $\{\tau \leq t\}$  is  $\mathcal{F}_t$  measurable for every  $t \geq 0$ , with  $\mathcal{F}_t$  the associated filtration to the space  $\mathcal{E}$ .

Let  $\mathcal{T}$  be the set consisting of all stopping times introduced in the above paragraph. With this in mind, for  $x \in \mathcal{O}$ ,  $f, \varphi \in C_p^w(\mathcal{O})$ ,  $\tau \in \mathcal{T}$ , and  $\alpha > \alpha_0 > 0$ , we define the following cost function

$$J(x, \tau) := \mathbb{E} \left[ \int_0^\tau f(y(t, x)) e^{-\alpha t} dt + \varphi(y(\tau, x)) e^{-\alpha \tau} \mathbf{1}_{\tau \infty} \right], \quad (48)$$

where as mentioned above,  $f$  and  $\varphi$  represent the running and stopping cost per unit of time respectively, and  $e^{-\alpha \cdot}$  denotes the discount factor at each instant of time.

The optimal cost, also known as the *value function*, is then defined as

$$\hat{u}(x) = \inf_{\tau \in \mathcal{T}} J(x, \tau). \quad (49)$$

We will say that the random variable  $\hat{\tau} \in \mathcal{T}$  is an *optimal stopping time* if it minimizes the cost (48) in the following way

$$\hat{u}(x) = J(x, \hat{\tau}). \quad (50)$$

One of the goals of this section will consist to showing that the value function  $\hat{u}$  defined in (49) does exist in  $C_p^w(\mathcal{O})$ . Furthermore, this function satisfies the next variational inequality (VI) in the integral (or weak) form:

$$\hat{u} \leq \varphi, \quad \hat{u} \leq \int_0^t \Phi_\alpha(s) f ds + \Phi_\alpha(t) \hat{u}, \quad \forall t \geq 0. \quad (51)$$

### 3.2 Penalized Method

We start our analysis by studying an ancillary problem so-called *penalized problem*. This problem consists of searching for a unique solution of the following *penalized equations*

$$\mathcal{A}_\alpha u_\varepsilon + \frac{1}{\varepsilon} (u_\varepsilon - \varphi)^+ = f, \quad \text{for each } \varepsilon > 0, \quad (52)$$

with

$$(u_\varepsilon - \varphi)^+ = \begin{cases} u_\varepsilon - \varphi, & \text{if } u_\varepsilon - \varphi \geq 0; \\ 0, & \text{if } u_\varepsilon - \varphi \leq 0. \end{cases}$$

Our goal is to prove that one subsolution of the inequality (51) can be characterized as the limit as  $\varepsilon \downarrow 0$  of the sequence of solutions  $u_\varepsilon$  associated to (52). This limit function will be the “good one” for us.

Note that  $(u_\varepsilon - \varphi)^+ = u_\varepsilon - (u_\varepsilon \wedge \varphi)$ . Hence, Proposition 4 together with (52), imply

$$\mathcal{A}_{\alpha + \frac{1}{\varepsilon}} u_\varepsilon = f + \frac{1}{\varepsilon} (u_\varepsilon \wedge \varphi). \quad (53)$$

Applying  $\mathcal{R}_{\alpha + \frac{1}{\varepsilon}}$  to the last equation we get

$$u_\varepsilon = \mathcal{R}_{\alpha + \frac{1}{\varepsilon}} (f + \frac{1}{\varepsilon} (u_\varepsilon \wedge \varphi)). \quad (54)$$

As mentioned earlier, we will prove that  $u_0 := \text{s-lim}_{\varepsilon \downarrow 0} u_\varepsilon$  verifies the VI (51) as well as its corresponding regularity. To this end, we need the following technical result.

**Lemma 5.** *The following inequality holds for any measurable functions  $f, g, h$  from  $\mathcal{O}$  to  $\mathbb{R}$ :*

$$|f \wedge h - g \wedge h| \leq |f - g|.$$

*Proof.* We have both  $-|f - g| + g \wedge h \leq f - g + g = f$  and  $-|f - g| + g \wedge h \leq h$  that imply  $-|f - g| + g \wedge h \leq f \wedge h$ . Analogously, we have  $-|f - g| + f \wedge h \leq g \wedge h$ , and joining the two obtained inequalities we get  $|f \wedge h - g \wedge h| \leq |f - g|$ .  $\square$

**Theorem 6.** *Assume that  $f, \varphi \in C_p^w(\mathcal{O})$ . Then, Assumption 1 implies the following.*

- (a) *There exists a unique solution  $u_\varepsilon \in D(\mathcal{A}_\alpha)$  of the penalized equation (52) for each  $\varepsilon > 0$ .*
- (b) *For all  $0 < \varepsilon' < \varepsilon$  we have that*

$$0 \leq u_\varepsilon - u_{\varepsilon'} \leq (u_\varepsilon - \varphi)^+ \leq |\mathcal{R}_{\alpha+\frac{1}{\varepsilon}} f + \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} \varphi - \varphi|. \quad (55)$$

Furthermore, there exists the limit  $u_0 := \text{s-lim}_{\varepsilon \downarrow 0} u_\varepsilon$  and therefore,  $u_0 \in C_p(\mathcal{O})$ .

*Proof.* First, we will show the existence of a unique solution  $u_\varepsilon$  of the penalized problem. Namely, based on (54), we define the nonlinear operator  $T_\varepsilon : B_w(\mathcal{O}) \rightarrow B_w(\mathcal{O})$  given by  $T_\varepsilon h := \mathcal{R}_{\alpha+1/\varepsilon}(f + \frac{1}{\varepsilon}(h \wedge \varphi))$ . We will prove that  $T_\varepsilon$  is a contraction map. Indeed, as  $h, g \in B_w(\mathcal{O})$ , we have

$$T_\varepsilon h - T_\varepsilon g = \frac{1}{\varepsilon} \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(h \wedge \varphi - g \wedge \varphi).$$

Using the monotony of the resolvent together with Lemma 5 we get

$$\frac{1}{\varepsilon} |\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(h \wedge \varphi - g \wedge \varphi)| \leq \frac{1}{\varepsilon} \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} |h \wedge \varphi - g \wedge \varphi| \leq \frac{1}{\varepsilon} \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} |h - g|.$$

Now use (31) to obtain

$$\|T_\varepsilon h - T_\varepsilon g\|_w \leq \frac{\frac{1}{\varepsilon}}{\alpha - \alpha_0 + \frac{1}{\varepsilon}} \|h - g\|_w.$$

We know that  $\frac{\frac{1}{\varepsilon}}{\alpha - \alpha_0 + \frac{1}{\varepsilon}} < 1$ . Then  $T_\varepsilon$  is a contraction map on the Banach space  $B_w(\mathcal{O})$ , so there exist a unique  $u_\varepsilon$  in  $B_w(\mathcal{O})$  such that  $T_\varepsilon u_\varepsilon = u_\varepsilon$ , this implies that  $u_\varepsilon$  solves (52). Moreover, we have that  $\lim_{n \rightarrow \infty} \|T_\varepsilon^n h - u_\varepsilon\|_w = 0$  that implies convergence in seminorm.

On the other hand, using the fact that  $f, \varphi \in C_p^w(\mathcal{O})$ , and taking  $h \in C_p^w(\mathcal{O})$ , all together allow us to apply Theorem 5 to claim that  $T_\varepsilon h = \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}(h \wedge \varphi)) \in D(\mathcal{A}_\alpha)$ . Iterating  $n$ -times the operator  $T_\varepsilon$ , it is easy to see that  $T_\varepsilon^n h \in$

$D(\mathcal{A}_\alpha)$  for all  $n \in \mathbb{N}$ . Hence, in virtue of Theorem 1 we have  $u_\varepsilon \in C_p^w(\mathcal{O})$ , yielding that  $u_\varepsilon = \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}(u_\varepsilon \wedge \varphi)) \in D(\mathcal{A}_\alpha)$ .

Let us prove now the inequalities (55). Namely, let  $0 < \varepsilon' < \varepsilon$ . Then from (53) we obtain

$$\begin{aligned} \mathcal{A}_{\alpha+\frac{1}{\varepsilon}} u_{\varepsilon'} &= \mathcal{A}_{\alpha+\frac{1}{\varepsilon'}} u_{\varepsilon'} + \left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon'}\right) u_{\varepsilon'} = f + \frac{1}{\varepsilon'}(u_{\varepsilon'} \wedge \varphi) + \left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon'}\right) u_{\varepsilon'} \\ &= f + \frac{1}{\varepsilon'} u_{\varepsilon'} - \frac{1}{\varepsilon'}(u_{\varepsilon'} - \varphi)^+ + \left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon'}\right) u_{\varepsilon'} \\ &= f - \frac{1}{\varepsilon'}(u_{\varepsilon'} - \varphi)^+ + \frac{1}{\varepsilon} u_{\varepsilon'} \leq f - \frac{1}{\varepsilon}(u_{\varepsilon'} - \varphi)^+ + \frac{1}{\varepsilon} u_{\varepsilon'} = f + \frac{1}{\varepsilon}(u_{\varepsilon'} \wedge \varphi). \end{aligned}$$

Applying  $\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}$  to the last inequality we obtain

$$u_{\varepsilon'} \leq T_\varepsilon u_{\varepsilon'}.$$

Iterating, we get  $u_{\varepsilon'} \leq T_\varepsilon^n u_{\varepsilon'}$ . Therefore, letting  $n \rightarrow \infty$  we obtain  $u_{\varepsilon'} \leq u_\varepsilon$ .

Next, we will show that  $u_\varepsilon - u_{\varepsilon'} \leq (u_\varepsilon - \varphi)^+ \leq |\mathcal{R}_{\alpha+\frac{1}{\varepsilon}} f + \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} \varphi - \varphi|$ . Namely, assuming  $u_{\varepsilon'} \geq \varphi$  we get  $u_\varepsilon - u_{\varepsilon'} \leq u_\varepsilon - \varphi \leq (u_\varepsilon - \varphi)^+$ . Otherwise, if  $\varphi \geq u_{\varepsilon'}$  then from (52) we obtain  $\mathcal{A}_\alpha(u_\varepsilon - u_{\varepsilon'}) = -\frac{1}{\varepsilon}(u_\varepsilon - \varphi)^+ \leq 0$ , and applying  $\mathcal{R}_\alpha$  to the last inequality we get  $u_\varepsilon - u_{\varepsilon'} \leq 0 \leq (u_\varepsilon - \varphi)^+$ . Moreover, from (54) we obtain

$$\begin{aligned} u_\varepsilon - \varphi &= \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}(u_\varepsilon \wedge \varphi)) - \varphi \\ &= \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}(u_\varepsilon \wedge \varphi) - \frac{1}{\varepsilon}\varphi) + \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\varphi - \varphi \\ &= \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f - \frac{1}{\varepsilon}(\varphi - u_\varepsilon)^+) + \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\varphi - \varphi \leq \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} f + \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\varphi - \varphi. \end{aligned} \quad (56)$$

Hence,

$$0 \leq u_\varepsilon - u_{\varepsilon'} \leq (u_\varepsilon - \varphi)^+ \leq |\mathcal{R}_{\alpha+\frac{1}{\varepsilon}} f + \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\varphi - \varphi|. \quad (57)$$

Let  $\varepsilon > \varepsilon' > 0$ . Using  $u_{\varepsilon'} \leq u_\varepsilon$  and  $0 \leq u_\varepsilon - u_{\varepsilon'} \leq (u_\varepsilon - \varphi)^+$ , we obtain that there exists the pointwise monotone limit  $u_0 := \lim_{\varepsilon \downarrow 0} u_\varepsilon$  and  $u_0 > -\infty$ . Letting  $\varepsilon' \downarrow 0$  in (57), we get  $0 \leq u_\varepsilon - u_0 \leq |\mathcal{R}_{\alpha+\frac{1}{\varepsilon}} f + \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\varphi - \varphi|$ . Thus, in virtue of the relations (30) and (45) we get

$$p(u_\varepsilon - u_0, x) \leq \frac{1}{\alpha + \frac{1}{\varepsilon} - \alpha_0} p(f, x) + p\left(\frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\varphi - \varphi, x\right) \rightarrow 0, \quad \text{as } \varepsilon \downarrow 0,$$

and so  $s - \lim_{\varepsilon \downarrow 0} u_\varepsilon = u_0$ ; this implies that  $u_0 \in C_p(\mathcal{O})$  after using Theorem 1(b).  $\square$

### 3.3 Variational Inequalities

Let  $f, \varphi \in C_p^w(\mathcal{O})$ . We say that  $u \in C_p^w(\mathcal{O})$  satisfies the variational inequalities (VI) if:

$$\begin{cases} u \leq \int_0^t \Phi_\alpha(s)f ds + \Phi_\alpha(t)u, \forall t \geq 0; \\ u \leq \varphi. \end{cases} \quad (58)$$

Any function  $u \in C_p^w(\mathcal{O})$  that satisfies the VI above, will be referred to as a *subsolution*.

On the other hand, by definition of  $w$  in (3), it is obvious that  $w \in B_w(\mathcal{O})$ . For the later purposes, we need the next assumption in order to guarantee that the subsolution of interest associated to (58) is regular enough.

**Assumption 2.** *We suppose that  $w$  defined in (3), belongs to  $C_p^w(\mathcal{O})$ .*

*Remark 6.* The above assumption is verified in particular models, see for instance, Menaldi [24, 25] or Menaldi and Sritharan [28], where the authors use a polynomial function of type  $w(x) = k_1(k_2 + |x|^2)^p$ , for some constants  $k_1 \geq 1$ ,  $k_2 \geq 0$ .

Now let  $u := \mathcal{R}_\alpha f - (\|\varphi\|_w + \frac{1}{\alpha - \alpha_0} \|f\|_w)w \in C_p^w(\mathcal{O})$ . Note that  $\mathcal{R}_\alpha f - \frac{1}{\alpha - \alpha_0} \|f\|_w w \leq 0$  because of (31), then  $u \leq -\|\varphi\|_w w \leq \varphi$ . We also have that

$$\Phi_\alpha(t)u \geq \Phi_\alpha(t)\mathcal{R}_\alpha f - (\|\varphi\|_w + \frac{1}{\alpha - \alpha_0} \|f\|_w)w.$$

Using (44), we obtain

$$\begin{aligned} \int_0^t \Phi_\alpha(s)f ds + \Phi_\alpha(t)u &= \mathcal{R}_\alpha f - \Phi_\alpha(t)\mathcal{R}_\alpha f + \Phi_\alpha(t)u \\ &\geq \mathcal{R}_\alpha f - (\|\varphi\|_w + \frac{1}{\alpha - \alpha_0} \|f\|_w)w = u. \end{aligned}$$

Therefore, we have proved that  $u \in C_p^w(\mathcal{O})$  defined in the previous paragraph satisfies the VI (58).

We will see next that the limit function  $u_0$  obtained in the past subsection, is the maximal subsolution on  $C_p^w(\mathcal{O})$  of the VI (58) and  $\|u_0\|_w < \infty$  as it is established in the following theorem.

**Theorem 7.** *Under Assumptions 1 and 2, the limit function  $u_0$  introduced in Theorem 6 verifies the VI (58). Moreover, every  $u \in C_p^w(\mathcal{O})$  that is also a subsolution of (58) satisfies  $u \leq u_0$ ; as a consequence  $u_0 \in C_p^w(\mathcal{O})$ .*

*Proof.* From (52) and (44), we obtain

$$\begin{aligned} u_\varepsilon &= \mathcal{R}_\alpha(f - \frac{1}{\varepsilon}(u_\varepsilon - \varphi)^+) = \int_0^t \Phi_\alpha(s)(f - \frac{1}{\varepsilon}(u_\varepsilon - \varphi)^+) ds + \Phi_\alpha(t)u_\varepsilon \\ &\leq \int_0^t \Phi_\alpha(s)f ds + \Phi_\alpha(t)u_\varepsilon. \end{aligned}$$

Moreover, for each  $t \geq 0$ , we have that  $\Phi_\alpha(t)u_\varepsilon \rightarrow \Phi_\alpha(t)u_0$  pointwise as  $\varepsilon \downarrow 0$ , because  $p(\Phi_\alpha(t)u_\varepsilon - \Phi_\alpha(t)u_0, x) \leq p(u_\varepsilon - u_0, x) \rightarrow 0$ , as  $\varepsilon \downarrow 0$ . So, letting  $\varepsilon \downarrow 0$  in the last inequality we get

$$u_0 \leq \int_0^t \Phi_\alpha(s)f ds + \Phi_\alpha(t)u_0.$$

On the other hand from (54) we have

$$u_\varepsilon = \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}(u_\varepsilon \wedge \varphi)) \leq \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}\varphi). \quad (59)$$

In virtue of (30) and (45), we have

$$\begin{aligned} & p(\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}\varphi) - \varphi, x) \\ & \leq \frac{1}{\alpha + \frac{1}{\varepsilon} - \alpha_0} p(f, x) + p\left(\frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\varphi - \varphi, x\right) \rightarrow 0, \quad \text{as } \varepsilon \downarrow 0. \end{aligned}$$

The last relation implies in particular that  $\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f + \frac{1}{\varepsilon}\varphi) \rightarrow \varphi$  pointwise, as  $\varepsilon \downarrow 0$ . Hence, letting  $\varepsilon \downarrow 0$  in (59) we get

$$u_0 \leq \varphi,$$

which implies that  $u_0$  satisfies (58).

It only remains to show that  $u_0$  the maximal subsolution. Indeed, take  $u \in C_p^w(\mathcal{O})$  that satisfies (58). Then,  $u$  satisfies:  $u - \Phi_\alpha(t) \leq \int_0^t \Phi_\alpha(s)f ds$ . Apply then  $\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}$  to both sides of the last inequality and hence multiply by  $\frac{1}{t}$ , so that

$$\frac{1}{t}(\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}u - \Phi_\alpha(t)\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}u) \leq \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}\frac{1}{t} \int_0^t \Phi_\alpha(s)f ds.$$

The commutative property between  $\Phi_\alpha(t)$  and  $\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}$  is due to (33). Using again (44), the fact that  $\alpha \mapsto \mathcal{R}_\alpha$  is a family of commutative operators given in (46), as well as the relation (33), we deduce

$$\frac{1}{t}(\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}u - \Phi_\alpha(t)\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}u) \leq \frac{1}{t}(\mathcal{R}_\alpha\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}f - \Phi_\alpha(t)\mathcal{R}_\alpha\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}f),$$

thus letting  $t \downarrow 0$  we get

$$\mathcal{A}_\alpha\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}u \leq \mathcal{A}_\alpha\mathcal{R}_\alpha\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}f = \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}f. \quad (60)$$

In virtue of Proposition 4, we know that  $(\frac{1}{\varepsilon}I + \mathcal{A}_\alpha)\mathcal{R}_{\alpha+\frac{1}{\varepsilon}} = \mathcal{A}_{\alpha+\frac{1}{\varepsilon}}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}} = I$ , then

$$\mathcal{A}_\alpha\mathcal{R}_{\alpha+\frac{1}{\varepsilon}} = I - \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}. \quad (61)$$

This last fact, together with the relation  $u = u \wedge \varphi$  (recall that  $u \leq \varphi$ ), and (60) yield that

$$u \leq \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}f + \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}u = \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}f + \frac{1}{\varepsilon}\mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(u \wedge \varphi) = T_\varepsilon u.$$

Iterating the last expression, we obtain that  $u \leq T_\varepsilon^n u$ , implying that  $u \leq u_\varepsilon$ . Letting  $\varepsilon \downarrow 0$ , we obtain  $u \leq u_0$ .

Finally, take  $u \in C_p^w(\mathcal{O})$  that satisfies the VI (58) (we know that there exist at least a function in  $C_p^w(\mathcal{O})$  satisfying the VI). Then we have  $u \leq u_0 \leq \varphi$  that implies  $|u_0| \leq |u_0 - u| + |u| \leq |\varphi - u| + |u|$ . Since  $\varphi - u$  and  $u$  belong to  $C_p^w(\mathcal{O})$  we get that  $\|u_0\|_w \leq \|\varphi - u\|_w + \|u\|_w < \infty$ . So, we conclude that  $u_0 \in C_p^w(\mathcal{O})$  is the maximal subsolution on  $C_p^w(\mathcal{O})$  of the VI (58).  $\square$

## 4 Solution of the Stopping Problem

In this section we will analyze the optimal control problem through the solution of the VI (58). In addition, we provide the way to find an optimal stopping time in terms of so-named *continuation region* or *contact set*.

To begin with, we will show the next result regarding the *strong Markov property* of the process  $y(t, x)$ . Its proof has been inspired from Proposition 8.9 and Theorem 19.17 in [21].

**Proposition 5.** *The Markov process  $\{y(t, x)\}_{t \geq 0}$  satisfies the strong Markov property in the following sense: for all stopping time  $\tau \in \mathcal{T}$  and  $h \in B(\mathcal{O})$  we have*

$$\mathbb{E}[h(y(s + \tau, x))|\mathcal{F}_\tau] = \mathbb{E}[h(y(s, y(\tau, x)))], \quad (62)$$

where  $\mathcal{F}_\tau$  is the  $\sigma$ -algebra generated of events  $A \in \mathcal{F}$  for which  $A \cap \{\tau \leq t\} \in \mathcal{F}_t$  for every  $t \geq 0$ .

*Proof.* First, let us suppose that  $\tau$  has a denumerable state space  $D$  in  $\bar{\mathbb{R}}$ . Then, we have that

$$\begin{aligned} \mathbb{E}[e^{-\alpha(t+\tau)}h(y(t + \tau, x))|\mathcal{F}_\tau] &= \sum_{s \in D} \mathbf{1}_{\tau=s} \mathbb{E}[e^{-\alpha(t+\tau)}h(y(t + \tau, x))|\mathcal{F}_\tau] \\ &= \sum_{s \in D} \mathbf{1}_{\tau=s} \mathbb{E}[e^{-\alpha(t+\tau)}h(y(t + s, x))|\mathcal{F}_s] \\ &= \sum_{s \in D} \mathbf{1}_{\tau=s} \mathbb{E}[e^{-\alpha(t+\tau)}h(y(t, y(s, x)))] = \mathbb{E}[e^{-\alpha(t+\tau)}h(y(t, y(\tau, x)))] . \end{aligned}$$

Note that every conditional expectation above is well defined since

$$\mathbb{E}[e^{-\alpha(t+\tau)}|h(y(t + \tau, x))|] \leq \mathbb{E}[\sup_{s \geq 0} e^{-\alpha s}|h(y(s, x))|] < \infty .$$

In the general case, by Lemma 7.4 in [21] we can take a sequence of stopping times  $\tau_n$  with denumerable state space such that  $\tau_n \downarrow \tau$ . So, we have that

$$\mathbb{E}[h(y(t + \tau_n, x))|\mathcal{F}_{\tau_n}] = \mathbb{E}[h(y(t, y(\tau_n, x)))]$$

which implies

$$\begin{aligned} \mathbb{E}[e^{-\tau_n}h(y(t + \tau_n, x))|\mathcal{F}_{\tau_n}] &= e^{-\alpha\tau_n} \mathbb{E}[h(y(t, y(\tau_n, x)))] \\ &= e^{-\alpha(\tau_n-t)} \Phi_\alpha(t) h(y(\tau_n, x)) . \end{aligned} \quad (63)$$

By the right continuity of  $s \mapsto \Phi_\alpha(t)h(y(s, x))$  and the fact that  $\tau_n \downarrow \tau$  we get  $e^{-\alpha(\tau_n-t)} \Phi_\alpha(t)h(y(\tau_n, x)) \rightarrow e^{-\alpha(\tau-t)} \Phi_\alpha(t)h(y(\tau, x))$  when  $n \rightarrow \infty$ . By Lemma 7.3 in [21], we have  $\mathcal{F}_\tau = \cap_{n \in \mathbb{N}} \mathcal{F}_{\tau_n}$  which together with Theorem 45 in [12] give us

$$\mathbb{E}[e^{-\tau_n}h(y(t + \tau_n, x))|\mathcal{F}_{\tau_n}] \rightarrow e^{-\alpha\tau} \mathbb{E}[h(y(t + \tau, x))|\mathcal{F}_\tau]$$

when  $n \rightarrow \infty$ . Using this last fact along with (63) we conclude that

$$\mathbb{E}[h(y(t + \tau, x))|\mathcal{F}_\tau] = \mathbb{E}[h(y(t, y(\tau, x)))] .$$

□

In order to characterize the optimal stopping time as the hitting time of certain region of the state space, we will also need the following property of our process  $\{y(t, x)\}_{t \geq 0}$ .

**Assumption 3.** *The process  $\{y(t, x)\}_{t \geq 0}$  is quasi-left continuous, that is, for every stopping time  $\tau$  and any sequence of stopping times  $\tau_1, \tau_2, \dots$  such that  $\tau_n \uparrow \tau$  we have that  $y(\tau_n, x) \rightarrow y(\tau, x)$   $\mathbb{P}$ -a.s. on  $\{\tau < \infty\}$ .*

*Remark 7.(a)* Assumption 3 is a little variation of the Hunt process definition.  
*(b)* It is well-known that a Markov process associated to a strong Feller semi-group is a Hunt process—for further details, see Chung [9], Chapter 3.

Let us now establish the main result of this section.

**Theorem 8.** *Under Assumptions 1, 2, and 3, the following statements hold true.*

- (a) *The optimal cost  $\hat{u}$  in (49) is equal to the limit function  $u_0$ .*
- (b) *The optimal stopping time can be regarded as the first hitting time of the so-called continuation region (a.k.a. contact set). That is, for all  $x \in \mathcal{O}$ ,*

$$\hat{\tau}(x) := \inf\{t \geq 0 : \hat{u}(y(t, x)) = \varphi(y(t, x))\} \quad (\text{continuation region}), \quad (64)$$

*satisfying  $\hat{u}(x) = J(x, \hat{\tau}(x))$ .*

- (c) *If the stopping cost  $\varphi \in D(\mathcal{A}_\alpha)$ , then*

$$\mathcal{R}_\alpha(f \wedge \mathcal{A}_\alpha \varphi) \leq \hat{u} \leq \mathcal{R}_\alpha f \wedge \varphi. \quad (65)$$

*Proof.* (a) Take  $\tau \in \mathcal{T}$ , where  $\mathcal{T}$  is the set of stopping times defined at the beginning of the section. Moreover, define  $u := f - \frac{1}{\varepsilon}(u_\varepsilon - \varphi)^+$ . Then, from (52) we have

$$\begin{aligned} u_\varepsilon(x) &= \mathcal{R}_\alpha u(x) = \int_0^\infty \mathbb{E}[e^{-\alpha s} u(y(s, x))] ds = \mathbb{E}\left[\int_0^\infty e^{-\alpha s} u(y(s, x)) ds\right] \\ &= \mathbb{E}\left[\int_0^\tau e^{-\alpha s} u(y(s, x)) ds\right] + \mathbb{E}\left[\int_\tau^\infty e^{-\alpha s} u(y(s, x)) ds\right]. \end{aligned} \quad (66)$$

Let us analyze the last term of (66). We have that  $\int_0^\infty \mathbb{E}[e^{-\alpha(s+\tau)} |u(y(s+\tau, x))|] ds \leq \int_0^\infty e^{-(\alpha-\alpha_0)s} p(h, x) ds < \infty$ . Then using Fubini Theorem and strong Markov property (62) we get

$$\begin{aligned} \mathbb{E}\left[\int_\tau^\infty e^{-\alpha s} u(y(s, x)) ds\right] &= \int_0^\infty \mathbb{E}[e^{-\alpha(s+\tau)} u(y(s+\tau, x))] ds \\ &= \int_0^\infty \mathbb{E}[\mathbb{E}[e^{-\alpha(s+\tau)} u(y(s+\tau, x)) | \mathcal{F}_\tau]] ds = \int_0^\infty \mathbb{E}[e^{-\alpha\tau} \Phi_\alpha(s) u(y(\tau, x))] ds \\ &= \mathbb{E}[e^{-\alpha\tau} \mathcal{R}_\alpha u(y(\tau, x))] = \mathbb{E}[e^{-\alpha\tau} u_\varepsilon(y(\tau, x))]. \end{aligned} \quad (67)$$

Hence, in virtue of (66) and (67), we have that

$$\begin{aligned} u_\varepsilon(x) &= \mathbb{E}\left[\int_0^\tau e^{-\alpha s} u(y(s, x)) ds\right] + \mathbb{E}\left[\int_\tau^\infty e^{-\alpha s} u(y(s, x)) ds\right] \\ &= \mathbb{E}\left[\int_0^\tau e^{-\alpha s} u(y(s, x)) ds\right] + \mathbb{E}[e^{-\alpha \tau} u_\varepsilon(y(\tau, x))] \\ &= \mathbb{E}\left[\int_0^\tau e^{-\alpha s} [f - \frac{1}{\varepsilon}(u_\varepsilon - \varphi)^+](y(s, x)) ds + e^{-\alpha \tau} u_\varepsilon(y(\tau, x))\right]. \end{aligned} \quad (68)$$

On the other hand, from the definition of the seminorm  $p$ , it is evident that  $\mathbb{E}[e^{-\alpha \tau}|u_\varepsilon(y(\tau, x)) - u_0(y(\tau, x))|] \leq p(u_\varepsilon - u_0, x) \rightarrow 0$  when  $\varepsilon \downarrow 0$ , where the last convergence is due to Theorem 6. Then, using this last fact along with (68) and Theorem 7, we obtain

$$\begin{aligned} u_0(x) &= \lim_{\varepsilon \downarrow 0} u_\varepsilon(x) \leq \lim_{\varepsilon \downarrow 0} \mathbb{E}\left[\int_0^\tau e^{-\alpha s} f(y(s, x)) ds + e^{-\alpha \tau} u_\varepsilon(y(\tau, x))\right] \\ &= \mathbb{E}\left[\int_0^\tau e^{-\alpha s} f(y(s, x)) ds + e^{-\alpha \tau} u_0(y(\tau, x))\right] \\ &\leq \mathbb{E}\left[\int_0^\tau e^{-\alpha s} f(y(s, x)) ds + e^{-\alpha \tau} \varphi(y(\tau, x))\right] = J(x, \tau). \end{aligned}$$

Therefore,  $u_0 \leq \hat{u}$ , after applying the infimum over all  $\tau$  in last rightmost term.

On the other hand, for each  $\varepsilon > 0$ , let us consider the stopping time

$$\tau_\varepsilon(x) := \inf\{t \geq 0 : u_\varepsilon(y(t, x)) \geq \varphi(y(t, x))\}.$$

Now take a sequence  $\{t_n\}_{n \in \mathbb{N}}$  in  $[0, \infty)$  such that  $t_n \downarrow \tau_\varepsilon(x)$  (pointwise w.r.t.  $\omega \in \Omega$ ) and  $u_\varepsilon(y(t_n, x)) \geq \varphi(y(t_n, x))$ . Since  $t \mapsto u_\varepsilon(y(t, x)) - \varphi(y(t, x))$  is continuous a.s., we obtain  $u_\varepsilon(y(\tau_\varepsilon(x), x)) \geq \varphi(y(\tau_\varepsilon(x), x))$  when  $t_n \downarrow \tau_\varepsilon(x)$ . Then by (68), we deduce

$$\begin{aligned} u_\varepsilon(x) &= \mathbb{E}\left[\int_0^{\tau_\varepsilon} e^{-\alpha s} [f - \frac{1}{\varepsilon}(u_\varepsilon - \varphi)^+](y(s, x)) ds + e^{-\alpha \tau_\varepsilon} u_\varepsilon(y(\tau_\varepsilon, x))\right] \\ &= \mathbb{E}\left[\int_0^{\tau_\varepsilon} e^{-\alpha s} f(y(s, x)) ds + e^{-\alpha \tau_\varepsilon} \varphi(y(\tau_\varepsilon, x))\right] = J(\tau_\varepsilon, x) \geq \hat{u}(x). \end{aligned}$$

This shows that  $u_0 = \lim_{\varepsilon \downarrow 0} u_\varepsilon \geq \hat{u}$ . Joining the pieces, we conclude that  $u_0 = \hat{u}$ .

(b) Given  $\varepsilon > \varepsilon'$ , we know by the proof of Theorem 6 that  $u_\varepsilon \geq u_{\varepsilon'}$ , then we have the expression

$$\{s \geq 0 : u_{\varepsilon'}(y(s, x)) \geq \varphi(y(s, x))\} \subseteq \{s \geq 0 : u_\varepsilon(y(s, x)) \geq \varphi(y(s, x))\},$$

implying  $\tau_\varepsilon \leq \tau_{\varepsilon'}$ . So, there exists the monotone limit  $\tau_\varepsilon \uparrow \tau_0$ , as  $\varepsilon \downarrow 0$ . Also, because of the continuity of  $s \mapsto u_0(y(s, x))$  on  $[0, \infty)$  a.s., we have that  $\varphi(y(\hat{\tau}, x)) = u_0(y(\hat{\tau}, x)) \leq u_\varepsilon(y(\hat{\tau}, x))$ , where  $\hat{\tau}$  was defined in (64). Hence, we obtain  $\tau_\varepsilon \leq \hat{\tau}$  that implies  $\tau_0 \leq \hat{\tau}$ .

On the other hand, the fact  $s - \lim u_\varepsilon = u_0$  gives us the existence of a sequence  $\varepsilon_n, n \in \mathbb{N}$ , such that  $\varepsilon_n \downarrow 0$  and

$$\lim_{n \rightarrow \infty} \sup_{s \geq 0} e^{-\alpha_0 s} |u_{\varepsilon_n}(y(s, x)) - u_0(y(s, x))| = 0, \quad a.s., \quad (69)$$

where this last assertion is due to Lemma 1. Also, because of  $u_{\varepsilon_n} \geq u_0$ , we have that

$$0 \leq u_{\varepsilon_n}(y(\tau_{\varepsilon_n}, x)) - u_0(y(\tau_{\varepsilon_n}, x)) \leq e^{\alpha_0 \tau_{\varepsilon_n}} \sup_{s \geq 0} e^{-\alpha_0 s} |u_{\varepsilon_n}(y(s, x)) - u_0(y(s, x))|. \quad (70)$$

If  $\tau_0 = \infty$  then  $\infty = \tau_0 \leq \hat{\tau}$ , so  $\tau_0 = \hat{\tau}$ . Now, suppose  $\tau_0 < \infty$  a.s., then we have that  $e^{\alpha_0 \tau_{\varepsilon_n}} \rightarrow e^{\alpha_0 \tau_0}$ , when  $n \rightarrow \infty$ . Hence, using (69), the right hand side of inequality (70) converges to 0 when  $n \rightarrow \infty$ . Using Assumption 3 we deduce

$$\varphi(y(\tau_0, x)) = \lim_{n \rightarrow \infty} \varphi(y(\tau_{\varepsilon_n}, x)) \leq \lim_{n \rightarrow \infty} u_{\varepsilon_n}(y(\tau_{\varepsilon_n}, x)) = u_0(y(\tau_0, x)), \quad a.s.$$

Thus, the definition of  $\hat{\tau}$  yields to  $\hat{\tau} \leq \tau_0$  and so,  $\hat{\tau} = \tau_0$ . It remains to show that  $u_0(x) = J(\hat{\tau}(x), x)$ . Namely, consider  $\varepsilon_0 > 0$  fixed. Given  $0 < \varepsilon \leq \varepsilon_0$  and  $t \leq \tau_{\varepsilon_0}$  we know that  $t \leq \tau_\varepsilon$  and  $u_\varepsilon(y(t, x)) < \varphi(y(t, x))$ . Then the relation (68) leads to

$$u_\varepsilon(x) = \mathbb{E}\left[\int_0^{\tau_{\varepsilon_0}} e^{-\alpha s} f(y(s, x)) ds + e^{-\alpha \tau_{\varepsilon_0}} u_\varepsilon(y(\tau_{\varepsilon_0}, x))\right].$$

By monotone convergence and quasi-left continuity of  $y(s, x)$ , letting  $\varepsilon \downarrow 0$  and hence  $\varepsilon_0 \downarrow 0$ , we obtain

$$u_0(x) = \mathbb{E}\left[\int_0^{\tau_0} e^{-\alpha s} f(y(s, x)) ds + e^{-\alpha \tau_0} \varphi(y(\tau_0, x))\right] = J(\tau_0, x).$$

Thus, we conclude that  $\hat{\tau}$  is the optimal stopping time for the problem (48)–(50).

(c) Suppose  $\varphi \in D(\mathcal{A}_\alpha)$  and let  $v_\varepsilon := \frac{1}{\varepsilon} \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f - \mathcal{A}_\alpha \varphi)^+$ . In virtue of (56) and a variation of (61) we obtain

$$u_\varepsilon - \varphi \leq \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} f + \frac{1}{\varepsilon} \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} \varphi - \varphi = \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} f - \mathcal{R}_{\alpha+\frac{1}{\varepsilon}} \mathcal{A}_\alpha \varphi,$$

which in turn gives  $(u_\varepsilon - \varphi)^+ \leq \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f - \mathcal{A}_\alpha \varphi)^+$ . Using this last inequality together with (52), we get

$$f - \mathcal{A}_\alpha u_\varepsilon = \frac{1}{\varepsilon} (u_\varepsilon - \varphi)^+ \leq \frac{1}{\varepsilon} \mathcal{R}_{\alpha+\frac{1}{\varepsilon}}(f - \mathcal{A}_\alpha \varphi)^+ = v_\varepsilon,$$

or equivalently,  $f - v_\varepsilon \leq \mathcal{A}_\alpha u_\varepsilon$ , yielding that

$$\mathcal{R}_\alpha(f - v_\varepsilon) \leq u_\varepsilon, \quad (71)$$

after applying the resolvent operator in both sides of this later expression. Also note that by (45), we know that  $\text{bs} - \lim_{\varepsilon \downarrow 0} v_\varepsilon = (f - \mathcal{A}_\alpha \varphi)^+$ . Using this last

property, we can let  $\varepsilon \downarrow 0$  at (71) to deduce  $\mathcal{R}_\alpha(f - (f - \mathcal{A}_\alpha \varphi)^+) = \mathcal{R}_\alpha(f \wedge \mathcal{A}_\alpha \varphi) \leq u_0$ .

On the other hand, using (52) again we have that  $\mathcal{A}_\alpha u_\varepsilon \leq f$ , or equivalently,  $u_\varepsilon \leq \mathcal{R}_\alpha f$ . Letting  $\varepsilon \downarrow 0$ , we obtain  $u_0 \leq \mathcal{R}_\alpha f$  but also we have that  $u_0 \leq \varphi$  because it is a subsolution of (58). Then  $\hat{u} = u_0 \leq \mathcal{R}_\alpha f \wedge \varphi$ . Hence, we conclude that

$$\mathcal{R}_\alpha(f \wedge \mathcal{A}_\alpha \varphi) \leq u_0 \leq \mathcal{R}_\alpha f \wedge \varphi.$$

□

**Acknowledgement.** This research was partially founded by CONACyT grant no. 87787.

## References

1. Anderson, W.: Continuous-Time Markov Chains. Springer, New York (1991)
2. Applebaum, D.: Lévy Processes and Stochastic Calculus. Cambridge University Press, Cambridge (2009)
3. Bensoussan, A., Lions, J.L.: Applications des Inéquations Variationnelles en Contrôle Stochastique. Dunod, Paris (1978)
4. Bensoussan, A.: Stochastic Control by Functional Analysis Methods. North-Holland Publishing Co., Amsterdam (1982)
5. Bensoussan, A., Lions, J.L.: Applications of Variational Inequalities in Stochastic Control. North-Holland Publishing Co., Amsterdam (1982)
6. Bensoussan, A.: Dynamic Programming and Inventory Control. IOS Press, Amsterdam (2011)
7. Bickel, P.J., El-Karoui, N., Yor, M.: Ecole d'Eté de Probabilités de Saint-Flour XI -1979. Springer, Berlin (1981)
8. Böttcher, B., Schilling, R., Wang, J.: Lévy Matters III. Springer, Cham (2013)
9. Chung, K.L.: Lectures from Markov Processes to Brownian Motion. Springer, New York (1982)
10. Da Prato, G.: An Introduction to Infinite-Dimensional Analysis. Springer, Berlin (2006)
11. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Cambridge University Press, Cambridge (2014)
12. Dellacherie, C., Meyer, P.A.: Probabilités et Potentiel. Chapitres V à VIII. Hermann, Paris (1980)
13. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. Wiley, New Jersey (1986)
14. El-Karoui, N., Kapoudjian, C., Pardoux, E., Peng, S., Quenez, M.C.: Reflected solutions of backward SDEs and related obstacle problems for PDEs. Ann. Probab. **25**, 702–737 (1997)
15. El-Karoui, N., Peng, S., Quenez, M.C.: Backward stochastic differential equation in finance. Math. Financ. **7**, 1–71 (1997)
16. Glowinski, R., Lions, J.L., Trémolières, R.: Numerical Analysis of Variational Inequalities. North-Holland Publishing Co., Amsterdam (1981)
17. Goran, P., Shiryaev, A.: Optimal Stopping and Free-Boundary Problems. Birkhäuser Verlag, Basel (2006)

18. Hamadène, S., Jeanblanc, M.: On the starting and stopping problem: application in reversible investments. *Math. Oper. Res.* **32**, 182–192 (2007)
19. Horiguchi, M.: Stopped Markov decision processes with a stopping time constraint. *Math. Meth. Oper. Res.* **53**, 279–295 (2001)
20. Jasso-Fuentes, H., Menaldi, J.L., Prieto-Rumeau, T.: Discrete-time control with non-constant discount factor. *Math. Meth. Oper. Res.* **92**, 377–399 (2020)
21. Kallenberg, O.: Foundations of Modern Probability. Springer, New York (2002)
22. Menaldi, J.L.: On the optimal stopping problem for degenerate diffusions. *SIAM J. Control. Optim.* **18**, 697–721 (1980)
23. Menaldi, J.L.: On the optimal impulse control problem for degenerate diffusions. *SIAM J. Control. Optim.* **18**, 722–739 (1980)
24. Menaldi, J.L.: Optimal impulse control problems for degenerate diffusions with jumps. *Acta Appl. Math.* **8**, 165–198 (1987)
25. Menaldi, J.L.: Stochastic hybrid optimal control models. In: Stochastic Models (Guanajuato, 2000), II Aportaciones Mat. Investig., pp. 205–250. Soc. Mat. Mexicana, México (2001)
26. Menaldi, J.L., Sritharan, S.S.: Stochastic 2-D Navier-Stokes equation. *Appl. Math. Optim.* **46**, 31–53 (2002)
27. Menaldi, J.L., Sritharan, S.S.: Remarks on impulse control problems for the stochastic Navier-Stokes equations. In: Differential Equations and Control Theory (Athens, OH). Lecture Notes in Pure and Applied Mathematics, vol. 225, pp. 245–255. Dekker, New York (2002)
28. Menaldi, J.L., Sritharan, S.S.: Impulse control of stochastic Navier-Stokes equations. *Nonlinear Anal.* **52**, 357–381 (2003)
29. Oksendal, B., Sulem, A.: Applied Stochastic Control of Jump Diffusions. Springer, Berlin (2005)
30. Rieder, U.: On stopped decision processes with discrete time parameter. *Stoch. Proc. Appl.* **3**, 365–383 (1975)
31. Robin, M.: Contrôle impulsif avec retard pour des processus de Markov. *Ann. Sci. Univ. Clermont* **14**, 115–128 (1976)
32. Robin, M.: Contrôle impulsif des processus de Markov. Thesis INRIA, TE-035, Paris France (1977)
33. Rogers, L.C.G., Williams, D.: Diffusion, Markov Processes and Martingales. Cambridge University Press, Cambridge (2000)
34. Stettner, L.: On some stopping and impulsive control problems with a general discount rate criterion. *Probab. Math. Statist.* **10**, 223–245 (1989)
35. Tudor, C.: Procesos estocásticos. Sociedad Matemática Mexicana, México (2002)



# Control of Continuous-Time Markov Jump Linear Systems with Partial Information

André Marcorin de Oliveira<sup>1</sup> and Oswaldo Luiz do Valle Costa<sup>2</sup>(✉)

<sup>1</sup> Institute of Science and Technology,  
Federal University of São Paulo (UNIFESP), São José dos Campos, SP, Brazil  
[andre.marcorin@unifesp.br](mailto:andre.marcorin@unifesp.br)

<sup>2</sup> Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil  
[oswaldo@lac.usp.br](mailto:oswaldo@lac.usp.br)

**Abstract.** In this paper we study the  $H_2$  state-feedback control of continuous-time Markov jump linear systems considering that the mode of operation cannot be directly measured. Instead we assume that there is a detector that provides the only information concerning the main jump process, so that the jump processes are modelled by a continuous-time exponential hidden Markov model. We present a new convex design condition for calculating a state-feedback controller depending only on the detector which guarantees stability in the mean-square sense of the closed-loop system, as well as a suitable bound on its  $H_2$  norm. We present an illustrative example in the context of systems subject to faults and compare our results with the current literature.

**Keywords:**  $H_2$  control · Hybrid systems · Continuous-time Markov chain · Linear matrix inequalities

**AMS(2020) Subject Classification:** Primary 93E03 · Secondary 90C25

## 1 Introduction

Lately a great deal of attention has been given to systems subject to sudden changes in their dynamic behavior. This is due in part to the fact that real world systems are subject to various alterations which can be caused internally or externally as, for instance, due to environmental conditions, faults in dynamical systems, or changes to new operation points. Bearing that in mind, modern control systems have to be designed with the capability of maintaining an acceptable behavior and meeting some performance requirements even in the presence of abrupt changes in the system dynamics. In order to derive treatable mathematical models for these situations, a class of systems that has been recently intensively studied in the literature is of linear systems in which the changes in their dynamics are modeled by a Markov chain (known as Markov

jump linear systems, MJLS). It has gained a great boost in the early 1990s when, among other applications, it was used to model fault-tolerant control systems (see, for instance, [21, 27]). We refer to [1, 5, 7, 14, 16, 22, 26] and references therein for a general overview on MJLS and [20] for the application of MJLS in active fault-tolerant control.

The literature on control of MJLS for the case in which the current state of the Markov process (mode of operation) is perfectly known is nowadays quite comprehensive but the case in which there is only a partial information on this parameter is more scarce. Recently, there have been proposed some approaches in the specialized literature to deal with the control problem for MJLS with partial observations of the Markov chain, under different names as the *detector-based approach* (see, e.g., [6, 30]); *MJLS with hidden Markov models* (see e.g. [4, 12, 18]); or *asynchronous MJLS* (see [19, 25]). It has a close connection to the so-called *active fault-tolerant control systems* (AFTCS) in the sense that it is assumed that the Markov chain  $\theta$  is a failure process and we would have access only to a type of *failure detector*  $\hat{\theta}$  for designing the controller. In this context, it was studied in [6] the  $H_2$ -control (or quadratic control) problem of discrete-time MJLS employing a *detector-based approach* for  $\hat{\theta}$ . It was shown in [6] that this approach encompasses the cases with perfect information, no information and the cluster observations of the Markov parameter. Analogously, the  $H_\infty$  control problem was studied in [30]. In [12], the mixed  $H_2/H_\infty$  dynamic output feedback control for discrete-time hidden MJLS was studied through a type of iterative separation procedure. The continuous-time counterpart of the  $H_\infty$  control problem was studied in [24], and [29], while the  $H_2$ -control problem was dealt with in [28], and the dynamic output control case for both the  $H_2$  as well as the  $H_\infty$  was analyzed in [13]. In all these cases, it was assumed that the dynamics of the *detector follows a probabilistic Markov type assumption* and an explicit analytical expression for that has been exhibited.

More specifically, the mathematical representation of the model considered in this chapter is given by a continuous-time linear system following the class of differential equations given by

$$\dot{x}(t) = A_{\theta(t)}x(t) + B_{\theta(t)}u(t), \quad x(0) = x_0, \quad \theta(0) = \theta_0. \quad (1)$$

where it is assumed that there is a continuous-time hidden Markov model (see, for instance, [17])  $Z(t) = (\theta(t), \hat{\theta}(t))$  in which the change on the mode of operation (due, for instance, to a component failure), is modeled by the unobserved component  $\theta(t)$ , while the observed component  $\hat{\theta}(t)$  plays the role of the detector, which provides the information on this change on the mode of operation (a failure detection and identification device in the case of failures). In this problem we are interested in controlling (1) under partial information on the mode of operation  $\theta(t)$ , that is, the goal is to find a state-feedback control  $u(t) = K_{\hat{\theta}(t)}x(t)$  such that the closed loop system

$$\dot{x}(t) = (A_{\theta(t)} + B_{\theta(t)}K_{\hat{\theta}(t)})x(t) \quad (2)$$

meets some stability and performance index requirements. It was proposed in [28] a linear matrix inequality (LMI) optimization formulation for the design of a stochastic stabilizing feedback control with guaranteed  $H_2$ -cost. For the perfect information case (that is,  $\hat{\theta}(t) = \theta(t)$ ) it was shown in [28] that these results recast the usual ones for the  $H_2$  control of continuous-time-time MJLS as presented in [5] provided a design parameter is made sufficiently large. It was also shown in [28] that this modeling encompasses the mode independent and cluster observation cases considered in [31] for the discrete-time case.

The goal of this chapter is re-visit the  $H_2$ -control problem studied in [28] and derive a new set of conditions to design a stochastic stabilizing feedback control with guaranteed  $H_2$ -cost. Notice that for the general hidden Markov model  $Z(t) = (\theta(t), \hat{\theta}(t))$  the set of conditions obtained here and in [28] are independent in the sense that it is not possible to say that one implies the other. But, as in [28], we show that for the perfect information case ( $\hat{\theta}(t) = \theta(t)$ ) these conditions also recast the usual ones for the  $H_2$  control of continuous-time-time MJLS as presented in [5] as long as a design parameter is made sufficiently large.

This chapter is organized as follows. In Sect. 2 we introduce some notation and auxiliary results that will be needed throughout this chapter. In Sect. 3 we present the stochastic model, the concept of mean square stability needed in this work, the quadratic performance index to be minimized, and the general optimization problem. The main result of this chapter is presented in Sect. 4. For the general partial observation case, Theorem 1 shows that if a set of LMI inequalities are satisfied then we get a state-feedback control such that the closed loop system is mean square stable and the associated quadratic performance index satisfies an upper bound value. Moreover it will be shown that whenever we assume the perfect information case (that is,  $\hat{\theta}(t) = \theta(t)$ ), we recast the optimal non-conservative solution for the control problem, provided that an input parameter of the LMI inequalities is made sufficiently large. Section 5 presents an illustrative numerical example and Sect. 6 concludes the chapter with some final comments.

## 2 Notation and Preliminaries

The real Euclidean space of dimension  $n$  is represented by  $\mathbb{R}^n$ , and the space of real matrices of dimension  $m \times n$ , by  $\mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$ , with  $\mathbb{B}(\mathbb{R}^n) \triangleq \mathbb{B}(\mathbb{R}^n, \mathbb{R}^n)$ . The identity matrix of size  $n \times n$  is given by  $I_n$  (or simply  $I$ ),  $(\cdots)'$  is the transpose operator and, for a square matrix  $G$ , we set  $Her(G) \triangleq G + G'$ , and  $Tr(\cdot)$  is the trace operator. Given positive integers  $N$  and  $M$ , we set  $\mathcal{N} \triangleq \{1, \dots, N\}$ ,  $\mathcal{M} \triangleq \{1, \dots, M\}$ , and  $\mathcal{V} \subseteq \mathcal{N} \times \mathcal{M}$ . The linear space composed by all sequence of matrices  $\mathbf{V} = (V_{ik} \in \mathbb{B}(\mathbb{R}^n, \mathbb{R}^m); (ik) \in \mathcal{V})$  is represented by  $\mathbb{H}^{n,m}$ , and for ease of notation we set  $\mathbb{H}^n \triangleq \mathbb{H}^{n,n}$  and  $\mathbb{H}^{n+} \triangleq \{\mathbf{V} \in \mathbb{H}^n : V_{ik} \geq 0, (ik) \in \mathcal{V}\}$ . Similarly we define the set  $\mathbb{M}^{n,m} \triangleq \{M_k \in \mathbb{B}(\mathbb{R}^n, \mathbb{R}^m), k \in \mathcal{M}\}$ ,  $\mathbb{M}^n \triangleq \mathbb{M}^{n,n}$ , and  $\mathbb{M}^{n+}$  accordingly. For  $V \in \mathbb{H}^{n+}$ , by  $V > 0$  we mean that  $V_{ik} > 0$  for all  $(ik) \in \mathcal{V}$  (similarly for  $P > 0$  in  $\mathbb{M}^{n+}$ ). We denote by  $o(h)$  any function  $f$  such that  $\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0$ .  $(\Omega, \mathcal{F}, Prob)$  is a probability space equipped with

a measurable right-continuous filtration  $\mathcal{F}_t$ . The expectation in this space is represented by  $E(\cdot)$ , and the conditional expectation, by  $E(\cdot | \cdot)$ .

We recall the following results that will be useful along this chapter. For given symmetric matrices  $F_i, i = 0, \dots, m$ , a strict linear matrix inequality (LMI) has the form

$$F(x) = F_0 + \sum_{i=1}^m x_i F_i > 0$$

where  $x = [x_1 \dots x_m]' \in \mathbb{R}^m$ ,  $x_i \in \mathbb{R}$ ,  $i = 1, \dots, m$  are the variables. A Semidefinite optimization programming (SDP optimization problem), also referred to as an LMI optimization problem, consists of finding a feasible  $x$  (that is, find  $x$  such that  $F(x) > 0$ ) which minimizes a linear function  $c'x$ . LMI optimization problems are tractable both from theoretical and numerical point of view (e.g. [2]). A key result for converting nonlinear convex inequalities into LMI formulation is the Schur complement, presented next.

**Proposition 1.** (*Schur's complement*) *The following affirmatives are equivalent:*

- a)  $Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q'_{12} & Q_{22} \end{pmatrix} > 0$ .
- b)  $Q_{22} > 0$  and  $Q_{11} - Q_{12}Q_{22}^{-1}Q'_{12} > 0$ .
- c)  $Q_{11} > 0$  and  $Q_{22} - Q'_{12}Q_{11}^{-1}Q_{12} > 0$ .

Notice that a) in Proposition 1 is in the form of a LMI, and b), c) is in the form of nonlinear convex inequalities. SDP optimization problems include several important standard classes of convex optimization problems, such as linear programming, quadratic programming, quadratically constrained quadratic program, and second-order cone programming problems.

Some important results that will be used in this chapter are as follows.

**Proposition 2** ([8,9]). *For  $G \in \mathbb{B}(\mathbb{R}^n)$  and  $P \in \mathbb{B}(\mathbb{R}^n)$  such that  $P > 0$ , we get that*

$$G'P^{-1}G \geq \text{Her}(G) - P. \quad (3)$$

**Proposition 3 (Projection Lemma [2]).** *Given  $P$ ,  $U$ , and  $V$ , there exists  $G$  such that*

$$P + \text{Her}(UGV') > 0$$

*if and only if*

$$\tilde{U}'P\tilde{U} > 0, \tilde{V}'P\tilde{V} > 0$$

*where  $\tilde{U}$  and  $\tilde{V}$  are, respectively, orthogonal complements of  $U$  and  $V$ .*

### 3 Problem Formulation

In a probability space  $(\Omega, \mathcal{F}, Prob)$  we consider a continuous-time hidden Markov model (CT-HMM)  $Z(t) = (\theta(t), \hat{\theta}(t))$ ,  $t \in \mathbb{R}^+$ , formed by two components, the hidden state  $\theta(t)$  taking values in the set  $\mathcal{N}$ , and the observation state  $\hat{\theta}(t)$  taking values in the set  $\mathcal{M}$ . It is assumed that  $Z(t)$  is a homogeneous Markov process taking values in  $\mathcal{N} \times \mathcal{M}$  and having transition rates  $\nu_{(ik)(j\ell)}$ , with  $\nu_{(ik)(j\ell)} \geq 0$  for  $(j\ell) \neq (ik)$  and  $-\nu_{(ik)(ik)} = \sum_{(j\ell) \neq (ik)} \nu_{(ik)(j\ell)}$ . We assume that the transition rates  $\nu_{(ik)(j\ell)}$  of  $Z(t) = (\theta(t), \hat{\theta}(t))$ , are given by

$$Prob(Z(t+h) = (j\ell) \mid Z(t) = (ik)) = \begin{cases} \frac{\nu_{(ik)(j\ell)}h + o(h)}{1 + \nu_{(ik)(ik)}h + o(h)}, & (j\ell) \neq (ik) \\ 1 + \nu_{(ik)(ik)}h + o(h), & (j\ell) = (ik) \end{cases}$$

where

$$\nu_{(ik)(j\ell)} = \begin{cases} \alpha_{j\ell}^k \lambda_{ij}, & i \neq j, \ell \in \mathcal{M}, \\ q_{k\ell}^i, & j = i, \ell \neq k, i \in \mathcal{N}, \\ \lambda_{ii} + q_{kk}^i, & j = i, \ell = k, \\ 0, & \text{otherwise} \end{cases}$$

and  $\sum_{\ell \in \mathcal{M}} \alpha_{j\ell}^k = 1$ ,  $\lambda_{ij} \geq 0$  for all  $i \neq j$ ,  $q_{k\ell}^i \geq 0$ ,  $\ell \neq k$ ,  $\lambda_{ii} = -\sum_{j \in \mathcal{N}^i} \lambda_{ij}$ ,  $q_{kk}^i = -\sum_{\ell \in \mathcal{M}^i} q_{k\ell}^i$ .

We represent by  $\mathcal{V} \subset \mathcal{N} \times \mathcal{M}$  an invariant set of  $Z(t)$ , that is,  $Prob(Z(t) \in \mathcal{V}) = 1$  provided that  $Z(0) \in \mathcal{V}$ .

*Remark 1.* Recalling that  $\lambda_{ij}$  represents the transition rate of  $\theta(t)$ , we get that  $\alpha_{j\ell}^k$  and  $q_{k\ell}^i$  models simultaneous and spontaneous jumps of  $\hat{\theta}(t)$ , that is, for small  $h > 0$ ,  $Prob(\hat{\theta}(t+h) = \ell \mid \theta(t+h) = j, Z(t) = (ik)) = \alpha_{j\ell}^k + r(h)$  for some function such that  $\lim_{h \rightarrow 0} r(h) = 0$ , and  $Prob(\hat{\theta}(t+h) = \ell \mid \theta(t+h) = i, Z(t) = (ik)) = q_{k\ell}^i h + o(h)$ , see [13, 28], and the references therein for more information.

As mentioned in Sect. 1 we consider the continuous-time MJLS (1) where  $A_i \in \mathbb{B}(\mathbb{R}^n)$ ,  $B_i \in \mathbb{B}(\mathbb{R}^m, \mathbb{R}^n)$ ,  $i \in \mathcal{N}$ , and with the state vector denoted by  $x(t) \in \mathbb{R}^n$  and the control input by  $u(t) \in \mathbb{R}^m$ . We aim at designing the following state-feedback controller

$$u(t) = K_{\hat{\theta}(t)} x(t) \tag{4}$$

that depends only on the observed variable  $\hat{\theta}(t)$  taking values in  $\mathcal{M}$ , such that the closed loop system (2) is mean square stable (see Definition 1 below) and has a guaranteed  $H_2$  cost (see Definition 2 below). In what follows we set  $\mathcal{K} = (K_1, \dots, K_M)$  and for  $i \in \mathcal{N}$ ,  $\ell \in \mathcal{M}$ ,

$$A_{i\ell} = A_i + B_i K_\ell. \tag{5}$$

*Remark 2.* The following cases found in the literature can be recasted from the approach presented above (see, for instance, [29]):

- *Mode-dependent case:*  $\mathcal{M} = \mathcal{N}$ ,  $q_{k\ell}^i = 0$ ,  $\alpha_{jj}^k = 1$ , and  $\alpha_{j\ell}^k = 0$  for  $j \neq \ell$ , with invariant set  $\mathcal{V} = \{(ii) \in \mathcal{N} \times \mathcal{N}\}$ . In this case, we get that  $\theta(t) = \hat{\theta}(t)$  almost surely (as).
- *Mode-independent case:*  $\mathcal{M} = \{1\}$ ,  $q_{k\ell}^i = 0$ , and  $\alpha_{j1}^1 = 1$ . Then,  $\hat{\theta}(t) = 1$  and  $\theta(t_1)$  and  $\hat{\theta}(t_2)$  jump with  $t_1 = t_2$  as.
- *No Mutual Jumps:*  $\alpha_{jk}^k = 1$  and  $\alpha_{j\ell}^k = 0$  for  $k \neq \ell$ .
- *The Cluster Case:* We partition the Markov chain states as the union of  $M \leq N$  disjoint sets (clusters)  $\mathcal{N}_i$  so that  $\mathcal{N} = \cup_{i \in \mathcal{M}} \mathcal{N}_i$ . Considering the function  $g : \mathcal{N} \rightarrow \mathcal{M}$  such that  $g(i) = j$  that represents the cluster where the Markov state belongs to, the controller would have access to  $g(i)$ . Equivalently, by taking  $q_{k\ell}^i = 0$  and  $\alpha_{ig(i)}^k = 1$ , so that whenever  $\theta(t)$  jumps to  $i$ ,  $\hat{\theta}(t)$  will jump simultaneously to  $g(i)$ .

We introduce next the concept of mean-square stability and the  $H_2$  norm.

**Definition 1 (Mean-square stability MSS, adapted from [5]).** System (2) is said to be MSS if  $\lim_{t \rightarrow \infty} E(\|x(t)\|^2) = 0$  for arbitrary  $x_0$  and  $Z(0)$ .

We now introduce conditions for verifying the MSS of (2). For that we define the linear operator  $\mathcal{T}$  from  $\mathbb{H}^n$  to  $\mathbb{H}^n$  such that

$$\mathcal{T}_{ik}(\mathbf{P}) \triangleq \text{Her}(A'_{ik} P_{ik}) + \sum_{(j\ell) \in \mathcal{V}} \nu_{(ik)(j\ell)} P_{j\ell} \quad (6)$$

for  $\mathbf{P} \in \mathbb{H}^n$ . We have the following lemma derived in Theorem 1 of [28].

**Lemma 1.** The system  $\dot{x}(t) = A_{\theta(t)\hat{\theta}(t)} x(t)$ ,  $x(0) = x_0 \in \mathbb{R}^n$ , is MSS if and only if there exists  $\mathbf{P} \in \mathbb{H}^{n+}$  such that

$$\mathbf{P} > 0, \mathcal{T}(\mathbf{P}) < 0. \quad (7)$$

The set of admissible controllers (4) is given by

$$\mathfrak{K} \triangleq \{\mathcal{K} = (K_1, \dots, K_M) : \text{such that (7) holds for } A_{i\ell} \text{ as in (5)}\}$$

We define next the concept of the  $H_2$  norm. In order to do that we consider the following MJLS in the probability space  $(\Omega, \mathcal{F}, \text{Prob})$

$$\mathcal{G} : \begin{cases} \dot{x}(t) = A_{\theta(t)} x(t) + B_{\theta(t)} u(t) + J_{\theta(t)} w(t) \\ z(t) = C_{\theta(t)} x(t) + D_{\theta(t)} u(t) \end{cases} \quad (8)$$

where, as before,  $x(t) \in \mathbb{R}^n$ ,  $u(t) \in \mathbb{R}^m$ , and  $z(t) \in \mathbb{R}^q$ ,  $w(t) \in \mathbb{R}^r$ . We also consider that  $x(0) = 0$  and that the initial probability distribution for  $\theta_0$  satisfies

$\text{Prob}(\theta_0 = i) = \mu_i > 0$ . By plugging (8) and (4), we get the closed-loop system yielding to

$$\mathcal{G}_K : \begin{cases} \dot{x}(t) = A_{\theta(t)\hat{\theta}(t)}x(t) + J_{\theta(t)}w(t) \\ z(t) = C_{\theta(t)\hat{\theta}(t)}x(t) \end{cases} \quad (9)$$

where  $A_{i\ell}$  is as in (5) and

$$C_{i\ell} = C_i + D_i K_\ell. \quad (10)$$

We introduce the definition of the  $H_2$  norm next. Notice that the  $H_2$  norm is associated to the minimization over  $K \in \mathfrak{K}$  of the infinite horizon quadratic cost  $\mathcal{J}_K(x_0, Z_0)$  defined by

$$\mathcal{J}_K(x_0, Z_0) \triangleq \int_0^\infty E(\|z(t)\|^2) dt, \quad (11)$$

where  $x(0) = x_0$  and  $Z_0 = (\theta_0, \hat{\theta}_0)$ , see [5] for further details.

**Definition 2 ( $H_2$  norm).** Considering that (9) is MSS and  $x(0) = 0$ , the  $H_2$  norm of (9) is given by  $\|\mathcal{G}_K\|_2^2 \triangleq \sum_{s=1}^r \|z_s\|_2^2$ , where  $z_s(k)$  is the controlled output of (9) for  $w(t) = v_s \delta(t)$ ,  $v_s$  is the  $s$ -th element of the standard basis of  $\mathbb{R}^r$ .

For calculating the  $H_2$  norm of (9) for a given stabilizing controller  $K$ , we can resort to the following (convex) optimization problem

$$\|\mathcal{G}_K\|_2^2 = \inf_{Q_{ik} > 0, \gamma} \gamma^2, \quad (12)$$

$$\sum_{(ik) \in \mathcal{V}} \mu_{ik} \text{Tr}(J'_{ik} Q_{ik} J_{ik}) < \gamma^2 \quad (13)$$

$$\text{Her}(Q_{ik} A_{ik}) + \sum_{(j\ell) \in \mathcal{V}} \nu_{(ik)(j\ell)} Q_{j\ell} + C'_{ik} C_{ik} < 0, \quad (14)$$

where  $\mathbf{Q} \in \mathbb{H}^{n+}$  and we recall that  $\text{Prob}(Z(0) = (ik)) = \mu_{ik} > 0$ ,  $(ik) \in \mathcal{V}$ .

We are now able to formally state the main goal of this work, that is, for a given  $\gamma > 0$ ,

$$\text{find } K \in \mathfrak{K} \text{ such that } \|\mathcal{G}_K\|_2 < \gamma. \quad (15)$$

For the perfect observation case, described in Remark 2 as the mode-dependent case (that is, we have perfect information of  $\theta(t)$  which corresponds to the situation  $\hat{\theta}(t) = \theta(t)$ ) we can obtain the optimal  $H_2$  controller by two methods, the classical Riccati equation approach and the LMI approach. Both methods are described in [5] as well as a connection between them, so that, the solution for this case is not conservative in the sense that the optimal controller can be numerically derived. On the other hand, for the more general case in which we

could have a mismatch between  $\hat{\theta}(t)$  and  $\theta(t)$ , optimality is lost at the expense of a tractable (convex) formulation to the control problem so that only a bound  $\gamma$  on the  $\mathcal{H}_2$  norm of (9), which can be minimized, is guaranteed.

A question that naturally arises is that if the numerical procedure that we derive for the general case recast the optimal solution whenever we assume the perfect information case. In [28] we derived a numerical procedure based on a LMI optimization problem that achieved this property. In the next section we present a different LMI optimization that also has this property, that is, the results in Sect. 4 yield to the optimal  $H_2$  control whenever the assumptions for the mode-dependent case described in Remark 2 are fulfilled. In this case, we show that the conditions presented in Sect. 4 are equivalent to (13)–(14) considering  $K$  as a decision variable in the (non-convex) optimization problem

$$\|\mathcal{G}_*\|_2^2 = \inf_{\kappa \in \mathfrak{K}, Q_{ik} > 0, \gamma} \{\gamma^2; \text{ such that the conditions in (13)–(14) hold}\}.$$

*Remark 3.* For  $\mathcal{K} \in \mathfrak{K}$ , we get that

$$\mathcal{J}_{\mathcal{K}}(x_0, Z_0) \leq E(x'_0 Q_{\theta_0 \hat{\theta}_0} x_0), \quad (16)$$

where  $\mathcal{J}_{\mathcal{K}}(x_0, \theta_0)$  is the quadratic cost defined in (11) and  $Q_{ik} > 0$  is any solution of (14). Considering that  $x_0 \in \mathbb{R}^n$  is a given known initial condition and recalling that  $\text{Prob}(Z(0) = (ik)) = \mu_{ik}$ , we get that

$$\mathcal{J}_{\mathcal{K}}(x_0, Z_0) \leq E(x'_0 Q_{\theta_0 \hat{\theta}_0} x_0) = x'_0 E(Q_{\theta_0 \hat{\theta}_0}) x_0 = x'_0 \sum_{(ik) \in \mathcal{V}} \mu_{ik} Q_{ik} x_0, \quad (17)$$

which is precisely the left-hand side of (13) for  $J_{ik} = x_0$ ,  $(ik) \in \mathcal{V}$ . In this case, it readily follows that  $\mathcal{J}_{\mathcal{K}}(x_0, Z_0) = \|\mathcal{G}_{\mathcal{K}}\|_2^2$ .

## 4 Main Result

In this section we present the main results of this chapter which consist of obtaining, through an LMI optimization problem, a state-feedback control  $u(t) = K_{\hat{\theta}(t)} x(t)$  such that the closed loop system (2) is MSS and the associated  $H_2$  norm satisfies an upper bound value. Moreover it will be shown that whenever we assume the perfect information case (that is,  $\hat{\theta}(t) = \theta(t)$ ), we recast the optimal solution for the  $H_2$  control problem, provided that a design parameter is made sufficiently large. These results will be presented in Theorem 1, while the LMI for the optimization problem will be defined next. Consider the following inequalities for  $(ik) \in \mathcal{V}$ ,

$$\sum_{(ik) \in \mathcal{V}} \mu_{ik} \text{Tr}(W_{ik}) < \varsigma \quad (18)$$

$$\begin{bmatrix} W_{ik} & \bullet \\ J_i & X_{ik} \end{bmatrix} > 0, \quad (19)$$

$$\mathcal{H}_{ik} + \text{Her}(\Psi_{ik} \Phi_{ik}) < 0 \quad (20)$$

$$\begin{bmatrix} Z_{(ik)(jl)} & \bullet \\ H_{ik} & X_{jl} \end{bmatrix} > 0 \quad (21)$$

$$X_{ik} > 0, \quad (22)$$

where

$$\mathcal{H}_{ik} \triangleq \begin{bmatrix} \nu_{(ik)(ik)} X_{ik} & \bullet & & \bullet & \bullet \\ X_{ik} & 0_{n \times n} & & \bullet & \bullet \\ X_{ik} & 0_{n \times n} - \text{Her}(H_{ik}) + \sum_{(j\ell) \in \mathcal{V}^{(ik)}} \nu_{(ik),(j\ell)} Z_{(ik),(j\ell)} & 0_{q \times n} & \bullet & \bullet \\ 0_{q \times n} & 0_{q \times n} & 0_{q \times n} & -I_q & \bullet \end{bmatrix},$$

$$\Psi'_{ik} \triangleq [I_n \zeta_{ik} I_n 0_{n \times n} 0_{n \times q}],$$

$$\Phi_{ik} \triangleq [(A_i G_k + B_i Y_k)' - G'_k 0_{n \times n} (C_i G_k + D_i Y_k)'],$$

with  $\mathcal{V}^{(ik)} = \{(j\ell) \in \mathcal{V} : \nu_{(ik),(j\ell)} > 0\}$ .

In the first part of the next theorem we have a design LMI procedure based on (18)–(22) for obtaining  $\mathcal{K}$  satisfying (15), while in the second part we show that for the perfect information case the existence of a solution for the LMI inequalities (18)–(22) is also necessary for (15) provided that the parameters  $\zeta_{ik}$  are made sufficiently large.

**Theorem 1.** *We have the following statements:*

1. *There exist  $\varsigma > 0$ ,  $\zeta_{ik} > \nu_{(ik),(ik)}/2$ ,  $Z_{(ik),(j\ell)}$ ,  $H_{ik}$ ,  $W_{ik}$ ,  $X_{ik}$ ,  $G_k$ ,  $Y_k$ ,  $(ik) \in \mathcal{V}$  such that (18)–(22) hold.*
2. *There exists  $\mathcal{K} \in \mathfrak{K}$  such that  $\|\mathcal{G}_{\mathcal{K}}\|_2 < \varsigma^{1/2}$ .*

We get that 1.  $\implies$  2. with  $\mathcal{K} = (K_1, \dots, K_M)$ ,  $K_k = Y_k G_k^{-1}$ ,  $k \in \mathcal{M}$ . Besides, if the complete observation assumption of Remark 2 is fulfilled, by taking  $\zeta_{ik}$  sufficiently large, we get that 2.  $\implies$  1.

*Proof.* 1.  $\implies$  2.: Given that (18)–(22) holds, we get, by setting  $\gamma^2 = \varsigma$  and  $Y_k = K_k G_k$ , that

$$\sum_{(ik) \in \mathcal{V}} \mu_{ik} \text{Tr}(W_{ik}) < \gamma^2, \quad (23)$$

$$\mathcal{H}_{ik} + \text{Her}(\Psi_{ik} G'_k \bar{\Phi}_{ik}) < 0 \quad (24)$$

holds, where

$$\bar{\Phi}_{ik} \triangleq [A'_{ik} - I_n 0_{n \times n} C'_{ik}].$$

By defining

$$\mathcal{N}_{ik} = \begin{bmatrix} I_n & 0 & 0 \\ A'_{ik} & 0 & C'_{ik} \\ 0 & I_n & 0 \\ 0 & 0 & I_q \end{bmatrix}$$

so that  $\text{Rank}(N_{ik}) = 2n + q$ , we get that  $\mathcal{N}_{ik}$  is the orthogonal complement of  $\bar{\mathcal{P}}'_{ik}$ . From Proposition 3 (or by directly multiplying (24) to the left-hand side by  $\mathcal{N}'_{ik}$  and to the right-hand side by  $\mathcal{N}_{ik}$ ), we obtain that

$$\mathcal{C}_{ik} \triangleq \mathcal{N}'_{ik} \mathcal{H}_{ik} \mathcal{N}_{ik} = \begin{bmatrix} \nu_{(ik)(ik)} X_{ik} + \text{Her}(A_{ik} X_{ik}) & X_{ik} & X_{ik} C'_{ik} \\ X_{ik} & M_{ik} & 0_{n \times q} \\ C_{ik} X_{ik} & 0_{q \times n} & -I_q \end{bmatrix} < 0 \quad (25)$$

holds, where  $M_{ik} \triangleq -\text{Her}(H_{ik}) + \sum_{(j\ell) \in \mathcal{V}^{(ik)}} \nu_{(ik),(j\ell)} Z_{(ik),(j\ell)}$ . Considering a similar reasoning as employed in [3], by the Schur complement (see Proposition 1), we get that (21) yields  $Z_{(ik)(j\ell)} > H'_{ik} X_{j\ell}^{-1} H_{ik}$  so that

$$\text{Her}(H_{ik}) - H'_{ik} \left( \sum_{(j\ell) \in \mathcal{V}^{(ik)}} \nu_{(ik),(j\ell)} X_{j\ell}^{-1} \right) H_{ik} \geq -M_{ik}.$$

By setting

$$G = H_{ik}, P = \left( \sum_{(j\ell) \in \mathcal{V}^{(ik)}} \nu_{(ik),(j\ell)} X_{j\ell}^{-1} \right)^{-1} \quad (26)$$

in Proposition 2, we get that  $P \geq \text{Her}(G) - G' P^{-1} G \geq -M_{ik}$ . Thus,

$$\begin{bmatrix} \nu_{(ik)(ik)} X_{ik} + \text{Her}(A_{ik} X_{ik}) & X_{ik} & X_{ik} C'_{ik} \\ X_{ik} & -\left( \sum_{(j\ell) \in \mathcal{V}^{(ik)}} \nu_{(ik),(j\ell)} X_{j\ell}^{-1} \right)^{-1} & 0_{n \times q} \\ C_{ik} X_{ik} & 0_{q \times n} & -I_q \end{bmatrix} < 0.$$

By applying the congruence transformation  $\text{diag}(X_{ik}^{-1}, I_n, I_q)$ , permuting some rows and columns, and using the Schur complement (Proposition 1) in the last inequality, we get that (14) holds for  $Q_{ik} = X_{ik}^{-1}$ . Similarly, by applying the Schur complement (see Proposition 1) to (19), we get that  $W_{ik} > J'_i X_{ik}^{-1} J_i$ , then by multiplying this equation by  $\mu_{ik}$ , summing everything up for all  $(ik) \in \mathcal{V}$  and considering (23), we get (13), and the claim follows.

2.  $\implies$  1.: If the complete observation hypothesis of Remark 2 is fulfilled (that is,  $\hat{\theta}(t) = \theta(t)$ ), we get that  $\mathcal{V} = \{(ii) : i = 1, \dots, N\}$ ,  $\nu_{(ii)(jj)} = \lambda_{ij}$  and

$$\sum_{(ii) \in \mathcal{V}} \mu_{ii} \text{Tr}(J'_{ii} Q_{ii} J_{ii}) < \gamma_2^2 \quad (27)$$

$$\text{Her}(Q_{ii} A_{ii}) + \sum_{(jj) \in \mathcal{V}} \nu_{(ii)(jj)} Q_{jj} + C'_{ii} C_{ii} < 0, \quad (28)$$

$$Q_{ii} > 0 \quad (29)$$

holds for some mode-dependent MS-stabilizing controller  $\mathcal{K} = (K_1, \dots, K_N)$ . Considering a similar reasoning as presented in [3] and the references therein, we define  $X_{ii} \triangleq Q_{ii}^{-1}$  along with

$$\begin{aligned} Z_{(ii)(jj)} &\triangleq \left( \sum_{(jj) \in \mathcal{V}} \nu_{(ii),(jj)} X_{jj}^{-1} \right)^{-1} X_{jj}^{-1} \left( \sum_{(jj) \in \mathcal{V}} \nu_{(ii),(jj)} X_{jj}^{-1} \right)^{-1} + I\epsilon \\ &> \left( \sum_{(jj) \in \mathcal{V}} \nu_{(ii),(jj)} X_{jj}^{-1} \right)^{-1} X_{jj}^{-1} \left( \sum_{(jj) \in \mathcal{V}} \nu_{(ii),(jj)} X_{jj}^{-1} \right)^{-1} \end{aligned} \quad (30)$$

for some small  $\epsilon > 0$ . Then, after applying the Schur complement (see Proposition 1) to (30) and setting  $H_{ii} = \left( \sum_{(jj) \in \mathcal{V}^{(ii)}} \nu_{(ii),(jj)} X_{jj}^{-1} \right)^{-1}$ , we get that (21) holds. By directly applying the Schur complement, Proposition 1, to (28), we get that

$$\begin{bmatrix} Q_{ii}\nu_{(ii)(jj)} + \text{Her}(Q_{ii}A_{ii}) & \bullet & \bullet \\ I & -\left(\sum_{j \in \mathcal{V}^{(ii)}} \nu_{(ii)(jj)} Q_{jj}^{-1}\right)^{-1} & \bullet \\ C_{ii} & 0 & -I \end{bmatrix} < 0 \quad (31)$$

Multiplying (31) by  $\text{diag}(X_{ii}, I, I)$ , we then get that

$$\begin{bmatrix} X_{ii}\nu_{(ii)(jj)} + \text{Her}(A_{ii}X_{ii}) & \bullet & \bullet \\ X_{ii} & -\left(\sum_{j \in \mathcal{V}^{(ii)}} \nu_{(ii)(jj)} X_{jj}^{-1}\right)^{-1} & \bullet \\ C_{ii}X_{ii} & 0 & -I \end{bmatrix} < 0 \quad (32)$$

holds. Besides,

$$\begin{aligned} -M_{ii} &= \text{Her}(H_{ii}) - \sum_{(jj) \in \mathcal{V}^{(ii)}} \nu_{(ii)(jj)} H_{ii} X_{jj}^{-1} H_{ii} + \nu_{(ii)} \epsilon I \\ &= H_{ii} + \nu_{(ii)} \epsilon I. \end{aligned} \quad (33)$$

By choosing the small perturbation  $\epsilon > 0$  to (32) such that

$$\begin{bmatrix} X_{ii}\nu_{(ii)(jj)} + \text{Her}(A_{ii}X_{ii}) & \bullet & \bullet \\ X_{ii} & (H_{ii} + \nu_{(ii)} \epsilon I) & \bullet \\ C_{ii}X_{ii} & 0 & -I \end{bmatrix} < 0 \quad (34)$$

still holds, and using (33), we get (25). The remainder of the proof is partially inspired in [23]. We now define

$$\mathcal{M}_i \triangleq \mathcal{C}_{ii} + \mathcal{D}_{ii}$$

where  $\mathcal{C}_{ii}$  is given by (25) and

$$\begin{aligned}\mathcal{D}_{ii} &\triangleq \frac{1}{2\zeta_{ii}} \begin{bmatrix} A_{ii} \\ 0 \\ C_{ii} \end{bmatrix} X_{ii} \begin{bmatrix} A'_{ii} & 0 & C'_{ii} \end{bmatrix} \\ &= \begin{bmatrix} A_{ii} \frac{X_{ii}}{\zeta_{ii}} \\ 0 \\ C'_{ii} \frac{X_{ii}}{\zeta_{ii}} \end{bmatrix} \frac{\zeta_{ii}}{2} X_{ii}^{-1} \begin{bmatrix} \frac{X_{ii}}{\zeta_{ii}} A'_{ii} & 0 & \frac{X_{ii}}{\zeta_{ii}} C'_{ii} \end{bmatrix} \geq 0,\end{aligned}\quad (35)$$

since  $X_{ii} > 0$ . Note that  $\lim_{\zeta_{ii} \rightarrow \infty} \mathcal{M}_i = \mathcal{C}_{ii} < 0$ , so that, by taking suitable  $\zeta_{ii}$  large enough, we get that  $\mathcal{M}_i < 0$ . Define  $G_i \triangleq X_{ii}/\zeta_{ii}$ , so that  $\text{Her}(G_i) = 2X_{ii}/\zeta_{ii}$ . For this suitable choice of  $\zeta_{ii} > 0$ , we get that

$$\mathcal{M}_{ii} = \mathcal{C}_{ii} + \begin{bmatrix} A_{ii} G_i \\ 0 \\ C_{ii} G_i \end{bmatrix} \text{Her}(G_i)^{-1} \begin{bmatrix} G'_i A'_{ii} & 0 & G'_i C'_{ii} \end{bmatrix} < 0.$$

By applying the Schur complement (see Proposition 1) to the last inequality, and recalling that  $X_{ii} = G_i \zeta_{ii}$ , we get that

$$\begin{bmatrix} \nu_{(ii)(ii)} X_{ii} + \zeta_{ii} \text{Her}(A_{ii} G_i) & \bullet & \bullet & \bullet \\ X_{ii} & M_{ii} & \bullet & \bullet \\ C_{ii} G_i \zeta_{ii} & 0_{q \times n} & -I_q & \bullet \\ G'_i A'_{ii} & 0 & G'_i C'_{ii} - \text{Her}(G_i) & \end{bmatrix} < 0 \quad (36)$$

holds. By commuting suitable rows and columns, we get that

$$\begin{bmatrix} \nu_{(ii)(ii)} X_{ii} + \zeta_{ii} \text{Her}(A_{ii} G_i) & \bullet & \bullet & \bullet \\ G'_i A'_{ii} & -\text{Her}(G_i) & \bullet & \bullet \\ X_{ii} & 0 & M_{ii} & \bullet \\ C_{ii} G_i \zeta_{ii} & C_{ii} G_i & 0_{q \times n} & -I_q \end{bmatrix} < 0. \quad (37)$$

Finally, by defining  $Y_i = K_i G_i$  and recalling that  $X_{ii} - G_i \zeta_{ii} = 0$ , we get that (20) holds, and the claim follows.  $\square$

The best upper bound for our main goal (15) can be calculated by solving the following LMI optimization problem

$$\inf_{\xi \in \Xi(\zeta)} \varsigma \quad (38)$$

where  $\xi = (\varsigma, Z_{(ik),(j\ell)}, H_{ik}, W_{ik}, X_{ik}, G_k, Y_k)$  and  $\Xi(\zeta)$  is the set of solutions of (18)–(22) for a given  $\zeta = (\zeta_{ik} : (ik) \in \mathcal{V})$ .

*Remark 4.* Note that if  $\zeta_{ik} \leq \nu_{(ik),(ik)}/2$ , then (19) is unfeasible. Indeed, through the Projection Lemma of Proposition 3, by setting  $U = \Phi_{ik}$  and taking the orthogonal complement  $\tilde{U}$  as

$$\tilde{U} = \begin{bmatrix} -I_n & 0 & 0 \\ I_n \zeta_{ik} & 0 & 0 \\ 0 & I_n & 0 \\ 0 & 0 & I_q \end{bmatrix},$$

we get that if (19) holds, then

$$\begin{bmatrix} \nu_{(ik)(ik)} X_{ik} - 2X_{ik}\zeta_{ik} & \bullet & \bullet \\ -X_{ik} & -\text{Her}(H_{ik}) + \sum_{(j\ell) \in \mathcal{V}^{(ik)}} \nu_{(ik),(j\ell)} Z_{(ik),(j\ell)} & \bullet \\ 0 & 0 & -I_q \end{bmatrix} < 0$$

also holds. Therefore, a necessary condition for the last inequality to hold is that  $X_{ik}(\nu_{(ik)(ik)} - 2\zeta_{ik}) < 0$ . Since  $X_{ik} > 0$ , we must have that  $\zeta_{ik} > \nu_{(ik)(ik)}/2$ , for all  $(ik) \in \mathcal{V}$ .

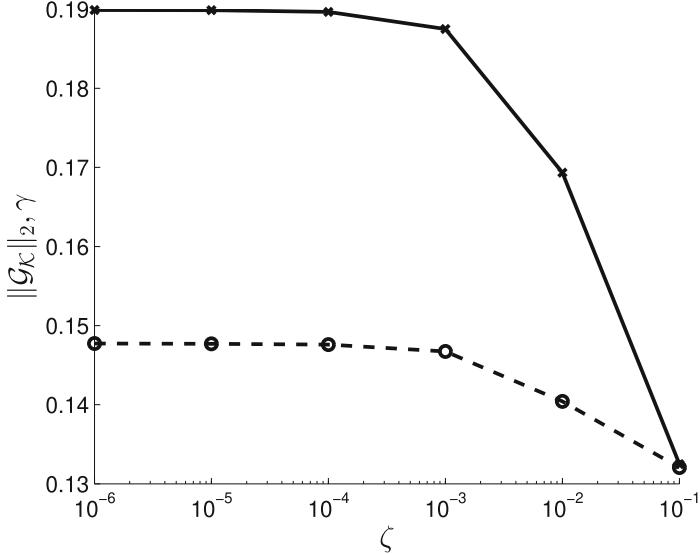
## 5 Illustrative Example

In this example, we consider the linearized model of the unstable lateral dynamics of an unmanned aircraft discussed in [15]. The original, nonlinear, model is obtained by considering a rigid-body motion, assuming that Earth is locally flat, so that centripetal acceleration caused by the its curvature is neglected, and also that Earth is an inertial (Galilean) frame so that the Coriolis acceleration is ignored. Then the nonlinear model follows by using classical (Newtonian) mechanics. The state  $x = [\bar{p} \bar{r} \beta \phi]$  is composed by variations on the roll rate  $\bar{p}$ , the yaw rate  $\bar{r}$ , the sideslip angle  $\beta$ , and the roll angle  $\phi$ . The control input  $u' = [\delta_a \delta_r]$  is given by variations on the aileron  $\delta_a$  and on the rudder  $\delta_r$ . The linearization is performed around the nominal conditions  $\bar{p}_{nom} = \bar{q}_{nom} = \bar{r}_{nom} = 0$ ,  $\theta_{nom} = \alpha_{nom}$ ,  $\beta_{nom} = 0$ , and  $\phi_{nom} = 0$ , where  $\bar{q}_{nom}$  is the nominal pitch rate,  $\theta_{nom}$  is the nominal pitch angle, and  $\alpha_{nom}$  is the nominal angle of attack, considering that the aircraft flies at a straight and level flight, a constant altitude of 500 above sea level, assuming a constant speed of 30 m/s. Therefore, the nominal matrices are given by

$$A_{nom} = \begin{bmatrix} -11.4540 & 2.7185 & -19.4399 & 0 \\ 0.5068 & -2.9875 & 23.3434 & 0 \\ 0.0922 & -0.9957 & -0.4680 & 0.3256 \\ 1 & 0.0926 & 0 & 0 \end{bmatrix}, B_{nom} = \begin{bmatrix} 78.4002 & -2.7282 \\ -3.4690 & 13.9685 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \quad (39)$$

We consider that the aircraft is subject to actuator faults that can be modeled by the Markov chain  $\theta(t)$  whose states represent three possible modes of operation: the nominal one  $= \theta(k) = 1$  so that  $B_1 = B_{nom}$ ; the case in which the actuator power is reduced to 50%  $\theta(t) = 0$ ,  $B_2 = 0.5B_{nom}$ ; and the case in which the actuator power is reduced to 10%  $B_3 = 0.1B_{nom}$ . That is,  $\mathcal{N} = \{1, 2, 3\}$ . Also,  $A_i = A_{nom}$ ,  $\forall i \in \mathcal{N}$ . Similarly to [13], we consider that the fault rates are given by

$$[\lambda_{ij}] = \begin{bmatrix} -0.3 & 0.2 & 0.1 \\ 1.1 & -1.5 & 0.4 \\ 1.0 & 1.0 & -2.0 \end{bmatrix}. \quad (40)$$



**Fig. 1.**  $\gamma$  and  $\|G_K\|_2$  against  $\bar{\zeta}$  for the complete observation case

The main goal is to investigate the  $H_2$  control through the lens of the LQR control as discussed in Remark 3. Then, we set

$$C_i = \begin{bmatrix} I_4 \\ 0_{2 \times 4} \end{bmatrix}, D_i = \begin{bmatrix} 0_{4 \times 2} \\ I_2 \end{bmatrix} \quad (41)$$

for all  $i \in \mathcal{N}$ , that is, we choose the same weights for all states and control inputs. We consider the initial condition

$$x_0 = [0 \ 0 \ 0.087 \ -0.087]^\top \quad (42)$$

so that  $J_i = x_0$ ,  $i \in \mathcal{V}$ , considering the reasoning of Remark 3.

Let us first assume that we have a perfect fault detector so that  $\hat{\theta}(t) = \theta(t)$  for all  $t$  and consider the invariant set

$$\mathcal{V} = \{(1, 1), (2, 2), (3, 3)\} \quad (43)$$

$\mu = [0.7808 \ 0.1502 \ 0.0691]$ . Then, we calculate the optimal  $H_2$  control by solving (38) for  $\bar{\zeta}_{ii} = \zeta \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ ,  $i \in \mathcal{N}$ . Finally, for the control  $K \in \mathcal{K}$  obtained by solving (38), we calculate the  $H_2$  norm of the closed-loop system resorting to (12)–(14). The upper bound  $\gamma$  and  $\|G_K\|_2$  is shown in Fig. 1.

In this example, we note that the conservatism between the upper bound yielded by (38) and the actual  $H_2$  norm is readily decreased by increasing  $\bar{\zeta}$ , as discussed in Theorem 1.

We now study the partial observation case and consider three possible detector outputs so that  $\mathcal{M} = \mathcal{N}$ . We consider that the detector can perfectly detect

the nominal mode of operation, that is,  $\hat{\theta}(t) = 1$  whenever  $\theta(t) = 1$ . However, the detector may have difficulties in distinguishing between  $\theta(t) = 2$  and  $\theta(t) = 3$ . In this case, the invariant set is given by

$$\mathcal{V} = \{(11), (22), (23), (32), (33)\} \quad (44)$$

and the transition rate matrix is given by

$$[\nu_{(ik),(j\ell)}] = \begin{bmatrix} \lambda_{11} & \lambda_{12}\alpha_{22}^1 & \lambda_{12}\alpha_{23}^1 & \lambda_{13}\alpha_{32}^1 & \lambda_{13}\alpha_{33}^1 \\ \lambda_{21} & \lambda_{22} + q_{22}^2 & q_{23}^2 & \lambda_{23}\alpha_{32}^2 & \lambda_{23}\alpha_{33}^2 \\ \lambda_{21} & q_{32}^2 & \lambda_{22} + q_{33}^2 & \lambda_{23}\alpha_{32}^3 & \lambda_{23}\alpha_{33}^3 \\ \lambda_{31} & \lambda_{32}\alpha_{22}^2 & \lambda_{32}\alpha_{23}^2 & \lambda_{33} + q_{22}^3 & q_{23}^3 \\ \lambda_{31} & \lambda_{32}\alpha_{22}^3 & \lambda_{32}\alpha_{23}^3 & q_{32}^3 & \lambda_{33} + q_{33}^3 \end{bmatrix}, \quad (45)$$

where the states sequence in the transition matrix is given by (11), (22), (23), (32), and (33). We note that, by restricting the invariant set to (44), we automatically impose that  $q_{11}^1 = 0$  and  $\alpha_{11}^k = 1$ ,  $k \in \mathcal{M}$ .

We first investigate the case in which only simultaneous jumps occur by varying  $\alpha_{j\ell}^k$ , that is, the probability of the detector going to  $\ell$  given that its current state is  $k$  and the next Markov state is  $j$ . We also consider that

$$\alpha_{22}^k = \bar{\alpha}_2, \alpha_{33}^k = \bar{\alpha}_3, k \in \mathcal{M},$$

for  $0 < \bar{\alpha}_i < 1$ ,  $i \in \{2, 3\}$ , along with the following regions

Region 1:  $\bar{\alpha}_i = 0$ ,  $\mathcal{V} = \{(11), (23), (32)\}$

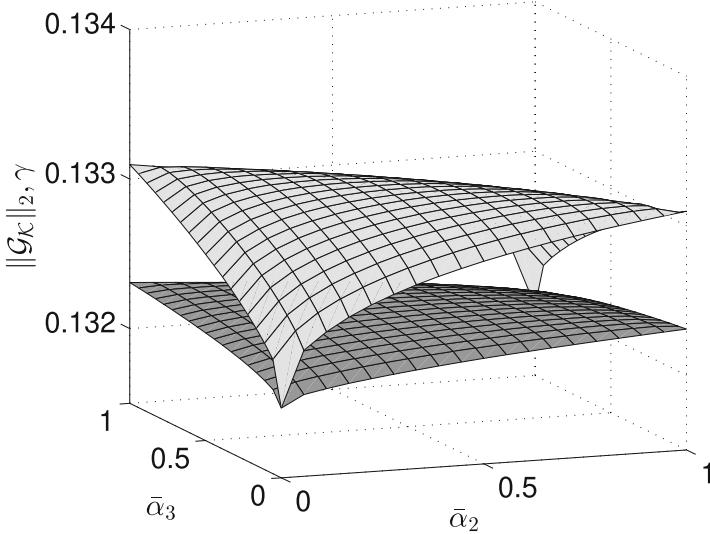
Region 2:  $0 < \bar{\alpha}_i < 1$ ,  $\mathcal{V}$  as in (44)

Region 3:  $\bar{\alpha}_i = 1$ ,  $\mathcal{V}$  as in (43)

for  $i \in \{2, 3\}$ . The spontaneous rates are set to zero, that is,  $q_{k\ell}^i = 0$ . We solve (38) by varying  $\bar{\alpha}_2$  and  $\bar{\alpha}_3$  and calculate the actual  $H_2$  norm for each case. In each iteration, we set the initial distribution of  $Z(t)$  as the stationary distribution and  $\zeta_{ik} = 10$ ,  $(ik) \in \{(11), (22), (23), (32), (33)\}$ . The upper bounds  $\gamma$  and  $\|\mathcal{G}_K\|_2$  are shown in Fig. 2 against  $\bar{\alpha}_3$  and  $\bar{\alpha}_2$ . The result of this simulation traces a parallel to the discrete-time hidden MJLS approach of [12] considering the behavior of  $\gamma$  and  $\|\mathcal{G}_K\|_2$  with respect to variations on  $\alpha_{j\ell}^k$  ( $\alpha_{il}$  for the discrete-time formulation). We note that we get the perfect observation case in Region 3. Interestingly the same configuration is obtained in Region 1: since we know for sure that the detector will jump to  $\hat{\theta}(t+h) = 3$  if  $\theta(t+h) = 2$  (and vice-versa), then we know which mode of operation we have in this situation. Finally, there is a worst-case line for  $\alpha_3 = 1 - \alpha_2$  in which all costs and controllers are numerically close and achieves their maximum value, that is,  $\gamma = 0.1345$  and  $\|\mathcal{G}_K\|_2 = 0.1329$ , with control gains given by

$$K_1 = \begin{bmatrix} -0.8814 & -0.0167 & -0.1129 & -1.0753 \\ -0.0087 & -0.8221 & -0.0084 & 0.0281 \end{bmatrix},$$

$$K_2 \approx K_3 \approx \begin{bmatrix} -0.5170 & -0.0437 & -0.0681 & -0.6962 \\ -0.1603 & -0.3536 & -0.1232 & -0.0985 \end{bmatrix}.$$



**Fig. 2.**  $\gamma$  and  $\|G_K\|_2$  against  $\bar{\alpha}_2$  and  $\bar{\alpha}_3$

By analysing the control gains, we note that there are two clusters (sets),  $\{1\}$  and  $\{2, 3\}$  that naturally arises from solving (38) with the given probabilities. All those cases we previously explained are similar to the ones studied in [10–12], and the references therein, for discrete-time hidden MJLS.

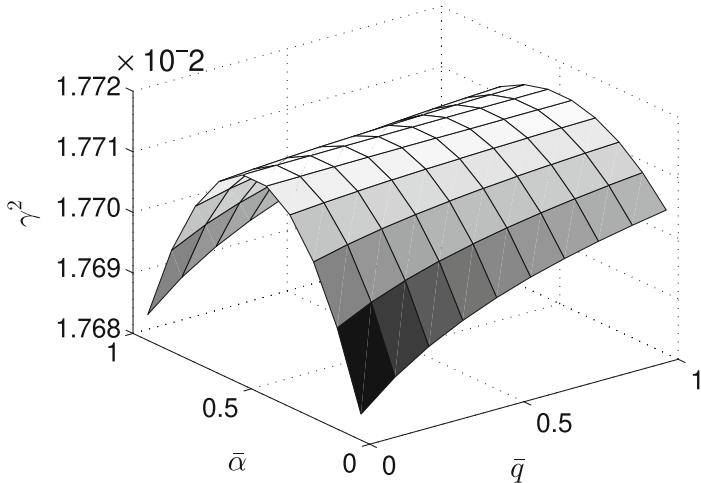
Let us now study the case in which only spontaneous jumps (no mutual jumps, see Remark 2) occur for the modes  $\{2, 3\}$  so that  $\alpha_{jk}^k = 1$  for all  $j \in \{2, 3\}$ ,  $k \in \{2, 3\}$  (recalling that  $\sum_{\ell \in \{2, 3\}} \alpha_{j\ell}^k = 1$ ,  $j \in \{2, 3\}$ ,  $k \in \{2, 3\}$ ). We set

$$q_{22}^2 = q_{33}^2 = q_{22}^3 = q_{33}^3 = -\bar{q}$$

so that

$$[\nu_{(ik),(j\ell)}] = \begin{bmatrix} \lambda_{11} & \lambda_{12}\alpha_{22}^1 & \lambda_{12}\alpha_{23}^1 & \lambda_{13}\alpha_{32}^1 & \lambda_{13}\alpha_{33}^1 \\ \lambda_{21} & \lambda_{22}-\bar{q} & \bar{q} & \lambda_{23} & 0 \\ \lambda_{21} & \bar{q} & \lambda_{22}-\bar{q} & 0 & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & 0 & \lambda_{33}-\bar{q} & \bar{q} \\ \lambda_{31} & 0 & \lambda_{32} & \bar{q} & \lambda_{33}-\bar{q} \end{bmatrix}, \quad (46)$$

and again, the states sequence in the transition matrix is given by (11), (22), (23), (32), and (33). By inspecting (46), we note that the choice of  $\mathcal{V}$  as in (44) imposes that simultaneous jumps will occur if  $\theta(t) = 1$  and  $\theta(t+h) = 2$  (or  $\theta(t+h) = 3$ ). In this case, we set  $\alpha_{22}^1 = \alpha_{33}^1 = \bar{\alpha}$ . By varying  $\bar{q} \in [0.01, 1.00]$  for  $\bar{\alpha} \in [0.05, 0.95]$ , we solve (38) and calculate the  $H_2$  norm of the resulting closed-loop system with (12). The upper bound  $\gamma^2$  against  $\bar{q}$  and  $\bar{\alpha}$  are shown in Fig. 3. By fixing  $\bar{q}$ , we note that the behavior of  $\gamma^2$  is similar to the one displayed in Fig. 2 for  $\bar{\alpha}_3 = \bar{\alpha}_2$ . That is, the smallest upper bounds are obtained if  $\bar{\alpha} \rightarrow 0$



**Fig. 3.**  $\gamma^2$  against  $\bar{q}$  and  $\bar{\alpha}$

or  $\bar{\alpha} \rightarrow 1$ . Conversely, the worst-case scenario is also given by  $\bar{\alpha} = 0.5$ . On the other hand, we note that, by increasing  $\bar{q}$ , we get that  $\gamma^2$  also increases, since  $\bar{q}$  increases the uncertainty of the detector, as discussed in [28].

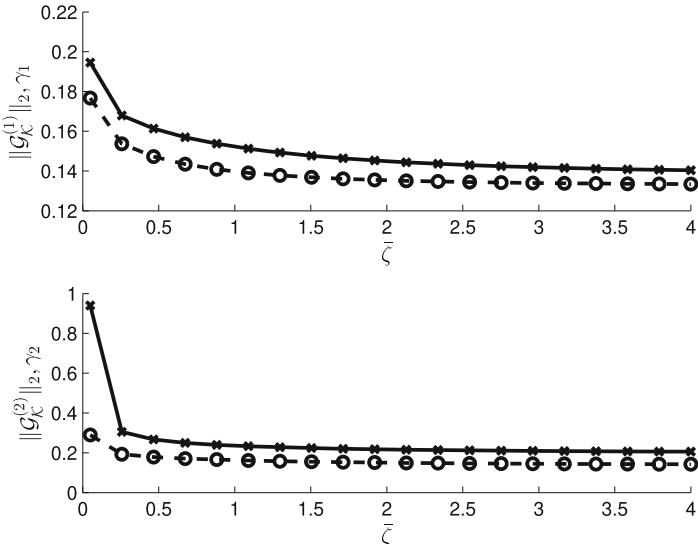
Let us now suppose a more general case in which  $\mathcal{V} = \{(ik), i \in \mathcal{N}, k \in \mathcal{M}\}$ , that is, we consider all possible combinations of  $i$  and  $k$ . We set

$$\alpha_{11}^k = 1, \alpha_{22}^k = \alpha_{33}^k = 0.7, k \in \mathcal{M}, \quad (47)$$

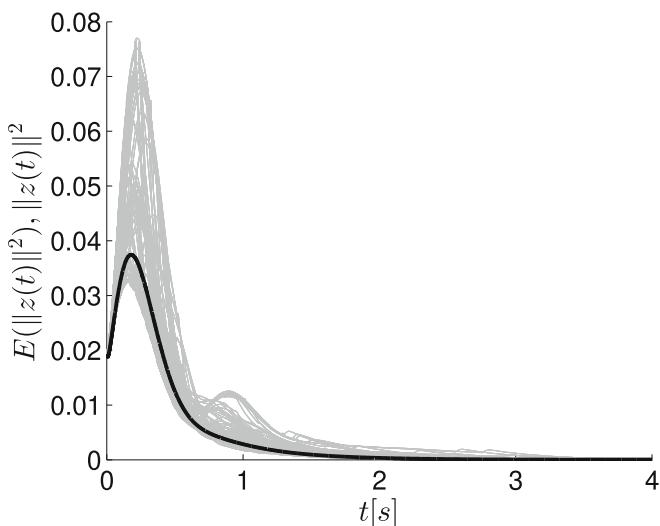
for all  $k \in \mathcal{M}$ , along with

$$[q_{k\ell}^i] = \begin{bmatrix} -1 & 1/3 & 2/3 \\ 1/3 & -1 & 2/3 \\ 1/3 & 2/3 & -1 \end{bmatrix} \quad (48)$$

for all  $i \in \mathcal{N}$ . We now compare our results to the ones given in [28]. By varying the parameter  $\zeta_{ik} = \bar{\zeta}$  of Theorem 1, for all  $(ik) \in \mathcal{V}$ , and  $\zeta_\ell = \bar{\zeta}$  of Theorem 5 of [28], for all  $\ell \in \mathcal{M}$ , for  $\zeta > 0$ , we obtain the upper bounds  $\gamma_1$  and  $\gamma_2$ , as well as  $\|\mathcal{G}_{\mathcal{K}}^{(1)}\|_2$  and  $\|\mathcal{G}_{\mathcal{K}}^{(2)}\|_2$ , shown in Fig. 4. In this example, we note that the upper bounds  $\gamma_1$  obtained through Theorem 1 are smaller compared to the ones,  $\gamma_2$ , yielded by Theorem 5 of [28]. The smallest upper bound obtained through (38) is given by  $\gamma_1^* = 0.1403$ , for an actual  $H_2$  norm of  $\|\mathcal{G}_{\mathcal{K}}^1\|_2 = 0.1334$  whereas we get that  $\gamma_2^* = 0.2059$  and  $\|\mathcal{G}_{\mathcal{K}}^2\|_2 = 0.1424$  obtained through Theorem 5 of [28], both for  $\bar{\zeta} = 4$ . Concerning the conservatism of both results, that is, the distance between the upper bounds and the actual  $H_2$  norm, we note that it tends to decrease as we increase  $\zeta$ , albeit not necessarily monotonically. Besides, we note that  $\gamma_1^*/\|\mathcal{G}_{\mathcal{K}}^1\|_2 = 1.0517$  and  $\gamma_2^*/\|\mathcal{G}_{\mathcal{K}}^2\|_2 = 1.4458$  for  $\bar{\zeta} = 4$ . Thus, for this example, the conservatism yielded by the conditions of Theorem 1 are smaller compared to Theorem 5 of [28].



**Fig. 4.** Top figure:  $\gamma_1$  (full line) and  $\|G_K^{(1)}\|_2, \gamma_1$  (dashed line) calculated through (38) against  $\bar{\zeta}$ ; Bottom figure:  $\gamma_2$  (full line) and  $\|G_K^{(2)}\|_2, \gamma_2$  (dashed line) against  $\bar{\zeta}$  calculated through Theorem 5 of [28].



**Fig. 5.**  $\|z(t)\|^2$  (grey lines) and  $E(\|z(t)\|^2)$  (black line) against  $t$  for a Monte Carlo simulation of 500 rounds.

Finally, we run a Monte Carlo simulation of 500 rounds and take the trajectories  $\|z(t)\|^2$  against time for the detector probabilities given in (47) and (48), respectively. The initial condition is given in (42) and we set  $\bar{\zeta} = \zeta_{ik} = 4$ . The  $\|z(t)\|^2$  curves, along with  $E(\|z(t)\|^2)$  are shown in Fig. 5.

By numerically integrating  $E(\|z(t)\|^2)$ , we get that

$$\mathcal{J}_K(x_0, \theta_0) \approx 0.0179. \quad (49)$$

Considering Remark 3, we get, by simulation, that  $\|\mathcal{G}_K\|_2 = \sqrt{\mathcal{J}_K(x_0, \theta_0)} \approx 0.134$ , whereas the actual  $H_2$  norm value is given by  $\|\mathcal{G}_K\|_2 = 0.133$ .

## 6 Conclusion

In this chapter, we revisited the  $\mathcal{H}_2$  state-feedback control of continuous-time Markov jump linear systems considering that the main Markov chain cannot be directly measured. We consider that the only information available of the main jump process comes from a detector. We assume that the joint process of the process of the plant and the detector follows an extended exponential Markov process, the so-called Exponential Hidden Markov Model. This modelling is appealing to represent systems subject to faults. We present new sufficient conditions for calculating state-feedback controllers depending on the detector that stabilize the closed-loop system while guaranteeing a bound on its  $H_2$  norm. In the case in which the detector is able to provide the correct information regarding the jump process of the plant, the so-called perfect observation case, our conditions also become necessary, leading to the optimal  $H_2$  state-feedback controller. We numerically compare our conditions to the ones already presented in the literature through illustrative examples in the context of networked control systems and systems subject to faults.

**Acknowledgement.** The second author was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), process No. 304149/2019 – 5, by FAPESP/Shell Research Center for Gas Innovation (RCGI) process FAPESP No. 2014/50279 – 4, and by Instituto Nacional de Ciência e Tecnologia para Sistemas Autônomos Cooperativos (INSAC), process CNPq/INCT – 465755/2014 – 3 and FAPESP/INCT-2014/50851 – 0.

## References

1. Boukas, E.K.: Stochastic Switching Systems: Analysis and Design. Birkhäuser, Basel (2006)
2. Boyd, S., El Ghaoui, L., Feron, E., Balakrishnan, V.: Linear Matrix Inequalities in System and Control Theory. SIAM Studies in Applied and Numerical Mathematics. SIAM, Philadelphia (1994)
3. Cardeliquio, C.B., Fioravanti, A.R., Gonçalves A.P.C.:  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  state-feedback control of continuous-time MJLS with uncertain transition rates. In: 2014 ECC, pp. 2237–2241 (2014)

4. Cheng, J., Ahn, C.K., Karimi, H.R., Cao, J., Qi, W.: An event-based asynchronous approach to Markov jump systems with hidden mode detections and missing measurements. *IEEE Trans. Syst. Man Cybern. Syst.* **49**(9), 1749–1758 (2019)
5. Costa, O.L.V., Fragoso, M.D., Todorov, M.G.: *Continuous-Time Markov Jump Linear Systems*. Springer, Berlin-Heidelberg-New York (2013)
6. Costa, O.L.V., Fragoso, M.D., Todorov, M.G.: A detector-based approach for the  $H_2$  control of Markov jump linear systems with partial information. *IEEE Trans. Automat. Contr.* **60**(5), 1219–1234 (2015)
7. Costa, O.L.V., Tuesta, E.F.:  $H_2$ -control and the separation principle for discrete-time Markovian jump linear systems. *Math. Control Signals Syst.* **16**(4), 320–350 (2004)
8. Daafouz, J., Bernussou, J.: Parameter dependent Lyapunov functions for discrete time systems with time varying parametric uncertainties. *Syst. Control Lett.* **43**(5), 355–359 (2001)
9. de Oliveira, M.C., Bernussou, J., Geromel, J.C.: A new discrete-time robust stability condition. *Syst. Control Lett.* **37**(4), 261–265 (1999)
10. de Oliveira, A.M., Costa, O.L.V.:  $H_2$ -filtering for discrete-time hidden Markov jump systems. *Int. J. Control* **90**(3), 599–615 (2017)
11. de Oliveira, A.M., Costa, O.L.V.: Mixed  $\mathcal{H}_2/\mathcal{H}_\infty$  control of hidden Markov jump systems. *Int. J. Robust Nonlinear Control* **28**(4), 1261–1280 (2018)
12. de Oliveira, A.M., Costa, O.L.V.: An iterative approach for the discrete-time dynamic control of Markov jump linear systems with partial information. *Int. J. Robust Nonlinear Control* **30**(2), 495–511 (2020)
13. de Oliveira, A.M., Costa, O.L.V., Fragoso, M.D., Stadtmann, F.: Dynamic output feedback control for continuous-time Markov jump linear systems with hidden Markov models. *Int. J. Control* (2021, in print)
14. Dragan, V., Morozan, T., Stoica, A.-M.: *Mathematical Methods in Robust Control of Linear Stochastic Systems (Mathematical Concepts and Methods in Science and Engineering)*. Springer, New York (2010)
15. Ducard, G.J.J.: *Fault-Tolerant Flight Control and Guidance Systems*. Springer, London-New York (2009)
16. Dufour, F., Elliott, R.J.: Adaptive control of linear systems with Markov perturbations. *IEEE Trans. Automat. Contr.* **43**(3), 351–372 (1998)
17. Elliott, R.J., Aggoun, L., Moore, J.B.: *Hidden Markov Models*. Springer, New York (1995)
18. Li, F., Xu, S., Zhang, B.: Resilient asynchronous  $H_\infty$  control for discrete-time Markov jump singularly perturbed systems based on hidden Markov model. *IEEE Trans. Syst. Man Cybern. Syst.* **50**(8), 2860–2869 (2020)
19. Liu, X., Ma, G., Pagilla, P.R., Ge, S.S.: Dynamic output feedback asynchronous control of networked Markovian jump systems. *IEEE Trans. Syst. Man Cybern. Syst.* **50**(7), 2705–2715 (2020)
20. Mahmoud, M., Jiang, J., Zhang, Y.: *Active Fault Tolerant Control Systems - Stochastic Analysis and Synthesis*. Springer, Germany (2003)
21. Mariton, M.: Detection delays, false alarm rates and the reconfiguration of control systems. *Int. J. Control* **49**, 981–992 (1989)
22. Mariton, M.: *Jump Linear Systems in Automatic Control*. CRC Press, New York (1990)
23. Morais, C.F., Braga, M.F., Oliveira, R.C.L.F., Peres, P.L.D.:  $H_2$  and  $H_\infty$  control design for polytopic continuous-time Markov jump linear systems with uncertain transition rates. *Int. J. Robust Nonlinear Control* **26**(3), 599–612 (2016)

24. Rodrigues, C.C.G., Todorov, M.G., Fragoso, D.M.:  $H_\infty$  control of continuous-time Markov jump linear systems with detector-based mode information. *Int. J. Control* **90**(10), 2178–2196 (2017)
25. Shen, Y., Wu, Z., Shi, P., Su, H., Huang, T.: Asynchronous filtering for Markov jump neural networks with quantized outputs. *IEEE Trans. Syst. Man Cybern. Syst.* **49**(2), 433–443 (2019)
26. Shi, P., Li, F.: A survey on Markovian jump systems: modeling and design. *Int. J. Control Autom. Syst.* **13**(1), 1–16 (2015)
27. Srichander, R., Walker, B.K.: Stochastic stability analysis for continuous-time fault tolerant control systems. *Int. J. Control* **57**(2), 433–452 (1993)
28. Stadtmann, F., Costa, O.L.V.:  $H_2$ -control of continuous-time hidden Markov jump linear systems. *IEEE Trans. Automat. Contr.* **62**(8), 4031–4037 (2017)
29. Stadtmann, F., Costa, O.L.V.: Exponential hidden Markov models for  $H_\infty$  control of jumping systems. *IEEE Control Syst. Lett.* **2**(4), 845–850 (2018)
30. Todorov, M.G., Fragoso, M.D., Costa, O.L.V.: Detector-based  $H_\infty$  results for discrete-time Markov jump linear systems with partial observations. *Automatica* **91**, 159–172 (2018)
31. Val, J.B.R., Geromel, J.C., Gonçalves, A.P.C.: The  $H_2$ -control for jump linear systems: cluster observations of the Markov state. *Automatica* **38**(2), 343–349 (2002)



# Q-Learning for Distributionally Robust Markov Decision Processes

Nicole Bäuerle<sup>(✉)</sup> and Alexander Glauner

Department of Mathematics, Karlsruhe Institute of Technology,  
76128 Karlsruhe, Germany  
[{nicole.baauerle,alexander.glauner}@kit.edu](mailto:{nicole.baauerle,alexander.glauner}@kit.edu)

**Abstract.** In this paper, we consider distributionally robust Markov Decision Processes with Borel state and action spaces and infinite time horizon. The problem is formulated as a Stackelberg game where nature as a second player chooses the least favorable disturbance density in each scenario. Under suitable assumptions, we prove that the value function is the unique fixed point of an operator and that minimizers respectively, maximizers lead to optimal policies for the decision maker and nature. Based on this result, we introduce a Q-learning approach to solve the problem via simulation-based techniques. We prove the convergence of the Q-learning algorithm and study its performance using a distributionally robust irrigation problem.

**Keywords:** Markov decision process · Robust optimization · Q-learning

**AMS (2020) Subject Classification:** Primary 90C40 · Secondary 68T05 · 90C17

## 1 Introduction

The theory of Markov Decision Processes (MDPs) which developed after the groundbreaking work by Richard Bellman (see e.g. [3] or the reprint [4]) has been shown to be extremely useful for solving stochastic dynamic decision problems. Areas of application are among others production planning, operations management, control of robots, scheduling in queueing networks, investment management and health care decisions. The starting point of the theory is a model where the state transition function, the cost function and the distribution of the disturbances are known or can be estimated with sufficient precision. Whereas the transition function is often given due to physical laws, in many cases it might not be possible or very costly to determine the true distribution of the disturbances. Hence, there is some kind of *model uncertainty* or *ambiguity* in the problem. There are various ways to deal with this uncertainty (for an overview in the field of economics see e.g. [13]). In this paper, we approach the problem by considering distributionally robust MDPs. More precisely, this

means that we consider a stochastic dynamic game against nature where nature as a second player tries to choose the least favorable disturbance distribution whereas the decision maker tries to minimize her expected discounted cost. We implement this as a Stackelberg game where the decision maker has to reveal her action first and then nature chooses the disturbance distribution. This can be seen as a *worst-case approach*.

Distributionally robust MDPs with finite state space have been considered before in [11, 19] on a theoretical basis, both for finite and infinite planning horizon. In [1] the finite horizon case has been extended to a situation with Borel state and action spaces and unbounded cost function. The major obstacle here is a sensible introduction of policies for nature. A similar situation is also considered in [7, 12]. In both papers, there is a classical game structure with a predetermined order of actions of both players. In particular, the model assumptions and the choice of the ambiguity set are different from our paper. In the present paper, we consider as ambiguity set the set of densities and thus use a different topology. The advantage is to obtain some relations to dynamic risk measures, see [1]. Indeed, relations like this have been discovered in the economic literature before. There, it is common to speak of *model ambiguity*. For an overview of the recent literature see [8]. We also use different, two-sided bounding functions. In [20] another approach is used, where nested uncertainty sets for the transition laws are considered which correspond to confidence sets.

In the current paper, we will extend the results in [1] to a setting with infinite time horizon. Under some assumptions on the continuity and compactness of the model data and under some growth conditions we will show that the value function of the model is the unique fixed point of a certain operator and that minimizers respectively, maximizers in the optimality equation lead to optimal policies for the decision maker and nature. Based on this result, we provide a Q-learning approach to solve the problem numerically via simulation-based techniques. To the best of our knowledge, this has not been done before. Q-learning can be seen as a combination of value iteration and simulation and also works in the case of a game. Other MDP algorithms like policy improvement cannot be generalized to games in an easy way. Q-learning determines the so-called Q-function from which we can derive the value function immediately. We prove the convergence of the algorithm and study its performance using a distributionally robust irrigation problem. The model is considered with different sizes of the state space and different learning rates. In this application, the state space and the action space of the decision maker are finite.

The paper is organized as follows: In the next section, we introduce our model and the optimization problem. We clarify in particular our ambiguity set. In Sect. 3, we summarize our assumptions and explain the solution theorem which consists of a fixed point statement. We use the weighted supremum norm to deal with the unbounded cost function and rely on Banach's fixed point theorem. In the subsequent section, we discuss the relation of our optimization criterion to risk measures. Section 5 contains the theory of the Q-learning algorithm and

proves in particular its convergence. In Sect. 6, the algorithm is applied to the irrigation example. In particular, the influence of the learning rate is discussed.

## 2 The Markov Decision Model

We consider the following stationary Markov Decision Process with state space  $E$ , action space  $A$  and *infinite planning horizon*. Both state and action space are assumed to be Borel spaces with Borel  $\sigma$ -algebras  $\mathcal{B}(E)$  and  $\mathcal{B}(A)$ , respectively. The possible state-action combinations are a measurable subset  $D \subset E \times A$  such that  $D$  contains the graph of a measurable mapping. The  $x$ -section

$$D(x) = \{a \in A : (x, a) \in D\}$$

is the set of admissible actions in state  $x \in E$ . We assume that the dynamics of the MDP depend on *disturbances*  $Z_1, Z_2, \dots$  which are i.i.d. random elements on a common probability space  $\otimes_{n=1}^{\infty}(\Omega, \mathcal{A}, \mathbb{P})$  with values in a measurable space  $(\mathcal{Z}, \mathfrak{Z})$ . W.l.o.g. we assume that  $Z_n((\omega_1, \omega_2, \dots)) = \omega_n$ . Let  $Z$  be a representative of the disturbance variables. When the current state is  $x_n$ , the controller chooses action  $a_n \in D(x_n)$  and  $z_{n+1}$  is the realization of  $Z_{n+1}$ , then the next state is given by

$$x_{n+1} = T(x_n, a_n, z_{n+1}),$$

where  $T : D \times \mathcal{Z} \rightarrow E$  is a measurable *transition function*. The *one-stage cost function*  $c : D \times E \rightarrow \mathbb{R}$  gives the cost  $c(x, a, x')$  for choosing action  $a$  if the system is in state  $x$  and the next state is  $x'$ .

In what follows we will restrict w.l.o.g. to deterministic Markovian policies, for more details see [1].

**Definition 1.** A measurable mapping  $d : E \rightarrow A$  with  $d(x) \in D(x)$  for every  $x \in E$  is called *decision rule*. A sequence  $\pi = (d_0, d_1, \dots)$  is called *policy*. The set of all policies is denoted by  $\Pi$ . A policy  $\pi$  is called *stationary* if  $\pi = (d, d, \dots)$  for some decision rule  $d$ .

We denote by  $(X_n), (A_n)$  the random state and action processes. In the sequel, we will require  $\mathbb{P}$  to be separable. The transition kernel is given by

$$Q(B|x, a) := \int 1_B(T(x, a, z))\mathbb{P}(dz), \quad B \in \mathcal{B}(E), (x, a) \in D. \quad (1)$$

We assume now that there is some uncertainty about  $\mathbb{P}$ , e.g. because it cannot be estimated properly. Moreover, the decision maker is very risk averse and tries to minimize the expected cost on a worst case basis. We denote by  $\mathcal{M}_1(\Omega, \mathcal{A}, \mathbb{P})$  the set of probability measures on  $(\Omega, \mathcal{A})$  which are absolutely continuous with respect to  $\mathbb{P}$  and define for  $q \in (1, \infty]$

$$\mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P}) := \left\{ \mathbb{Q} \in \mathcal{M}_1(\Omega, \mathcal{A}, \mathbb{P}) : \frac{d\mathbb{Q}}{d\mathbb{P}} \in L^q(\Omega, \mathcal{A}, \mathbb{P}) \right\}.$$

Henceforth, we fix a non-empty subset  $\mathcal{Q} \subseteq \mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P})$  which is referred to as *ambiguity set*. This can be seen as the set of probability measures which may reflect the law of motion. Due to absolute continuity, we can identify  $\mathcal{Q}$  with the set of corresponding densities w.r.t.  $\mathbb{P}$

$$\mathcal{Q}^d := \left\{ \frac{d\mathbb{Q}}{d\mathbb{P}} \in L^q(\Omega, \mathcal{A}, \mathbb{P}) : \mathbb{Q} \in \mathcal{Q} \right\}.$$

Accordingly, we view  $\mathcal{Q}$  as a subset of  $L^q(\Omega, \mathcal{A}, \mathbb{P})$  and endow it with the trace topology of the weak\* topology  $\sigma(L^q, L^p)$  on  $L^q(\Omega, \mathcal{A}, \mathbb{P})$  where  $\frac{1}{p} + \frac{1}{q} = 1$ . The weak\* topology in turn induces a Borel  $\sigma$ -algebra on  $\mathcal{Q}$  making it a measurable space. We obtain the following result (for a proof see the appendix of [1]).

**Lemma 1.** *Let the ambiguity set be norm-bounded (see (A)(vi) below) and the probability measure  $\mathbb{P}$  on  $(\Omega, \mathcal{A})$  be separable. Then  $\mathcal{Q}$  endowed with the weak\* topology  $\sigma(L^q, L^p)$  is a separable metrizable space. If  $\mathcal{Q}$  is additionally weak\* closed, it is even a compact Borel space.*

The controller only knows that the transition kernel (1) at each stage is defined by some  $\mathbb{Q} \in \mathcal{Q}$  instead of  $\mathbb{P}$  but not which one exactly. For example it could be known that the disturbances are normally distributed, but mean and variance are not precisely known, i.e.

$$\mathcal{Q} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in [\mu_1, \mu_2], \sigma \in [\sigma_1, \sigma_2]\}.$$

Since all moments of the normal distribution exist, such an ambiguity set with compact parameter intervals is bounded in the  $L^q$ -norm and Lemma 1 applies.

The controller's worst-case approach can be interpreted as a dynamic game against nature. This means that nature reacts to the controller's action  $a \in D(x)$  at time  $n$  with a measurable decision rule  $\gamma_n : D \rightarrow \mathcal{Q}$ . A *policy of nature* is a sequence of such decision rules  $\gamma = (\gamma_0, \gamma_1, \dots)$ . Let  $\Gamma$  be the set of all policies of nature. Thus, we are faced with a *Stackelberg game* where the controller is the mover and nature is the follower. A proof of the next lemma can be found in the appendix of [1].

**Lemma 2.** *A decision rule  $\gamma : D \rightarrow \mathcal{Q}$  induces a stochastic kernel from  $D$  to  $\Omega$  by*

$$\gamma(B|x, a) := \gamma(x, a)(B), \quad B \in \mathcal{A}, (x, a) \in D.$$

As in the case without ambiguity, the Theorem of Ionescu-Tulcea yields that each initial state  $x \in E$  and pair of policies of the controller and nature  $(\pi, \gamma) \in \Pi \times \Gamma$  induce a unique law of motion

$$\mathbb{Q}_x^{\pi\gamma} := \delta_x \otimes \gamma_0(\cdot|x_0, d_0(x_0)) \otimes \gamma_1(\cdot|x_1, d_1(x_1)) \otimes \dots$$

with corresponding expectation operator  $\mathbb{E}_x^{\pi\gamma}$ .

The value of a policy pair  $(\pi, \gamma) \in \Pi \times \Gamma$  under an infinite planning horizon is defined as

$$J_{\infty\pi\gamma}(x) := \mathbb{E}_x^{\pi\gamma} \left[ \sum_{k=0}^{\infty} \beta^k c(X_k, d_k(X_k), X_{k+1}) \right], \quad x \in E. \quad (2)$$

The corresponding robust value of a policy  $\pi \in \Pi$  of the controller is then the worst case cost

$$J_{\infty\pi}(x) := \sup_{\gamma \in \Gamma} J_{\infty\pi\gamma}(x), \quad x \in E.$$

Hence, the optimality criterion is to minimize this worst case cost

$$J_{\infty}(x) := \inf_{\pi \in \Pi} J_{\infty\pi}(x), \quad x \in E. \quad (3)$$

### 3 Solution Theory for the Distributionally Robust MDP

In order to solve the problem we make the following assumptions:

#### Assumptions (A)

- (i) The set-valued mapping  $E \ni x \mapsto D(x)$  is upper semicontinuous and compact-valued.
- (ii) The transition function  $T$  is continuous in  $(x, a)$ .
- (iii) The one-stage cost function  $c$  is lower semicontinuous.
- (iv) There exist  $\alpha, \underline{\epsilon}, \bar{\epsilon} \geq 0$  with  $\underline{\epsilon} + \bar{\epsilon} = 1$  and measurable functions  $\underline{b} : E \rightarrow (-\infty, -\underline{\epsilon}]$  and  $\bar{b} : E \rightarrow [\bar{\epsilon}, \infty)$  such that for all  $\mathbb{Q} \in \mathcal{Q}$  and  $(x, a) \in D$

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}} [-c^-(x, a, T(x, a, Z))] &\geq \underline{b}(x), & \mathbb{E}^{\mathbb{Q}} [\underline{b}(T(x, a, Z))] &\geq \alpha \underline{b}(x), \\ \mathbb{E}^{\mathbb{Q}} [c^+(x, a, T(x, a, Z))] &\leq \bar{b}(x), & \mathbb{E}^{\mathbb{Q}} [\bar{b}(T(x, a, Z))] &\leq \alpha \bar{b}(x). \end{aligned}$$

- (v) We define  $b : E \rightarrow [1, \infty)$ ,  $b(x) := \bar{b}(x) - \underline{b}(x)$ . For all  $(\bar{x}, \bar{a}) \in D$  there exists an  $\epsilon > 0$  and measurable functions  $\Theta_1^{\bar{x}, \bar{a}}, \Theta_2^{\bar{x}, \bar{a}} : \mathcal{Z} \rightarrow \mathbb{R}_+$  such that  $\Theta_1^{\bar{x}, \bar{a}}(Z), \Theta_2^{\bar{x}, \bar{a}}(Z) \in L^p(\Omega, \mathcal{A}, \mathbb{P})$  and

$$|c(x, a, T(x, a, z))| \leq \Theta_1^{\bar{x}, \bar{a}}(z), \quad b(T(x, a, z)) \leq \Theta_2^{\bar{x}, \bar{a}}(z)$$

for all  $z \in \mathcal{Z}$  and  $(x, a) \in B_\epsilon(\bar{x}, \bar{a}) \cap D$ . Here,  $B_\epsilon(\bar{x}, \bar{a})$  is the closed ball around  $(\bar{x}, \bar{a})$  w.r.t. an arbitrary product metric on  $E \times A$ .

- (vi) The ambiguity set  $\mathcal{Q}$  is norm bounded, i.e.  $\exists K \in [1, \infty)$  such that

$$\mathbb{E} \left| \frac{d\mathbb{Q}}{d\mathbb{P}} \right|^q \leq K$$

for all  $\mathbb{Q} \in \mathcal{Q}$ .

- (vi) The discount factor  $\beta$  satisfies  $\alpha\beta < 1$  with  $\alpha$  from (iv).

- Remark 1.* a) Conditions (i)–(iii) and (v) are needed to ensure the existence of optimal policies. Condition (iv) guarantees that the value functions we are interested in have a finite weighted supremum norm with weight function  $b$ . Condition (vi) is a requirement for Lemma 1 and the last condition ensures the contraction property of the optimality operator.  
b) Note that when  $E$  and  $A$  are finite, conditions (i)–(vi) are automatically satisfied. In particular  $b$  can be chosen as a constant and  $\alpha = 1$ .

It is convenient to introduce the corresponding finite horizon problems. For horizon  $N \in \mathbb{N}$  and policies  $\pi \in \Pi$ ,  $\gamma \in \Gamma$ , we set

$$J_{N\pi\gamma}(x) := \mathbb{E}_x^{\pi\gamma} \left[ \sum_{k=0}^{N-1} \beta^k c(X_k, d_k(X_k), X_{k+1}) \right], \quad x \in E.$$

Moreover, let  $J_{N\pi} = \sup_{\gamma \in \Gamma} J_{N\pi\gamma}$  and  $J_N = \inf_{\pi \in \Pi} J_{N\pi}$ . We first make sure that (2) is well-defined.

**Lemma 3.** *Under Assumptions (A) the sequences  $\{J_{N\pi\gamma}\}_{N \in \mathbb{N}}$ ,  $\{J_{N\pi}\}_{N \in \mathbb{N}}$  and  $\{J_N\}_{N \in \mathbb{N}}$  converge pointwise for every policy pair  $(\pi, \gamma) \in \Pi \times \Gamma$  to limits which are bounded by  $\frac{1}{1-\alpha\beta}b$  from below and by  $\frac{1}{1-\alpha\beta}\bar{b}$  from above. Moreover, it holds*

$$\lim_{N \rightarrow \infty} J_{N\pi\gamma} = J_{\infty\pi\gamma}(x), \quad x \in E.$$

*Proof.* We have for  $1 \leq m \leq N$

$$\begin{aligned} J_{N\pi\gamma}(x) &= J_{m\pi\gamma}(x) + \sum_{k=m+1}^{N-1} \beta^k \mathbb{E}_x^{\pi\gamma} [c(X_k, d_k(X_k), X_{k+1})] \\ &\geq J_{m\pi\gamma}(x) + \sum_{k=m+1}^{N-1} \beta^k \mathbb{E}_x^{\pi\gamma} [-c^-(X_k, d_k(X_k), X_{k+1})] \\ &\geq J_{m\pi\gamma}(x) + \underline{b}(x) \sum_{k=m+1}^{N-1} (\alpha\beta)^k \\ &\geq J_{m\pi\gamma}(x) + \underline{b}(x) \sum_{k=m}^{\infty} (\alpha\beta)^k \\ &=: J_{m\pi\gamma}(x) + \delta_m(x) \end{aligned} \tag{4}$$

where  $\delta_m$  is a non-positive function with  $\lim_{m \rightarrow \infty} \delta_m(x) = 0$  for all  $x \in E$ . Hence, the sequence of functions  $\{J_{N\pi\gamma}\}_{N \in \mathbb{N}}$  is weakly increasing. Taking the supremum over  $\gamma$  (and infimum over  $\pi$ ) on both sides of (4), yields that the sequences  $\{J_{N\pi}\}_{N \in \mathbb{N}}$  and  $\{J_N\}_{N \in \mathbb{N}}$  are weakly increasing, too. By Theorem A.1.6 in [2] all three sequences are convergent. Moreover, we can apply Theorem A3 in [10] which yields

$$J_{\infty\pi\gamma}(x) = \lim_{N \rightarrow \infty} \mathbb{E}_x^{\pi\gamma} \left[ \sum_{k=0}^{N-1} \beta^k c(X_k, d_k(X_k), X_{k+1}) \right] = \lim_{N \rightarrow \infty} J_{N\pi\gamma}(x).$$

In the same way as (4) we can prove that

$$J_{N\pi\gamma}(x) \leq J_{m\pi\gamma}(x) + \bar{b}(x) \sum_{k=m}^{\infty} (\alpha\beta)^k \quad (5)$$

Choosing  $m = 0$  and taking the limit  $N \rightarrow \infty$  in (4) and (5) yields

$$\frac{1}{1 - \alpha\beta} b(x) \leq J_{\infty\pi\gamma}(x) \leq \frac{1}{1 - \alpha\beta} \bar{b}(x).$$

For the other limits the same bounds obviously hold, too.  $\square$

The pointwise limits

$$J_\pi(x) := \lim_{N \rightarrow \infty} J_{N\pi}(x) \quad \text{and} \quad J(x) := \lim_{N \rightarrow \infty} J_N(x), \quad x \in E,$$

are referred to as *limit robust policy value* of  $\pi \in \Pi$  and *limit value function*, respectively.

*Remark 2.* The infinite horizon and limit robust policy values and value functions have the following relations.

- (i) It holds for any policy pair  $(\pi, \gamma) \in \Pi \times \Gamma$  that  $J_{N\pi\gamma} \leq J_{N\pi}$ . By taking the limit  $N \rightarrow \infty$  it follows that  $J_{\infty\pi\gamma} \leq J_\pi$  and finally by taking the supremum over  $\gamma \in \Gamma$

$$J_{\infty\pi}(x) \leq J_\pi(x), \quad x \in E.$$

- (ii) It holds for any policy  $\pi \in \Pi$  that  $J_N \leq J_{N\pi}$ . Taking limits yields

$$J(x) \leq J_\pi(x), \quad x \in E.$$

With the bounding function  $b = \bar{b} - \underline{b}$  we define the function space

$$\mathbb{B}_b := \{v : E \rightarrow \mathbb{R} \mid v \text{ measurable, } \exists \lambda \in \mathbb{R}_+ \text{ s.t. } |v(x)| \leq \lambda b(x) \forall x \in E\}.$$

Endowing  $\mathbb{B}_b$  with the weighted supremum norm

$$\|v\|_b := \sup_{x \in E} \frac{|v(x)|}{b(x)}$$

makes  $(\mathbb{B}_b, \|\cdot\|_b)$  a Banach space, cf. Proposition 7.2.1 in [9]. Note that according to Lemma 3 and Theorem 3.6 and 3.10 in [1] we have  $J, J_\pi, J_{\infty\pi\gamma} \in \mathbb{B}_b$ . To ease the notation we introduce the following operators.

**Definition 2.** For functions  $v \in \mathbb{B}_b$  and for all  $(x, a) \in D, \mathbb{Q} \in \mathcal{Q}$  and decision rules  $d, \gamma$  let

$$Lv(x, a, \mathbb{Q}) := \int c(x, a, T(x, a, z)) + \beta v(T(x, a, z)) \mathbb{Q}(dz),$$

$$\mathcal{T}_{d,\gamma}v(x) := Lv(x, d(x), \gamma(x, d(x))),$$

$$\mathcal{T}_d v(x) := \sup_{\mathbb{Q} \in \mathcal{Q}} Lv(x, d(x), \mathbb{Q}),$$

$$\mathcal{T}v(x) := \inf_{a \in D(x)} \sup_{\mathbb{Q} \in \mathcal{Q}} Lv(x, a, \mathbb{Q}).$$

For the next result define

$$\mathbb{B} := \{v \in \mathbb{B}_b \mid v \text{ lower semicontinuous}\}$$

which is again a complete metric space.

**Lemma 4.** *Given Assumptions (A), the Bellman operator  $\mathcal{T}$  is a contraction on  $\mathbb{B}$  with modulus  $\alpha\beta \in (0, 1)$ .*

*Proof.* Let  $v \in \mathbb{B}$ . It has been established in the proof of Theorem 3.10 in [1] that  $\mathcal{T}v$  is lower semicontinuous. Furthermore,

$$\begin{aligned} |\mathcal{T}v(x)| &= \left| \inf_{a \in D(x)} \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [c(x, a, T(x, a, Z)) + \beta v(T(x, a, Z))] \right| \\ &\leq \inf_{a \in D(x)} \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [|c(x, a, T(x, a, Z))|] + \beta \mathbb{E}^{\mathbb{Q}} [|v(T(x, a, Z))|] \\ &\leq \inf_{a \in D(x)} \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [|c(x, a, T(x, a, Z))|] + \beta \lambda \mathbb{E}^{\mathbb{Q}} [|b(T(x, a, Z))|] \\ &\leq (1 + \alpha\beta\lambda)b(x), \end{aligned}$$

Hence, the operator  $\mathcal{T}$  is an endofunction on  $\mathbb{B}$  and it remains to verify the Lipschitz constant  $\alpha\beta$ . It holds for  $v_1, v_2 \in \mathbb{B}$

$$\begin{aligned} \mathcal{T}v_1(x) - \mathcal{T}v_2(x) &\leq \sup_{a \in D(x)} \left( \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [c(x, a, T(x, a, Z)) + \beta v_1(T(x, a, Z))] \right. \\ &\quad \left. - \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [c(x, a, T(x, a, Z)) + \beta v_2(T(x, a, Z))] \right) \\ &\leq \beta \sup_{a \in D(x)} \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [|v_1(T(x, a, Z)) - v_2(T(x, a, Z))|] \\ &\leq \beta \|v_1 - v_2\|_b \sup_{a \in D(x)} \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [|b(T(x, a, Z))|] \\ &\leq \alpha\beta \|v_1 - v_2\|_b b(x). \end{aligned}$$

Interchanging the roles of  $v_1$  and  $v_2$  yields

$$|\mathcal{T}v_1(x) - \mathcal{T}v_2(x)| \leq \alpha\beta \|v_1 - v_2\|_b b(x).$$

Now, dividing by  $b(x)$  and taking the supremum over  $x \in E$  on the left hand side completes the proof.  $\square$

The following theorem is a consequence of Proposition 3.5, Theorem 3.6 and Theorem 3.10 in [1]. It is crucial for our main result.

**Theorem 1.** *Let Assumptions (A) be satisfied and policies  $\pi = (d_0, d_1, \dots) \in \Pi$  and  $\gamma = (\gamma_0, \gamma_1, \dots) \in \Gamma$  be given with  $\bar{\pi} := (d_1, d_2, \dots)$ ,  $\bar{\gamma} := (\gamma_1, \gamma_2, \dots)$ .*

- a) For all  $N \in \mathbb{N}$  we have  $J_{N\pi\gamma} = \mathcal{T}_{d_0, \gamma_0} J_{N-1\bar{\pi}\bar{\gamma}}$ .
- b) For all  $N \in \mathbb{N}$  we have  $J_{N\pi} = \mathcal{T}_{d_0} J_{N-1\bar{\pi}}$ .

c) For all  $N \in \mathbb{N}$  we have  $J_N = \mathcal{T}J_{N-1}$  and  $J_N \in \mathbb{B}$ .

The next theorem is our main result. It characterizes the value function and explains how optimal policies for the decision maker and nature can be obtained.

**Theorem 2.** *Let Assumptions (A) be satisfied.*

- a) *The limit value function  $J$  is the unique fixed point of the Bellman operator  $\mathcal{T}$  in  $\mathbb{B}$ .*
- b) *There exists a decision rule  $d^* : E \rightarrow A$  of the controller such that*

$$\mathcal{T}_{d^*} J(x) = \mathcal{T}J(x), \quad x \in E.$$

*Moreover, for every  $\epsilon > 0$  there exists an  $\epsilon$ -optimal decision rule  $\hat{\gamma}_0 : D \rightarrow \mathcal{Q}$  of nature such that*

$$\mathcal{T}_{d^* \hat{\gamma}_0} J(x) + \epsilon \geq \mathcal{T}J(x), \quad x \in E.$$

- c) *If the ambiguity set  $\mathcal{Q}$  is weak\* closed, there exists a decision rule  $\gamma_0^* : D \rightarrow \mathcal{Q}$  of nature such that*

$$\mathcal{T}_{d^* \gamma_0^*} J(x) = \mathcal{T}J(x), \quad x \in E.$$

- d) *Each stationary policy  $\pi^* = (d^*, d^*, \dots)$  induced by a decision rule  $d^*$  as in part b) is optimal for optimization problem (3) and it holds  $J_\infty = J$ .*
- e) *If the ambiguity set  $\mathcal{Q}$  is weak\* closed, each stationary policy  $\gamma^* = (\gamma_0^*, \gamma_0^*, \dots)$  induced by a decision rule  $\gamma_0^*$  as in part c) is an optimal response of nature to  $\pi^*$ , i.e.  $J_{\infty \pi^* \gamma^*} = J_\infty$ .*

*Proof.* a) The fact that  $J$  is the unique fixed point of the operator  $\mathcal{T}$  in  $\mathbb{B}$  follows directly from Banach's Fixed Point Theorem using Lemma 4.

- b) The existence of a minimizing decision rule of the controller and an  $\epsilon$ -optimal decision rule of nature follow from the respective results in the finite horizon case, cf. Theorem 3.10 a) in [1].
- c) This follows analogously from Theorem 3.10 b) in [1].
- d) Let  $d^*$ ,  $\hat{\gamma}_0$  be decision rules as in part b) and  $\pi^* := (d^*, d^*, \dots)$ ,  $\hat{\gamma} := (\hat{\gamma}_0, \hat{\gamma}_0, \dots)$ . It has to be shown that

$$J_{\infty \pi^*}(x) = J_\infty(x) = J(x), \quad x \in E. \tag{6}$$

We proceed in two steps. Firstly, we prove that

$$J(x) \geq J_{\infty \pi^*}(x), \quad x \in E \tag{7}$$

and secondly we prove that

$$J(x) \leq J_{\infty \pi}(x), \quad x \in E, \quad \text{for all } \pi \in \Pi. \tag{8}$$

From (7) we get  $J \geq J_{\infty \pi^*} \geq J_\infty$ . On the other hand, taking the infimum over  $\pi \in \Pi$  in (8) yields  $J \leq J_\infty$ . Together, these inequalities imply (6) and

the assertion is proven.

*Step 1:* We show by induction that for all  $N \in \mathbb{N}_0$

$$J(x) \geq J_{N\pi^*}(x) + \frac{(\alpha\beta)^N}{1-\alpha\beta} b(x), \quad x \in E.$$

Then letting  $N \rightarrow \infty$  yields (7). The case  $N = 0$  follows from Lemma 3. For  $N \geq 1$  it follows from the induction hypothesis

$$\begin{aligned} J(x) &= \mathcal{T}_{d^*} J(x) \\ &= \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} [c(x, d^*(x), T(x, d^*(x), Z)) + \beta J(T(x, d^*(x), Z))] \\ &\geq \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} \left[ c(x, d^*(x), T(x, d^*(x), Z)) + \beta J_{N-1\pi^*}(T(x, d^*(x), Z)) \right. \\ &\quad \left. + \beta \frac{(\alpha\beta)^{N-1}}{1-\alpha\beta} b(T(x, d^*(x), Z)) \right] \\ &\geq \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} \left[ c(x, d^*(x), T(x, d^*(x), Z)) + \beta J_{N-1\pi^*}(T(x, d^*(x), Z)) \right. \\ &\quad \left. + \frac{(\alpha\beta)^N}{1-\alpha\beta} b(x) \right] \\ &= J_{N\pi^*}(x) + \frac{(\alpha\beta)^N}{1-\alpha\beta} b(x). \end{aligned}$$

Note that the last inequality is by Assumption (A) (ii) and the last equality by Theorem 1 b).

*Step 2:* Let  $\pi = (d_0, d_1, \dots) \in \Pi$  be arbitrary. We show by induction for  $\varepsilon$  and  $\hat{\gamma}$  from b) that for all  $N \in \mathbb{N}_0$

$$J(x) \leq J_{N\pi\hat{\gamma}}(x) + \frac{\epsilon}{1-\beta} + \frac{(\alpha\beta)^N}{1-\alpha\beta} \bar{b}(x), \quad x \in E.$$

Then letting  $N \rightarrow \infty$  yields  $J \leq J_{\infty\pi\hat{\gamma}} + \frac{\epsilon}{1-\beta}$ . Since  $\epsilon > 0$  is arbitrarily small, it follows that  $J \leq J_{\infty\pi}$ , i.e. (8) holds. The case  $N = 0$  follows again from

Lemma 3. For  $N \geq 1$  we have

$$\begin{aligned}
J(x) &= \mathcal{T}J(x) \leq \mathcal{T}_{d^*\hat{\gamma}_0}J(x) + \epsilon \leq \mathcal{T}_{d_0\hat{\gamma}_0}J(x) + \epsilon \\
&\leq \mathcal{T}_{d_0\hat{\gamma}_0}\left(J_{N-1\bar{\pi}\hat{\gamma}}(x) + \frac{\epsilon}{1-\beta} + \frac{(\alpha\beta)^{N-1}}{1-\alpha\beta}\bar{b}(x)\right) + \epsilon \\
&= \int c(x, d_0(x), T(x, d_0(x), z)) + \beta J_{N-1\bar{\pi}\hat{\gamma}}(T(x, d_0(x), z)) \\
&\quad + \beta \frac{(\alpha\beta)^{N-1}}{1-\alpha\beta}\bar{b}(T(x, d_0(x), z))\hat{\gamma}_0(dz|x, d_0(x)) + \left(1 + \frac{\beta}{1-\beta}\right)\epsilon \\
&= J_{N\pi\hat{\gamma}}(x) + \beta \frac{(\alpha\beta)^{N-1}}{1-\alpha\beta}\int \bar{b}(T(x, d_0(x), z))\hat{\gamma}_0(dz|x, d_0(x)) + \frac{\epsilon}{1-\beta} \\
&\leq J_{N\pi\hat{\gamma}}(x) + \beta \frac{(\alpha\beta)^{N-1}}{1-\alpha\beta} \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\bar{b}(T(x, d_0(x), Z))] + \frac{\epsilon}{1-\beta} \\
&\leq J_{N\pi\hat{\gamma}}(x) + \frac{(\alpha\beta)^N}{1-\alpha\beta}\bar{b}(x) + \frac{\epsilon}{1-\beta}.
\end{aligned}$$

We used that  $\pi \in \Pi$  is arbitrary, so it is no problem to apply the induction hypothesis to the shifted policy  $\bar{\pi}$ . The third equality is by Theorem 1 a).

- e) Replacing the  $\epsilon$ -optimal decision rule  $\hat{\gamma}_0$  by the optimal one  $\gamma_0^*$  in step 2 of part d) yields  $J \leq J_{\infty\pi\gamma^*}$  for all  $\pi \in \Pi$ , so especially  $J \leq J_{\infty\pi^*\gamma^*}$ . Combining this with (6), we get

$$J \leq J_{\infty\pi^*\gamma^*} \leq J_{\infty\pi^*} = J_{\infty} = J,$$

which concludes the proof.  $\square$

*Remark 3.* Note that we do not have a classical game here. In particular it is not possible in general to interchange sup and inf. Additional properties like convexity are required to achieve this. For a discussion and examples, see [1].

## 4 Connection to Risk Measures

In this section, we outline how distributionally robust MDPs are related to the minimization of coherent risk measures. This provides another interpretation of the optimality criterion (3) in addition to the worst-case approach and the dynamic Stackelberg game. A *risk measure* is a functional  $\rho : L^p(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \mathbb{R}$  which determines the necessary capital to make holding a risky position  $X \in L^p(\Omega, \mathcal{A}, \mathbb{P})$  acceptable. The following properties are important.

**Definition 3.** A risk measure  $\rho : L^p(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \mathbb{R}$  is

- a) monotone if  $X \leq Y$  implies  $\rho(X) \leq \rho(Y)$ .
- b) translation invariant if  $\rho(X + m) = \rho(X) + m$  for all  $m \in \mathbb{R}$ .
- c) positive homogeneous if  $\rho(\lambda X) = \lambda\rho(X)$  for all  $\lambda \in \mathbb{R}_+$ .

- d) subadditive if  $\rho(X + Y) \leq \rho(X) + \rho(Y)$  for all  $X, Y$ .
- e) coherent if it has properties a)-d).
- f) said to have the Fatou property, if for every sequence  $\{X_n\}_{n \in \mathbb{N}} \subseteq L^p$  with  $|X_n| \leq Y \mathbb{P}\text{-a.s.}$  for some  $Y \in L^p$  and  $X_n \rightarrow X \mathbb{P}\text{-a.s.}$  for some  $X \in L^p$  it holds

$$\liminf_{n \rightarrow \infty} \rho(X_n) \geq \rho(X).$$

Recall that an extended real-valued convex functional is called *proper* if it never attains  $-\infty$  and is strictly smaller than  $+\infty$  in at least one point. Coherent risk measures have the following dual or robust representation, cf. Theorem 7.20 in [16].

**Theorem 3.** *A functional  $L^p(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \bar{\mathbb{R}}$  is a proper coherent risk measure with the Fatou property if and only if there exists a subset  $\mathcal{Q} \subseteq \mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P})$  such that*

$$\rho(X) = \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X], \quad X \in L^p. \quad (9)$$

The supremum is attained since the subset  $\mathcal{Q} \subseteq \mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P})$  can be chosen  $\sigma(L^q, L^p)$ -compact and the functional  $\mathbb{Q} \mapsto \mathbb{E}^{\mathbb{Q}}[X]$  is  $\sigma(L^q, L^p)$ -continuous.

With this duality result we can reformulate the right hand side of the fixed point equation  $J = \mathcal{T}J$  from Theorem 2 to

$$J(x) = \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z)) + \beta J(T(x, a, Z))\right) \quad (10)$$

for some proper coherent risk measure  $\rho$  with the Fatou property if and only if the ambiguity set  $\mathcal{Q}$  is weak\* closed. Note that we already require  $\mathcal{Q}$  to be norm bounded, cf. Assumption (A) (vi), and by the Theorem of Banach-Alaoglu weak\* compact is equivalent to norm bounded and weak\* closed. Equation (10) shows that for a weak\* closed ambiguity set the distributionally robust optimality criterion is equivalent to the stage-wise minimization of a coherent risk measure.

Due to this connection, the dual representations of coherent risk measures are a natural source for ambiguity sets. A particular advantage is that there are often explicit formulas for nature's maximizing probability measure. We present two examples. Since the probability measures in  $\mathcal{Q}$  are absolutely continuous w.r.t.  $\mathbb{P}$ , we can consider the set of densities  $\mathcal{Q}^d$ .

(i) *Expected Shortfall* is defined on  $L^1(\Omega, \mathcal{A}, \mathbb{P})$  as

$$\text{ES}_\alpha(X) := \frac{1}{1-\alpha} \int_\alpha^1 F_X^{-1}(u) du, \quad \alpha \in [0, 1),$$

with  $F_X^{-1}$  denoting the quantile function of  $X$ . Its dual representation is based on the set of densities

$$\mathcal{Q}^d = \left\{ Y \in L^\infty(\Omega, \mathcal{A}, \mathbb{P}) : \mathbb{E}[Y] = 1, Y \leq \frac{1}{1-\alpha} \right\}.$$

The supremum (9) is attained in

$$Y = \frac{\mathbf{1}\{X > F_X^{-1}(\alpha)\} + \kappa \mathbf{1}\{X = F_X^{-1}(\alpha)\}}{1 - \alpha}$$

with  $\kappa = \frac{1 - \alpha - \mathbb{P}(X > F_X^{-1}(\alpha))}{\mathbb{P}(X = F_X^{-1}(\alpha))} \mathbf{1}\{\mathbb{P}(X = F_X^{-1}(\alpha)) > 0\}$ , see Remark 8.15 in [14].

- (ii) A superclass are the *spectral risk measures*  $\rho_\phi : L^p(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \bar{\mathbb{R}}$ . They are of the form

$$\rho_\phi(X) := \int_0^1 F_X^{-1}(u)\phi(u)du,$$

where  $\phi : [0, 1] \rightarrow \mathbb{R}_+$  is an increasing function with  $\|\phi\|_q < \infty$  and  $\int_0^1 \phi(u)du = 1$  called *spectrum*. Expected Shortfall is a special case with spectrum  $\phi(u) = \frac{\mathbf{1}\{u \geq \alpha\}}{1 - \alpha}$ . The dual representation of spectral risk measures is given by the set of densities

$$\mathcal{Q}^d = \{Y \in L^q(\Omega, \mathcal{A}, \mathbb{P}) : Y \leq_{cx} \phi(U), U \sim \mathcal{U}(0, 1)\}.$$

The maximizing density in (9) is  $\phi(U_X)$ , where  $U_X$  is the generalized distributional transform of  $X$ , i.e. a uniformly distributed random variable satisfying almost surely  $F_X^{-1}(U_X) = X$ , see Corollary 12 in [15].

The connection of distributionally robust MDPs to coherent risk measures goes beyond the stage-wise perspective of (10). The optimality criterion (3) can be written as

$$J_\infty(x) = \inf_{\pi \in \Pi} \sup_{\mathbb{Q} \in \mathfrak{Q}_\pi} \mathbb{E}_x^\mathbb{Q} \left[ \sum_{k=0}^{\infty} \beta^k c(X_k, d_k(X_k), X_{k+1}) \right], \quad x \in E,$$

where  $\mathfrak{Q}_\pi = \{\mathbb{Q}_x^{\pi\gamma} : \gamma \in \Gamma\}$ . By direct verification of the axioms one can see that for a fixed policy  $\pi \in \Pi$  of the controller  $\tilde{\rho}(X) = \sup_{\mathbb{Q} \in \mathfrak{Q}_\pi} \mathbb{E}^\mathbb{Q}[X]$ ,  $X \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ , defines a coherent risk measure. I.e. in some sense the stage-wise connection (10) holds also globally. If the ambiguity set  $\mathcal{Q}$  is induced by a spectral risk measure and the model data has certain monotonicity properties,  $\mathfrak{Q}$  is independent of  $\pi$ , cf. Lemma 6.8 and subsequent remarks in [1]. In this case, the distributionally robust expected cost optimization is equivalent to the minimization of a coherent risk measure applied to the total cost.

## 5 Q-Learning for Distributionally Robust Models

We want to obtain  $J = J_\infty$  and the optimal policy numerically. In order to achieve this, we use a  $Q$ -learning algorithm. For simplicity let us assume now that state and action space are finite as well as the ambiguity set. Thus Assumptions (A) (i)–(vi) are automatically satisfied (see Remark 1). We only have to assume

that  $\beta < 1$ . In what follows it will be more convenient to work with the densities  $\mathcal{Q}^d$  instead of  $\mathcal{Q}$ . The fixed point equation of Theorem 2 a) reads

$$J(x) = \mathcal{T}J(x) = \inf_{a \in D(x)} \sup_{Y \in \mathcal{Q}^d} \sum_z \mathbb{P}(z) Y(z) \left( c(x, a, T(x, a, z)) + \beta J(T(x, a, z)) \right).$$

The *Q-function* of the problem is for  $(x, a, Y) \in D \times \mathcal{Q}^d$  given by

$$Q(x, a, Y) := LJ(x, a, Y).$$

It is the value when we take the pair  $(a, Y)$  as the first action of the decision maker and nature and act optimally afterwards. In particular, we have  $J(x) = \mathcal{T}J(x) = \inf_{a \in D(x)} \sup_{Y \in \mathcal{Q}^d} Q(x, a, Y)$ . Thus, we obtain

$$\begin{aligned} Q(x, a, Y) &= \sum_z \mathbb{P}(z) Y(z) \left( c(x, a, T(x, a, z)) + \beta \inf_{a' \in D(x)} \sup_{Y' \in \mathcal{Q}^d} Q(T(x, a, z), a', Y') \right) \\ &=: HQ(x, a, Y) \end{aligned} \quad (11)$$

The operator  $H$  is slightly different to  $\mathcal{T}$ , however they share the following important property. In what follows we denote by  $\|\cdot\|_\infty$  the supremum norm on the Banach space of bounded functions.

**Theorem 4.** *The operator  $H$  is a contraction on the space of bounded functions with modulus  $\beta \in (0, 1)$  and  $Q$  is the unique fixed point of the  $H$ -operator in the set of bounded functions.*

*Proof.* First note that when  $Q$  is bounded,  $HQ$  is bounded, too. Now take two bounded functions  $Q_1, Q_2$  on  $D \times \mathcal{Q}^d$ . Then

$$\begin{aligned} &(HQ_1 - HQ_2)(x, a, Y) \\ &= \beta \sum_z \mathbb{P}(z) Y(z) \left( \inf_{a'} \sup_{Y'} Q_1(T(x, a, z), a', Y') - \inf_{a'} \sup_{Y'} Q_2(T(x, a, z), a', Y') \right) \\ &\leq \beta \sum_z \mathbb{P}(z) Y(z) \sup_{a'} \sup_{Y'} \left( Q_1(T(x, a, z), a', Y') - Q_2(T(x, a, z), a', Y') \right) \\ &\leq \beta \|Q_1 - Q_2\|_\infty \end{aligned}$$

Interchanging  $Q_1$  and  $Q_2$  finally yields  $\|HQ_1 - HQ_2\|_\infty \leq \beta \|Q_1 - Q_2\|_\infty$  and implies that  $H$  is contracting. Thus, it follows from Banach's fixed point theorem that the fixed point of  $H$  in the set of bounded functions is unique. That  $Q$  is a fixed point follows from (11).  $\square$

We consider the following iteration with numbers  $\alpha_t \geq 0$  called *learning rate* and satisfying  $\lim_{t \rightarrow \infty} \alpha_t = 0$ . We start with  $Q^{(0)} \equiv 0$ . In each step, we choose randomly a feasible pair  $(x, a, Y)$ , generate  $z$  according to  $\mathbb{P}$  and update  $Q^{(t)}$ .

**Algorithm:**

1. Set  $Q^{(0)} \equiv 0$ .
2. Choose a pair  $(x, a, Y)$  at random (uniformly over  $D \times \mathcal{Q}^d$ ) and generate  $z$  according to  $\mathbb{P}$ .
3. Update the value at  $(x, a, Y)$  by

$$\begin{aligned} Q^{(t+1)}(x, a, Y) = & (1 - \alpha_t)Q^{(t)}(x, a, Y) + \alpha_t Y(z) \left( c(x, a, T(x, a, z)) \right. \\ & \left. + \beta \min_{a'} \max_{Y'} Q^{(t)}(T(x, a, z), a', Y') \right) \end{aligned}$$

and set  $Q^{(t+1)}(\cdot) = Q^{(t)}(\cdot)$  for all other arguments.

It is now possible to prove that the iteration converges to the  $Q$ -function.

**Theorem 5.** *If the numbers  $(\alpha_t)$  are chosen such that*

$$\sum_{t=0}^{\infty} \alpha_t = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty,$$

*then  $\{Q^{(t)}\}_{t \in \mathbb{N}_0}$  converges with probability 1 to  $Q$  for  $t \rightarrow \infty$ .*

*Proof.* Note that we can write the iteration as

$$\begin{aligned} Q^{(t+1)}(x, a, Y) = & (1 - \alpha_t)Q^{(t)}(x, a, Y) \\ & + \alpha_t Y(z) \left( c(x, a, T(x, a, z)) + \beta \min_{a'} \max_{Y'} Q^{(t)}(T(x, a, z), a', Y') \right) \\ = & (1 - \alpha_t)Q^{(t)}(x, a, Y) + \alpha_t \left( HQ^{(t)}(x, a, Y) + w_t(x, a, Y) \right) \end{aligned}$$

where

$$\begin{aligned} w_t(x, a, Y) = & Y(z) \left( c(x, a, T(x, a, z)) + \beta \min_{a'} \max_{Y'} Q^{(t)}(T(x, a, z), a', Y') \right) \\ & - HQ^{(t)}(x, a, Y). \end{aligned}$$

The statement follows from Proposition 4.4 in [5] since  $H$  is contracting and the random variables  $W_t(x, a, Y)$  which are obtained from  $w_t(x, a, Y)$  by replacing the realisation  $z$  by its random counterpart  $Z$  satisfy

- (i)  $\mathbb{E}W_t(x, a, Y) = 0$  by definition of the  $H$ -operator.
- (ii)  $\mathbb{E}W_t^2(x, a, Y)$  is bounded.

Thus, we can apply Proposition 4.4 in [5]. □

Once we have obtained  $Q$  we can compute  $J$  and the minimizer  $d^*$  and maximizer  $\gamma^*$  which yields the optimal policies.

*Remark 4.* Note that the  $Q$ -learning algorithm is model-free in the sense that it is not necessary to know the probability law  $\mathbb{P}$ . Instead of simulating  $z$  one can of course use observed model data if available.

## 6 Numerical Example

In this section, we apply the distributionally robust Q-learning algorithm to an agricultural irrigation management problem. With progressing climate change, water becomes a scarce resource in many regions of the world which must be carefully managed. Therefore, mathematical optimization may be needed where simple rules of thumb have been sufficient in the past. The stylized setting of this example is designed to illustrate the performance of our algorithm. It can be easily extended to a practical model. We refer the interested reader to [17, 18] for some approaches in continuous time.

Consider a greenhouse that is irrigated from a water reservoir with capacity  $\bar{s} \in \mathbb{N}$ . One unit of water is needed for every irrigation procedure. The crops rot when irrigated on two consecutive days and wither if they are not watered again within  $\bar{x} \geq 2$  days. Both events destroy the harvest and a fixed cost  $c > 0$  is incurred. Precipitation may occur on each day independently with probability  $p \in [p_1, p_2] \subset (0, 1)$  and add one unit of water to the reservoir. The true rain probability is unknown and it is therefore prudent to work with the confidence interval  $[p_1, p_2]$  instead of a single estimate. I.e.  $\mathcal{Q}$  consists of all Bernoulli distributions with parameter between  $p_1$  and  $p_2$ . Thus, we can identify  $\mathcal{Q}^d$  with the parameter set  $[p_1, p_2]$ . If the maximal capacity of the reservoir is exceeded, the spillover goes into the greenhouse like a regular irrigation. The corresponding Markov decision model is given by the following data.

- (i) The state space is  $(\{0, \dots, \bar{x}\} \times \{0, \dots, \bar{s}\}) \cup \{\infty\}$ . The first component of a state  $(x, s)$  gives the days since the last irrigation and the second one the current level of the water reservoir. The absorbing state  $\infty$  corresponds to a destroyed harvest.
- (ii) The action space is  $\{0, 1\}$ . Action  $a = 1$  means that the crops are watered and  $a = 0$  that they are not. The decision maker faces no constraint.
- (iii) The i.i.d. disturbances  $Z_1, Z_2, \dots \sim \text{Bin}(1, p)$ ,  $p \in [p_1, p_2]$  model the amount of daily precipitation.
- (iv) The transition function  $T(x, s, a, z)$  is given by

$$\begin{aligned}
 T(x, s, 0, 0) &= \begin{cases} (x + 1, s), & x < \bar{x} \\ \infty, & x = \bar{x} \end{cases} \\
 T(x, s, 1, 0) &= \begin{cases} (0, s - 1), & x > 0, s > 0 \\ (x + 1, 0), & x < \bar{x}, s = 0 \\ \infty, & x = \bar{x}, s = 0 \text{ or } x = 0, s > 0 \end{cases} \\
 T(x, s, 0, 1) &= \begin{cases} (x + 1, s + 1), & x < \bar{x}, s < \bar{s} \\ \infty, & x = \bar{x}, s < \bar{s} \text{ or } x = 0, s = \bar{s} \\ (0, \bar{s}), & x > 0, s = \bar{s} \end{cases} \\
 T(x, s, 1, 1) &= \begin{cases} (0, s), & x > 0 \\ \infty, & x = 0 \end{cases}
 \end{aligned}$$

and  $T(\infty, a, z) = \infty$ .

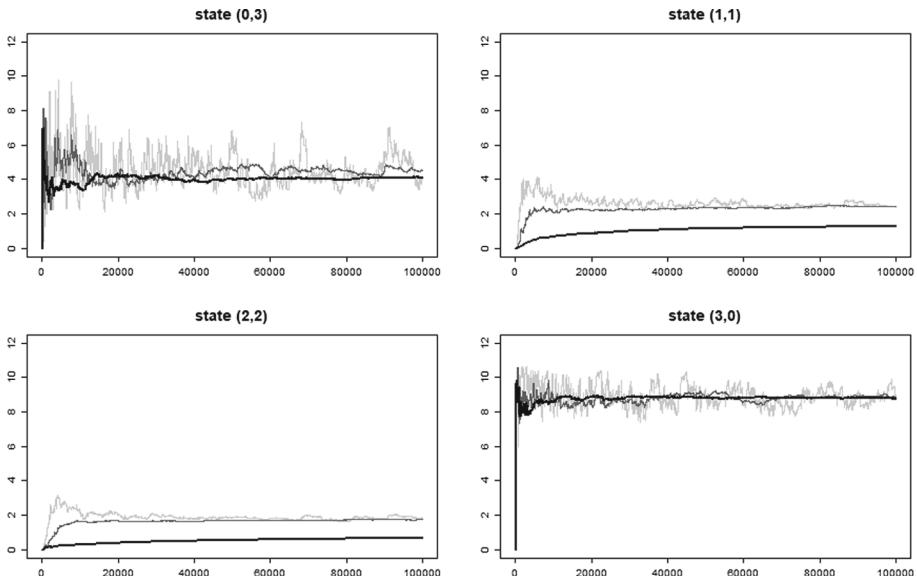
(v) The one-stage cost function is  $c(x, s, x', s') = c \mathbf{1}\{(x, s) \neq \infty, (x', s') = \infty\}$ .

The model clearly satisfies Assumptions (A). The target of the decision maker is to minimize the expected discounted cost

$$J_\infty(x, s) = \inf_{\pi \in \Pi} \sup_{\gamma \in \Gamma} \mathbb{E}_{x, s}^{\pi \gamma} \left[ \sum_{k=0}^{\infty} \beta^k c(X_k, S_k, X_{k+1}, S_{k+1}) \right]$$

under the assumption of being confronted with the most adverse precipitation probability  $p$  on each day. This means that the decision maker tries to avoid ruin if ever possible or to delay it to a later time point. His opponent in the dynamic Stackelberg game is nature in the proper meaning of the word. She selects the rain probability knowing the current state and the decision maker's action. Since expectation is linear in the measure, her optimal action can only be at the boundary, i.e.  $p_1$  or  $p_2$ . So we have a robust point of view here.

For the implementation of the Q-learning algorithm we selected  $\beta = 0.9$  as discount factor,  $c = 10$  as fixed cost,  $\bar{x} = 3$  as time until withering,  $p_1 = 0.2$ ,  $p_2 = 0.3$  and 0.25 as reference probability for the two densities  $Y_1(z) = \frac{0.2}{0.25} \mathbf{1}\{z = 1\} + \frac{0.8}{0.75} \mathbf{1}\{z = 0\}$  and  $Y_2(z) = \frac{0.3}{0.25} \mathbf{1}\{z = 1\} + \frac{0.7}{0.75} \mathbf{1}\{z = 0\}$  that nature may select. At first, the maximal capacity of the reservoir is  $\bar{s} = 3$ .



**Fig. 1.** Approximation of the value function in different states as a function of the number of iterations.

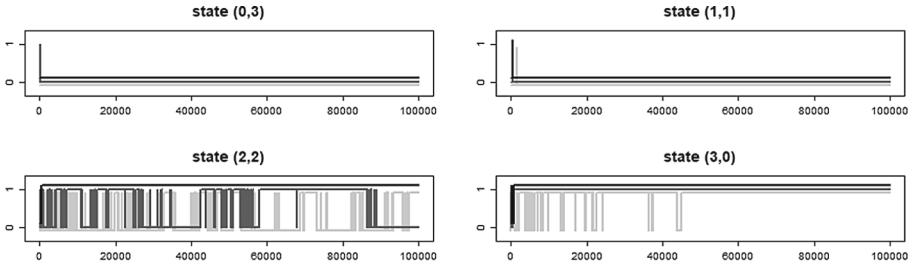
Figure 1 shows the convergence of the approximated value function

$$J^{(t)}(x, s) = \min_a \max_Y Q^{(t)}(x, s, a, Y), \quad t = 0, \dots, 100000$$

in four exemplary states. State  $(0, 3)$  represents an imminent spillover,  $(3, 0)$  imminent withering and  $(1, 1)$ ,  $(2, 2)$  are two moderate situations. We compared three different learning rates.

black curve:	dark grey curve:	light grey curve:
$\alpha_t = \frac{0.5}{1 + 0.01t}$	$\alpha_t = \frac{0.5}{1 + 0.001t}$	$\alpha_t = \frac{0.5}{1 + 0.0001t}$

The same color code is used in all other figures, too. The faster the learning rate goes to zero, the earlier the approximate cost stabilizes. In the two extreme states  $(0, 3)$  and  $(3, 0)$ , where the optimal action of both players is obvious, even the learning rate with the strongest decay yields a good approximation. In the two moderate states, where the path to ruin is longer, the strong decay essentially terminates the approximation too early. On the other hand, the slowest decay works well in case of a long path to ruin while convergence in the two extreme states takes unnecessarily long. The medium decay seems to be a suitable compromise for all states.



**Fig. 2.** Approximation of the decision maker's optimal policy in different states as a function of the number of iterations.

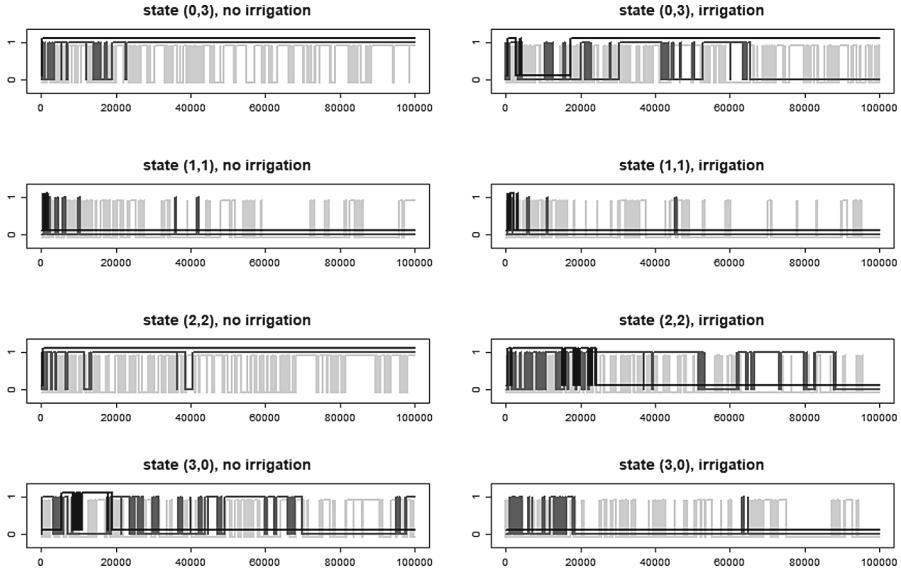
Figure 2 shows the convergence of the approximated optimal policy of the decision maker

$$\pi^{(t)}(x, s) = \arg \min_a \max_Y Q^{(t)}(x, s, a, Y), \quad t = 0, \dots, 100000$$

in the same states and for the same learning rates. In  $(0, 3)$  and  $(1, 1)$  the learning rates are indistinguishable which is also true in  $(3, 0)$  from iteration 50000 onward. Only in state  $(2, 2)$  the minimizing argument remains rather unstable despite the fast stabilization of the minimal value shown in Fig. 1. I.e. here the two actions lead to almost the same cost.

Figure 3 displays the convergence of the approximated optimal policy of nature

$$\gamma^{(t)}(x, s, a) = \arg \max_Y Q^{(t)}(x, s, a, Y), \quad t = 0, \dots, 100000$$



**Fig. 3.** Approximation of nature's optimal policy in different state-action combinations as a function of the number of iterations.

again in the same states and for the same learning rates. With the third learning rate, nature's optimal action does not stabilize during the first 100000 iterations in all four states. The other two learning rates perform better. In the relevant scenarios given optimal behavior of the decision maker  $(x, s, a) = (0, 3, 0), (1, 1, 0), (3, 0, 1)$  we observe an early stabilization under the two learning rates with faster decay. In state  $(2, 2)$  the stabilization is good at least for action  $a = 0$ .

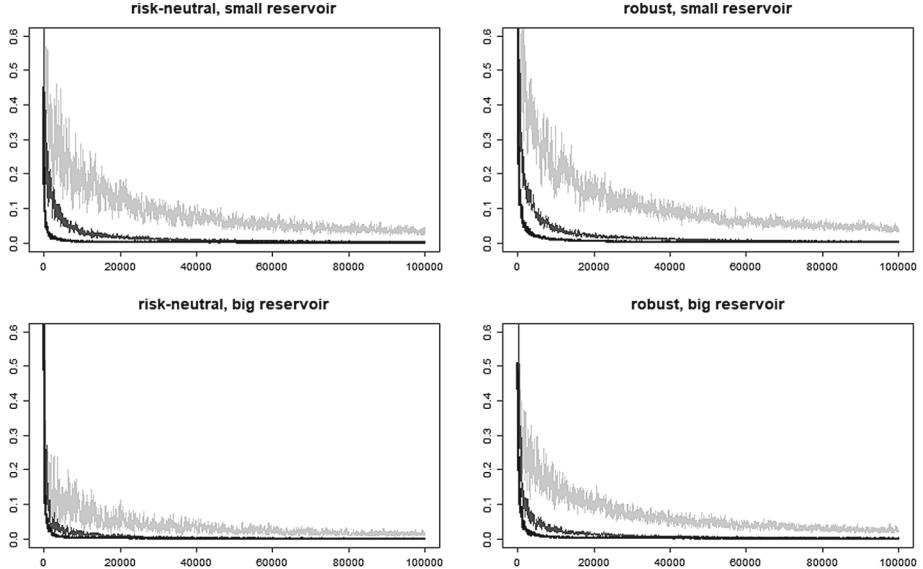
All in all, the second learning rate appears to be the best choice in this application with fast convergence of the value function to the true optimal cost and a relatively good stabilization of the optimizing arguments.

In Fig. 4, we compare the convergence of the distributionally robust Q-learning algorithm with the classical risk-neutral version (with rain probability  $p = 0.25$ ) in terms of the absolute step sizes

$$\begin{aligned} \delta^{(t)} &= \|Q^{(t+1)} - Q^{(t)}\|_\infty \\ &= \alpha_t Y_t(z_t) \left| c(x_t, s_t, T(x_t, s_t, a_t, z_t)) + \beta \min_{a'} \max_{Y'} Q^{(t)}(T(x_t, s_t, a_t, z_t), a', Y') \right. \\ &\quad \left. - Q^{(t)}(x_t, s_t, a_t, Y_t) \right|. \end{aligned}$$

Here,  $(x_t, s_t, a_t, Y_t, z_t)$  is the state-action-disturbance combination sampled in iteration  $t$ . The plots show the moving averages

$$\Delta^{(t)} = \frac{1}{100} \sum_{k=t-99}^t \delta^{(k)}, \quad t = 99, \dots, 100000$$



**Fig. 4.** Moving average of the absolute step sizes as a function of the number of iterations.

of the step sizes both for the risk-neutral and the distributionally robust algorithm as well as the small reservoir  $\bar{s} = 3$  and a larger one with  $\bar{s} = 10$ . First, we observe that the distributionally robust algorithm performs as good as its classical counterpart. Besides, the fast convergence also holds for larger models. We can also note that the learning rate with intermediate decay combines fast convergence with the good approximation results shown above.

**Table 1.** Robust optimal policy of the decision maker in all states  $(x, s)$  with additional 1's compared to the risk-neutral case in bold print.

$x \setminus s$	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	<b>1</b>	1	1	1	1
2	0	0	0	0	0	0	<b>1</b>	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1

To illustrate the difference between the classical risk-neutral and the distributionally robust cost minimization criterion for the decision maker, Table 1 shows his optimal policy for the model with larger reservoir. Optimal actions that differ under the two optimization targets are in bold print. The two bold 1's belong to the robust case and must be zero in the risk-neutral case. In order to prevent

a spillover destroying the harvest, the more conservative decision maker in the robust model irrigates the crops already at water level 6 where a risk-neutral controller would not take action yet.

## References

1. Bäuerle, N., Glauner, A.: Distributionally robust Markov decision processes and their connection to risk measures. [arXiv:2007.13103](https://arxiv.org/abs/2007.13103) (2020)
2. Bäuerle, N., Rieder, U.: *Markov Decision Processes with Applications to Finance*. Springer, Heidelberg (2011)
3. Bellman, R.: *Dynamic Programming*. Princeton University Press, Princeton (1957)
4. Bellman, R.: *Dynamic Programming*. Dover Publications, Mineola (2003)
5. Bertsekas, D., Tsitsiklis, J.N.: *Neuro-Dynamic Programming*. Athena Scientific, Belmont (1996)
6. Glauner, A.: Robust and Risk-sensitive Markov decision processes with applications to dynamic optimal reinsurance. Ph.D. thesis, Karlsruhe Institute of Technology (2020). <https://doi.org/10.5445/IR/1000126170>
7. González-Trejo, J.I., Hernández-Lerma, O., Hoyos-Reyes, L.F.: Minimax control of discrete-time stochastic systems. *SIAM J. Control Optim.* **41**(5), 1626–1659 (2002)
8. Guidolin, M., Rinaldi, F.: Ambiguity in asset pricing and portfolio choice: a review of the literature. *Theory Decis.* **74**(2), 183–217 (2013)
9. Hernández-Lerma, O., Lasserre, J.B.: *Further Topics on Discrete-Time Markov Control Processes*. Springer, New York (1999)
10. Hinderer, K.: *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*. Springer, Heidelberg (1970)
11. Iyengar, G.N.: Robust dynamic programming. *Math. Oper. Res.* **30**(2), 257–280 (2005)
12. Jaśkiewicz, A., Nowak, A.S.: Robust Markov control processes. *J. Math. Anal. Appl.* **420**(2), 1337–1353 (2014)
13. Maccheroni, F., Marinacci, M., Rustichini, A.: Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica* **74**(6), 1447–1498 (2006)
14. McNeil, A.J., Frey, R., Embrechts, P.: *Quantitative Risk Management: Concepts, Techniques and Tools*, revised Princeton University Press, Princeton and Oxford (2015)
15. Pichler, A.: Premiums and reserves, adjusted by distortions. *Scand. Actuar. J.* **2015**(4), 332–351 (2015)
16. Rüschendorf, L.: *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios*. Springer, Heidelberg (2013)
17. Unami, K., Mohawesh, O., Sharifi, E., Takeuchi, J., Fujihara, M.: Stochastic modelling and control of rainwater harvesting systems for irrigation during dry spells. *J. Clean. Prod.* **88**, 185–195 (2015)
18. Unami, K., Yangyuoru, M., Alam, A.H.M.B., Kranjac-Berisavljevic, G.: Stochastic control of a micro-dam irrigation scheme for dry season farming. *Stoch. Environ. Res. Risk Assess.* **27**(1), 77–89 (2013)
19. Wiesemann, W., Kuhn, D., Rustem, B.: Robust Markov Decision Processes. *Math. Oper. Res.* **38**(1), 153–183 (2012)
20. Xu, H., Mannor, S.: Distributionally robust Markov decision processes. *Adv. Neural Inform. Process. Syst.* **23**, 2505–2513 (2010)



# State Estimation in Partially Observed Stochastic Networks with Queueing Applications

Konstantin V. Semenikhin<sup>(✉)</sup>

Department Probability Theory and Computer Modeling, Moscow Aviation Institute,  
Volokolamskoye shosse, 4, Moscow, Russia  
[siemenkv@mail.ru](mailto:siemenkv@mail.ru)

[https://www.researchgate.net/profile/Konstantin\\_Semenikhin](https://www.researchgate.net/profile/Konstantin_Semenikhin)

**Abstract.** The problem of filter-based state estimation for a partially observed stochastic network is considered in this paper, using the measure change approach. The network is assumed to have two types of nodes: observed and hidden. Their dynamics are defined by a set of counting processes with state-dependent intensities. The goal is to derive the nonlinear optimal filter and to propose a numerical scheme for its practical implementation. Network models that allow the optimal filter to be finite-dimensional are also considered. The theoretical results are applied to a retrial queueing system to track changes in two hidden stations: one accumulates blocked customers and the other contains unsatisfied customers.

**Keywords:** Partially observed stochastic network · Point process · Filtering · Martingale · Change of measure · Retrial queueing system

**AMS (2020) Subject Classification:** Primary 93E11 · Secondary 90B15

## 1 Introduction

First publications on stochastic filtering in queueing systems and networks were aimed at proving and enhancing classical results of the queueing theory (such as Burke's output theorem and Arrivals-See-Time-Averages properties) to a wider class of point processes, using martingale methods [5, 8, 18]. Martingale theory together with a reference probability approach has obtained numerous applications in estimation, control and optimization for stochastic systems described by jump Markov processes [6, 7, 14–16]. However in the field of queueing systems there has been little work on the applications of filtering theory [3, 4, 13, 16]. This can be explained, in part, by the opinion that rational queueing does not need complicated estimation algorithms even if dealing with strategic customers who can observe the queue length [9]. Nevertheless, recovery of unknown parameters

and hidden states based on partially observed dynamics constitutes an important class of inverse problems in the queueing theory [2]. In communication network applications, especially in wireless congestion control, filter-based estimates have recently received considerable attention to cope with time-varying behavior of packet arrival rates [12, 17]. This problem known as bandwidth estimation is formulated in the form of a nonlinear filtering problem to track changes in incoming data flows given measurements of buffer occupancy.

In this paper, we consider a Jackson-type stochastic network with observed and hidden nodes. The number of units at each hidden node is to be estimated from changes in states of the observed nodes. Instead of using the infinite-dimensional differential system for conditional probabilities, we adopt the reference probability method to derive underlying equations for the conditional expectation and covariances. Although these equations, in general, have no closed-form solution we present a particular class of the network model that provides a finite-dimensional filter. For practical implementation of the state estimation method we propose a numerical scheme based on regularization of the optimal filtering equations. To justify the estimation algorithm we consider a call center model described by the main station (a single-server finite queueing system) and two additional stations (“orbits”) whose states are to be estimated given the observed queue length at the main system.

## 2 Model Description and Problem Formulation

We study a stochastic network with the set of nodes  $S = \{1, 2, \dots, d\}$ . Each node receives units (jobs, customers and so on) from other nodes and from outside. An additional node 0 is used as a source of external arrivals or a sink in the case of service completion. Network dynamics are determined by a continuous-time process  $X(t) = (X_1(t), \dots, X_d(t))$  defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , where  $X_i(t)$  denotes the number of units at node  $i$  at time  $t \geq 0$ . Any change in the network state is caused by one of three possible single-unit movements:

- a) a unit moves from one node  $i \in S$  to another  $j \in S$ ;
- b) a unit finishes a service at node  $i \in S$ ;
- c) a unit arrives to node  $j \in S$  from outside.

These transitions are described by the respective point processes  $N_{i,j}(t)$ ,  $N_{i,0}(t)$ , and  $N_{0,j}(t)$  which have right-continuous sample paths and unit jumps. We do not consider instantaneous transitions within the same node, so the processes  $N_{i,i}$  or  $N_{0,0}$  are not used in the paper.

Assume all these processes  $\{N_{\alpha,\beta}\}$  are adapted to some right-continuous complete filtration  $\mathbf{F} = \{\mathcal{F}_t\}_{t \geq 0}$  and have the following representation:

$$N_{\alpha,\beta}(t) = \int_0^t \nu_{\alpha,\beta}(s) ds + M_{\alpha,\beta}(t), \quad (1)$$

where  $M_{\alpha,\beta}$  is a square-integrable  $\mathbf{F}$ -martingale and  $\nu_{\alpha,\beta}$  is a nonnegative  $\mathbf{F}$ -predictable function [6].

We suppose the martingales  $M_{\alpha,\beta}$  and  $M_{\alpha',\beta'}$  are orthogonal for  $(\alpha, \beta) \neq (\alpha', \beta')$ . This is equivalent to the condition that jumps of  $N_{\alpha,\beta}$  and  $N_{\alpha',\beta'}$  (i.e., any two different transitions including arrivals and departures) do not occur at the same time.

Then the state of node  $k \in S$  can be expressed as follows:

$$X_k(t) = X_k(0) + \sum_{\alpha} N_{\alpha,k}(t) - \sum_{\beta} N_{k,\beta}(t),$$

where  $\alpha$  and  $\beta$  run over  $S \cup \{0\}$ .

To complete the description of the network model, it remains to define how the transition intensities  $\nu_{\alpha,\beta}$  depend on the current state or previous evolution of the network.

For Jackson networks, given an  $\cdot/M_{\mu_i}/m_i$  queueing system at each node  $i$ , constant arrival rates  $\lambda_j$ , service rates  $\mu_i$ , and routing probabilities  $r_{i,j}$ ,  $i, j \in S$ , we obtain the transition intensities:  $\nu_{0,j} = \lambda_j$ ,  $\nu_{i,j}(t) = \mu_i(X_i(t-) \wedge m_i)r_{i,j}$ , and  $\nu_{i,0}(t) = \mu_i(X_i(t-) \wedge m_i)(1 - \sum_{j \in S} r_{i,j})$ . In the case of loss networks, there is a station  $j$  with finite capacity  $K_j$ , so the routing probabilities  $\{r_{i,j}\}_{j \in S}$  must be multiplied by the indicator  $I\{X_j(t-) < K_j\}$ . If the network is considered in a control setting, all three sets of parameters  $\{\lambda_j\}$ ,  $\{\mu_i\}$ , and  $\{r_{i,j}\}$  can be defined by access, service, and routing control policies, respectively.

In this paper, we study a partially observed stochastic network. To this end, let us split the set of nodes into two subsets:  $S = J \sqcup H$ , where  $J$  will denote the set of all observed nodes while  $H$  will contain hidden nodes of the network except for the fictitious node 0 which will also be treated as unobservable. The only information about the network evolution is given by the state of the observed nodes  $Y(t) = \{X_i(t)\}_{i \in J}$  including the initial state of the entire network  $X(0)$ . Then, write

$$\mathcal{Y}_t = \sigma\{X(0), Y(s): s \leq t\} \quad \text{and} \quad \mathcal{Y}_{t-} = \sigma\{X(0), Y(s): s < t\}$$

for complete sigma-algebras generated by the observations and  $\mathbf{Y} = \{\mathcal{Y}_t\}_{t \geq 0}$  for the corresponding filtration.

We make an additional assumption on transitions from the observed nodes: the intensities  $\nu_{i,\beta}$  must be  $\mathbf{Y}$ -predictable for all  $i \in J$  and  $\beta \in S \cup \{0\}$ . This means that we not only know the true state of nodes  $i \in J$  at each time; we also have direct information on the rate at which units move from these nodes. This condition is fulfilled for Jackson-type stochastic networks whenever service rates  $\mu_i$  and routing probabilities  $r_{i,\beta}$  are  $\mathbf{Y}$ -predictable for all observed nodes  $i \in J$ . In contrast, it does not hold for loss networks if there is a transition from one observed node  $i \in J$  to some hidden station  $k \in H$  with finite capacity.

The goal of the optimal filtering problem for the partially observed network is to find the conditional expectation  $\hat{Z}(t) = E\{Z(t) | \mathcal{Y}_t\}$  of the network's hidden part  $Z(t) = \{X_k(t)\}_{k \in H}$  given the observations available up to the current time  $t$ . Since we are going to solve this problem without finding the whole posterior distribution  $\{P\{Z(t) = z | \mathcal{Y}_t\}: z = \{x_k\}_{k \in H}\}$ , we will use the conditional covariance matrix  $Q(t) = \text{cov}\{Z(t), Z(t) | \mathcal{Y}_t\}$  to characterize the estimation accuracy.

Our aim is to determine  $\hat{X}(t)$  and  $Q(t)$  in a recursive manner which is suitable for practical implementation including approximation schemes.

This setting is motivated by optimization problems that arise in the design of queueing systems, such as contact centers. The lack of exact information about how many customers are blocked by the system or unsatisfied with the quality of service makes difficult to improve the efficiency of the system. Thus, filtered estimates of unobservable interactions can be used to tune the tradeoff between customer satisfaction and personnel-related operating costs.

Another application where partially observed stochastic networks can be useful is related to a bottleneck link problem in data transmission. Some nodes of wireless communication networks, especially over a multi-hop path, are often hidden from direct measurements of service rate and buffer occupancy, so to adequately track end-to-end throughput one needs to develop recursive algorithms for on-line estimating actual states and parameters of unobservable nodes.

### 3 Optimal Filter for a Process with Network Dynamics

We start with a simple but important remark on the observable dynamics: the filtration  $\mathbf{Y}$  can be defined as that generated only by the point processes

$$N_{i,j}, \quad N_j^a = \sum_{k \notin J} N_{k,j}, \quad \text{and} \quad N_i^d = \sum_{k \notin J} N_{i,k} \quad (i, j \in J)$$

together with the initial state  $X(0)$ . Note that  $\{N_{i,j}\}$  describe transitions inside the set of observed nodes  $J$ , while  $N_j^a$  and  $N_i^d$  count arrivals to  $j \in J$  from any unobservable node  $k \notin J$  and departures from  $i \in J$  to any  $k \notin J$ , respectively. The intensities of observed arrivals and departures are the following:

$$\nu_j^a = \sum_{k \notin J} \nu_{k,j} \quad \text{and} \quad \nu_i^d = \sum_{k \notin J} \nu_{i,k}.$$

To derive equations for the optimal filter we will use the reference probability method [6]. To this end, define a measure  $\bar{\mathbb{P}}$  on  $(\Omega, \mathcal{F})$  such that under  $\bar{\mathbb{P}}$ , all point processes  $\{N_{\alpha,\beta}\}$  are mutually independent Poisson processes with unit intensity. The measure  $\bar{\mathbb{P}}$  is called a reference probability and the corresponding expectation is denoted by  $\bar{\mathbb{E}}$ .

The lemma below shows that expectations under  $\bar{\mathbb{P}}$  are computed in an easy way. To simplify notation, we write  $\bar{\mathbb{E}}\{\xi dt \mid \mathcal{Y}_t\} = d\eta$  as shorthand for the integral equation  $\bar{\mathbb{E}}\{\int_0^t \xi(s) ds \mid \mathcal{Y}_t\} = \int_0^t d\eta(s)$  if it holds for all  $t \in (0, \infty)$ .

**Lemma 1.** *Suppose that  $\mathbf{F}$  is a complete filtration generated by all point processes  $\{N_{\alpha,\beta}\}$ . If  $\xi(t)$  is an  $\mathbf{F}$ -predictable process such that  $\int_0^t \bar{\mathbb{E}}|\xi(s)| ds < \infty$  for any  $t < \infty$ , then*

$$\bar{\mathbb{E}}\{\xi dt \mid \mathcal{Y}_t\} = \bar{\xi} dt, \tag{2}$$

$$\bar{\mathbb{E}}\{\xi dN_{i,j} \mid \mathcal{Y}_t\} = \bar{\xi} dN_{i,j}, \quad i, j \in J, \quad (3)$$

$$\bar{\mathbb{E}}\{\xi dN_{i,k} \mid \mathcal{Y}_t\} = \frac{\bar{\xi}}{p} dN_i^d, \quad i \in J, \quad k \notin J, \quad (4)$$

$$\bar{\mathbb{E}}\{\xi dN_{k,j} \mid \mathcal{Y}_t\} = \frac{\bar{\xi}}{p} dN_j^a, \quad k \notin J, \quad j \in J, \quad (5)$$

$$\bar{\mathbb{E}}\{\xi dN_{k,l} \mid \mathcal{Y}_t\} = \bar{\xi} dt, \quad k, l \notin J, \quad (6)$$

where  $\bar{\xi}(t)$  denotes a  $\mathbf{Y}$ -predictable version of  $\bar{\mathbb{E}}\{\xi(t) \mid \mathcal{Y}_{t-}\}$  and  $p$  equals the number of unobservable nodes  $H \cup \{0\}$ . Furthermore, after replacing each point process with the centered counterpart

$$\overset{\circ}{N}_{\alpha,\beta}(t) = N_{\alpha,\beta}(t) - t, \quad \overset{\circ}{N}_i^d(t) = N_i^d(t) - pt, \quad \overset{\circ}{N}_j^a(t) = N_j^a(t) - pt,$$

(3), (4), and (5) remain to be valid whereas (6) yields zero.

To return to the “real-world” model, one needs to define a probability measure  $\mathsf{P}$ , under which the stochastic network will have the original transition intensities  $\nu_{\alpha,\beta}$  given by (1). To do this, we first consider a stochastic exponential

$$d\Theta(t) = \Theta(t-) dM(t), \quad t > 0, \quad \Theta(0) = 1,$$

$$M(t) = \int_0^t \sum_{\alpha,\beta} (\nu_{\alpha,\beta}(s) - 1) d\overset{\circ}{N}_{\alpha,\beta}(s),$$

and then put

$$\mathsf{P}(A) = \bar{\mathbb{E}}\{\mathbb{I}_A \Theta(t)\}, \quad A \in \mathcal{F}_t, \quad t \geq 0. \quad (7)$$

The next lemma confirms the fact that (7) determines the original model described in Sect. 2.

**Lemma 2.** Assume  $\mathcal{F} = \sigma\left\{\bigcup_{t \geq 0} \mathcal{F}_t\right\}$  and the intensities satisfy two conditions:

$$\nu_{\alpha,\beta} > 0 \quad \text{whenever} \quad \Delta N_{\alpha,\beta} > 0; \quad (8)$$

$$\exists C = const: \quad \sum_{\alpha,\beta} \nu_{\alpha,\beta} \leq C \sum_{\alpha,\beta} N_{\alpha,\beta}. \quad (9)$$

Then

1.  $\Theta(t)$  is a positive  $\bar{\mathbb{P}}$ -martingale with  $\bar{\mathbb{E}}\Theta(t) = 1$ ;
2. (7) is a probability measure uniquely defined on  $\mathcal{F}$ ;
3. Any  $\mathbf{F}$ -adapted process  $\xi(t)$  with  $\mathbb{E}|\xi(t)| < \infty$  is a  $\mathbb{P}$ -martingale if and only if  $\xi(t)\Theta(t)$  is a  $\bar{\mathbb{P}}$ -martingale;
4. Under  $\mathbb{P}$ , conditional expectations are calculated using Bayes' rule

$$\mathbb{E}\{\xi | \mathcal{Y}_t\} = \bar{\mathbb{E}}\{\xi\Theta(t) | \mathcal{Y}_t\}/\theta(t), \quad \theta(t) = \bar{\mathbb{E}}\{\Theta(t) | \mathcal{Y}_t\}, \quad (10)$$

where  $\xi$  is a random variable such that  $\mathbb{E}|\xi| < \infty$ ;

5. Under  $\mathbb{P}$ , each point process  $N_{\alpha,\beta}$  has the martingale representations (1).

Let us consider a process  $\xi(t)$  with jumps generated by the stochastic network:

$$d\xi = \eta dt + \sum_{i \in J, k \notin J} \zeta_{i,k} dN_{i,k} + \sum_{j \in J, k \notin J} \zeta_{k,j} dN_{k,j} + \sum_{k,l \notin J} \zeta_{k,l} dN_{k,l} \quad (11)$$

where  $\eta(t)$  and  $\{\zeta_{\alpha,\beta}(t)\}$  are  $\mathbf{F}$ -predictable processes and  $\xi(0)$  is a  $\mathcal{Y}_0$ -measurable initial state. The terms related to transitions within the observable part of the network  $\{\zeta_{i,j} dN_{i,j}, i, j \in J\}$  are not used in the paper, so they are omitted in (11).

Our goal now is to obtain equations for the unnormalized estimate  $\tilde{\xi}(t)$  and the conditional expectation  $\hat{\xi}(t)$ :

$$\tilde{\xi}(t) = \bar{\mathbb{E}}\{\xi(t)\Theta(t) | \mathcal{Y}_t\} \quad \text{and} \quad \hat{\xi}(t) = \mathbb{E}\{\xi(t) | \mathcal{Y}_t\}.$$

From now on we use this notation for the estimates of any  $\mathbf{F}$ -adapted *corlol* process  $\xi$  [19]. In the case of an  $\mathbf{F}$ -predictable process, say  $\eta$ , we denote  $\mathbf{Y}$ -predictable versions of  $\bar{\mathbb{E}}\{\eta(t)\Theta(t-) | \mathcal{Y}_{t-}\}$  and  $\mathbb{E}\{\eta(t) | \mathcal{Y}_{t-}\}$  by  $\tilde{\eta}(t)$  and  $\hat{\eta}(t)$ , respectively. In the case of the product, we write  $\tilde{\xi}\eta(t)$  and  $\hat{\xi}\eta(t)$  for the corresponding estimates of the  $\mathbf{F}$ -predictable process  $\xi(t-) \eta(t)$ .

The theorem below is the main tool for deriving filtered estimates of any process governed by the network dynamics.

**Theorem 1.** *Under the assumptions of Lemmas 1 and 2, the estimates of (11) satisfy the following equations:*

$$d\hat{\xi} = \left( \hat{\eta} + \sum_{k,l \notin J} \widehat{\zeta_{k,l}\nu_{k,l}} - \sum_{j \in J} \hat{c}_j^a \right) dt + \sum_{i \in J} \frac{\hat{\zeta}_i^d}{\nu_i^d} dN_i^d + \sum_{j \in J} \frac{\hat{\zeta}_j^a + \hat{c}_j^a}{\hat{\nu}_j^a} dN_j^a, \quad (12)$$

$$\zeta_i^d = \sum_{k \notin J} \zeta_{i,k} \nu_{i,k}, \quad \zeta_j^a = \sum_{k \notin J} \zeta_{k,j} \nu_{k,j}, \quad \hat{c}_j^a = \widehat{\xi\nu_j^a} - \hat{\xi}(t-) \hat{\nu}_j^a$$

and

$$\begin{aligned} d\tilde{\xi} &= \tilde{\eta} dt + \tilde{\xi}(t-) dM' + \sum_{j \in J} (\widetilde{\xi\nu_j^a}/p - \tilde{\xi}(t-)) d\overset{\circ}{N}_j^a \\ &\quad + \frac{1}{p} \sum_{i \in J} \tilde{\zeta}_i^d dN_i^d + \frac{1}{p} \sum_{j \in J} \tilde{\zeta}_j^a dN_j^a + \sum_{k,l \notin J} \widetilde{\zeta_{k,l}\nu_{k,l}} dt, \\ dM' &= \sum_{i,j \in J} (\nu_{i,j} - 1) d\overset{\circ}{N}_{i,j} + \sum_{i \in J} (\nu_i^d/p - 1) d\overset{\circ}{N}_i^d. \end{aligned} \quad (13)$$

with initial conditions  $\hat{\xi}(0) = \tilde{\xi}(0) = \xi(0)$  and  $M'(0) = 0$ .

*Proof.* We first apply Ito's rule:

$$\begin{aligned} d(\xi\Theta) &= \Theta(t-) d\xi + \xi(t-) d\Theta + \Delta\xi\Delta\Theta \\ &= \Theta(t-) \eta dt + \sum \Theta(t-) \zeta_{\alpha,\beta} dN_{\alpha,\beta} + \sum_{i,j \in J} \xi(t-) \Theta(t-) (\nu_{i,j} - 1) d\overset{\circ}{N}_{i,j} \\ &\quad + \sum \xi(t-) \Theta(t-) (\nu_{\alpha,\beta} - 1) d\overset{\circ}{N}_{\alpha,\beta} + \sum \zeta_{\alpha,\beta} \Theta(t-) (\nu_{\alpha,\beta} - 1) dN_{\alpha,\beta} \end{aligned}$$

where all sums without subscripts are taken over  $(\alpha, \beta) \notin J \times J$ . Then, using Lemma 1, we obtain

$$\begin{aligned} d\tilde{\xi} &= \tilde{\eta} dt + \sum_{i,j \in J} \tilde{\xi}(t-) (\nu_{i,j} - 1) d\overset{\circ}{N}_{i,j} \\ &\quad + \frac{1}{p} \sum_{i \in J, k \notin J} \tilde{\xi}(t-) (\nu_{i,k} - 1) d\overset{\circ}{N}_i^d + \frac{1}{p} \sum_{j \in J, k \notin J} (\widetilde{\xi\nu_{k,j}} - \tilde{\xi}(t-)) d\overset{\circ}{N}_j^a \\ &\quad + \frac{1}{p} \sum_{i \in J, k \notin J} \tilde{\zeta}_i \nu_{i,k} dN_i^d + \frac{1}{p} \sum_{j \in J, k \notin J} \widetilde{\zeta_{k,j}\nu_{k,j}} dN_j^a + \sum_{k,l \notin J} \widetilde{\zeta_{k,l}\nu_{k,l}} dt \end{aligned}$$

which coincides with (13).

In particular, we can now write the equation for  $\theta(t) = \bar{E}\{\Theta(t) | \mathcal{Y}_t\}$ :

$$d\theta = \theta(t-) dM' + \sum_{j \in J} (\tilde{\nu}_j^a/p - \theta(t-)) d\overset{\circ}{N}_j^a. \quad (14)$$

To derive the estimate  $\hat{\xi}(t) = \tilde{\xi}(t)/\theta(t)$ , we use the expression

$$d\hat{\xi} = \frac{d\tilde{\xi}}{\theta(t-)} - \frac{\hat{\xi}(t-) d\theta}{\theta(t-)} - \frac{\Delta\hat{\xi}\Delta\theta}{\theta(t-)}. \quad (15)$$

Then, from (13) and (14) it follows that

$$\begin{aligned} \frac{d\tilde{\xi}}{\theta(t-)} - \frac{\hat{\xi}(t-) d\theta}{\theta(t-)} &= \hat{\eta} dt + \frac{1}{p} \sum_{j \in J} (\widehat{\xi\nu_j^a} - \hat{\xi}(t-)\hat{\nu}_j^a) d\overset{\circ}{N}_j^a \\ &\quad + \frac{1}{p} \sum_{i \in J} \hat{\zeta}_i^d dN_i^d + \frac{1}{p} \sum_{j \in J} \hat{\zeta}_j^a dN_j^a + \sum_{k,l \notin J} \widetilde{\zeta_{k,l}\nu_{k,l}} dt \end{aligned} \quad (16)$$

The last term in (15) may be nonzero only in three cases:

- a)  $\Delta N_{i,j} \neq 0$  ( $i, j \in J$ ),      b)  $\Delta N_i^d \neq 0$  ( $i \in J$ ),      c)  $\Delta N_j^a \neq 0$  ( $j \in J$ ).

Then,  $\Delta \tilde{\xi} = \xi(t-) - g$  and  $\Delta \theta = \theta(t-) - h$ , where

$$\text{a) } g = a, \quad b = 0, \quad \text{b) } g = a = \frac{\nu_i^d}{p}, \quad b = \frac{\tilde{\xi}_i^d}{p}, \quad \text{c) } g = \frac{\tilde{\xi} \nu_j^a}{p \tilde{\xi}(t-)}, \quad a = \frac{\hat{\nu}_j^a}{p}, \quad b = \frac{\tilde{\xi}_j^a}{p}.$$

It is easy to show that

$$\frac{\Delta \hat{\xi} \Delta \theta}{\theta(t-)} = (a - 1) \left( \hat{\xi}(t-) \left( \frac{g}{a} - 1 \right) + \frac{b}{a \theta(t-)} \right).$$

This yields zero in case a) and

$$\frac{\Delta \hat{\xi} \Delta \theta}{\theta(t-)} = \begin{cases} \left( \frac{1}{p} - \frac{1}{\nu_i^d} \right) \hat{\zeta}_i^d & \text{in case b),} \\ \left( \frac{1}{p} - \frac{1}{\hat{\nu}_j^a} \right) \left( \tilde{\xi} \nu_j^a - \hat{\xi}(t-) \hat{\nu}_j^a + \hat{\zeta}_j^a \right) & \text{in case c).} \end{cases} \quad (17)$$

Subtracting (17) from (16), we obtain (12).  $\square$

*Remark 1.* The structure of the estimate (12) can be explained as follows. The unobservable dynamics  $\{\zeta_{k,l} dN_{k,l}, k, l \notin J\}$  affect only the drift coefficient of the filter. The term  $\hat{\zeta}_i^d / \nu_i^d$  defines the average effect of jumps  $\{\zeta_{i,k}\}_{k \notin J}$  caused by transitions from node  $i$  to the unobservable part of the network. Analogously,  $\hat{\zeta}_j^a / \hat{\nu}_j^a$  is a mixture of terms  $\{\zeta_{k,j}\}_{k \notin J}$  related to transitions from unobservable nodes to station  $j$ . The only difference between these two types of transitions is the correction term

$$\hat{c}_j^a = \text{cov}\{\xi(t-), \nu_j^a | \mathcal{Y}_{t-}\} \quad (18)$$

which is added to the coefficient of jump  $dN_j^a$  and subtracted from the drift.

## 4 State Estimation for Hidden Nodes

In this section, we focus on deriving a filtering algorithm for state estimation of unobservable nodes in the stochastic network.

For any node  $k \in H$ , its state can be represented in the form of (11):

$$dX_k = \sum_{i \in J} dN_{i,k} - \sum_{j \in J} dN_{k,j} + \sum_{m \notin J} (dN_{m,k} - dN_{k,m}).$$

From Theorem 1, we have immediately

$$d\hat{X}_k = \left\{ \hat{\nu}_k^a - \hat{\nu}_k^d - \sum_{j \in J} \hat{c}_{k,j} \right\} dt + \sum_{i \in J} \xi_{i,k}^d dN_i^d + \sum_{j \in J} \xi_{k,j}^a dN_j^a, \quad (19)$$

$$\hat{\nu}_k^a = \sum_{m \notin J} \hat{\nu}_{m,k}, \quad \hat{\nu}_k^d = \sum_{m \notin J} \hat{\nu}_{k,m}, \quad (20)$$

$$\xi_{i,k}^d = \frac{\nu_{i,k}}{\nu_i^d}, \quad \xi_{k,j}^a = \frac{\hat{c}_{k,j} - \hat{\nu}_{k,j}}{\hat{\nu}_j^a}, \quad (21)$$

where the coefficients  $\{\hat{c}_{k,j}\}$  are analogous to (18):

$$\hat{c}_{k,j} = \text{cov}\{X_k(t-), \nu_j^a \mid \mathcal{Y}_{t-}\} = \widehat{X_k \nu_j^a} - \hat{X}(t-) \hat{\nu}_j^a. \quad (22)$$

Since by assumption  $\{\nu_{i,k}\}$  are  $\mathbf{Y}$ -predictable for any observed node  $i$ , the terms  $\{\xi_{i,k}^d\}$  related to departures from  $i \in J$  do not require to be estimated.

In addition to the estimates  $\{\hat{X}_k(t)\}_{k \in H}$ , we also describe their errors

$$\varepsilon_k(t) = X_k(t) - \hat{X}_k(t)$$

using conditional variances and covariances.

**Theorem 2.** *Under the conditions of Lemmas 1 and 2, the following statements hold:*

1. For  $k \in H$ , the estimate  $\hat{X}_k(t) = \mathbb{E}\{X_k(t) \mid \mathcal{Y}_t\}$  satisfies (19);
2. For  $k \in H$ , the conditional error variance  $Q_{k,k}(t) = \mathbb{E}\{\varepsilon^2(t) \mid \mathcal{Y}_t\}$  has the form

$$\begin{aligned} dQ_{k,k} &= \left( \hat{\nu}_k^a + \hat{\nu}_k^d + 2\hat{b}_{k,k} - \sum_{j \in J} \hat{\tau}_{k,k,j} \right) dt + \sum_{i \in J} (1 - \xi_{i,k}^d) \xi_{i,k}^d dN_i^d \\ &\quad + \sum_{j \in J} \left\{ \frac{1}{\hat{\nu}_j^a} (\hat{\tau}_{k,k,j} + \hat{\nu}_{k,j} - 2\hat{\varkappa}_{k,k,j}) - (\xi_{k,j}^a)^2 \right\} dN_j^a; \end{aligned} \quad (23)$$

3. For  $k, l \in H$  such that  $k \neq l$ , the conditional error covariance  $Q_{k,l}(t) = \mathbb{E}\{\varepsilon_k(t)\varepsilon_l(t) \mid \mathcal{Y}_t\}$  is given by the equation

$$\begin{aligned} dQ_{k,l} &= (\hat{b}_{k,l} + \hat{b}_{l,k} - \hat{\nu}_{k,l} - \hat{\nu}_{l,k} - \sum_{j \in J} \hat{\tau}_{k,l,j}) dt - \sum_{i \in J} \xi_{i,k}^d \xi_{i,l}^d dN_i^d \\ &\quad + \sum_{j \in J} \left\{ \frac{1}{\hat{\nu}_j^a} (\hat{\tau}_{k,l,j} - \hat{\varkappa}_{k,l,j} - \hat{\varkappa}_{l,k,j}) - \xi_{k,j}^a \xi_{l,j}^a \right\} dN_j^a. \end{aligned} \quad (24)$$

The above coefficients  $\hat{\varkappa}_{k,l,j}$ ,  $\hat{b}_{k,l}$ , and  $\hat{\tau}_{k,l,j}$  are  $\mathbf{Y}$ -predictable versions of the conditional covariances:

$$\hat{\varkappa}_{k,l,j} = \text{cov}\{X_k(t-), \nu_{l,j} \mid \mathcal{Y}_{t-}\}, \quad (25)$$

$$\hat{b}_{k,l} = \text{cov}\{X_k(t-), \nu_l^a - \nu_l^d \mid \mathcal{Y}_{t-}\} = \sum_{m \notin J} (\hat{\varkappa}_{k,m,l} - \hat{\varkappa}_{k,l,m}), \quad (26)$$

$$\hat{\tau}_{k,l,j} = \text{cov}\{\varepsilon_k(t-) \varepsilon_l(t-), \nu_j^a \mid \mathcal{Y}_{t-}\}. \quad (27)$$

*Proof.* We start by representing the estimation error in the form (11):

$$\begin{aligned} d\varepsilon_k = dX_k - d\hat{X}_k &= \eta_k dt + \sum_{m \notin J} (dN_{m,k} - dN_{k,m}) \\ &+ \sum_{i \in J} \sum_{m \notin J} (\delta_{m,k} - \xi_{i,k}^d) dN_{i,m} - \sum_{j \in J} \sum_{m \notin J} (\delta_{m,k} + \xi_{k,j}^a) dN_{m,j}, \end{aligned}$$

where  $\eta_k$  is some  $\mathbf{Y}$ -predictable coefficient and  $\delta_{m,k}$  is a Kronecker's symbol.

Given any  $k, l \in H$ , we apply Ito's product formula

$$\begin{aligned} d(\varepsilon_k \varepsilon_l) &= \varepsilon_k(t-) d\varepsilon_l + \varepsilon_l(t-) d\varepsilon_k + \Delta \varepsilon_k \Delta \varepsilon_l \\ &= (\varepsilon_k(t-) \eta_l + \varepsilon_l(t-) \eta_k) dt + \Delta \varepsilon_k \Delta \varepsilon_l + \varepsilon_l(t-) \Delta \varepsilon_k + \varepsilon_k(t-) \Delta \varepsilon_l. \end{aligned}$$

Since  $\widehat{\varepsilon_k \eta_l} = \widehat{\varepsilon_k}(t-) \eta_l = 0$ , the drift term in  $dQ_{k,l}$  can be omitted. So we are interested in calculating only the discontinuous component  $\Delta(\varepsilon_k \varepsilon_l)$ . It consists of three parts. The first is related to completely unobservable jumps:

$$\begin{aligned} &\delta_{k,l} \sum_{m \notin J} (dN_{m,k} + dN_{k,m}) - (1 - \delta_{k,l})(dN_{k,l} + dN_{l,k}) \\ &+ \sum_{m \notin J} \{ \varepsilon_l(t-) (dN_{m,k} - dN_{k,m}) + \varepsilon_k(t-) (dN_{m,l} - dN_{l,m}) \}. \end{aligned}$$

The second part is a sum of  $\{dN_{i,m}, i \in J, m \notin J\}$  with the coefficients:

$$(\delta_{m,k} - \xi_{i,k}^d)(\delta_{m,l} - \xi_{i,l}^d) + \varepsilon_l(t-) (\delta_{m,k} - \xi_{i,k}^d) + \varepsilon_k(t-) (\delta_{m,l} - \xi_{i,l}^d).$$

The third part contains  $\{dN_{m,j}, j \in J, m \notin J\}$  with the coefficients:

$$(\delta_{m,k} + \xi_{k,j}^a)(\delta_{m,l} + \xi_{l,j}^a) - \varepsilon_l(t-) (\delta_{m,k} + \xi_{k,j}^a) - \varepsilon_k(t-) (\delta_{m,l} + \xi_{l,j}^a).$$

From Theorem 1 it follows that the first two parts can be estimated separately. The estimate of the first part is

$$\begin{aligned} &\left\{ \delta_{k,l} \sum_{m \notin J} (\hat{\nu}_{m,k} + \hat{\nu}_{k,m}) - (1 - \delta_{k,l})(\hat{\nu}_{k,l} + \hat{\nu}_{l,k}) \right. \\ &\left. + \sum_{m \notin J} (\hat{\varkappa}_{l,m,k} - \hat{\varkappa}_{l,k,m} + \hat{\varkappa}_{k,m,l} - \hat{\varkappa}_{k,l,m}) \right\} dt \end{aligned}$$

which coincides with the drift of (23) and (24) except the correction term  $\sum_{j \in J} \hat{\tau}_{k,l,j}$ .

Using  $\widehat{\varepsilon_l \nu_{i,m}} = \widehat{\varepsilon_l}(t-) \nu_{i,m} = 0$ , we obtain the estimate of the second part

$$\sum_{i \in J} \sum_{m \notin J} (\delta_{m,k} - \xi_{i,k}^d)(\delta_{m,l} - \xi_{i,l}^d) \frac{\nu_{i,m}}{\nu_i^d} dN_i^d. \quad (28)$$

Taking into account the correction term, the estimate of the third part takes the form:

$$\begin{aligned} & - \sum_{j \in J} \hat{\tau}_{k,l,j} dt + \sum_{j \in J} \frac{1}{\hat{\nu}_a} \left[ \hat{\tau}_{k,l,j} + \sum_{m \notin J} \{ (\delta_{m,k} + \xi_{k,j}^a) (\delta_{m,l} + \xi_{l,j}^a) \hat{\nu}_m, j \right. \\ & \quad \left. - (\delta_{m,k} + \xi_{k,j}^a) \hat{\varkappa}_{l,m,j} - (\delta_{m,l} + \xi_{l,j}^a) \hat{\varkappa}_{k,m,j} \} \right] dN_j^a. \end{aligned} \quad (29)$$

Simple calculations show that (28) and (29) yield the corresponding terms in (23) and (24).  $\square$

The following proposition describes a class of stochastic networks that admit a closed-form optimal filter for state estimates of hidden nodes.

**Corollary 1.** *If a) transitions from hidden to observed nodes have  $\mathbf{Y}$ -predictable intensities  $\{\nu_{m,j}, m \notin J, j \in J\}$ ; b) transitions within the unobservable part of the network are linear functions of the states:*

$$\nu_{m,n} = \mu_{m,n,0} + \sum_{\alpha \in H} \mu_{m,n,\alpha} X_{\alpha}(t-) \quad (m, n \notin J)$$

with  $\mathbf{Y}$ -predictable coefficients  $\{\mu_{m,n,\alpha}\}$ , then the optimal estimates  $\{\hat{X}_k\}_{k \in H}$  are described by a finite-dimensional filter:

$$d\hat{X}_k = (\hat{\nu}_k^a - \hat{\nu}_k^d) dt + \sum_{i \in J} \frac{\nu_{i,k}}{\nu_i^d} dN_i^d - \sum_{j \in J} \frac{\nu_{k,j}}{\nu_j^a} dN_j^a.$$

Furthermore, taking into account

$$\begin{aligned} \hat{\tau}_{k,l,j} &= \hat{\varkappa}_{k,l,j} = 0 \quad (k, l \in H, j \in J) \\ \hat{\varkappa}_{k,m,n} &= \sum_{\alpha \in H} \mu_{m,n,\alpha} Q_{k,\alpha}(t-) \quad (k \in H, m, n \notin J) \end{aligned}$$

the conditional error covariance matrix  $\{Q_{k,l}\}_{k,l \in H}$  satisfies the closed-form system (23), (24).

Assumption a) and b) look rather restrictive in view of applications to queueing models. Even if we have a simple tandem system  $M_{\lambda}|M_{\mu_1}|m_1 \rightarrow \cdot |M_{\mu_1}|m_2$  where station 1 is to be estimated given the observed state of station 2, both conditions a) and b) are violated.

So we need an approximation scheme to practically implement filtering equations derived above. To this end, we consider a stochastic network that has Jackson-like transition intensities at least for hidden nodes:

$$\nu_{k,\beta} = \mu_{k,\beta} X_k(t-), \quad k \in H, \quad \nu_{0,\beta} = \lambda_{\beta}$$

where  $\mu_{k,\beta}$  and  $\lambda_{\beta}$  are  $\mathbf{Y}$ -predictable coefficients.

For such a network, we have  $\hat{c}_{k,j} = \sum_{m \in H} Q_{k,m}(t-) \mu_{m,j}$ ,  $\hat{\nu}_{k,\beta} = \mu_{k,\beta} \hat{X}_k(t-)$ . So all coefficients of the optimal filter (19) can be expressed in terms of the state estimates  $\{\hat{X}_k\}_{k \in H}$  and conditional covariances  $\{Q_{k,l}\}_{k,l \in H}$ .

To simplify equations for  $\{Q_{k,l}\}$ , we propose to exclude third-order terms

$$M_{k,l}^\tau(t) = \int_0^t \sum_{j \in J} \hat{\tau}_{k,l,j} \left( \frac{dN_j^a}{\hat{\nu}_j^a} - ds \right).$$

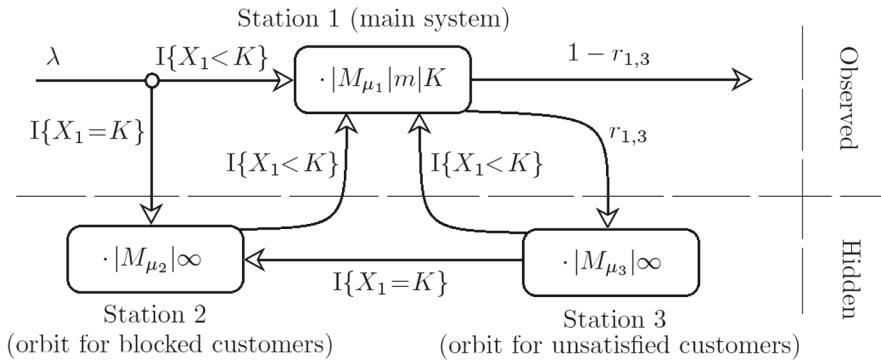
Since  $\{M_{k,l}^\tau\}$  are zero-mean martingales, this operation can be considered as a projection. Other coefficients of (23) and (24) are represented via the state estimates and error covariances (e.g.,  $\hat{\gamma}_{k,l,\beta} = Q_{k,l}(t-\mu_{l,\beta})$ ).

So after this simplification we obtain a closed-form counterpart of the system (19), (23), (24). Between jump times of  $\{N_i^d\}$  and  $\{N_j^a\}$ , it is described by the system of linear ordinary differential equations:

$$\begin{aligned} \dot{\hat{Z}} &= \Lambda^\top \hat{Z} + \lambda - Q\gamma, \\ \dot{Q} &= (Q - \text{diag}[\hat{Z}])\Lambda + \Lambda^\top (Q - \text{diag}[\hat{Z}]) + \text{diag}[\Lambda^\top \hat{Z} + \lambda], \end{aligned}$$

where the column vector  $\hat{Z}$  and the matrix  $Q$  are approximations of the state estimate and conditional error covariance, respectively;  $\lambda = \{\lambda_k\}_{k \in H}$  and  $\gamma = \{\gamma_k\}_{k \in H}$  are column vectors and  $\Lambda = \{\lambda_{k,l}\}_{k,l \in H}$  is a square matrix such that

$$\gamma_k = \sum_{j \in J} \mu_{k,j}, \quad \lambda_{k,l} = \mu_{k,l} - \delta_{k,l} \sum_{m \notin J} \mu_{k,m}.$$



**Fig. 1.** Retrial queueing system as a stochastic network.

## 5 State Estimation in a Retrial Queueing System

In this section, we study a partially observed network model of inbound call centers.

Figure 1 depicts a call center model in the form of a queueing network with three stations. Station 1 is the main queueing system providing service for incoming customers by  $m$  independent agents. For each agent, the processing time

is exponentially distributed with mean  $1/\mu_1$ . Customers arrive at the system according to a Poisson stream with rate  $\lambda$ . The maximum number of customers in the system is finite and denoted by  $K$ .

Station 2 contains blocked customers: they are not served in the main system because all agents are busy, so they try to call again after a random time exponentially distributed with mean  $1/\mu_2$ .

Station 3 includes unsatisfied customers: after being served, they try to call again to get additional information or extra service from the agents; such retrials occur after a random delay exponentially distributed with mean  $1/\mu_3$ . The probability that a customer will remain unsatisfied with the service is  $r_{1,3}$ . If the main system is busy, unsatisfied customers join station 2.

The initial state is assumed to be zero for all stations of the network.

Since the network belongs to the class of retrial queueing systems [1], we refer to stations 2 and 3 as the orbits. The number of customers in the both orbits are not observed directly; rather the state of the main system is known exactly at each time.

Our goal is to apply the filtering scheme designed above to state estimation for two unobservable stations given the on-line information on the main queueing system.

The queueing network we study has one observed node  $J = \{1\}$  and two hidden nodes  $H = \{2, 3\}$ . The number of customers at node  $i$  is denoted by  $X_i$  ( $i = 1, 2, 3$ ). The transition intensities are as follows:

$$\begin{aligned} \nu_{0,1} &= \lambda(1 - \beta), & \nu_{0,2} &= \lambda\beta, \\ \nu_{1,0} &= \mu_1(1 - r_{1,3})(X_1(t-) \wedge m), & \nu_{1,3} &= \mu_1 r_{1,3}(X_1(t-) \wedge m), \\ \nu_{2,1} &= \mu_2 X_2(t-)(1 - \beta), \\ \nu_{3,1} &= \mu_3 X_3(t-)(1 - \beta), & \nu_{3,2} &= \mu_3 X_3(t-) \beta, \end{aligned}$$

where  $\beta(t) = \mathbb{I}\{X_1(t-) = K\}$ .

Using point processes  $\{N_{i,j}\}$ , we can write the state dynamics

$$\begin{aligned} dX_1 &= dN_{0,1} + dN_{2,1} + dN_{3,1} - (dN_{1,0} + dN_{1,3}), \\ dX_2 &= dN_{0,2} + dN_{3,2} - dN_{2,1}, \\ dX_3 &= dN_{1,3} - (dN_{3,1} + dN_{3,2}). \end{aligned}$$

We have the two observed point processes with the corresponding intensities:

$$\begin{aligned} N_1^a &= N_{0,1} + N_{2,1} + N_{3,1}, & N_1^d &= N_{1,0} + N_{1,3}, \\ \nu_1^a &= (\lambda + \mu_2 X_2(t-) + \mu_3 X_3(t-))(1 - \beta), & \nu_1^d &= \mu_1(X_1(t-) \wedge m). \end{aligned}$$

Due to (20), (21), and (22), the coefficients of the state estimates  $\hat{X}_2, \hat{X}_3$  take the form:

$$\begin{aligned}\hat{c}_{k,1} &= \text{cov}\{X_k(t-), \nu_1^a | \mathcal{Y}_{t-}\} = (Q_{k,2}(t-) \mu_2 + Q_{k,3}(t-) \mu_3)(1 - \beta), \quad k = 2, 3, \\ \hat{\nu}_2^a &= \hat{\nu}_{0,2} + \hat{\nu}_{3,2} = (\lambda + \mu_3 \hat{X}_3(t-)) \beta, \quad \hat{\nu}_3^d = \hat{\nu}_{3,2} = \mu_3 \hat{X}_3(t-) \beta, \quad \hat{\nu}_2^d = \hat{\nu}_3^a = 0, \\ \xi_{k,1}^a &= \frac{Q_{k,2}(t-) \mu_2 + Q_{k,3}(t-) \mu_3 - \hat{X}_k(t-) \mu_k}{\lambda + \mu_2 \hat{X}_2(t-) + \mu_3 \hat{X}_3(t-)}, \quad \xi_{1,2}^d = 0, \quad \xi_{1,3}^d = r_{1,3}.\end{aligned}$$

From (25) and (26) we obtain the coefficients of equations for the error covariances  $\{Q_{k,l}\}$ :

$$\begin{aligned}\hat{\varkappa}_{k,l,1} &= \text{cov}\{X_k(t-), \nu_{l,1} | \mathcal{Y}_{t-}\} = Q_{k,l}(t-) \mu_l (1 - \beta), \quad k, l = 2, 3, \\ \hat{b}_{2,2} &= \text{cov}\{X_2(t-), \nu_2^a | \mathcal{Y}_{t-}\} = Q_{2,3}(t-) \mu_3 \beta, \\ \hat{b}_{3,3} &= \text{cov}\{X_3(t-), -\nu_3^d | \mathcal{Y}_{t-}\} = -Q_{3,3}(t-) \mu_3 \beta, \\ \hat{b}_{3,2} + \hat{b}_{2,3} &= (Q_{3,3}(t-) - Q_{2,3}(t-)) \mu_3 \beta.\end{aligned}$$

Now we are ready to present the filtering equations:

$$\begin{aligned}d\hat{X}_2 &= ((\lambda + \mu_3 \hat{X}_3) \beta - (Q_{2,2} \mu_2 + Q_{2,3} \mu_3)(1 - \beta)) dt + \xi_{2,1}^a dN_1^a, \\ d\hat{X}_3 &= (-\mu_3 \hat{X}_3 \beta - (Q_{3,2} \mu_2 + Q_{3,3} \mu_3)(1 - \beta)) dt + \xi_{3,1}^a dN_1^a + r_{1,3} dN_1^d, \\ dQ_{2,2} &= \mu_3 (\lambda/\mu_3 + \hat{X}_3 + 2Q_{2,3}) \beta dt + \left\{ (\hat{\nu}_{2,1} - 2\hat{\varkappa}_{2,2,1})/\hat{\nu}_1^a - (\xi_{2,1}^a)^2 \right\} dN_1^a, \\ dQ_{3,3} &= \mu_3 (\hat{X}_3 - 2Q_{3,3}) \beta dt + \left\{ (\hat{\nu}_{3,1} - 2\hat{\varkappa}_{3,3,1})/\hat{\nu}_1^a - (\xi_{3,1}^a)^2 \right\} dN_1^a \\ &\quad + (1 - r_{1,3}) r_{1,3} dN_1^d, \\ dQ_{2,3} &= \mu_3 (Q_{3,3} - Q_{2,3} - \hat{X}_3) \beta dt - \left\{ (\hat{\varkappa}_{2,3,1} + \hat{\varkappa}_{3,2,1})/\hat{\nu}_1^a + \xi_{2,1}^a \xi_{3,1}^a \right\} dN_1^a.\end{aligned}$$

These equations will be referred to as the suboptimal filter (SF).

It is worth noting that just before the jump  $\Delta N_1^a > 0$  the main system has a vacant place, so that  $\beta = 0$  in all terms related to  $dN_1^a$ . In contrast, each error (co)variance  $Q_{k,l}$  has a non-zero drift only if the main system is full, i.e.  $\beta = 1$ .

To provide a comparative analysis of the estimation accuracy, we also propose two additional filtering schemes. The first is called the truncated filter (TF) because it is obtained by truncation of the filtering equations, specifically, by letting  $\hat{c}_{k,j} = 0$  in (19). The TF estimates denoted by  $\{\check{X}_k\}$  are described by the following equations:

$$\begin{aligned}d\check{X}_2 &= (\lambda + \mu_3 \check{X}_3) \beta dt - \frac{\check{X}_2(t-) \mu_2}{\lambda + \mu_2 \check{X}_2(t-) + \mu_3 \check{X}_3(t-)} dN_1^a, \\ d\check{X}_3 &= -\mu_3 \check{X}_3 \beta dt - \frac{\check{X}_3(t-) \mu_3}{\lambda + \mu_2 \check{X}_2(t-) + \mu_3 \check{X}_3(t-)} dN_1^a + r_{1,3} dN_1^d.\end{aligned}$$

The second filter used for comparison is the drift-based filter (DF). The DF estimates are denoted by  $\{\bar{X}_k\}$ . To define them, we replace each point process

$dN_{k,l}$  in the dynamics of  $dX_k$  with the drift term  $\bar{\nu}_{k,l} dt$ . So we obtain a system of linear ODEs:

$$\begin{aligned}\dot{\bar{X}}_2 &= (\lambda + \mu_3 \bar{X}_3)\beta - \mu_2 \bar{X}_2(1 - \beta), \\ \dot{\bar{X}}_3 &= \mu_1 r_{1,3}(X_1(t-) \wedge m) - \mu_3 \bar{X}_3.\end{aligned}$$

For numerical experiments we choose the following parameters:

$$m = 20, \quad K = 25, \quad \lambda = 21, \quad \mu_1 = 2, \quad \mu_2 = 10.5, \quad \mu_3 = 4.2, \quad r_{1,3} = 0.45.$$

We take  $\lambda$  less than but close to  $\mu_1(1 - r_{1,3})m$  in order for the main system to be near the loaded state. In this case, customers are blocked more frequently but the load of station 2 behaves stable.

Table 1 contains root-mean square errors (RMSEs) obtained in one experiment. The estimation accuracy has been evaluated on 10 time intervals (with 1000 jumps of the network process in each interval). Figure 2 shows sample paths of the states and suboptimal estimates on two time intervals.

**Table 1.** Estimation errors over several segments along one sample path

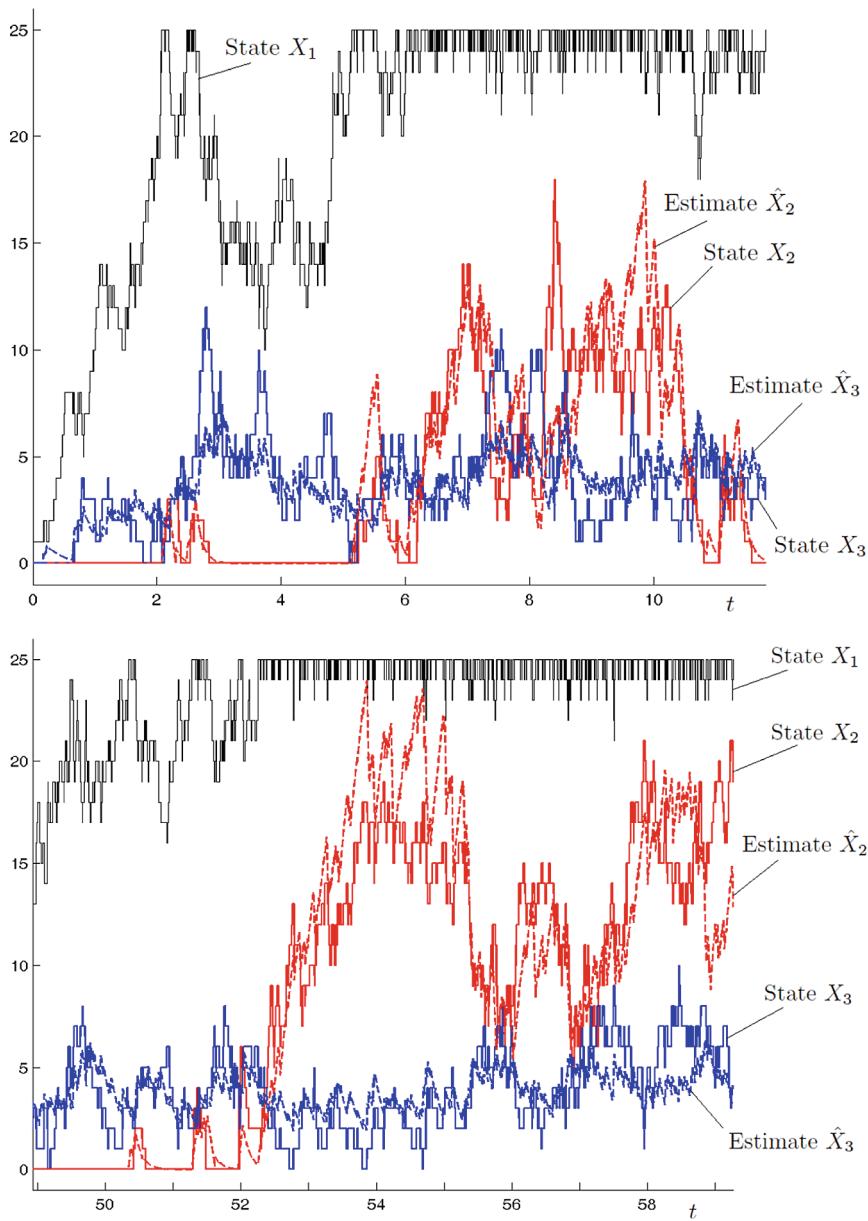
Segment:	1	2	3	4	5	6	7	8	9	10	Total
RMSE of $\check{X}_2$ :	3.976	0.733	1.337	5.107	8.097	9.849	0.167	2.310	2.401	3.497	4.631
RMSE of $\bar{X}_2$ :	2.396	1.043	1.617	1.891	3.806	2.815	0.201	1.424	2.465	2.267	2.145
RMSE of $\hat{X}_2$ :	2.145	0.840	1.305	1.933	2.755	2.763	0.188	1.191	2.118	2.193	1.867
RMSE of $\check{X}_3$ :	1.939	1.830	1.648	1.772	1.926	2.283	2.042	1.949	1.977	1.935	1.934
RMSE of $\bar{X}_3$ :	2.104	1.716	1.975	1.906	1.973	2.112	1.986	2.166	1.691	1.959	1.962
RMSE of $\hat{X}_3$ :	1.692	1.609	1.771	1.569	1.489	1.830	1.663	1.947	1.555	1.852	1.705

Our experiment shows the superiority of the suboptimal scheme over two other filtering algorithms. However it should be noted that the drift-based scheme demonstrates relatively close results: its RMSE ranges within 10–15% in comparison with the suboptimal filter for both hidden stations. In contrast, the truncated scheme turns to be much worse in estimating the number of blocked customers.

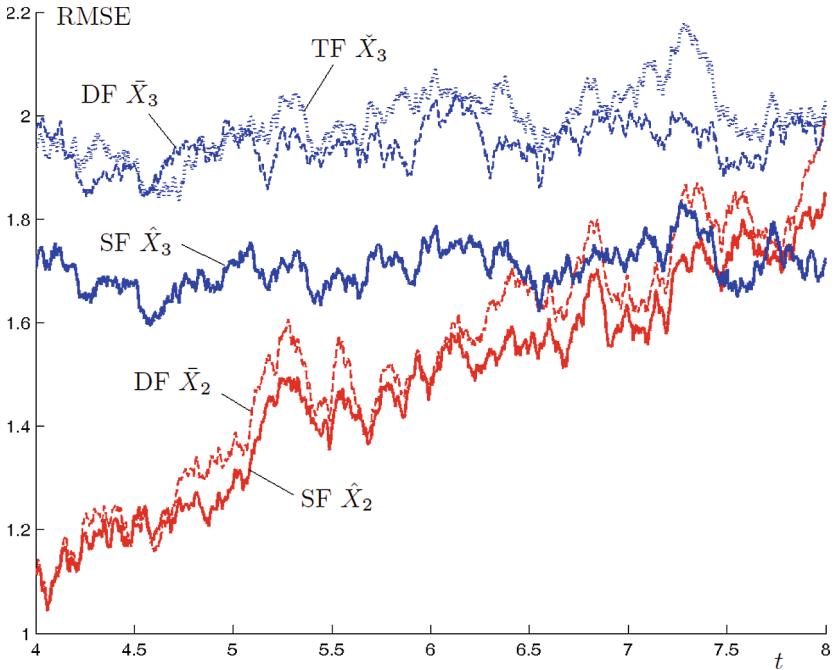
Figure 3 depicts RMSE trajectories evaluated on the basis of 1000 Monte Carlo runs. Basically, this experiment confirms the results obtained along one sample path, though the accuracy of SF and DF estimates become more similar for station 2 before achieving the steady state mode.

## 6 Appendix

*Proof of Lemma 1.* Due to the monotone class theorem and the dominated-convergence theorem, it suffices to consider an  $\mathbf{F}$ -predictable step process



**Fig. 2.** Sample paths of states (shown as solid lines) and SF-estimates (shown as dashed lines).



**Fig. 3.** RMSE of three filters SF (solid), DF (dashed), and TF (dotted) for two states  $X_2$  (red) and  $X_3$  (blue).

$\xi(s) = U \mathbf{I}_{(t_1, t_2]}(s)$ , where  $U$  is a bounded  $\mathcal{F}_{t_1}$ -measurable random variable. Then,

$$\bar{\mathbb{E}}\left\{\int_0^t \xi dN_{\alpha, \beta} \mid \mathcal{Y}_t\right\} = \bar{\mathbb{E}}\{U(N_{\alpha, \beta}(t_2) - N_{\alpha, \beta}(t_1)) \mid \mathcal{Y}_t\} \quad \forall t \geq t_2. \quad (30)$$

We introduce two  $\sigma$ -algebras  $\mathcal{F}_{t_1, t_2}$  and  $\mathcal{Y}_{t_1, t_2}$ . Both of them are generated by the increments  $\{N(s) - N(t_1) : t_1 \leq s \leq t_2\}$ , where for  $\mathcal{F}_{t_1, t_2}$ ,  $N$  is any of  $N_{\alpha, \beta}$ , whereas for  $\mathcal{Y}_{t_1, t_2}$ ,  $N$  is any observed process  $N_{i,j}$ ,  $N_i^d$ , or  $N_j^a$  ( $i, j \in J$ ).

It is important that  $\mathcal{F}_{t_1}$  and  $\mathcal{F}_{t_1, t_2}$  are independent under  $\bar{\mathbb{P}}$ .

Note that  $\mathcal{Y}_{s-}$  is generated by events  $AB$  such that  $A \in \mathcal{Y}_{t_1}$  and  $B \in \mathcal{Y}_{t_1, s'}$  for some  $s' \in (t_1, s)$ . Since  $A, B$  are independent, we have

$$\bar{\mathbb{E}}\{\bar{\mathbb{E}}\{U \mid \mathcal{Y}_{t_1}\} \mathbf{I}_{AB}\} = \bar{\mathbb{E}}\{\bar{\mathbb{E}}\{U \mid \mathcal{Y}_{t_1}\} \mathbf{I}_A\} \bar{\mathbb{E}}\{\mathbf{I}_B\} = \bar{\mathbb{E}}\{U \mathbf{I}_A\} \bar{\mathbb{E}}\{\mathbf{I}_B\} = \bar{\mathbb{E}}\{U \mathbf{I}_{AB}\}$$

and hence

$$\bar{\mathbb{E}}\{U \mid \mathcal{Y}_{t_1}\} = \bar{\mathbb{E}}\{U \mid \mathcal{Y}_{s-}\} \quad \forall s > t_1.$$

Therefore, the right-hand side of the integral equalities in (6) is

$$\bar{\mathbb{E}}\{U \mid \mathcal{Y}_{t_1}\} (t_2 - t_1). \quad (31)$$

For the right-hand side of (3) and (4)–(5) we have

$$\bar{E}\{U | \mathcal{Y}_{t_1}\} (N_{i,j}(t_2) - N_{i,j}(t_1)) \quad \text{and} \quad \bar{E}\{U | \mathcal{Y}_{t_1}\} (N(t_2) - N(t_1))/p, \quad (32)$$

respectively, where  $N$  stands for  $N_i^d$  or  $N_j^a$ .

To prove that (30) equals (31) or (32), we consider an event  $AB \in \mathcal{Y}_t$  such that  $A \in \mathcal{Y}_{t_1}$  and  $B \in \mathcal{Y}_{t_1,t}$ .

The random variable  $D_{k,l} = N_{k,l}(t_2) - N_{k,l}(t_1)$  ( $k, l \notin J$ ) and  $\sigma$ -algebras  $\mathcal{F}_{t_1}$  and  $\mathcal{Y}_{t_1,t}$  are mutually independent. This implies

$$\begin{aligned} \bar{E}\{UD_{k,l} I_{AB}\} &= \bar{E}\{U I_A\} \bar{E}\{I_B\} \bar{E}\{D_{k,l}\} \\ &= \bar{E}\{\bar{E}(U | \mathcal{Y}_{t_1}) I_A\} \bar{E}\{I_B\} (t_2 - t_1) = \bar{E}\{\bar{E}(U | \mathcal{Y}_{t_1})(t_2 - t_1) I_{AB}\}. \end{aligned}$$

In the case  $i, j \in J$ , pairs  $\{U, A\}$  and  $\{D_{i,j}, B\}$  are independent. Therefore, we obtain

$$\begin{aligned} \bar{E}\{UD_{i,j} I_{AB}\} &= \bar{E}\{U I_A\} \bar{E}\{D_{i,j} I_B\} = \bar{E}\{\bar{E}(U | \mathcal{Y}_{t_1}) I_A\} \bar{E}\{D_{i,j} I_B\} \\ &= \bar{E}\{\bar{E}(U | \mathcal{Y}_{t_1}) D_{i,j} I_{AB}\}. \end{aligned}$$

In the case  $i \in J, k \notin J$ , we use the same independence:

$$\bar{E}\{UD_{i,k} I_{AB}\} = \bar{E}\{\bar{E}(U | \mathcal{Y}_{t_1}) I_A\} \bar{E}\{D_{i,k} I_B\}.$$

It remains to note that  $\bar{E}\{D_{i,k} | \mathcal{Y}_{t_1,t}\} = D_i/p$ , where  $D_i = N_i^d(t_2) - N_i^d(t_1)$ . This follows from two facts: 1)  $N_i^d$  is a sum of the processes  $\{N_{i,l}, l \notin J\}$  that are independent of all observed processes except for  $N_i^d$ ; 2)  $D_{i,k}$  and  $D_i - D_{i,k}$  are independent Poisson variables with parameters proportional to 1 and  $p-1$ , respectively, and hence  $\bar{E}\{D_{i,k} | D_i\} = D_i/p$ .

Thus, we have established (6), (3), and (4). Equality (5) can be verified similarly to (4). A proof of (2) can be found in [19, Lemma 7.3.2].  $\square$

*Proof of Lemma 2.* The exponential  $\Theta(t)$  is positive due to condition (8) [10, 4.62]. To prove the martingale property for  $\Theta(t)$ , we can apply [11, Th.5.1]: it suffices to note that  $\Theta(t)$  is defined by a local  $\bar{P}$ -martingale  $M(t)$  with the integrand that grows no faster than a linear function of the state  $X(t)$ . The last condition coincides with (9). Statements 2–4 can be proved similarly to [19].

To prove the last part, we need to verify that the process  $M_{\alpha,\beta}$  satisfying (1) is a  $P$ -martingale. To do this, we will prove that  $M_{\alpha,\beta}\Theta$  is a  $\bar{P}$ -martingale. Applying Ito's rule, we obtain

$$d(M_{\alpha,\beta}\Theta) = M_{\alpha,\beta}(t-) d\Theta + \Theta(t-) dM_{\alpha,\beta} + \Delta M_{\alpha,\beta} \Delta\Theta.$$

Since the first term in the right-hand side defines a  $\bar{P}$ -martingale, it remains to see that the other terms yield a  $\bar{P}$ -martingale as well:

$$\Theta(t-)(dN_{\alpha,\beta} - \nu_{\alpha,\beta} dt) + \Theta(t-)(\nu_{\alpha,\beta} - 1) dN_{\alpha,\beta} = \Theta(t-) \nu_{\alpha,\beta} \overset{\circ}{dN}_{\alpha,\beta}.$$

$\square$

## References

1. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Berlin (2008)
2. Baccelli, F., Kauffmann, B., Veitch, D.: Inverse problems in queueing theory and Internet probing. *Queueing Syst.* **63**, 59–107 (2009)
3. Bensoussan, A., Cakanyildirim, M., Sethi, S.P., Shi, R.: An incomplete information inventory model with presence of inventories or backorders as only observations. *J. Optim. Theory Appl.* **146**(3), 544–580 (2010)
4. Borisov, A.V.: Application of optimal filtering methods for on-line of queueing network states. *Autom. Remote. Control.* **77**, 277–296 (2016)
5. Bremaud, P.: On the output theorem of queueing theory, via filtering. *J. Appl. Probab.* **15**(2), 397–405 (1978)
6. Elliott, R.J., Aggoun, L., Moore, J.B.: Hidden Markov Models, Estimation and Control. Springer, New York (2008)
7. Elliott, R.J., Dufour, F., Malcolm, W.P.: State and mode estimation for discrete-time jump Markov systems. *SIAM J. Control Optim.* **44**(3), 1081–1104 (2005)
8. El-Taha, M., Stidham, S.: A filtered ASTA property. *Queueing Syst.* **11**, 211–222 (1992)
9. Hassin, R.: Rational Queueing. CRC Press, Boca Raton (2016)
10. Jacod, J., Shiryaev, A.N.: Limit Theorems for Stochastic Processes, 2nd edn. Springer, New York (2003)
11. Klebaner, F., Liptser, R.: When a stochastic exponential is a true martingale. Extension of the Beneš method. *Theory Probab. Appl.* **58**(1), 38–62 (2014)
12. Li, X., Yousefi'zadeh, H.: Robust EKF-based wireless congestion control. *IEEE Trans. Commun.* **61**(12), 5090–5102 (2013)
13. Lukashuk, L.I., Semenchenko, Y.A.: Filtering of a semi-Markov queueing system with retrials. *Cybern. Syst. Anal.* **27**(4), 627–631 (1991)
14. Miller, B.M., Avrachenkov, K.E., Stepanyan, K.V., Miller, G.B.: Flow control as a stochastic optimal control problem with incomplete information. *Probl. Inf. Transm.* **41**(2), 150–170 (2005)
15. Miller, B.M., Miller, G.B., Semenikhin, K.V.: Optimal channel choice for lossy data flow transmission. *Autom. Remote. Control.* **79**(1), 66–77 (2018)
16. Rieder, U., Winter, J.: Optimal control of Markovian jump processes with partial information and applications to a parallel queueing model. *Math. Meth. Oper. Res.* **70**, 567–596 (2009)
17. Stuckey, N., Vasquez, J., Graham, S., Maybeck, P.: Stochastic control of computer networks. *IET Control Theory Appl.* **6**(3), 403–411 (2012)
18. Walrand, J., Varaiya, P.: Flows in queueing networks: a martingale approach. *Math. Oper. Res.* **6**(3), 387–404 (1981)
19. Wong, E., Hajek, B.: Stochastic Processes in Engineering Systems. Springer, New York (1985)



# Estimation of Equilibria in an Advertising Game with Unknown Distribution of the Response to Advertising Efforts

Alan D. Robles-Aguilar<sup>1</sup>, David González-Sánchez<sup>2</sup>,  
and J. Adolfo Minjárez-Sosa<sup>3</sup>(✉)

<sup>1</sup> Instituto Tecnológico de Sonora, Cd. Obregón, Sonora, Mexico  
[alan\\_daniel@yahoo.com](mailto:alan_daniel@yahoo.com)

<sup>2</sup> CONACYT–Universidad de Sonora, Rosales s/n,  
83000 Hermosillo, Sonora, Mexico  
[david.glzsnz@gmail.com](mailto:david.glzsnz@gmail.com)

<sup>3</sup> Departamento de Matemáticas, Universidad de Sonora, Rosales s/n,  
83000 Hermosillo, Sonora, Mexico  
[aminjare@gauss.mat.uson.mx](mailto:aminjare@gauss.mat.uson.mx)

**Abstract.** We study a class of discrete-time advertising game with random responses to the advertising efforts made by a duopoly. The firms are assumed to observe the values of the random responses but they do not know their distributions. With the recorded values, firms estimate distributions and play estimated equilibrium strategies. Under suitable assumptions, we prove that the estimated equilibrium strategies converge to equilibria of the advertising game with the true distributions. Our results are numerically illustrated for specific cases.

**Keywords:** Advertising games · Lanchester model · Markov games · Empirical distribution

**AMS (2020) Subject Classification:** Primary 91A15 · Secondary 91A80

## 1 Introduction

We consider a dynamic noncooperative game of advertising where the market shares of the firms follow a stochastic difference equation. The stochastic behavior in the market shares comes from the uncertain responses to advertising efforts modeled by a sequence of random variables. Further, we assume that firms can observe the values of such random variables *a posteriori* but they do not know the distributions. In this sense, by using appropriate statistical estimation methods to approximate the distributions of the random variables, firms can play Nash equilibrium strategies of the estimated games. When these equilibrium strategies converge, the question we aim to answer is whether the limit strategies are equilibria for the game with the true distributions of the responses to advertising efforts.

The literature about dynamic models of advertising and marketing games is very large; we can mention the papers [4, 6, 7, 21] and the books [2, 8]. Most of these references mainly focus on deterministic differential game models; instead there are few works that deal with stochastic differential game models and deterministic discrete-time models, we can cite, for instance, [1, 18]. On the other hand, discrete-time stochastic zero-sum games with incomplete information have been studied under several context, see, e.g., [5, 10, 12–16, 22, 23], which include the case when the transition law among states is unknown. However, to the best of our knowledge, the only work dealing on estimation problem for nonzero-sum Markov games is [19]. Specifically, in [19] is used the empirical distribution of the disturbance process to obtain an almost surely convergent procedure to approximate Nash equilibria under the discounted criterion.

In this chapter we analyze the stochastic version of the advertising Lanchester model introduced in [1]. Additionally, we assume that the random variables modeling the uncertainty in responses to advertising efforts have unknown distributions. Under this scenario, using the empirical distribution as an estimator and considering finite action sets for players, we apply similar ideas to [19] to simulate values of the advertising responses, estimate equilibrium strategies, and prove that these equilibria converge in some sense to an equilibrium of the advertising game with full information. In order to introduce the model and compare our results, previously we analyze the advertising game with full information, where we numerically compute the Nash equilibria in mixed stationary strategies.

The remaining of the paper is organized as follows. The stochastic advertising game we deal with is described in Sect. 2 as well as the numerical algorithm we use to compute the Nash equilibria. Section 3 is devoted to the stochastic game with unknown distributions of the advertising responses. Finally, in Sect. 4, we give some conclusions.

## 2 A Discrete-Time Stochastic Game of Advertising

Essentially, Lanchester model is an ordinary-differential-equation model of warfare [11]. Over time, this model has been adapted to study different conflict situations, including advertising models. In this section, we introduce a discrete-time stochastic version of the Lanchester model in the context of the models that appear in [1] and [8, pp. 29–31]. We also give a numerical algorithm to find Nash equilibria in stationary strategies of the proposed model.

### 2.1 The Advertising Game Model

Consider a duopoly competing for the market share by making advertising efforts. Let  $x$  be the market share of Firm 1 and let  $a$  and  $b$  be the advertising efforts of Firm 1 and Firm 2, respectively, at some decision epoch. The market share of Firm 2 is  $1 - x$ . Then the market share of Firm 1 at the beginning of the next decision epoch is determined by the mapping

$$(x, a, b) \mapsto x + (1 - x)d(\xi, a) - xe(\zeta, b) \quad (1)$$

where  $d(\xi, a)$  and  $e(\zeta, b)$  are the advertising responses to  $a$  and  $b$ , respectively, and  $(\xi, \zeta)$  is a pair of random variables. The functions  $d(i, \cdot)$  and  $e(j, \cdot)$ —for fixed values of  $i$  and  $j$ —are *production functions*, that is, they are increasing, have diminishing marginal effects, and take nonnegative values. Typical advertising responses are

$$d(\xi, a) = \xi\sqrt{a}, \quad e(\zeta, b) = \zeta\sqrt{b}. \quad (2)$$

The evolution of the state system is given by the mapping (1) and has the following interpretation: the advertising of Firm 1 aims to attract customers from Firm 2, thus the increment of the market share is proportional to  $(1 - x)$ , and analogously for the advertising made by Firm 2.

For the purposes of this paper, we assume that the triples  $(x, a, b)$  belong to a finite set  $\mathbb{X} \times \mathbb{A} \times \mathbb{B}$ . Thus the image of the mapping (1)—with the advertising responses (2), for instance—is not necessarily a subset of  $\mathbb{X}$ . In such a case, we map  $x + (1 - x)d(\xi, a) - xe(\zeta, b)$  to the nearest state in  $\mathbb{X}$ . Although, for simplicity, we write

$$x_{k+1} = x_k + (1 - x_k)d(\xi_k, a_k) - x_k e(\zeta_k, b_k), \quad k = 0, 1, \dots, \quad (3)$$

where  $x_0 \in \mathbb{X}$  is given. In addition, the so-called disturbance processes  $\{\xi_k\}$  and  $\{\zeta_k\}$  consist of independent and identically distributed (i.i.d.) random variables, which take values in the finite sets  $\mathbb{S}_1$  and  $\mathbb{S}_2$  respectively. The process  $\{(\xi_k, \zeta_k)\}$  is defined on some underlying probability space  $(\Omega, \mathcal{F}, P)$ . The common probability functions of the random variables  $\{\xi_k\}$  and  $\{\zeta_k\}$  are, respectively,  $\theta$  and  $\vartheta$ , that is,

$$\begin{cases} \theta(i) = P[\xi_k = i] & \forall i \in \mathbb{S}_1, k \in \mathbb{N}_0, \\ \vartheta(j) = P[\zeta_k = j] & \forall j \in \mathbb{S}_2, k \in \mathbb{N}_0. \end{cases} \quad (4)$$

We use the notation  $\mathbb{K} := \{(x, a, b) : x \in \mathbb{X}, a \in \mathbb{A}, b \in \mathbb{B}\}$ . Combining (3) and (4), we obtain the transition law among the states as follows. For each  $(x, a, b) \in \mathbb{K}$ ,

$$P_{x,y}[a, b] := P[x_{k+1} = y \mid x_k = x, a_k = a, b_k = b] = \sum_{(i,j) \in S_F} \theta(i)\vartheta(j), \quad y \in \mathbb{X} \quad (5)$$

where

$$S_F := \{(s, t) \in \mathbb{S}_1 \times \mathbb{S}_2 : x + (1 - x)d(s, a) - xe(t, b) = y\}.$$

Finally,  $r_i : \mathbb{K} \rightarrow \mathbb{R}$  is the one-stage payoff function for the Firm  $i = 1, 2$ ,

$$\begin{cases} r_1(x, a, b) = p_1 x - a \\ r_2(x, a, b) = p_2(1 - x) - b \end{cases} \quad (6)$$

where  $p_1$  and  $p_2$  are the gross profit rate of Firms 1 and 2 respectively. In what follows, the probability space  $(\Omega, \mathcal{F}, P)$  is fixed and *a.s.* means almost surely with respect to  $P$ .

Putting together all the elements described above, we define the advertising game model as

$$\mathcal{G}_{\theta, \vartheta} := (\mathbb{X}, \mathbb{A}, \mathbb{B}, \mathbb{S}_1, \mathbb{S}_2, \theta, \vartheta, r_1, r_2) \quad (7)$$

The model is a representation of a dynamic game which is played as follows. At each stage  $k \in \mathbb{N}_0$ , when the game is in state  $x_k \in \mathbb{X}$ , the firms independently choose actions  $a_k = a \in \mathbb{A}$  and  $b_k = b \in \mathbb{B}$ . Consequently, the following happens: first, Firm  $i$  receives payoffs of  $r_i(x, a, b)$ ,  $i = 1, 2$ ; and second, the system moves to the next state  $x_{k+1} \in \mathbb{X}$  according to probability transition (5). Once the system reaches the next state, the process repeats. In addition, the payoffs are accumulated according to a discounted criterion, as we will define below.

Let  $P_{\mathbb{A}}$  and  $P_{\mathbb{B}}$  consist of the set of all probability functions on  $\mathbb{A}$  and  $\mathbb{B}$  respectively. That is,  $P_{\mathbb{A}}$  is the set of functions  $\sigma : A \rightarrow [0, 1]$  such that  $\sum_{a \in \mathbb{A}} \sigma(a) = 1$ . Similarly for  $P_{\mathbb{B}}$ . By convention, for each  $\sigma \in P_{\mathbb{A}}$ ,  $\tau \in P_{\mathbb{B}}$ , we denote

$$v(x, \sigma, \tau) := \sum_{a \in \mathbb{A}} \sum_{b \in \mathbb{B}} v(x, a, b) \sigma(a) \tau(b), \quad x \in \mathbb{X} \quad (8)$$

for any function  $v : \mathbb{K} \rightarrow \mathbb{R}$ . Likewise, for  $\sigma \in P_{\mathbb{A}}$ ,  $\tau \in P_{\mathbb{B}}$

$$[x + (1 - x)d(s, \sigma) - xe(t, \tau)] := \sum_{a \in \mathbb{A}} \sum_{b \in \mathbb{B}} [x + (1 - x)d(s, a) - xe(t, b)] \sigma(a) \tau(b), \quad (9)$$

where  $x \in \mathbb{X}$ ,  $s \in \mathbb{S}_1$ , and  $t \in \mathbb{S}_2$ .

A strategy played by Firm 1 is a sequence  $\pi = \{\pi_k\}$  where  $\pi_k$  is a probability function over  $\mathbb{A}$  conditioned on the history  $h_k := (x_0, a_0, b_0, \dots, a_{k-1}, b_{k-1}, x_k)$ . That is, for each history  $h_k$ ,  $\pi_k(\cdot | h_k) \in P_{\mathbb{A}}$ . The set of all strategies for Firm 1 is denoted by  $\Pi$ . A strategy  $\pi \in \Pi$  is said to be a Markov strategy if there is a probability function  $f_k$  over  $\mathbb{A}$  such that  $\pi_k(\cdot | h_k) = f_k(\cdot | x_k)$  for all  $k \in \mathbb{N}_0$ . Further, a Markov strategy  $\pi = \{f_k\}$  is stationary if  $f_k = f$  for all  $k \in \mathbb{N}_0$ ; in this case, we use this notation

$$f^\infty := \{f, f, f, \dots\}.$$

We denote by  $\Pi_M$  and  $\mathbb{F}$  the sets of Markov strategies and stationary strategies, respectively, for Firm 1. The sets  $\Gamma$ ,  $\Gamma_M$ , and  $\mathbb{G}$  of all strategies, Markov strategies, and stationary strategies for Firm 2 are defined similarly.

Let  $\pi = \{\pi_k\} \in \Pi$  and  $\gamma = \{\gamma_k\} \in \Gamma$  be a pair of strategies. For each initial state  $x \in \mathbb{X}$ , we define the discounted criterion, also known as expected discounted payoff, for Firm  $i = 1, 2$ , as

$$\begin{cases} J_1^{\theta, \vartheta} = E_x^{(\pi, \gamma)} \left[ \sum_{k=0}^{\infty} \beta^k \{p_1 x_k - a_k\} \right] \\ J_2^{\theta, \vartheta} = E_x^{(\pi, \gamma)} \left[ \sum_{k=0}^{\infty} \beta^k \{p_2(1 - x_k) - b_k\} \right] \end{cases} \quad (10)$$

where  $\beta \in (0, 1)$  is the discount factor and  $E_x^{(\pi, \gamma)}$  denotes the expectation operator corresponding to the unique probability measure  $P_x^{(\pi, \gamma)}$  induced by  $x \in \mathbb{X}$  and  $(\pi, \gamma) \in \Pi \times \Gamma$ , (see [3]).

## 2.2 Stationary Nash Equilibrium in Discounted Games

**Definition 1.** A pair of strategies  $(\pi^*, \gamma^*) \in \Pi \times \Gamma$  is a Nash equilibrium if, for all  $x \in \mathbb{X}$ ,

$$J_1^{\theta, \vartheta}(x, \pi^*, \gamma^*) \geq J_1^{\theta, \vartheta}(x, \pi, \gamma^*), \quad \forall \pi \in \Pi$$

and

$$J_2^{\theta, \vartheta}(x, \pi^*, \gamma^*) \geq J_2^{\theta, \vartheta}(x, \pi^*, \gamma), \quad \forall \gamma \in \Gamma.$$

The equilibrium payoffs of the game, with initial state  $x$ , are  $J_1^{\theta, \vartheta}(x, \pi^*, \gamma^*)$  and  $J_2^{\theta, \vartheta}(x, \pi^*, \gamma^*)$ .

The following lemma about the existence of Nash equilibria in Markov strategies for this model is well known. For instance, see [17, Theorem 5.1].

**Lemma 1.** The game model, with discounted payoffs  $J_1^{\theta, \vartheta}$  and  $J_2^{\theta, \vartheta}$ , has a Nash equilibrium in stationary strategies. That is, there exists  $(f^\infty, g^\infty) \in \mathbb{F} \times \mathbb{G}$  such that for each  $x \in \mathbb{X}$ ,

$$J_1^{\theta, \vartheta}(x, f^\infty, g^\infty) \geq J_1^{\theta, \vartheta}(x, \pi, g^\infty), \quad \forall \pi \in \Pi$$

and

$$J_2^{\theta, \vartheta}(x, f^\infty, g^\infty) \geq J_2^{\theta, \vartheta}(x, f^\infty, \gamma), \quad \forall \gamma \in \Gamma.$$

Observe that once  $f^\infty \in \mathbb{F}$  and  $g^\infty \in \mathbb{G}$  are fixed,

$$\bar{J}_1(x, \pi) := J_1^{\theta, \vartheta}(x, \pi, g^\infty), \quad \pi \in \Pi, \quad x \in \mathbb{X}$$

and

$$\bar{J}_2(x, \gamma) := J_2^{\theta, \vartheta}(x, f^\infty, \gamma), \quad \gamma \in \Gamma, \quad x \in \mathbb{X}$$

constitute performance indices, where each of them corresponds to an optimal control problem. Hence, the *value functions*

$$V(x) := \max_{\pi \in \Pi} \bar{J}_1(x, \pi), \quad x \in \mathbb{X} \quad (11)$$

and

$$W(x) := \max_{\gamma \in \Gamma} \bar{J}_2(x, \gamma), \quad x \in \mathbb{X}, \quad (12)$$

satisfy, respectively, the Dynamic Programming equations

$$V(x) = \max_{\mu \in P_A} \left[ [p_1 x - \mu] + \beta \sum_{(i,j) \in \mathbb{S}_1 \times \mathbb{S}_2} V[x + (1-x)d(i, \mu) - xe(j, g)]\theta(i)\vartheta(j) \right] \quad (13)$$

$$= [p_1 x - f] + \beta \sum_{(i,j) \in \mathbb{S}_1 \times \mathbb{S}_2} V[x + (1-x)d(i, f) - xe(j, g)]\theta(i)\vartheta(j), \quad \forall x \in \mathbb{X}, \quad (14)$$

and

$$W(x) = \max_{\lambda \in P_B} \left[ [p_2(1-x) - \lambda] + \beta \sum_{(i,j) \in \mathbb{S}_1 \times \mathbb{S}_2} W[x + (1-x)d(i, f) - xe(j, \lambda)]\theta(i)\vartheta(j) \right] \quad (15)$$

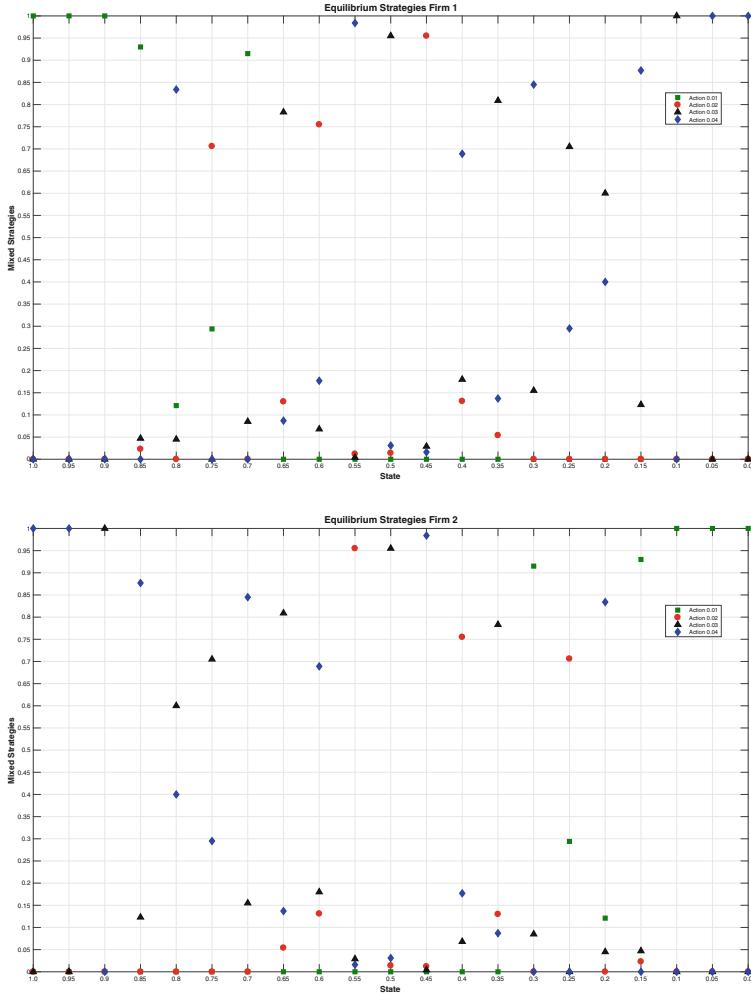
$$= [p_2(1-x) - g] + \beta \sum_{(i,j) \in \mathbb{S}_1 \times \mathbb{S}_2} W[x + (1-x)d(i, f) - xe(j, g)]\theta(i)\vartheta(j), \quad \forall x \in \mathbb{X}. \quad (16)$$

*Remark 1.* By considering standard dynamic programming arguments, if there are functions  $V$  and  $W$  and a pair  $(f, g)$  satisfying (13)–(16), then  $(f^\infty, g^\infty) \in \mathbb{F} \times \mathbb{G}$  is a stationary Nash equilibrium for the game with discounted payoffs (10). Further, the equilibrium payoffs are  $J_1^{\theta, \vartheta}(x, f^\infty, g^\infty) = V(x)$  and  $J_2^{\theta, \vartheta}(x, f^\infty, g^\infty) = W(x)$ .

## 2.3 Numerical Examples

We compute the equilibria in Markov strategies for an advertising game with the data of Table 1.

The equilibrium strategies are found using and adaptation of the well-known value iteration algorithm from discounted dynamic programming. In each

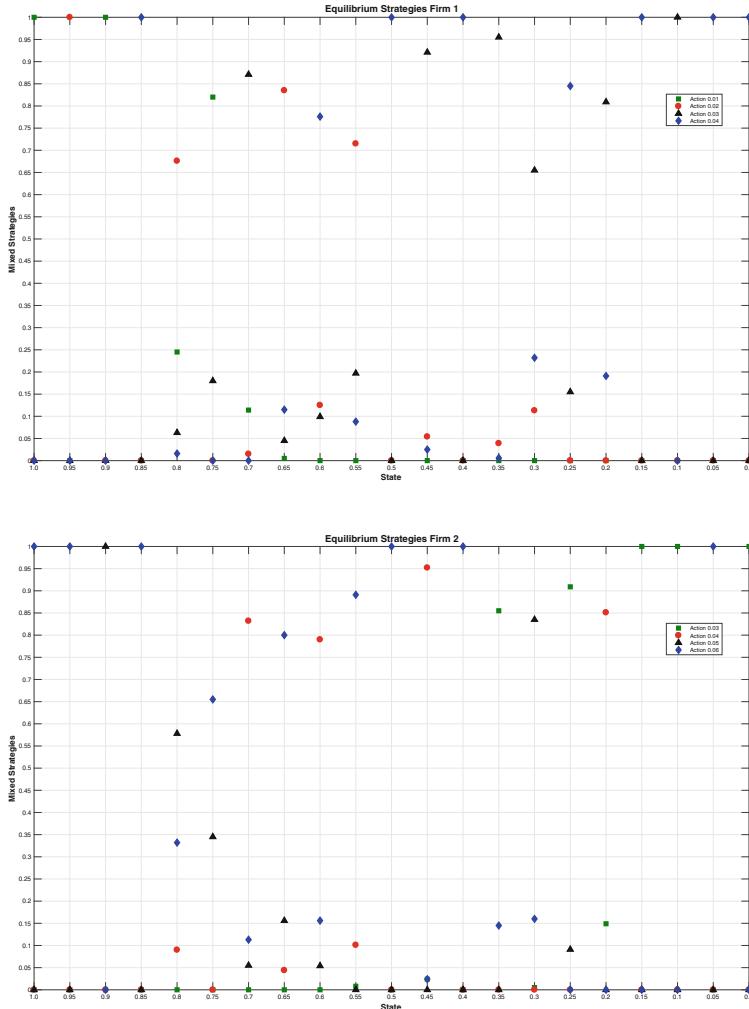


**Fig. 1.** Equilibrium strategies  $f$  and  $g$  in the full-information game with data of Table 1. The height of each action is the probability it is played with.

iteration we get the equilibrium by minimizing McKelvey's function, see [9, p. 133]. For the parameters given above, the iteration algorithm converges. The algorithm is implemented in Python and the code is available at

<https://github.com/adra1973/>

The limit strategies  $(f, g)$ , that form the stationary equilibrium  $(f^\infty, g^\infty)$ , are plotted in Fig. 1 and 2. Since we are using exactly the same parameters for both firms, in Fig. 1 we can observe for each state an effect of "mirror" in the strategies for both firms.



**Fig. 2.** Equilibrium strategies  $f$  and  $g$  in the full-information game with data from Table 1 but the set of actions for Firm 2 is replaced by (17).

In Fig. 2 we plot the equilibrium strategies for the game with the same data of Table 1 but the set of actions for Firm 2 now is

$$\mathbb{B} = \{0.03, 0.04, 0.05, 0.06\} \quad (17)$$

and thus the behavior of the strategies breaks the “mirror” observed before.

**Table 1.** Data for the advertising game.

Variable	Description
$\mathbb{X}$	Space of 21 states of market shares, $\{0.0, 0.05, 0.1, 0.15, 0.2, \dots, 0.8, 0.85, 0.9, 0.95, 1.0\}$ ,
$\mathbb{A}$	Set of 4 actions for advertising effort of Firm 1, $\mathbb{A} = \{0.01, 0.02, 0.03, 0.04\}$ ,
$\mathbb{B}$	Set of 4 actions for advertising effort of Firm 2, $\mathbb{B} = \{0.01, 0.02, 0.03, 0.04\}$ ,
$\mathbb{S}_1$	Set of 10 values of Firm 1, $\mathbb{S}_1 = \{0.95, \dots, 1.05\}$
$\mathbb{S}_2$	Set of 10 values of Firm 2, $\mathbb{S}_2 = \{0.95, \dots, 1.05\}$
$\xi$	Random variable of Firm 1 that take values in $\mathbb{S}_1$ with probability $\theta(i)$ , $i \in \mathbb{S}_1$ , $\xi \sim \text{Binomial}(10, 0.4)$
$\zeta$	Random variable of Firm 2 that take values in $\mathbb{S}_2$ with probability $\vartheta(j)$ , $j \in \mathbb{S}_2$ , $\zeta \sim \text{Binomial}(10, 0.4)$ .
$d$	Advertising response function of Firm 1, $d(\xi, a) = \xi\sqrt{a}$ , $a \in \mathbb{A}$
$e$	Advertising response function of Firm 2, $e(\zeta, b) = \zeta\sqrt{b}$ , $b \in \mathbb{B}$
$p_1$	Gross profit for each product sold by Firm 1, $p_1 = 1.2$
$p_2$	Gross profit for each product sold by Firm 2, $p_2 = 1.2$
$\beta$	The discount factor $\beta = 0.95$

### 3 The Advertising Game with Unknown Distribution

In this section, we study the advertising game when the distributions of the random variables  $(\xi, \zeta)$  are unknown for the players. We assume that, after the  $n$ -th stage, players have recorded the values  $\underline{\xi}_n := (\xi_0, \xi_1, \dots, \xi_n)$  and  $\bar{\zeta}_n := (\zeta_0, \zeta_1, \dots, \zeta_n)$  and use the empirical distributions

$$\theta_n(i) := \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{1}_i(\xi_t), \quad i \in \mathbb{S}_1, \quad n \in \mathbb{N}$$

and

$$\vartheta_n(j) := \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{1}_j(\zeta_t), \quad j \in \mathbb{S}_2, \quad n \in \mathbb{N}$$

to estimate equilibrium strategies. More precisely, for each  $n \in \mathbb{N}$ , consider the *empirical advertising game*

$$\mathcal{G}_{\theta_n, \vartheta_n} := (\mathbb{X}, \mathbb{A}, \mathbb{B}, \mathbb{S}_1, \mathbb{S}_2, \theta_n, \vartheta_n, r_1, r_2) \tag{18}$$

with dynamics (3) and payoffs (10), where  $\theta$  and  $\vartheta$  are replaced by  $\theta_n$  and  $\vartheta_n$ , respectively. Given a stationary Nash equilibrium  $(f_n^\infty, g_n^\infty)$  for the empirical advertising game (18), by well-known dynamic programming results, there exist functions  $V_n$  and  $W_n$  that satisfy the optimality equations

$$\begin{aligned} V_n(x) &= \max_{\mu \in P_{\mathbb{A}}} \left[ [p_1 x - \mu] + \beta \sum_{i,j} V_n[x + (1-x)d(i, \mu) - xe(j, g_n)]\theta_n(i)\vartheta_n(j) \right] \\ &= [p_1 x - f_n] + \beta \sum_{i,j} V_n[x + (1-x)d(i, f_n) - xe(j, g_n)]\theta_n(i)\vartheta_n(j), \quad x \in \mathbb{X}, \end{aligned} \quad (19)$$

and

$$\begin{aligned} W_n(x) &= \max_{\lambda \in P_{\mathbb{B}}} \left[ [p_2(1-x) - \lambda] + \beta \sum_{i,j} W_n[x + (1-x)d(i, f_n) - xe(j, \lambda)]\theta_n(i)\vartheta_n(j) \right] \\ &= [p_2(1-x) - g_n] + \beta \sum_{i,j} W[x + (1-x)d(i, f_n) - xe(j, g_n)]\theta_n(i)\vartheta_n(j), \quad x \in \mathbb{X}. \end{aligned} \quad (20)$$

*Remark 2.* Notice that  $V_n$  and  $W_n$  are defined on  $\mathbb{X} \times \Omega$ , thus  $V_n(x)$  and  $W_n(x)$  are random variables for each  $x \in \mathbb{X}$ . The strategies  $f_n$  and  $g_n$  are also random vectors.

The following proposition is based on [19]; for completeness, we outline a proof in the scenario of the present work.

**Proposition 1.** For each  $n \in \mathbb{N}$ , let  $f_n, g_n, V_n$ , and  $W_n$  satisfy (19) and (20). If

$$\lim_{n \rightarrow \infty} (f_n, g_n) = (f, g) \quad P - a.s. \quad (21)$$

and

$$\lim_{n \rightarrow \infty} (V_n, W_n) = (V, W) \quad P - a.s., \quad (22)$$

then  $(f^\infty, g^\infty)$  is  $P - a.s.$  a Nash equilibrium for the advertising game with dynamics (3) and payoffs (10).

*Proof.* It is well known that from the strong law of large numbers,

$$(\theta_n, \vartheta_n) \rightarrow (\theta, \vartheta) \quad P - a.s. \quad (23)$$

Now, fix  $\omega$  in  $\Omega$  such that the convergence in (21), (22), and (23) holds. Then, for each  $\mu \in P_{\mathbb{A}}$ ,  $x \in \mathbb{X}$ , and  $n \in \mathbb{N}$ ,

$$\begin{aligned} &\sum_{i,j} \left| V_n[x + (1-x)d(i, \mu) - xe(j, g_n)] \right. \\ &\quad \left. - V[x + (1-x)d(i, \mu) - xe(j, g_n)] \right| \theta_n(i)\vartheta_n(j) \\ &\leq \sum_{i,j} \max_{x \in \mathbb{X}} \left| V_n(x) - V(x) \right| \theta_n(i)\vartheta_n(j) \\ &\leq \max_{x \in \mathbb{X}} \left| V_n(x) - V(x) \right|. \end{aligned} \quad (24)$$

and

$$\begin{aligned}
& \sum_{i,j} \left| V[x + (1-x)d(i, \mu) - xe(j, g_n)] \right. \\
& \quad \left. - V[x + (1-x)d(i, \mu) - xe(j, g)] \theta_n(i) \vartheta_n(j) \right| \\
& \leq \sum_{i,j} \sum_{b \in \mathbb{B}} \left| V[x + (1-x)d(i, \mu) - xe(j, b)] \right| \left| g_n(b|x) - g(b|x) \right| \theta_n(i) \vartheta_n(j) \\
& \leq \max_{x \in \mathbb{X}} |V(x)| \sum_{b \in \mathbb{B}} \left| g_n(b|x) - g(b|x) \right|
\end{aligned} \tag{25}$$

Thus

$$\begin{aligned}
& \sum_{i,j} \left| V_n[x + (1-x)d(i, \mu) - xe(j, g_n)] \theta_n(i) \vartheta_n(j) \right. \\
& \quad \left. - V[x + (1-x)d(i, \mu) - xe(j, g)] \theta(i) \vartheta(j) \right| \\
& \leq \sum_{j \in \mathbb{S}} \left| V_n[x + (1-x)d(i, \mu) - xe(j, g_n)] \theta_n(i) \vartheta_n(j) \right. \\
& \quad \left. - V[x + (1-x)d(i, \mu) - xe(j, g_n)] \theta_n(i) \vartheta_n(j) \right| \\
& \quad + \sum_{j \in \mathbb{S}} \left| V[x + (1-x)d(i, \mu) - xe(j, g_n)] \theta_n(i) \vartheta_n(j) \right. \\
& \quad \left. - V[x + (1-x)d(i, \mu) - xe(j, g)] \theta(i) \vartheta_n(j) \right| \\
& \quad + \sum_{j \in \mathbb{S}} \left| V[x + (1-x)d(i, \mu) - xe(j, g)] \theta(i) \vartheta(j) \right. \\
& \quad \left. - V[x + (1-x)d(i, \mu) - xe(j, g)] \theta_n(i) \vartheta_n(j) \right| \\
& \quad + \sum_{j \in \mathbb{S}} \left| V[x + (1-x)d(i, \mu) - xe(j, g)] \theta_n(i) \vartheta_n(j) \right. \\
& \quad \left. - V[x + (1-x)d(i, \mu) - xe(j, g)] \theta(i) \vartheta(j) \right|.
\end{aligned}$$

Then, (24), (25), and (23) imply

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \sum_{i,j} V_n[x + (1-x)d(i, \mu) - xe(j, g_n)] \theta_n(i) \vartheta_n(j) \\
& = \sum_{i,j} V[x + (1-x)d(i, \mu) - xe(j, g)] \theta(i) \vartheta(j) \quad P - a.s.
\end{aligned} \tag{26}$$

for each  $\mu \in P_{\mathbb{A}}$  and  $x \in \mathbb{X}$ . We can also show that

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \sum_{i,j} V_n[x + (1-x)d(i, f_n) - xe(j, g_n)] \theta_n(i) \vartheta_n(j) \\
& = \sum_{i,j} V[x + (1-x)d(i, f) - xe(j, g)] \theta(i) \vartheta(j) \quad P - a.s.
\end{aligned} \tag{27}$$

On the other hand, from (24) and (26), we have

$$V_n(x) \geq [p_1x - \mu] + \beta \sum_{i,j} V_n[x + (1-x)d(i,\mu) - xe(j,g_n)]\theta(i)\vartheta(j) \quad \forall \mu \in P_{\mathbb{A}}$$

and hence, by letting  $n \rightarrow \infty$ ,

$$V(x) \geq [p_1x - \mu] + \beta \sum_{i,j} V[x + (1-x)d(i,\mu) - xe(j,g)]\theta(i)\vartheta(j) \quad \forall \mu \in P_{\mathbb{A}}.$$

Furthermore, the second equality in (24) and (27) yield

$$\begin{aligned} V(x) &= \max_{\mu \in P_{\mathbb{A}}} \left[ [p_1x - \mu] + \beta \sum_{i,j} V[x + (1-x)d(i,\mu) - xe(j,g)]\theta(i)\vartheta(j) \right] \\ &= [p_1x - f] + \beta \sum_{i,j} V[x + (1-x)d(i,f) - xe(j,g)]\theta(i)\vartheta(j), \quad P-a.s. \end{aligned}$$

The following equalities are analogously proved

$$\begin{aligned} W(x) &= \max_{\lambda \in P_{\mathbb{B}}} \left[ [p_2(1-x) - \lambda] + \beta \sum_{i,j} W[x + (1-x)d(i,f) - xe(j,\lambda)]\theta(i)\vartheta(j) \right] \\ &= [p_2(1-x) - g] + \beta \sum_{i,j} W[x + (1-x)d(i,f) - xe(j,g)]\theta(i)\vartheta(j), \quad P-a.s. \end{aligned}$$

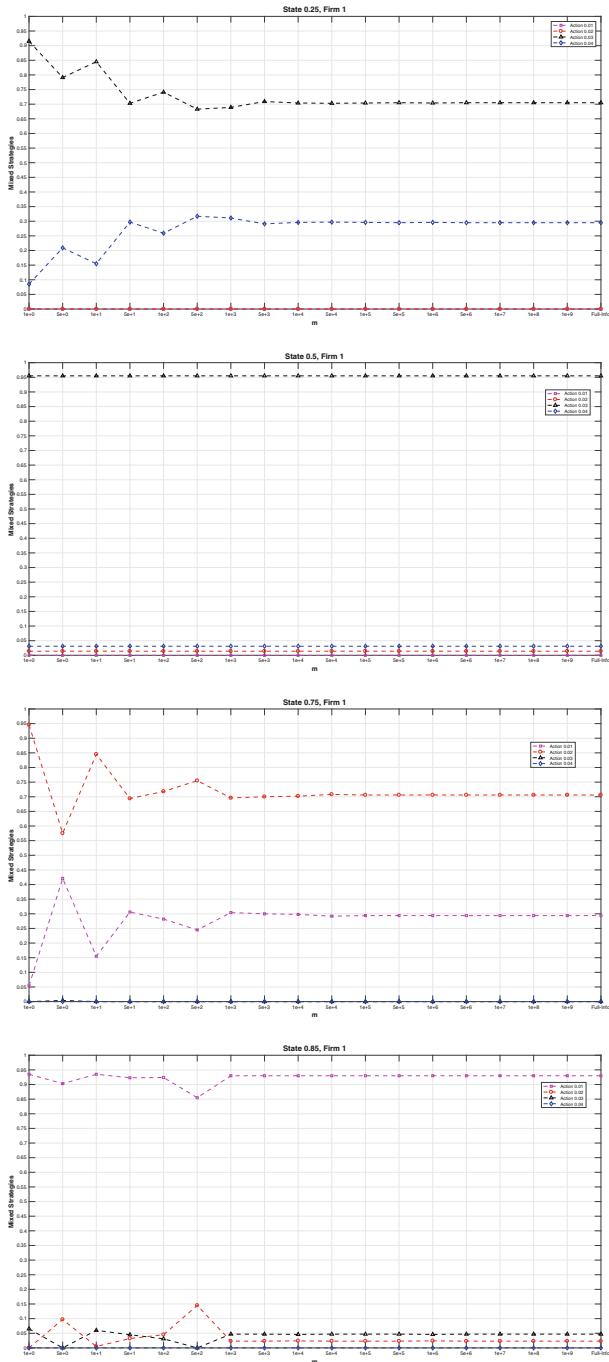
These optimality equations prove that  $(f^\infty, g^\infty)$  is a stationary Nash equilibrium  $P-a.s.$  for the advertising game.  $\square$

### 3.1 Numerical Examples for the Empirical Game Model

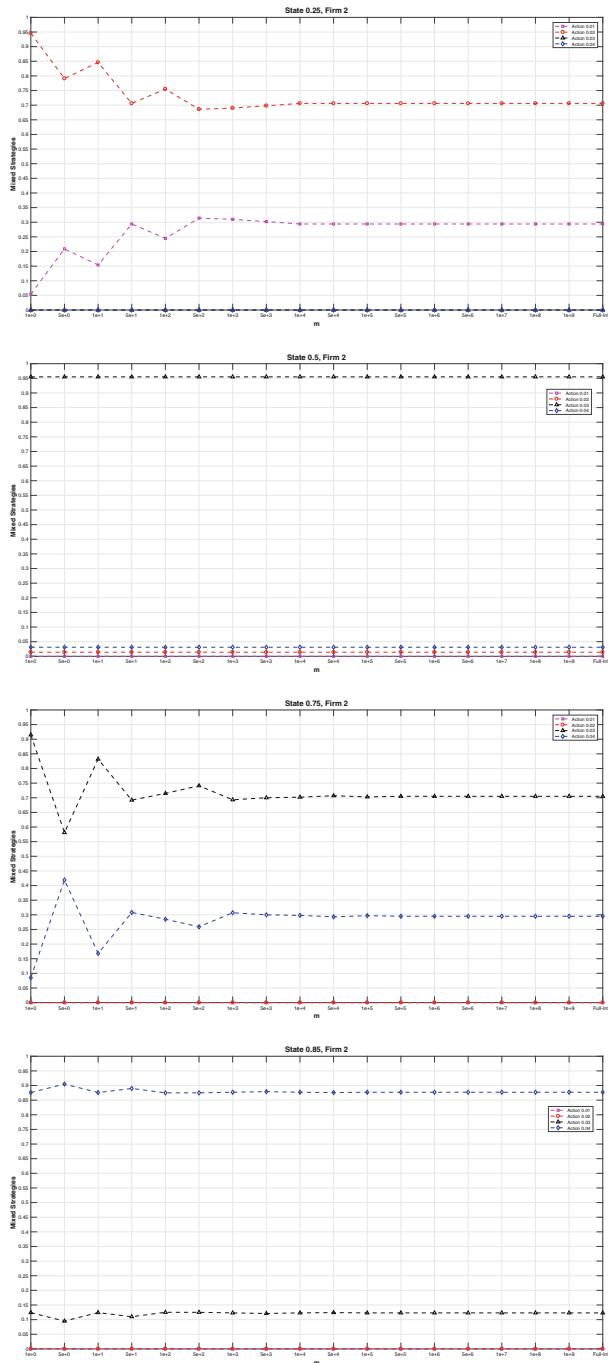
In order to generate simulations of the empirical games  $\mathcal{G}_{\theta_m, \vartheta_m}$ , we use the algorithm in [20, p. 56] to produce values from a Binomial random variable. All parameters are exactly the same as in Table 1 but the pair  $(\theta, \vartheta)$  is replaced by  $(\theta_m, \vartheta_m)$ . As in Subsection 2.3, we compute the stationary Nash equilibrium  $(f_m^\infty, g_m^\infty)$  for each empirical game  $\mathcal{G}_{\theta_m, \vartheta_m}$ , with  $m \in \mathbb{N}_0$ .

For a realization  $\omega \in \Omega$  and different values of  $m$ , the equilibrium strategies  $(f_m, g_m)$  are plotted in Fig. 3 and 4, and equilibrium payoffs  $(V_m, W_m)$  are shown in Fig. 5 and 6. By looking at the proof of Proposition 1, if (21) and (22) hold for a given value of  $\omega$ , then the limit strategy pair  $(f, g)$  determines a stationary Nash equilibrium of the full information game. The equilibrium strategy  $(f, g)$  or equilibrium payoffs  $(V, W)$  for the full-information model (7) are also plotted on the right of each figure.

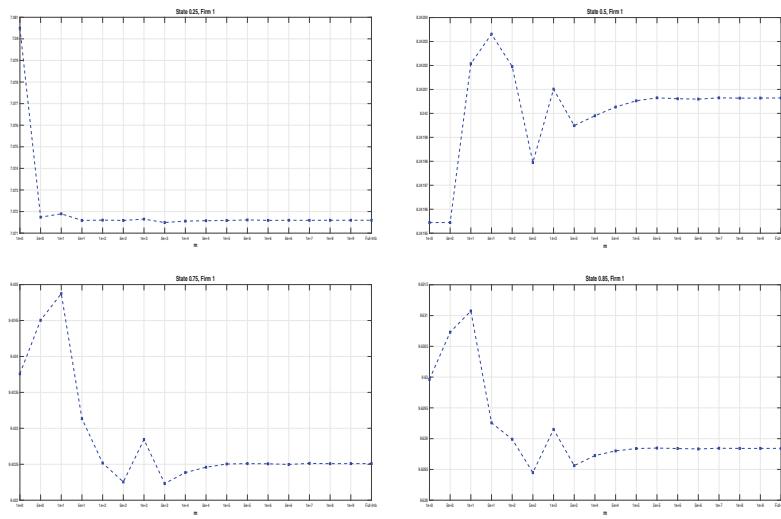
A numerical validation of the hypotheses in Proposition 1 would consist in simulating empirical games for infinitely many realizations of  $\omega$ , computing the equilibria along with the payoffs, and verifying (21) and (22). From a practical point of view, however, firms record the values of the random variables—and



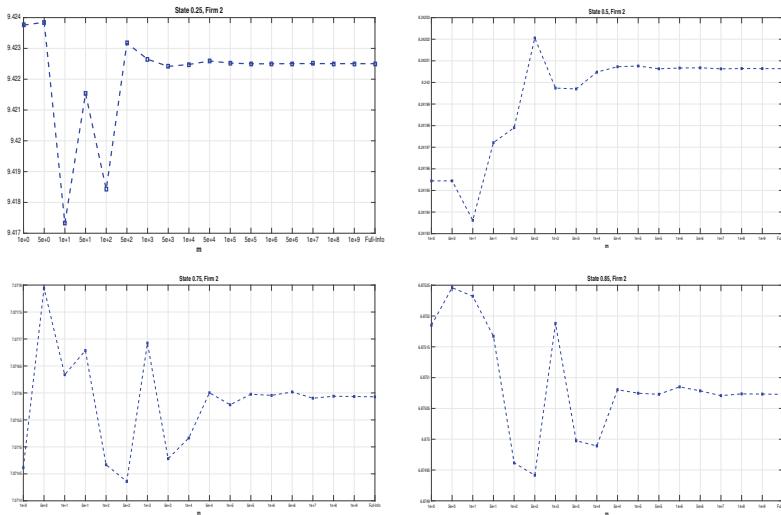
**Fig. 3.** Estimated equilibrium strategies of Firm 1 for different values of  $m$  at the states 0.25, 0.5, 0.75, and 0.85.



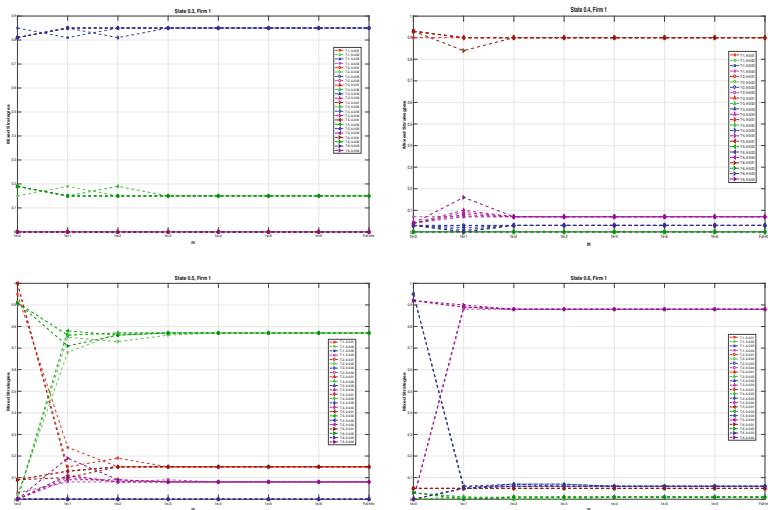
**Fig. 4.** Estimated equilibrium strategies of Firm 2 for different values of  $m$  at the states 0.25, 0.5, 0.75, and 0.85.



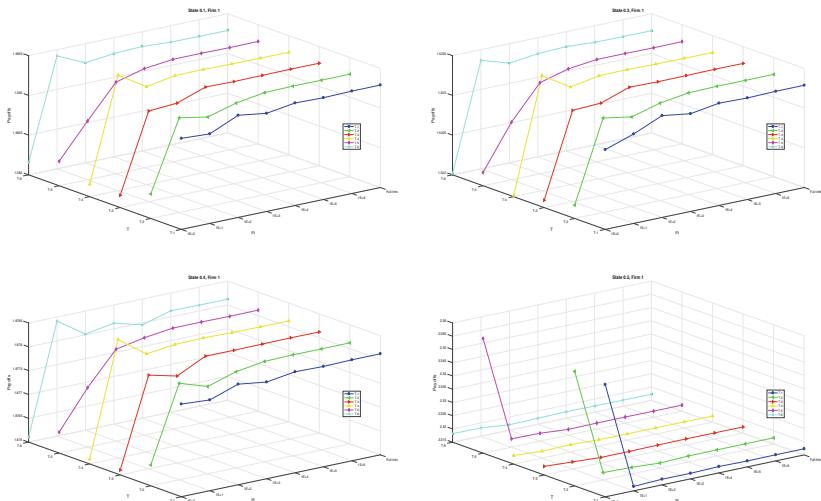
**Fig. 5.** Estimated equilibrium payoffs of Firm 1 for different values of  $m$  at the states 0.25, 0.5, 0.75, and 0.85.



**Fig. 6.** Estimated equilibrium payoffs of Firm 2 for different values of  $m$  at the states 0.25, 0.5, 0.75, and 0.85.



**Fig. 7.** Estimated equilibrium strategies of Firm 1 for six realizations of  $\omega$  and different values of  $m$  at states 0.3, 0.4, 0.5, and 0.6.



**Fig. 8.** Estimated equilibrium payoffs of Firm 1 for six realizations of  $\omega$  and different values of  $m$  at states 0.1, 0.2, 0.4, and 0.5.

play the corresponding equilibrium strategies—of a single realization  $\omega$ . If the strategies converge, then Proposition 1 asserts that, with probability 1, the estimated equilibrium strategies are close to an equilibrium of the full-information game.

For illustrative purposes, in Fig. 7, we plot the equilibrium strategies corresponding to six different realizations of  $\omega$ . The game model components are given in Table 1, except for  $\beta = 0.75$  and  $\mathbb{X} = \{0.0, 0.1, 0.2, 0.3, \dots, 1.0\}$ . The associated payoffs are shown in Fig. 8. We plot data for some states of Firm 1 only. An interesting feature we can observe in this numerical experiment, possibly due to the uniqueness of equilibrium in the full-information game, is that the limits of the estimated equilibrium strategies and the estimated payoffs are independent of  $\omega$ .

## 4 Conclusions

We have shown how to estimate equilibrium strategies in a stochastic advertising game with unknown distributions of the response to advertising efforts. From the numerical results, it is worth remarking some features of our model. First, since we deal with a finite game, the equilibrium strategies are mixed instead of pure strategies—obtained in most of the deterministic differential games of advertising—because the corresponding action spaces in those models are convex. Second, the qualitative behavior of the equilibrium strategies we found corresponds to that in the existing literature, namely, for higher market shares the advertising efforts are also higher. Third, we assume that at the  $m$ -th decision epoch, firms have recorded  $m$  values of the advertising responses; hence firms have good estimators  $(\theta_n, \vartheta_n)$  only when  $m$  is large enough. However, firms can improve the estimators by using information of previous advertising campaigns as well as information acquired between decision epochs. With such improved estimators, the conclusion of Proposition 1 does not change. Finally, the problem of multiple equilibria and/or the non convergence of the estimated equilibrium strategies can be overcome by passing to a subsequence as is shown in [19].

**Acknowledgement.** This work was partially supported by Consejo Nacional de Ciencia y Tecnología (CONACYT-México) under grant Ciencia Frontera 2019-87787.

## References

1. Breton, M., Jarrar, R., Zaccour, G.: A note on feedback sequential equilibria in a Lanchester model with empirical application. *Manage. Sci.* **52**(5), 804–811 (2006)
2. Dockner, E.J., Jørgensen, S., Van Long, N., Sorger, G.: *Differential Games in Economics and Management Science*. Cambridge University Press, Cambridge (2000)
3. Dynkin, E.B., Yushkevich, A.A.: *Controlled Markov Processes*. Springer, New York (1979)
4. Feichtinger, G., Hartl, R.F., Sethi, S.P.: Dynamic optimal control models in advertising: recent developments. *Manage. Sci.* **40**(2), 195–226 (1994)

5. Ghosh, M., McDonald, D., Sinha, S.: Zero-sum stochastic games with partial information. *J. Optim. Theory Appl.* **121**(1), 99–118 (2004)
6. Huang, J., Leng, M., Liang, L.: Recent developments in dynamic advertising research. *European J. Oper. Res.* **220**, 591–609 (2012)
7. Jørgensen, S., Zaccour, G.: A survey of game-theoretic models of cooperative advertising. *European J. Oper. Res.* **237**, 1–14 (2014)
8. Jørgensen, S., Zaccour, G.: Differential Games in Marketing. Kluwer Academic Publishers, Springer, Boston (2004)
9. Judd, K.L.: Numerical Methods in Economics. MIT Press, Cambridge (1998)
10. Krausz, A., Rieder, U.: Markov games with incomplete information. *Math. Meth. Oper. Res.* **46**(2), 263–279 (1997)
11. Lanchester, F.W.: Mathematics in warfare. In: J. Newman, J. (ed.) *The World of Mathematics*, vol. 4, pp. 2138–2157. Simon and Schuster, New York (1956)
12. Luque-Vásquez, F., Minjárez-Sosa, J.A.: Empirical approximation in Markov games under unbounded payoff: discounted and average criteria. *Kybernetika* **53**(4), 694–716 (2017)
13. Minjárez-Sosa, J.A.: Zero-Sum Discrete-Time Markov Games with Unknown Disturbance Distribution. Springer, Cham (2020)
14. Minjárez-Sosa, J.A., Luque-Vásquez, F.: Two person zero-sum semi-Markov games with unknown holding times distribution on one side: a discounted payoff criterion. *Appl. Math. Optim.* **57**(3), 289–305 (2008)
15. Minjárez-Sosa, J.A., Vega-Amaya, Ó.: Asymptotically optimal strategies for adaptive zero-sum discounted Markov games. *SIAM J. Control. Optim.* **48**(3), 1405–1421 (2009)
16. Minjárez-Sosa, J.A., Vega-Amaya, Ó.: Optimal strategies for adaptive zero-sum average Markov games. *J. Math. Anal. Appl.* **402**(1), 44–56 (2013)
17. Parthasarathy, T.: Discounted, positive, and noncooperative stochastic games. *Int. J. Game Theory* **2**(1), 25–37 (1973)
18. Prasad, A., Sethi, S.P.: Competitive advertising under uncertainty: a stochastic differential game approach. *J. Optim. Theory Appl.* **123**, 163–185 (2004)
19. Robles-Aguilar, A.D., González-Sánchez, D., Minjárez-Sosa, J.A.: Empirical approximation of Nash equilibria in finite Markov games with discounted payoffs. Submitted for publication (2021)
20. Ross, S.M.: Simulation, 5th edn. Elsevier/Academic Press, Amsterdam (2013)
21. Sethi, S.P.: Dynamic optimal control models in advertising: a survey. *SIAM Rev.* **19**, 685–725 (1977)
22. Shimkin, N., Shwartz, A.: Asymptotically efficient adaptive strategies in repeated games Part I: certainty equivalence strategies. *Math. Oper. Res.* **20**, 743–767 (1995)
23. Shimkin, N., Shwartz, A.: Asymptotically efficient adaptive strategies in repeated games Part II: asymptotic optimality. *Math. Oper. Res.* **21**(2), 487–512 (1996)



# Robustness to Approximations and Model Learning in MDPs and POMDPs

Ali Devran Kara<sup>(✉)</sup> and Serdar Yüksel

Department of Mathematics and Statistics, Queen's University,  
Kingston, ON, Canada  
[{16adk,yuksel}@queensu.ca](mailto:{16adk,yuksel}@queensu.ca)

**Abstract.** In stochastic control applications, typically only an ideal model (controlled transition kernel) is assumed and the control design is based on the given model, raising the problem of performance loss due to the mismatch between the assumed model and the actual model. In some further setups, an exact model may be known, but this model may entail computationally challenging optimality analysis leading to the solution of some approximate model being implemented. With such a motivation, we study continuity properties of discrete-time stochastic control problems with respect to system models and robustness of optimal control policies designed for incorrect models applied to the true system. We study both fully observed and partially observed setups under an infinite horizon discounted expected cost criterion. We show that continuity can be established under total variation convergence of the transition kernels under mild assumptions and with further restrictions on the dynamics and observation model under weak and setwise convergence of the transition kernels. Using these, we establish convergence results and error bounds due to mismatch that occurs by the application of a control policy which is designed for an incorrectly estimated system model to the actual system, thus establishing results on robustness. These entail implications on empirical learning in (data-driven) stochastic control since often system models are learned through empirical training data where typically the weak convergence criterion applies but stronger convergence criteria do not. We finally view and establish approximation as a particular instance of robustness.

**Keywords:** Markov decision processes · Robust stochastic control · Approximate models · Empirical learning · POMDPs

**AMS(2020) subject classification:** Primary 93E20 · Secondary 90C40 · 90C39

## 1 Introduction and Problem Definition

In this article, we study the robustness problem of Markov Decision Processes (MDPs) and partially observed Markov decision processes (POMDPs) with

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

A. Piunovskiy and Y. Zhang (Eds.): *Modern Trends in Controlled Stochastic Processes*,  
ECC 41, pp. 166–191, 2021. [https://doi.org/10.1007/978-3-030-76928-4\\_9](https://doi.org/10.1007/978-3-030-76928-4_9)

incomplete/incorrect characterization, and view learning and approximate modeling as instances of the robustness problem. The article builds on some recent work of the authors but the models considered here are more general (involving changing cost functions also in the MDP models), and the unifying relationship between robustness and finite model approximations involving standard Borel models has not been studied elsewhere, to our knowledge.

Let  $\mathbb{X} \subset \mathbb{R}^m$  denote a Borel set which is the state space of a partially observed controlled Markov process. Here and throughout the paper  $\mathbb{Z}_+$  denotes the set of non-negative integers and  $\mathbb{N}$  denotes the set of positive integers. Let  $\mathbb{Y} \subset \mathbb{R}^n$  be a Borel set denoting the observation space of the model, and let the state be observed through an observation channel  $Q$ . The observation channel,  $Q$ , is defined as a stochastic kernel (regular conditional probability) from  $\mathbb{X}$  to  $\mathbb{Y}$ , such that  $Q(\cdot|x)$  is a probability measure on the (Borel)  $\sigma$ -algebra  $\mathcal{B}(\mathbb{Y})$  of  $\mathbb{Y}$  for every  $x \in \mathbb{X}$ , and  $Q(A|\cdot) : \mathbb{X} \rightarrow [0, 1]$  is a Borel measurable function for every  $A \in \mathcal{B}(\mathbb{Y})$ . A decision maker (DM) is located at the output of the channel  $Q$ , and hence it only sees the observations  $\{Y_t, t \in \mathbb{Z}_+\}$  and chooses its actions from  $\mathbb{U}$ , the action space which is a Borel subset of some Euclidean space. An *admissible policy*  $\gamma$  is a sequence of control functions  $\{\gamma_t, t \in \mathbb{Z}_+\}$  such that  $\gamma_t$  is measurable with respect to the  $\sigma$ -algebra generated by the information variables

$$I_t = \{Y_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \quad I_0 = \{Y_0\},$$

where

$$U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+, \tag{1}$$

are the  $\mathbb{U}$ -valued control actions and

$$Y_{[0,t]} = \{Y_s, 0 \leq s \leq t\}, \quad U_{[0,t-1]} = \{U_s, 0 \leq s \leq t-1\}.$$

We define  $\Gamma$  to be the set of all such admissible policies. The update rules of the system are determined by (1) and the following:

$$\Pr((X_0, Y_0) \in B) = \int_B P(dx_0) Q(dy_0|x_0), \quad B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}),$$

where  $P$  is the (prior) distribution of the initial state  $X_0$ , and

$$\begin{aligned} & \Pr\left((X_t, Y_t) \in B \mid (X, Y, U)_{[0,t-1]} = (x, y, u)_{[0,t-1]}\right) \\ &= \int_B T(dx_t|x_{t-1}, u_{t-1}) Q(dy_t|x_t), B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}), t \in \mathbb{N}, \end{aligned}$$

where  $T$  is the transition kernel of the model. The objective of the agent (decision maker) is the minimization of the infinite horizon discounted cost,

$$J_\beta(c, T, \gamma) = E_P^{T, \gamma} \left[ \sum_{t=0}^{\infty} \beta^t c(X_t, U_t) \right]$$

for some discount factor  $\beta \in (0, 1)$ , over the set of admissible policies  $\gamma \in \Gamma$ , where  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is a Borel-measurable stage-wise cost function and  $E_P^{\mathcal{T}, \gamma}$  denotes the expectation with initial state probability measure  $P$  and transition kernel  $\mathcal{T}$  under policy  $\gamma$ . Note that we write the infinite horizon discounted cost as a function of the transition kernels and the stage-wise cost function since we will analyze the cost under the changes on those variables.

We define the optimal cost for the discounted infinite horizon setup as a function of the stage-wise cost function and the transition kernels as

$$J_\beta^*(c, \mathcal{T}) = \inf_{\gamma \in \Gamma} J_\beta(c, \mathcal{T}, \gamma).$$

**Problem P1: Continuity of  $J_\beta^*(c, \mathcal{T})$  under the Convergence of the Models.** Let  $\{\mathcal{T}_n, n \in \mathbb{N}\}$  be a sequence of transition kernels which converges in some sense to another transition kernel  $\mathcal{T}$  and  $\{c_n, n \in \mathbb{N}\}$  be a sequence of stage-wise cost functions corresponding to  $\mathcal{T}_n$  which converge in some sense to another cost function  $c$ . Does that imply that

$$J_\beta^*(c_n, \mathcal{T}_n) \rightarrow J_\beta^*(c, \mathcal{T})?$$

**Problem P2: Robustness to Incorrect Models.** A problem of major practical importance is robustness of an optimal controller to modeling errors. Suppose that an optimal policy is constructed according to a model which is incorrect: how does the application of the control to the true model affect the system performance and does the error decrease to zero as the models become closer to each other? In particular, suppose that  $\gamma_n^*$  is an optimal policy designed for  $\mathcal{T}_n$  and  $c_n$ , an incorrect model for a true model  $\mathcal{T}$  and  $c$ . Is it the case that if  $\mathcal{T}_n \rightarrow \mathcal{T}$  and  $c_n \rightarrow c$ , then  $J_\beta(c, \mathcal{T}, \gamma_n^*) \rightarrow J_\beta^*(c, \mathcal{T})$ ?

**Problem P3: Empirical Consistency of Learned Probabilistic Models and Data-Driven Stochastic Control.** Let  $\mathcal{T}(\cdot|x, u)$  be a transition kernel given previous state and action variables  $x \in \mathbb{X}, u \in \mathbb{U}$ , which is unknown to the decision maker (DM). Suppose the DM builds a model for the transition kernels,  $\mathcal{T}_n(\cdot|x, u)$ , for all possible  $x \in \mathbb{X}, u \in \mathbb{U}$  by collecting training data (e.g. from the evolving system). Do we have that the cost calculated under  $\mathcal{T}_n$  converges to the true cost (i.e., do we have that the cost obtained from applying the optimal policy for the empirical model converges to the true cost as the training length increases)?

**Problem P4: Approximation by Finite MDPs as an Instance of Robustness to Incorrect Models.** Can we view the approximation problem of a continuous space MDP model with a finite model (in particular [22, Theorem 2.2], [22, Theorem 4.1] or [23, Theorem 3.2]) as an instance of the robustness problem?

**Brief Literature Review.** Robustness is a desired property for the optimal control of stochastic or deterministic systems when a given model does not reflect the actual system perfectly, as is usually the case in practice. This is a classical problem, and there is a very large literature on robust stochastic control and its application to learning-theoretic methods; see e.g. [1, 2, 7, 8, 14, 16, 18, 20, 21, 25, 26]. A rather comprehensive literature review is presented in [18]. The article builds on [16, 18], but the models considered here are more general (involving changing cost functions also in the MDP models), and the unifying relationship between robustness and finite model approximations involving standard Borel models has not been studied elsewhere, to our knowledge.

### 1.1 Some Examples and Convergence Criteria for Transition Kernels

**Convergence Criteria for Transition Kernels.** Before presenting convergence criteria for controlled transition kernels, we first review the convergence of probability measures. Three important notions of convergences for sets of probability measures to be studied in the paper are weak convergence, setwise convergence, and convergence under total variation. For  $N \in \mathbb{N}$ , a sequence  $\{\mu_n, n \in \mathbb{N}\}$  in  $\mathcal{P}(\mathbb{R}^N)$  is said to converge to  $\mu \in \mathcal{P}(\mathbb{R}^N)$  *weakly* if

$$\int_{\mathbb{R}^N} c(x) \mu_n(dx) \rightarrow \int_{\mathbb{R}^N} c(x) \mu(dx) \quad (*)$$

for every continuous and bounded  $c : \mathbb{R}^N \rightarrow \mathbb{R}$ .  $\{\mu_n\}$  is said to converge *setwise* to  $\mu \in \mathcal{P}(\mathbb{R}^N)$  if  $(*)$  holds for all measurable and bounded  $c : \mathbb{R}^N \rightarrow \mathbb{R}$ . For probability measures  $\mu, \nu \in \mathcal{P}(\mathbb{R}^N)$ , the *total variation* metric is given by

$$\|\mu - \nu\|_{TV} = 2 \sup_{B \in \mathcal{B}(\mathbb{R}^N)} |\mu(B) - \nu(B)| = \sup_{f: \|f\|_\infty \leq 1} |\int f(x) \mu(dx) - \int f(x) \nu(dx)|,$$

where the supremum is taken over all measurable real  $f$  such that  $\|f\|_\infty = \sup_{x \in \mathbb{R}^N} |f(x)| \leq 1$ . A sequence  $\{\mu_n\}$  is said to converge in total variation to  $\mu \in \mathcal{P}(\mathbb{R}^N)$  if  $\|\mu_n - \mu\|_{TV} \rightarrow 0$ . Total variation defines a stringent metric for convergence; for example, a sequence of discrete probability measures does not converge in total variation to a probability measure which admits a density function. Setwise convergence, though, induces a topology on the space of probability measures which is not metrizable [10, p. 59]. However, the space of probability measures on a complete, separable, metric (Polish) space endowed with the topology of weak convergence is itself complete, separable, and metric [19]. We also note here that relative entropy convergence, through Pinsker's inequality [11, Lemma 5.2.8], is stronger than even total variation convergence, which has also been studied in robust stochastic control. Another metric for probability measures is the Wasserstein distance: For compact spaces, the Wasserstein distance of order 1 metrizes the weak topology and for non-compact spaces convergence

in the  $W_1$  metric implies weak convergence. Considering these relations, our results in this paper can be directly generalized to the relative entropy distance or the Wasserstein distance. Building on the above, we introduce the following convergence notions for (controlled) transition kernels.

**Definition 1.** For a sequence of transition kernels  $\{\mathcal{T}_n, n \in \mathbb{N}\}$ , we say that

- $\mathcal{T}_n \rightarrow \mathcal{T}$  weakly if  $\mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u)$  weakly, for all  $x \in \mathbb{X}$  and  $u \in \mathbb{U}$ ,
- $\mathcal{T}_n \rightarrow \mathcal{T}$  setwise if  $\mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u)$  setwise, for all  $x \in \mathbb{X}$  and  $u \in \mathbb{U}$ ,
- $\mathcal{T}_n \rightarrow \mathcal{T}$  under the total variation distance if  $\mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u)$  under total variation for all  $x \in \mathbb{X}$  and  $u \in \mathbb{U}$ .

**Examples [18].** Let a controlled model be given as  $x_{t+1} = F(x_t, u_t, w_t)$ , where  $\{w_t\}$  is an i.i.d. noise process. The uncertainty on the transition kernel for such a system may arise from lack of information on  $F$  or the i.i.d. noise process  $w_t$  or both:

- (i) Let  $\{F_n\}$  denote an approximating sequence for  $F$ , so that  $F_n(x, u, w) \rightarrow F(x, u, w)$  pointwise. Assume that the probability measure of the noise is known. Then, corresponding kernels  $\mathcal{T}_n$  converge weakly to  $\mathcal{T}$ : If we denote the probability measure of  $w$  with  $\mu$ , for any  $g \in C_b(\mathbb{X})$  and for any  $(x_0, u_0) \in \mathbb{X} \times \mathbb{U}$  using the dominated convergence theorem we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \int g(x_1) \mathcal{T}_n(dx_1|x_0, u_0) &= \lim_{n \rightarrow \infty} \int g(F_n(x_0, u_0, w)) \mu(dw) \\ &= \int g(F(x_0, u_0, w)) \mu(dw) = \int g(x_1) \mathcal{T}(dx_1|x_0, u_0). \end{aligned}$$

- (ii) Much of the robust control literature deals with deterministic systems where the nominal model is a deterministic perturbation of the actual model (see e.g. [24]). The considered model is in the following form:  $\tilde{F}(x_t, u_t) = F(x_t, u_t) + \Delta F(x_t, u_t)$ , where  $F$  represents the nominal model and  $\Delta F$  is the model uncertainty satisfying some norm bounds. For such deterministic systems, pointwise convergence of  $\tilde{F}$  to the nominal model  $F$ , i.e.  $\Delta F(x_t, u_t) \rightarrow 0$ , can be viewed as weak convergence for deterministic systems by the discussion in (i). It is evident, however, that total variation convergence would be too strong for such a convergence criterion, since  $\delta_{\tilde{F}(x_t, u_t)} \rightarrow \delta_{F(x_t, u_t)}$  weakly but  $\|\delta_{\tilde{F}(x_t, u_t)} - \delta_{F(x_t, u_t)}\|_{TV} = 2$  for all  $\Delta F(x_t, u_t) \neq 0$  where  $\delta$  denotes the Dirac measure.
- (iii) Let  $F(x_t, u_t, w_t) = f(x_t, u_t) + w_t$  be such that the function  $f$  is known and  $w_t \sim \mu$  is not known correctly and an incorrect model  $\mu_n$  is assumed. If  $\mu_n \rightarrow \mu$  weakly, setwise, or in total variation, then the corresponding transition kernels  $\mathcal{T}_n$  converge in the same sense to  $\mathcal{T}$ . Observe the following:

$$\begin{aligned} &\int g(x_1) \mathcal{T}_n(dx_1|x_0, u_0) - \int g(x_1) \mathcal{T}(dx_1|x_0, u_0) \\ &= \int g(w_0 + f(x_0, u_0)) \mu_n(dw_0) - \int g(w_0 + f(x_0, u_0)) \mu(dw_0). \end{aligned} \quad (2)$$

- (a) Suppose  $\mu_n \rightarrow \mu$  weakly. If  $g$  is a continuous and bounded function, then  $g(\cdot + f(x_0, u_0))$  is a continuous and bounded function for all  $(x_0, u_0) \in \mathbb{X} \times \mathbb{U}$ . Thus, (2) goes to 0. Note that  $f$  does not need to be continuous. (b) Suppose  $\mu_n \rightarrow \mu$  setwise. If  $g$  is a measurable and bounded function, then  $g(\cdot + f(x_0, u_0))$  measurable and bounded for all  $(x_0, u_0) \in \mathbb{X} \times \mathbb{U}$ . Thus, (2) goes to 0. (c) Finally, assume  $\mu_n \rightarrow \mu$  in total variation. If  $g$  is bounded, (2) converges to 0, as in item (b). As a special case, assume that  $\mu_n$  and  $\mu$  admit densities  $h_n$  and  $h$ , respectively; then the pointwise convergence of  $h_n$  to  $h$  implies the convergence of  $\mu_n$  to  $\mu$  in total variation by Scheffé's theorem.
- (iv) Suppose now neither  $F$  nor the probability model of  $w_t$  is known perfectly. It is assumed that  $w_t$  admits a measure  $\mu_n$  and  $\mu_n \rightarrow \mu$  weakly. For the function  $F$  we again have an approximating sequence  $\{F_n\}$ . If  $F_n(x, u, w_n) \rightarrow F(x, u, w)$  for all  $(x, u) \in \mathbb{X} \times \mathbb{U}$  and for any  $w_n \rightarrow w$ , then the transition kernel  $\mathcal{T}_n$  corresponding to the model  $F_n$  converges weakly to the one of  $F$ ,  $\mathcal{T}$ : For any  $g \in C_b(\mathbb{X})$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} \int g(x_1) \mathcal{T}_n(dx_1 | x_0, u_0) &= \lim_{n \rightarrow \infty} \int g(F_n(x_0, u_0, w)) \mu_n(dw) \\ &= \int g(F(x_0, u_0, w)) \mu(dw) = \int g(x_1) \mathcal{T}(dx_1 | x_0, u_0). \end{aligned}$$

- (v) Let again  $\{F_n\}$  denote an approximating sequence for  $F$  and suppose now  $F_{x_0, u_0, n}(\cdot) := F_n(x_0, u_0, \cdot) : \mathbb{W} \rightarrow \mathbb{X}$  is invertible for all  $x_0, u_0 \in \mathbb{X} \times \mathbb{U}$  and  $F_{(x_0, u_0), n}^{-1}(\cdot)$  denotes the inverse for fixed  $(x_0, u_0)$ . It is assumed that  $F_{(x_0, u_0), n}^{-1}(x_1) \rightarrow F_{x_0, u_0}^{-1}(x_1)$  pointwise for all  $(x_0, u_0)$ . Suppose further that the noise process  $w_t$  admits a continuous density  $f_W(w)$ . The Jacobian matrix,  $\frac{\partial x_1}{\partial w}$ , is the matrix whose components are the partial derivatives of  $x_1$ , i.e. with  $x_1 \in \mathbb{X} \subset \mathbb{R}^m$  and  $w \in \mathbb{W} \subset \mathbb{R}^m$ , it is an  $m \times m$  matrix with components  $\frac{\partial(x_1)_i}{\partial w_j}$ ,  $1 \leq i, j \leq m$ . If the Jacobian matrix of derivatives  $\frac{\partial x_1}{\partial w}(w)$  is continuous in  $w$  and nonsingular for all  $w$ , then we have that the density of the state variables can be written as

$$\begin{aligned} f_{X_1, n, (x_0, u_0)}(x_1) &= f_W(F_{x_0, u_0, n}^{-1}(x_1)) \left| \frac{\partial x_1}{\partial w}(F_{x_0, u_0, n}^{-1}(x_1)) \right|^{-1}, \\ f_{X_1, (x_0, u_0)}(x_1) &= f_W(F_{x_0, u_0}^{-1}(x_1)) \left| \frac{\partial x_1}{\partial w}(F_{x_0, u_0}^{-1}(x_1)) \right|^{-1}. \end{aligned}$$

With the above,  $f_{X_1, n, (x_0, u_0)}(x_1) \rightarrow f_{X_1, (x_0, u_0)}(x_1)$  pointwise for all fixed  $(x_0, u_0)$ . Therefore, by Scheffé's theorem, the corresponding kernels  $\mathcal{T}_n(\cdot | x_0, u_0) \rightarrow \mathcal{T}(\cdot | x_0, u_0)$  in total variation for all  $(x_0, u_0)$ .

- (vi) These examples will be utilized in Sect. 5.1, where data-driven stochastic control problems will be considered where estimated models are obtained through empirical measurements of the state action variables.

## 1.2 Summary

We now introduce the main assumptions that will be occasionally used for our technical results in the article.

- Assumption 1.**(a) *The sequence of transition kernels  $T_n$  satisfies the following:  $\{T_n(\cdot|x_n, u_n), n \in \mathbb{N}\}$  converges weakly to  $\mathcal{T}(\cdot|x, u)$  for any sequence  $\{x_n, u_n\} \subset \mathbb{X} \times \mathbb{U}$  and  $x, u \in \mathbb{X} \times \mathbb{U}$  such that  $(x_n, u_n) \rightarrow (x, u)$ .*
- (b) *The stochastic kernel  $\mathcal{T}(\cdot|x, u)$  is weakly continuous in  $(x, u)$ .*
- (c) *The sequence of stage-wise cost functions  $c_n$  satisfies the following:  $c_n(x_n, u_n) \rightarrow c(x, u)$  for any sequence  $\{x_n, u_n\} \subset \mathbb{X} \times \mathbb{U}$  and  $x, u \in \mathbb{X} \times \mathbb{U}$  such that  $(x_n, u_n) \rightarrow (x, u)$ .*
- (d) *The stage-wise cost function  $c(x, u)$  is non-negative, bounded, and continuous on  $\mathbb{X} \times \mathbb{U}$ .*
- (e)  $\mathbb{U}$  is compact.

**Assumption 2.** *The observation channel  $Q(\cdot|x)$  is continuous in total variation i.e., if  $x_n \rightarrow x$ , then  $Q(\cdot|x_n) \rightarrow Q(\cdot|x)$  in total variation (only for partially observed models).*

- Assumption 3.**(a) *The sequence of transition kernels  $T_n$  satisfies the following:  $\{T_n(\cdot|x, u_n), n \in \mathbb{N}\}$  converges setwise to  $\mathcal{T}(\cdot|x, u)$  for any sequence  $\{u_n\} \subset \mathbb{U}$  and  $x, u \in \mathbb{X} \times \mathbb{U}$  such that  $u_n \rightarrow u$ .*
- (b) *The stochastic kernel  $\mathcal{T}(\cdot|x, u)$  is setwise continuous in  $u$ .*
- (c) *The sequence of stage-wise cost functions  $c_n$  satisfies the following:  $c_n(x, u_n) \rightarrow c(x, u)$  for any sequence  $\{u_n\} \subset \mathbb{U}$  and  $x, u \in \mathbb{X} \times \mathbb{U}$  such that  $u_n \rightarrow u$ .*
- (d) *The stage-wise cost function  $c(x, u)$  is non-negative, bounded, and continuous on  $\mathbb{U}$ .*
- (e)  $\mathbb{U}$  is compact.

**Assumption 4.**(a) *The sequence of transition kernels  $T_n$  satisfies the following:  $\|\mathcal{T}_n(\cdot|x, u_n) - \mathcal{T}(\cdot|x, u)\|_{TV} \rightarrow 0$  for any sequence  $\{u_n\} \subset \mathbb{U}$  and  $x, u \in \mathbb{X} \times \mathbb{U}$  such that  $u_n \rightarrow u$ .*

(b) *The stochastic kernel  $\mathcal{T}(\cdot|x, u)$  is continuous in total variation in  $u$ .*

(c) *The sequence of stage-wise cost functions  $c_n$  satisfies the following:  $c_n(x, u_n) \rightarrow c(x, u)$  for any sequence  $\{u_n\} \subset \mathbb{U}$  and  $x, u \in \mathbb{X} \times \mathbb{U}$  such that  $u_n \rightarrow u$ .*

(d) *The stage-wise cost function  $c(x, u)$  is non-negative, bounded, and continuous on  $\mathbb{U}$ .*

(e)  $\mathbb{U}$  is compact.

In Sects. 2 and 3 we study continuity (Problem P1) and robustness (Problem P2) for partially observed models. In particular we show the following:

- Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 1).
- Under Assumptions 1 and 2, continuity and robustness hold (Theorem 4, Theorem 8).
- Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 5).

- (d) Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 1).
- (f) Under Assumption 4, continuity and robustness hold (Theorem 6, Theorem 7).

In Sect. 4, we study continuity (Problem P1) and robustness (Problem P2) for fully observed models. In particular we show the following

- (a) Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 9, Example 1).
- (b) Under Assumption 1, continuity holds (Theorem 10), under Assumption 1, robustness holds if the optimal policies for every initial point are identical (Theorem 11).
- (c) Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 12, Theorem 14).
- (d) Under Assumption 3, continuity holds (Theorem 13), and under Assumption 3, robustness holds if the optimal policies for every initial point are identical (Theorem 15).
- (e) Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 1).
- (f) Under Assumption 4, continuity and robustness hold (Subsect. 4.3).

In Sect. 5, we study applications to empirical learning (in Sect. 5.1) where we establish the positive relevance of Theorem 10, and then applications to finite model approximations under the perspective of robustness in Sect. 5.2. Here, we restrict the analysis to the case with weakly continuous kernels.

## 2 Continuity of Optimal Cost in Convergence of Models (POMDP Case)

In this section, we will study continuity of the optimal discounted cost under the convergence of transition kernels and cost functions.

### 2.1 Weak Convergence

**Absence of Continuity Under Weak Convergence.** The following shows that the optimal cost may not be continuous under weak convergence of transition kernels.

**Theorem 1** [18]. *Let  $\mathcal{T}_n \rightarrow \mathcal{T}$  weakly, then it is not necessarily true that  $J_\beta^*(c, \mathcal{T}_n) \rightarrow J_\beta^*(c, \mathcal{T})$  even when the prior distributions are the same, the measurement channel  $Q$  is continuous in total variation, and  $c(x, u)$  is continuous and bounded on  $\mathbb{X} \times \mathbb{U}$ .*

We prove the result with a counterexample [18]. Letting  $\mathbb{X} = \mathbb{U} = \mathbb{Y} = [-1, 1]$  and  $c(x, u) = (x - u)^2$ , the observation channel is chosen to be uniformly distributed over  $[-1, 1]$ ,  $Q \sim U([-1, 1])$ , the initial distributions of the state

variable are chosen to be same as  $P \sim \delta_1$ , where  $\delta_x(A) := 1_{\{x \in A\}}$  for Borel  $A$ , and the transition kernels are:

$$\begin{aligned}\mathcal{T}(\cdot|x, u) &= \delta_{-1}(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] \\ &\quad + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot) \\ \mathcal{T}_n(\cdot|x, u) &= \delta_{-1}(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) \\ &\quad + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot).\end{aligned}$$

It can be seen that  $\mathcal{T}_n \rightarrow \mathcal{T}$  weakly according to Definition 1(i). Note that the cost function is continuous, and the measurement channel is continuous in total variation. The optimal discounted costs can be found as

$$\begin{aligned}J_\beta^*(c, \mathcal{T}) &= \sum_{k=1}^{\infty} E_P^\mathcal{T}[\beta^k X_k^2] = \sum_{k=1}^{\infty} \beta^k = \frac{\beta}{1-\beta} \\ J_\beta^*(c, \mathcal{T}_n) &= \sum_{k=1}^{\infty} E_P^{\mathcal{T}_n}[\beta^k X_k^2] = \beta[\frac{1}{2}(1 - \frac{1}{n})^2 + \frac{1}{2}(-1 + \frac{1}{n})^2].\end{aligned}$$

Then we have  $J_\beta^*(c, \mathcal{T}_n) \rightarrow \beta \neq \frac{\beta}{1-\beta}$ .

## 2.2 A Sufficient Condition for Continuity Under Weak Convergence

In the following, we will establish and utilize some regularity properties for the optimal cost with respect to the convergence of transition kernels.

- Assumption 5.(a)** *The stochastic kernel  $\mathcal{T}(\cdot|x, u)$  is weakly continuous in  $(x, u)$ , i.e. if  $(x_n, u_n) \rightarrow (x, u)$ , then  $\mathcal{T}(\cdot|x_n, u_n) \rightarrow \mathcal{T}(\cdot|x, u)$  weakly.*
- (b) *The observation channel  $Q(\cdot|x)$  is continuous in total variation, i.e., if  $x_n \rightarrow x$ , then  $Q(\cdot|x_n) \rightarrow Q(\cdot|x)$  in total variation.*
- (c) *The stage-wise cost function  $c(x, u)$  is non-negative, bounded and continuous on  $\mathbb{X} \times \mathbb{U}$*
- (d)  $\mathbb{U}$  is compact.

It is a well known result that, any POMDP can be reduced to a (completely observable) MDP, whose states are the posterior state distributions or *beliefs* of the observer; that is, the state at time  $t$  is  $Z_t(\cdot) := \Pr\{X_t \in \cdot | Y_0, \dots, Y_t, U_0, \dots, U_{t-1}\} \in \mathcal{P}(\mathbb{X})$ . We call this equivalent MDP the belief-MDP. The belief-MDP has state space  $Z = \mathcal{P}(\mathbb{X})$  and action space  $\mathbb{U}$ . Under the topology of weak convergence, since  $\mathbb{X}$  is a Borel space,  $Z$  is metrizable with the Prokhorov metric which makes  $Z$  into a Borel space [19]. The transition probability  $\eta$  of the belief-MDP can be constructed through non-linear filtering equations.

The one-stage cost function  $c$  of the belief-MDP is given by  $\tilde{c}(z, u) := \int_{\mathbb{X}} c(x, u) z(dx)$ . Under the regularity of the belief-MDP, we have that the *discounted cost optimality operator*  $T : C_b(Z) \rightarrow C_b(Z)$

$$(T(f))(z) = \min_u (\tilde{c}(z, u) + \beta E[f(z_1)|z_0 = z, u_0 = u]) \quad (3)$$

is a contraction from  $C_b(Z)$  to itself under the supremum norm. As a result, there exists a fixed point, the value function, and an optimal control policy exists. In view of this existence result, in the following we will consider optimal policies.

The following result is key to proving the main result of this section whose detailed analysis can be found in [18].

**Theorem 2.** Suppose we have a uniformly bounded family of functions  $\{f_n^\gamma : \mathbb{X} \rightarrow \mathbb{R}, \gamma \in \Gamma, n > 0\}$  such that  $\|f_n^\gamma\|_\infty < C$  for all  $\gamma \in \Gamma$  and for all  $n > 0$  for some  $C < \infty$ .

Further suppose we have another uniformly bounded family of functions  $\{f^\gamma : \mathbb{X} \rightarrow \mathbb{R}, \gamma \in \Gamma\}$  such that  $\|f^\gamma\|_\infty < C$  for all  $\gamma \in \Gamma$  for some  $C < \infty$ . Under the following assumptions,

(i) For any  $x_n \rightarrow x$

$$\sup_{\gamma \in \Gamma} |f_n^\gamma(x_n) - f^\gamma(x)| \rightarrow 0, \quad \sup_{\gamma \in \Gamma} |f^\gamma(x_n) - f^\gamma(x)| \rightarrow 0,$$

(ii)  $\sup_{\gamma} \rho(\mu_n^\gamma, \mu^\gamma) \rightarrow 0$  where  $\rho$  is some metric for the weak convergence topology,

we have

$$\sup_{\gamma \in \Gamma} \left| \int f_n^\gamma(x) \mu_n^\gamma(dx) - \int f^\gamma(x) \mu^\gamma(dx) \right| \rightarrow 0.$$

**Theorem 3.** Under Assumptions 1 and 2,

$$\sup_{\gamma \in \Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \rightarrow 0.$$

*Proof Sketch.*

$$\begin{aligned} & \sup_{\gamma \in \Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \\ &= \sup_{\gamma \in \Gamma} \left| \sum_{t=0}^{\infty} \beta^t \left( E_P^{\mathcal{T}_n} \left[ c_n(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{\mathcal{T}} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right) \right| \\ &\leq \sum_{t=0}^{\infty} \beta^t \sup_{\gamma \in \Gamma} \left| E_P^{\mathcal{T}_n} \left[ c_n(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{\mathcal{T}} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right|. \end{aligned}$$

Recall that an *admissible policy*  $\gamma$  is a sequence of control functions  $\{\gamma_t, t \in \mathbb{Z}_+\}$ . In the last step above, we make a slight abuse of notation; the sup at the first step is over all sequence of control functions  $\{\gamma_t, t \in \mathbb{Z}_+\}$  whereas the sup at the last step is over all sequence of control functions  $\{\gamma_{t'}, t' \leq t\}$ , but we will use the same notation,  $\gamma$ , in the rest of the proof.

For any  $\epsilon > 0$ , we choose a  $K < \infty$  such that  $\sum_{t=K+1}^{\infty} \beta^t 2\|c\|_\infty \leq \epsilon/2$ . For the chosen  $K$ , we choose an  $N < \infty$  such that

$$\sup_{\gamma \in \Gamma} \left| E_P^{\mathcal{T}_n} \left[ c_n(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{\mathcal{T}} \left[ c(X_t, \gamma(Y_{[0,t]})) \right] \right| \leq \epsilon/2K$$

for all  $t \leq K$  and for all  $n > N$ . We note that in [18] a fixed  $c$  function was considered, but by considering the additional term

$$\sup_{\gamma \in \Gamma} \left| E_P^{\mathcal{T}_n} \left[ c_n(X_t, \gamma(Y_{[0,t]})) \right] - E_P^{\mathcal{T}} \left[ c_n(X_t, \gamma(Y_{[0,t]})) \right] \right|$$

and noting that  $\sup_{\gamma} |\int Q(dy|x_n) c_n(x_n, \gamma(y)) - \int Q(dy|x) c(x, \gamma(y))| \rightarrow 0$ , for every  $x_n \rightarrow x$ , by a generalized dominated convergence theorem as  $Q$  is continuous in total variation, a triangle inequality argument shows that the same result applies. This follows from a generalized dominated convergence theorem as stated in Theorem 2 whose detailed analysis can be found in [18]. Thus,  $\sup_{\gamma \in \Gamma} |J_{\beta}(c_n, \mathcal{T}_n, \gamma) - J_{\beta}(c, \mathcal{T}, \gamma)| \rightarrow 0$  as  $n \rightarrow \infty$ .  $\square$

**Theorem 4.** *Suppose the conditions of Theorem 3 hold. Then  $\lim_{n \rightarrow \infty} |J_{\beta}^*(c_n, \mathcal{T}_n) - J_{\beta}^*(c, \mathcal{T})| = 0$ .*

*Proof Sketch.* We start with the following bound:

$$\begin{aligned} & |J_{\beta}^*(c_n, \mathcal{T}_n) - J_{\beta}^*(c, \mathcal{T})| \\ & \leq \max \left( J_{\beta}(c_n, \mathcal{T}_n, \gamma^*) - J_{\beta}(c, \mathcal{T}, \gamma^*), J_{\beta}(c, \mathcal{T}, \gamma_n^*) - J_{\beta}(c_n, \mathcal{T}_n, \gamma_n^*) \right), \end{aligned} \quad (4)$$

where  $\gamma^*$  and  $\gamma_n^*$  are the optimal policies, respectively, for  $\mathcal{T}$  and  $\mathcal{T}_n$ . Both terms go to 0 by Theorem 3.  $\square$

## 2.3 Absence of Continuity Under Setwise Convergence

We now show that continuity of optimal costs may fail under the setwise convergence of transition kernels. Theorem 12 in the next section establishes this result for fully observed models, which serves as a proof for this setup also.

**Theorem 5.** *Let  $\mathcal{T}_n \rightarrow \mathcal{T}$  setwise. Then, it is not true in general that  $J_{\beta}^*(c, \mathcal{T}_n) \rightarrow J_{\beta}^*(c, \mathcal{T})$ , even when  $\mathbb{X}, \mathbb{Y}$ , and  $\mathbb{U}$  are compact and  $c(x, u)$  is continuous and bounded in  $\mathbb{X} \times \mathbb{U}$ .*

## 2.4 Continuity Under Total Variation

**Theorem 6.** *Under Assumption 4,  $J_{\beta}^*(c_n, \mathcal{T}_n) \rightarrow J_{\beta}^*(c, \mathcal{T})$ .*

*Proof Sketch.* We start with the following bound:

$$\begin{aligned} |J_{\beta}^*(c_n, \mathcal{T}_n) - J_{\beta}^*(c, \mathcal{T})| & \leq \max \left( J_{\beta}(c_n, \mathcal{T}_n, \gamma^*) - J_{\beta}(c, \mathcal{T}, \gamma^*), J_{\beta}(c_n, \mathcal{T}_n, \gamma_n^*) \right. \\ & \quad \left. - J_{\beta}(c, \mathcal{T}, \gamma_n^*) \right), \end{aligned}$$

where  $\gamma^*$  and  $\gamma_n^*$  are the optimal policies, respectively, for  $\mathcal{T}$  and  $\mathcal{T}_n$ .

We now study the following:

$$\begin{aligned} & \sup_{\gamma \in \Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \\ &= \sup_{\gamma \in \Gamma} \left| \sum_{t=0}^{\infty} \beta^t \left( E_P^{\mathcal{T}_n} [c_n(X_t, \gamma(Y_{[0,t]}))] - E_P^{\mathcal{T}} [c(X_t, \gamma(Y_{[0,t]}))] \right) \right| \\ &\leq \sum_{t=0}^{\infty} \beta^t \sup_{\gamma \in \Gamma} \left| E_P^{\mathcal{T}_n} [c_n(X_t, \gamma(Y_{[0,t]}))] - E_P^{\mathcal{T}} [c(X_t, \gamma(Y_{[0,t]}))] \right|. \end{aligned}$$

It can be shown that [18]

$$\sup_{\gamma \in \Gamma} \left| E_P^{\mathcal{T}_n} [c_n(X_t, \gamma(Y_{[0,t]}))] - E_P^{\mathcal{T}} [c(X_t, \gamma(Y_{[0,t]}))] \right| \rightarrow 0. \quad (5)$$

This was shown in [18] for fixed  $c$ . The extension to varying  $c_n$  follows from a triangle inequality step with the assumption that  $\mathcal{T}_n(\cdot|x, u_n) \rightarrow \mathcal{T}(\cdot|x, u)$  setwise, and  $c_n(x, u_n) \rightarrow c(x, u)$  for any  $u_n \rightarrow u$ . Therefore, using identical steps as in the proof of Theorem 3 we have  $\sup_{\gamma \in \Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \rightarrow 0$ .  $\square$

### 3 Robustness to Incorrect Models (POMDP Case)

Here, we consider the robustness problem **P2**: Suppose we design an optimal policy,  $\gamma_n^*$ , for a transition kernel,  $\mathcal{T}_n$  and a cost function  $c_n$ , assuming they are the correct model and apply the policy to the true model whose transition kernel is  $\mathcal{T}$  and whose cost function is  $c$ . We study the robustness of the sub-optimal policy  $\gamma_n^*$ .

#### 3.1 Total Variation

The next theorem gives an asymptotic robustness result.

**Theorem 7.** *Under Assumption 4*

$$|J_\beta(c_n, \mathcal{T}, \gamma_n^*) - J_\beta^*(c, \mathcal{T})| \rightarrow 0,$$

where  $\gamma_n^*$  is the optimal policy designed for the kernel  $\mathcal{T}_n$ .

*Proof Sketch.* We write the following:

$$|J_\beta(c, \mathcal{T}, \gamma_n^*) - J_\beta^*(c, \mathcal{T})| \leq |J_\beta(c, \mathcal{T}, \gamma_n^*) - J_\beta^*(c_n, \mathcal{T}_n)| + |J_\beta^*(c_n, \mathcal{T}_n) - J_\beta^*(c, \mathcal{T})|.$$

Both terms can be shown to go to 0 using (5).  $\square$

#### 3.2 Setwise Convergence

Theorem 14 in the next section establishes the lack of robustness under the setwise convergence of kernels. As we note later, a fully observed system can be viewed as a partially observed system with the measurement being the state itself, (see (6)).

### 3.3 Weak Convergence

**Theorem 8.** Under Assumptions 1 and 2,  $|J_\beta(c, \mathcal{T}, \gamma_n^*) - J_\beta^*(c, \mathcal{T})| \rightarrow 0$ , where  $\gamma_n^*$  is the optimal policy designed for the transition kernel  $\mathcal{T}_n$ .

*Proof Sketch.* We write

$$\begin{aligned} |J_\beta(c, \mathcal{T}, \gamma_n^*) - J_\beta^*(c, \mathcal{T})| &\leq |J_\beta(c, \mathcal{T}, \gamma_n^*) - J_\beta(c_n, \mathcal{T}_n, \gamma_n^*)| + |J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \\ &\quad - J_\beta(\mathcal{T}, \gamma^*)|. \end{aligned}$$

The first term goes to 0 by Theorem 3. For the second term we use Theorem 4.  $\square$

## 4 Continuity and Robustness in the Fully Observed Case

In this section, we consider the fully observed case where the controller has direct access to the state variables. We present the results for this case separately, since here we cannot utilize the regularity properties of measurement channels which allows for stronger continuity and robustness results. Under measurable selection conditions due to weak or strong (setwise) continuity of transition kernels [13, Section 3.3], for infinite horizon discounted cost problems optimal policies can be selected from those which are stationary and deterministic. Therefore we will restrict the policies to be stationary and deterministic so that  $U_t = \gamma(X_t)$  for some measurable function  $\gamma$ . Notice also that fully observed models can be viewed as partially observed with the measurement channel thought to be

$$Q(\cdot|x) = \delta_x(\cdot), \tag{6}$$

which is only weakly continuous, thus it does not satisfy Assumption 2.

### 4.1 Weak Convergence

**Absence of Continuity Under Weak Convergence.** We start with a negative result.

**Theorem 9.** For  $\mathcal{T}_n \rightarrow \mathcal{T}$  weakly, it is not necessarily true that  $J_\beta^*(c, \mathcal{T}_n) \rightarrow J_\beta^*(c, \mathcal{T})$  even when the prior distributions are the same and  $c(x, u)$  is continuous and bounded in  $\mathbb{X} \times \mathbb{U}$ .

*Proof.* We prove the result with a counterexample, similar to the model used in the proof of Theorem 1. Letting  $\mathbb{X} = [-1, 1]$ ,  $\mathbb{U} = \{-1, 1\}$  and  $c(x, u) = (x - u)^2$ , the initial distributions are given by  $P \sim \delta_1$ , that is,  $X_0 = 1$ , and the transition kernels are

$$\begin{aligned} \mathcal{T}(\cdot|x, u) &= \delta_{-1}(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] \\ &\quad + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot), \end{aligned}$$

$$\begin{aligned} \mathcal{T}_n(\cdot|x, u) &= \delta_{-1}(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) \\ &\quad + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + (1 - \delta_{-1}(x))(1 - \delta_1(x))\delta_0(\cdot). \end{aligned}$$

It can be seen that  $\mathcal{T}_n \rightarrow \mathcal{T}$  weakly according to Definition 1(i). Under this setup we can calculate the optimal costs as follows:

$$J_\beta^*(c, \mathcal{T}_n) = \frac{1}{n^2} + \sum_{k=2}^{\infty} \beta^k = \frac{1}{n^2} + \frac{\beta^2}{1-\beta},$$

and  $J_\beta^*(c, \mathcal{T}) = 0$ . Thus, continuity does not hold.  $\square$

We now present another counter example emphasizing the importance of continuous convergence in the actions. The following counter example shows that without the continuous convergence and regularity assumptions on the kernel  $\mathcal{T}$ , continuity fails even when  $\mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u)$  pointwise (for  $x, u$ ) in total variation (also setwise and weakly) and even when the cost function  $c(x, u)$  is continuous and bounded. Notice that this example also holds for both setwise and weak convergence.

*Example 1.* Assume that the kernels are given by

$$\begin{aligned} \mathcal{T}_n(\cdot|x, u) &\sim U([u^n, 1+u^n]), \\ \mathcal{T}(\cdot|x, u) &\sim \begin{cases} U([0, 1]) & \text{if } u \neq 1, \\ U([1, 2]) & \text{if } u = 1, \end{cases} \end{aligned}$$

where  $\mathbb{U} = [0, 1]$  and  $\mathbb{X} = \mathbb{R}$ . We note first that  $\mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u)$  in total variation for every fixed  $x$  and  $u$ .

The cost function is in the following form:

$$c(x, u) = \begin{cases} 2 & \text{if } x \leq \frac{1}{e}, \\ 2 - \frac{x - \frac{1}{e}}{0.1} & \text{if } \frac{1}{e} < x \leq 0.1 + \frac{1}{e}, \\ 1 & \text{if } 0.1 + \frac{1}{e} < x \leq 1 + \frac{1}{e} - 0.1, \\ 2 - \frac{1 + \frac{1}{e} - x}{0.1} & \text{if } 1 + \frac{1}{e} - 0.1 < x \leq 1 + \frac{1}{e}, \\ 2 & \text{if } 1 + \frac{1}{e} < x. \end{cases}$$

Notice that  $c(x, u)$  is a continuous function.

With this setup,  $\gamma^*(x) = 0$  is an optimal policy for  $\mathcal{T}$  since on the  $[0, 1]$  interval the induced cost is less than the cost induced on the  $[1, 2]$  interval. The cost under this policy is

$$J_\beta^*(c, \mathcal{T}) = \sum_{t=0}^{\infty} \beta^t \left( 2 \times \frac{1}{e} + \frac{0.3}{2} + 0.9 - \frac{1}{e} \right) = \frac{1}{1-\beta} \left( 1.05 + \frac{1}{e} \right).$$

For  $\mathcal{T}_n$ ,  $\gamma_n^*(x) = e^{-\frac{1}{n}}$  is an optimal policy for every  $n$  as  $e^{-\frac{1}{n} \times n} = \frac{1}{e}$  and thus the state is distributed between  $\frac{1}{e} < x \leq 1 + \frac{1}{e}$  in which interval the cost is the least. Hence, we can write

$$\begin{aligned} \lim_{n \rightarrow \infty} J_\beta(c, \mathcal{T}_n, \gamma_n^*) &= \sum_{t=0}^{\infty} \beta^t \left( 0.3 + 1 - 0.2 \right) = \frac{1.1}{1-\beta} \neq \frac{1}{1-\beta} \left( 1.05 + \frac{1}{e} \right) \\ &= J_\beta^*(c, \mathcal{T}). \end{aligned}$$

**A Sufficient Condition for Continuity Under Weak Convergence.** We will now establish that if the kernels and the model components have some further regularity, continuity does hold. The assumptions of the following result are the same as the assumptions for the partially observed case (Theorem 4) except for the assumption on the measurement channel  $Q$ .

**Theorem 10.** *Under Assumption 1,  $J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \rightarrow J_\beta(c, \mathcal{T}, \gamma^*)$  for any initial state  $x_0$ , as  $n \rightarrow \infty$ .*

*Proof.* We will use the successive approximations for an inductive argument.

Recall *discounted cost optimality operator*  $T : C_b(Z) \rightarrow C_b(Z)$  from (3)

$$(T(v))(x) = \inf_{u \in \mathbb{U}} \left( c(x, u) + \beta E[v(x_1)|x_0 = x, u_0 = u] \right),$$

which is a contraction from  $C_b(\mathbb{X})$  to itself under the supremum norm and has a fixed point, the value function. For the kernel  $\mathcal{T}$ , we will denote the approximation functions by

$$v^k(x) = T(v^{k-1})(x),$$

and for the kernel  $\mathcal{T}_n$  we will use  $v_n^k(x)$  to denote the approximation functions, notice that the operator  $T$  also depends on  $n$  for the model  $\mathcal{T}_n$ , but we will continue using it as  $T$  in what follows.

We wish to show that the approximation functions for  $\mathcal{T}_n$  continuously converge to the ones for  $\mathcal{T}$ . Then, for the first step of the induction we have

$$v^1(x) = c(x, u^*), \quad v_n^1(x_n) = c_n(x_n, u_n^*),$$

and thus we can write,

$$|v^1(x) - v_n^1(x_n)| \leq \sup_{u \in \mathbb{U}} |c(x, u) - c_n(x_n, u)|$$

since  $c_n(x_n, u_n) \rightarrow c(x, u)$  for all  $(x_n, u_n) \rightarrow (x, u)$  and the action space,  $\mathbb{U}$ , is compact, the first step of the induction holds, i.e.  $\lim_{n \rightarrow \infty} |v^1(x) - v_n^1(x_n)| = 0$ .

For the  $k^{th}$  step we have

$$\begin{aligned} v^k(x) &= T(v^{k-1})(x) = \inf_u [c(x, u) + \beta \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x, u)], \\ v_n^k(x_n) &= T(v_n^{k-1})(x_n) = \inf_u [c_n(x_n, u) + \beta \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n, u)]. \end{aligned}$$

Note that the assumptions of the theorem satisfy the measurable selection criteria and hence we can choose minimizing selectors [13, Section 3.3]. If we denote

the selectors by  $u^*$  and  $u_n^*$ , we can write

$$\begin{aligned} & |v^k(x) - v_n^k(x_n)| \\ & \leq \max \left( \left[ |c(x, u^*) - c_n(x_n, u^*)| \right. \right. \\ & \quad + \beta \left| \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x, u^*) - \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n, u^*) \right| \left. \right] \\ & \quad \left[ |c(x, u_n^*) - c_n(x_n, u_n^*)| \right. \\ & \quad + \beta \left| \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x, u_n^*) - \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n, u_n^*) \right| \left. \right] \). \end{aligned}$$

Hence, we can write

$$\begin{aligned} & |v^k(x) - v_n^k(x_n)| \\ & \leq \sup_{u \in \mathbb{U}} \left[ |c(x, u) - c_n(x_n, u)| \right. \\ & \quad + \beta \left| \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x, u) - \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n, u) \right| \left. \right], \end{aligned} \tag{7}$$

above, the first term goes to 0 as  $c_n(x_n, u_n) \rightarrow c(x, u)$  for all  $(x_n, u_n) \rightarrow (x, u)$  and the action space,  $\mathbb{U}$ , is compact. For the second term we write,

$$\begin{aligned} & \sup_{u \in \mathbb{U}} \left| \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x, u) - \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n, u) \right| \\ & \leq \sup_{u \in \mathbb{U}} \left| \int_{\mathbb{X}} (v^{k-1}(x^1) - v_n^{k-1}(x^1)) \mathcal{T}_n(dx^1|x_n, u) \right| \\ & \quad + \sup_{u \in \mathbb{U}} \left| \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x, u) - \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n, u) \right| \end{aligned}$$

above, for the first term, by the induction argument for any  $x_n^1 \rightarrow x^1$ ,  $|v^{k-1}(x^1) - v_n^{k-1}(x_n^1)| \rightarrow 0$  (i.e., we have continuous convergence). We also have that  $\mathcal{T}_n(\cdot|x_n, u) \rightarrow \mathcal{T}(\cdot|x, u)$  weakly uniformly over  $u \in \mathbb{U}$  as  $\mathbb{U}$  is compact. Therefore, using Theorem 2 the first term goes to 0. For the second term we again use that  $\mathcal{T}_n(\cdot|x_n, u)$  converges weakly to  $\mathcal{T}(\cdot|x, u)$  uniformly over  $u \in \mathbb{U}$ . With an almost identical induction argument it can also be shown that  $v^{k-1}(x^1)$  is continuous in  $x^1$ , thus the second term also goes to 0.

So far, we have showed that for any  $k \in \mathbb{N}$ ,  $\lim_{n \rightarrow \infty} |v_n^k(x_n) - v^k(x)| = 0$  for any  $x_n \rightarrow x$ , in particular it is also true that  $\lim_{n \rightarrow \infty} |v_n^k(x) - v^k(x)| = 0$  for any  $x$ .

As we have stated earlier, it can be shown that the approximation operator,  $T$  is a contractive operator under supremum norm with modulus  $\beta$  and it converges

to a fixed point which is the value function. Thus, we have

$$|J_\beta(c, \mathcal{T}, \gamma^*) - v^k(x)| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}, \quad |J_\beta^*(c_n, \mathcal{T}_n, \gamma_n^*) - v_n^k(x)| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}. \quad (8)$$

Combining the results,

$$\begin{aligned} |J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) - J_\beta(c, \mathcal{T}, \gamma^*)| &\leq |J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) - v_n^k(x)| + |v_n^k(x) - v^k(x)| \\ &\quad + |J_\beta(c, \mathcal{T}, \gamma^*) - v^k(x)|. \end{aligned}$$

Note that the first and the last term can be made arbitrarily small since (8) holds for all  $k \in \mathbb{N}$ ; the second term goes to 0 with an inductive argument for all  $k \in \mathbb{N}$ .  $\square$

**A Sufficient Condition for Robustness Under Weak Convergence.** We now present a result that establishes robustness if the optimal policies for every initial point are identical. That is, for every  $n$ ,  $\gamma_n^*$  is optimal for every  $x_0 \in \mathbb{X}$  (under the model  $\mathcal{T}_n$ ). A sufficient condition for this property is that  $\gamma_n^*$  solves the discounted cost optimality equation (DCOE) for every initial point.

A policy  $\gamma^* \in \Gamma$  solves the discounted cost optimality equation and is optimal if it satisfies

$$J_\beta^*(c, \mathcal{T}, x) = c(x, \gamma^*(x)) + \beta \int J_\beta^*(c, \mathcal{T}, x_1) \mathcal{T}(dx_1|x, \gamma^*(x)).$$

Thus, a policy is optimal for every initial point if it satisfies the DCOE for all initial points  $x \in \mathbb{X}$ . The following generalizes [18].

**Theorem 11.** *Under Assumption 1,  $J_\beta(c, \mathcal{T}, \gamma_n^*) \rightarrow J_\beta(c, \mathcal{T}, \gamma^*)$  for any initial point  $x_0$  if  $\gamma_n^*$  is optimal for any initial point for the kernel  $\mathcal{T}_n$  and for the stage-wise cost function  $c_n$ .*

*Remark 1.* For the partially observed case, the proof approach we use makes use of policy exchange (e.g. (4)) and for this approach the total variation continuity of channel  $Q(\cdot|x)$  is a key step to deal with the uniform convergence over policies. As we stated before, the channel for fully observed models can be considered in the form of (6) which is only weakly continuous and not continuous in total variation. Thus, obtaining a result uniformly over all policies may not be possible. However, for the fully observed models we can reach continuity and robustness (Theorem 10, Theorem 11) using a value iteration approach. With this approach, instead of exchanging policies and analyzing uniform convergence over all policies, we can exchange control actions (e.g. (7)) and analyze uniform convergence over the action space  $\mathbb{U}$  by using the discounted optimality operator (3). Hence, we are only able to show convergence over optimal policies for the fully observed case, i.e.  $J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \rightarrow J_\beta(c, \mathcal{T}, \gamma^*)$  or  $J_\beta(c, \mathcal{T}, \gamma_n^*) \rightarrow J_\beta(c, \mathcal{T}, \gamma^*)$  where  $\gamma_n^*$  and  $\gamma^*$  are optimal policies, whereas, for partially observed models, regularity of the channel allows us to show convergence over any sequence of policies, i.e.  $\sup_{\gamma \in \Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \rightarrow 0$ .

*Remark 2.* As we have discussed in Subsect. 2.2, a partially observed model can be reduced to a fully observed process where the state process (beliefs) becomes probability measure valued. Consider the partially observed models with transition kernels  $\mathcal{T}_n$  and  $\mathcal{T}$  (with a channel  $Q$ ) and their corresponding fully observed transition kernels  $\eta_n$  and  $\eta$ : following the discussions and techniques in [9] and [15], one can show that  $\eta_n$  and  $\eta$  satisfy the conditions of Theorem 11 and Theorem 10 that is  $\eta_n(\cdot|z_n, u_n) \rightarrow \eta(\cdot|z, u)$  for any  $(z_n, u_n) \rightarrow (z, u)$  under the following set of assumptions

- $\mathcal{T}_n(\cdot|x_n, u_n) \rightarrow \mathcal{T}(\cdot|x, u)$  for any  $(x_n, u_n) \rightarrow (x, u)$ ,
- $Q(\cdot|x)$  is continuous on total variation in  $x$ .

We remark that these conditions also agree with the conditions presented for continuity and robustness of the partially observed models (Theorem 4 and Theorem 8).

## 4.2 Setwise Convergence

**Absence of Continuity Under Setwise Convergence.** We give a negative result similar to Theorem 5, via Example 1:

**Theorem 12.** *Letting  $\mathcal{T}_n \rightarrow \mathcal{T}$  setwise, then it is not necessarily true that  $J_\beta^*(c, \mathcal{T}_n) \rightarrow J_\beta^*(c, \mathcal{T})$  even when  $c(x, u)$  is continuous and bounded in  $\mathbb{X} \times \mathbb{U}$ .*

### A Sufficient Condition for Continuity Under Setwise Convergence.

**Theorem 13.** *Under Assumption 3  $J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \rightarrow J_\beta(c, \mathcal{T}, \gamma^*)$ , for any initial state  $x_0$ , as  $n \rightarrow \infty$ .*

*Proof.* We use the same value iteration technique that we used to prove Theorem 10. See [18].  $\square$

**Absence of Robustness Under Setwise Convergence.** Now, we give a result showing that even if the continuity holds under the setwise convergence of the kernels, the robustness may not be satisfied (see [18, Theorem 4.7]).

**Theorem 14.** *Supposing  $\mathcal{T}_n(\cdot|x_n, u_n) \rightarrow \mathcal{T}(\cdot|x, u)$  setwise for every  $x \in \mathbb{X}$  and  $u \in \mathbb{U}$  and  $(x_n, u_n) \rightarrow (x, u)$ , then it is not true in general that  $J_\beta(c, \mathcal{T}, \gamma_n^*) \rightarrow J_\beta(c, \mathcal{T}, \gamma^*)$ , even when  $\mathbb{X}$  and  $\mathbb{U}$  are compact and  $c(x, u)$  is continuous and bounded in  $\mathbb{X} \times \mathbb{U}$ .*

**A Sufficient Condition for Robustness Under Setwise Convergence.** We now present a similar result to Theorem 11 that is we show that under the conditions of Theorem 13, if further for every  $n$ ,  $\gamma_n^*$  is optimal for every  $x_0 \in \mathbb{X}$  (under the model  $\mathcal{T}_n$ ) then robustness holds under setwise convergence.

**Theorem 15.** *Supposing Assumption 3 holds, if further we have that for every  $n$ ,  $\gamma_n^*$  is optimal for every  $x_0 \in \mathbb{X}$  (under the model  $\mathcal{T}_n$ ) then  $J_\beta(c, \mathcal{T}, \gamma_n^*) \rightarrow J_\beta(c, \mathcal{T}, \gamma^*)$ .*

### 4.3 Total Variation

The continuity result in Theorem 6 and the robustness result in Theorem 7 apply to this case since the fully observed model may be viewed as a partially observed model with the measurement channel  $Q$  given in (6).

## 5 Applications to Data-Driven Learning and Finite Model Approximations

### 5.1 Application of Robustness Results to Data-Driven Learning

In practice, one may estimate the kernel of a controlled Markov chain using empirical data; see e.g. [3, 12] for some related literature in the control-free and controlled contexts.

Let us briefly review the basic case where an i.i.d. sequence of random variables is repeatedly observed, but its probability measure is not known apriori. Let  $\{(X_i), i \in \mathbb{N}\}$  be an  $\mathbb{X}$ -valued i.i.d. random variable sequence generated according to some distribution  $\mu$ . Defining for every (fixed) Borel  $B \subset \mathbb{X}$ , and  $n \in \mathbb{N}$ , the empirical occupation measures  $\mu_n(B) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \in B\}}$ , one has  $\mu_n(B) \rightarrow \mu(B)$  almost surely by the strong law of large numbers. It then follows that  $\mu_n \rightarrow \mu$  weakly with probability one [6, Theorem 11.4.1]. However,  $\mu_n$  does not converge to  $\mu$  in total variation or setwise, in general. On the other hand, if we know that  $\mu$  admits a density, we can find estimators to estimate  $\mu$  under total variation [5, Chapter 3]. For a more detailed discussion, see [17, pp. 1950–1951]. In the previous sections, we established robustness results under the convergence of transition kernels in the topology of weak convergence and total variation. We build on these observations.

**Corollary 1 (to Theorem 6 and Theorem 7).** *Suppose we are given the following dynamics for finite state space,  $\mathbb{X}$ , and finite action space,  $\mathbb{U}$ ,*

$$x_{t+1} = f(x_t, u_t, w_t), \quad y_t = g(x_t, v_t)$$

*where  $\{w_t\}$  and  $\{v_t\}$  are i.i.d. noise processes and the noise models are unknown. Suppose that there is an initial training period so that under some policy, every  $x, u$  pair is visited infinitely often if training were to continue indefinitely, but that the training ends at some finite time. Let us assume that, through this training, we empirically learn the transition dynamics such that for every (fixed) Borel  $B \subset \mathbb{X}$ , for every  $x \in \mathbb{X}$ ,  $u \in \mathbb{U}$  and  $n \in \mathbb{N}$ , the empirical occupation measures are*

$$\mathcal{T}_n(B|x_0 = x, u_0 = u) = \frac{\sum_{i=1}^n 1_{\{X_i \in B, X_{i-1} = x, U_{i-1} = u\}}}{\sum_{i=1}^n 1_{\{X_{i-1} = x, U_{i-1} = u\}}}.$$

*Then we have that  $J_\beta^*(\mathcal{T}_n) \rightarrow J_\beta^*(\mathcal{T})$  and  $J_\beta(\mathcal{T}, \gamma_n^*) \rightarrow J_\beta^*(\mathcal{T})$ , where  $\gamma_n^*$  is the optimal policy designed for  $\mathcal{T}_n$ . Since the channel model  $g$  has no restrictions, this result also applies to the fully observed model setup by taking  $g(x_t, v_t) = x_t$ .*

*Proof.* We have that  $\mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u)$  weakly for every  $x \in \mathbb{X}$ ,  $u \in \mathbb{U}$  almost surely by law of large numbers. Since the spaces are finite, we also have  $\mathcal{T}_n(\cdot|x, u) \rightarrow \mathcal{T}(\cdot|x, u)$  under total variation. By Theorem 6 and Theorem 7, the results follow.  $\square$

The following holds for more general spaces.

**Corollary 2 (to Theorems 8, 4, 10 and 11).** *Suppose we are given the following dynamics with state space  $\mathbb{X}$  and action space  $\mathbb{U}$ ,*

$$x_{t+1} = f(x_t, u_t, w_t), \quad y_t = g(x_t, v_t),$$

where  $\{w_t\}$  and  $\{v_t\}$  are i.i.d. noise processes and the noise models are unknown. Suppose that  $f(x, u, \cdot) : \mathbb{W} \rightarrow \mathbb{X}$  is invertible for all fixed  $(x, u)$  and  $f(x, u, w)$  is continuous and bounded on  $\mathbb{X} \times \mathbb{U} \times \mathbb{W}$ . We construct the empirical measures for the noise process  $w_t$  such that for every (fixed) Borel  $B \subset \mathbb{W}$ , and for every  $n \in \mathbb{N}$ , the empirical occupation measures are

$$\mu_n(B) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{f_{x_{i-1}, u_{i-1}}^{-1}(x_i) \in B\}} \quad (9)$$

where  $f_{x_{i-1}, u_{i-1}}^{-1}(x_i)$  denotes the inverse of  $f(x_{i-1}, u_{i-1}, w) : \mathbb{W} \rightarrow \mathbb{X}$  for given  $(x_{i-1}, u_{i-1})$ . Using the noise measurements, we construct the empirical transition kernel estimates for any  $(x_0, u_0)$  and Borel  $B$  as

$$\mathcal{T}_n(B|x_0, u_0) = \mu_n(f_{x_0, u_0}^{-1}(B)).$$

- (i) If the measurement channel (represented by the function  $g$ ) is continuous in total variation then  $J_\beta^*(\mathcal{T}_n) \rightarrow J_\beta^*(\mathcal{T})$  and  $J_\beta(\mathcal{T}, \gamma_n^*) \rightarrow J_\beta^*(\mathcal{T})$ , where  $\gamma_n^*$  is the optimal policy designed for  $\mathcal{T}_n$  for all initial points.
- (ii) If the measurement channel is in the form  $g(x_t, v_t) = x_t$  (i.e. fully observed) then  $J_\beta^*(\mathcal{T}_n) \rightarrow J_\beta^*(\mathcal{T})$  and if further for every  $n$ ,  $\gamma_n^*$  is optimal for every  $x_0 \in \mathbb{X}$  (under the model  $\mathcal{T}_n$ ) then  $J_\beta(\mathcal{T}, \gamma_n^*) \rightarrow J_\beta^*(\mathcal{T})$ .

*Proof.* We have  $\mu_n \rightarrow \mu$  weakly with probability one where  $\mu$  is the model. We claim that the transition kernels are such that  $\mathcal{T}_n(\cdot|x_n, u_n) \rightarrow \mathcal{T}(\cdot|x, u)$  weakly for any  $(x_n, u_n) \rightarrow (x, u)$ . To see that observe the following for  $h \in C_b(\mathbb{X})$

$$\begin{aligned} & \int h(x_1) \mathcal{T}_n(dx_1|x_n, u_n) - \int h(x_1) \mathcal{T}(dx_1|x, u) \\ &= \int h(f(x_n, u_n, w)) \mu_n(dw) - \int h(f(x, u, w)) \mu(dw) \rightarrow 0, \end{aligned}$$

where  $\mu_n$  is the empirical measure for  $w_t$  and  $\mu$  is the true measure again. For the last step, we used that  $\mu_n \rightarrow \mu$  weakly and  $h(f(x_n, u_n, w))$  continuously converge to  $h(f(x, u, w))$  i.e.  $h(f(x_n, u_n, w_n)) \rightarrow h(f(x, u, w))$  for some  $w_n \rightarrow w$  since  $f$  and  $g$  are continuous functions. Similarly, it can be also shown that  $\mathcal{T}_n(\cdot|x, u)$  and  $\mathcal{T}(\cdot|x, u)$  are weakly continuous on  $(x, u)$ . Thus, for the case where the channel is

continuous in total variation by Theorem 8 and Theorem 4 if  $c(x, u)$  is bounded and  $\mathbb{U}$  is compact the result follows.

For the fully observed case,  $J_\beta^*(\mathcal{T}_n) \rightarrow J_\beta^*(\mathcal{T})$  by Theorem 10 and  $J_\beta(\mathcal{T}, \gamma_n^*) \rightarrow J_\beta^*(\mathcal{T})$  by Theorem 11.  $\square$

*Remark 3.* We note here that the moment estimation method can also lead to consistency. Suppose that the distribution of  $W$  is determined by its moments, such that estimate models  $W_n$  have moments of all orders and  $\lim_n = E[W_n^r] = E[W^r]$  for all  $r \in \mathbb{Z}_+$ . Then, we have that [4, Theorem 30.2]  $W_n \rightarrow W$  weakly and thus  $\mathcal{T}_n(\cdot|x_n, u_n) \rightarrow \mathcal{T}(\cdot|x, u)$  weakly for any  $(x_n, u_n) \rightarrow (x, u)$  under the assumptions of above corollary. Hence, we reach continuity and robustness using the same arguments as in the previous result (Corollary 2).

Now, we give a similar result with the assumption that the noise process of the dynamics admits a continuous probability density function.

**Corollary 3 (to Theorem 6 and Theorem 7).** *Suppose we are given the following dynamics for real vector state space  $\mathbb{X}$  and action space  $\mathbb{U}$*

$$x_{t+1} = f(x_t, u_t, w_t), \quad y_t = g(x_t, v_t),$$

where  $\{w_t\}$  and  $\{v_t\}$  are i.i.d. noise processes and the noise models are unknown but it is known that the noise  $w_t$  admits a continuous probability density function. Suppose that  $f(x, u, \cdot) : \mathbb{W} \rightarrow \mathbb{X}$  is invertible for all  $(x, u)$ . We collect i.i.d. samples of  $\{w_t\}$  as in (9) and use them to construct an estimator,  $\tilde{\mu}_n$ , as described in [5] which consistently estimates  $\mu$  in total variation. Using these empirical estimates, we construct the empirical transition kernel estimates for any  $(x_0, u_0)$  and Borel  $B$  as

$$\mathcal{T}_n(B|x_0, u_0) = \tilde{\mu}_n(f_{x_0, u_0}^{-1}(B)).$$

Then independent of the channel,  $J_\beta^*(\mathcal{T}_n) \rightarrow J_\beta^*(\mathcal{T})$  and  $J_\beta(\mathcal{T}, \gamma_n^*) \rightarrow J_\beta^*(\mathcal{T})$ , where  $\gamma_n^*$  is the optimal policy designed for  $\mathcal{T}_n$ . Since the channel model  $g$  has no restrictions, this result also applies to the fully observed model setup by taking  $g(x_t, v_t) = x_t$ .

*Proof.* By [5] we can estimate  $\mu$  in total variation so that almost surely  $\lim_{n \rightarrow \infty} \|\tilde{\mu}_n - \mu\|_{TV} = 0$ . We claim that the convergence of  $\tilde{\mu}_n$  to  $\mu$  under total variation metric implies the convergence of  $\mathcal{T}_n$  to  $\mathcal{T}$  in total variation uniformly over all  $x \in \mathbb{X}$  and  $u \in \mathbb{U}$  i.e.  $\lim_{n \rightarrow \infty} \sup_{x, u} \|\mathcal{T}_n(\cdot|x, u) - \mathcal{T}(\cdot|x, u)\|_{TV} = 0$ . Observe the following:

$$\begin{aligned} & \sup_{x, u} \|\mathcal{T}_n(\cdot|x, u) - \mathcal{T}(\cdot|x, u)\|_{TV} \\ &= \sup_{x, u} \sup_{\|h\|_\infty \leq 1} \left| \int h(x_1) \mathcal{T}_n(dx_1|x, u) - \int h(x_1) \mathcal{T}(dx_1|x, u) \right| \\ &= \sup_{x, u} \sup_{\|h\|_\infty \leq 1} \left| \int h(f(x, u, w)) \tilde{\mu}_n(dw) - \int h(f(x, u, w)) \mu(dw) \right| \\ &\leq \|\tilde{\mu}_n - \mu\|_{TV} \rightarrow 0. \end{aligned}$$

Thus, by Theorem 6 and Theorem 7, the result follows.  $\square$

The following example presents some system and channel models which satisfy the requirements of the above corollaries.

*Example 2.* Let  $\mathbb{X}, \mathbb{Y}, \mathbb{U}$  be real vector spaces with

$$x_{t+1} = f(x_t, u_t) + w_t, \quad y_t = h(x_t, v_t)$$

for unknown i.i.d. noise processes  $\{w_t\}$  and  $\{v_t\}$ .

1. Suppose the channel is in the following form;  $y_t = h(x_t, v_t) = x_t + v_t$  where  $v_t$  admits a density (e.g. Gaussian density). It can be shown by an application of Scheffé's theorem that the channels in this form are continuous in total variation. If further  $f(x_t, u_t)$  is continuous and bounded then the requirements of Corollary 2 hold for partially observed models.
2. If the channel is in the following form;  $x_t = h(x_t, v_t)$  then the system is fully observed. If further  $f(x_t, u_t)$  is continuous and bounded then the requirements of Corollary 2 holds for fully observed models.
3. Suppose the function  $f(x_t, u_t)$  is known, if the noise process  $w_t$  admits a continuous density, then one can estimate the noise model in total variation in a consistent way (see [5]). Hence, the conditions of Corollary 3 holds independent of the channel model.

## 5.2 Application to Approximations of MDPs and POMDPs with Weakly Continuous Kernels

We now discuss **Problem P4**, that is whether approximation of an MDP model with a standard Borel space with a finite MDPs can be viewed an instance of robustness problem to incorrect models and whether our results can be applied.

**Review of Finitely Quantized Approximations to Standard Borel MDPs.** Consider an MDP which is quantized as follows.

**Finite State Approximate MDP: Quantization of the State Space.** Let  $d_{\mathbb{X}}$  denote the metric on  $\mathbb{X}$ . For each  $n \geq 1$ , there exists a finite subset  $\{x_{n,i}\}_{i=1}^{k_n}$  of  $\mathbb{X}$  such that

$$\min_{i \in \{1, \dots, k_n\}} d_{\mathbb{X}}(x, x_{n,i}) < 1/n \text{ for all } x \in \mathbb{X}.$$

Let  $\mathbb{X}_n := \{x_{n,1}, \dots, x_{n,k_n}\}$  and define  $Q_n$  mapping any  $x \in \mathbb{X}$  to the nearest element of  $\mathbb{X}_n$ , i.e.,

$$Q_n(x) := \arg \min_{x_{n,i} \in \mathbb{X}_n} d_{\mathbb{X}}(x, x_{n,i}).$$

For each  $n$ , a partition  $\{\mathbb{S}_{n,i}\}_{i=1}^{k_n}$  of the state space  $\mathbb{X}$  is induced by  $Q_n$  by setting

$$\mathbb{S}_{n,i} = \{x \in \mathbb{X} : Q_n(x) = x_{n,i}\}.$$

Let  $\psi$  be a probability measure on  $\mathbb{X}$  which satisfies

$$\psi(\mathbb{S}_{n,i}) > 0 \text{ for all } i, n,$$

and define probability measures  $\psi_{n,i}$  on  $\mathbb{S}_{n,i}$  by restricting  $\psi$  to  $\mathbb{S}_{n,i}$ :

$$\psi_{n,i}(\cdot) := \psi(\cdot)/\psi(\mathbb{S}_{n,i}).$$

Using  $\{\psi_{n,i}\}$ , we define a sequence of finite-state MDPs, denoted as f-MDP $_m$ , to approximate the compact-state MDP.

For each  $m$ , f-MDP $_m$  is defined as:  $(\mathbb{X}_n, \mathbb{U}, \mathcal{T}_n, c_n)$ , and the one-stage cost function  $c_n : \mathbb{X}_n \times \mathbb{U} \rightarrow [0, \infty)$  and the transition probability  $\mathcal{T}_n$  on  $\mathbb{X}_n$  given  $\mathbb{X}_n \times \mathbb{U}$  are given by

$$\begin{aligned} c_n(x_{n,i}, a) &:= \int_{\mathbb{S}_{n,i}} c(x, a) \psi_{n,i}(dx) \\ \mathcal{T}_n(\cdot | x_{n,i}, a) &:= \int_{\mathbb{S}_{n,i}} Q_n * \mathcal{T}(\cdot | x, a) \psi_{n,i}(dx), \end{aligned}$$

where  $Q_n * \mathcal{T}(\cdot | x, a) \in \mathcal{P}(\mathbb{X}_n)$  is the pushforward of the measure  $\mathcal{T}(\cdot | x, a)$  with respect to  $Q_n$ ; that is,

$$Q_n * \mathcal{T}(z_{n,j} | x, a) = \mathcal{T}(\{y \in \mathbb{X} : Q_n(y) = x_{n,j}\} | x, a),$$

for all  $x_{n,j} \in \mathbb{X}_n$ .

**Finite Action Approximate MDP: Quantization of the Action Space.** Let  $d_{\mathbb{U}}$  denote the metric on  $\mathbb{U}$ . Since the action space  $\mathbb{U}$  is compact and thus totally bounded, one can find a sequence of finite sets  $\Lambda_n = \{a_{n,1}, \dots, a_{n,k_n}\} \subset \mathbb{U}$  such that for all  $n$ ,

$$\min_{i \in \{1, \dots, k_n\}} d_{\mathbb{U}}(a, a_{n,i}) < 1/n \text{ for all } a \in \mathbb{U}.$$

In other words,  $\Lambda_n$  is a  $1/n$ -net in  $\mathbb{U}$ . Let us assume that the sequence  $\{\Lambda_n\}_{n \geq 1}$  is fixed. To ease the notation in the sequel, let us define the mapping  $\mathcal{Y}_n$

$$\mathcal{Y}_n(f)(x) := \arg \min_{a \in \Lambda_n} d_{\mathbb{U}}(f(x), a), \quad (10)$$

where ties are broken so that  $\mathcal{Y}_n(f)(x)$  is measurable.

It is known that finite quantization policies are nearly optimal under the conditions to be presented below, see [23, Theorem 3.2]. Thus, to make the presentation shorter, we will either assume that the action set is finite, or it has been approximated by a finite action space through the construction above. Assuming finite action sets will help us avoid measurability issues (see universal measurability discussions in [22]) as well as issues with existence of optimal policies.

- Assumption 6.** (a) The one stage cost function  $c$  is nonnegative and continuous.  
(b) The stochastic kernel  $\mathcal{T}(\cdot|x, a)$  is weakly continuous in  $(x, a) \in \mathbb{X} \times \mathbb{U}$ .  
(c)  $\mathbb{U}$  is finite.  
(d)  $\mathbb{X}$  is compact.

We note that condition (d) in Assumption 6 as presented in [22] was more general, but we have used the simpler version here for clarity in exposition.

One can write the following fixed point equation for the finite MDP

$$J_\beta^n(x) = \min_{a \in \mathbb{U}} \left\{ c_n(x, a) + \beta \sum_{x_1 \in \mathbb{X}_n} J_\beta^n(x_1) \mathcal{T}_n(x_1|x, a) \right\}$$

where  $\mathcal{T}_n$  is the transition model for the finite MDP and  $c_n$  is the cost function defined on the finite model. Since the action space is finite, we can find an optimal policy, say  $f_n^*$  for this fixed point equation. One can also simply extend  $J_\beta^n$  and  $f_n^*$ , which are defined on  $\mathbb{X}_n$  to the entire state space  $\mathbb{X}$  by taking them constant over the quantization bins  $\mathbb{S}_{n,i}$ . If we call the extended versions  $\hat{J}_\beta^n$  and  $\hat{f}_n$ , the following result can be established:

**Theorem 16.** [22, Theorem 2.2 and 4.1] Suppose Assumption 6 holds. Then, for any  $\beta \in (0, 1)$  the discounted cost of the deterministic stationary policy  $\hat{f}_n$ , obtained by extending the discounted optimal policy  $f_n^*$  of f-MDP<sub>m</sub> to  $\mathbb{X}$  (i.e.,  $\hat{f}_n = f_n^* \circ Q_n$ ), converges to the discounted value function  $J^*$  of the compact-state MDP:

$$\lim_{n \rightarrow \infty} \|\hat{J}_\beta^n(\cdot) - J_\beta^*(\cdot)\| = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \|J_\beta(\hat{f}_n, \cdot) - J_\beta^*\| = 0. \quad (11)$$

Theorems 16 shows that under Assumption 6, an optimal solution can be approximated via the solutions of finite models. We now show that the above approximation scheme can be viewed in relation to our robustness results.

*Proof Sketch of Theorem 16 via results from Sect. 4.* With the introduced setup, one can see that the extended value function and optimal policy for the finite model satisfy the following:

$$\hat{J}_\beta^n(x) = \min_{a \in \mathbb{U}} \left\{ \hat{c}_n(x, u) + \beta \int \hat{J}_\beta^n(x_1) \hat{\mathcal{T}}_n(dx_1|x, u) \right\}$$

where  $\hat{c}_n$  is the extended version of  $c_n$  to the state space  $\mathbb{X}$  by making it constant over the quantization bins  $\{\mathbb{S}_{n,i}\}_i$  and  $\hat{\mathcal{T}}_n$  is such that for any function  $f$

$$\int f(x_1) \hat{\mathcal{T}}_n(dx_1|x, u) := \int_{x_1 \in \mathbb{X}} \int_{z \in \mathbb{S}_{n,i}} f(x_1) \mathcal{T}(dx_1|z, u) \psi_{n,i}(dz)$$

where  $\mathbb{S}_{n,i}$  is the quantization bin that  $x$  belongs to.

With this setup, one can see that for any  $x_n \rightarrow x$  we have  $\hat{c}_n(x_n, u) \rightarrow c(x, u)$  and for any continuous and bounded  $f$

$$\begin{aligned} \int f(x_1) \hat{T}_n(dx_1|x_n, u) &:= \int_{x_1 \in \mathbb{X}} \int_{z \in \mathbb{S}_{n,i}} f(x_1) \mathcal{T}(dx_1|z, u) \psi_{n,i}(dz) \\ &\rightarrow \int f(x_1) \mathcal{T}(dx_1|x, u). \end{aligned}$$

Hence, Assumption 1 holds under Assumption 6, and we can conclude the proof using Theorem 11 and Theorem 10.  $\square$

## 6 Concluding Remarks

We studied regularity properties of optimal stochastic control on the space of transition kernels, and applications to robustness of optimal control policies designed for an incorrect model applied to an actual system. We also presented applications to data-driven learning and related the robustness problem to finite MDP approximation techniques. For the problems presented in this article, our focus was on infinite horizon discounted cost setup. However, we note that the results can be extended to the infinite horizon average cost setup under various forms of ergodicity properties on the state process.

## References

1. Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., Eder, M.: Adapted Wasserstein distances and stability in mathematical finance. *Financ. Stoch.* **24**, 3601–632 (2020)
2. Bayraktar, E., Dolinsky, Y., Guo, J.: Continuity of utility maximization under weak convergence. *Math. Financial Econ.* **14**(4), 1–33 (2020)
3. Billingsley, P.: Statistical methods in Markov chains. *Ann. Math. Statist.* **32**, 12–40 (1961)
4. Billingsley, P.: Probability and Measure, 3rd edn. Wiley, New York (1995)
5. Devroye, L., Györfi, L.: Non-parametric Density Estimation: The  $L_1$  View. Wiley, New York (1985)
6. Dudley, R.M.: Real Analysis and Probability, 2nd edn. Cambridge University Press, Cambridge (2002)
7. Dupuis, P., James, M.R., Petersen, I.: Robust properties of risk-sensitive control. *Math. Control Signals Syst.* **13**(4), 318–332 (2000)
8. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.* **171**(1), 1–52 (2018)
9. Feinberg, E., Kasyanov, P., Zgurovsky, M.: Partially observable total-cost Markov decision process with weakly continuous transition probabilities. *Math. Oper. Res.* **41**(2), 656–681 (2016)
10. Ghosh, J.K., Ramamoorthi, R.V.: Bayesian Nonparametrics. Springer, New York (2003)
11. Gray, R.M.: Entropy and Information Theory. Springer-Verlag, New York (1990)

12. Györfi, L., Kohler, M.: Nonparametric estimation of conditional distributions. *IEEE Trans. Inf. Theory* **53**(5), 1872–1879 (2007)
13. Hernandez-Lerma, O., Lasserre, J.: Discrete-Time Markov Control Processes. Springer, New York (1996)
14. Jacobson, D.: Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Automat. Contr.* **18**(2), 124–131 (1973)
15. Kara, A.D., Saldi, N., Yüksel, S.: Weak Feller property of non-linear filters. *Syst. Control Lett.* **134**, 104–112 (2019)
16. Kara, A. D., Yüksel, S.: Robustness to incorrect system models in stochastic control and application to data-driven learning. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 2753–2758 (2018)
17. Kara, A.D., Yüksel, S.: Robustness to incorrect priors in partially observed stochastic control. *SIAM J. Control. Optim.* **57**(3), 1929–1964 (2019)
18. Kara, A.D., Yüksel, S.: Robustness to incorrect system models in stochastic control. *SIAM J. Control. Optim.* **58**(2), 1144–1182 (2020)
19. Parthasarathy, K.: Probability Measures on Metric Spaces. AMS, Providence (2005)
20. Petersen, I., James, M.R., Dupuis, P.: Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Automat. Contr.* **45**(3), 398–412 (2000)
21. Pra, P.D., Meneghini, L., Runggaldier, W.J.: Connections between stochastic control and dynamic games. *Math. Control Signals Syst.* **9**(4), 303–326 (1996)
22. Saldi, N., Yüksel, S., Linder, T.: On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. *Math. Oper. Res.* **42**(4), 945–978 (2017)
23. Saldi, N., Yüksel, S., Linder, T.: Near optimality of quantized policies in stochastic control under weak continuity conditions. *J. Math. Anal. Appl.* **435**(1), 321–337 (2015)
24. Savkin, A.V., Petersen, I.R.: Robust control of uncertain systems with structured uncertainty. *J. Math. Syst. Est. Control* **6**(3), 1–14 (1996)
25. Sun, H., Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. *Math. Oper. Res.* **41**(2), 377–401 (2016)
26. Ugrinovskii, V.A.: Robust H-infinity control in the presence of stochastic uncertainty. *Int. J. Control.* **71**(2), 219–237 (1998)



# Full Gradient DQN Reinforcement Learning: A Provably Convergent Scheme

Konstantin E. Avrachenkov<sup>1(✉)</sup>, Vivek S. Borkar<sup>2</sup>, Hars P. Dolhare<sup>2</sup>,  
and Kishor Patil<sup>1</sup>

<sup>1</sup> INRIA Sophia Antipolis, Valbonne 06902, France  
`{k.avrachenkov,kishor.patil}@inria.fr`

<sup>2</sup> Indian Institute of Technology Bombay, Mumbai 400076, India  
`borkar.vs@gmail.com,harshdolhare99@gmail.com`

**Abstract.** We analyze the DQN reinforcement learning algorithm as a stochastic approximation scheme using the o.d.e. (for ‘*ordinary differential equation*’) approach and point out certain theoretical issues. We then propose a modified scheme called Full Gradient DQN (FG-DQN, for short) that has a sound theoretical basis and compare it with the original scheme on sample problems. We observe a better performance for FG-DQN.

**Keywords:** Markov decision process (MDP) · Approximate dynamic programming · Deep Reinforcement Learning (DRL) · Stochastic approximation · Deep Q-network (DQN) · Full Gradient DQN · Bellman error minimization

**AMS(2000) Subject Classification:** Primary 93E35 · Secondary 68T05 · 90C40 · 93E35

## 1 Introduction

Recently we have witnessed tremendous success of Deep Reinforcement Learning algorithms in various application domains. Just to name a few examples, DRL has achieved superhuman performance in playing Go [41], Chess [42] and many Atari video games [31,32]. In Chess, DRL algorithms have also beaten the state of the art computer programs, which are based on more or less brute-force enumeration of moves. Moreover, playing Go and Chess, DRL surprised experts with new insights and beautiful strategies [41,42]. We would also like to mention the impressive progress of DRL applications in robotics [23,24,33], telecommunications [29,36,51] and medicine [26,34].

The use of Deep Neural Networks is of course an essential part of DRL. However, there are other paramount elements that contributed to the success of DRL. A starting point for DRL was the Q-learning algorithm of Watkins [49], which in its original form can suffer from the proverbial curse of dimensionality. In [25,45] the convergence of Q-learning has been rigorously established.

Then, in [21, 22] Gordon has proposed and analyzed fitted Q-learning using a novel architecture based on what he calls ‘averager’ maps. In [38] Riedmiller has proposed using a neural network for approximating Q-values. There he has also suggested that we treat the right hand side of the dynamic programming equation for Q-values (see Eq. (5) below) as the ‘target’ to be chased by the left hand side, i.e., the Q-value itself, and then seek to minimize the mean squared error between the two. The right hand side in question also involves the Q-value approximation and *ipso facto* the parameter itself, which is treated as a ‘given’ for this purpose, as a part of the target, and the minimization is carried out only over the same parameter appearing in the left hand side. This leads to a scheme reminiscent of temporal difference learning, albeit a nonlinear variant of it. The parameter dependence of the target leads to some difficulties because of the permanent shifting of the target itself, what one might call the ‘dog chasing its own tail’ phenomenon. Already in [38], frequent instability of the algorithm has been reported.

The next big step in improvement of DRL performance was carried out by DeepMind researchers, who elaborated the Deep Q-Network (DQN) scheme [31], [32]. Firstly, to improve the stability of the algorithm in [38], they suggested freezing the parameter value in the target network for several iterates. Thus in DQN, the target network evolves on a slower timescale. The second successful tweak for DQN has been the use of ‘experience replay’, or averaging over some relevant traces from the past, a notion introduced in [27, 28]. Then, in [47, 48] it was suggested that we introduce a separation of policy estimation and evaluation to further improve stability. The latter scheme is called Double DQN. While various success stories of DQN and Double DQN schemes have been reported, this does not completely fix the theoretical and practical issues.

Let us mention that apart from Q-value based methods in DRL, there is another large family of methods based on policy gradient. Each family has its own positive and negative features (for background on RL and DRL methods we recommend the texts [7, 20, 43]). While there has been a notable progress in the theoretical analysis of the policy gradient methods [1, 2, 8, 13, 30, 44], there are no works establishing convergence of the neural Q-value based methods to the best of our knowledge.

In this work, we revisit DQN and scrutinize it as a stochastic approximation algorithm, using the ‘o.d.e.’ (for ‘ordinary differential equation’) approach for its convergence analysis (see [11] for a textbook treatment). In fact, we go beyond the basic o.d.e. approach to its generalization based on differential inclusions, involving in particular non-smooth analysis. This clarifies the underlying difficulties regarding theoretical guarantees of convergence and also suggests a modification, which we call the Full Gradient DQN, or FG-DQN. We establish theoretical convergence guarantees for FG-DQN and compare it empirically with DQN on sample problems (forest management [14, 16] and cartpole [5, 19]), where it gives better performance at the expense of some additional computational overhead per iteration.

As was noticed above, another successful tweak for DQN has been the use of ‘experience replay’. We too incorporate this in our scheme. Many advantages

of experience replay have been cited in literature, which we review later in this article. We also unearth an interesting additional advantage of ‘experience replay’ for Bellman error minimization using gradient descent and compare it with the ‘double sampling’ technique of [3]. See Sects. 4.2 and 5.1 below.

## 2 DQN Reinforcement Learning

### 2.1 Q-learning

We begin by recalling the derivation of the original Q-learning scheme [49] to set up the context. Consider a Markov chain  $\{X_n\}$  on a finite state space  $S := \{1, 2, \dots, s\}$ , controlled by a control process  $\{U_n\}$  taking values in a finite action space  $A = \{1, 2, \dots, a\}$ . Its transition probability function is denoted by  $(x, y, u) \in S^2 \times A \mapsto p(y|x, u) \in [0, 1]$  such that  $\sum_y p(y|x, u) = 1 \forall x, u$ . The controlled Markov property then is

$$P(X_{n+1} = y|X_m, U_m, m \leq n) = p(y|X_n, U_n) \quad \forall n \geq 0, y \in S.$$

We call  $\{U_n\}$  an admissible control policy. It is called a stationary policy if  $U_n = v(X_n) \forall n$  for some  $v : S \rightarrow A$ . A more general notion is that of a stationary randomized policy wherein one chooses the control  $U_n$  at time  $n$  probabilistically with a conditional law given the  $\sigma$ -field  $\mathcal{F}_n := \sigma(X_m, U_m, m < n; X_n)$  that depends only on  $X_n$ . That is,

$$\varphi(u|X_n) := P(U_n = u|\mathcal{F}_n) = P(U_n = u|X_n)$$

for a prescribed map  $x \in S \mapsto \varphi(\cdot|x) \in \mathcal{P}(A) :=$  the simplex of probability vectors on  $A$ . One identifies such a policy with the map  $\varphi$ . Denote the set of stationary randomized policies by  $\mathcal{U}_{SR}$ . In anticipation of the learning schemes we discuss, we impose the ‘frequent updates’ or ‘sufficient exploration’ condition

$$\liminf_{n \uparrow \infty} \frac{1}{n} \sum_{m=0}^{n-1} I\{X_m = x, U_m = u\} > 0 \quad \text{a.s.} \quad \forall x, u. \quad (1)$$

Given a per stage reward  $(x, u) \mapsto r(x, u)$  and a discount factor  $\gamma \in (0, 1)$ , the objective is to maximize the infinite horizon expected discounted reward

$$E \left[ \sum_{m=0}^{\infty} \gamma^m r(X_m, U_m) \right].$$

The ‘value function’  $V : S \rightarrow \mathcal{R}$  defined as

$$V(x) = \max E \left[ \sum_{m=0}^{\infty} \gamma^m r(X_m, U_m) \middle| X_0 = x \right], \quad x \in S, \quad (2)$$

then satisfies the dynamic programming equation

$$V(x) = \max_u \left[ r(x, u) + \gamma \sum_y p(y|x, u) V(y) \right], \quad x \in S. \quad (3)$$

Furthermore, the maximizer  $v^*(x)$  on the right (chosen arbitrarily if not unique) defines a stationary policy  $v^* : S \rightarrow A$  that is optimal, i.e., achieves the maximum in (2). Equation (3) is a fixed point equation of the form  $V = F(V)$  (which defines the map  $F : \mathcal{R}^s \rightarrow \mathcal{R}^s$ ) and can be solved by the ‘value iteration’ algorithm

$$V_{n+1}(x) = \max_u \left[ r(x, u) + \gamma \sum_y p(y|x, u) V_n(y) \right], \quad n \geq 0, \quad (4)$$

beginning with any  $V_0 \in \mathcal{R}^s$ .  $F$  can be shown to satisfy

$$\|F(x) - F(y)\|_\infty \leq \gamma \|x - y\|_\infty,$$

i.e., it is an  $\|\cdot\|_\infty$ -norm contraction. Then (4) is a standard fixed point iteration of a contraction map and converges exponentially to its unique fixed point  $V$ .

Now define Q-values as the expression in square brackets in (3), i.e.,

$$Q(x, u) = r(x, u) + \gamma \sum_y p(y|x, u) V(y), \quad x \in S, u \in A.$$

If the function  $Q(\cdot, \cdot)$  is known, then the optimal control at state  $x$  is found by simply maximizing  $Q(x, \cdot)$  without requiring the knowledge of reward or transition probabilities. This makes it suitable for data-driven algorithms of reinforcement learning. By (3),  $V(x) = \max_u Q(x, u)$ . The Q-values then satisfy their own dynamic programming equation

$$Q(x, u) = r(x, u) + \gamma \sum_y p(y|x, u) \max_v Q(y, v), \quad (5)$$

which in turn can be solved by the ‘Q-value iteration’

$$Q_{n+1}(x, u) = r(x, u) + \gamma \sum_y p(y|x, u) \max_v Q_n(y, v), \quad x \in S, u \in A. \quad (6)$$

What we have gained at the expense of increased dimensionality is that the nonlinearity is now inside the conditional expectation w.r.t. the transition probability function. This facilitates a stochastic approximation algorithm [11] where we first replace this conditional expectation by actual evaluation at a real or simulated random variable  $\zeta_{n+1}(x, u)$  with law  $p(\cdot|x, u)$ , and then make an incremental correction to the current guess based on it. That is, replace (6) by

$$Q_{n+1}(x, u) = (1 - a(n))Q_n(x, u) + a(n) \left( r(x, u) + \gamma \max_v Q_n(\zeta_{n+1}(x, v), v) \right) \quad (7)$$

for some  $a(n) > 0$ . The Q-learning algorithm does so using a single run of a real or simulated controlled Markov chain  $(X_n, U_n), n \geq 0$ , so that:

- at each time instant  $n$ ,  $(X_n, U_n)$  are observed and the  $(X_n, U_n)$ th component of  $Q$  is updated, leaving other components of  $Q_n(\cdot, \cdot)$  unchanged,

- this update follows (7) where  $\zeta_{n+1}(x, u)$  with  $x = X_n, u = U_n$ , gets replaced by  $X_{n+1}$ , which indeed has the conditional law  $p(\cdot | X_n, U_n)$  as required,
- $\{a(n)\}$  are positive scalars in  $(0, 1)$  chosen to satisfy the standard Robbins-Monro conditions of stochastic approximation [11], i.e.,

$$\sum_n a(n) = \infty, \quad \sum_n a(n)^2 < \infty. \quad (8)$$

It is more convenient to write the resulting Q-learning algorithm as

$$Q_{n+1}(x, u) = Q_n(x, u) + a(n) I\{X_n = x, U_n = u\} \left( r(x, u) + \gamma \max_v Q_n(X_{n+1}, v) - Q_n(x, u) \right) \quad \forall x, u, \quad (9)$$

where  $I\{\dots\}$  := the indicator random variable that equals 1 if ‘ $\dots$ ’ holds and 0 if not. The fact that only one component is being updated at a time makes this an asynchronous stochastic approximation. Nevertheless, it exhibits the well known ‘averaging effect’ of stochastic approximation whereby it is a data-driven scheme that emulates (6) and exhibits convergence a.s. to the same limit, viz.,  $Q$ . For formal proofs, see [25, 45, 50].

## 2.2 DQN Learning

The raw Q-learning scheme (9), however, does inherit the ‘curse of dimensionality’ of MDPs. One common fix is to replace  $Q$  by a parametrized family  $(x, u, \theta) \mapsto Q(x, u; \theta)$  (where we again use the notation  $Q(\cdot, \cdot; \cdot)$  by abuse of terminology so as to match standard usage). Here  $\theta \in \Theta \subset \mathcal{R}^d$  for a moderate  $d \geq 1$  and the objective is to learn the ‘optimal’ approximation  $Q(\cdot, \cdot; \theta^*)$  by iterating in  $\Theta$ . For simplicity, we take  $\Theta = \mathcal{R}^d$ . One natural performance measure is the ‘DQN Bellman error’

$$\mathcal{E}(\theta) := E \left[ (Z_n - Q(X_n, U_n; \theta))^2 \right], \quad (10)$$

where

$$Z_n := r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \theta_n)$$

is the ‘target’ that is taken as a *given quantity* and the expectation is w.r.t. the stationary law of  $(X_n, U_n)$ . For later reference, note that this is different from the ‘true Bellman error’

$$\bar{\mathcal{E}}(\theta) := E \left[ \left( r(X_n, U_n) + \gamma \sum_y p(y | X_n, U_n) \max_v Q(y, v; \theta) - Q(X_n, U_n; \theta) \right)^2 \right]. \quad (11)$$

The stochastic gradient type scheme based on the empirical semi-gradient of  $\mathcal{E}(\cdot)$  then becomes

$$\theta_{n+1} = \theta_n + a(n)(Z_n - Q(X_n, U_n; \theta_n)) \nabla_\theta Q(X_n, U_n; \theta_n), \quad n \geq 0. \quad (12)$$

### 2.3 Experience Replay

An important modification of the DQN scheme has been the incorporation of ‘experience replay’. The idea is to replace the term multiplying  $a(n)$  on the right hand side of (12) by an empirical average over traces of transitions from past that are stored in memory. The algorithm then becomes

$$\begin{aligned}\theta_{n+1} = \theta_n + \frac{a(n)}{M} \\ \times \sum_{m=1}^M \left( (Z_{n(m)} - Q(X_{n(m)}, U_{n(m)})) \nabla_{\theta} Q(X_{n(m)}, U_{n(m)}; \theta_{n(m)}) \right), \quad n \geq 0,\end{aligned}\quad (13)$$

where  $(X_{n(m)}, U_{n(m)}), 1 \leq m \leq N$ , are samples from past. This has multiple advantages. Some that have been cited in literature are as follows.

1. As in the mini-batch stochastic gradient descent for empirical risk minimization in machine learning, it helps reduce variance. It also diminishes effects of anomalous transitions.
2. Training based on only the immediate experiences ( $\approx$  samples) tends to overfit the model to current data. This is prevented by experience replay. In particular, if past samples are randomly picked, they are less correlated.
3. The re-use of data leads to data efficiency.
4. Experience replay is better suited for delayed rewards or costs, e.g., when the latter are realized only at the end of a long episode or epoch.

There are also variants of basic experience replay, e.g., [40], which replaces purely random sampling from past by a non-uniform sampling which picks a sample with probability proportional to its absolute Bellman error.

We shall be implementing experience replay a little differently in the variant we describe next, which has yet another major advantage from a theoretical standpoint in the specific context of our scheme.

### 2.4 Double DQN Learning

One more modification of the vanilla DQN scheme is doing the policy selection according the local network [47, 48]. The target network is still used in  $Z_n$  and is updated on a slower time scale. The latter can be represented with another set of parameters  $\bar{\theta}_n$ . Thus, the iterate for the Double DQN scheme can be written as follows:

$$\theta_{n+1} = \theta_n + a(n)(Z_n - Q(X_n, U_n; \theta_n)) \nabla_{\theta} Q(X_n, U_n; \theta_n), \quad n \geq 0, \quad (14)$$

with

$$Z_n := r(X_n, U_n) + \gamma Q(X_{n+1}, v; \bar{\theta}_n) \Big|_{v=\operatorname{argmax}_{v'} Q(X_{n+1}, v'; \theta_n)}.$$

For the sake of comparison, in the vanilla DQN one has:

$$Z_n := r(X_n, U_n) + \gamma Q(X_{n+1}, v; \bar{\theta}_n) \Big|_{v=\operatorname{argmax}_{v'} Q(X_{n+1}, v'; \bar{\theta}_n)}.$$

Note that in Double DQN, the selection and evaluation of the policy is done separately. According to [47, 48] this modification improves the stability of the DQN learning. One can also combine Double DQN with experience replay [48].

### 3 The Issues with DQN Learning

The expression for DQN learning scheme is appealing because of its apparent similarity with the very successful temporal difference learning for policy evaluation [46], not to mention its empirical successes, including some high profile ones such as [32]. Nevertheless, a good theoretical justification seems lacking. The difficulty arises from the fact that the ‘target’  $Z_n$  is not something extraneous, but is also a function of the operative parameter  $\theta_n$ . In fact, this becomes apparent once we expand  $Z_n$  in (12) to write

$$\begin{aligned} \theta_{n+1} = \theta_n + a(n)(r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \theta_n) - Q(X_n, U_n; \theta_n)) \\ \times \nabla_\theta Q(X_n, U_n; \theta_n), \quad n \geq 0. \end{aligned} \quad (15)$$

Write

$$\tilde{\mathcal{E}}(\theta, \bar{\theta}) := E \left[ \left( r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \bar{\theta}) - Q(X_n, U_n; \theta) \right)^2 \right], \quad (16)$$

where  $E[\cdot]$  is the stationary expectation as before. Consider the ‘off-policy’ case, i.e.,  $\{(X_n, U_n)\}$  is the state-action sequence of a controlled Markov chain satisfying (1) with a pre-specified stationary randomized policy that does not depend on the iterates. (As we point out later, the ‘on-policy’ version, which allows for the latter adaptation, has additional issues.) If we apply the ‘o.d.e. approach’ for analysis of stochastic approximation (see, e.g., [11] for a textbook treatment), we get the limiting o.d.e. as

$$\dot{\theta}(t) = -\nabla_1 \tilde{\mathcal{E}}(\theta(t), \theta(t)),$$

where  $\nabla_i$  denotes gradient with respect to the  $i$ th argument of  $\tilde{\mathcal{E}}(\cdot, \cdot)$  for  $i = 1, 2$ . Thus it is a partial stochastic gradient descent wherein only the gradient with respect to the first occurrence of the variable is used. Unlike gradient dynamics, there is no reason why such dynamics should converge. It was already mentioned that in case of linear function approximation, the DQN iteration bears a similarity with TD(0), except for the nonlinear ‘max’ term. The o.d.e. proof of convergence for TD(0) does not carry over to DQN precisely because the stochastic approximation version leads to the interchange of the conditional expectation and max operators. The other issue is that in TD(0), the linear operator in question is a contraction w.r.t. the weighted  $L_2$ -norm weighted by the stationary distribution. That argument also fails for DQN because of presence of the max operator.

That said, there is already a tweak that treats the first occurrence of  $\theta$  on the RHS, i.e., that inside the maximizer, as the ‘target’ being followed, and updates

it only after several (say,  $K$ ) iterates. In principle, this implies a delay in the corresponding input to the iteration and with decreasing stepsizes, introduces only an asymptotically negligible additional error, so that the limiting o.d.e. remains the same ([11], Chapter 6). This is also the case for Double DQN.

Suppose on the other hand that in DQN or Double DQN we consider a small constant stepsize  $a(n) \equiv a > 0$  and let  $K$  be large, so that with a fixed target value, the algorithm nearly minimizes the Bellman error before the target is updated. Then, assuming the simpler ‘off-policy’ case again, the limiting o.d.e. *for the target*, treating the multiple iterates between its successive iterates as a subroutine, is

$$\dot{\theta}(t) = -\nabla_1 \tilde{\mathcal{E}}(x, \theta(t)) \Big|_{x=\text{argmax}(\tilde{\mathcal{E}}(\cdot, \theta(t)))}. \quad (17)$$

There is no obvious reason why this should converge either. In fact the right hand side would be  $\approx$  the zero vector near the current maximizer and the evolution of the o.d.e. and the iteration would be very slow. Of course, this is a limiting case of academic interest only, stated to underscore the fact that it is difficult to get convergent dynamics out of the DQN learning scheme. This motivates our modification, which we state in the next section.

## 4 Full Gradient DQN

We propose the obvious, viz., to treat both occurrences of the variable  $\theta$  on equal footing, i.e., treat it as a single variable, and then take the full gradient with respect to it. The iteration now is

$$\begin{aligned} \theta_{n+1} = \theta_n - a(n) & \left( r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \theta_n) - Q(X_n, U_n; \theta_n) \right) \\ & \times (\gamma \nabla_\theta Q(X_{n+1}, v_n; \theta_n) - \nabla_\theta Q(X_n, U_n; \theta_n)) \end{aligned} \quad (18)$$

for  $n \geq 0$ , where  $v_n \in \text{Argmax}Q(X_{n+1}, \cdot; \theta_n)$  chosen according to some tie-breaking rule when necessary. Note that when the maximizer in the term involving the max operator is not unique, one may lose its differentiability, but the expression above still makes sense in terms of the Frechet sub-differential, see Appendix. We assume throughout that  $\{X_n\}$  is a Markov chain controlled by the control process  $\{U_n\}$  generated according to a fixed stationary randomized policy  $\varphi \in \mathcal{U}_{SR}$ . Other simulation scenarios are possible for the off-policy set-up. For example, we may replace the triplets  $(X_n, U_n, X_{n+1})$  on the right hand side by triplets  $(X'_n, U'_n, Y'_n)$  where  $\{X'_n\}$  are generated i.i.d. according to some distribution with full support and  $(U'_n, Y'_n)$  are generated with conditional law  $P(U'_n = u, Y'_n = y | X'_n = x) = \varphi(u|x)p(y|x, u)$ , conditionally independent of all other random variables generated till  $n$  given  $X'_n$ . The analysis will be similar. Yet another possibility is that of going through the relevant pairs  $(x, u)$  in a round robin fashion.

We modify (18) further by replacing the right hand side as follows:

$$\begin{aligned} \theta_{n+1} = \theta_n - a(n) & \left( \overline{(r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \theta_n) - Q(X_n, U_n; \theta_n))} \right. \\ & \times \left. (\gamma \nabla_\theta Q(X_{n+1}, v_n; \theta_n) - \nabla_\theta Q(X_n, U_n; \theta_n)) + \xi_{n+1} \right) \end{aligned} \quad (19)$$

for  $n \geq 0$ , where  $\{\xi_n\}$  is extraneous i.i.d. noise componentwise distributed independently and uniformly on  $[-1, 1]$ , and the overline stands for a modified form of experience replay which comprises of averaging at time  $n$  over past traces sampled from  $(X_k, U_k, X_{k+1}), k \leq n$ , for which  $X_k = X_n, U_k = U_n$ . We analyze the asymptotic behavior of this scheme in the remainder of this section in the ‘off-policy’ case, i.e., we use a prescribed stationary randomized policy  $\varphi \in \mathcal{U}_{SR}$ .

We make the following key assumptions:

**(C1)** (Assumptions regarding the function  $Q(\cdot, \cdot; \cdot)$ )

1. The map  $(x, u; \theta) \mapsto Q(x, u; \theta)$  is bounded and twice continuously differentiable in  $\theta$  with bounded first and second derivatives;
2. For each choice of  $x \in S$ , the set of  $\theta$  for which the maximizer of  $Q(x, \cdot; \theta)$  is not unique, is the complement of an open and dense set and has Lebesgue measure zero;
3. Call  $\hat{\theta}$  a critical point of  $\mathcal{E}(\cdot)$  (which is defined in terms of  $Q$ ) if the zero vector is contained in the (Frechet) subdifferential  $\partial^-\mathcal{E}(\hat{\theta})$  (see the Appendix for a definition). We assume that there are at most finitely many such points.

We also assume:

**(C2)** (Stability assumption)

The iterates remain a.s. bounded, i.e.,

$$\sup_n \|\theta_n\| < \infty \text{ a.s.} \quad (20)$$

Our final assumption is a bit more technical. Rewrite the term

$$\overline{(r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \theta_n) - Q(X_n, U_n; \theta_n))}$$

as

$$\begin{aligned} & \sum_y p(y|X_n, U_n) \left( r(X_n, U_n) + \gamma \max_v Q(y, v; \theta_n) - Q(X_n, U_n; \theta_n) \right) \\ & + \varepsilon(X_n, U_n, \theta_n) \end{aligned}$$

where the error term  $\varepsilon(X_n, U_n, \theta_n)$  captures the difference between the empirical conditional expectation using experience replay and the actual conditional expectation. We assume that:

**(C3)** (Assumption regarding the residual error in experience replay)

The error terms  $\{\varepsilon(X_n, U_n, \theta_n)\}$  satisfy

$$\varepsilon(X_n, U_n, \theta_n) \rightarrow 0 \text{ a.s. and } \sum_n a(n) E[\|\varepsilon(X_n, U_n, \theta)\|] |_{\theta=\theta_n} < \infty \text{ a.s.},$$

where the expectation is taken w.r.t. the stationary distribution of the state-action pairs.

We comment on these assumptions later. Recall the true Bellman error  $\bar{\mathcal{E}}(\cdot)$  defined in (11).

**Theorem 1.** The sequence  $\{\theta_n\}$  generated by FG-DQN converges a.s. to a sample path dependent critical point of  $\bar{\mathcal{E}}(\cdot)$ .

**Proof:** For notational ease, write

$$\epsilon(n) := -\varepsilon(X_n, U_n, \theta_n) (\gamma \nabla_\theta Q(X_{n+1}, v_n; \theta_n) - \nabla_\theta Q(X_n, U_n; \theta_n)),$$

where  $v_n$  is chosen from Argmax  $Q(X_{n+1}, \cdot; \theta_n)$  as described earlier. Consider the iteration

$$\begin{aligned} \theta_{n+1} &= \theta_n - a(n) \\ &\times \left( \left( \sum_y p(y|X_n, U_n) (r(X_n, U_n) + \gamma \max_v Q(y, v; \theta_n) - Q(X_n, U_n; \theta_n)) \right) \right. \\ &\quad \left. \times (\gamma \nabla_\theta Q(X_{n+1}, v_n; \theta_n) - \nabla_\theta Q(X_n, U_n; \theta_n)) + \epsilon(n) + \xi_{n+1} \right) \end{aligned} \quad (21)$$

for  $n \geq 0$ . Adding and subtracting the one step conditional expectation of the RHS with respect to  $\mathcal{F}'_n := \sigma(X_m, U_m, m \leq n)$ , we have

$$\begin{aligned} \theta_{n+1} &= \theta_n - a(n) \\ &\times \left( \sum_y p(y|X_n, U_n) (r(X_n, U_n) + \gamma \max_v Q(y, v; \theta_n) - Q(X_n, U_n; \theta_n)) \right) \\ &\times \left( \sum_y p(y|X_n, U_n) (\gamma \nabla_\theta Q(y, u_n(y); \theta_n) - \nabla_\theta Q(X_n, U_n; \theta_n)) \right) \\ &+ a(n)\epsilon(n) + a(n)M_{n+1}(\theta_n) \end{aligned} \quad (22)$$

where  $u_n(y) \in \text{Argmax } Q(y, \cdot; \theta_n)$  is chosen as described earlier, and  $\{M_n(\theta_{n-1})\}$  is a martingale difference sequence w.r.t. the sigma fields  $\{\mathcal{F}'_n\}$ , given by

$$\begin{aligned} M_{n+1}(\theta_n) &= \\ &\left( \left( \sum_y p(y|X_n, U_n) (r(X_n, U_n) + \gamma \max_v Q(y, v; \theta_n) - Q(X_n, U_n; \theta_n)) \right) \right. \\ &\quad \left. \times \left( \gamma \nabla_\theta Q(X_{n+1}, v_n; \theta_n) - \sum_y p(y|X_n, U_n) \gamma \nabla_\theta Q(y, u_n(y); \theta_n) \right) + \xi_{n+1} \right). \end{aligned}$$

Because of our assumptions on  $Q(\cdot, \cdot; \cdot)$  and  $\{\xi_n\}$ ,  $M_n(\cdot)$  will have derivatives uniformly bounded in  $n$  and therefore a uniform linear growth w.r.t.  $\theta$ . The same holds for the expression multiplying  $a(n)$  in the first term on the right. We shall analyze this iteration as a stochastic approximation with Markov noise  $(X_n, U_n), n \geq 0$ , and martingale difference noise  $M_{n+1}, n \geq 0$  ([11], Chapter 6).

The difficult terms are those of the form  $\gamma \nabla_\theta Q(y, u; \theta)$  above, because all we can say about them is that:

$$\begin{aligned} \nabla_\theta Q(y, u; \theta) &\in G(y, \theta) := \\ &\left\{ \sum_v \psi(v|y) \nabla_\theta Q(y, v; \theta) : \psi(\cdot|y) \in \text{Argmax}_{\phi(\cdot|y)} \left( \sum_u \phi(u|y) Q(y, u; \theta) \right) \right\}. \end{aligned}$$

Define correspondingly the set-valued map

$$(x, u, \theta) \mapsto H(x, u, \theta)$$

by

$$\begin{aligned} H(x, u; \theta) &:= \overline{\text{co}} \left( \left\{ \left( \sum_y p(y|x, u) (r(x, u) + \gamma \max_v Q(y, v; \theta) - Q(x, u; \theta)) \right) \right. \right. \\ &\times \sum_y p(y|x, u) (\gamma \nabla_\theta Q(y, v_j; \theta) - \nabla_\theta Q(x, u; \theta)) : v_j \in \text{Argmax} Q(y, \cdot; \theta) \left. \right\} \Bigg) \\ &= \left\{ \left( \sum_y p(y|x, u) (r(x, u) + \gamma \max_v Q(y, v; \theta) - Q(x, u; \theta)) \right) \right. \\ &\times \sum_y p(y|x, u) (\gamma \nabla_\theta \bar{Q}(y, \pi_y; \theta) - \nabla_\theta Q(x, u; \theta)) : \pi_y \in \text{Argmax} \bar{Q}(y, \cdot; \theta) \left. \right\} \end{aligned}$$

where  $\bar{Q}(y, \psi; \theta) := \sum_u \psi(u|y) Q(y, u; \theta)$  for  $\psi \in \mathcal{U}_{SR}$ . Then (22) can be written in the more convenient form as the stochastic recursive inclusion ([11], Chapter 5) given by

$$\theta_{n+1} \in \theta_n - a(n) \left( H(X_n, U_n; \theta_n) + \epsilon(n) + M_{n+1}(\theta_n) \right). \quad (23)$$

We shall now use Theorem 7.1 of [52], pp. 355, for which we need to verify the assumptions (A1)–(A5), pp. 331–2, therein. We do this next.

- (A1) requires  $H(y, \phi, \theta)$  to be nonempty convex compact valued and upper semicontinuous, which is easily verified. It is also bounded by our assumptions on  $Q(\cdot, \cdot; \cdot)$ .
- $S_n$  of [52] corresponds to our  $(X_n, U_n)$  and (A2) can be verified easily.
- (A3) are the standard conditions on  $\{a(n)\}$  also used here.
- $M_{n+1}(\theta_n), n \geq 0$ , defined above, has linear growth in  $\|\theta_n\|$  as observed above. Thus (20) implies that for some  $K \in (0, \infty)$ ,

$$\sum_{m=0}^n a(m)^2 E [\|M_{m+1}(\theta_m)\|^2 | \mathcal{F}_m] \leq K(1 + \sup_m \|\theta_m\|^2) \sum_m a(n)^2 < \infty \text{ a.s.}$$

This implies that  $\sum_{m=0}^{n-1} a(m) M_{m+1}(\theta_m)$  is an a.s. convergent martingale by Theorem 3.3.4, pp. 53-4, [10]. This verifies (A4).

- (A5) is the same as (20) above.

Let  $\mu(x, u) :=$  the stationary probability  $P(X_n = x, U_n = u)$  under  $\varphi$ . Then Theorem 7.1 of [52] applies and allows us to conclude that the iterates will track the asymptotic behavior of the differential inclusion

$$\dot{\theta}(t) \in - \sum_{x,u} \mu(x, u) H(x, u, \theta(t)). \quad (24)$$

Now we make the important observation that under our hypotheses on the function  $Q(\cdot, \cdot; \cdot)$  (see 2. of (C1)), for all  $x, u$  and Lebesgue-a.e.  $\theta$  belonging to some open dense set  $O$ ,  $H(x, u, \theta)$  is the singleton corresponding to Argmax  $Q(x, \cdot; \theta) = \{u\}$  for some  $u \in A$ . Furthermore, in this case, the RHS of (24) reduces to  $-\nabla \mathcal{E}(\theta(t))$ . Since  $\{\xi_n\}$  has density w.r.t. the Lebesgue measure, so will  $\{\theta_n\}$  and therefore by (C1),  $\theta_n \in O \ \forall n$ , a.s. Let

$$L(x, u; \theta) := \frac{1}{2} \left( r(x, u) + \gamma \sum_y p(y|x, u) \max_v Q(y, v; \theta) - Q(x, u; \theta) \right)^2$$

denote the instantaneous Bellman error. Then

$$\bar{\mathcal{E}}(\theta) = \sum_{x,u} \mu(x, u) L(x, u; \theta).$$

Write  $\hat{\mathcal{E}}(\theta')$  for  $\bar{\mathcal{E}}(\theta)$  evaluated at a possibly random  $\theta'$ , in order to emphasize the fact that while  $\bar{\mathcal{E}}(\cdot)$  is defined in terms of an expectation, a random argument of  $\hat{\mathcal{E}}(\cdot)$  is not being averaged over. We use an analogous notation for other quantities in what follows. Applying the Taylor formula to  $\bar{\mathcal{E}}(\cdot)$ , we have,

$$\hat{\mathcal{E}}(\theta_{n+1}) = \hat{\mathcal{E}}(\theta_n) + \sum_{x,u} \mu(x, u) \langle \nabla_\theta L(x, u; \theta), \theta_{n+1} - \theta_n \rangle + O(a(n)^2).$$

But by (22), a.s.,

$$\begin{aligned}\theta_{n+1} - \theta_n &= a(n) \left( -\nabla_\theta L(X_n, U_n; \theta_n) + \epsilon(n) + M_{n+1}(\theta_n) \right) \\ &= a(n) \left( -\sum_{x,u} \mu(x, u) \nabla_\theta L(x, u; \theta_n) + \epsilon(n) + \widetilde{M}_{n+1}(\theta_n) + O(a(n)^2) \right),\end{aligned}$$

where we have replaced  $\nabla_\theta L(X_n, U_n; \theta_n)$  with  $\sum_{i,u} \mu(i, u) \nabla_\theta L(i, u; \theta_n)$ , i.e., with the state-action process  $(X_n, Z_n)$  averaged w.r.t. its stationary distribution (recall that under our randomized stationary Markov policy, it is a Markov chain). This uses a standard (though lengthy) argument for stochastic approximation with Markov noise that converts it to a stochastic approximation with martingale difference noise using the associated parametrized Poisson equation, at the expense of: (i) adding an additional martingale difference noise term that we have added to  $M_{n+1}(\theta_n)$  to obtain the combined martingale difference noise  $\widetilde{M}_{n+1}(\theta_n)$ , and, (ii) another  $O(a(n)^2)$  term that comes from the difference of the solution of the Poisson equation evaluated at  $\theta_n$  and  $\theta_{n+1}$ , which is  $O(\|\theta_{n+1} - \theta_n\|) = O(a(n))$ , multiplied further by an additional  $a(n)$  from (22) to give a net error that is  $O(a(n)^2)$ . See [6] for a classical treatment of this passage.

Hence for suitable constants  $0 < K_1, K'_1 < \infty$ ,

$$\begin{aligned}E[\hat{\mathcal{E}}(\theta_{n+1}) | \mathcal{F}'_n] &\leq \hat{\mathcal{E}}(\theta_n) + a(n) \left( -\left\| \sum_{x,u} \mu(x, u) \nabla_\theta L(x, u; \theta_n) \right\|^2 \right. \\ &\quad \left. + K_1 \sum_{x,u} \mu(x, u) |\varepsilon(x, u, \theta_n)| + K_2 a(n)^2 \right) \\ &\leq \hat{\mathcal{E}}(\theta_n) + a(n) \left( K_1 \sum_{x,u} \mu(x, u) |\varepsilon(x, u, \theta_n)| + K_2 a(n)^2 \right), (25)\end{aligned}$$

where we have used (C1). In view of (C3) and the fact  $\sum_n a(n)^2 < \infty$ , the ‘almost supermartingale’ convergence theorem (Theorem 3.3.6, p. 54, [10]) implies that  $\hat{\mathcal{E}}(\theta_n)$  converges a.s. This is possible only if

$$\begin{aligned}\theta_n &\rightarrow \left\{ \theta : \text{the zero vector is in } \sum_{x,u} \mu(x, u) H(x, u; \theta) \right\} \\ &= \left\{ \theta : \theta \text{ is a critical point of } \sum_{x,u} \mu(x, u) H(x, u; \theta) \right\}.\end{aligned}$$

By property (P4) of the Appendix, it follows that  $H(i, u; \theta) \subset \partial^- L(i, u; \theta)$ . By property (P3) of the Appendix, it then follows that  $\sum_{i,u} \mu(i, u) H(i, u; \theta) \subset \partial^- \bar{\mathcal{E}}(\theta)$ . The claim follows from item 3 in (C1) given that any limit point of  $\theta_n$  as  $n \uparrow \infty$  must be a critical point of  $\partial^- \bar{\mathcal{E}}(\cdot)$  in view of the foregoing.  $\square$

Some comments regarding our assumptions are in order.

1. The vanilla Q-learning iterates, being convex combinations of previous iterates with a bounded quantity, remain bounded. Thus the boundedness assumption on  $Q$  in (C1) is reasonable. The twice continuous differentiability of  $Q$  in  $\theta$  is reasonable when the neural network uses a smooth nonlinearity such as SmoothReLU, GELU or a sigmoid function. As we point out later, using standard ReLU adds another layer of non-smooth analysis which we avoid here for the sake of simplicity of exposition. The last condition in (C1) is also reasonable, e.g., when the graphs of  $Q(x, u; \cdot), Q(x, u'; \cdot)$  cross along a finite union of lower dimensional submanifolds.
2. (C2) assumes stability of iterates, i.e.,  $\sup_n \|\theta_n\| < \infty$  a.s. There is an assortment of tests to verify this. See, e.g., [11], Chapter 3. Also, one can enforce this condition by projection onto a convenient large convex set every time the iterates exit this set, see *ibid.*, Chapter 7.
3. (C3) entails that we perform successive experience replays over larger and larger batches of past samples so that the error in applying the strong law of large numbers decreases sufficiently fast. While this is possible in principle because of the increasing pool of past traces with time, this will be an idealization in practice. It seems possible that the additional error in absence of this can be analyzed as in [37]. Note also that for deterministic control problems, experience replay is not needed for our purposes. The cartpole model studied in the next section is an example of this.

It is worth noting that bulk of the argument above is indeed the classical argument for convergence of stochastic gradient descent with both Markov and martingale difference noise, except that our iteration fits this paradigm only ‘a.s.’. The missing piece is that the (possibly random) point it converges to need not be a point of differentiability of  $\bar{\mathcal{E}}(\cdot)$ , and therefore not a classical critical point thereof. This is what calls for the back and forth between the classical proof and the differential inclusion limit for stochastic gradient descent to minimize a non-smooth objective function.

Before we proceed, we would like to underscore a subtle point, viz., the role of experience replay here. Consider the scheme without the experience replay as above, given by

$$\begin{aligned} \theta_{n+1} = \theta_n - a(n)(r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \theta_n) - Q(X_n, U_n; \theta_n)) \\ \times \left( \gamma \nabla_\theta Q(X_{n+1}, v; \theta_n) \Big|_{v=\operatorname{argmax} Q(X_{n+1}, \cdot; \theta_n)} - \nabla_\theta Q(X_n, U_n; \theta_n) \right). \end{aligned} \quad (26)$$

The limiting o.d.e. for this is

$$\begin{aligned} \dot{\theta}(t) = E \left[ \sum_y p(y|X_n, U_n) \left( (r(X_n, U_n) + \gamma \max_v Q(y, v; \theta(t)) - Q(X_n, U_n; \theta_n)) \right. \right. \\ \times \left. \left. \left( \gamma \nabla_\theta Q(y, v; \theta(t)) \Big|_{v=\operatorname{argmax} Q(y, \cdot; \theta(t))} - \nabla_\theta Q(X_n, U_n; \theta(t)) \right) \right) \right], \end{aligned} \quad (27)$$

where  $E[\cdot]$  denotes the stationary expectation. This is again not in a form where the convergence is apparent. The problem, typical of naive Bellman error gradient methods, is that we have a conditional expectation (w.r.t.  $p(\cdot | X_n, U_n)$ ) of a product instead of a product of conditional expectations, as warranted by the actual Bellman error formula. The experience replay suggested above does one of the conditional expectations ahead of time, albeit approximately, and therefore renders (approximately) the expression a product of conditional expectations. Observe that this is so because we average over past traces  $(X_m, U_m, X_{m+1})$  where  $X_m, U_m$  are fixed at the current  $X_n, U_n$ , so that it is truly a Monte Carlo evaluation of a conditional expectation. If we were to average over such traces without fixing  $X_n, U_n$ , we would get the o.d.e.

$$\begin{aligned} \dot{\theta}(t) &= E \left[ r(X_n, U_n) + \gamma \max_v Q(X_{n+1}, v; \theta(t)) - Q(X_n, U_n; \theta_n) \right] \\ &\quad \times E \left[ \gamma \nabla_\theta Q(X_{n+1}, v; \theta(t)) \Big|_{v=\operatorname{argmax}_{Q(X_{n+1}, \cdot; \theta(t))}} - \nabla_\theta Q(X_n, U_n; \theta(t)) \right], \end{aligned} \quad (28)$$

where  $E[\cdot]$  denotes the stationary expectation. Here the problem is that the desired ‘expectation of a product of conditional expectations’ has been split into a product of expectations, which too is wrong. This discussion underscores an additional advantage of experience replay in the context of Bellman error gradient methods, over and above its traditional advantages listed earlier.

#### 4.1 Comments About ‘On-Policy’ Schemes

An ‘on-policy’ scheme has an additional complication, viz., the expectation operator in the definition of  $\bar{\mathcal{E}}(\cdot)$  itself depends on the parameter  $\theta$ . This is because the policy with which the state-action pairs  $(X_n, U_n)$  are being sampled depends at time  $n$  on the current iterate  $\theta_n$ . Therefore there is explicit  $\theta$  dependence for the probability measure  $\mu(\cdot, \cdot)$ , now written as  $\mu_\theta(\cdot, \cdot)$ . The framework of [52] is broad enough to allow this ‘iterate dependence’ and we get the counterpart of (24) with  $\mu(\cdot, \cdot)$  replaced by  $\mu_{\theta(t)}(\cdot, \cdot)$ , leading to the limiting differential inclusion

$$\dot{\theta}(t) \in -\nabla^* \bar{\mathcal{E}}_{\theta(t)}(\theta(t)). \quad (29)$$

Here  $\nabla^*$  denotes the Frechet subdifferential with respect to only the argument in parentheses, not the subscript. Hence it is not the full subdifferential and the theoretical issues we pointed out regarding DQN come back to haunt us. This is true, e.g., when you use the  $\epsilon$ -greedy policy that picks the control  $\operatorname{argmax}(Q(X_n, \cdot; \theta_n))$  with probability  $1 - \epsilon$ , and chooses a control independently and with uniform probability from  $A$ , with probability  $\epsilon$ .

Clearly, a scheme such as (29) that performs gradient descent for the stationary expectation of a parametrized cost function w.r.t. the parameter, but ignoring the parameter dependence of the stationary law itself on the parameter, is not guaranteed to converge. There are special situations such as the EM algorithm [18] where additional structure of the problem makes it work. In general, policy gradient methods based on suitable sensitivity formulas for Markov decision processes seem to provide the most flexible approach in such situations, see, e.g., [30].

## 4.2 Comparison with Double Sampling

To recapitulate, DQN can be viewed as an instance of a broader class of schemes known as Bellman error minimization or Bellman residual minimization [3]. The commonality between such schemes is that they first replace the candidate value function by a parsimoniously parametrized family of functions, e.g., linear combinations of basis functions or neural networks. The original equation then need not hold, so one seeks to minimize the ‘Bellman error’, i.e., the squared difference between the right and left hand sides of the approximate Bellman quation. Its gradient involves a product of conditional expectations. If one uses the naive strategy of replacing them by evaluation at actual samples, the gradient of the resulting ‘empirical Bellman error’ leads to an (approximate) expectation of a product in the averaged dynamics where it should have been the expectation of a product of conditional expectations. That is, product and conditional expectation get interchanged, causing bias to creep in. In fact, [3] already contains a way to avoid this. This is the ‘double sampling’ scheme that simulates two transitions simultaneously at each time instant for the current state-action pair. These are simulated independently with the same conditional law. One then performs the function evaluations for next state in the two terms of the product in Bellman error gradient using the two different samples thus generated. While this has been used subsequently (see, e.g., [9, 35]), it can be very awkward to implement in some simulation environments and is certainly untenable in on-line mode. Also, it increases the variance as we note below in numerical experiments. One of the contributions of the present work is to circumvent this by using a variant of experience replay. This can be executed with a single simulation run with buffered data and also has the advantage of lower variance due to averaging.

As for the mathematical analysis, the error process  $\{\epsilon(n)\}$  in the application of the strong law of large numbers to experience replay drops out and assumption (C3) becomes redundant if no experience replay is used. With pure double sampling without experience replay, we have only the martingale difference noise obtained by subtracting from the right hand side of the iteration its one step conditional expectation. This will be a little different from the martingale difference noise  $\{M_{n+1}(\theta_n)\}$  above due to the additional simulated transition and performance will have higher variance.

A recent work [39] treats the empirical Bellman error as a deterministic function of the parameter and minimizes it using the full gradient as described here. It does not, however, use either double sampling or experience replay and therefore retains the problem of replacing a product of conditional expectations by conditional expectation of a product.

For deterministic systems, double sampling is redundant as there is no conditional expectation in the Bellman equation. Experience replay may still be desirable for its other advantages mentioned earlier, but is not required on above grounds.

## 5 Numerical Results

In this section, we compare on two realistic examples the performance of FG-DQN with respect to that of the standard DQN scheme [32]. In particular, we investigate the behaviour of Bellman error, Hamming distance from the optimal policy (if the optimal policy is known) and the average reward. The pseudo-code for FG-DQN is described in Algorithm 1.

### 5.1 Forest Management Problem

Consider a Markov decision process framework for a simple forest management problem [14, 16]. The objective is to maintain an old forest for wildlife and make money by selling the cut wood. We consider discounted infinite horizon discrete-time problem. The state of the forest at time  $n$  is represented by  $X_n \in \{0, 1, 2, 3, \dots, M\}$  where the value of the state represents the age of the forest; 0 being the youngest and  $M$  being the oldest. The forest is managed by two actions: ‘Wait’ and ‘Cut’. An action is applied at each time at the beginning of the time slot. If we apply the action ‘Cut’ at any state, the forest will return to its youngest age, i.e., state 0. On the other hand, when the action ‘Wait’ is applied, the forest will grow and move to the next state if no fire occurred. Otherwise, with probability  $p$ , the fire burns the forest after applying the ‘Wait’ action, leaving it at its youngest age (state 0). Note that if the forest reaches its maximum age, it will remain there unless there is a fire or action ‘Cut’ is performed. Lastly, we only get a reward when the ‘Cut’ action is performed. In this case, the reward is equal to the age of the forest. There is no reward for the action ‘Wait’.

**Algorithm 1:** Full Gradient DQN (FG-DQN)

---

**Input:** replay memory  $\mathcal{D}$  of size  $M$ , minibatch size  $B$ , number of episodes  $N$ , maximal length of an episode  $T$ , discount factor  $\gamma$ , exploration probability  $\epsilon$ .

Initialise the weights  $\theta$  randomly for the Q-Network.

**for**  $Episode = 1$  to  $N$  **do**

- Receive initial observation  $s_1$ .
- for**  $n = 1$  to  $T$  **do**

  - if**  $Uni[0,1] < \epsilon$  **then**
  - | Select action  $U_n$  at random.
  - else**
  - |  $U_n = Argmax_u Q(X_n, u; \theta)$
  - end**
  - Execute the action and take a step in the RL environment.
  - Observe the reward  $R_n$  and obtain next state  $X_{n+1}$ .
  - Store the tuple  $(X_n, U_n, R_n, X_{n+1})$  in  $\mathcal{D}$ .
  - Sample random minibatch of  $B$  tuples from  $\mathcal{D}$ .
  - for**  $k = 1$  to  $B$  **do**

    - | Sample all tuples  $(X_j, U_j, R_j, X_{j+1})$  with a fix state-action pair  $(X_j = X_k, U_j = U_k)$  from  $\mathcal{D}$
    - |  $Set Z_j = \begin{cases} R_j, & \text{for terminal state,} \\ R_j + \gamma \max_u Q(X_{j+1}, u; \theta), & \text{otherwise.} \end{cases}$
    - | Compute gradients and using Eq. (19) update parameters  $\theta$ .

  - end**

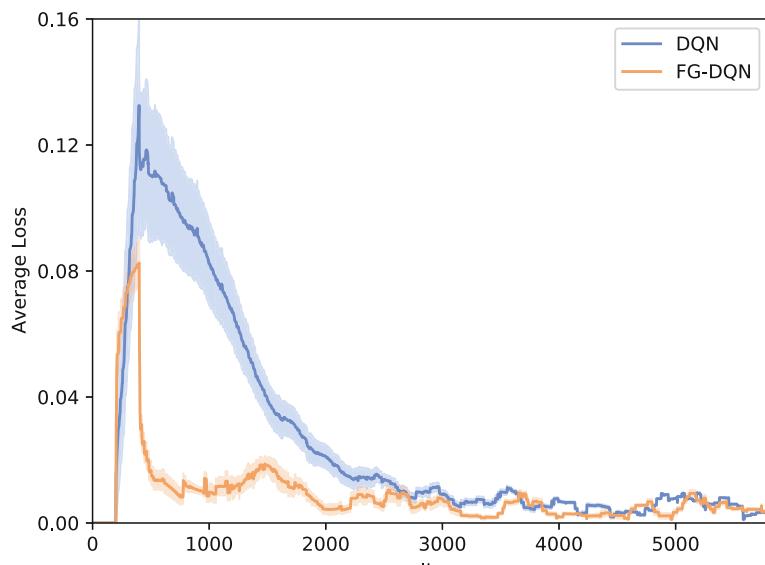
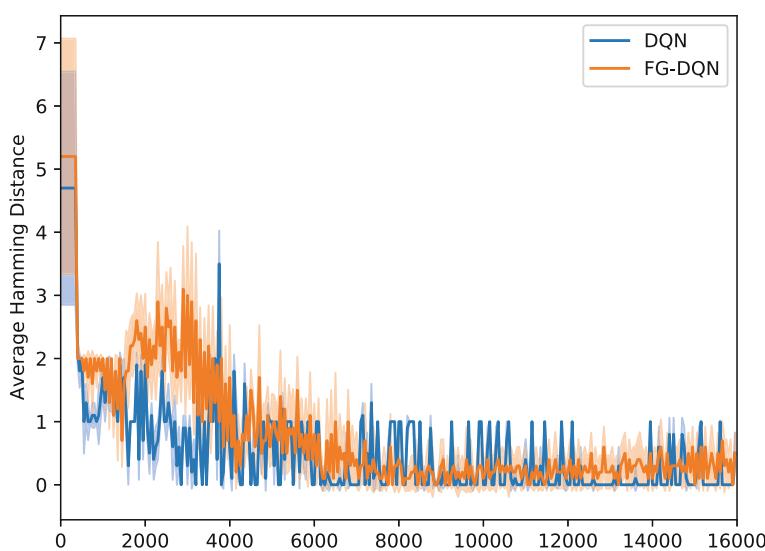
- end**

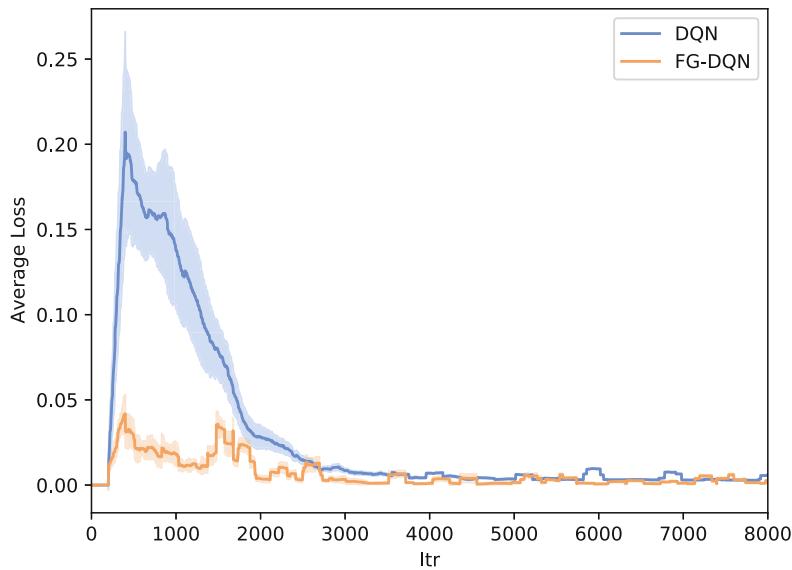
---

Since the objective is to maximize the discounted profit obtained by selling wood, we may want to keep waiting to get the maximum possible reward, but there is an increasing chance that the forest will get burned down.

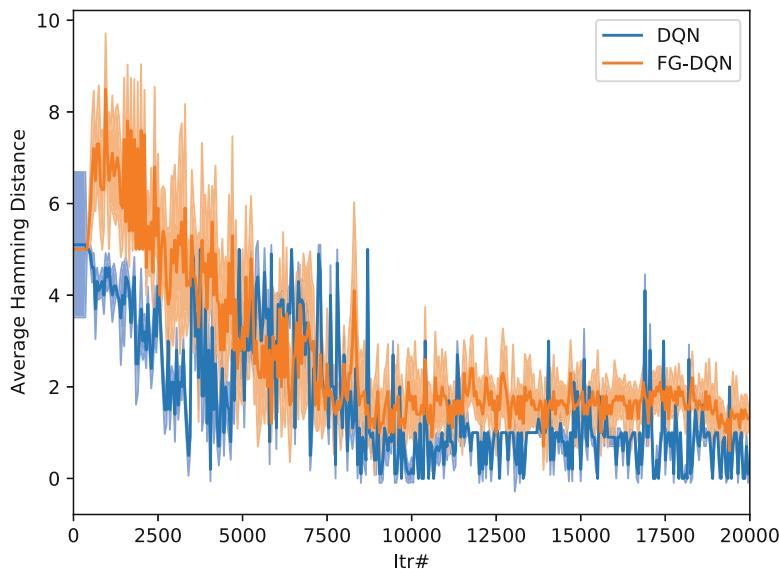
For numerical simulations, we assume that the maximum age of the forest is  $M = 10$ . Then, we implement standard DQN and FG-DQN to analyse the policy obtained from the algorithm and the Bellman error. We use a neural network with one hidden layer to approximate the Q-value. The number of neurons for this hidden layer is 2000, and we use ReLU for nonlinear activation. It has been recently advocated to use a neural network with one but very wide hidden layer [2,15]. The input to the neural network is the state of the forest and the action. Furthermore, the batch size to draw the samples for the experience replay is fixed to 25. We test both the algorithms for the off-policy scheme, i.e., we run through all possible state-action pairs in round-robin fashion to train the neural network.

We run two different simulations - i) with low discounting factor  $\gamma = 0.8$  and ii) with high discounting factor  $\gamma = 0.95$ . Figure 1 depicts the simulation

(a) Average loss and 95% confidence interval for  $\gamma = 0.8$  and  $p = 0.05$ .(b) Average Hamming distance from the optimal policy for  $\gamma = 0.8$  and  $p = 0.05$ .**Fig. 1.** Forest management problem with  $\gamma = 0.8$  and  $p = 0.05$



(a) Average loss and 95% confidence interval for  $\gamma = 0.95$  and  $p = 0.01$ .



(b) Average Hamming distance from the optimal policy for  $\gamma = 0.95$  and  $p = 0.01$ .

**Fig. 2.** Forest management problem with  $\gamma = 0.95$  and  $p = 0.01$

results for case i) with forest fire probability  $p = 0.05$ . We run the experiment 10 times and plot the running average of Bellman error across iterations in Fig. 1(a). We also calculate the standard deviation of the Bellman error. The shaded region in the plot denotes the 95% confidence interval. We observe that FG-DQN converges much faster than DQN. Furthermore, the variance for FG-DQN is relatively low.

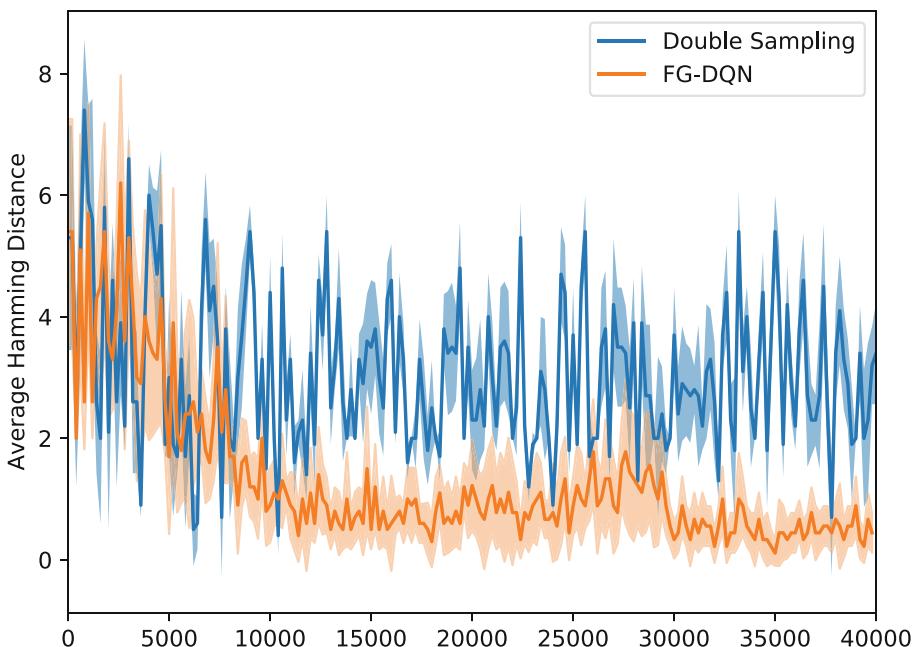
We now analyse how far the answer of each algorithm is from the optimal policy. To do this, we first find the optimal policy for this setting by policy iteration algorithm. The optimal policy has a threshold structure as follows:  $\pi^* = [0, 0, 1, 1, 1, 1, 1, 1, 1, 1]$  for  $\gamma = 0.8$  and  $p = 0.05$ .

After each iteration, we now evaluate the Q-network and calculate the Hamming distance between the current policy and the optimal policy  $\pi^*$ , which gives us the count of the number of states where optimal action is not taken. We run the simulations 10 times and plot the average Hamming distance for DQN and FG-DQN in Fig. 1(b). Note that we plot every 50th value of the average Hamming distance in the figure. It is to avoid the squeezing of rare spikes obtained at later time steps of the simulations. The shaded region denotes the 95% confidence interval for the averaged Hamming distance. We observe from the figure that the policy obtained by FG-DQN starts converging to the optimal policy at around 8000 iterations. In comparison, for DQN, we observe a lot of spikes during later iterations. The occurrence of these spikes means that there is a one-bit error in the policy obtained by DQN. Further analysis shows that the DQN policy in this case which has one bit error resembles to myopic policy  $[0, 1, 1, 1, 1, 1, 1, 1, 1, 1]$ .

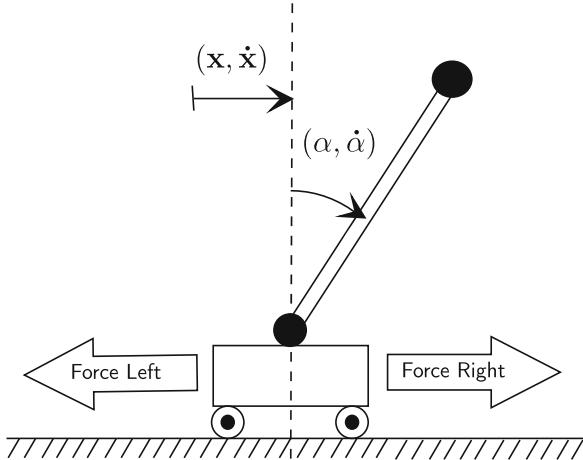
We next observe the impact of a high discounting factor on the performance of our algorithm and how well it performs as compared to the standard DQN scheme. We set  $\gamma = 0.95$  and forest fire probability  $p = 0.01$ . The optimal policy obtained by exact policy iteration for this case is  $\pi^* = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]$ . Figure 2(a) shows the mean loss for 10 simulations and the corresponding 95% confidence interval. We observe similar behaviour as before, i.e., the variance for FG-DQN is low. Figure 2(b) shows the averaged Hamming distance between the policy obtained by the algorithm and the optimal policy. It is clear from the figure that the variance for DQN is very high throughout the simulation. It means we may end up with a policy that can have a 3 or 4 bits error at the end of our simulation runs. On the other hand, FG-DQN is more stable since it shows fewer variations with the increasing number of iterations. Thus, we are more likely to get the policy with a 2 bits error on average. The shaded region in the plot shows the 95% confidence interval for 10 simulations which demonstrates that the behaviour is consistent across the multiple simulations.

Let us present an additional simulation to evaluate the performance of FG-DQN versus double sampling scheme [3]. We note that the double sampling scheme requires to generate two independent samples at each time step. This becomes difficult in many simulation environments and impossible in on-policy mode. We further note that if the underlying environment is deterministic, both these schemes become exactly identical. Therefore, in order to investigate the difference in their performance, we slightly modify the forest management

problem to have more stochasticity in its dynamics. Namely, the dynamics remain the same except for the following change. With probability  $p$ , the fire burns a fraction of the forest after applying the ‘Wait’ action. The fraction of the forest burnt follows a uniform distribution. In this simulation, we set  $p = 0.2$  and the discount factor  $\gamma = 0.9$ . The optimal policy obtained by the exact policy iteration for this case is  $\pi^* = [0, 0, 1, 1, 1, 1, 1, 1, 1, 1]$ . Figure 3 shows the comparison of averaged Hamming distance between the policy obtained by respective algorithm and the optimal policy. Note that we run the simulations 10 times and also plot the 95% confidence intervals. We observe that the policy obtained from FG-DQN approaches quicker the optimal policy and the performance is more stable. On the other hand, the double sampling policy has significant fluctuations even after 30000 iterations. The figure also shows that the double sampling policy has a 2–4 bits error at the end of our simulation runs.



**Fig. 3.** Comparison with the double sampling scheme. Average Hamming distance from the optimal policy for the forest management problem with resetting to a uniform value and with  $\gamma = 0.9$  and  $p = 0.2$ .



**Fig. 4.** Cartpole system

## 5.2 Cartpole - OpenAI Gym Model

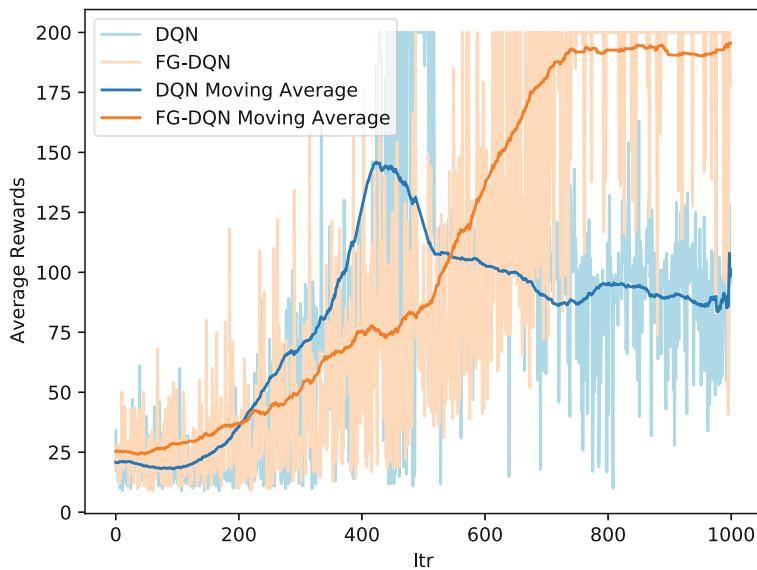
We now test our algorithm for a more complex example, the Cartpole-v0 model from OpenAI gym [12]. The environment description is as follows. The state of the system is defined by a four dimensional tuple that represents cart position  $x$ , cart velocity  $\dot{x}$ , pole angle  $\alpha$  and angular velocity  $\dot{\alpha}$  (See Fig. 4). The pole starts upright and the aim is to prevent it from falling over by pushing the cart to the left or to the right (binary action space). The cart moves without friction along the  $x$ -axis.

We run multiple simulations, each with 1500 episodes for DQN and FG-DQN. For every time-step while an episode is running, we get the reward of +1. The episode ends if any of the following conditions holds: the pole is more than  $12^\circ$  from the vertical axis, the cart moves more than 2.4 units from the centre, or the episode length is more than 200. The model is considered to be trained well when the discounted reward is greater than or equal to 195.0 over 100 consecutive trials.

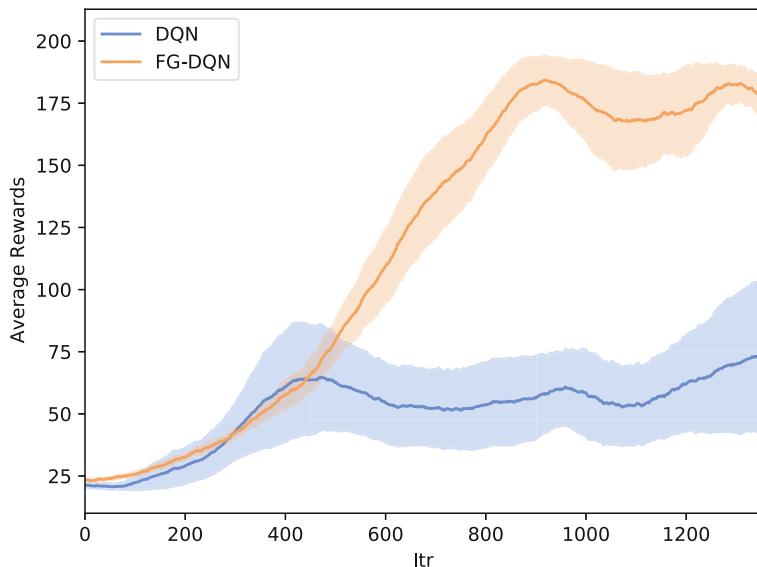
In this example, we used the ‘on-policy’ version with the popular ‘ $\epsilon$ -greedy’ scheme which picks the current guess for the optimal (i.e., the control that maximizes  $Q(X_n, \cdot; \theta_n)$ ) with probability  $1 - \epsilon$  and chooses a control uniformly with probability  $\epsilon$  for a prescribed  $\epsilon > 0$ . We use  $\epsilon = 0.1$ . As we see below, FG-DQN continues to do much better than DQN even in this on-policy scheme for which we do not have a convergence proof as yet.

We use a neural network with three hidden layers. The number of nodes for the hidden layers are 16, 32, and 32, respectively. For non-linearity, we use ReLU activation after each hidden layer.

We now compare the performances of FG-DQN and DQN for a very high discounting factor of 0.99. Note that the Cartpole example is deterministic,



(a) Average rewards for a typical single simulation run.



(b) Rewards averaged over simulations for Cartpole and 95% confidence intervals.

**Fig. 5.** Cartpole example with  $\gamma = 0.99$

meaning that for a fixed state-action pair  $(X_n, U_n)$ , the pole moves to state  $X'_n$  with probability 1. As a result, there will be no averaging in Eq. (18) and no need for ‘experience replay’. Since this example is complex with significant non-linearity, we use the batch size of 128 for both DQN and FG-DQN to update the parameters of the neural network inside one iteration.

Figure 5(a) depicts the reward behaviour for a single typical run of DQN and FG-DQN. We see that the fluctuations for reward per episode for both the algorithms are high, and thus, we also plot the moving average of rewards with a window of 100 episodes. It is clear from the figure that FG-DQN starts achieving the maximum reward of 200 after 800 episodes regularly, however, DQN hardly attains the maximum reward during 1000 episodes. To check the consistency of the behaviour of our algorithm, we run the experiment 10 times and plot the average reward and 95% confidence interval in Fig. 5(b). We see that FG-DQN performs much better than DQN with an average reward after 1000 episodes lying around 175. In comparison, the average reward for DQN is between 50 and 75.

## 6 Conclusions and Future Directions

We proposed and analyzed a variant of the popular DQN algorithm that we call Full Gradient DQN or FG-DQN wherein we also include the parametric gradient (in a generalized sense) of the target. This leads to a provably convergent scheme with sound theoretical basis which also shows improved performance over test cases. There is ample opportunity for further research in this direction, both theoretically and in terms of actual implementations. To highlight opportunities, we state here some additional remarks, which also contain a few pointers to future research directions.

1. Since the critical points are isolated, we get a.s. convergence to a single sample path dependent critical point. This situation is generic in the sense that it holds true for the problem parameters in an open dense set thereof, by a standard fact from Morse theory in the smooth case. However, connected sets of non-isolated equilibria can occur due to overparametrization and it will be interesting to develop sufficient conditions for point convergence.
2. Thanks to the addition of  $\{\xi_n\}$ , the noise in FG-DQN is ‘rich enough’ in all directions in a certain sense. One then expects it to ensure that under reasonable assumptions, the unstable equilibria, here the critical points other than local minima of the Bellman error, will be avoided with probability one. That is, a.s. convergence to a local minimum can be claimed. See Sect. 4.3 of [11] for a result of this flavor under suitable technical conditions. We expect a similar result to hold here. In practice, the extraneous noise  $\{\xi_n\}$  is usually unnecessary and the inherent numerical errors and noise of the iterations suffice.
3. We can also use approximation of the ‘max’ operator by ‘softmax’, i.e., by picking the control with a probability distribution that concentrates on

- argmax and depends smoothly on the parameter  $\theta$ . Then we can work with a legitimate gradient in place of a set-valued map in the o.d.e. limit, at the expense of picking up an additional bounded error term. Then the convergence to a small neighborhood of an equilibrium may be expected, the size of which will be dictated in turn by the bound on this error, see, e.g., [37]. There is a similar issue if we drop (C3) and let a persistent small error due to the use of approximate conditional expectation by experience replay remain.
4. Working with nondifferentiable nonlinearities such as ReLU raises further technical issues in analysis that need to be explored. This will require further use of non-smooth analysis.
  5. As we have pointed out while describing our numerical experiments on the cartpole example, FG-DQN gives a significantly better performance than DQN, in an ‘on-policy’ scenario for which we do not have rigorous theory yet. This is another promising and important research direction for the future.

**Acknowledgement.** The authors are greatly obliged to Prof. K. S. Mallikarjuna Rao for pointers to the relevant literature on non-smooth analysis. The work of VSB was supported in part by an S. S. Bhatnagar Fellowship from the Council of Scientific and Industrial Research, Government of India. The work of KP and KA is partly supported by ANSWER project PIA FSN2 (P15 9564-266178/DOS0060094) and the project of Inria - Nokia Bell Labs “Distributed Learning and Control for Network Analysis”. This work is also partly supported by the project IFC/DST-Inria-2016-01/448 “Machine Learning for Network Analytics”.

## Appendix: Elements of Non-smooth Analysis

The (Frechet) sub/super-differentials of a map  $f : \mathcal{R}^d \mapsto \mathcal{R}$  are defined by

$$\begin{aligned}\partial^- f(x) &:= \left\{ z \in \mathcal{R}^d : \liminf_{y \rightarrow x} \frac{f(y) - f(x) - \langle z, y - x \rangle}{|x - y|} \geq 0 \right\}, \\ \partial^+ f(x) &:= \left\{ z \in \mathcal{R}^d : \limsup_{y \rightarrow x} \frac{f(y) - f(x) - \langle z, y - x \rangle}{|x - y|} \leq 0 \right\},\end{aligned}$$

respectively. Assume  $f, g$  is Lipschitz. Some of the properties of  $\partial^\pm f$  are as follows.

- **(P1)** Both  $\partial^- f(x), \partial^+ f(x)$  are closed convex and are nonempty on dense sets.
- **(P2)** If  $f$  is differentiable at  $x$ , both equal the singleton  $\{\nabla f(x)\}$ . Conversely, if both are nonempty at  $x$ ,  $f$  is differentiable at  $x$  and they equal  $\{\nabla f(x)\}$ .
- **(P3)**  $\partial^- f + \partial^- g \subset \partial^-(f + g)$ ,  $\partial^+ f + \partial^+ g \subset \partial^+(f + g)$ .

The first two are proved in [4], pp. 30-1. The third follows from the definition. Next consider a continuous function  $f : \mathcal{R}^d \times B \mapsto \mathcal{R}$  where  $B$  is a compact metric space. Suppose  $f(\cdot, y)$  is continuously differentiable uniformly w.r.t.  $y$ . Let

$\nabla_x f(x, y)$  denote the gradient of  $f(\cdot, y)$  at  $x$ . Let  $g(x) := \max_y f(x, y)$ ,  $h(x) := \min_y f(x, y)$  with

$$M(x) := \{\nabla_x f(x, y), y \in \text{Argmax} f(x, \cdot)\}$$

and

$$N(x) := \{\nabla_x f(x, y), y \in \text{Argmin} f(x, \cdot)\}.$$

Then  $N(x)$ ,  $M(x)$  are compact nonempty subsets of  $B$  which are upper semi-continuous in  $x$  as set-valued maps. We then have the following general version of Danskin's theorem [17]:

- **(P4)**  $\partial^- g(x) = \overline{\text{co}}(M(x))$ ,  $\partial^+ g(x) = y$  if  $M(x) = \{y\}$ ,  $= \phi$  otherwise, and  $g$  has a directional derivative in any direction  $z$  given by  $\max_{y \in M(x)} \langle y, z \rangle$ .
- **(P5)**  $\partial^+ h(x) = \overline{\text{co}}(N(x))$ ,  $\partial^- h(x) = y$  if  $N(x) = \{y\}$ ,  $= \phi$  otherwise, and  $h$  has a directional derivative in any direction  $z$  given by  $\min_{y \in N(x)} \langle y, z \rangle$ .

The latter is proved in [4], pp. 44–6, the former follows by a symmetric argument.

## References

1. Agarwal, A., Kakade, S.M., Jason D.L., Mahajan, G.: Optimality and approximation with policy gradient methods in Markov decision processes. In: Conference on Learning Theory, PMLR, pp. 64–66 (2020)
2. Agazzi, A., Lu, J.: Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. arXiv preprint [arXiv:2010.11858](https://arxiv.org/abs/2010.11858) (2020)
3. Baird, L.: Residual algorithms: reinforcement learning with function approximation. In: Machine Learning Proceedings, vol. 30–37 (1995)
4. Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Birkhäuser, Boston (2018)
5. Barto, A.G., Sutton, R.S., Anderson, C.W.: Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans. Syst. Man Cybern. Syst. **5**, 834–846 (1983)
6. Benveniste, A., Metivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations. Springer, Heidelberg (1991). [https://doi.org/10.1007/978-3-642-75894-2\\_9](https://doi.org/10.1007/978-3-642-75894-2_9)
7. Bertsekas, D.P.: Reinforcement Learning and Optimal Control. Athena Scientific (2019)
8. Bhandari, J., Russo, D.: Global optimality guarantees for policy gradient methods. arXiv preprint [arXiv:1906.01786](https://arxiv.org/abs/1906.01786) (2019)
9. Bhatnagar, S., Borkar, V.S., Prabuchandran, K.J.: Feature search in the Grassmannian in online reinforcement learning. IEEE J. Sel. Top. Signal Process. **7**(5), 746–758 (2013)
10. Borkar, V.S.: Probability Theory: An Advanced Course. Springer, New York (1995)
11. Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Hindustan Publishing Agency, New Delhi, and Cambridge University Press, Cambridge, UK (2008)

12. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: OpenAI Gym. ArXiv preprint [arXiv:1606.01540](https://arxiv.org/abs/1606.01540) (2016)
13. Cai, Q., Yang, Z., Lee, J.D., Wang, Z.: Neural temporal-difference learning converges to global optima. *Adv. Neural Inf. Process. Syst.* **32** (2019)
14. Chadès, I., Chapron, G., Cros, M.J., Garcia, F., Sabbadin, R.: MDPtoolbox: a multi-platform toolbox to solve stochastic dynamic programming problems. *Ecography* **37**, 916–920 (2014)
15. Chizat, L., Bach, F.: On the global convergence of gradient descent for over-parameterized models using optimal transport. In: *Proceedings of Neural Information Processing Systems*, pp. 3040–3050 (2018)
16. Couture, S., Cros, M.J., Sabbadin, R.: Risk aversion and optimal management of an uneven-aged forest under risk of windthrow: a Markov decision process approach. *J. For. Econ.* **25**, 94–114 (2016)
17. Danskin, J.M.: The theory of max-min, with applications. *SIAM J. Appl. Math.* **14**, 641–664 (1966)
18. Delyon, B., Lavielle, M., Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm. *Ann. Stat.* **27**, 94–128 (1999)
19. Florian, R.V.: Correct equations for the dynamics of the cart-pole system. Romania, Center for Cognitive and Neural Studies (Coneural) (2007)
20. François-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G., Pineau, J.: An introduction to deep reinforcement learning. *Found. Trends Mach. Learn.* **11**(3–4), 219–354 (2018)
21. Gordon, G. J.: Stable fitted reinforcement learning. In: *Advances in Neural Information Processing Systems*, pp. 1052–1058 (1996)
22. Gordon, G. J.: Approximate solutions to Markov decision processes. Ph.D. Thesis, Carnegie-Mellon University (1999)
23. Gu, S., Holly, E., Lillicrap, T., Levine, S.: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 3389–3396 (2017)
24. Haarnoja, T., Ha, S., Zhou, A., Tan, J., Tucker, G., Levine, S.: Learning to walk via deep reinforcement learning. ArXiv preprint [arXiv:1812.11103](https://arxiv.org/abs/1812.11103) (2018)
25. Jaakola, T., Jordan, M.I., Singh, S.P.: On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.* **6**, 1185–1201 (1994)
26. Jonsson, A.: Deep reinforcement learning in medicine. *Kidney Diseases* **5**, 18–22 (2019)
27. Lin, L.J.: Self-improving reactive agents based on reinforcement learning, planning and teaching. *Mach. Learn.* **8**(3–4), 293–321 (1992)
28. Lin, L.-J.: Reinforcement learning for robots using neural networks. Ph.D. Thesis School of Computer Science, Carnegie-Mellon University, Pittsburgh (1993)
29. Luong, N.C., Hoang, D.T., Gong, S., Niyato, D., Wang, P., Liang, Y.C., Kim, D.I.: Applications of deep reinforcement learning in communications and networking: a survey. *IEEE Commun. Surv. Tutor.* **21**, 3133–3174 (2019)
30. Marbach, P., Tsitsiklis, J.N.: Simulation-based optimization of Markov reward processes. *IEEE Trans. Automat. Contr.* **46**, 191–209 (2001)
31. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with deep reinforcement learning. arXiv preprint [arXiv:1312.5602](https://arxiv.org/abs/1312.5602) (2013)
32. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S.: Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015)

33. Peng, X.B., Berseth, G., Yin, K., van de Panne, M.: Deeploco: dynamic locomotion skills using hierarchical deep reinforcement learning. *ACM Trans. Graph.* **36**, 1–13 (2017)
34. Popova, M., Isayev, O., Tropsha, A.: Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**(7), eaap7885 (2018)
35. Prabuchandran, K.J., Bhatnagar, S., Borkar, V.S.: Actor-critic algorithms with online feature adaptation. *ACM Trans. Model. Comput. Simul. (TOMACS)* **26**(4), 1–26 (2016)
36. Qian, Y., Wu, J., Wang, R., Zhu, F., Zhang, W.: Survey on reinforcement learning applications in communication networks. *J. Commun. Netw.* **4**, 30–39 (2019)
37. Ramaswamy, A., Bhatnagar, S.: Analysis of gradient descent methods with nondiminishing bounded errors. *IEEE Trans. Automat. Contr.* **63**, 1465–1471 (2018)
38. Riedmiller, M.: Neural fitted Q iteration—first experiences with a data efficient neural reinforcement learning method. *Machine Learning: ECML*, pp. 317–328 (2005)
39. Saleh, E., Jiang, N.: Deterministic Bellman residual minimization. In: *Proceedings of Optimization Foundations for Reinforcement Learning Workshop at NeurIPS* (2019)
40. Schaul, T., Quan, J., Antonoglou, I., Silver, D.: Prioritized experience replay. arXiv preprint [arXiv:1511.05952](https://arxiv.org/abs/1511.05952) (2015)
41. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016)
42. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Hassabis, D.: A general reinforcement learning algorithm that masters chess, Shogi, and Go through self-play. *Science* **362**, 1140–1144 (2018)
43. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*, 2nd edn. MIT Press, Cambridge (2018)
44. Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: *Neural Information Processing Systems Proceedings*, pp. 1057–1063 (1999)
45. Tsitsiklis, J.N.: Asynchronous stochastic approximation and Q-learning. *Mach. Learn.* **16**, 185–202 (1994)
46. Tsitsiklis, J.N., Van Roy, B.: An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Contr.* **42**, 674–690 (1997)
47. van Hasselt, H.: Double Q-learning. *Adv. Neural. Inf. Process. Syst.* **23**, 2613–2621 (2010)
48. van Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double Q-learning. In: *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, vol. 30, pp. 2094–2100 (2016)
49. Watkins, C.J.C.H.: Learning from delayed rewards. Ph.D. Thesis, King's College, University of Cambridge, UK (1989)
50. Watkins, C.J., Dayan, P.: Q-learning. *Mach. Learn.* **8**(3–4), 279–292 (1992)
51. Xiong, Z., Zhang, Y., Niyato, D., Deng, R., Wang, P., Wang, L.C.: Deep reinforcement learning for mobile 5G and beyond: fundamentals, applications, and challenges. *IEEE Veh. Technol. Mag.* **14**, 44–52 (2019)
52. Yaji, V.G., Bhatnagar, S.: Stochastic recursive inclusions with non-additive iterate-dependent Markov noise. *Stochastics* **90**, 330–363 (2018)



# On Finite Approximations to Markov Decision Processes with Recursive and Nonlinear Discounting

Fan Deng<sup>1</sup>, Xin Guo<sup>2</sup>, and Yi Zhang<sup>1(✉)</sup>

<sup>1</sup> Department of Mathematical Sciences, University of Liverpool, Liverpool, UK  
`{Fan.Deng,yi.zhang}@liverpool.ac.uk`

<sup>2</sup> School of Economics and Management, Tsinghua University, Beijing, China  
`guoxin5@sem.tsinghua.edu.cn`

**Abstract.** In this paper, finite approximation schemes are justified for Markov decision processes in Borel spaces with recursive and nonlinear discounting. Explicit error bounds are obtained in terms of the system primitives. This allows one to solve the original problem approximately up to any given accuracy, by solving a sequence of problems in finite spaces.

**Keywords:** Finite approximations · Error bound · Markov decision processes · Nonlinear discounting

**AMS (2020) Subject Classification:** Primary 90C40 · Secondary 90C59

## 1 Introduction

In this paper, we justify a finite approximation scheme to solve numerically Markov decision processes (MDP) with recursive and nonlinear discounting. The deterministic dynamic programming problem (as a special MDP model) with recursive and nonlinear discounting was considered in [11, 12], which found many applications to economics. A more recent development is [5], which also demonstrates the connections of this model with several other relevant problems.

There are two possible ways of extending this model from the deterministic to the stochastic dynamic programming setup. The latter term is used interchangeably below with an MDP. One way of extension was carried out in [10], where the total cost is discounted firstly along each sample path and then the expectation is applied. For the resulting MDP problem, in general, stationary policies do not form a sufficient class. A second possible extension was published more recently, see [3], where the conditional expected discounted cost is aggregated recursively. In [3], it was shown that stationary optimal policies exist under the conditions imposed therein, along with some meaningful examples in e.g., optimal growth problems. We mention that the models in both [3, 10] cover the standard linear discounting as a special case.

The purpose of this paper is to justify a finite approximation scheme with an explicit error bound to the MDP problem considered in [3]. Finite approximations to MDP models with standard discounting have been considered intensively in the literature, and we confine ourselves to the most relevant ones here. Most early literature provides convergence results without an explicit estimate of the error bound. For models with denumerable state and action spaces, see [2, 15, 17] and the discussion therein. A most recent development is [13]. Finite approximations to MDP models in Borel spaces with standard linear discounting were considered and justified in e.g., [4], where an explicit error bound was provided for the underlying approximation scheme. More recent developments in this direction can be found in e.g., [6, 7, 16]. An error bound is desirable for practical implementations and computations, but establishing it usually requires stronger conditions on the model.

In the present paper, we extend the method in [6, 7] to MDP problems with recursive and nonlinear discounting. The model considered here is with state and action spaces being both Borel spaces. Besides the compactness-continuity and growth conditions imposed in [3], which are needed for establishing basic optimality results (solvability and the dynamic programming equation), we assume further that the model has Lipschitz continuous initial data. Like in [4, 6, 7], this allows us to obtain an explicit error bound. The imposed conditions are satisfied by a version of the stochastic optimal growth problem formulated in [3], which can also serve as a motivation of this paper.

The rest of the paper is organized as follows. In Sect. 2, we describe the model, impose the conditions on it, as well as briefly present the relevant facts established in [3]. In Sect. 3 we present the main results, whose proofs are postponed to Sect. 5. An example is presented in Sect. 4 to illustrate the verification of the imposed conditions.

## 2 Model Description

In this section, we present the concerned model, and introduce the conditions on the system primitives. To ease the reading, we also formulate the relevant statements and facts, primarily from [3], which will be referred to in the subsequent sections. In what follows, unless stated otherwise, measurability is understood with respect to underlying Borel  $\sigma$ -algebra, and  $\delta_x$  denotes the Dirac measure.

The system primitives of our model are as follows:

- $X$  is the state space, assumed to be a locally compact (topological) Borel space. A (topological) Borel space is a Borel subset (endowed with the relative topology) of a complete separable metric space. Let  $d_X$  be the metric on  $X$ , and we endow  $X$  with its Borel  $\sigma$ -algebra  $\mathcal{B}(X)$ .
- $A$  is the action space, assumed to be a locally compact Borel space, with the metric  $d_A$  and the Borel  $\sigma$ -algebra  $\mathcal{B}(A)$ .
- $\mathbf{A}(x) \in \mathcal{B}(A)$  is the nonempty set of admissible actions at the state  $x \in X$ . That is,  $\mathbf{A}(x)$  defines a multifunction on  $X$ , denoted by  $\mathbf{A}$ . Assume that

$$D := \{(x, a) : x \in X, a \in \mathbf{A}(x)\}$$

is a Borel subset of  $X \times A$  such that it contains the graph of some measurable mapping from  $X$  to  $A$ , say  $f^\infty$ . Here and below,  $X \times A$  is endowed with the metric  $d_X + d_A$  defined by  $d_X(x_1, x_2) + d_A(a_1, a_2)$  for all  $x_1, x_2 \in X$  and  $a_1, a_2 \in A$ .

- $q(dy|x, a)$  is a stochastic kernel on  $X$  given  $(x, a) \in D$ , representing the controlled transition probability.
- $u$  is an  $\mathbb{R}$ -valued measurable function on  $D$  with  $u(x, a)$  representing the utility associated with the current state  $x \in X$  and action  $a \in \mathbf{A}(x)$ .
- $\delta$  is an  $\mathbb{R}$ -valued increasing (and thus measurable) function on  $\mathbb{R}$  such that  $\delta(0) = 0$ , with  $\delta(v)$  representing the discounted value if the continuing value of the utility at the next stage is  $v$ . (The standard linear discounting with a constant discount factor  $\beta \in (0, 1)$  is retrieved if  $\delta(v) = \beta v$ .)

Let us describe the controlled and controlling processes as follows. Let  $H_0 := X$  and  $H_n := D^n \times X$  for all  $1 \leq n < \infty$ . We put  $H := D^\infty$  as the countably infinite product. For each  $1 \leq n < \infty$ ,  $H_n$  is a Borel space and is endowed with the corresponding Borel  $\sigma$ -algebra  $\mathcal{B}(H_n)$ . The similar assertion applies to  $H$ .

- Definition 1.** (a) A policy  $\pi = \{\pi_n\}_{n \geq 0}$  is given by a sequence of  $A$ -valued measurable mappings  $\pi_n$  on  $H_n$  such that  $\pi_n(h_n) \in A(x_n)$  for each  $h_n = (x_0, a_0, x_1, a_1, \dots, x_n) \in H_n$ .
- (b) A policy  $\pi = \{\pi_n\}_{n \geq 0}$  is called Markov and is written as  $\{f_n\}_{n \geq 0}$  with  $f_n$  being measurable on  $X$  if  $\pi_n(h_n) = f_n(x_n)$  for all  $n \geq 0$  and  $h_n = (x_0, a_0, x_1, a_1, \dots, x_n) \in H_n$ .
- (c) A policy  $\pi = \{\pi_n\}_{n \geq 0}$  is called stationary and is written as  $f$  if for some measurable mapping  $f$  on  $X$ ,  $\pi_n(h_n) = f(x_n)$  for all  $n \geq 0$  and  $h_n = (x_0, a_0, x_1, a_1, \dots, x_n) \in H_n$ .

The above policies are called pure or deterministic. For simplicity we do not consider randomized strategies, in which case  $\pi_n$  would be stochastic kernels on  $A$  given  $h_n = (x_0, a_0, \dots, x_n) \in H_n$  concentrated on  $A(x_n)$ .

Take  $(H, \mathcal{B}(H))$  as the sample space. Given an initial state  $x \in X$  and policy  $\pi = \{\pi_n\}_{n \geq 0}$ , by the Ionescu-Tulcea theorem, there is a unique probability measure  $P_x^\pi$  defined thereon such that

$$P_x^\pi(x_0 \in dy) = \delta_x(dy);$$

$$P_x^\pi(x_{n+1} \in dy|h_n, a_n) = q(dy|x_n, a_n); P_x^\pi(a_n \in da|h_n) = \delta_{\pi_n(h_n)}(da) \quad \forall n \geq 0.$$

Here we use interchangeably  $x_n$  and the random element defined by  $X_n(h) = x_n$  for each  $h = (x_0, a_0, \dots, x_n, a_n, \dots) \in H$ , and the same concerns the use of  $a_n$ . The context excludes any confusion.

We shall impose the following conditions to guarantee the performance measure introduced below to be well defined.

**Condition 1.** There is some  $[1, \infty)$ -valued measurable function  $w$  on  $X$  such that the following are verified.

- (a) For some constant  $b \geq 0$ ,  $|u(x)| \leq bw(x)$  for all  $x \in X$ .

- (b) There is some  $[0, \infty)$ -valued increasing and continuous function  $\gamma$  on  $[0, \infty)$  satisfying
- (i)  $\gamma(0) = 0$  and  $\gamma(x) < x$  for all  $x \in (0, \infty)$ .
  - (ii)  $|\delta(x_1) - \delta(x_2)| \leq \gamma(|x_1 - x_2|)$  for all  $x_1, x_2 \in \mathbb{R}$ .
  - (iii)  $\gamma(x_1 + x_2) \leq \gamma(x_1) + \gamma(x_2)$  for all  $x_1, x_2 \in [0, \infty)$ .
  - (iv)  $\gamma(w(x)y) \leq w(x)\gamma(y)$  for all  $x \in X$  and  $y \in [0, \infty)$ .
  - (v) For some  $\alpha \in (0, \infty)$ ,  $\int_X w(y)q(dy|x, a) \leq \alpha w(x)$  for all  $(x, a) \in D$ , and  $\alpha\gamma(y) < y$  for all  $y \in (0, \infty)$ .

In the forthcoming discussions, we assume that Condition 1 is satisfied unless stated otherwise. Let us list down some immediate consequences of the above condition.

Condition 1(b, i) implies that

$$\lim_{n \rightarrow \infty} \gamma^{(n)}(y) = 0 \quad \forall y \in [0, \infty),$$

where  $\gamma^{(n)}(y) := \gamma(\gamma^{(n-1)})(y)$  for each  $n \geq 2$ . Indeed, this is automatic if  $y = 0$  for  $\gamma(0) = 0$ . Consider  $y > 0$ . Since  $\gamma^{(n)}(y)$  decreases in  $n$ ,  $\lim_{n \rightarrow \infty} \gamma^{(n)}(y) = c \geq 0$  exists. If  $c > 0$ , then  $c > \gamma(c) = \gamma(\lim_{n \rightarrow \infty} \gamma^{(n)}(y)) = \lim_{n \rightarrow \infty} \gamma^{(n+1)}(y) = c$ , which is a contradiction. Now Condition 1(b, i, v) implies

$$\lim_{n \rightarrow \infty} \tilde{\gamma}^{(n)}(y) = 0 \quad \forall y \in [0, \infty), \tag{1}$$

for

$$\tilde{\gamma} := \alpha\gamma. \tag{2}$$

Condition 1(b, iii) asserts that  $\gamma$  is a sub-additive function on  $[0, \infty)$ , which together with Condition 1(b, i), implies that the next result applies to  $\gamma$  and  $\tilde{\gamma}$ .

**Proposition 1.** *Let  $\psi$  be a  $[0, \infty)$ -valued increasing sub-additive continuous function on  $[0, \infty)$  satisfying  $\psi(y) < y$  for all  $y \in (0, \infty)$  (so that  $\psi(0) = 0$ ). Define for all  $y \in [0, \infty)$ ,*

$$\underline{\psi}_0(y) := 0, \quad \underline{\psi}_1(y) := \underline{\psi}(y) := y; \quad \underline{\psi}_{n+1}(y) := y + \psi(\underline{\psi}_n(y)) \quad \forall n \geq 0, \tag{3}$$

*Then for each  $y \in [0, \infty)$ ,  $\underline{\psi}_n(y)$  is increasing in  $n$ , and*

$$\underline{\psi}_\infty(y) := \lim_{n \rightarrow \infty} \underline{\psi}_n(y)$$

*exists and is finite. In particular,  $\underline{\psi}_\infty(y) = y + \psi(\underline{\psi}_\infty(y))$  for all  $n \geq 0$ . Moreover,  $\underline{\psi}_\infty$  is continuous on  $[0, \infty)$ .*

*Proof.* See Lemma 4.6 of [3]. (For the last assertion, by inspecting the proof of Lemma 4.6 of [3], we see that  $\underline{\psi}_n$  converges to  $\underline{\psi}_\infty$  uniformly on each compact subset, and thus the continuity of  $\underline{\psi}_\infty$  follows from the continuity of  $\underline{\psi}_n$ .)  $\square$

One can recognize that  $\underline{\psi}_n(z) = z + \psi(z + \psi(z + \dots + \psi(z \dots)))$ , where  $z$  appears  $n$  times. Proposition 1 will be instrumental on several occasions in the main text below. In particular, we may legitimately consider

$$\begin{aligned}\underline{\gamma}_\infty(z) &:= \lim_{n \rightarrow \infty} \underline{\gamma}_n(z) = \sup_{n \geq 1} \underline{\gamma}_n(z) \in [0, \infty) \quad \forall z \in [0, \infty); \\ \tilde{\underline{\gamma}}_\infty(z) &:= \lim_{n \rightarrow \infty} \tilde{\underline{\gamma}}_n(z) = \sup_{n \geq 1} \tilde{\underline{\gamma}}_n(z) \in [0, \infty) \quad \forall z \in [0, \infty)\end{aligned}$$

with  $\underline{\gamma}_n(z)$  and  $\tilde{\underline{\gamma}}_n(z)$  as defined in (3) with  $\gamma$  and  $\tilde{\gamma}$  in lieu of  $\psi$ .

For any  $[1, \infty)$ -valued measurable function  $w$  on a (measurable) space  $E$ , let  $\mathbb{B}_w(E)$  be the collection of measurable functions  $v$  on  $E$  such that  $\|v\|_w := \sup_{x \in E} \frac{|v(x)|}{w(x)} < \infty$ . Such a function  $v$  will be called  $w$ -bounded (on  $E$ ). Condition 1(a) asserts that  $u$  is  $w$ -bounded with  $\|u\|_w \leq b$ .

To introduce the performance measure of a policy  $\pi = \{\pi_n\}_{n \geq 0}$ , for each  $n \geq 0$ , we consider the operators  $T_{\pi_n}$  and  $Q_{\pi_n}^\gamma$  defined as follows. For each  $w$ -bounded function  $v$  on  $H_{n+1}$  ( $n \geq 0$ ),

$$\begin{aligned}T_{\pi_n} v(h_n) &:= u(x_n, \pi_n(h_n)) + \int_X \delta(v(h_n, \pi_n(h_n), x_{n+1})) q(dx_{n+1} | x_n, \pi_n(h_n)), \\ Q_{\pi_n}^\gamma v(h_n) &:= \int_X \gamma(v(h_n, \pi_n(h_n), x_{n+1})) q(dx_{n+1} | x_n, \pi_n(h_n)) \quad \forall h_n \in H_n.\end{aligned}\quad (4)$$

Condition 1 implies that  $T_{\pi_n}|v|$  is  $w$ -bounded on  $H_n$ . Consequently,

$$U_1^\pi(x) := T_{\pi_0} 0(x); \quad U_n^\pi(x) := T_{\pi_0} T_{\pi_1} \dots T_{\pi_{n-1}} 0(x) \quad \forall n \geq 2, \quad x \in X$$

are well defined and in  $\mathbb{B}_w(X)$ . In fact, the next upper bound of the  $w$ -norm of  $U_n^\pi$  will be needed below.

**Proposition 2.** Suppose Condition 1 is satisfied. For each  $n \geq 1$  and policy  $\pi$ ,

$$|U_n^\pi(x)| \leq w(x) \tilde{\underline{\gamma}}_n(\|u\|_w) \leq w(x) \tilde{\underline{\gamma}}_\infty(\|u\|_w) \quad \forall x \in X.$$

*Proof.* See the proof of Lemma 5.3 of [3]. □

The above-defined  $U_n^\pi$  is called the  $n$ -stage total recursively discounted utility of the policy  $\pi$ , or say the total recursively discounted utility for the  $n$ -stage problem. The discounting is non-linear. In case  $\pi = \{f_n\}_{n \geq 0}$  is a Markov policy, it is informative to write down that

$$\begin{aligned}U_3^\pi(x) &= u(x, f_0(x)) \\ &+ \int_X \delta \left( u(x_1, f_1(x_1)) + \int_X \delta(u(x_2, f_2(x_2))) q(dx_2 | x_1, f_1(x_1)) \right) q(dx_1 | x, f_0(x)).\end{aligned}$$

The next proposition allows us to define the infinite horizon total recursively discounted utility of a policy as the limit of the  $n$ -stage performance measure.

**Proposition 3.** Suppose Condition 1 is satisfied. Then

$$U^\pi(x) := \lim_{n \rightarrow \infty} U_n^\pi(x) \quad \forall x \in X,$$

exists for each policy  $\pi$ , so that  $|U^\pi(x)| \leq w(x)\tilde{\gamma}_\infty(\|u\|_w)$  for all  $x \in X$ . In particular,  $U^\pi \in \mathbb{B}_w(X)$  for each policy  $\pi$ . Moreover, the convergence also holds in  $\mathbb{B}_w(X)$ : in fact,  $\|U^\pi - U_n^\pi\|_w \leq \tilde{\gamma}^{(n)}(\tilde{\gamma}_\infty(\|u\|_w)) \rightarrow 0$  as  $n \rightarrow \infty$ .

*Proof.* See Lemma 5.3 of [3] for the convergence, and Proposition 1 and (5.5) of [3] for the bound of  $|U^\pi|$  and  $\|U^\pi - U_n^\pi\|_w$ .  $\square$

According to Proposition 3, we may legitimately consider

$$U^\pi(x) = \lim_{n \rightarrow \infty} T_{\pi_0} T_{\pi_1} \dots T_{\pi_{n-1}} 0(x) \quad \forall x \in X.$$

It is useful to observe that in the above definition of  $U^\pi$ , we may replace 0 with any function  $v \in \mathbb{B}_w(X)$ , as stated in the next lemma.

**Lemma 1.** Suppose Condition 1 is satisfied. Then for any policy  $\pi$  and  $v \in \mathbb{B}_w(X)$ ,

$$U^\pi(x) = \lim_{n \rightarrow \infty} T_{\pi_0} T_{\pi_1} \dots T_{\pi_{n-1}} v(x) \quad \forall x \in X.$$

*Proof.* Note that

$$\begin{aligned} & |T_{\pi_0} T_{\pi_1} \dots T_{\pi_{n-1}} 0(x) - T_{\pi_0} T_{\pi_1} \dots T_{\pi_{n-1}} v(x)| \leq Q_{\pi_0}^\gamma Q_{\pi_1}^\gamma \dots Q_{\pi_{n-1}}^\gamma |v|(x) \\ & \leq Q_{\pi_0}^\gamma Q_{\pi_1}^\gamma \dots Q_{\pi_{n-1}}^\gamma (w\|v\|_w)(x) \leq \tilde{\gamma}^{(n)}(\|v\|_w)w(x), \end{aligned}$$

where the operator  $Q_{\pi_n}^\gamma$  was defined in (4), the first inequality is by Condition 1(b, i, ii), and the last inequality is by applying Condition 1(b, iv, v); recall that  $\tilde{\gamma}$  was defined by (2). It remains to recall that  $\lim_{n \rightarrow \infty} \tilde{\gamma}^{(n)}(\|v\|_w) = 0$  by (1).  $\square$

The concerned optimal control problem can be now stated as

$$\text{Maximize over all } \pi: U^\pi(x). \quad (5)$$

The value function  $U$  is defined by  $U(x) := \sup_\pi U^\pi(x)$  for all  $x \in X$ .

**Definition 2.** We call a policy  $\pi$  uniformly optimal if  $U^\pi(x) = U(x)$  for all  $x \in X$ , and uniformly optimal on a subset  $E \subseteq X$  if  $U^\pi(x) = U(x)$  for all  $x \in E$ . If  $E = \{x\}$  is a singleton, we call the policy optimal at  $x$ . For a given  $\epsilon > 0$ , we call a policy uniformly  $\epsilon$ -optimal on a subset  $E \subseteq X$  if  $U^\pi(x) + \epsilon \geq U(x)$  for all  $x \in E$ . If  $E = X$ , then it is called uniformly  $\epsilon$ -optimal.

The objective here is to provide an implementable scheme to obtain a uniformly  $\epsilon$ -optimal policy on a given compact subset of the state space. To this end, we impose further conditions on the model.

**Condition 2.(a)** The multifunction  $\mathbf{A}$  is compact-valued, i.e.,  $\mathbf{A}(x)$  is a compact subset of  $A$  for each  $x \in X$ .

(b) For some constant  $L_{\mathbf{A}} \in [0, \infty)$ , the multifunction  $\mathbf{A}$  satisfies

$$d_H(\mathbf{A}(x), \mathbf{A}(y)) \leq L_{\mathbf{A}} d_X(x, y) \quad \forall x, y \in X,$$

where  $d_H$  is the Hausdorff metric on the space of nonempty compact subsets of  $A$ , so that

$$d_H(\mathbf{A}(x), \mathbf{A}(y)) := \sup_{a \in A(x)} \inf_{b \in A(y)} d_A(a, b) \vee \sup_{b \in A(y)} \inf_{a \in A(x)} d_A(a, b).$$

(c) The function  $w$  from Condition 1 is continuous on  $X$ , and the function  $u$  is Lipschitz continuous on  $D$ , i.e., for some constant  $L_u \in [0, \infty)$ ,

$$|u(x, a) - u(y, b)| \leq L_u(d_X(x, y) + d_A(a, b)) \quad \forall x, y \in X, a \in \mathbf{A}(x), b \in \mathbf{A}(y).$$

(d)  $\int_X v(y)q(dy|x, a)$  is continuous in  $(x, a) \in D$  for each bounded continuous function  $v$  on  $X$ , and for  $v = w$ .

Condition 2(a, b) implies that the multifunction  $\mathbf{A}$  is upper semicontinuous (in fact, continuous), according to Lemma 2.6 of [7]. Therefore, Condition 2 is stronger than Condition (W) in [3], which together with Condition 1, in turn implies the following result.

**Proposition 4.** Suppose Conditions 1 and 2 are satisfied. Then the following assertions hold.

- (a) There is a stationary uniformly optimal policy for problem (5).
- (b)  $|U(x)| \leq w(x)\tilde{\gamma}_{\infty}(\|u\|_w)$  for all  $x \in X$ ,  $U$  is upper semicontinuous on  $X$ , and is the unique solution to  $TU = U$  out of the set of upper semicontinuous functions in  $\mathbb{B}_w(X)$ , where  $T$  is defined for each  $v \in \mathbb{B}_w(X)$  by

$$Tv(x) := \sup_{a \in \mathbf{A}(x)} \left\{ u(x, a) + \int_X \delta(v(y))q(dy|x, a) \right\} \quad \forall x \in X.$$

Moreover,  $U = \lim_{n \rightarrow \infty} T^{(n)}v$  for any upper semicontinuous  $v \in \mathbb{B}_w(X)$ , where the convergence is in  $\mathbb{B}_w(X)$ .

- (c) Define the functions  $\{U_n\}_{n=0}^N$  on  $X$  by

$$U_0 \equiv 0; \quad U_n(x) := \sup_{a \in \mathbf{A}(x)} \left\{ u(x, a) + \int_X \delta(U_{n-1}(y))q(dy|x, a) \right\}$$

$\forall x \in X, 1 \leq n \leq N.$

Then for each  $0 \leq n \leq N$ ,  $U_n = \sup_{\pi} U_n^{\pi}$ ,  $U_n \in \mathbb{B}_w(X)$  and is upper semicontinuous on  $X$ , and  $\|U_n\|_w \leq \tilde{\gamma}_n(\|u\|_w) \leq \tilde{\gamma}_{\infty}(\|u\|_w)$ ,  $\|U_n - U\|_w \leq \tilde{\gamma}^{(n)}(\tilde{\gamma}_{\infty}(\|u\|_w))$ .

*Proof.* For parts (a,b), see Theorem 5.1 of [3]. For part (c),  $U_n = \sup_{\pi} U_n^{\pi}$  is by a standard dynamic programming argument. The rest was established in the proof of Theorem 5.1 of [3].  $\square$

The operator  $T$  will be referred to frequently below. Under the conditions of Proposition 4, by an extension of the Berge theorem, see [8, 9], it maps any upper semicontinuous function  $v \in \mathbb{B}_w(X)$  to an upper semicontinuous function in  $\mathbb{B}_w(X)$ . We impose an additional condition, under which it will be verified below that  $T$  is an operator from the space of Lipschitz continuous function  $v \in \mathbb{B}_w(X)$  to itself, and  $U$  is a Lipschitz continuous function.

**Condition 3.** *There is some constant  $L_q \in [0, \infty)$  such that the following are satisfied.*

(a) *For each Lipschitz continuous function  $v \in \mathbb{B}_w(X)$  with a Lipschitz constant  $L_v$ ,*

$$\begin{aligned} & \left| \int_X \delta(v(z))q(dz|x, a) - \int_X \delta(v(z))q(dz|y, b) \right| \\ & \leq \gamma(L_q L_v)(d_X(x, y) + d_A(a, b)) \quad \forall x, y \in X, a \in \mathbf{A}(x), b \in \mathbf{A}(y). \end{aligned}$$

(b)  $\gamma(L_q y)(1 + L_{\mathbf{A}}) < y$  for all  $y > 0$ .

For brevity, we put

$$\varphi(y) := \gamma(L_q y)(1 + L_{\mathbf{A}}), \quad y \geq 0, \quad (6)$$

so that  $\varphi(0) = 0$ . Under Conditions 1 and 3, Proposition 1 applies to  $\varphi$  in lieu with  $\psi$ , so that for each  $y \geq 0$ ,  $\underline{\varphi}_{\infty}(y)$  is defined, and is finite.

Observe that when  $\delta(x) = \beta x = \gamma(x)$  for all  $x \in X$  and some  $\beta \in [0, 1)$ , Condition 3(a) is the same as the next condition.

**Condition 4.** *There is some constant  $L_q \in [0, \infty)$  such that for each Lipschitz continuous function  $v \in \mathbb{B}_w(X)$  with a Lipschitz constant  $L_v$ ,*

$$\begin{aligned} & \left| \int_X v(z)q(dz|x, a) - \int_X v(z)q(dz|y, b) \right| \leq L_q L_v(d_X(d, y) + d_A(a, b)) \\ & \quad \forall x, y \in X, a \in \mathbf{A}(x), b \in \mathbf{A}(y). \end{aligned}$$

### 3 Main Statement

In what follows, let  $K_0$  be a fixed compact subset of  $X$ , and we present schemes for obtaining stationary and Markov policies that are uniformly  $\epsilon$ -optimal on the arbitrarily fixed set  $K_0$ . The schemes are similar to those in [6, 7] for linearly discounted model and finite horizon model. They are based on solving a sequence of models in finite state and action spaces, and are implementable in the sense of Remark 1. In particular, the expression of the Markov policy that is uniformly  $\epsilon$ -optimal on  $K_0$  can be explicitly obtained.

Let  $\zeta, \zeta_X, \zeta_A \in (0, \infty)$  be fixed. Then according to the proof of Lemma 2.9 of [7], there is a sequence of compact subsets  $\{K_n\}_{n \geq 1}$  of  $X$  satisfying

$$\sup_{x \in K_n, a \in \mathbf{A}(x)} \int_{X \setminus K_{n+1}} w(y)q(dy|x, a) < \zeta \quad \forall n \geq 0. \quad (7)$$

For each  $n \geq 0$ , since  $K_n$  is compact, it has a finite  $\zeta_X$ -net

$$X_n := \{z_1, \dots, z_{k_n}\}$$

of  $K_n$ , and an associated measurable partition  $\{K_n^i\}_{i=1}^{k_n}$  of  $K_n$  such that  $z_i \in K_n^i$ , and for each  $z \in K_n^i$ ,  $d_X(x, z_i) < \zeta_X$ . Let  $p_{X_n}^{K_n}(x) = z_i$  for each  $x \in K_n^i$ . Similarly, for each  $x \in X$ , since  $\mathbf{A}(x)$  is compact, it has a finite  $\zeta_A$ -net

$$\mathbf{B}(x) := \{b_1, \dots, b_{k(x)}\}.$$

Let  $N \geq 1$  be fixed. Define recursively the following functions:

$$\hat{U}_{N-1,N}(x) := 0 \quad \forall x \in X_N,$$

$$\hat{U}_{N-1,n}(x) := \max_{a \in \mathbf{B}(x)} \left\{ u(x, a) + \sum_{y \in X_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q((p_{X_{n+1}}^{K_{n+1}})^{-1}(y) | x, a) \right\}$$

$$\forall x \in X_n, \quad 0 \leq n \leq N - 1.$$

For each  $x \in X_n$ ,  $0 \leq n \leq N - 1$ , there is some  $c_{N,n}(x) \in \mathbf{B}(x)$  such that

$$\hat{U}_{N-1,n}(x) := u(x, c_{N,n}(x)) + \sum_{y \in X_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q((p_{X_{n+1}}^{K_{n+1}})^{-1}(y) | x, c_{N,n}(x)).$$

For each  $N \geq 0$ , we define a Markov policy  $g^N = \{f_n^N\}_{n \geq 0}$  by

$$\begin{aligned} f_n^N(x) &:= \operatorname{argmin}_{a \in \mathbf{A}(x)} \{d_A(a, c_{N,n}(p_{X_n}^{K_n}(x)))\} \quad \forall x \in K_n, \\ f_n^N(x) &:= f^\infty(x) \quad \forall x \in X \setminus K_n \end{aligned}$$

for all  $0 \leq n \leq N - 1$ , and  $f_n^N(x) := f^\infty(x)$  for all  $x \in X$  and  $n \geq N$ , where  $f^\infty$  is an arbitrarily fixed stationary policy. The above definition is legitimate because  $d_A(a, c_{N,n}(p_{X_n}^{K_n}(x)))$  is continuous in  $a \in \mathbf{A}(x)$  and measurable in  $x \in K_n$  and thus jointly measurable in  $(x, a)$  by [1, Lem.4.51] or [14, Prop.B.1.38], and  $\mathbf{A}$  is compact-valued and upper semicontinuous by [7, Lem.2.6]. In particular,  $f_n^N(x) := c_{N,n}(x)$  for all  $x \in X_n$  and  $n \leq N - 1$ . This Markov policy  $g^N$  will be shown to be a required uniformly  $\epsilon$ -optimal policy on the given compact set  $K_0$  when  $\zeta, \zeta_X, \zeta_A$  and  $N$  are suitably chosen.

Finally, we define a stationary policy  $f^N$  that will be shown to be a required uniformly  $\epsilon$ -optimal policy on the given compact set  $K_0$  when  $\zeta, \zeta_X, \zeta_A$  and  $N$  are suitably chosen. Let  $C_0 := K_0$ ,

$$C_n := \bigcap_{i=0}^{n-1} (X \setminus C_i) \cap K_n \quad n \geq 1,$$

and  $C_\infty := X \setminus (\bigcup_{n \geq 0} C_n)$ . Then  $\{C_n\}_{n=0,1,\dots,\infty}$  is a (disjoint) partition of  $X$  satisfying  $\bigcup_{n \geq 0} C_n = \bigcup_{n \geq 0} K_i$ . For the fixed  $N \geq 1$ , define a stationary policy  $f^N$  as follows:

$$f^N(x) := f_n^{N+n}(x) \quad \forall x \in C_n, \quad f^N(x) := f^\infty(x) \quad \forall x \in C_\infty.$$

**Theorem 1.** Suppose Conditions 1, 2 and 3 are satisfied. Let  $\zeta, \zeta_X, \zeta_A \in (0, \infty)$  and an integer  $N \geq 1$  be fixed. Let  $K_0$  be any compact subset of  $X$ . Then, for the stationary policy  $f^N$  defined above, the following holds:

$$\begin{aligned} \sup_{x \in K_0} |U^{f^N}(x) - U(x)| &\leq \underline{\gamma}_\infty(L_U \zeta_X + 3\zeta\gamma(\tilde{\underline{\gamma}}_\infty(\|u\|_w)) + 2\underline{\gamma}_\infty(\tilde{\Lambda})) \\ &\quad + \tilde{\underline{\gamma}}_\infty(2\tilde{\gamma}^{(N)}(\tilde{\underline{\gamma}}_\infty(\|u\|_w))) \sup_{x \in K_0} w(x), \end{aligned}$$

where

$$L_U := \underline{\varphi}_\infty(\Lambda), \quad \tilde{\Lambda} = L_U(\zeta_A + \zeta_X) + \gamma(\tilde{\underline{\gamma}}_\infty(\|u\|_w))\zeta$$

with  $\varphi$  being defined by (6), and

$$\Lambda := L_u(1 + L_{\mathbf{A}}) \geq 0. \quad (8)$$

The proofs of this theorem and Theorem 2 below are postponed to Sect. 5.

*Remark 1.(a)* As  $\zeta_A, \zeta_X, \zeta \rightarrow 0$ ,  $\tilde{\Lambda} \rightarrow 0$ , and by Proposition 1,  $\lim_{\tilde{\Lambda} \rightarrow 0} \underline{\gamma}_\infty(\tilde{\Lambda}) = 0$ . It follows that, for any given  $\epsilon > 0$ , one may take small enough constants  $\zeta, \zeta_X, \zeta_A \in (0, \infty)$  and a large enough integer  $N \geq 1$  such that the right-hand side of the inequality in Theorem 1 is majorized by  $\epsilon$ . The corresponding stationary policy  $f^N$  is uniformly  $\epsilon$ -optimal on the given compact set  $K_0$ . Given the current state  $x \in X$ , there is a unique  $n \in \{0, 1, \dots, \infty\}$  such that  $x \in C_n$ , and according to that  $n$ , one can compute  $f^N(x) = f_n^{N+n}(x)$  as the action that should be chosen.

(b) The proof of the previous statement reveals that, for each  $x \in \bigcup_{n \geq 0} K_n$ ,

$$\begin{aligned} |U^{f^N}(x) - U(x)| &\leq \underline{\gamma}_\infty(L_U \zeta_X + 3\zeta\gamma(\tilde{\underline{\gamma}}_\infty(\|u\|_w)) + 2\underline{\gamma}_\infty(\tilde{\Lambda})) \\ &\quad + \tilde{\underline{\gamma}}_\infty(2\tilde{\gamma}^{(N)}(\tilde{\underline{\gamma}}_\infty(\|u\|_w)))w(x). \end{aligned}$$

The next statement asserts that the Markov policy  $g^N$  is a required uniformly  $\epsilon$ -optimal policy on the given compact set  $K_0$  when  $\zeta, \zeta_X, \zeta_A$  and  $N$  are suitably chosen.

**Theorem 2.** Suppose Conditions 1, 2 and 4 are satisfied. Let  $\zeta, \zeta_X, \zeta_A \in (0, \infty)$  and an integer  $N \geq 1$  be fixed. Let  $K_0$  be any compact subset of  $X$ . For the Markov policy  $g^N$  defined above, the following holds: Then

$$\sup_{x \in K_0} |U^{g^N}(x) - U(x)| \leq 2\tilde{\gamma}^{(N)}(\tilde{\underline{\gamma}}_\infty(\|u\|_w)) \sup_{x \in K_0} w(x) + \underline{\gamma}_N(G)$$

with  $G := 2\underline{\gamma}_N(\tilde{\Lambda}_N) + \underline{\varphi}'_N(\Lambda)\zeta_X + 3\zeta\tilde{\underline{\gamma}}_N(\|u\|_w)$ , where

$$\varphi'(y) := (1 + L_{\mathbf{A}})L_q y \quad \forall y \geq 0, \quad \tilde{\Lambda}_N := \underline{\varphi}'_N(\Lambda)(\zeta_A + \zeta_X) + \gamma(\tilde{\underline{\gamma}}_N(\|u\|_w))\zeta.$$

Obviously, a similar remark to Remark 1(a) can be formulated.

## 4 Example

We take a stochastic optimal growth model from [3] (see Example 7.1 therein) as an example, to which the approximation schemes in this paper can be applied.

*Example 1.* The state  $x \in X = [0, \infty)$  represents the wealth. At each stage, one has to decide the amount  $a \in \mathbf{A}(x) = [0, x]$  to be consumed. Let  $A = [0, \infty)$ . The unconsumed wealth will be invested. If  $y$  is invested in this stage, then the wealth in the next stage is  $yS$ , where  $S$ , representing the random shock, is a  $[0, \infty)$ -valued random variable, whose distribution is  $\nu$ . We assume that the random shocks are all independent and identically distributed and with a finite mean

$$\bar{s} := \int_{[0, \infty)} s\nu(ds) < \infty.$$

Therefore, we may take

$$q(dy|x, a) = \int_{[0, \infty)} \delta_{(x-a)s}(dy)\nu(ds).$$

**Proposition 5.(a)** Consider  $u(x, a) = \sqrt{1+a}$  for all  $x \in X$  and  $a \in \mathbf{A}(x)$ , and

$$\delta(x) = ((1-\varepsilon)x + \varepsilon \ln(1+\varepsilon))I\{x \geq 0\}$$

with  $\varepsilon \in (0, 1)$  being a constant. Then Conditions 1, 2 and 4 are satisfied with  $\gamma = \delta$  on  $[0, \infty)$ ,  $w(x) = \sqrt{1+x}$ ,  $\alpha = 1$ ,  $L_{\mathbf{A}} = L_u = 1$ ,  $L_q = \bar{s}$ .

(b) Consider  $u(x, a) = \sqrt{1+a} - 2$  for all  $x \in X$  and  $a \in \mathbf{A}(x)$ , and

$$\delta(x) = \begin{cases} \beta_1 x & x \leq 0 \\ \beta_2 x & x \geq 0 \end{cases}$$

for some constant  $\beta_1, \beta_2 \in (0, 1)$ . Assume  $2\beta\bar{s} < 1$ . Then Conditions 1, 2 and 3 are satisfied with  $\gamma(x) = \beta x$ ,  $\beta = \max\{\beta_1, \beta_2\}$ ,  $w(x) = \sqrt{1+x}$ ,  $\alpha = 1$ ,  $L_{\mathbf{A}} = L_u = 1$ ,  $L_q = \bar{s}$ .

*Proof.* Condition 1(a) and (b, i), as well as Condition 2(a, b) are evidently satisfied, whereas Condition 1(b, ii–v) and Condition 2(d) were verified by the given function  $w$  and constant  $\alpha$  in Example 7.1 of [3]. For example, Condition 1(v) holds according to the calculation

$$\int_X w(y)q(dy|x, a) = \int_{[0, \infty)} \sqrt{(x-a)s+1}\nu(ds) \leq \sqrt{x+1}.$$

Condition 2(c) holds because the derivative of  $\sqrt{1+a}$  with respect to  $a$  is bounded by 1, and  $|\sqrt{1+a} - 1| \leq a - 0$  for all  $a \in [0, \infty)$ . Finally, regarding Condition 4, for a Lipschitz continuous function  $v \in \mathbb{B}_w(X)$  with a Lipschitz

constant  $L_v$ , we note that

$$\begin{aligned} & \left| \int_X v(z)q(dz|x, a) - \int_X v(z)q(dz|y, b) \right| \\ &= \left| \int_X v((x-a)s)\nu(ds) - \int_X v((y-b)s)\nu(ds) \right| \\ &\leq \int_X L_v(|x-y| + |a-b|)s\nu(ds) = L_v\bar{s}(|x-y| + |a-b|) \end{aligned}$$

so that we may take  $L_q = \bar{s}$ .

(b) Conditions 1 and 2 can be seen to be satisfied as in part (a). Regarding Condition 3(a), for a Lipschitz continuous function  $v \in \mathbb{B}_w(X)$  with a Lipschitz constant  $L_v$ ,

$$\begin{aligned} & \left| \int_X \delta(v(z))q(dz|x, a) - \int_X \delta(v(z))q(dz|y, b) \right| \\ &\leq \int_{[0, \infty)} \gamma(L_v(|x-y| + |a-b|)s)\nu(ds) \\ &= \beta L_v\bar{s}(|x-y| + |a-b|) = \gamma(\bar{s}L_v)(|x-y| + |a-b|) \end{aligned}$$

and so we may take  $L_q = \bar{s}$ . Condition 3(b) holds because  $2\beta\bar{s} < 1$ .  $\square$

## 5 Proof of Main Statements

In this section, we provide the detailed proof of Theorem 1. The proof of Theorem 2 is similar to the proof of Theorem 1, and will be sketched.

### 5.1 Proof of Theorem 1

Throughout this subsection, we suppose that Conditions 1, 2 and 3 are satisfied, without explicit indications.

**Lemma 2.** *Let  $v \in \mathbb{B}_w(X)$  be Lipschitz continuous with a Lipschitz constant  $L_v$ . Then  $Tv \in \mathbb{B}_w(X)$  is also Lipschitz continuous with a Lipschitz constant*

$$L_{Tv} = (L_u + \gamma(L_qL_v))(1 + L_{\mathbf{A}}).$$

*Proof.* In view of the remarks below Proposition 4, we only need to check the claimed Lipschitz continuity of  $Tv$  as follows. Let  $x, z \in X$  and some Lipschitz continuous  $v \in \mathbb{B}_w(X)$  with a Lipschitz constant  $L_v$  be fixed. Then

$$\begin{aligned}
& |Tv(x) - Tv(z)| \\
& \leq \max \left\{ \sup_{a \in \mathbf{A}(x)} \inf_{b \in \mathbf{A}(z)} \{|u(x, a) - u(y, b)| \right. \\
& \quad + \left| \int_X \delta(v(y)) q(dy|x, a) - \int_X \delta(v(y)) q(dy|z, b) \right| \left. \right\}, \\
& \quad \sup_{b \in \mathbf{A}(z)} \inf_{a \in \mathbf{A}(x)} \{|u(x, a) - u(y, b)| \right. \\
& \quad + \left| \int_X \delta(v(y)) q(dy|x, a) - \int_X \delta(v(y)) q(dy|z, b) \right| \left. \right\}.
\end{aligned} \tag{9}$$

Indeed, in case  $|Tv(x) - Tv(z)| = Tv(x) - Tv(z)$ , for any fixed  $\epsilon > 0$ , there is some  $a^* \in \mathbf{A}(x)$  such that  $Tv(x) \leq u(x, a^*) + \int_X \delta(v(y)) q(dy|x, a^*) + \epsilon$  and thus

$$\begin{aligned}
& |Tv(x) - Tv(z)| \leq u(x, a^*) + \int_X \delta(v(y)) q(dy|x, a^*) + \epsilon \\
& \quad + \inf_{b \in \mathbf{A}(z)} \left\{ -u(z, b) - \int_X \delta(v(y)) q(dy|z, b) \right\} \\
& \leq \sup_{a \in \mathbf{A}(x)} \inf_{b \in \mathbf{A}(z)} \{|u(x, a) - u(z, b)| \right. \\
& \quad + \left| \int_X \delta(v(y)) q(dy|x, a) - \int_X \delta(v(y)) q(dy|z, b) \right| \left. \right\} + \epsilon.
\end{aligned}$$

Since  $\epsilon > 0$  was arbitrarily fixed,

$$\begin{aligned}
& |Tv(x) - Tv(z)| \leq \sup_{a \in \mathbf{A}(x)} \inf_{b \in \mathbf{A}(z)} \{|u(x, a) - u(z, b)| \right. \\
& \quad + \left| \int_X \delta(v(y)) q(dy|x, a) - \int_X \delta(v(y)) q(dy|z, b) \right| \left. \right\}.
\end{aligned}$$

In case  $|Tv(x) - Tv(z)| = Tv(z) - Tv(x)$ , we analogously see

$$\begin{aligned}
& |Tv(x) - Tv(z)| \leq \sup_{b \in \mathbf{A}(z)} \inf_{a \in \mathbf{A}(x)} \{|u(x, a) - u(z, b)| \right. \\
& \quad + \left| \int_X \delta(v(y)) q(dy|x, a) - \int_X \delta(v(y)) q(dy|z, b) \right| \left. \right\},
\end{aligned}$$

and hence (9) holds. By Conditions 2 and 3

$$\begin{aligned}
& |u(x, a) - u(z, b)| + \left| \int_X \delta(v(y)) q(dy|x, a) - \int_X \delta(v(y)) q(dy|z, b) \right| \\
& \leq L_u(d_X(x, z) + d_A(a, b)) + \gamma(L_q L_v)(d_X(x, z) + d_A(a, b)) \\
& = (L_u + \gamma(L_q L_v))(d_X(x, z) + d_A(a, b))
\end{aligned}$$

and so by (9),

$$\begin{aligned}
& |Tv(x) - Tv(z)| \\
& \leq (L_u + \gamma(L_q L_v))(d_X(x, z) + \sup_{a \in \mathbf{A}(x)} \inf_{b \in \mathbf{A}(z)} d_A(a, b) \vee \sup_{b \in \mathbf{A}(z)} \inf_{a \in \mathbf{A}(x)} d_A(a, b)) \\
& = (L_u + \gamma(L_q L_v))(d_X(x, z) + d_H(\mathbf{A}(x), \mathbf{A}(z))) \\
& \leq (L_u + \gamma(L_q L_v))(1 + L_{\mathbf{A}})d_X(x, z),
\end{aligned}$$

where the last inequality is by Condition 2(b).  $\square$

As a consequence of the previous lemma, we deduce the Lipschitz continuity of the value function  $U$ .

**Lemma 3.** *Let  $v \in \mathbb{B}_w(X)$  be a Lipschitz continuous function with a Lipschitz constant  $L_v$ . Then the following assertions hold.*

- (a) *For each  $n \geq 1$ ,  $T^n v \in \mathbb{B}_w(X)$  is with a Lipschitz constant  $\underline{\varphi}_n(\Lambda) + \varphi^{(n)}(L_v)$ , where  $\underline{\varphi}_n$  is defined by (3) with  $\varphi$  in lieu of  $\psi$ . In particular,  $U_n = T^n 0$  is Lipschitz continuous with a Lipschitz constant  $\underline{\varphi}_n(\Lambda) \leq \underline{\varphi}_{\infty}(\Lambda)$ .*
- (b) *The value function  $U$  is Lipschitz continuous with a Lipschitz constant  $L_U = \underline{\varphi}_{\infty}(\Lambda)$ .*

*Proof.* (a) By Lemma 2,  $T^n v \in \mathbb{B}_w(X)$  and is Lipschitz continuous for each  $n \geq 0$ , and we may take the following as a Lipschitz constant of  $Tv$ :

$$L_u(1 + L_{\mathbf{A}}) + \gamma(L_q L_v)(1 + L_{\mathbf{A}}) = \Lambda + \varphi(L_v),$$

and thus the claimed relation holds for  $n = 1$ . Assume it holds for  $n$ . Now, by Lemma 2 and the inductive supposition, we may take the following as a Lipschitz constant of  $T^{n+1}v = T(T^n v)$ :

$$\begin{aligned}
& \Lambda + \varphi(\underline{\varphi}_n(\Lambda) + \varphi^{(n)}(L_v)) \leq \Lambda + \varphi(\underline{\varphi}_n(\Lambda)) + \varphi^{(n+1)}(L_v) \\
& = \underline{\varphi}_{n+1}(\Lambda) + \varphi^{(n+1)}(L_v),
\end{aligned}$$

where the inequality is by the sub-additivity of  $\varphi$ . The statement follows from this and the induction.

(b) For each  $x, y \in X$ , by Proposition 4 and the assertion in (a) with  $v \equiv 0 = \varphi(0) = L_v$ ,

$$\begin{aligned}
|U(x) - U(y)| & \leq \lim_{n \rightarrow \infty} |T^n 0(x) - T^n 0(y)| \leq \lim_{n \rightarrow \infty} \underline{\varphi}_n(\Lambda) d_X(x, y) \\
& = \underline{\varphi}_{\infty}(\Lambda) d_X(x, y),
\end{aligned}$$

where the limit  $\underline{\varphi}_{\infty}(\Lambda)$  is finite and exists by applying Proposition 1 to  $\varphi$ , which is valid under Conditions 1 and 3. The statement follows now.  $\square$

For the forthcoming discussions and statements, for each fixed  $N \geq 1$  and  $0 \leq n \leq N$ , we extend the definition of  $\hat{U}_{N-1,n}$  from  $X_n$  to  $K_n$  by putting

$\hat{U}_{N-1,n}(x) := \hat{U}_{N-1,n}(p_{X_n}^{K_n}(x))$  for all  $x \in K_n \setminus X_n$ . Then for all  $x \in X_n$  and  $0 \leq n \leq N-1$ ,

$$\hat{U}_{N-1,n}(x) = \max_{a \in \mathbf{B}(x)} \left\{ u(x, a) + \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|x, a) \right\}. \quad (10)$$

**Lemma 4.** Let  $N \geq 1$  and  $0 \leq n \leq N$  be fixed. Then  $\sup_{x \in K_n} |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| \leq \underline{\gamma}_\infty(\tilde{\Lambda})$  with  $\tilde{\Lambda} = L_U(\zeta_A + \zeta_X) + \gamma(\underline{\gamma}_\infty(\|u\|_w))\zeta$ .

*Proof.* The case of  $n = N$  is trivial. Let  $0 \leq n \leq N-1$  be fixed, and consider firstly some  $x \in X_n$ . Then

$$\begin{aligned} |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| &= \left| \max_{a \in \mathbf{B}(x)} \left\{ u(x, a) + \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|x, a) \right\} \right. \\ &\quad \left. - \sup_{b \in \mathbf{A}(x)} \left\{ u(x, b) + \int_X \delta(U_{N-n-1}(y)) q(dy|x, b) \right\} \right|. \end{aligned}$$

The same argument as in the justification of (9) shows

$$\begin{aligned} &|\hat{U}_{N-1,n}(x) - U_{N-n}(x)| \\ &\leq \max \left\{ \sup_{b \in \mathbf{A}(x)} \inf_{a \in \mathbf{B}(x)} \left\{ |u(x, b) - u(x, a)| + \left| \int_X \delta(U_{N-n-1}(y)) q(dy|x, b) \right. \right. \right. \\ &\quad \left. \left. \left. - \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|x, a) \right| \right\}, \right. \\ &\quad \left. \sup_{a \in \mathbf{B}(x)} \inf_{b \in \mathbf{A}(x)} \left\{ |u(x, b) - u(x, a)| + \left| \int_X \delta(U_{N-n-1}(y)) q(dy|x, b) \right. \right. \right. \\ &\quad \left. \left. \left. - \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|x, a) \right| \right\} \right\}. \end{aligned}$$

Recall from Condition 2(c) that  $|u(x, a) - u(x, b)| \leq L_u d_A(a, b)$  for each  $a \in \mathbf{B}(x)$  and  $b \in \mathbf{A}(x)$ . Also, for each  $a \in \mathbf{B}(x)$  and  $b \in \mathbf{A}(x)$ ,

$$\begin{aligned} &\left| \int_X \delta(U_{N-n-1}(y)) q(dy|x, b) - \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|x, a) \right| \\ &\leq \left| \int_X \delta(U_{N-n-1}(y)) q(dy|x, b) - \int_X \delta(U_{N-n-1}(y)) q(dy|x, a) \right| \\ &\quad + \left| \int_X \delta(U_{N-n-1}(y)) q(dy|x, a) - \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|x, a) \right|. \quad (11) \end{aligned}$$

For the first summand, since  $U_{N-n-1} \in \mathbb{B}_w(X)$  and is Lipschitz continuous with a Lipschitz constant  $\underline{\varphi}_{N-n-1}(\Lambda)$  by Lemma 3 and Proposition 4, applying

Condition 3 to it gives

$$\begin{aligned} & \left| \int_X \delta(U_{N-n-1}(y))q(dy|x, b) - \int_X \delta(U_{N-n-1}(y))q(dy|x, a) \right| \\ & \leq \gamma(L_q \underline{\varphi}_{N-n-1}(\Lambda))d_A(a, b). \end{aligned}$$

For the second summand in (11),

$$\begin{aligned} & \left| \int_X \delta(U_{N-n-1}(y))q(dy|x, a) - \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y))q(dy|x, a) \right| \\ & \leq \int_{K_{n+1}} |\delta(U_{N-n-1}(y)) - \delta(\hat{U}_{N-1,n+1}(y))|q(dy|x, a) \\ & \quad + \int_{X \setminus K_{n+1}} |\delta(U_{N-n-1}(y))|q(dy|x, a) \\ & \leq \sup_{y \in K_{n+1}} |\delta(U_{N-n-1}(y)) - \delta(\hat{U}_{N-1,n+1}(y))| \\ & \quad + \gamma(\tilde{\gamma}_{N-n-1}(\|u\|_w)) \int_{X \setminus K_{n+1}} w(y)q(dy|x, a) \\ & \leq \sup_{y \in K_{n+1}} \gamma(|U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) + \gamma(\tilde{\gamma}_{N-n-1}(\|u\|_w))\zeta, \end{aligned}$$

where the second inequality holds because

$$\begin{aligned} & |\delta(U_{N-n-1})| \leq \gamma(|U_{N-n-1}|) \leq w\gamma(\|U_{N-n-1}\|_w) \\ & \leq w\gamma(\tilde{\gamma}_{N-n-1}(\|u\|_w)) \in \mathbb{B}_w(X) \end{aligned}$$

by Proposition 4 and Condition 1, and the last inequality holds by Condition 1 and (7).

Now

$$\begin{aligned} & |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| \\ & \leq \max\{ \sup_{b \in \mathbf{B}(x)} \inf_{a \in \mathbf{B}(x)} \{L_u d_A(a, b) + \gamma(L_q \underline{\varphi}_{N-n-1}(\Lambda))d_A(a, b) \\ & \quad + \sup_{y \in K_{n+1}} \gamma(|U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) + \gamma(\tilde{\gamma}_{N-n-1}(\|u\|_w))\zeta\}, \\ & \quad \sup_{a \in \mathbf{B}(x)} \inf_{b \in \mathbf{B}(x)} \{L_u d_A(a, b) + \gamma(L_q \underline{\varphi}_{N-n-1}(\Lambda))d_A(a, b) \\ & \quad + \sup_{y \in K_{n+1}} \gamma(|U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) + \gamma(\tilde{\gamma}_{N-n-1}(\|u\|_w))\zeta\}\}, \\ & = (L_u + \gamma(L_q \underline{\varphi}_{N-n-1}(\Lambda)))\zeta_A + \sup_{y \in K_{n+1}} \gamma(|U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \\ & \quad + \gamma(\tilde{\gamma}_{N-n-1}(\|u\|_w))\zeta \\ & \leq (L_u + \gamma(L_q \underline{\varphi}_\infty(\Lambda)))(1 + L_\mathbf{A})\zeta_A + \sup_{y \in K_{n+1}} \gamma(|U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \\ & \quad + \gamma(\tilde{\gamma}_\infty(\|u\|_w))\zeta. \end{aligned}$$

Having recognized  $(L_u + \gamma(L_q \varphi_\infty(\Lambda)))(1 + L_{\mathbf{A}}) = \Lambda + \varphi(\varphi_\infty(\Lambda)) = \varphi_\infty(\Lambda) = L_U$  (recall (8) and (6), Proposition 1 and Lemma 3), we see now

$$\begin{aligned} & |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| \leq L_U \zeta_A + \gamma(\tilde{\gamma}_\infty(\|u\|_w)) \zeta \\ & + \sup_{y \in K_{n+1}} \gamma(|U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \\ & = L_U \zeta_A + \gamma(\tilde{\gamma}_\infty(\|u\|_w)) \zeta + \gamma(\sup_{y \in K_{n+1}} |U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \quad \forall x \in X_n, \end{aligned}$$

where the last equality holds because  $\gamma$  is increasing.

Next, we arbitrarily fix some  $x \in K_n$  and  $z = p_{X_n}^{K_n}(x) \in X_n$ . Then

$$\begin{aligned} & |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| = |\hat{U}_{N-1,n}(z) - U_{N-n}(x)| \\ & \leq |\hat{U}_{N-1,n}(z) - U_{N-n}(z)| + |U_{N-n}(z) - U_{N-n}(x)| \\ & \leq L_U \zeta_A + \gamma(\tilde{\gamma}_\infty(\|u\|_w)) \zeta + \gamma(\sup_{y \in K_{n+1}} |U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \\ & \quad + \varphi_\infty(\Lambda) d_X(x, z) \\ & \leq \varphi_\infty(\Lambda)(\zeta_A + \zeta_X) + \gamma(\tilde{\gamma}_\infty(\|u\|_w)) \zeta + \gamma(\sup_{y \in K_{n+1}} |U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \end{aligned}$$

where the second inequality is by (12) and Lemma 3. Hence,

$$\begin{aligned} & \sup_{x \in K_n} |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| \leq \varphi_\infty(\Lambda)(\zeta_A + \zeta_X) + \gamma(\tilde{\gamma}_\infty(\|u\|_w)) \zeta \\ & + \gamma(\sup_{y \in K_{n+1}} |U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \\ & = \tilde{\Lambda} + \gamma(\sup_{y \in K_{n+1}} |U_{N-n-1}(y) - \hat{U}_{N-1,n+1}(y)|) \\ & \leq \tilde{\Lambda} + \gamma(\tilde{\Lambda} + \gamma(\sup_{y \in K_{n+2}} |U_{N-n-2}(y) - \hat{U}_{N-1,n+2}(y)|)), \end{aligned}$$

and by iteration, we see from the sub-additivity of  $\gamma$  that

$$\begin{aligned} & \sup_{x \in K_n} |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| \\ & \leq \underline{\gamma}_{N-n}(\tilde{\Lambda}) + \gamma^{(N-n)}(\sup_{x \in K_N} |\hat{U}_{N-1,N}(x) - U_0(x)|) = \underline{\gamma}_{N-n}(\tilde{\Lambda}) \leq \underline{\gamma}_\infty(\tilde{\Lambda}) \end{aligned}$$

with the last equality following from  $\gamma(0) = 0$  and that  $\underline{\gamma}_n(\tilde{\Lambda})$  increases in  $n$ .  $\square$

**Corollary 1.** *For each  $N \geq 1$ ,*

$$\sup_{x \in K_0} |\hat{U}_{N-1,0}(x) - U(x)| \leq \underline{\gamma}_\infty(\tilde{\Lambda}) + w(x) \tilde{\gamma}^{(N)}(\tilde{\gamma}_\infty(\|u\|_w)),$$

where  $\tilde{\Lambda} := \underline{\gamma}_\infty(\Lambda)(\zeta_A + \zeta_X) + \gamma(\tilde{\gamma}_\infty(\|u\|_w)) \zeta$ .

*Proof.* This follows from

$$|\hat{U}_{N-1,0}(x) - U(x)| \leq |\hat{U}_{N-1,0}(x) - U_N(x)| + |U_N(x) - U(x)|,$$

Lemma 4 and Proposition 4(c).  $\square$

**Lemma 5.** Let  $N \geq 1$ ,  $0 \leq n \leq N-1$  and  $x \in K_n$  be fixed. Then

$$\begin{aligned} & U(x) - L_U \zeta_X - 2\zeta \gamma(\tilde{\gamma}_\infty(\|u\|_w)) - 2\gamma_\infty(\tilde{\Lambda}) - 2w(x)\tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(\|u\|_w)) \\ & \leq u(x, f_n^N(x)) + \int_{K_{n+1}} \delta(U(y))q(dy|x, f_n^N(x)). \end{aligned}$$

*Proof.* Let  $x \in K_n$  and  $z = p_{X_n}^{K_n}(x) \in X_n$  be fixed.

Recall from Proposition 4 that

$$\begin{aligned} & U(x) \leq U_{N-n}(x) + w(x)\tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(\|u\|_w)) \\ & \leq \hat{U}_{N-1,n}(x) + \gamma_\infty(\tilde{\Lambda}) + w(x)\tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(\|u\|_w)) \\ & = \hat{U}_{N-1,n}(z) + \gamma_\infty(\tilde{\Lambda}) + w(x)\tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(\|u\|_w)), \end{aligned}$$

where the inequality is by Lemma 4 and the last equality is by the definition of  $\hat{U}_{N-1,n}(x)$  for  $x \in K_n$ . For  $\hat{U}_{N-1,n}(z)$ , recall from (10) and the definition of  $f_n^N$ ,

$$\begin{aligned} & \hat{U}_{N-1,n}(z) = u(z, f_n^N(z)) + \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y))q(dy|z, f_n^N(z)) \\ & \leq u(z, f_n^N(z)) + \int_{K_{n+1}} \delta(U_{N-(n+1)}(y) + \gamma_\infty(\tilde{\Lambda}))q(dy|z, f_n^N(z)) \\ & \leq u(z, f_n^N(z)) + \int_{K_{n+1}} \delta(U_{N-(n+1)}(y))q(dy|z, f_n^N(z)) \\ & \quad + \int_{K_{n+1}} \gamma(\gamma_\infty(\tilde{\Lambda}))q(dy|z, f_n^N(z)) \\ & \leq u(z, f_n^N(z)) + \int_{K_{n+1}} \delta(U_{N-(n+1)}(y))q(dy|z, f_n^N(z)) + \gamma_\infty(\tilde{\Lambda}), \end{aligned}$$

where the first inequality is by Lemma 4, the second inequality is by the following consequence of Condition 1(b, ii):  $|\delta(x_1 + x_2) - \delta(x_1)| \leq \gamma(|x_2|)$  for all  $x_1, x_2 \in \mathbb{R}$ , and the last inequality is by Condition 1(b,i). Now

$$\begin{aligned} & U(x) \leq u(z, f_n^N(z)) + \int_{K_{n+1}} \delta(U_{N-(n+1)}(y))q(dy|z, f_n^N(z)) + 2\gamma_\infty(\tilde{\Lambda}) \\ & \quad + w(x)\tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(\|u\|_w)), \end{aligned}$$

and so

$$\begin{aligned}
& U(x) - 2\underline{\gamma}_\infty(\tilde{\Lambda}) - w(x)\tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(||u||_w)) - |u(x, f_n^N(x)) - u(z, f_n^N(z))| \\
& - \left| \int_X \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) - \int_X \delta(U_{N-n-1}(y))q(dy|z, f_n^N(z)) \right| \\
& \leq u(z, f_n^N(z)) + \int_{K_{n+1}} \delta(U_{N-(n+1)}(y))q(dy|z, f_n^N(z)) \\
& + u(x, f_n^N(x)) - u(z, f_n^N(z)) \\
& + \int_X \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) - \int_X \delta(U_{N-n-1}(y))q(dy|z, f_n^N(z)) \\
& = u(x, f_n^N(x)) - \int_{X \setminus K_{n+1}} \delta(U_{N-n-1}(y))q(dy|z, f_n^N(z)) \\
& + \int_X \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)). \tag{12}
\end{aligned}$$

Note that

$$\begin{aligned}
& |u(x, f_n^N(x)) - u(z, f_n^N(z))| \\
& + \left| \int_X \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) - \int_X \delta(U_{N-n-1}(y))q(dy|z, f_n^N(z)) \right| \\
& \leq L_u(d_X(x, z) + d_A(f_n^N(x), f_n^N(z))) \\
& + \gamma(L_q L_{U_{N-n-1}})(d_X(x, z) + d_A(f_n^N(x), f_n^N(z))) \\
& = (L_u + \gamma(L_q L_{U_{N-n-1}}))(d_X(x, z) + \inf_{a \in \mathbf{A}(x)} d_A(a, f_n^N(z))) \\
& \leq (L_u + \gamma(L_q L_{U_{N-n-1}}))(d_X(x, z) + d_H(\mathbf{A}(x), \mathbf{B}(z))) \\
& \leq (L_u + \gamma(L_q L_{U_{N-n-1}}))(1 + L_{\mathbf{A}})d_X(x, z) \\
& \leq (L_u + \gamma(L_q L_{U_{N-n-1}}))(1 + L_{\mathbf{A}})\zeta_X \leq (L_u(1 + L_{\mathbf{A}}) + (1 + L_{\mathbf{A}})\gamma(L_q L_U))\zeta_X \\
& = (\Lambda + \varphi(\underline{\varphi}_\infty(\Lambda)))\zeta_X,
\end{aligned}$$

where the first inequality is by Condition 2 applied to  $u$  and Condition 3 applied to  $U_{N-n-1}$ , which is Lipschitz and in  $\mathbb{B}_w(X)$  by Lemma 3, the first equality holds by the definition of  $f_n^N$ , the second inequality holds by the definition of the Hausdorff metric, the third inequality is by Condition 2 regarding the multifunction  $\mathbf{A}$ , the fourth inequality holds because of the definition of  $z$ , and the fifth inequality is by Lemma 3. That is, applying Proposition 1 to  $\varphi$ , we recognize

$$\begin{aligned}
& |u(x, f_n^N(x)) - u(z, f_n^N(z))| \\
& + \left| \int_X \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) - \int_X \delta(U_{N-n-1}(y))q(dy|z, f_n^N(z)) \right| \\
& \leq \underline{\varphi}_\infty(\Lambda)\zeta_X = L_U\zeta_X.
\end{aligned}$$

Consequently, from (12) we see

$$\begin{aligned}
& U(x) - 2\underline{\gamma}_\infty(\tilde{A}) - w(x)\tilde{\gamma}^{(N-n)}(\underline{\gamma}_\infty(||u||_w)) - L_U \zeta_X \\
& \leq u(x, f_n^N(x)) - \int_{X \setminus K_{n+1}} \delta(U_{N-n-1}(y))q(dy|x, f_n^N(z)) \\
& \quad + \int_{K_{n+1}} \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) \\
& \quad + \int_{X \setminus K_{n+1}} \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) \\
& \leq u(x, f_n^N(x)) + \gamma(\tilde{\gamma}_\infty(||u||_w)) \int_{X \setminus K_{n+1}} w(y)q(dy|z, f_n^N(z)) \\
& \quad + \int_{K_{n+1}} \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) \\
& \quad + \gamma(\tilde{\gamma}_\infty(||u||_w)) \int_{X \setminus K_{n+1}} w(y)q(dy|x, f_n^N(x)) \\
& \leq u(x, f_n^N(x)) + 2\gamma(\tilde{\gamma}_\infty(||u||_w))\zeta_X + \int_{K_{n+1}} \delta(U_{N-n-1}(y))q(dy|x, f_n^N(x)) \\
& \leq u(x, f_n^N(x)) + 2\gamma(\tilde{\gamma}_\infty(||u||_w))\zeta_X + \int_{K_{n+1}} \delta(U(y))q(dy|x, f_n^N(x)) \\
& \quad + \int_{K_{n+1}} |\delta(U_{N-n-1}(y)) - \delta(U(y))|q(dy|x, f_n^N(x)),
\end{aligned}$$

where the third inequality is by (7), and the second inequality follows from the calculation

$$\begin{aligned}
& \delta(U_{N-n-1}(y)) \leq \gamma(||U_{N-n-1}||_w w(y)) \leq w(y)\gamma(||U_{N-n-1}||_w) \\
& \leq w(y)\gamma(\tilde{\gamma}_\infty(||u||_w))
\end{aligned}$$

by Proposition 4. That is,

$$\begin{aligned}
& U(x) - 2\underline{\gamma}_\infty(\tilde{A}) - 2\gamma(\tilde{\gamma}_\infty(||u||_w))\zeta_X - w(x)\tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(||u||_w)) - L_U \zeta_X \\
& \leq u(x, f_n^N(x)) + \int_{K_{n+1}} \delta(U(y))q(dy|x, f_n^N(x)) \\
& \quad + \int_{K_{n+1}} |\delta(U_{N-n-1}(y)) - \delta(U(y))|q(dy|x, f_n^N(x)) \\
& \leq u(x, f_n^N(x)) + \int_{K_{n+1}} \delta(U(y))q(dy|x, f_n^N(x)) \\
& \quad + \int_{K_{n+1}} \gamma(|(U_{N-n-1}(y)) - U(y)|)q(dy|x, f_n^N(x))
\end{aligned}$$

$$\begin{aligned}
&\leq u(x, f_n^N(x)) + \int_{K_{n+1}} \delta(U(y)) q(dy|x, f_n^N(x)) \\
&\quad + \int_X \gamma(\tilde{\gamma}^{(N-n-1)}(\tilde{\gamma}_\infty(\|u\|_w)) w(y)) q(dy|x, f_n^N(x)) \\
&\leq u(x, f_n^N(x)) + \int_{K_{n+1}} \delta(U(y)) q(dy|x, f_n^N(x)) \\
&\quad + \gamma(\tilde{\gamma}^{(N-n-1)}(\tilde{\gamma}_\infty(\|u\|_w))) \int_X w(y) q(dy|x, f_n^N(x)) \\
&\leq u(x, f_n^N(x)) + \int_{K_{n+1}} \delta(U(y)) q(dy|x, f_n^N(x)) \\
&\quad + w(x) \tilde{\gamma}^{(N-n)}(\tilde{\gamma}_\infty(\|u\|_w)),
\end{aligned}$$

where the last two inequalities hold by Condition 1. Now the statement follows.  $\square$

In the next statement, let the function  $\bar{U}$  on  $X$  be defined by

$$\bar{U}(x) := U(x) \forall x \in \bigcup_{n \geq 0} K_n, \quad \bar{U}(x) := -w(x) \tilde{\gamma}_\infty(\|u\|_w) \forall x \in X \setminus \bigcup_{n \geq 0} K_n.$$

**Lemma 6.** *For each  $N \geq 1$ ,*

$$\begin{aligned}
&\bar{U}(x) - (L_U \zeta_X + 3\zeta \gamma(\tilde{\gamma}_\infty(\|u\|_w)) + 2\tilde{\gamma}_\infty(\tilde{\Lambda})) - 2w(x) \tilde{\gamma}^{(N)}(\tilde{\gamma}_\infty(\|u\|_w)) \\
&\leq u(x, f^N(x)) + \int_X \bar{U}(y) q(dy|x, f^N(x)) \quad \forall x \in X.
\end{aligned}$$

*Proof.* Note that  $|\bar{U}(x)| \leq w(x) \tilde{\gamma}_\infty(\|u\|_w)$  for all  $x \in X$ , according to Proposition 4, and consequently  $\bar{U} \in \mathbb{B}_w(X)$ .

For  $x \in X \setminus \bigcup_{n \geq 0} K_n = C_\infty$ , it holds that

$$\begin{aligned}
&u(x, f^N(x)) + \int_X \delta(\bar{U}(y)) q(dy|x, f^N(x)) \\
&\geq -\|u\|_w w(x) - \int_X \gamma(\tilde{\gamma}_\infty(\|u\|_w)) w(y) q(dy|x, f^N(x)) \\
&\geq -\|u\|_w w(x) - \tilde{\gamma}(\tilde{\gamma}_\infty(\|u\|_w)) w(x) = -w(x) \{ \|u\|_w + \tilde{\gamma}(\tilde{\gamma}_\infty(\|u\|_w)) \} \\
&= -w(x) \tilde{\gamma}_\infty(\|u\|_w) = \bar{U}(x),
\end{aligned}$$

where the second inequality is by Condition 1, and the last inequality is by Proposition 1. Therefore, the claimed relation in the lemma holds for  $x \in X \setminus \bigcup_{n \geq 0} K_n = C_\infty$ .

Now let  $x \in C_n$  be fixed for some  $n \in \{0, 1, \dots\}$ . Since  $C_n \subseteq K_n$ ,  $f^N(x) = f_n^{N+n}(x)$ , and we have from the definition of  $\bar{U}$  and Lemma 5 with  $N+n$  in lieu

of  $N$  therein that

$$\begin{aligned} \bar{U}(x) - L_U \zeta_X - 2\zeta \gamma(\tilde{\underline{\gamma}}_\infty(||u||_w)) - 2\underline{\gamma}_\infty(\tilde{A}) - 2w(x)\tilde{\gamma}^{(N)}(\tilde{\underline{\gamma}}_\infty(||u||_w)) \\ \leq u(x, f^N(x)) + \int_{K_{n+1}} \delta(\bar{U}(y))q(dy|x, f^N(x)) \\ = u(x, f^N(x)) + \int_{K_{n+1}} \delta(\bar{U}(y))q(dy|x, f^N(x)) - \int_X \delta(\bar{U}(y))q(dy|x, f^N(x)) \\ + \int_X \delta(\bar{U}(y))q(dy|x, f^N(x)), \end{aligned}$$

and so

$$\begin{aligned} \bar{U}(x) - L_U \zeta_X - 2\zeta \gamma(\tilde{\underline{\gamma}}_\infty(||u||_w)) - 2\underline{\gamma}_\infty(\tilde{A}) - 2w(x)\tilde{\gamma}^{(N)}(\tilde{\underline{\gamma}}_\infty(||u||_w)) \\ - \int_{X \setminus K_{n+1}} |\delta(\bar{U}(y))|q(dy|x, f^N(x)) \leq u(x, f^N(x)) + \int_X \delta(\bar{U}(y))q(dy|x, f^N(x)). \end{aligned}$$

Observe that on the left hand side of the above inequality,

$$\begin{aligned} & \int_{X \setminus K_{n+1}} |\delta(\bar{U}(y))|q(dy|x, f^N(x)) \\ & \leq \gamma(\tilde{\underline{\gamma}}_\infty(||u||_w)) \int_{X \setminus K_{n+1}} w(y)q(dy|x, f^N(x)) \leq \zeta \gamma(\tilde{\underline{\gamma}}_\infty(||u||_w)). \end{aligned}$$

Now the statement follows.  $\square$

**Lemma 7.** *If for some stationary policy  $f$  and constants  $R, Q \in [0, \infty)$ ,*

$$\bar{U}(x) \leq u(x, f(x)) + \int_X \delta(\bar{U}(y))q(dy|x, f(x)) + R + Qw(x) \quad \forall x \in X,$$

then

$$\bar{U}(x) \leq T_f^n \bar{U}(x) + \underline{\gamma}_n(R) + \tilde{\gamma}_n(Q)w(x) \quad \forall x \in X$$

for all  $n \geq 1$ .

*Proof.* Let  $x \in X$  be fixed, and we prove the statement by induction, as follows. When  $n = 1$ , the claimed relation holds because  $\underline{\gamma}_1(R) = R$  and  $\tilde{\gamma}_1(Q) = Q$ . Assume the claimed relation holds for  $n$ . Then

$$\begin{aligned} \bar{U}(x) & \leq T_f \bar{U}(x) + R + Qw(x) \\ & \leq T_f(T_f^n \bar{U} + \underline{\gamma}_n(R) + \tilde{\gamma}_n(Q)w)(x) + R + Qw(x) \\ & = u(x, f(x)) + \int_X \delta(T_f^n \bar{U}(y) + \underline{\gamma}_n(R) + \tilde{\gamma}_n(Q)w(y))q(dy|x, f(x)) + R + Qw(x) \end{aligned}$$

$$\begin{aligned}
&\leq u(x, f(x)) + \int_X (\delta(T_f^n \bar{U}(y)) + \gamma(\underline{\gamma}_n(R) + \tilde{\underline{\gamma}}_n(Q)w(y)))q(dy|x, f(x)) \\
&\quad + R + Qw(x) \\
&\leq u(x, f(x)) + \int_X \delta(T_f^n \bar{U}(y))q(dy|x, f(x)) + \int_X (\gamma(\underline{\gamma}_n(R)) \\
&\quad + \gamma(\tilde{\underline{\gamma}}_n(Q)w(y)))q(dy|x, f(x)) + R + Qw(x) \\
&\leq u(x, f(x)) + \int_X \delta(T_f^n \bar{U}(y))q(dy|x, f(x)) + (R + \gamma(\underline{\gamma}_n(R))) \\
&\quad + (Q + \tilde{\gamma}(\tilde{\underline{\gamma}}_n(Q)))w(x),
\end{aligned}$$

where the second inequality is by the inductive supposition, the third, fourth and fifth inequalities are all by Condition 1. That is, by (3) applied to  $\gamma$  and  $\tilde{\gamma}$ ,

$$\bar{U}(x) \leq T_f^{n+1} \bar{U}(x) + \underline{\gamma}_{n+1}(R) + \tilde{\underline{\gamma}}_{n+1}(Q)w(x),$$

as required.  $\square$

*Proof of Theorem 1.* Lemma 6 asserts that

$$\bar{U}(x) \leq u(x, f^N(x)) + \int_X \delta(\bar{U}(y))q(dy|x, f^N(x)) + R + Qw(x) \quad \forall x \in X$$

with

$$R = (L_U \zeta_X + 3\zeta \gamma(\tilde{\underline{\gamma}}_\infty(\|u\|_w)) + 2\underline{\gamma}_\infty(\tilde{\Lambda})), \quad Q = 2\tilde{\gamma}^{(N)}(\tilde{\underline{\gamma}}_\infty(\|u\|_w)).$$

By Lemma 7, for each  $x \in \bigcup_{n \geq 0} K_n$ ,

$$\begin{aligned}
U(x) &= \bar{U}(x) \leq \lim_{n \rightarrow \infty} \left\{ T_f^{n+1} \bar{U}(x) + \underline{\gamma}_{n+1}(R) + \tilde{\underline{\gamma}}_{n+1}(Q)w(x) \right\} \\
&= U^{f^N}(x) + \underline{\gamma}_\infty(R) + \tilde{\underline{\gamma}}_\infty(Q)w(x),
\end{aligned}$$

where the first equality is by the definition of  $\bar{U}$ , and the last equality is by Lemma 1 and Proposition 1. The statement follows now because  $U^{f^N}(x) \leq U(x)$  for each  $x \in X$ .  $\square$

## 5.2 Proof of Theorem 2

We now sketch the proof of Theorem 2.

*Proof of Theorem 2.* One can show that

$$|Tv(x) - Tv(y)| \leq (L_u + L_q L_v)(1 + L_A)d_X(x, z) \quad \forall x, z \in X,$$

$U_n$  has a Lipschitz constant  $L'_{U_n} := \underline{\varphi}'_n(\Lambda)$ , and

$$\sup_{x \in K_n} |\hat{U}_{N-1,n}(x) - U_{N-n}(x)| \leq \underline{\gamma}_{N-n}(\tilde{\Lambda}_N) \quad \forall N \geq 1, 0 \leq n \leq N. \quad (13)$$

The above relations correspond to and can be established as in Lemma 2, Lemma 3(a) and Lemma 4, and  $\varphi'$  and  $\tilde{A}_N$  correspond to  $\varphi$  and  $\tilde{A}$ .

Let  $0 \leq n \leq N - 1$  be fixed, and consider some  $x \in K_n$  for now. Let  $z = p_{Z_n}^{K_n}(x)$ . Then  $\hat{U}_{N-1,n}(x) = \hat{U}_{N-1,n}(z)$ , and

$$\begin{aligned}
& U_{N-n}^{g^N}(x) - \hat{U}_{N-1,n}(x) \\
&= u(x, f_n^N(x)) + \int_X \delta(U_{N-(n+1)}^{g^N}(y)) q(dy|x, f_n^N(x)) - u(z, f_n^N(z)) \\
&\quad - \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|z, f_n^N(z)) \\
&= u(x, f_n^N(x)) + \int_X \delta(U_{N-(n+1)}(y)) q(dy|x, f_n^N(x)) \\
&\quad - \int_X \delta(U_{N-(n+1)}(y)) q(dy|x, f_n^N(x)) \\
&\quad + \int_X \delta(U_{N-(n+1)}^{g^N}(y)) q(dy|x, f_n^N(x)) - u(z, f_n^N(z)) \\
&\quad - \int_{K_{n+1}} \delta(\hat{U}_{N-1,n+1}(y)) q(dy|z, f_n^N(z)) \\
&\quad + \int_{K_{n+1}} \delta(U_{N-1,n+1}(y)) q(dy|z, f_n^N(z)) \\
&\quad - \int_X \delta(U_{N-1,n+1}(y)) q(dy|z, f_n^N(z)) \\
&\quad + \int_{X \setminus K_{n+1}} \delta(U_{N-1,n+1}(y)) q(dy|z, f_n^N(z)).
\end{aligned}$$

Consequently,

$$\begin{aligned}
& |U_{N-n}^{g^N}(x) - \hat{U}_{N-1,n}(x)| \\
&\leq |u(x, f_n^N(x)) - u(z, f_n^N(z))| \\
&\quad + \left| \int_X \delta(U_{N-(n+1)}(y)) q(dy|x, f_n^N(x)) - \int_X \delta(U_{N-1,n+1}(y)) q(dy|z, f_n^N(z)) \right| \\
&\quad + \int_X \left| \delta(U_{N-(n+1)}^{g^N}(y)) - \delta(U_{N-(n+1)}(y)) \right| q(dy|x, f_n^N(x)) \\
&\quad + \int_{K_{n+1}} \left| \delta(\hat{U}_{N-1,n+1}(y)) - \delta(U_{N-1,n+1}(y)) \right| q(dy|z, f_n^N(z)) \\
&\quad + \int_{X \setminus K_{n+1}} |\delta(U_{N-1,n+1}(y))| q(dy|z, f_n^N(z)),
\end{aligned}$$

where

$$\begin{aligned}
& \int_X \left| \delta(U_{N-(n+1)}^{g^N}(y)) - \delta(U_{N-(n+1)}(y)) \right| q(dy|x, f_n^N(x)) \\
&= \int_{X \setminus K_{n+1}} \left| \delta(U_{N-(n+1)}^{g^N}(y)) - \delta(U_{N-(n+1)}(y)) \right| q(dy|x, f_n^N(x)) \\
&\quad + \int_{K_{n+1}} \left| \delta(U_{N-(n+1)}^{g^N}(y)) - \delta(U_{N-(n+1)}(y)) \right| q(dy|x, f_n^N(x)) \\
&\leq \gamma(||U_{N-(n+1)}^{g^N} - U_{N-(n+1)}||_w) \zeta + \gamma(\sup_{x \in K_{n+1}} |U_{N-(n+1)}(x) - U_{N-(n+1)}^{g^N}(x)|)
\end{aligned}$$

by (7). Applying Condition 2 to  $u$  and Condition 4, we see

$$\begin{aligned}
& |U_{N-n}^{g^N}(x) - \hat{U}_{N-1,n}(x)| \\
&\leq L_u(1 + L_{\mathbf{A}})\zeta_X + L_q L'_{U_{N-(n+1)}}(d_X(x, z) + d_A(f_n^N(x), f_n^N(z))) \\
&\quad + \gamma(||U_{N-(n+1)}^{g^N} - U_{N-(n+1)}||_w)\zeta + \gamma(\sup_{x \in K_{n+1}} |U_{N-(n+1)}(x) - U_{N-(n+1)}^{g^N}(x)|) \\
&\quad + \gamma(\sup_{x \in K_{n+1}} |\hat{U}_{N-(n+1)}(y) - U_{N-(n+1)}(y)|) + \gamma(||U_{N-(n+1)}||_w)\zeta \\
&\leq \zeta_X(L_u + L_q L'_{U_{N-(n+1)}})(1 + L_{\mathbf{A}}) + \zeta(\gamma(||U_{N-(n+1)}^{g^N} - U_{N-(n+1)}||_w) \\
&\quad + \gamma(||U_{N-(n+1)}||_w)) \\
&\quad + \gamma(\underline{\gamma}_{N-(n+1)}(\tilde{A}_N)) + \gamma(\sup_{x \in K_{n+1}} |U_{N-(n+1)}(x) - U_{N-(n+1)}^{g^N}(x)|) \\
&\leq \underline{\varphi}'_{N-n}(A)\zeta_X + 3\underline{\zeta}\tilde{\gamma}_N(||u||_w) + \underline{\gamma}_N(\tilde{A}_N) \\
&\quad + \gamma(\sup_{x \in K_{n+1}} |U_{N-(n+1)}(x) - U_{N-(n+1)}^{g^N}(x)|),
\end{aligned}$$

where the second inequality is by (13), and the third inequality is by Proposition 2.

Now the previous inequality and (13) imply

$$\begin{aligned}
& \sup_{x \in K_n} |U_{N-n}^{g^N}(x) - U_{N-n}(x)| \\
&\leq \sup_{x \in K_n} |U_{N-n}(x) - \hat{U}_{N-1,n}(x)| + \sup_{x \in K_n} |\hat{U}_{N-1,n}(x) - U_{N-n}^{g^N}(x)| \\
&\leq \underline{\varphi}'_N(A)\zeta_X + 3\underline{\zeta}\tilde{\gamma}_N(||u||_w) + 2\underline{\gamma}_N(\tilde{A}_N) \\
&\quad + \gamma(\sup_{x \in K_{n+1}} |U_{N-(n+1)}(x) - U_{N-(n+1)}^{g^N}(x)|) \\
&= G + \gamma(\sup_{x \in K_{n+1}} |U_{N-(n+1)}(x) - U_{N-(n+1)}^{g^N}(x)|),
\end{aligned}$$

and by iteration, we see

$$\sup_{x \in K_n} |U_{N-n}^{g^N}(x) - U_{N-n}(x)| \leq \underline{\gamma}_{N-n}(G) \leq \underline{\gamma}_N(G).$$

Finally,

$$\begin{aligned}
& \sup_{x \in K_0} |U^{g^N}(x) - U(x)| \\
& \leq \sup_{x \in K_0} |U^{g^N}(x) - U_N^{g^N}(x)| + \sup_{x \in K_0} |U_N^{g^N}(x) - U_N(x)| + \sup_{x \in K_0} |U_N(x) - U(x)| \\
& \leq 2\tilde{\gamma}^{(N)}(\tilde{\gamma}_\infty(||u||_w)) \sup_{x \in K_0} w(x) + \underline{\gamma}_N(G),
\end{aligned}$$

where the last inequality is by Propositions 3 and 4.  $\square$

## References

1. Aliprantis, C., Border, K.: Infinite Dimensional Analysis. Springer, Heidelberg (2006)
2. Altman, E.: Constrained Markov Decision Processes. Chapman and Hall/CRC, Boca Raton (1999)
3. Băuerle, N., Jaśkiewicz, A., Nowak, A.: Stochastic dynamic programming with non-linear discounting. *Appl. Math. Optim.* (2020). <https://doi.org/10.1007/s00245-020-09731-x>
4. Bertsekas, D.: Convergence of discretization procedures in dynamic programming. *IEEE Trans. Autom. Control* **20**, 415–419 (1975)
5. Cioletti, L., Oliveira, E.: Applications of variable discounting dynamic programming to iterated function systems and related problems. *Nonlinearity* **32**, 853–883 (2019)
6. Dufour, F., Prieto-Rumeau, T.: Approximation of infinite horizon discounted cost Markov decision processes. In: Hernández-Hernández, D., Minjárez-Sosa, J. (eds.) Optimization, Control, and Applications of Stochastic Systems, pp. 59–76. Birkhäuser, Boston (2012)
7. Dufour, F., Prieto-Rumeau, T.: Approximation of Markov decision processes with general state space. *J. Math. Anal. Appl.* **388**, 1254–1267 (2012)
8. Feinberg, E.A., Kasyanov, P., Zadoianchuk, N.: Berge's theorem for noncompact image sets. *J. Math. Anal. Appl.* **397**, 255–259 (2013)
9. Hernández-Lerma, O., Lasserre, J.: Discrete-Time Markov Control Processes. Springer, New York (1996)
10. Jaśkiewicz, A., Matkowski, J., Nowak, A.: Persistently optimal policies in stochastic dynamic programming with generalized discounting. *Math. Oper. Res.* **38**, 108–121 (2013)
11. Jaśkiewicz, A., Matkowski, J., Nowak, A.: On variable discounting in dynamic programming: applications to resource extraction and other economic models. *Ann. Oper. Res.* **220**, 263–278 (2014)
12. Jaśkiewicz, A., Matkowski, J., Nowak, A.: Generalized discounting in dynamic programming with unbounded returns. *Oper. Res. Lett.* **42**, 231–233 (2014)
13. Kuntz, J., Thomas, P., Stan, G., Barahona, M.: Approximations of countably-infinite linear programs over bounded measure spaces. *SIAM J. Optim.* (2020). Preprint available via [arXiv:1810.03658v3](https://arxiv.org/abs/1810.03658v3)
14. Piunovskiy, A., Zhang, Y.: Continuous-Time Markov Decision Processes. Springer, Cham (2020)

15. Puterman, M.: *Markov Decision Processes*. Wiley, New York (1994)
16. Saldi, N., Linder, T., Yüksel, S.: *Finite Approximations in Discrete-Time Stochastic Control*. Springer, Cham (2018)
17. Sennott, L.: *Stochastic Dynamic Programming and the Control of Queueing Systems*. Wiley, New York (1999)



# Locks, Bombs and Testing: The Case of Independent Locks

Li Liu<sup>(✉)</sup> and Isaac M. Sonin

University of North Carolina at Charlotte, Charlotte, NC 28223, USA  
[{lliu26,imsonin}@uncc.edu](mailto:{lliu26,imsonin}@uncc.edu)  
<https://webpages.uncc.edu/imsonin/>

**Abstract.** We present a Defense/Attack resource allocation model, where Defender has some number of “locks” to protect  $n$  vulnerable boxes (sites), and Attacker is trying to destroy these boxes, having  $m$  “bombs,” which can be placed into boxes. Similar models were studied in game theory - (Colonel) Blotto games, but our model has a feature absent in previous literature. Attackers test the vulnerability of all sites before allocating their resources, and these tests are not perfect, i.e., a test can give plus for a box without a lock and minus for a box with a lock. We describe the optimal strategies for a version of this Locks-Bombs-Testing (LBT) model when locks appear independently in each box with the same probability.

**Keywords:** Defense/attack model · Blotto game · Search · Testing

**AMS(2020) Subject Classification:** Primary 91A27 · Secondary 90B40

## 1 Introduction

The problem of allocation of limited resources between different tasks is a classical problem in many areas of Operations Research, Economics, Finance and Engineering. This problem with a few players (participants) is an important field in Game Theory. In a classical Blotto game, two players distribute limited resources between different sites (battlefields) with the goal to win more sites, winning a site if you have more resources on this site than your opponent. There is substantial literature on this topic, with classic paper [13] and more recent publications, such as [12], where a complete solution of the “continuous” version was given, as well as [4], where some interesting extensions are discussed. In a comprehensive and detailed survey [5] dedicated to Search Games, the Blotto game is classified as an attack-defense game. There are even more papers dedicated to these games and as in all of Operations Research all classifications have many overlapping parts. As an example of an important paper on an attack-defense game we mention [11].

The inspiration for the model in this paper and some basic ideas can be traced to the paper by K. Sonin and A. Wright [14], where they provided a model of intelligence gathering in combat and used highly detailed data about Afghan rebel attacks, insurgent-led spy networks, and counterinsurgency operations. This theoretical model was a novel version of the Colonel Blotto's game. First, the government allocates its scarce defense resources across possible targets. Then, each target is independently tested for vulnerability. Finally, the rebels base their choice of the targets on the results of these tests. Empirically, the paper demonstrated a robust link between local economic conditions and the patterns of rebel attacks.

A more general and abstract mathematical model called the Locks, Bombs and Testing (LBT) model was described in [15, 16], where one important special case was solved. The solution for the other important case was obtained in the PhD thesis of Liu Li [8]. This thesis in a modified form is a substantial part of this paper.

First, we describe a Symmetrical LBT model, where all boxes are identical. As in most attack-defense games, the two players play quite different roles. We call one of them Defender (DF) and the other, Attacker (AT). There are  $n$  “boxes” (sites, battlefields, cells, targets, time slots, etc.) with equal values for both players. AT is trying to destroy these boxes by placing “bombs” that can result in explosions (destructions). One or more bombs can be placed into the same box. AT has  $m$ ,  $m = 1, 2, \dots$ , bombs to allocate among  $n$  boxes. A box is destroyed if at least one explosion occurs, and the explosions of different bombs in the same site or in different sites are independent. We denote by  $p$  the probability of explosion.

DF is trying to protect the boxes by distributing “locks” among them. A lock is a protection device which, when placed in a box, prevents its destruction with any number of bombs in it. Obviously, locks and bombs are just the names of discrete units of resources of protection and destruction. The number of locks  $k$ ,  $k < n$ , can be fixed, in which case it is an  $A(n, k)$  problem, the subject of paper [16], or it can be a random variable obtained when a lock appears in site  $i$  with probability  $\lambda$  independently of other boxes, in which case it is a  $B(n, \lambda)$  problem, the main subject of this paper. The latter assumption can represent either the uncertainty of DF about resources that will be available to her or the uncertainty of AT about how many locks will be distributed.

The important feature of both models in contrast to classical Blotto games is that AT can and will *test* every box, trying to find boxes without locks. This testing is not perfect: a test of site  $i$  may have a positive result,  $S_i = 1$ , even if there is no lock at the site,  $T_i = 0$ , and negative,  $S_i = 0$ , even if there is a lock,  $T_i = 1$ . The probabilities of correct identification of both types, in statistical language the *sensitivity* and *specificity*,  $P(S_i = 1|T_i = 1) = a$  and  $P(S_i = 0|T_i = 0) = b$ , are known to both players. The result of testing is a vector of signals  $s = (s_1, \dots, s_n)$ , where each  $s_i = 0, 1$  is known to AT. Hereafter we refer to this vector as *signal*  $s$ .

When the number of available locks  $k$ ,  $0 \leq k \leq n$ , became known to DF, then her *strategy* is a probability distribution  $b_k(\gamma)$  on a set of all possible positions

of locks  $\{\gamma\}$ , where  $\gamma = (i_1, i_2, \dots, i_k)$  with  $1 \leq i_1 < \dots < i_k \leq n$ . The case  $k = 0$  means that no locks are allocated. The collection of  $b_k(\gamma)$ ,  $k = 0, 1, 2, \dots, n$  defines the strategy of DF, hereafter  $b(\gamma)$ .

In our Bayesian setting we assume that the parameter  $k$  in problem A or parameter  $\lambda$  specifying the distribution of the random number of locks  $K$ , and prior distribution  $b(\gamma)$  are known to AT, although the positions of locks and their actual number  $k$  in problem B are not. After the locks are allocated by DF, AT tests all boxes, receives signal  $s$ , and then, using prior distribution  $b(\gamma)$  and the probabilities of signals  $p(s) \equiv p(s|b(\gamma))$ , calculates the posterior distribution of the positions of locks (ADL)  $b(\gamma|s)$ . Then for each signal  $s$  and each  $m$ , AT solves the problem of optimal allocation of  $m$  bombs  $u_{opt}(s|b(\gamma)) = (u_1(s), \dots, u_n(s))|b(\gamma))$ ,  $\sum_i u_i = m$ , trying to maximize the expected number of destroyed sites.

Note that our model for both problems has one special and important feature. AT's analysis and the solution consists of two parts. In the first part AT considers a *statistical problem* to find the posterior distribution of locks, given signal  $s$  and prior information. In this statistical problem the probability of explosion  $p$  and the bombs allocation do not participate at all.

The second part is to optimize the allocation of  $m$  bombs among  $n$  sites. Such allocation can be deterministic or use some randomization. WLOG, we can assume that the allocation of bombs is deterministic and an *optimal strategy of AT*  $\pi_{opt}(b(\gamma))$ , with respect to the strategy of DF  $b(\gamma)$ , is a collection of her optimal deterministic responses  $u_{opt}(s|b(\gamma)) \equiv u_{opt}(s) = (u_1(s), \dots, u_n(s))$  to each signal  $s$ , where  $u_i(s)$  is the number of bombs placed into site  $i$ ,  $i = 1, \dots, n$ ,  $\sum_i u_i(s) = m$ . This optimal strategy  $\pi_{opt}(b(\gamma))$  together with the prior distribution  $b(\gamma)$  results in the corresponding total expected damage (loss),  $L_{opt}(b(\gamma))$ .

The goal of DF is to select a prior distribution of locks  $b_*(\gamma)$  to minimize this loss. We assume that DF knows the parameters of testing  $a$  and  $b$  and the number of bombs  $m$ . Then the pair  $(b_*(\gamma), \pi_*)$ , where  $\pi_* = \pi_{opt}(b_*(\gamma))$  is an optimal strategy of AT with respect to strategy  $b_*(\gamma)$ , forms a classical Nash equilibrium (NE) point. The corresponding value of the game is  $v_* = L(b_*(\gamma), \pi_*)$ . Though  $b_*(\gamma)$  are not unique, they all have common properties that result in a unique (up to some randomization) AT strategy  $\pi_*$ , and thus a specific value of  $v_*$ .

We call this game the symmetrical LBT game (model)  $A(n, k)$  or  $B(n, \lambda)$  with parameters  $(n, k, m, a, b)$ , or correspondingly  $(n, \lambda, m, a, b)$ , where  $n$  is the number of sites,  $k$  is the fixed number of locks, and  $\lambda$  is the probability of a lock being present in the box.

In a more general setting parameter  $\lambda$  can be replaced by vector  $\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ , where  $\lambda_i$  is the probability of presence of a lock in box  $i$ , parameters  $a$  and  $b$  are replaced by vectors  $a = (a_i)$ ,  $1 \leq i \leq n$  and  $b = (b_i)$ ,  $1 \leq i \leq n$ , and parameter  $c_i = 1$  by an  $n$ -dimensional vector  $c = (c_i)$ ,  $1 \leq i \leq n$ , where vectors  $a, b, c$  represent the sensitivities and the specificities of testing, and the values of all sites, i.e.  $P(S_i = 1|T_i = 1) = a_i$  and  $P(S_i = 0|T_i = 0) = b_i$ ,  $1 \leq i \leq n$ . Hereafter we also refer to the general model as the general LBT model. The additional justification to limit the consideration to the symmetrical case is the

following. The general LBT game and the straightforward approach to solving it, described above, have two basic drawbacks. First, the set of possible positions for locks, i.e., the set of subsets of an  $n$  element set, generally has order  $2^n$ , and so does the set of potential signals. As a result, the calculations of posterior distributions  $b(\gamma|s)$  and their marginal distributions  $\alpha_i(s) = P(T_i = 0|s)$  which play a crucial role in the description of optimal strategies, become cumbersome for large  $n$ . The second problem is that the knowledge of detailed information about the values of  $c_i, a_i$ , and  $b_i$  in many cases is unrealistic. As a result, the main focus in [8, 16] was on the analysis of a simpler S-LBT model, where all sites have identical values  $c_i = 1$ , and all  $a_i = a, b_i = b$ .

The main goal of our paper is to present the complete solution of the  $B(n, \lambda)$  S-LBT game. One of our main results about S-LBT is that the optimal strategy of AT  $\pi(\cdot|m, s)$  depends only on probability of an explosion  $p$ , value  $x$  of rv  $N$ , the number of minuses in a signal, and the ratio  $r \equiv r_B = \frac{P(T=0|S=0)}{P(T=0|S=1)}$ . A similar statement is true for problem A with the ratio  $r(x) \equiv r_A(x) = \frac{P(T=0|S=0,x)}{P(T=0|S=1,x)}$ . Both ratios depend on the parameters of the model,  $n, k, \lambda, a, b$ .

When the parameters of sensitivity  $a$  and specificity  $b$  are “informative”, these ratios are more than one. We show later that “informative” means that  $a + b > 1$ . This immediately implies that if there is only one bomb and signal  $s$  has pluses and minuses then a bomb goes to a minus box. When the number of bombs  $m$  exceeds  $x$ , optimal strategies can be expressed through  $r_B, r_A(x)$ , and other parameters.

As a result, the optimal strategy in both problems will have a much simpler structure than in the general case, symmetrical with respect to all sites with minus signals, and correspondingly for sites with plus signals. We describe this strategy on a heuristic level immediately.

The optimal strategy in both problems depends on the number  $m$  of bombs available and, given  $N = x, 0 \leq x \leq n$ , has the following structure. Initially, all bombs are placed one by one into each of  $x$  minus boxes until the threshold level  $d$ ,  $d_A(x)$  in Problem A or level  $d_B$  in Problem B is reached in each of them or the bombs are exhausted. Afterwards, the bombs are added one by one to plus boxes until there is a bomb in each of them. Then, bombs are added one by one into minus boxes until each of these boxes has  $d + 1$  bombs in each of them, then back to plus boxes until each has at least 2 bombs, etc. This “fill and switch” process stops when AT runs out of bombs. We will call such a strategy a *d-uniform as possible* strategy (hereafter, a “*d-UAP* strategy”). If  $x = 0$  or  $n$ , then all boxes are simply filled sequentially, and this is a 0-UAP strategy. The outcome of this process will be an allocation in which either all plus boxes will have the same number of bombs in each of them, and then all minus boxes either also have the same number of bombs in each of them, or some minus boxes have one extra bomb in comparison with the other minus boxes. A similar symmetrical situation takes place when all minus boxes have the same number of bombs in each of them. If all boxes with plus signals have no bombs, then the number of bombs in minus boxes does not exceed  $d$ . Note that though in problem B the ratio  $r_B$  and the threshold value  $d_B$  do not depend on the number

of minus boxes  $x$ , the allocation of bombs and then the value function for each signal do depend on that parameter. In a sense, the values  $N = x$  and  $r_B(\lambda)$  ( $r_A(x)$ ) play the role of sufficient statistics in the optimization problem. The value of the threshold  $d_B$  represents the “advantage level” of a minus box over a plus box. A similar interpretation can be given to the threshold  $d_A(x)$  given that  $x$  minuses were observed.

In an example with  $n = 5$ ,  $x = 3$ ,  $m = 12$ , and  $d = 3$ , the 3-UAP strategy is to place 3 bombs into 2 minus boxes, 4 into the third minus box, and 1 bomb into each of the 2 plus boxes. When  $d = 2$ , each of the minus boxes has 3 bombs, 1 plus box has 1, and the second plus box has 2.

Though the solutions of both problems have certain similarities, some of their features are very distinct. For example, an interesting and even counter intuitive property is that in the problem  $A(n, k)$ , the function  $r_A(x)$  and therefore the optimal strategy and the value function, depend on  $a$  and  $b$  only through the value  $c = \frac{a}{1-a} \frac{b}{1-b}$ , a combined characteristic of the quality of testing. In the problem  $B(n, \lambda)$ , this property does not hold with respect to the value  $r_B(\lambda)$ . The other important distinction between these two problems is that in the former problem the minuses and pluses in different boxes are not independent, but in the latter problem they are.

Note that *sites, locks, bombs and testing* in this and more general models are rather abstract terms and may have very different interpretations beyond our initial exposition of the DF and AT defense-attack game. We refer to [14, 15] for a more detailed exposition. It is easy to extend the LBT model in many different directions. Here we mention only that the *dynamic version* of the LBT model will have common features and in a sense will be a very broad generalization of the well-known model in Applied Probability—the Multi Armed Bandit problems. This model was studied in many papers and a few books—D. Berry and B. Fristedt (1985), E. Presman and I. Sonin (1987, 1990), J. Gittins (1989), J. Gittins, K. Glazebrook and R. Weber (2011), and the current internet version by T. Lattimore and C. Szepesvari (2019). The full solution of the general LBT game is a difficult task though some special cases were presented in [15].

The structure of our paper is as follows. In Sect. 2 we present some preliminary formulas and auxiliary results. In Sect. 3 we obtain optimal strategies for problem B, and in Sect. 4 we consider corresponding examples.

We thank Michael Grabchak, Ernst Presman, Mark Whitmeyer and Fedor Sandomirsky for their valuable remarks and helpful discussions, and patience with reading numerous drafts.

## 2 Preliminary Formulas and Auxiliary Results

Though the main focus of our paper is on the problem  $B \equiv B(n, \lambda)$ , we also provide for the comparison some details from [16] about model  $A \equiv A(n, k)$ . Also of possible interest is to compare the strategies and the value functions for both

problems for the “matching” values of  $k$  and  $\lambda$ , i.e. when the *expected number* of locks in  $n$  boxes is the same,  $\lambda n = k$ . We present some numerical results in Sect. 4.

The following notation is used throughout the paper. Define random variable  $T_i, S_i, C_i$ , each taking two values 0 and 1:  $T_i = 1$  if and only if the  $i$ th box is protected;  $S_i = 1$  if and only if the  $i$ th box test is positive, i.e. the protection is present, ( $s_i = 1$ ), and  $C_i = 1$  if and only if the  $i$ th box is destroyed. The absence of subindex  $i$  means that the formula applies to any box. Our assumptions imply the following basic equations:

$$\begin{aligned} P(S_i = 1 | T_i = 1) &= a, \quad P(S_i = 0 | T_i = 0) = b, \\ P(C_i = 1 | T_i = 1) &= 0, \quad P(C_i = 1 | T_i = 0, u_i) = p(u_i), \end{aligned} \quad (1)$$

where  $u_i$  is the number of bombs in box  $i$ , and  $p(u)$  is the success function, the probability of at least one explosion in a box with  $u$  bombs. As we assumed that the success is independent across bombs,  $p(u) = 1 - (1 - p)^u$ . The function  $p(u)$  is increasing and upward concave, and the function  $\Delta p(u) \equiv p(u + 1) - p(u)$  is decreasing. The diminishing effect of each extra bomb will play an important role in determining the optimal strategy.

## 2.1 Basic Notation and Lemma 1

Of possible interest in both models are the aposterior probabilities  $P(T_i = 0|s), s = (s_1, \dots, s_n)$  and the *aposterior distribution of locks* (ADL)  $b(\gamma|s) = P(T_i = 1, i \in \gamma, T_i = 0, i \notin \gamma | S_i = s_i, i = 1, \dots, n)$ .

To describe these distributions, given that the number of locks  $k$  is fixed, let us introduce rvs  $N_1$ , the number of minuses in locked boxes, i.e. the number of *false minuses*, or equivalently, the number of locks in boxes with minuses,  $N_2$ , the number of minuses in unlocked boxes, i.e. the number of *correct minuses*, and  $N = N_1 + N_2$ , the total number of minuses after testing. The rv  $N_1$  is a binomial rv with  $k$  trials and probability of success  $1 - a$ , the rv  $N_2$  is a binomial rv with  $n - k$  trials and probability of success  $b$ . These two random variables are independent, and unless  $b \neq 1 - a$ , rv  $N = N_1 + N_2$ , taking values  $0, 1, \dots, n$ , is not a binomial rv. Sometimes the distribution of  $N$  is called the Poisson binomial distribution. The signal  $s = (s_1, \dots, s_n)$  and the value  $N = x$  are *observable* in contrast to the values of  $N_1$  and  $N_2$ , which are not. To stress this point, sometimes we use the notation  $t = N_1(\gamma, s), x = N(s)$ .

We denote by  $p_i(j)$  the pmf (probability mass function) of the binomial rvs  $N_i, i = 1, 2$  and  $p(j|r, p), j = 0, 1, \dots, r$ , the pmf of a binomial distribution with  $r$  trials and probability of success  $p$ . Thus  $p_1(j) = p(j|k, 1 - a)$  and  $p_2(j) = p(j|n - k, b)$ . Then the pmf of rv  $N$  in problem A,  $g_A(x) \equiv g_{n,k}(x), 0 \leq x \leq n$ , can be calculated by standard *discrete convolution formula*, the first formula below.

In problem B the number of locks is rv  $K$  with a binomial distribution with  $n$  trials and probability of success  $\lambda$ . Thus rv  $K$  has distribution  $p(k|n, \lambda), k = 0, 1, \dots, n$ . Given  $K = k$ , rv  $N$  has conditional distribution  $g_{n,k}(x)$ , and then  $g_B(x) \equiv P(N = x)$  can be calculated by the second formula below

$$\begin{aligned} g_A(x) &\equiv g_{n,k}(x) = \sum_j p_1(j)p_2(x-j) \equiv \sum_t p_1(x-t)p_2(t), \\ g_B(x) &= \sum_{k=0}^n p(k|n, \lambda)g_{n,k}(x). \end{aligned} \quad (2)$$

The summation over  $j$  in the convolution formula above is taken over values  $j$  such that  $0 \leq j \leq k$ ,  $0 \leq x - j \leq n - k$ . Similar holds for the summation over  $t$ , where  $0 \leq x - t \leq k$ ,  $0 \leq t \leq n - k$ . Further in all convolution formulas we may omit the exact range of summation assuming that all probabilities involved in the sums are well defined.

The distribution of locks and the testing may be viewed as a two-stage random experiment with outcomes represented by pairs  $(\gamma, s)$ , where  $\gamma_k \equiv \gamma = (i_1, i_2, \dots, i_k)$  with  $1 \leq i_1 < \dots < i_k \leq n$  is a (vector) *allocation of  $k$  locks* and  $s = (s_1, \dots, s_n)$  is a (vector) *signal* about boxes vulnerability. In Problem A  $k$  is a fixed number, and in Problem B,  $0 \leq k \leq n$  is a result of a binomial experiment. The probability of each outcome is  $P(\gamma, s) = b(\gamma)P(s|\gamma)$ , where  $b(\gamma)$  is prior distribution of locks, and  $P(s|\gamma) = P(S_1 = s_1, \dots, S_n = s_n|\gamma)$ . Given  $\gamma = (i_1, i_2, \dots, i_k)$  and the prior distributions of locks, AT, using Bayes' formula can obtain the posterior distributions of locks  $b_k(\gamma|s) = P(T_i = 1, i \in \gamma, T_i = 0, i \notin \gamma|s)$ . The collections of these probabilities for different signals  $s$  and  $k$  in Problem B give  $b(\gamma)$  and  $b(\gamma|s)$ .

In our model, DF has no information about the allocation of bombs by AT. Therefore, her strategy in a Nash equilibrium point is straightforward: distribute  $k$  available locks between  $n$  boxes uniformly, i.e. the prior distribution  $b_k(\gamma)$  is uniform. In statistical physics, this distribution is called the Fermi-Dirac statistics: any combination of  $k$  protected boxes has the same probability  $1/\binom{n}{k}$ . It is easy to see that the probability of protection for each individual box is  $t = \frac{k}{n}$ . The similar probability for Problem B is  $\lambda$ . The substantial difference between models is that the rvs  $T_i$  are independent in B but not in A.

Thus, our main interest is in AT's strategy. Given signal  $s$  and prior distribution of locks  $b(\gamma)$ , AT can obtain *aposterior distribution* of locks (ADL)  $b(\gamma|s)$ . To construct an optimal allocation for each signal  $s$ , AT has to obtain values  $P(T_i = 0|s_i = 0, s_{-i})$  and  $P(T_i = 0|s_i = 1, s_{-i})$  and their ratios, where  $s_{-i}$  is an  $n - 1$ -dimensional vector  $s = (s_1, \dots, s_n)$  without coordinate  $s_i$ . In both problems the symmetry of minus and plus boxes gives a hint that for S-LBT the only information necessary besides the signal in a particular box is the total number of minuses. To justify this assertion we need a few results.

We start with the following lemma with two intuitively appealing observations. First, the posterior probability of signal distribution is uniform conditional on the number  $x$  of minus signals, and second, the posterior probability that box  $i$  has no lock conditional on the full vector signal  $s = (s_1, \dots, s_n)$  is equal to the conditional probability that box  $i$  has no lock conditional only on the individual signal  $s_i$  and the total number of minus signals.

**Lemma 1.** a) For problems A and B, for any signal  $s$  and any  $x = 0, 1, \dots, n$

$$P(s|N = x) = 1/\binom{n}{x}. \quad (3)$$

b) For problem B, for any signal  $s$  and any  $x = 0, 1, \dots, n$

$$P(T_i = 0|s, N = x) = P(T_i = 0|s_i). \quad (4)$$

c) For problem A, for any signal  $s$  and any  $x = 0, 1, \dots, n$

$$P(T_i = 0|s, N = x) = P(T_i = 0|s_i, N = x), \quad (5)$$

*Proof.* a) The symmetry of signals and boxes implies that  $P(s|N = x) = c(x)$ , i.e. all signals with the same number  $x$  of signals  $s = 0$  have the same probability. Let  $\Sigma(x) = \{s : N(s) = x\}$ . Then, since  $|\Sigma(x)| = \binom{n}{x}$  and  $\sum_{s \in \Sigma(x)} P(s|N = x) = 1$ , we obtain that  $P(s|N = x)$  is given by the equality in (3).

- b) For Problem B, the equality in (4) is obvious since the result of the test of box  $i$  does not depend on the presence of locks and the results of the testing in other boxes.
- c) The formal, non-trivial proof of the equality in (5), can be found in paper [16].  $\square$

The formulas in Lemma 1 are at the heart of the intuition behind our main results.

## 2.2 Lemma 2 and Key Ratio $r_B$

To find the optimal strategy of AT we need to know the probabilities of a destruction of a minus and plus boxes with  $u$  bombs in a box, i.e.  $P(C_i = 1|s_i, u)$ ,  $s_i = 0, 1$ . These probabilities in turn depend on the probabilities of an absence of a lock in minus and plus boxes, i.e.  $P(T_i = 0|s_i)$ ,  $s_i = 0, 1$ . The optimal strategy will be defined by the likelihood ratio of the latter probabilities,  $r_B = P(T_i = 0|s_i = 0)/P(T_i = 0|s_i = 1)$ . They are all described in Lemma 2. This lemma shows that in Problem B, in contrast to Problem A, the ratio  $r_B$  does not depend on  $x$  but does depend on parameters  $a, b, \lambda$  and is given by an explicit formula, where we use the shorthand notation  $h = a + b - 1$ .

**Lemma 2.** a) The probabilities of an absence of a lock in a minus and plus boxes, and the corresponding probabilities of destruction of a minus and plus boxes with  $u$  bombs are

$$P(T_i = 0|s_i = 0) = p^- = \frac{(1 - \lambda)b}{\lambda(1 - a) + (1 - \lambda)b} = \frac{(1 - \lambda)b}{b - \lambda h}, \quad (6)$$

$$P(T_i = 0|s_i = 1) = p^+ = \frac{(1 - \lambda)(1 - b)}{\lambda a + (1 - \lambda)(1 - b)} = \frac{(1 - \lambda)(1 - b)}{1 - b + \lambda h}, \quad (7)$$

$$P(C_i = 1|s_i = 0, u) = p^- p(u), \quad P(C_i = 1|s_i = 1, u) = p^+ p(u); \quad (8)$$

b) The ratio  $r_B$  for Problem B is given by the formula

$$r_B = \frac{p^-}{p^+} = \frac{b}{1-b} \frac{1-b+\lambda h}{b-\lambda h}, \quad 0 < \lambda < 1; \quad (9)$$

c) If  $a+b > 1$  then function  $r_B(\lambda)$  is increasing from 1 to  $\frac{a}{1-a} \frac{b}{1-b} = c_1 c_2 = c > 1$ , when  $\lambda$  is increasing from 0 to 1;  
if  $a+b < 1$  then function  $r_B(\lambda)$  is decreasing from 1 to  $c < 1$ ; and if  $a+b = 1$  then  $r_B(\lambda) = 1$ .

*Proof.* The conditional independence of testing and explosions, and equalities in (1) and (4) imply that the probability of destruction of box  $i$  with  $u_i = u \geq 1$  bombs, given signal  $s$  with  $N = x, S_i = s_i$ , is

$$P(C_i = 1|s, x, u) = P(T_i = 0|s, x)P(C_i = 1|T_i = 0, u) = P(T_i = 0|s_i)p(u). \quad (10)$$

We also have the equalities

$$\begin{aligned} p^- &= P(T = 0|S = 0) = P(T = 0)P(S = 0|T = 0)/P(S = 0) \\ &= (1 - \lambda)b/P(S = 0), \\ P(S = 0) &= P(T = 1)P(S = 0|T = 1) + P(T = 0)P(S = 0|T = 0) \\ &= \lambda(1 - a) + (1 - \lambda)b, \\ p^+ &= P(T = 0|S = 1) = P(T = 0)P(S = 1|T = 0)/P(S = 1) \\ &= (1 - \lambda)(1 - b)/P(S = 1), \\ P(S = 1) &= P(T = 1)P(S = 1|T = 1) + P(T = 0)P(S = 1|T = 0) \\ &= \lambda a + (1 - \lambda)(1 - b). \end{aligned}$$

Using these equalities and formula (10) with  $s_i = 0$  and  $s_i = 1$ , we obtain all formulas in points (a) and (b).

To prove c), note that it is easy to check that  $\frac{d}{d\lambda} r_B(\lambda) = \frac{b}{1-b} \frac{h}{f(\lambda)^2}$ , where  $f(\lambda) = b - \lambda h$  for all  $0 < \lambda < 1$ , and that  $r_B(0) = 1, r_B(1) = \frac{a}{1-a} \frac{b}{1-b} = c_1 c_2 = c$ . It is easy to check that the inequality  $a+b > 1$  is equivalent to  $c = r_B(1) > 1$ . Hereafter we assume that  $h = a+b-1 > 0$  and therefore  $r_B(\lambda) > 1$  for all  $\lambda > 0$ .  $\square$

Note that  $c_1$  and  $c_2$  represent the quality of sensitivity and specificity, and  $c$  represent the combined quality of testing.

*Remark 1.* Lemma 2 implies that testing is not very informative if locks are rare, i.e.  $\lambda$  is small, even when the parameters of the testing are very good, i.e.  $a$  and  $b$  are close to 1. When there are many locks, i.e.  $\lambda$  is close to one, the “amount of information” is limited by the combined characteristic of the quality of testing, parameter  $c$ .

Note also that parameters  $a$  and  $b$  in function  $r_B$  are not symmetrical, i.e., though  $r_B(.5|a,b)r_B(.5|b,a) = c$  and  $r_B(\lambda|a,b) \approx r_B(\lambda|b,a) \approx c$  for  $\lambda$  close to 1, generally  $r_B(a,b) \neq r_B(b,a)$  for all  $\lambda < 1$ . This asymmetry property is in sharp contrast to the symmetry of  $a$  and  $b$  for  $r_A(x)$  in Problem A.

### 3 Main Results and Their Proofs

Let  $B^-(s) = \{i : s_i = 0\}$  and  $B^+(s) = \{i : s_i = 1\}$  denote the sets of minus and plus boxes for signal  $s$ . Using the equality in (4), we obtain that, given a strategy  $\pi = (u_1, \dots, u_n)$  for  $m$  bombs, and any signal  $s$  with  $N(s) = x$ , the value of a strategy  $\pi$ , i.e. the expected number of destroyed boxes, is

$$\begin{aligned} w^\pi(s) \equiv w^\pi(s, x) &= \sum_{i=1}^n P(T_i = 0 | s_i) p(u_i) \\ &= p^- \sum_{i \in B^-(s)} p(u_i) + p^+ \sum_{i \in B^+(s)} p(u_i). \end{aligned} \quad (11)$$

Let  $U^- \equiv U^-(\pi|s) = \{u_j, j \in B^-(s)\}$  and  $U^+ \equiv U^+(\pi|s) = \{u_j, j \in B^+(s)\}$  be the two possible sets of the values of  $u_j$  at minus and plus boxes. Formula (11) immediately implies that all strategies obtained by permutations of sets  $(U^-, U^+)$  among corresponding boxes have the same value denoted as  $w^\pi(x) \equiv w^\pi(x, m)$ , where  $m$  is the number of available bombs. Hereafter we use notation  $w^\pi(s) = w^\pi(x)$ , where  $x = N(s)$ .

We denote  $v(x, m) = \sup_\pi w^\pi(x, m)$ , the *value function* over all strategies, given  $m$  and  $x$ , and  $v(m)$ , the value function over all strategies and all possible values of  $x$ , i.e.  $v(m) = \sum_x g_B(x)v(x, m)$ , where  $g_B(x) = P(N = x)$  is given by the formula in (2).

Formula (11) gives a hint that the proportion of the number of bombs placed into a minus site to the number of bombs placed into a plus site is defined by ratio  $r_B = p^-/p^+$ , given in formula (9).

Let us assume for simplicity that the number of bombs is the same in all minus boxes, and a similar statement holds for all plus boxes, i.e.:  $u_i = u^-, i \in B^-(s)$  and  $u_i = u^+, i \in B^+(s)$ . Then the role of ratio  $r = r_B$  becomes clear, since formula (11) takes the form of

$$w^\pi(s, x) \equiv w^\pi(x) = p^+[r_B x p(u^-) + (n - x)p(u^+)],$$

where  $p^+$ ,  $p^-$ ,  $r$  are given by the formulas in (6), (7) and (9).

To obtain an optimal strategy, we use a natural and obviously necessary equilibrium condition: *with an optimal allocation of bombs it is impossible to increase the payoff by moving essentially a bomb from one box to another*. Essentially means changing sets of the values of  $u_j$  at minus and plus boxes, i.e. sets  $U^-(\pi|s)$  and  $U^+(\pi|s)$ .

We remind that in the Introduction we heuristically described the potential structure of optimal strategies in both problems, namely that they should be  $d$ -UAP strategies with some values of  $d = 0, 1, 2, \dots$ . We will prove the optimality of a

$d$ -UAP strategy by showing that any other strategy does not satisfy this condition. Later we give formulas for the optimal value  $d = d_B$ .

The following two lemmas describe the properties of optimal strategies. Lemma 3 proves that an optimal strategy is nearly uniform inside of minus and plus boxes: if the signals in two boxes have the same sign, then the optimal number of bombs can differ at most by 1.

**Lemma 3.** *Let  $\pi(x) = (u_i, i = 1, 2, \dots, n)$  be an optimal strategy. Then  $|u_r - u_t| \leq 1$  when the signals in boxes  $r, t$  have the same sign, i.e.  $s_r = s_t$ .*

*Proof.* In proof of Lemma 2, see formula (10) and after, we obtained the equalities

$$P(C_t = 1 | s_t = 1, u) = p^+ p(u), \quad P(C_t = 1 | s_t = 0, u) = p^- p(u) = r_B p^+ p(u). \quad (12)$$

Suppose that Lemma 3 is not true—say for boxes 1 and 2 with  $s_1 = s_2 = 1$ ,  $u_1 = i, u_2 = j$  and  $j - i \geq 2$ . The concavity of function  $p(u)$  implies that  $p(i+1) + p(j-1) > p(i) + p(j)$ . Then, using the first equality in (12) for  $t = 1$  and  $t = 2$ , we obtain

$$\begin{aligned} & P(C_1 = 1 | s_1 = 1, u_1 = i+1, x) + P(C_2 = 1 | s_2 = 1, u_2 = j-1, x) \\ &= p^+[p(i+1) + p(j-1)] > p^+[p(i) + p(j)] \\ &= P(C_1 = 1 | s_1 = 1, u_1 = i, x) + P(C_2 = 1 | s_2 = 1, u_2 = j-1, x). \end{aligned} \quad (13)$$

Thus the initial allocation of bombs is not optimal. The proof for  $s_1 = s_2 = 0$  is similar, using the second equality in (12) and replacing  $p^+$  by  $p^- = r_B(\lambda)p^+$ .  $\square$

Once we have established that the optimal numbers of bombs in boxes with the same signal can differ by not more than 1, our next step is to find the optimal values of  $m^-, m^+$ , the numbers of bombs in minus and plus boxes. Given  $N = x$ ,  $0 \leq x \leq n$ , the number  $m$  of bombs available and a  $d$ -UAP strategy, there is a unique allocation of bombs, given by tuple  $(l^-, e^-, l^+, e^+)$ , where  $l^-, l^+$  are the numbers of “complete layers” of bombs in minus and plus boxes, respectively, and  $e^-, e^+$  are the numbers of “extra” bombs in the “incomplete layers”. Hereafter we use shorthand notation  $l^- = l, e^- = e$ . Note that  $e$  indicates also how many minus boxes have an extra bomb among all minus boxes. The same is true for plus boxes. Thus  $0 \leq e < x, 0 \leq e^+ < n - x$  and  $e \times e^+ = 0$ . All these terms depend on  $m, x$  and  $d = d_B$  but we do not indicate this explicitly. We have

$$m^- = l \times x + e, \quad m^+ = l^+ \times (n - x) + e^+$$

Thus, if  $e^+ > 0$  then  $e = 0$ , and  $l - l^+ = d$ , and if  $e > 0$  then  $e^+ = 0$  and either  $l^+ = 0, l - l^+ < d$  or  $l^+ > 0, l - l^+ = d$ .

Let us define the threshold levels  $d$  for Problem B by the formula

$$d_B = \min_{i \geq 1} \left\{ i : r_B (1 - p)^i < 1 \right\}. \quad (14)$$

The optimality of this level, and hence the optimality of the corresponding strategy is proved in Lemma 4.

**Lemma 4.** Let  $\pi(x) = (u_i, i = 1, 2, \dots, n)$  be an optimal strategy,  $0 < x < n$ ,  $(u^-, u^+)$  a pair of bombs in some pair (minus box, plus box), and  $d = d_B$  be defined by formula (14). Then a  $d$ -UAP strategy with  $d = d_B$  defined by formula (14) is an optimal strategy.

*Proof.* As always, we assume that  $a + b > 1$  and then  $r_B > 1$ , and hence  $u^- \geq u^+$ . We will show that if  $u^- - u^+ > d$  for this pair of minus and plus boxes, then a transfer of one bomb from a minus box from this pair to a plus box will increase the value of a strategies. Similarly, if  $u^+ \geq 1$  and  $u^- - u^+ < d - 1$  for such pair, then the inverse transfer will also increase the value. Let  $u^- = i, u^+ = j, P(\cdot|N = x) = P(\cdot|x)$ , and denote the incremental utilities for minus and plus boxes as  $\Delta C^-(i|x) = P(C = 1|i+1, S = 0, x) - P(C = 1|i, S = 0, x)$ ,  $\Delta C^+(j|x) = P(C = 1|j+1, S = 1, x) - P(C = 1|j, S = 1, x)$ . Then, using the formulas in (12) and (13) with  $u = i$  and  $u = j$ , we obtain that their difference for  $0 \leq j \leq i$  is with  $q = 1 - p$ ,

$$\begin{aligned}\Delta(i, j|x) &= \Delta C^-(i|x) - \Delta C^+(j|x) = pq^i r_B p^+ - pq^j p^+ \\ &= pq^j p^+ (r_B q^{i-j} - 1).\end{aligned}\quad (15)$$

The definition of  $d = d_B$  in (14) implies that  $\Delta(i, j|x)$  is positive if  $j = 0, i < d$ , or if  $j \geq 1, i - j < d$ . Similarly,  $\Delta(i, j|x)$  is negative if  $j = 0, i \geq d$ , or if  $j \geq 1, i - j \geq d$ . The optimality of  $d$ -strategy is proven.  $\square$

Note also, that if  $p = 1$ , i.e.  $q = 0$ , then  $d = 1$  for all  $0 < x < n$ , and if  $p$  is decreasing to zero, then  $d$  tends to infinity. Now we are ready to formulate our main theorem for Problem B. As usual we assume that  $a + b > 1$  and hence  $r_B > 1$  and then  $0 < x < n$ .

**Theorem 1 (Value function for  $B(n, \lambda)$ ).** Suppose that, given signals, the total number of minus boxes, with the total number of minuses  $N = x$ ,  $0 \leq x \leq n$ .

a) If  $x = n$ , (or  $x = 0$ ), then the optimal strategy is 0-UAP and the value function  $v(m|n) = v(m|0)$  for  $m = n \times l + e$ ,  $l = 0, 1, \dots$ ,  $0 \leq e < n$ , ( $l = l^-, e = e^-$ ) is given by the formula

$$v(n|m) = v(0|m) = (1 - \lambda)[ep(l + 1) + (n - e)p(l)]. \quad (16)$$

b) If  $0 < x < n$ , then the optimal strategy is  $d$ -UAP strategy, where  $d = d_B$  is defined by formula (14) and  $r_B$  is defined by formula (9). The value function  $v(x, m)$  for  $m = m^- + m^+ = l \times x + e + l^+ \times (n - x) + e^+$ , is given by formula  $v(x, m) = p^+(\lambda)[r_B(\lambda)(ep(l+1)+(x-e)p(l))+(e^+p(l^++1)+(n-x-e^+)p(l^+))]$ .  $\quad (17)$

(c) The value function  $v(m)$ ,  $m = 1, 2, \dots$  is given by formula

$$v(m) = \sum_{x=0}^n g_B(x)v(x, m), \text{ where } g_B(x) = P(N = x), \quad (18)$$

is given by formula in (2).

*Proof.* (a) If  $x = 0$  or  $n$ , i.e. all boxes have the same minus or plus sign, and signal  $s$  brings no information, Lemma 4 implies that an optimal strategy is 0-UAP. When  $m = n \times l + e$ , where  $0 \leq e < n$ , then 0-UAP means that  $e$  boxes have  $l + 1$  bombs each, and  $n - e$  boxes have  $l$  bombs each. The probability that a particular box has no lock is  $1 - \lambda$ . Then the expected damage in all  $n$  boxes is

$$(1 - \lambda)[eP(C = 1|l + 1, T = 0) + (n - e)P(C = 1|l, T = 0)] \\ = (1 - \lambda)[ep(l + 1) + (n - e)p(l)].$$

i.e.,  $v(n|m) = v(0|m)$  is given by formula (16).

- (b) If  $0 < x < n$ , then by Lemma 4, an optimal strategy is  $d$ -UAP, and hence  $m^-$ ,  $m^+$  satisfy the equalities  $m^- = l \times x + e$ ,  $m^+ = l^+ \times (n - x) + e^+$ ,  $0 \leq e < x$ ,  $0 \leq e^+ < n - x$ ,  $ee^+ = 0$ . Then each of  $e$  minus boxes has  $l + 1$  bombs each, and  $x - e$  minus boxes have  $l$  bombs each, and in plus boxes  $e^+$  boxes have  $l^+ + 1$  bombs each, and  $n - x - e^+$  boxes have  $l^+$  bombs each. Then using formulas in Lemma 2 for the probabilities of destruction for minus and plus boxes we obtain formula (17). We proved b). The proof of (c) is straightforward.  $\square$

*Remark 2.* For computational purpose, the formulas in (16), (17), and (18) can be represented recursively in  $m$ .

*Remark 3.* By definition of  $d_B$ , let  $d_B = d$ , we have  $r_B q^{d-1} \geq 1$ . If  $r_B q^{d-1} > 1$ , then the  $d$ -UAP strategy is the unique optimal strategy. If there is an equality, then there are other optimal strategies with the allocation of bombs obtained as follows. When all minus sites are filled with  $d - 1$  full layers, the next bomb, if available, can be placed either in a minus site or in a plus site. The incremental utility will be the same. And so on with other extra bombs. As a result, the difference between the full layers in the minus and the plus sites can be either  $d$  or  $d - 1$ .

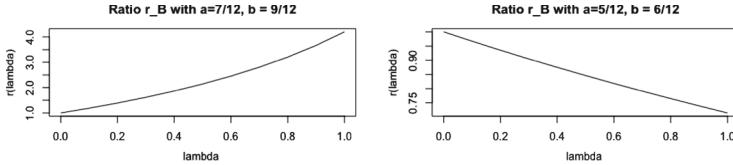
A theorem similar to Theorem 1 holds for the case A. The full version of this theorem, with formulas for the value functions  $v(m, x)$  and  $v(m)$  for all  $0 \leq x \leq n$ , and  $m$ , can be found in [16].

## 4 Examples

We will analyze the following pairs of examples when the expected number of locks in  $B(n, \lambda)$  matches the fixed number  $k$  in  $A(n, k)$ , i.e.  $n\lambda = k$ .

*Example 1 (Ratio r for  $B(n, \lambda)$  and  $A(n, k)$ ).* Let  $n = 2$ ,  $a = \frac{7}{12}$ ,  $b = \frac{9}{12}$ . For  $B(2, \lambda)$ , by formula (9), when  $\lambda = 1/2$ , we obtain  $r_B(1/2) = \frac{15}{7} \approx 2.143$ ; when  $\lambda = 1$ ,  $r_B(1) = 4.2 = c$ . With  $a = \frac{9}{12}$ ,  $b = \frac{7}{12}$ , we obtain  $r_B(1/2) = \frac{49}{25} = 1.96$ ,  $r_B(1) = 4.2 = c$ . Hence  $r_B(1/2|a, b)r_B(1/2|b, a) = \frac{21}{5} = 4.2 = c$ . When  $a + b > 1$ , ratio  $r_B$  as a function of  $\lambda$  is increasing, while when  $a + b < 1$ , the ratio is decreasing, as shown in Fig. 1.

Let  $n = 2$ ,  $k = 1$ ,  $a = \frac{7}{12}$ ,  $b = \frac{9}{12}$ . For  $A(2, 1)$ , we have  $r_A(1) = \frac{21}{5} = 4.2 = c$ ,  $r_A(0)$  is not defined, and  $r_A(2)$  is also not defined. For any problem  $A(n, n - 1)$  we have  $r_A(1) = c = \frac{21}{5} = 4.2$ .



**Fig. 1.** Ratio  $r_B$ .

We skip the proof of the formula for  $r_A(x) \equiv r(x)$  in Problem A, and the properties of this coefficient. Some details can be found in [9] and [16].

*Example 2 (Optimal strategy for  $A(n, k)$  and  $B(n, \lambda)$ ).* In an example with  $n = 5$ ,  $x = 4$ ,  $m = 7$ ,  $a = 7/12$ ,  $b = 9/12$ , for  $B(5, 0.4)$ , we have  $d_B = 1$ ,  $l = 1$ ,  $e = 2$ ,  $m^- = 6$ ,  $l^+ = 1$ ,  $e^+ = 0$ ,  $m^+ = 1$ . Hence the optimal strategy is to place 1 bomb in each of the 2 minus boxes, 2 bombs in each of the remaining 2 minus boxes and 1 bomb goes to the 1 plus box. Thus  $m^- = (x - e)l + e(l + 1) = 2 * 1 + 2 * 2 = 6$ ,  $m^+ = (n - x - e^+)l^+ + e^+(l^+ + 1) = 1 * 1 + 0 = 1$ .

However, for  $A(5, 2)$ , we have  $d_A = 2$ ,  $l = 1$ ,  $e = 3$ ,  $m^- = 7$ ,  $l^+ = e^+ = m^+ = 0$ . Hence our optimal strategy is to put 1 bomb in the 1 minus box, 2 bombs in each of the 3 minus boxes, and no bomb goes in the plus box. Thus  $m^- = (x - e)l + e(l + 1) = 1 * 1 + 3 * 2 = 7$ .

Assuming that the signals for the first 4 boxes are all minus, the signal for the remaining box is plus, and the two locks are allocated among the boxes. The following table illustrates the idea of the bombs placement.

	1	2	3	4	5
$\gamma$ (Lock position)	$\otimes$				$\otimes$
$s$ (Observed signal)	-	-	-	-	+
Bomb placement for $B(5, 0.4)$	2	2	1	1	1
Bomb placement for $A(5, 2)$	2	2	2	1	0

*Example 3 (Value function for  $A(n, k)$  and  $B(n, \lambda)$ ).* Let  $a = 7/12$ ,  $b = 9/12$ , number of bombs  $m = 1, 2, \dots, 7$ , and  $p = 0.6$ .

For  $B(n, \lambda)$ , from Lemma 2, we know that  $p^+$ ,  $p^-$  and ratio  $r_B$  only depend on  $\lambda$ ,  $a$  and  $b$ , while  $d_B$  only depends on  $r_B$  and  $p$ . Hence we calculate these values based on different  $\lambda$  (see Table 1).

Now we can compare the value function for:

1.  $A(2, 1)$  and  $B(2, 0.5)$ .

For  $B(2, 0.5)$ , according to Table 1, for  $\lambda = 0.5$ , we have  $d = 1$ , and  $r_B = 2.143$ . For  $A(2, 1)$ ,  $d$  is changing with respect to  $x$ , and so is  $r(x)$ . The conditional value function  $v(x, m)$  is shown in Table 2. With different number of bombs, the value function is shown in Table 3. We also generate a comparison plot for  $A(2, 1)$  and  $B(2, 0.5)$  with respect to different  $m$ , as shown in Fig. 2.

**Table 1.**  $B(n, \lambda)$ : ratio and  $d$  for different  $\lambda$ .

$\lambda$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$r_B(\lambda)$	1	1.186	1.39	1.615	1.865	2.143	2.455	2.806	3.207	3.667	4.2
$p^-$	1	0.942	0.878	0.808	0.73	0.643	0.545	0.435	0.31	0.167	0
$p^+$	1	0.794	0.632	0.5	0.391	0.3	0.222	0.155	0.097	0.046	0
$d_B$	1	1	1	1	1	1	1	2	2	2	2

**Table 2.** Value function for  $B(2, 0.5)$  and  $A(2, 1)$  when  $m = 7$ .

Bomb placement and value function												
$x$	Problem	$d$	$r$	$l$	$e$	$m^-$	$l^+$	$e^+$	$m^+$	$v(x, m)$	$g(x)$	$v(x, m)g(x)$
0	$B(2, 0.5)$	1	2.143	0	0	0	3	1	7	0.955	0.174	0.166
	$A(2, 1)$	1	None	0	0	0	3	1	7	0.955	0.312	0.298
1	$B(2, 0.5)$	1	2.143	4	0	4	3	0	3	0.907	0.486	0.441
	$A(2, 1)$	2	4.2	4	0	4	3	0	3	0.967	0.542	0.524
2	$B(2, 0.5)$	1	2.143	3	1	7	0	0	0	0.955	0.34	0.325
	$A(2, 1)$	1	None	3	1	7	0	0	0	0.955	0.146	0.139

**Table 3.** The value function for different number of bombs.

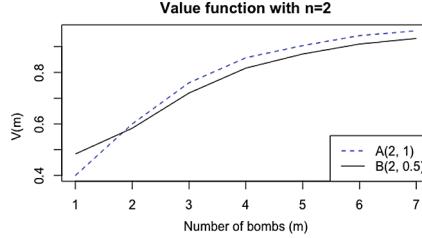
	$m$	1	2	3	4	5	6	7
$B(2, 0.5)$	$v_B(m)$	0.483	0.583	0.72	0.817	0.871	0.91	0.932
$A(2, 1)$	$v_A(m)$	0.4	0.6	0.76	0.857	0.904	0.943	0.962

Thus, when  $m = 7$ , for  $B(2, 0.5)$ , the value function equals

$$v(m) = \sum_{x=0}^2 v(x, m)g(x) = 0.166 + 0.441 + 0.325 = 0.932;$$

for  $A(2, 1)$ , the value function equals

$$v(m) = \sum_{x=0}^2 v(x, m)g(x) = 0.298 + 0.524 + 0.139 = 0.962.$$



**Fig. 2.** The value function for  $A(2, 1)$  and  $B(2, 0.5)$ .

## 2. $A(3, 1)$ and $B(3, 0.33)$ .

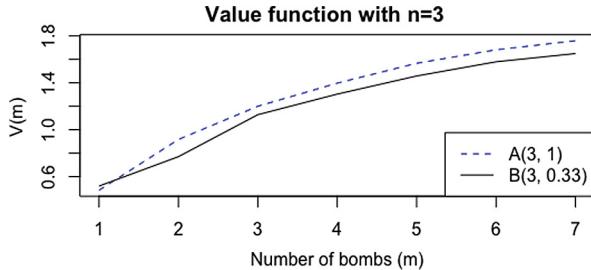
The conditional value function  $v(x, m)$  is shown in Table 4. With different number of bombs, we have the value function shown in Table 5. We also generate a comparison plot for  $A(3, 1)$  and  $B(3, 0.33)$  with respect to different  $m$ , as shown in Fig. 3. The expected damage is relatively higher in  $A(3, 1)$  than in  $B(3, 0.33)$ , except the case when there's only one bomb.

**Table 4.** Value function for  $B(3, 0.33)$  and  $A(3, 1)$  when  $m = 7$ .

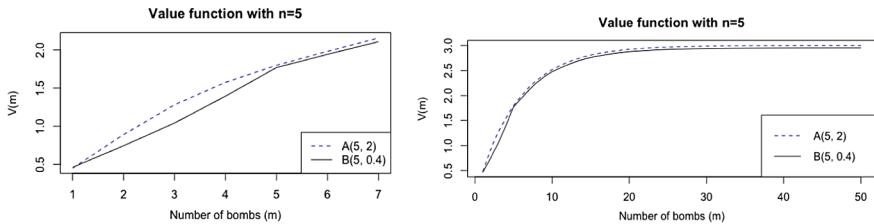
Bomb Placement and Value Function												
$x$	Problem	$d$	$r$	$l$	$e$	$m^-$	$l^+$	$e^+$	$m^+$	$v(x, m)$	$g(x)$	$v(x, m)g(x)$
0	$B(3, 0.33)$	1	1.688	0	0	0	2	1	7	1.753	0.047	0.082
	$A(3, 1)$	1	None	0	0	0	2	1	7	1.744	0.13	0.227
1	$B(3, 0.33)$	1	1.688	3	0	3	2	0	4	1.517	0.249	0.377
	$A(3, 1)$	1	1.615	3	0	3	2	0	4	1.766	0.408	0.72
2	$B(3, 0.33)$	1	1.688	2	1	5	2	0	2	1.785	0.442	0.79
	$A(3, 1)$	2	2.6	3	0	6	1	0	1	1.764	0.377	0.664
3	$B(3, 0.33)$	1	1.688	2	1	7	0	0	0	1.753	0.262	0.459
	$A(3, 1)$	1	None	2	1	7	0	0	0	1.744	0.085	0.148

**Table 5.** Value function for different number of bombs.

	$m$	1	2	3	4	5	6	7
$B(3, 0.33)$	$v(m)$	0.519	0.77	1.128	1.304	1.458	1.579	1.65
$A(3, 1)$	$v(m)$	0.483	0.917	1.2	1.397	1.567	1.681	1.759



**Fig. 3.** Value function for  $A(3, 1)$  and  $B(3, 0.33)$ .



**Fig. 4.** Value function for  $A(5, 2)$  and  $B(5, 0.4)$ .

### 3. $A(5, 2)$ and $B(5, 0.4)$ .

Let's check the value function for  $A(5, 2)$  and  $B(5, 0.4)$  for different  $m$ . The value function  $v(m)$  is shown in Table 6 and the plot of  $v(m)$  is on the left in Fig. 4. We can see that  $A(5, 2)$  has a higher expected damage value than  $B(5, 0.4)$ , but then they are becoming more and more closer as  $m$  gets larger.

The comparison plot of the value function for relatively large  $m$  is shown on the right in Fig. 4. The value function is clearly getting closer as  $m$  gets larger.

**Table 6.** Value function for different number of bombs

	$m$	1	2	3	4	5	6	7
$B(5, 0.4)$	$v(m)$	0.462	0.744	1.042	1.396	1.77	1.943	2.109
$A(5, 2)$	$v(m)$	0.451	0.891	1.282	1.576	1.8	1.982	2.158

Conclusion: According to all three comparison plots, for the same small amount of bombs  $m$ ,  $A(n, k)$  usually has a higher expected damage value than  $B(n, k/n)$ , but when  $m$  is large, the difference becomes smaller and smaller.

## References

1. Gittins, J., Glazebrook, K., Weber, R.: Multi-Armed Bandit Allocation Indices. John Wiley & Sons, Chichester (2011)
2. Gross, O., Wagner, R.: A continuous Colonel Blotto game. RAND Corporation RM-408 (1950)
3. Hart, S.: Discrete colonel blotto and general lotto games. Int. J. Game Theory **36**, 441–460 (2008). <https://doi.org/10.1007/s00182-007-0099-9>
4. Hart, S.: Allocation games with caps: from Captain Lotto to all-pay auctions. Int. J. Game Theory **45**, 37–61 (2016)
5. Hohzaki, R.: Search games: literature and survey. J. Oper. Res. Soc. Jpn. **59**(1), 1–34 (2016)
6. Kvasov, D.: Contests with limited resources. J. Econ. Theory **136**, 738–748 (2007)
7. Lattimore, T., Szepesvari, C.: Bandit Algorithms. Cambridge University Press, Cambridge (2020)
8. Liu, L.: Optimal strategies in “Locks, Bombs, and Testing” (LBT) problem for the case of independent protection. PhD thesis, UNCC (2019). <https://math.uncc.edu/preprint-archive/optimal-strategies-locks-bombs-and-testing-lbt-problem-case-independent-protection>
9. Presman, E.: On one property of an important characteristic in a defense/attack problem. Herald of CEMI **1**(1), 81–85 (2018). <https://cemi.jes.su/s11111110000118-2-1>. (in Russian)
10. Presman, E., Sonin, I.: Sequential Control with Incomplete Information: the Bayesian Approach to Multi-armed Bandit Problems. Academic Press, London (1990)
11. Powell, R.: Defending against terrorist attacks with limited resources. Am. Political Sci. Rev. **101**(3), 527–541 (2007)
12. Robertson, B.: The colonel blotto game. Econ. Theory **29**, 1–24 (2006)
13. Shubik, M., Weber, R.J.: Systems defense games: Colonel Blotto, command and control. Nav. Res. Log. Quart. **28**, 281–287 (1981)
14. Sonin, K., Wright, A.: Rebel capacity and combat tactics. (2018) <http://dx.doi.org/10.2139/ssrn.3030736>
15. Sonin, I.: Bayesian game of locks, bombs and testing. [arXiv:1906.01163](https://arxiv.org/abs/1906.01163) (2019)
16. Sonin, I., Sonin, K.: Bayesian game of locks, bombs and testing. (2019) *Unpublished manuscript.*



# A Survey of Stability Results for Redundancy Systems

Elene Anton<sup>1,3(✉)</sup>, Urtzi Ayesta<sup>1,2,3,4</sup>, Matthieu Jonckheere<sup>5</sup>,  
and Ina Maria Verloop<sup>1,3</sup>

<sup>1</sup> CNRS, IRIT, 2 Rue Charles Camichel, 31071 Toulouse, France  
[{elene.anton,urtzi.ayesta,verloop}@irit.fr](mailto:{elene.anton,urtzi.ayesta,verloop}@irit.fr)

<sup>2</sup> IKERBASQUE - Basque Foundation for Science, 48011 Bilbao, Spain  
<sup>3</sup> Université de Toulouse, INP, 31071 Toulouse, France

<sup>4</sup> UPV/EHU, University of the Basque Country, 20018 Donostia, Spain

<sup>5</sup> Instituto de Cálculo - Conicet, Facultad de Ciencias Exactas y Naturales,  
Universidad de Buenos Aires (1428) Pabellón II, Buenos Aires, Argentina  
[mjonckhe@dm.uba.ar](mailto:mjonckhe@dm.uba.ar)

**Abstract.** Redundancy mechanisms consist in sending several copies of a same job to a subset of servers. It constitutes one of the most promising ways to exploit diversity in multi-servers applications. However, its pros and cons are still not sufficiently understood in the context of realistic models with generic statistical properties of service-times distributions and correlation structures of copies. We aim at giving a survey of recent results concerning the stability - arguably the first benchmark of performance - of systems with cancel-on-completion redundancy. We also point out open questions and conjectures.

**Keywords:** Redundancy · Load balancing · Stability

**AMS(2020) Subject Classification:** Primary 60K25 · Secondary 68M20

## 1 Introduction

While there are several variants of redundancy-based systems, the general notion of redundancy is to dispatch multiple copies of each job to a subset of servers and to consider the result of whichever copy completes service first. By allowing for redundant copies, the aim is to minimize the system latency by exploiting the variability in the queue lengths of the different queues. The potential of redundancy mechanisms lies in finding the right trade-off between exploiting variability and the waste of resources induced by having redundant copies.

Several empirical [2, 3, 11, 12, 38, 41] and numerical studies [15, 16, 26, 29, 30] suggest that redundancy might potentially improve the performance of real-world computer system applications. In particular, Vulimiri et al. [41] consider a 10 DNS servers system and compare the system where each arriving query

dispatches 10 copies to all the 10 DNS servers, to an alternative system where queries are assigned to a single server chosen uniformly at random. The authors observe that the fraction of queries with a service time exceeding 500 ms is reduced by a factor 6.5, and the fraction exceeding 1.5 is reduced by a factor 50. Another interesting study is provided by Dean and Barroso [12] who underline that several redundancy techniques are applied in Google's BigTable in order to improve the latency of incoming queries. They show that a redundancy system with two copies reduces the median response time by 16% and the 99.9th-percentile of the tail of the response time distribution by nearly 40% compared to the non-redundant system.

Broadly speaking, depending on when replicas are deleted, we can consider two classes of redundancy systems: cancel-on-start (*c.o.s.*) and cancel-on-completion (*c.o.c.*). In redundancy systems with *c.o.c.*, once one of the copies has completed service, the other copies are deleted and the job is said to have received service. In redundancy systems with *c.o.s.*, copies are deleted as soon as one copy starts being served, and as a consequence, *c.o.s.* does not waste any computation resources.

In this survey, we will provide an overview on stability results in redundancy systems. From the point of view of stability, *c.o.s.* does not have any negative impact, and for this reason we focus on stability results when *c.o.c.* is implemented.

Let us illustrate through a simple example how redundancy affects the stability region. Consider a system with  $K$  homogeneous servers in which copies of each arriving job are dispatched to  $d \leq K$  servers chosen uniformly at random. We assume that jobs arrive according to a Poisson process of rate  $\lambda$  and jobs have general service times with unit mean. Without redundancy, i.e.  $d = 1$ , the stability condition under any work-conserving policy is given by  $\lambda < \mu K$ , where  $\mu$  is the capacity of the servers. Now, let us assume that the service times are exponentially distributed, that copies are i.i.d. and that  $d = K$ . In this case, the system behaves as a single server system with arrival rate  $\lambda$  and server capacity  $\mu K$ , and the stability condition is again  $\lambda < \mu K$ . However, if all the copies had the same service time as the original job (identical copies), servers are synchronized and the instantaneous departure rate is just  $\mu$ . Therefore, the system behaves as a single server system with arrival rate  $\lambda$  and server capacity  $\mu$ , for which the stability condition is  $\lambda < \mu$ . This simple example illustrates how the modeling assumptions and the degree of redundancy can dramatically impact the stability condition of the system.

One of the main lessons we draw from the results available in the literature, is that the stability region depends strongly on the scheduling policy employed at the servers and the correlation structure of copies. Somewhat surprisingly, we also identify situations for which it was shown that adding redundant copies does not reduce the stability region. Overall, we believe more research is needed in order to design efficient redundancy algorithms.

The rest of the survey is organized as follows. Section 2 describes the main model assumptions and notation, Sect. 3 deals with the case in which the service

times of the copies are *i.i.d.*, and Sect. 4 with identical and correlated copies. In Sect. 5, we present a brief account of results on redundancy that, even though not directly related to stability, are relevant from the performance point of view. We conclude with Sect. 6 where we discuss several open problems and state various conjectures.

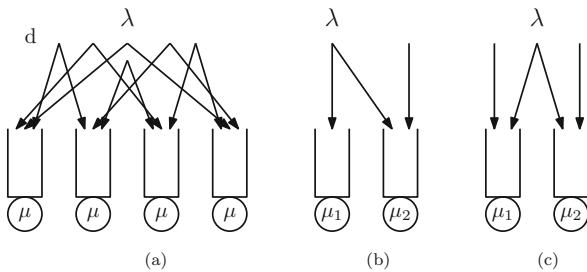
## 2 Model Description and Preliminaries

We consider a  $K$  parallel heterogeneous server system. That is, we have a set of servers  $S = \{1, \dots, K\}$  and server  $s$  has capacity  $\mu_s$ , for  $s \in S$ . Jobs arrive to the system according to a Poisson process of rate  $\lambda$ . Arriving jobs have service times that are independent across jobs and are identically distributed with mean 1.

Jobs are labeled by types  $c = \{s_1, \dots, s_i\} \subset S$ , where  $i$  is the number of copies and  $c$  is the set of servers to which this job will dispatch copies. We let  $\mathcal{C}$  be the set of all possible types. A job is of type  $c$  with probability  $p_c$ , where  $\sum_{c \in \mathcal{C}} p_c = 1$ .

We consider redundancy models that are *c.o.c.*, that is, as soon as a copy is fully served, the additional copies of that job are removed from the system. This cancellation process induces a correlation in the departure process at the servers. Thus, within a server  $s$  there is a departure of a copy due to the following two events: *i*) a local copy departs due to completion in server  $s$ , or *ii*) a copy in another server completes that induces a departure in server  $s$ .

**Model Topology.** A well-known symmetric topology is the one in which each job sends a copy to  $d$  out of  $K$  servers. In case the servers are chosen uniformly at random, that is,  $p_c = 1/\binom{K}{d}$ , and servers have the same capacity  $\mu$ , we refer to this model as the redundancy- $d$  model, see Fig. 1 (a). The number of copies,  $d$ , is referred to as the redundancy degree.



**Fig. 1.** (a) The redundancy- $d$  model for  $K = 4$  and  $d = 2$ . (b) The  $N$ -model. (c) The  $W$ -model.

Two other examples of redundancy topologies are the so-called  $N$ -model and  $W$ -model, see Fig. 1 (b) and (c). Both models are non-symmetric, with two servers. The set of possible job types is  $\mathcal{C} = \{\{2\}, \{1, 2\}\}$  in the  $N$ -model, and  $\mathcal{C} = \{\{1\}, \{2\}, \{1, 2\}\}$  in the  $W$ -model.

When no specific structure is assumed, we refer to it in the sequel as a general topology.

**Scheduling Policy.** A scheduling policy determines how copies are served within each server. As we will see, the choice of the scheduling policy can have a dramatic impact on the stability region. First-Come-First-Served (FCFS) and Processor Sharing (PS) are widely implemented in real-world computer systems [21], and are thus common policies considered in the literature on redundancy. Random-Order-of-Service (ROS) is not a common discipline in systems, but as we will see in the ensuing, it yields very good performance in terms of stability for a redundancy system. These three policies represent the main focus of our survey. To the best of our knowledge, other policies such as Last-Come-First-Served (LCFS), Shortest-Remaining-Processing-Time (SRPT), and Least-Attained-Service (LAS) have not been considered so far.

**Correlation Structure Among Copies.** This describes how the service times of the copies of a given job are related. Formally, the service times  $X_1, \dots, X_k$  of the copies of one job can be sampled from a joint distribution  $F(x_1, \dots, x_k)$ . Two extreme cases are *i.i.d. copies* and *identical copies*. Under i.i.d. copies, all copies have independent service times sampled from the same distribution, whereas with identical copies, all the copies of a job have the same service time. Another interesting framework is the so-called *S&X* model introduced in [16]. Here, the service time of each copy is decomposed into two components; the inherent job size, which is identical for all the copies of a job, and the experienced slowdown on the server it is being served.

**Existing Stability Results.** Table 1 summarizes the main stability results for *c.o.c.* redundancy models available in the literature and discussed in this survey. The table is organized by scheduling policy, service time distribution, redundancy topology and correlation structure. In brackets we specify the additional assumptions that the authors consider in their respective paper. The term “red- $d$ ” refers to the redundancy- $d$  system and the term “gen.” refers to a general redundancy topology.

**Table 1.** Stability results for *c.o.c* redundancy models.

	Service time dist.	i.i.d. copies		identical copies		General correlation	
		red- $d$	gen.	red- $d$	gen.	red- $d$	gen.
FCFS	Exp.	[17]	[8, 20]	[24] (Mean field) [4]			
	General	[35] (Scaled Bernoulli)				[32] (Suf. cond.)	[34] (Comparison result)
PS	Exp.	[4]	X	[4]			
	General			[36]	[5]	[36] (Nec. cond.)	
ROS	Exp.	[4]	X	[4]			X
	General						

The stability condition when jobs have i.i.d. copies is the main topic of Sect. 3, first for exponential service times (Sect. 3.1) and then for scaled Bernoulli distributions (Sect. 3.2). These are the results in the first two columns of Table 1. Correlated copies are discussed in Sect. 4, first for identical copies (Sect. 4.1, middle two columns in Table 1) and then for general correlation structures (Sect. 4.2, last two columns of Table 1). In Sect. 6, we discuss open problems and state various conjectures regarding stability conditions. In Table 1, these conjectures are indicated with a **X**.

### 3 Independent and Identically Distributed Copies

In this section we assume that jobs have i.i.d. copies.

#### 3.1 Exponential Service Times

We first discuss results on FCFS and exponentially distributed service times, a setting studied by Gardner et al. [17, 20] and Bonald and Comte [8]. It was shown in [8] that this model fits the framework of Order Independent queues (see [28, Chapter 2]), which is a large class of systems that have a product-form steady-state distribution. This can be seen as follows. Since copies are i.i.d., we can describe the system through the Markovian state descriptor  $(c_n, c_{n-1}, \dots, c_2, c_1)$ . Here,  $n$  is the number of jobs in the system,  $c_1$  is the type of the eldest job in the system and  $c_i$  is the type of the  $i$ th eldest job. Because of FCFS, the eldest job is served in all of its compatible servers  $c_1$ . The  $i$ -th eldest job is in service at servers  $s \in c_i \setminus \bigcup_{j=1}^{i-1} c_j$ , for  $i = 1, \dots, n$ . Due to the exponentially distributed service times and i.i.d. copies, the instantaneous departure rate of the  $i$ th job is given by the sum of the rates in the servers where the job is in service, that is,  $\sum_{s \in c_i \setminus c_1, \dots, c_{i-1}} \mu_s$ . Hence, the total instantaneous departure rate out of state  $(c_n, c_{n-1}, \dots, c_2, c_1)$  is  $\sum_{s \in \bigcup_{j=1}^n c_j} \mu_s$ , which depends on the set of classes present in the system, but not on their ordering in the state descriptor, i.e., the so-called order independent property.

The characterization of the steady-state distribution facilitates the derivation of performance measures such as the stability condition and mean response times. The proposition below states the stability result for this model.

**Proposition 1** ([8, 20]). *For a redundancy system with general topology under FCFS with exponentially distributed service times and i.i.d. copies, the system is stable if for all  $C \subseteq \mathcal{C}$ ,*

$$\lambda \sum_{c \in C} p_c < \sum_{s \in S(C)} \mu_s, \quad (1)$$

where  $S(C) = \bigcup_{c \in C} \{s \in c\}$ . The system is unstable if there exists  $\tilde{C} \subseteq \mathcal{C}$  such that

$$\lambda \sum_{c \in \tilde{C}} p_c > \sum_{s \in S(\tilde{C})} \mu_s.$$

Informally, Eq. (1) states that the arrival rate to any subset of job types must be less than the total capacity of the associated compatible servers. For exponential service times, this is the *maximum stability* condition, i.e., the system cannot be stable if one of these inequalities were not satisfied. Thus, we conclude that the stability region is not reduced due to adding redundant copies. The latter might seem counter-intuitive at first, since even if servers waste resources serving copies that are not fully served, the stability condition is as large as if there was no redundancy (see also the simple example in the introduction).

Extending Proposition 1 to other scheduling policies is an important open problem (see Sect. 6 for more details). To the best of our knowledge, this has only been achieved for the redundancy- $d$  model. In this case, it is easy to see that Eq. (1) reduces to  $\lambda < \mu K$ , and it has been shown that this stability condition remains valid when either PS or ROS is implemented.

**Proposition 2 ([4]).** *For the redundancy- $d$  model under either PS or ROS with exponentially distributed service times and i.i.d. copies, the system is stable when  $\lambda < K\mu$  and unstable when  $\lambda > K\mu$ .*

Hence, under PS, ROS and FCFS, the redundancy- $d$  model is maximum stable. This however does not hold true in general. In the example below (originally in [4]), we describe priority policy that is not maximum stable, i.e., the system can become unstable even though  $\lambda < K\mu$ .

**Example: Priority Policy.** Consider the redundancy- $d$  system with  $K = 3$ ,  $d = 2$  and  $\mu = 1$ . There are three different types of jobs:  $\mathcal{C} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ . In server 1, FCFS is implemented. In server 2 and server 3, jobs of types  $\{1, 2\}$  and  $\{1, 3\}$  have preemptive priority over jobs of type  $\{2, 3\}$ , respectively. Additionally, within a type, jobs are served in order of arrival.

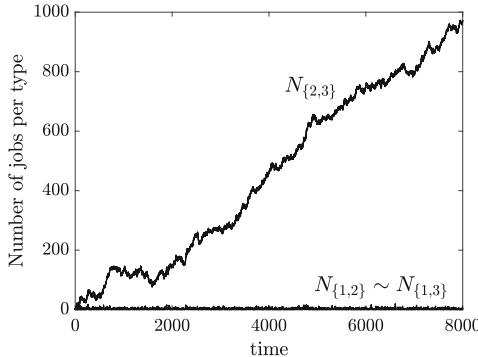
In Fig. 2 we plot the sample-path of the number of jobs when  $\lambda = 2.9 < 3 = \mu K$ . We observe that the number of type- $\{2, 3\}$  jobs in the system grows large, while the number of type- $\{1, 2\}$  and type- $\{1, 3\}$  jobs stay close to 0. Hence, the system is clearly unstable, even though  $\lambda < \mu K$ . This can intuitively be explained by the inefficiency induced by the priority mechanism as the type- $\{2, 3\}$  jobs are preempted by type- $\{1, 2\}$  and type- $\{1, 3\}$  jobs in servers 2 and 3, respectively. We refer to [4] for more details.

### 3.2 General Service Times

To the best of our knowledge, no stability results exist for general service times with i.i.d. copies. In this section, we present the stability result obtained for scaled Bernoulli service times, defined as

$$\begin{cases} X \cdot M, & \text{with probability } 1/M \\ 0, & \text{with probability } 1 - 1/M, \end{cases}$$

where  $M > 0$  and  $X$  is a strictly positive random variable with  $E[X] = 1$ . In this setting, Raaijmakers et al. [35] characterize the stability condition for the redundancy- $d$  model where FCFS is implemented and the number of servers grows large.



**Fig. 2.** The trajectory of the number of jobs per type when  $\lambda = 2.9$ .

**Proposition 3 ([35]).** Consider the redundancy- $d$  model under FCFS with scaled Bernoulli service times and i.i.d. copies. Then,  $\lambda < \frac{M^{d-1}}{E[\min(X_1, \dots, X_d)]}$  is a sufficient stability condition for any  $M$ . In addition, for any  $\epsilon$ , it holds that  $(1 - \epsilon)\lambda < \frac{M^{d-1}}{E[\min(X_1, \dots, X_d)]}$  is a necessary condition, for  $M$  sufficiently large.

We observe that the stability condition is independent of the number of servers, but strongly depends on the number of copies  $d$ . The latter is in contrast to the exponentially distributed service times, where the stability condition does depend on the number of servers but is independent of  $d$  (see Proposition 2). Thus, we observe that when copies are i.i.d., the stability condition strongly depends on the service time distribution. In addition, we observe that as  $M$  grows large (and hence the variance of the service times grows large), the stability region increases by a factor  $M^{d-1}$ , by taking advantage of a greater diversity in service times.

## 4 Correlated Copies

Several studies (e.g., [42]) have shown that the i.i.d. copies assumption can be unrealistic, since large jobs remain large when replicated. Hence, having additional copies could lead to high response times and even instability. Motivated by the latter, stability results with correlated copies have been the focus of recent literature.

### 4.1 Identical Copies

In this section, we assume that jobs have identical copies, i.e., all copies belonging to one job have the same size. This correlation makes that a job can only depart due to its copy that has received most service so far. Thus, the instantaneous departure rate of a job depends on its copy that has currently attained most service.

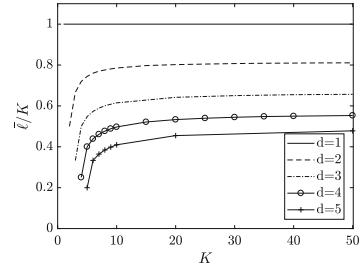
**FCFS Policy.** With FCFS, the eldest job in the system will be served at all of its compatible servers. A job later in the queue will be served at its compatible servers that are not engaged by earlier jobs in the queue.

The stability condition for the redundancy- $d$  system with FCFS and exponentially distributed service times is characterized in Anton et al. [4], through the average departure rate per type in the so-called *saturated system*. The latter assumes an infinite backlog of jobs waiting for service. The long-run time-average number of jobs in service in the saturated system is denoted by  $\bar{\ell}$ . A detailed description of the saturated system and the characterization of  $\bar{\ell}$  can be found in [4].

**Proposition 4 ([4]).** *For the redundancy- $d$  system under FCFS with exponentially distributed service times and identical copies, the system is stable if  $\lambda < \bar{\ell}\mu$  and unstable if  $\lambda > \bar{\ell}\mu$ .*

The value of  $\bar{\ell}$ , and hence the stability region, can be numerically obtained by solving the balance equations of the saturated system, see [4] for more details. We note that the instantaneous departure rate in the saturated system strongly depends on the types in service. As a consequence, no expression has been derived so far for  $\bar{\ell}$  for general  $K$  and  $d$  values.

$\bar{\ell}/K$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$
$d = 1$	1	1	1	1	1	1	1
$d = 2$	0.5	0.66	0.71	0.74	0.76	0.77	0.77
$d = 3$		0.33	0.5	0.54	0.57	0.58	0.60
$d = 4$			0.25	0.4	0.43	0.46	0.47
$d = 5$				0.2	0.33	0.36	0.38
$d = 6$					0.16	0.28	0.31
$d = 7$						0.14	0.25



**Fig. 3.** The table and figure show the values of  $\bar{\ell}/K$  for different values of  $d$  and  $K$ .

Note that the stability condition can equivalently be written as  $\frac{\lambda}{K\mu} < \frac{\bar{\ell}}{K}$ , where  $\frac{\lambda}{K\mu}$  is the traffic load. In Fig. 3 (originally in [4]), we provide numerical values for  $\bar{\ell}/K$ , that is, the traffic supported by the system. The table (*left*) shows  $\bar{\ell}/K$  for small values of  $K$  and the figure (*right*) plots the value of  $\bar{\ell}/K$  as  $K$  grows large. To obtain the value of  $\bar{\ell}$  for  $d \neq 1, K - 2, K - 1, K$ , the authors simulate the saturated system, rather than solving the balance equations<sup>1</sup>. It was proven in [4] that  $\bar{\ell}/K$ , hence the amount of supported traffic, increases

<sup>1</sup> When  $d = K - 1$ , there are  $d$  servers that process copies of one job, and the remaining  $K - d = 1$  server serves one additional job, hence,  $\bar{\ell} = 2$ . When instead  $d = 1$ , there is no redundancy and each server serves one job in the saturated system, i.e.,  $\bar{\ell} = K$ . When  $d = K$ , the system behaves as a single server with capacity  $\mu$ , that is,  $\bar{\ell} = 1$ .

when the number of servers ( $K$ ) grows large, a property that can be observed in Fig. 3.

**PS Policy.** Under PS and identical copies, the stability condition is characterized in [5]. There it is shown that the stability condition coincides with that of a  $K$  parallel server system where each type- $c$  job is only dispatched to its so-called least-loaded servers. In order to state this result, we first need to define several sets of servers and customer types. The first subsystem includes all servers, that is  $S_1 = S$ . We denote by  $\mathcal{L}_1$  the set of least-loaded servers in the system  $S_1 = S$ . Thus,

$$\mathcal{L}_1 = \left\{ s \in S_1 : s = \arg \min_{\bar{s} \in S_1} \left\{ \frac{1}{\mu_{\bar{s}}} \sum_{c \in \mathcal{C}(\bar{s})} p_c \right\} \right\}.$$

For  $i = 2, \dots, K$ , we define recursively

$$\begin{aligned} S_i &:= S \setminus \cup_{l=1}^{i-1} \mathcal{L}_l, \\ \mathcal{C}_i &:= \{c \in \mathcal{C} : c \subset S_i\}, \\ \mathcal{C}_i(s) &:= \mathcal{C}_i \cap \mathcal{C}(s), \\ \mathcal{L}_i &:= \left\{ s \in S_i : s = \arg \min_{\bar{s} \in S_i} \left\{ \frac{1}{\mu_{\bar{s}}} \sum_{c \in \mathcal{C}_i(\bar{s})} p_c \right\} \right\}. \end{aligned}$$

The  $S_i$ -subsystem refers to the system consisting of the servers in  $S_i$ , with only jobs of types in the set  $\mathcal{C}_i$ . The  $\mathcal{C}_i(s)$  is the subset of types that are served in server  $s$  in the  $S_i$ -subsystem. We let  $\mathcal{C}_1 = \mathcal{C}$ . The set  $\mathcal{L}_i$  represents the set of least-loaded servers in the  $S_i$ -subsystem. Finally, we denote by  $i^* := \arg \max_{i=1, \dots, K} \{\mathcal{C}_i : \mathcal{C}_i \neq \emptyset\}$  the last index  $i$  for which the subsystem  $S_i$  is not empty of job types.

The stability condition is now characterized in [5] by the least-loaded servers that can serve each job type.

**Proposition 5 ([5]).** *Assume that the service time distribution is such that it has no atoms and is light-tailed in the following sense,*

$$\lim_{r \rightarrow \infty} \sup_{a \geq 0} \mathbf{E}[(X - a) \mathbf{1}_{\{X-a>r\}} | X > a] = 0. \quad (2)$$

*For a redundancy system with a general topology under PS with identical copies, the system is stable if  $\lambda \sum_{c \in C_i(s)} p_c < \mu_s$ , for all  $s \in \mathcal{L}_i$ ,  $i = 1, \dots, i^*$ . The redundancy system is unstable if there exists  $i \leq i^*$  and  $s \in \mathcal{L}_i$  such that  $\lambda \sum_{c \in C_i(s)} p_c > \mu_s$ .*

It can be seen (as observed in [33]) that the light-tailed condition in (2) also implies

$$\sup_{a \geq 0} \mathbf{E}[(X - a) | X > a] \leq \Phi < \infty, \quad (3)$$

which is a usual light-tailed condition (see [14]). Hence, (2) and (3) exclude heavy tailed distributions like Pareto, but include large sets of distributions such as phase type (which are dense in the set of all distributions on  $\mathbb{R}^+$ ),

exponential and hyper-exponential distributions, as well as distributions with bounded support.

For the redundancy- $d$  model, the above stability condition simplifies into  $\lambda < K\mu/d$ . The latter coincides with the stability condition of a system where all the copies need to be served, that is, the worst possible stability condition.

**ROS Policy.** When ROS is implemented in the servers, it was shown in [4] that the stability condition is not reduced when adding redundant copies. This was proved for exponentially distributed service times and identical copies for the redundancy- $d$  model. However, as stated in Sect. 6, we believe that this holds true for any redundancy structure and any correlation structure.

**Proposition 6 ([4]).** *For the redundancy- $d$  model under ROS with exponentially distributed service times and identical copies, the system is stable if  $\lambda < K\mu$ .*

The intuition behind the above result is as follows. Whenever there are many jobs in a server, the probability that this server serves a copy of a job that has also a copy elsewhere in service will be close to zero. Hence, with a probability close to 1, all highly-loaded servers are serving copies of different jobs and their instantaneous departure rate equals the sum of their capacities.

## 4.2 Generally Correlated Copies

In this section, we consider redundancy models where the service times of the copies of each job are correlated according to some general structure.

For FCFS, Raaijmakers et al. [34] consider a general workload model, which subsumes the S&X model, introduced in [17]. The main difference is that in [34] the server capacities are not fixed, but each job samples server capacities from a discrete and finite distribution. The authors assume that the server speed variations (slowdowns) are either distributed according to New-Better-than-Used (NBU) or New-Worse-than-Used (NWU). See [37] for more details on NBU and NWU distributions<sup>2</sup>.

Depending on the random variation in the server speed, the authors prove that either no replication ( $d = 1$ ) or full replication ( $d = K$ ) provides a larger stability region. Note that here the stability region refers to a wider concept than what we considered before. That is, it refers to the set of arrival rates such that there exists a static assignment rule that makes the system stable.

**Proposition 7 ([34]).** *Consider the following model. Each job is routed to  $d$  servers according to some static probabilistic assignment. Servers implement*

---

<sup>2</sup>  $X$  is said to be *New-Better-than-Used (NBU)* if for all  $t_1, t_2 \in \mathbb{R}$ ,  $\bar{F}_X(t_1 + t_2) \leq \bar{F}_X(t_1)\bar{F}_X(t_2)$ .  $X$  is said to be *New-Worse-than-Used (NWU)* if for all  $t_1, t_2 \in \mathbb{R}$ ,  $\bar{F}_X(t_1 + t_2) \geq \bar{F}_X(t_1)\bar{F}_X(t_2)$ . A sufficient condition for  $X$  to be NBU (NWU) is to have an increasing (a decreasing) hazard rate, i.e.,  $r(x)$  is increasing (decreasing) in  $x$ .

*FCFS. Every time a server starts serving a new copy, it samples a speed variation, which is independent across servers. The type of a job is determined by the capacities it would obtain in each server. A job has a generally distributed service time.*

- If the probabilistic assignment can depend on the job type, and the speed variation follows an NBU distribution, then the stability region for  $d = 1$  is larger or equal than that for  $d > 1$ .
- If the probabilistic assignment does not depend on the job type, and the speed variation follows an NWU distribution, then the stability region for  $d = K$  is larger or equal than that for  $d = 1$ .

From the above we observe that the optimal redundancy degree does not depend on the job size distributions, but rather on the random variation in the server speeds for a given job among the servers.

A sufficient stability condition for the redundancy- $d$  model with FCFS has been obtained in Mendelson [32]. He considers that the service times of the copies  $X_1, \dots, X_d$  are identically distributed with mean 1 and sampled from a joint distribution  $F(x_1, \dots, x_d)$ .

**Proposition 8** ([32]). *Consider the redundancy- $d$  model where FCFS is implemented and the service times of the copies are sampled from a general joint distribution  $F(x_1, \dots, x_d)$ . Then,  $\lambda < \lambda_{lb}$  is a sufficient stability condition, where*

$$\lambda_{lb} := \frac{\mu K}{\sum_{m=0}^d \left( \sum_{j=1}^{d-m} E[\min(X_1, \dots, X_j)] + mE[\min(X_1, \dots, X_d)] \right) P_m},$$

and  $P_m = \binom{K-d}{d-m} \binom{d}{m} / \binom{K}{d}$ .

For the special cases  $d = 1$  and  $d = K$ , the sufficient condition simplifies to  $\lambda < \lambda_{lb} = K\mu$  and  $\lambda < \lambda_{lb} = \mu/E[\min(X_1, \dots, X_d)]$ , respectively, which are in fact also the necessary stability conditions.

We now consider the redundancy- $d$  model where PS is implemented. Raaijmakers et al. [36] characterize the stability condition under any service time distribution through the minimum of the service times of the copies of a job. The latter can be heuristically explained as follows: assume that all servers are equally loaded. Then, due to PS, the copy that completes first is the one with the smallest service time among all copies of the job.

**Proposition 9** ([36]). *For the redundancy- $d$  model under PS where the service times of the copies are sampled from a general joint distribution  $F(x_1, \dots, x_d)$ , a necessary stability condition is  $\lambda d E[\min(X_1, \dots, X_d)] < K\mu$ .*

In the particular case where copies are identical, the authors in [36] prove that Proposition 9 gives a sufficient and necessary stability condition, which is given by  $\lambda d < K\mu$ . We note that the latter coincides with the stability condition for light-tailed service times distributions provided in Proposition 5. Moreover, [36] shows that the stability condition under NWU service time distributions, respectively NBU service time distributions, is larger, respectively smaller, than that for exponential service times.

## 5 Related Work

In this section, we briefly overview relevant papers on redundancy. Even though the results do not deal directly with stability, they are important pointers for the reader who wishes to work on redundancy.

### 5.1 Response Time

The response time (a.k.a. delay) measures the time elapsed between arrival and departure. It is together with stability the main performance measure, and it has received considerable attention. The first performance analysis of a redundancy model was for *cancel-on-complete* (*c.o.c.*) with exponentially distributed service times, independent and identically distributed (i.i.d.) copies and FCFS. As discussed in Sect. 3.1, Gardner et al. [17, 20] and Bonald and Comte [8] exploit the link between this redundancy system and the Order Independent queue [28], in order to show that the steady-state distribution has a product form. The paper [17] showed that the mean response time in the system reduces as the redundancy degree  $d$  increases. Redundancy *c.o.s.* with FCFS and exponentially distributed job sizes has been analyzed in Ayesta et al. [7], where it was shown that the steady-state distribution also has a product form. This was achieved by showing that this model fits within the framework of multi-type jobs and multi-type servers studied in Visschers et al. [40]. The above results have motivated researchers to develop unifying frameworks to explain the emergence of product form distributions in redundancy models. This is done in Ayesta et al. [6] and Gardner and Righter [19] by extending the frameworks of Visschers et al. [40] and Order Independent queues [28], respectively.

Comte and Dorsman [10] introduce the Pass-and-Swap queue, not included in the above unifying frameworks, but for which the product-form of the steady-state distribution is preserved. The authors provide several examples that fall into this framework, including a loss variant of the *c.o.s.* redundancy model.

The response time has also been studied in limiting regimes such as heavy traffic and mean field. Cardinaels et al. [9] consider both *c.o.c.* and *c.o.s.* and establish that in heavy traffic the joint distribution of the number of jobs of the various types converges to the product of an exponentially distributed random variable times a deterministic vector, a phenomenon known as state-space collapse. Hellemans et al. [24, 25] consider the mean-field regime and characterize the stationary workload distribution of *c.o.c.* with FCFS, general service times and both identical and i.i.d. copies. In Hellemans et al. [22] the authors generalize the previous result to other redundancy scheduling implementations such as replication if above certain threshold, delayed replication policy or replicate small jobs. Another mean-field result can be found in Hellemans et al. [23] where the authors analyze the stationary response time and workload distributions of JSW( $d$ ), JSQ( $d$ ) and redundancy- $d$  under FCFS and general service times.

## 5.2 Optimizing Redundancy

The stability results presented in this survey show that both the scheduling policy and the degree of redundancy can have a big impact on the stability region and hence on the performance of the system. Motivated by this, researchers have aimed at *i*) characterizing what is the optimal scheduling policy in the servers and *ii*) determining what is the optimum number of copies that should be created.

One of the first papers studying redundancy was by Koole and Righter [27], which considered a system where jobs can dispatch i.i.d. copies to any subset of servers in the system. The authors showed that with FCFS and NWU service time distributions, the best policy is to replicate to all the servers.

Several optimality results have been derived for the Least-Redundant-First (LRF) scheduling policy, which serves jobs in lower priority as their number of copies increases (jobs with the same number of copies are served according to FCFS). In particular, Gardner et al. [15, 18] consider nested redundancy models with exponential service times and i.i.d. copies, and show that the mean response time is minimized under LRF. We note that a redundancy model is nested if for all  $c, c' \in \mathcal{C}$ , either *i*)  $c \subset c'$  or *ii*)  $c' \subset c$  or *iii*)  $c \cap c' = \emptyset$ .

Akgun et al. [1] consider a two-server system in which each server has dedicated traffic, that is, each server is a unique compatible server for one job type. The authors consider the DCF (Dedicated-Customers-First) scheduling policy and analyze the efficiency and fairness for both dedicated and redundant jobs.

Sun et al. [39] consider various low-complexity redundancy scheduling techniques for systems where jobs have i.i.d. copies, and investigate when these are delay-optimal (or nearly-delay optimal) with respect to the stochastic ordering. These new scheduling techniques are based on job replication and job cancellation decision features. For instance, the authors show that the *fewest unassigned task first with low-priority replication* and *earliest due date first with replication* policies are nearly delay-optimal with NBU and NWU distributions, respectively.

## 5.3 Related Models

Redundancy as considered in this chapter is closely related to the  $(n, k)$  fork-join system. In the latter, there exist  $n$  servers each one receiving one of the blocks, and the job is completed once  $k < n$  blocks are served. If  $k = 1$ , this model becomes equivalent to the redundancy- $n$  model with *c.o.c.*

For the  $(n, k)$  fork-join model, Lee et al. [30] provide sequences of systems that upper and lower bound the original one, and that converge to the original system. Through these bounds, the authors characterize the mean response time of the system. Li et al. [31] derive that in the mean-field regime, coding always improves the mean response time compared to the redundancy model, i.e.,  $(n, 1)$ .

In [26], the authors consider the  $(n, r, k)$  partial fork-join system, where the job is sent to  $r$  out of  $n$  servers uniformly chosen at random and waits for the first  $k \leq r$  to complete. They study effective replication strategies for various scenarios. The authors show that both latency and cost are minimized when  $r$

increases for log-convex (high variable) service time distributions. Duffy et al. [13] compare the tail response time of the  $(n, r, k)$  model to that of the redundancy- $d$  model (with batch arrivals of size  $r$ ). The authors show that the tail distribution of the response time under  $(n, r, k)$  partial fork-join is smaller than under the redundancy- $d$  model as long as  $r - k \geq d$ , as the number of servers tend to infinity.

In a recent paper, Zubeldia [43] considers the  $S\&X$  model where the slowdown experienced by each copy in service is independent across servers, but not necessarily independent from the job's service time. The author provides a lower-bound on the mean delay for the  $(n, r, k)$  partial fork-join system, and shows that when slowdowns are exponentially distributed and independent of the service time of the job, the expected delay is minimized in the mean-field limit for a constant  $r$  that only depends on the arrival rate and mean slowdown.

## 6 Conclusions, Open Problems and Conjectures

The literature on the stability analysis of redundancy systems is recent and growing. However, there are many important cases that have not been analyzed yet. In this section, we address some of the open problems related to stability, and state several conjectures that are based on our intuitive understanding of the system. It is our hope that this survey might encourage more research on this relevant and timely topic.

### 6.1 I.i.d. Copies.

As shown in Proposition 1, FCFS is maximum stable with exponential service times and i.i.d. copies. We believe that this result should remain valid for any work-conserving scheduling policy with non-preferential treatment across types, for instance PS, ROS, LCFS, LAS and SRPT. The reason for this is that the i.i.d assumption combined with the non-preferential treatment across types permits to take advantage of diversity when the system is close to saturation.

**Conjecture 1.** Consider a redundancy system with a general topology with exponentially distributed service times and i.i.d. copies. For any work-conserving non-preferential scheduling policy, the system is stable if for all  $C \subseteq \mathcal{C}$ ,

$$\lambda \sum_{c \in C} p_c < \sum_{s \in S(C)} \mu_s,$$

where  $S(C) = \bigcup_{c \in C} \{s \in c\}$ .

**Open Problem 1.** If we relax the exponential service times to general service time distribution, the stability condition is unknown.

## 6.2 FCFS Scheduling Policy with Identical Copies

In Sect. 4.1, we saw that  $\lambda/\mu K < \bar{\ell}/K$  is the stability condition of the redundancy- $d$  system where jobs have identical copies and exponential service times.

**Open Problem 2.** If we relax the redundancy- $d$  structure to general topologies, or the exponential service times to general service times, the stability condition is unknown.

For exponential service times with the redundancy- $d$  structure, we observed in Fig. 3 that for a given number of copies  $d$ ,  $\lim_{K \rightarrow \infty} \bar{\ell}/K < 1$ . Note that  $\lambda/\mu K < 1$  is the stability condition for a system with no redundancy. Hence, if it can be proved that  $\lim_{K \rightarrow \infty} \bar{\ell}/K < 1$ , this would imply that as the number of servers grows large, the traffic load that a redundancy system can support is smaller than if no redundancy was implemented.

**Conjecture 2.** Consider the redundancy- $d$  model where FCFS is implemented and jobs have exponentially distributed service times and identical copies. Then, for fixed  $d$ ,  $\lim_{K \rightarrow \infty} \bar{\ell}/K < 1$ .

The limit should coincide with the stability condition given in [24], where the authors develop a numerical method to derive the stability condition in the mean-field limit.

We also observed the following monotonicity property in the number of redundant copies. More precisely, we conjecture that as the degree of redundancy increases, the stability region becomes smaller.

**Conjecture 3.** Consider the redundancy- $d$  model where FCFS is implemented and jobs have exponentially distributed service times and identical copies. Then, for fixed  $K$ ,  $\bar{\ell}$  is decreasing in  $d$ , and hence, the stability region is decreasing in  $d$ .

## 6.3 ROS Scheduling Policy with Generic Correlation Structure

In the particular case of ROS, we believe that Conjecture 1 will remain valid even if copies follow a general correlation structure, including identical copies. So far, this was only proved for the redundancy- $d$  model with exponential distributed service times with identical copies, see Proposition 6.

**Conjecture 4.** Consider a redundancy system with a general topology with exponentially distributed service times and an arbitrary correlation structure among copies. ROS is stable if for all  $C \subseteq \mathcal{C}$ ,

$$\lambda \sum_{c \in C} p_c < \sum_{s \in S(C)} \mu_s,$$

where  $S(C) = \bigcup_{c \in C} \{s \in c\}$ .

The intuition would be the following. In principle, multiple copies of the same job could be served simultaneously at various of its compatible servers. Due to the heterogeneous capacities and the correlation among the copies, the departure rate of that job depends on the residual service time of each copy. However, when the number of jobs in the system grows large, the probability that more than one copy of the same job is simultaneously in service goes to zero. Using fluid-limit techniques, as done in [4], one then obtains that the fluid limit of the system equals that of the system where jobs have i.i.d. copies. Hence, if Conjecture 1 is valid, this would imply that Conjecture 4 is true as well.

## 6.4 Redundancy-Aware Scheduling

Another interesting, and so far unexplored area, is the impact of redundancy-aware scheduling policies on the stability region and the performance of the system. By redundancy-aware we refer to policies like LRF or Most-Redundant-First that can use information on the number of copies when choosing which copy to serve in a server. As discussed in Sect. 5.2, the authors of [15, 18] consider the nested model with exponentially distributed service times and i.i.d. copies and show that LRF minimizes the mean response time. It would be interesting to explore this further for more general redundancy settings.

**Acknowledgement.** Research of E. Anton supported and research of M. Jonckheere partially supported by the French “Agence Nationale de la Recherche (ANR)” through the project ANR-15-CE25-0004 (ANR JCJC RACON). U. Ayesta has received funding from the Department of Education of the Basque Government through the Consolidated Research Group MATHMODE (IT1294-19).

## References

1. Akgun, O., Righter, R., Wolff, R.: Partial flexibility in routing and scheduling. *Adv. Appl. Probab.* **45**(3), 673–691 (2013)
2. Ananthanarayanan, G., Ghodsi, A., Shenker, S., Stoica, I.: Why let resources idle? Aggressive cloning of jobs with dolly. In: Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Computing, HotCloud’ 12, Article 17, p. 6 (2012)
3. Ananthanarayanan, G., Ghodsi, A., Shenker, S., Stoica, I.: Effective straggler mitigation: attack of the clones. In: Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation vol. 13, pp. 185–198 (2013)
4. Anton, E., Ayesta, U., Jonckheere, M., Verloop, I.M.: On the stability of redundancy models. *Oper. Res.* (2021). <https://doi.org/10.1287/opre.2020.2030>
5. Anton, E., Ayesta, U., Jonckheere, M., Verloop, I.M.: Improving the performance of heterogeneous data centers through redundancy. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems – SIGMETRICS 4(3), Article 48, p. 29 (2020)
6. Ayesta, U., Bodas, T., Dorsman, J., Verloop, I.M.: A token-based central queue with order-independent service rates. *Oper. Res., to appear* (2021)
7. Ayesta, U., Bodas, T., Verloop, I.M.: On a unifying product form framework for redundancy models. *Perform. Eval.* **127–128**, 93–119 (2018)

8. Bonald, T., Comte, C.: Balanced fair resource sharing in computer clusters. Perform. Eval. **116**, 70–83 (2017)
9. Cardinaels, E., Borst, S.C., van Leeuwaarden, J.S.H.: Redundancy scheduling with locally stable compatibility graphs. arXiv:2005.14566 (2020)
10. Comte, C., Dorsman, J.: Pass-and-swap queues. arXiv:2009.12299 (2020)
11. Dean, J.: Achieving rapid response times in large online services. Google Research (2012). <http://research.google.com/people/jeff/latency.html>
12. Dean, J., Barroso, L.A.: The tail at scale. Commun. ACM **56**, 74–80 (2013)
13. Duffy, K.R., Shneer, S.: MDS coding is better than replication for job completion times. arXiv:1907.11052 (2019)
14. Foss, S., Korshunov, D., Zachary, S.: An Introduction to Heavy-Tailed and Subexponential Distributions. Springer, NY (2013)
15. Gardner, K., Harchol-Balter, M., Hyytia, E., Righter, R.: Scheduling for efficiency and fairness in systems with redundancy. Perform. Eval. **116**, 1–25 (2017)
16. Gardner, K., Harchol-Balter, M., Scheller-Wolf, A., van Houdt, B.: A better model for job redundancy: decoupling server slowdown and job size. IEEE ACM Trans. Netw. **25**(6), 3353–3367 (2017)
17. Gardner, K., Harchol-Balter, M., Scheller-Wolf, A., Velednitsky, M., Zbarsky, S.: Redundancy-d: the power of d choices for redundancy. Oper. Res. **65**, 1078–1094 (2017)
18. Gardner, K., Hyytiä, E., Righter, R.: A little redundancy goes a long way: convexity in redundancy systems. Perform. Eval. **131**, 22–42 (2019)
19. Gardner, K., Righter, R.: Product forms for FCFS queueing models with arbitrary server-job compatibilities: an overview. Queueing Syst. **96**(1), 3–51 (2020)
20. Gardner, K., Zbarsky, S., Doroudi, S., Harchol-Balter, M., Hyytiä, E., Scheller-Wolf, A.: Queueing with redundant requests: exact analysis. Queueing Syst. **83**(3–4), 227–259 (2016)
21. Harchol-Balter, M.: Performance Modeling and Design of Computer Systems: Queueing Theory in Action. Cambridge University Press, NY (2013)
22. Hellemans, T., Bodas, T., van Houdt, B.: Performance analysis of workload dependent load balancing policies. In: International Conference on Measurement and Modeling of Computer Systems vol. 3(2), Article 35, p. 35 (2019)
23. Hellemans, T., van Houdt, B.: On the Power-of-d-choices with least loaded server selection. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems – SIGMETRICS vol. 2(2), Article 27, p. 22 (2018)
24. Hellemans, T., van Houdt, B.: Analysis of redundancy(d) with identical replicas. ACM Sigmetrics Perform. Eval. Rev. **46**(3), 74–79 (2018)
25. Hellemans, T., van Houdt, B.: Performance of redundancy(d) with identical/independent replicas. In: ACM Transaction on Modeling and Performance Evaluation of Computing Systems (TOMPECS), vol. 4(2), Article 9, p. 28 (2019)
26. Joshi, G., Soljanin, E., Wornell, G.: Efficient redundancy techniques for latency reduction in cloud systems. In: ACM Transaction on Modeling and Performance Evaluation of Computing Systems (TOMPECS), vol. 2(2), Article 12, p. 30 (2017)
27. Koole, G., Righter, R.: Resource allocation in grid computing. J. Sched. **11**, 163–173 (2007)
28. Krzesinski, A.E.: Order independent queues. In: Boucherie, R.J., van Dijk, N.M. (eds.) Queueing Networks: a Fundamental Approach, pp. 85–120. Springer, Boston, MA (2011)
29. Lee, K., Shah, N.B., Huang, L., Ramchandran, K.: When do redundant requests reduce latency? IEEE Trans. Commun. **64**(2), 715–722 (2016)

30. Lee, K., Shah, N.B., Huang, L., Ramchandran, K.: The mds queue: analysing the latency performance of erasure codes. *IEEE Trans. Inf. Theory* **63**(5), 2822–2842 (2017)
31. Li, B., Ramamoorthy, A., Srikant, R.: Mean-field-analysis of coding versus replication in cloud storage systems. In: *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9 (2016)
32. Mendelson, G.: A lower bound on the stability region of redundancy-d with FIFO service discipline. *Oper. Res. Lett.* **49**(1), 113–120 (2021)
33. Paganini, F., Tang, A., Ferragut, A., Andrew, L.: Network stability under alpha fair bandwidth allocation with general file size distribution. *IEEE Trans. Automat. Contr.* **57**, 579–591 (2012)
34. Raaijmakers, Y., Borst, S.C.: Achievable stability in redundancy systems. In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems – SIGMETRICS*, vol. 4(3), Article 46, p. 21 (2020)
35. Raaijmakers, Y., Borst, S.C., Boxma, O.: Redundancy scheduling with scaled Bernoulli service requirements. *Queueing Syst.* **93**, 67–82 (2019)
36. Raaijmakers, Y., Borst, S.C., Boxma, O.: Stability of redundancy systems with processor sharing. In: *Proceedings of the 13th EAI International Conference on Performance Evaluation Methodologies and Tools, Valuetools 20*, pp. 120–127 (2020)
37. Ross, S.M.: *Stochastic Processes*. Wiley & Sons, NY (1996)
38. Sieber, C., Blenk, A., Hinteregger, M., Kellerer, W.: The cost of aggressive http adaptive streaming: quantifying youtube’s redundant traffic. In: *2015 IFIP/IEEE Intern. Symp. on Integrated Network Management (IM)*, pp. 1261–1267 (2015)
39. Sun, Y., Koksal, C.E., Shroff, N.B.: On delay-optimal scheduling in queueing systems with replications. [arXiv:1603.07322](https://arxiv.org/abs/1603.07322) (2016)
40. Visschers, J., Adan, I., Weiss, G.: A product form solution to a system with multi-type jobs and multi-type servers. *Queueing Syst.* **70**, 269–298 (2012)
41. Vulimiri, A., Godfrey, P.B., Mittal, R., Sherry, J., Ratnasamy, S., Shenker, S.: Low latency via redundancy. In: *Proceedings of the ACM Conference on Emerging Networking Experiments and Technologies*, pp. 283–294 (2013)
42. Vulimiri, A., Michel, O., Godfrey, P.V., Shenker, S.: More is less: reducing latency via redundancy. In: *Proceedings of the 11th ACM Workshop on Hot Topics in Networks, HotNets’11*, vol. 11, pp. 13–18 (2012)
43. Zubeldia, M.: Delay-optimal policies in partial fork-join systems with redundancy and random slowdowns. In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems – SIGMETRICS*, vol. 4(1), Article 2, p. 49 (2020)



# IBM Crew Pairing and Rostering Optimization (C-PRO) Technology with MDP for Optimization Flow Orchestration

Vladimir Lipets and Alexander Zadorojniy<sup>(✉)</sup>

IBM Research - Haifa, Haifa, Israel  
[{lipets,zalex}@il.ibm.com](mailto:{lipets,zalex}@il.ibm.com)

**Abstract.** We created the IBM Crew Pairing and Rostering Optimization (C-PRO) solution for air crew scheduling. It was deployed at El Al in 2013 and at Aeroflot in 2020. The core of the system is an optimization flow, which models the problem using mixed integer linear programming (MILP) with millions of integer variables. The solution is derived iteratively using heuristics. Most recently, we applied Markov Decision Process (MDP) in place of the heuristics orchestrator and realized a 30% improvement in performance.

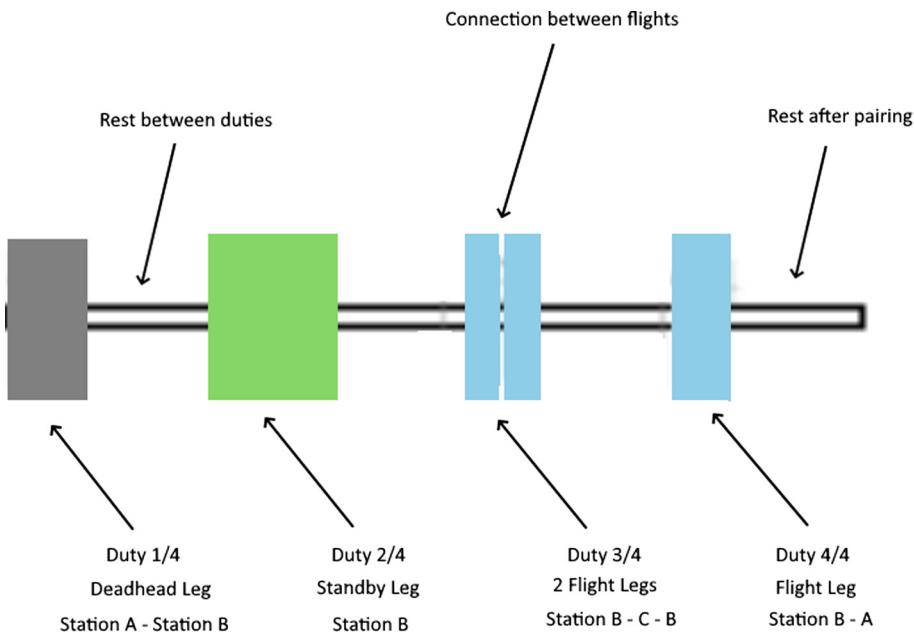
**Keywords:** Markov decision process (MDP) · Mixed integer linear programming (MILP) · Scheduling

**AMS(2020) Subject Classification:** Primary 90C40 · Secondary 90C11

## 1 Introduction

### 1.1 Airline Crew Pairing and Rostering Problem

Assigning airline crews to flights – what is commonly referred as pairing and rostering – is an extremely complex problem that is also very well-studied [2, 9, 10]. A pairing is a sequence of flight legs that start and end at the same location where the crew members live (Fig. 1). It typically spans between one and five days; however, in some cases it can be more than one week in duration. To create assignments for crew members, airline planners start by generating pairings to cover as many flight legs as possible, with a cost as low as possible. For each pairing, they specify which types of crew members (e.g., captains, first officers, flight attendants, etc.) are required and at what quantity, which is known as a “crew complement” (Fig. 2). Once optimal pairings and their crew complements are defined, the next phase is to assign crew members to the pairings, while maintaining compliance with a variety of work regulations

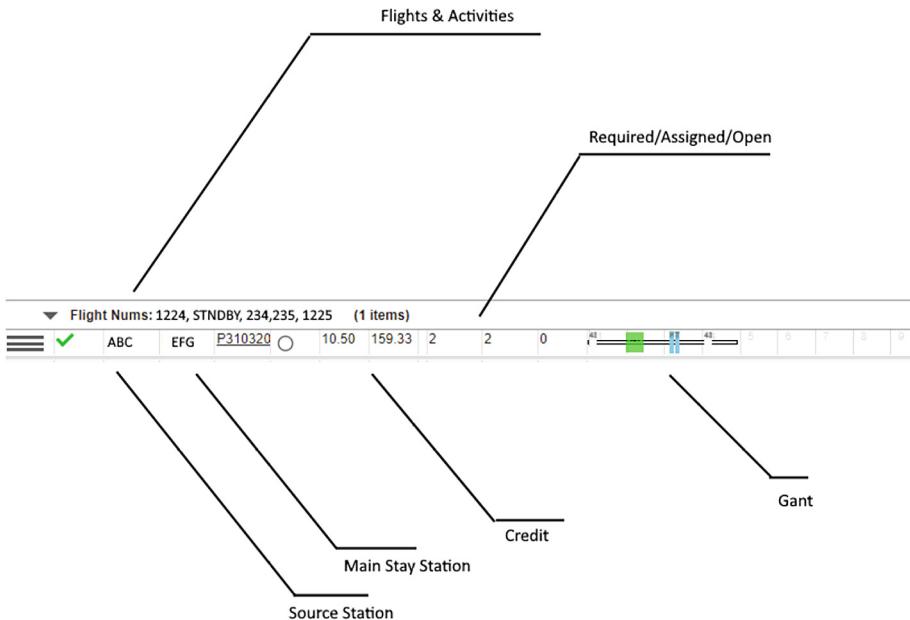


**Fig. 1.** Pairing Diagram. A ‘duty’ consists of four duties. In the ‘Deadhead Leg’, crew members fly from their point of origin as regular passengers to the actual start of their duty. In the “Standby Leg,” the crew members are in standby mode. Next, in the ‘2 Flight Legs’ duty, the crew members fly two flights with a short connection between them. Finally, in the ‘Flight Leg’ duty, the crew arrives back to their point of origin.

and collective agreements. This part of the problem is called crew assignment. To solve this problem properly, pairings, rest periods, training periods, annual leaves, and so forth, must be taken into account to create working schedules (rosters) for crew members (Fig. 3).

Technology that creates an optimized schedule for flight crews can provide significant savings to airlines. The benefits go beyond cost savings. An equitable, well-planned and efficient crew roster contributes to flight safety and employee satisfaction. The challenge is that mainstream airline rostering solutions typically take days or weeks to generate a single plan for pairing and rostering, which is typically generated ahead of every month. They also require optimization experts to implement any changes to the optimization logic. Researchers at IBM developed a solution that delivers optimized pairing and rostering (product named, ‘C-PRO’) with unparalleled speed and flexibility. Planners can create assignments and implement many types of changes without relying on optimization experts. It is also faster: IBM’s C-PRO can execute multiple iterations in a single day. This makes it easier to create the best option by enabling customers to rapidly improve in quick increments. They can also support changing business requirements, such as complying with new regulations, and meeting spe-

cial requests from crew personnel. The optimized crew scheduling also supports “what if” analysis to predict the impact of various changes; for example, working hours, vacations, etc. It also provides intuitive explanations, which are presented in terms of business objects and logic that ordinary users can understand.



**Fig. 2.** Pairing Diagram from C-PRO. The three fields ‘Required/Assigned/Open’ show the values of ‘2/2/0’. This means crew complement requires two crew members, two crew members have been assigned by the optimization engine already, and there are zero open positions. Also shown here: 10.50 h is a flight duration, 159.33 is the total pairing time in hours, P310320 is the pairing id, and ABC is the starting point.

## 1.2 Challenges We Faced Working on C-PRO

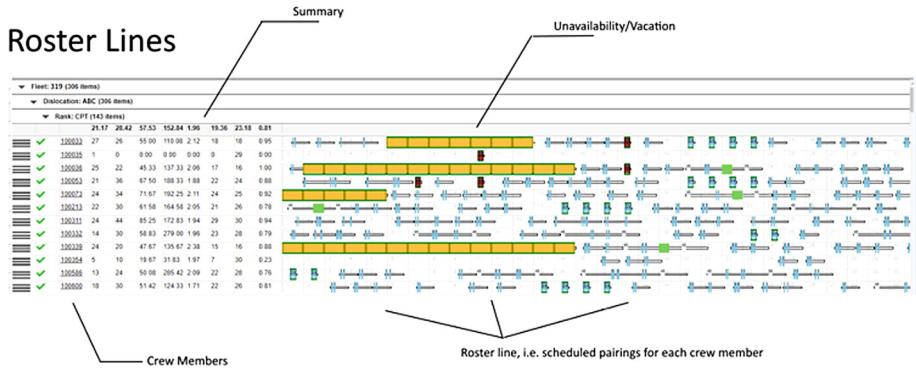
### 1. Extremely large size of the problem.

Crew scheduling is a very large problem in many aspects:

- There are hundreds of domain-related objects (e.g., stations, regions, legs, shifts, etc.) that needed to be described mathematically.
- There are dozens of different types of requirements that must be described mathematically and satisfied.
- MILP formulation of the problem consist of many millions of integer variables and constraints; and therefore, cannot be solved ‘as is’ by available solvers.

### 2. Rapid turnaround of requests for update by clients.

Customers want to make changes quickly, but in existing solutions, modifying rules and constraints typically can take weeks or even months. Optimization



**Fig. 3.** Rostering from C-PRO. This figure shows the rostering for each crew member by crew member ID. The summary fields represent (in order from left to right) number of assignments, number of flights, flight time (credit), staying abroad time, average flights durations performed, number of flight days, available days to work, and utilization per scheduling period.

experts need to formulate new business rules as new constraints and incorporate the new model into the application. For the customer, this means waiting until the next update to obtain the desired modification.

### 3. Explainability.

In the Enterprise Optimization domain, in order to be useful in real world applications, optimization solution must be interpretable. The system which cannot ‘explain’ to the end-user (who are domain experts) how tradeoffs were made will not be able to earn trust and will not win adoption by customers.

### 4. Multi-problem solution capabilities.

Crew pairing and rostering is not a single type of problem. To address effectively, several types of problems need to be considered, including:

- Coverage problem, which finds optimized set of pairings for covering all activities.
- Assignment problem, which assigns crew members to the pairings.
- Routing problem, which finds optimal route for the airplanes, taking into account their location and required maintenance.
- Shift scheduling problem, which schedules crew member to weekly shifts.
- Personal requests biddings, which optimizes assignment of crew members according to personal preferences resolving conflicts in an optimized manner.

### 5. Reusability.

Development of enterprise level optimization solution is a significant investment. Reusability of the solution for other clients in the same domain (or clients in adjacent domains) is vital for the market success of the solution.

## 2 MILP Problem Formulation

While creating pairings and assignments, we require that

- the cost of pairings should be kept in the minimum,
- pairings must be legal according to specified regulations,
- all flights and other activities must be assigned exactly once.

This optimization problem is modeled as a MILP problem where pairings are the variables of the problem, which are either assigned zero (not selected) or one (selected). Costs of the pairings are the coefficients of the variables. Covering the flights exactly once is formulated as a constraint of the optimization. Therefore,

$$\begin{aligned} \min & \sum_{j \in P} c_j \cdot x_j \\ \text{s.t.} \\ x_j & \in \{0, 1\}, \forall j \in P \\ \sum_{j \in PF_i} x_j & = 1, \forall i \in F \end{aligned}$$

where  $P$  is the set of all possible pairings,  $F$  is the set of all flights.  $\forall i \in F$ ,  $PF_i$  denotes set of pairings containing flight  $i \in F$ .  $c_j$  is the cost of the pairing  $j \in P$ . Notice that when  $x_j$  is assigned to one, it means that pairing  $j$  was selected. The objective is to minimize the total cost of selected pairings. The constraints guarantee that each flight is covered only once. Although this is a ‘vanilla’ problem formulation, the problem is complicated. There is a significant challenge to effectively build the set  $P$  of legal pairings. As the number of flights increases, the number of potential pairings grows exponentially. To deal with this, we apply different graph theory-based algorithms in the earlier stages to remove candidates that have a small chance to be selected. Also, each pairing can consist of a set of duties, each duty may consist of flights, and each flight may appear only in one duty. Moreover, the crew (captain, first officer, etc.) required for each duty may change according to different properties of the duty such as flight duration, period of the day, airplane type, etc. For instance, duties with durations more than 12 h that start in the morning may require two captains and one first officer; whereas, similar duties at night require two captains and two first officers. The formulation in this case is as follows:

$$\begin{aligned} \min & \sum_{j \in P, r \in R} c_{j,r} \cdot z_{j,r} \\ \text{s.t.} \\ z_{j,r} & \in \mathbb{Z}^+, \forall j \in P, \forall r \in R \\ y_k & \in \{0, 1\}, \forall k \in D \\ \sum_{k \in D_i} y_k & = 1, \forall i \in F \\ \sum_{j \in PD_k} z_{j,r} & = q_{k,r} \cdot y_k, \forall k \in D, \forall r \in R \end{aligned}$$

where  $R$  is the set of all ranks,  $D$  is the set of all duties,  $D_i$  denotes set of duties containing flight  $i \in F$ ,  $PD_k$  denotes set of pairings containing duty  $k \in D$ ,  $c(j, r)$  is the cost of the pairing  $j \in P$  for rank  $r$ ,  $q(k, r)$ , is the number (constant) of crew members of rank  $r$  required for duty  $k \in D$ . Notice that when  $y_k$  is assigned to one, it means that duty  $k$  was selected, and  $z_{j,r}$  obtains a positive value, which refers to the number of crew members of rank  $r$  scheduled to pairing  $j$ . The objective is to minimize the total cost of the selected pairings with respect to the ranks. The constraints guarantee that each flight is covered exactly once by the duty and that each selected duty is covered by the pairings according to the duty and rank requirements. Extended business rules may be applied on top of these rules, which for example, may require certain number of pairings with given properties. For instance, assuming property ‘start on base B’ we may require a proper balance between staffing level on base B and the number of created pairings starting from this base. Moreover, if we introduce the conception of acclimatization, which means that  $q_{k,r}$  will depend not only on duty  $k$ , but also will be a function of the time passed after the previous duty of the pairing.

After determining pairings, the next phase is to assign crew members to execute these pairing for a given time period, while complying with a variety of work regulations and collective agreements. This is the crew assignment problem, where pairings, rest periods, training periods, annual leaves, etc., are combined to form working schedules for crew members. The classical MILP problem formulation for solving assignment problem is based on the assignment of  $a_{e,p}$  decision variable for one or zero, if employee  $e$  is assigned to pairing  $p$  or not, respectively. The problem becomes more complicated when we need to satisfy more advanced rules and constraints. For example, decision variables  $a_{e,p}$  are replaced by variables  $a_{e,p,r,k}$ , where additional parameters of rank  $r$  and role  $k$  of the assignment are introduced. In another example, we have a rule defining duration of non-working period within a floating time window, or set of rules asserting that if some crew member did pairing of type A in the middle of the month, then they have to execute pairings of type C within the last day of the same month. This problem is formulated as MILP. For the sake of compactness, we skip its detailed formulation.

### 3 The Solution

In this paper we mostly address Challenge number 1. The core of the system is a complex optimization flow which models a problem as a largescale mixed integer linear programming problem (MILP) with millions of integer variables and solves it using different type of algorithms. No MILP solver on the market can solve a problem of this size in a reasonable time. To create a solution, we used an approach [5] which allows incorporation of multiple heuristics, including ‘business heuristics’ and ‘business decompositions’. The uniqueness of this approach, as opposed to the well-known column generation, is that ‘business decompositions’ take into account a business characteristic of a client objects while the column

generation would not. ‘Business heuristics’ leverage the structure of input data to build the schedule in a more efficient way.

The novelty in our approach is applying machine learning as an automated tool to continuously seek the best problem decomposition and modification strategy. Each of the heuristics in the flow is important for the solve to succeed overall. It starts from a heuristic that finds an initial feasible solution. Next, a cruncher heuristic [6] improves the feasible solution iteratively as much as possible. Finally, polishing is run to improve the solution even further using more ‘delicate’ operations (see Appendix B for more details).

The entire flow is controlled by an orchestrator. The orchestration of the optimization flow (e.g., which heuristic, when, and with what parameters to run) is crucial. Finding the right strategy can be a very complex and time-consuming process. Moreover, a new strategy may be required for each new deployment in a new domain or even for a different type of problem in the same domain. To address this, we automated the process by using Markov Decision Process (MDP) framework (see Appendix A for more details). In doing so, we changed our approach to the heuristics from the rules-based approach which we implemented initially. In the rule-based approach, the three aforementioned heuristics are run in series: feasible solution finder, the cruncher, and the polisher. When we applied MDP to the process instead, the cruncher and the polisher are modeled as a single MDP model with an extended set of actions. In this case, there is no order between the cruncher and the polisher: the order is solely prescribed by the policy.

The MDP state variables consists of CPLEX time per iteration, gap to optimality, current objective value, and convergence rate. Action space consists of the number of unfixed unassigned integers, the number of unfixed assigned integers and per iteration time upper bound. Immediate cost is defined as relative objective improvement between consecutive iterations. Transition probability matrix is estimated from the C-PRO runs using real data (see Appendix B for the details) from clients, which is interpolated using math properties of the state features (e.g., continuity, absorbing state knowledge). Reward per state-action is the relative objective improvement between consecutive iterations. During the run, the state of the run is estimated, and an optimal action is applied for the next iteration using pre-solved MDP policy (Fig. 4).

## 4 Results and Conclusions

We applied the MDP flow orchestrator to the C-PRO deployment at Aeroflot (see Appendix C). The MDP flow orchestrator was run on Airbus A320 and A321 fleets, which consists of over 100 aircraft and over 1,000 employees to be scheduled. The MDP flow orchestrator outperformed the existing rule-based orchestrator by roughly 30% in speed. Additionally, it is more flexible to apply to new domains or new problems in the same domain. Our future research directions are focused on building more effective optimization flow concepts, including: a multi-flow parallelism, incorporating advanced constraints formulations into MDP and

Iter	Coverage	CplexTime	Gap	Obj	Conv. Rate	FreeZeros	FreeOnes	TimeUB
0	0.58	350	0.001	5771107	0	2000	0	800
1	0.58	1448	0.3499	198370565	10	72000	1	400
2	0.72	348	0.0006	69613241	3337.3	6000	0.6	200
3	0.78	301	0.0007	45915428	64.9	12000	0.5	200
4	0.82	320	0.0004	38653341	34	12000	0.5	200
5	0.84	300	0.0003	35482205	15.8	12000	0.5	200
6	0.86	305	0.0002	34599949	8.2	12000	0.5	200
7	0.87	302	0.0004	33437291	2.5	12000	0.5	200
8	0.88	271	0.0006	32993209	3.4	12000	0.5	200
9	0.89	313	0.0002	32903230	1.3	12000	0.5	200
10	0.9	379	0.0009	32517656	0.3	24000	0.3	200
11	0.91	335	0.0004	32324484	1.2	24000	0.3	200
12	0.91	331	0.0004	32087260	0.6	24000	0.3	200
13	0.92	343	0.0005	32025903	0.7	24000	0.3	200
14	0.94	989	0.0008	31824439	0.2	72000	2	400
15	0.94	372	0.0001	31771205	0.6	24000	0.3	200
16	0.94	1002	0.0692	31770422	0.2	72000	2	400

**Fig. 4.** Snapshot of Running Trace. This snapshot showing a trace of the run using MDP policy for orchestration applied for A320 and A321 instance of Aeroflot. The first two columns show number of iteration and coverage of the requirements, respectively. The next four columns correspond to state variables and last three columns correspond to action variables.

Deep RL, and more automation (with less skill required) for optimization flow generation and orchestration.

## 5 Discussion

In this section we describe what we learned from working with real customers in enterprise optimization area.

- Lesson 1. Improving optimization means improving the entire stack.  
Every component of the system needs to be tuned to deliver an effective solution in enterprise optimization. There is a big gap between an ‘academic’ optimization solution and the real-world. Solving real-world enterprise optimization problem is a ‘multi-dimensional task’. It’s not just optimization, it’s also: rules description, explainability, short turnaround process and user interface. This is different from a strictly academic approach which mainly focuses on optimization algorithms. To improve the whole solution, one must improve the whole ‘stack’: extract, transform, load (ETL) process, flexibility rules description, and explainability features. Flexibility on rules means that domain expert can add and modify rules that significantly change the

optimization problem. For example, summer and winter have different rules for crew scheduling. In C-PRO, we enabled configurable rules which end-users (domain experts) could modify that would automatically update the optimization problem.

- Lesson 2. Explainability.

The solution must be self-explainable. Customers will adopt systems not only because they provide optimal solutions. Customers also demand that the system can explain and “defend” its choices. Enterprise optimization cannot be a black box. To achieve this, we applied multiple techniques, including a verbosity engine, key performing indicators (KPIs) and monitoring tools.

- Lesson 3. Separation between business and algorithmic logic.

We need modularity for large-scale optimization. Input can frequently change (e.g., format, new features) but this should be transparent to the math model. Otherwise, the system becomes brittle in the face of change. In C-PRO, we accomplished this with modules for ‘business objects’, ‘business logic’, and ‘math objects’. We also enabled interim explainability on these objects to enable customers to see the results of the business logic separate from the entire optimization. By adding simple declarations in our code, we make the results of the business logic reviewable by users.

- Lesson 4. Optimization model manageability.

To handle the math model more effectively, we implemented the math model as a type of database that includes constraints, variables, equations, collectors (predefined construction for defining sums), indicators (predefined construction for defining lower and upper bounds), and penalties (predefined construction for defining objectives). We used business objects as a key for accessing these math objects. We defined an Application Programming Interface (API) to access and modify the math model. For instance, we can apply a query which selects all variables associated with optimization flow orchestration.

- Lesson 5. Reusability.

From the beginning, reusability has been a goal. To recall, in the rule-based flow orchestration all three heuristics – feasible solution finder, the cruncher, and the polisher – must run and in a particular order. When moving to a new problem, we would need to reconfigure a rule-based flow accordingly. However, with the machine learning orchestration, the algorithm itself adjusts automatically to the new problem using available data.

**Acknowledgment.** The authors would like to thank Donny Rose for numerous fruitful discussions.

## 7 Appendix A - MDP Framework

### 7.1 Definition of MDP

An MDP [8] is a 4-tuple  $\langle X, U, P, c \rangle$ , where  $X = \{0, \dots, n - 1\}$  is a finite set of *states*,  $U = \{0, \dots, k - 1\}$  is a finite set of *actions*,  $P : X^2 \times U \rightarrow [0, 1]$  is a *transition probability function*, and  $c : X \times U \rightarrow \mathbb{R}$  is a *cost function*. The

probability of transition from state  $x$  to state  $y$  when action  $u$  is chosen is specified by the function  $P$  and denoted by  $P(y|x, u)$ . The cost associated with selecting the action  $u$  when in state  $x$  equals  $c(x, u)$ . We often refer to the cost function as a vector  $c \in \mathbb{R}^{nk}$ . We denote initial states by  $x_0$ . In fact [8], implies that the initial state does not affect the optimal policy.

Time is discrete, and in each time unit  $t$ , let  $x_t$  denote the random variable that equals the state at time  $t$ . Similarly, let  $u_t$  denote the random variable that equals the action selected at time  $t$ . A non-stationary policy is a function  $\pi : X \times U \times t \rightarrow [0, 1]$ , such that  $\sum_u \pi(x, t, u) = 1$  for every  $x \in X$  for each time unit  $t$ . A stationary policy is a function  $\pi : X \times U \rightarrow [0, 1]$ , such that  $\sum_u \pi(x, u) = 1$  for every  $x \in X$ . A policy controls the action selected in each state as follows: the probability of selecting action  $u$  in state  $x$  equals  $\pi(x, u)$ . A policy can be either randomized or deterministic. A randomized policy is a policy with a state  $x_i$  for which  $\pi(x_i, u) > 0$  for more than one action  $u$ . A deterministic policy is a policy where for all states  $x \in X$ , there is exactly one action  $u \in U$  such that  $\pi(x, u) = 1$ . The initial state together with a policy determine a probability measure on states and actions. The goal is to find a policy that minimizes the cost  $C(\pi)$  defined below. We consider a discounted cost model with infinite horizon throughout the paper.

*Discounted Cost Model.* In the discounted cost model, the parameter  $\beta \in (0, 1)$  specifies the rate by which future costs are reduced. Let  $P^\pi(x_t = x, u_t = u)$  denote the probability of the event  $x_t = x$  and  $u_t = u$  when the initial state equals  $x_0$  (once set, remains unchanged and omitted from the notation) and the policy is  $\pi$ . The infinite horizon discounted expected cost  $C(\pi)$  is defined by

$$C(\pi) \triangleq (1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot E^\pi[c(x_t, u_t)].$$

*Occupation Measures.* Every policy  $\pi$  induces a probability measure over the state-action pairs. We call this probability measure the *occupation measure* corresponding to  $\pi$  and denote it by  $\rho^\pi$  such that  $\rho^\pi(x, u) \triangleq (1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot P^\pi(x_t = x, u_t = u)$  (for simplicity we will omit  $\pi$  from the denotation of  $\rho$ ).

Given an occupation measure  $\rho(x, u)$  over  $X \times U$ , the policy  $\pi^\rho$  induced by  $\rho$  is defined by  $\pi^\rho(x, u) \triangleq \rho(x, u) / \sum_{u'} \rho(x, u')$ . (Note that if  $\sum_{u'} \rho(x, u') = 0$ , then one may define  $\pi^\rho(x, u)$  arbitrarily as long as  $\sum_u \pi^\rho(x, u) = 1$ .) A cost can be rewritten using occupation measure notations such as

$$C(\pi) \triangleq \sum_{x \in X, u \in U} c(x, u) \cdot \rho^\pi(x, u).$$

## 7.2 Transition Probability Matrix Estimation

We start with a data set from a client. The problems are large and take time to solve (every iteration takes between 10 to 20 min); and therefore, the amount of

data samples that can be collected in a reasonable time is limited. We interpolate the client data using domain knowledge information. The augmentation is used to generate a transition probability matrix [11] of MDP, combining available historical data and domain knowledge information. Mainly due to continuity of state variables, we use neighboring state interpolation for the domain knowledge augmentation. We use an absorbing state check to help eliminate misleading samples subject to shortage in data availability. We collect new data samples every time the orchestrator is applied, which we use to augment the historical batch of data and recalculate an MDP transition probability matrix. In Aeroflot use-case, discretizing the variables, we got 24 states and 92 actions. This resulted in  $24 \times 2208$  matrix with total of around 53000 entries. These entries were estimated by using just of around 500 ‘real’ samples from C-PRO.

### 7.3 Linear Programming (LP) Formulation

Publications from the 1960s [3, 4, 7] proved that MDP can be formulated as an LP problem. They also proved that there is a stationary optimal deterministic policy for MDP. Below we show an LP dual formulation for MDP, that appears to be useful in real-life applications. To formulate the LP, we switch to vectorized representation such that  $c$  and  $\rho$  are vectors of length of  $|X| \cdot |U|$ ,  $P$  is a transition probabilities matrix with  $|X|$  rows and  $|X| \cdot |U|$  columns,  $I$  is an identity matrix,  $(1 - \beta, 0, 0, 0 \dots, 0)$  is a vector that represents the initial states distribution where  $1 - \beta$  corresponds to state  $x_0$ , and  $|X| \cdot |U|$  rows. To solve the LP problem we used CPLEX solver.

$$\begin{aligned} & \min_{\rho} c^T \cdot \rho \\ & \text{s.t.} \\ & (I - \beta \cdot P) \cdot \rho = (1 - \beta, 0, 0, 0 \dots, 0)^T \\ & \rho \geq 0 \end{aligned}$$

## 8 Appendix B - Building Blocks of the Optimization Flow

There are three custom heuristics: feasible solution finder, the cruncher, and the polisher.

*The Feasible Solution Finder.* The feasible solution finder consists of two parts:

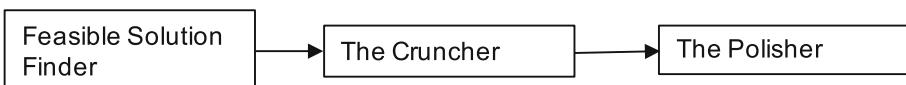
- Labeling different types of constraints into groups (e.g., pairing cover constraints, fairness constraints, etc.).
- Iteratively adding constraints by label to MILP such that all are incorporated, and a feasible solution is found.

The Feasible Solution Finder is rule-based (rules how to apply labels and when to add them to MILP) have been coded based on trial-and-error in creating the C-PRO solution).

*The Cruncher.* The cruncher heuristic improves the feasible solution from the previous stage. It fixes and unfixes assigned/unassigned integer variables iteratively. In C-PRO, the optimization problems is represented as matrix, in which each resource (pilot, flight attendant) is either ‘assigned’ or ‘unassigned’ to a duty. The ‘original’ Cruncher is rule-based and the number of fixed and unfixed variables, run time per iteration, etc., is set by a rule-based orchestrator created based on trial-and-error in creating the C-PRO solution.

*Polisher.* The polisher heuristic works similarly to the cruncher but with the number of variables that are unfixed per iteration is generally smaller than in the cruncher. Instead of working by percentage, it works with tasks which are more ‘gentle’ (e.g., instead of ‘20% of assigned variables to be unfixed to the next iteration’, the Polisher would say, ‘2 assigned tasks per employee to be unfixed for the next iteration’). Like the Cruncher, the ‘original’ Polisher is operated by a rule-based orchestrator.

In the rule-based application of three types of heuristics, called one after the other as it appears on Fig. 5.



**Fig. 5.** Rule-based order of heuristics applied. First: Feasible solution finder, Second: The Cruncher, Third: Polisher.

When MDP is used for orchestration, heuristics type and input configuration are determined by the MDP.

## 9 Appendix C - Aeroflot Letter

See Fig. 6.



Aeroflot is a flag carrier and the largest airline of the Russian Federation. It operates 240 aircraft in 6 fleets to more than 150 destinations in the world. Aeroflot employs over 30,000 employees including 3,000 pilots and 9,000 flight attendants.

In 2018, IBM demonstrated the ability to be most effective and flexible at scheduling large number of personnel, especially flight attendants.

In 2019, Aeroflot contracted with IBM to deploy a crew pairing and rostering optimization system (C-PRO). Starting 2020, Aeroflot has been using C-PRO for optimizing pairing and rostering for pilots flying to over 150 destinations in the world.

Fig. 6. Aeroflot letter.

## References

1. Altman, E.: Constrained Markov Decision Processes. CRC Press, Boca Raton (1999)
2. Andersson, E., Housos, E., Kohl, N., Wedelin, D.: Crew pairing optimization. In: Yu, G. (ed.) OR in Airline Industry, pp. 228–258. Kluwer Academic Publishers, Boston (1999)

3. de Ghellinck, G.: Les problemes de decisions sequentielles. Cahiers Centre d'Etudes Recherche Operationnelle **2**, 161–179 (1960)
4. Depenoux, F.: A probabilistic production and inventory problem. Manage. Sci. **10**(1), 98–108 (1963)
5. Katz, M., Lipets, V., Masin, M., Moshkovich, D., Wasserkrug, S.E.: Reusable Modeling for Solving Problems. US Patent US20170337042A1 (2017)
6. Lipets, V., Schiloach, Y.: Reusable Modeling for Solving Problems. US Patent US8554704B2 (2010)
7. Manne, A.: Linear programming and sequential decisions. Manage. Sci. **6**(3), 259–267 (1960)
8. Puterman, M.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)
9. Ye, X.: Airlines' crew pairing optimization: a brief review technical report. Department of Applied Mathematics and Statistics, Johns Hopkins University (2007)
10. Yu, G., Thengvall, B.G.: Airline optimization. In: Pardalos, P.M., Resende, M.G.C. (eds.) Handbook of Applied Optimization, pp. 689–703. Oxford University Press, New York (2002)
11. Zadorojniy, A., Shwartz, A., Wasserkrug, S., Zeltyn, S.: Operational optimization of wastewater treatment plants: a CMDP based decomposition approach. Ann. Oper. Res. (2016). <https://doi.org/10.1007/s10479-016-2146-z>



# A Regulatory Principle for Robust Reciprocal-Time Decay of the Adaptive Immune Response

Anthony Almudevar<sup>(✉)</sup>

Department of Biostatistics and Computational Biology, University of Rochester,  
Rochester, NY 14642, USA  
[anthony.almudevar@urmc.rochester.edu](mailto:anthony.almudevar@urmc.rochester.edu)

**Abstract.** Follicular dendritic cells (FDC) play a crucial role in the regulation of immunity. They are believed to be responsible for long-term persistence of humoral antibody following vaccination or infection, due to their role in antibody response induction and their ability to retain antigen for long periods. In this paper, a regulatory control model is described which links persistence of humoral immunity with cellular processes associated with FDCs. The model predicts universal and stable reciprocal-time ( $= 1/t$ ) decay of humoral antibody levels, which has been widely reported over a range of ages, observation times and vaccine types.

**Keywords:** Control model · Power-law decay · Homeostasis · Adaptive immune response

**AMS(2020) Subject Classification:** Primary 93C95 · Secondary 37N25 · 92C37

## 1 Introduction

The observation of humoral antibody (Ab) concentrations following vaccination permits the estimation of post-challenge Ab kinetics, and many such studies are reported in the literature. One important advantage of these studies is that observations can be time synchronized to measure Ab decay from a common challenge starting time. The dynamics of Ab response  $C_t$  in time  $t \geq t_{min}$  are commonly observed to be driven by a period of rapid increase to peak levels, followed by prolonged periods of decay. This decay process is widely reported to conform to a power-law decay model

$$\frac{C_t}{C_s} = \left[ \frac{t}{s} \right]^k, \quad s, t \geq t_{min}, \quad (1)$$

for some  $k < 0$  (see [4, 10, 21, 22]). Power-law decay of Ab response was formulated as a model in some detail in [12], where it was noted that  $k$  was close to  $-1$  in the several examples given for which statistical estimates were available.

In some cases, Ab decay appears to include a nonzero asymptote, probably due to long-term Ab production by plasma cells which have migrated to bone marrow (see [5, 7, 9]). In this case, observed Ab decay is probably a superposition of two processes,

$$\frac{C_t}{C_s} = h_s(t) + \nu, \quad (2)$$

where  $h_s(t)$  represents the adaptive immune response, which decays to zero, and  $\nu$  represents longer term Ab production due to plasma cells. Then the power-law decay proposed in [12] would be represented by the component  $h_s(t)$ .

Thus, estimation of decay rates must anticipate a nonzero asymptote  $\nu$  using appropriate statistical methods. In [1] a literature review of Ab response studies was undertaken with the purpose of validating the power-law decay model with the inclusion of a nonzero asymptote  $\nu$ . Of the 13 Ab time series examined, two exhibited no variation over time (as a consequence of a poor vaccine response), while the remaining 11 conformed very closely to the power-law decay model, with exponent  $k = -1$ .

Identifying the regulatory principle by which the immune response terminates is an important open problem, since unregulated Ab persistence is physiologically harmful, which is what characterizes auto-immune disease (for a recent discussion of the issue see [15]). If the adaptive immune response possesses a single decay rate for all infection types, this may be a fingerprint of an important regulatory principle. On the other hand, empirical observations of power-law decay are often the result of some artifact, rather than the direct observation of a dynamic law deducible from first principles.

We give a brief outline of this paper. The question of the empirical observation of power-law decay as artifact is considered in Sect. 2. If we were to accept power-law or reciprocal-time decay as a true model of decay, the question then arises as to the type of model that would be needed to predict that form of decay. In Sect. 3 we argue that the properties of power-law decay force a careful consideration of the class of model which would be appropriate. In Sect. 4 we describe a control model for the regulation of the adaptive immune response proposed in [1]. The model is based on the functionality of follicular dendritic cells (FDC), which are found in the B-cell follicles of secondary lymph nodes, the primary site of the adaptive immune response. The model possesses reciprocal-time decay as a stable attractor. While the attractor is maintained by homeostatic control, at a higher level the control model does not rely on feedback. Rather, an FDC population provides open-loop control by functioning as a timer. The model is demonstrated by computer simulations in Sect. 5, with a discussion following in Sect. 6.

## 2 Empirical Observations of Power-Law Decay

In the literature the term “power-law decay” is used to describe both decay in a dynamic process  $C_t$  and a probability distribution. Of course, the two can be equated. Given normalization  $C_{t_0} = 1$ , we may set survival curve  $P(T > t) = C_t$

as the population proportion surviving beyond time  $t$ , where  $T$  is the cell survival time. The density function  $f_T$  of  $T$  is the derivative of  $-P(T > t)$ , so given  $r > 0$  in (1), we have  $f_T(t) \propto 1/t^{r+1}$ . Therefore, under power-law decay of rate  $r$  survival times possess a Pareto density with parameter  $r + 1$ . This means the commentary on power-law frequencies is generally relevant to power-law decay, with respective decay rates  $r + 1$  and  $r$ .

There are many explanations of empirical power-law offered in the literature. We next review a few of these.

**A Consequence of Underpowered Statistical Analysis.** A power-law between two quantities  $y \propto x^\alpha$  can be discerned from paired observations  $(x_i, y_i)$ ,  $i = 1, \dots, m$  by the double log transform  $\log y = \alpha \log x + C$ , which under the power-law hypothesis will be a straight line. Thus, a linear log-log plot is widely accepted as a fingerprint of power-law decay. This means that the power-law is widely accepted as a null hypothesis, if only implicitly. This is analogous to the common practice of testing for normality in statistical modeling. The difference here is that normality is theoretically justified by the central limit theorem as the aggregation of additive noise, whereas there is often little first principles justification for accepting the power-law as a null hypothesis. As argued in [6], when data is compatible with a power-law, it may be compatible with any number of alternative heavy tailed distributions. Therefore, widespread reports of power-law decay may be partly explained by its acceptance as a null hypothesis. In [6] 24 data sets reported to conform to the power-law distribution were reanalyzed. The power-law was ruled out in 7 of these using a goodness-of-fit test. Of the remaining data sets, only one (distribution of frequencies of word occurrence in the English language) was convincingly power-law, in the sense that a set of alternative densities could be rejected.

**A Consequence of Aggregation.** The exponential model of decay is given by

$$\frac{C_t}{C_s} = e^{-\mu(t-s)} \quad t > s, \quad (3)$$

where  $\mu$  is a positive constant. Some models accept (3), but explain empirical power-law decay as an artifact of observation. One widely reported version of this effect is the rate mixture model. If  $f(\mu)$  is a gamma density with shape and rate parameters  $\alpha, \tau$  then

$$\int_{\mu=0}^{\infty} e^{-\mu t} f(\mu) d\mu = \frac{1}{(1 + \tau t)^\alpha}. \quad (4)$$

In [17] it is assumed that human memory decays exponentially for individuals, but with some population variation in rates. Thus, observed power-law decay at the population level is simply an artifact of statistical averaging over a population. Another version of this process is reported in [19], based on observations of exponentially growing processes at random times. It is first noted that if we

are given a deterministic exponential growth process  $X(t) = \exp(\mu t)$  and  $T$  is an exponentially distributed observation time, then  $X(T)$  has a power-law distribution. The idea is extended to a number of stochastic processes commonly used in modeling which exhibit exponential growth. This raises the possibility that empirical power-law decay in Ab response studies is due to the averaging of individual times series with imperfectly synchronized observation times. In [9] the aggregation model of Eq. (4) was proposed to explain empirical power-law decay in post-vaccine Ab concentrations, assuming significant heterogeneity of decay rates within the immune response.

**A Consequence of Partial Decay.** Another reason that exponential decay may resemble power-law decay is that some portion  $c$  of the original concentration  $C_{t_{min}}$  is protected from decay. This yields the relationship

$$(C_t - c) = (C_s - c)e^{-\mu(t-s)} \quad (5)$$

which would yield exponential decay with an asymptote other than 0. It is easy to see how this may empirically resemble the much slower power-law decay. This model was suggested as one of several possible explanations for the empirical power-law decay of memory function in [26]. In that article (which can be especially recommended), it is also proposed that power-law observations may follow from nonlinear measurement of memory function, in particular, that a measurement scale may be less sensitive at lower function. Heterogeneous aggregation of the type described above is also given as a putative explanation, as well as the possibility that power-law decay truly is a first principles model of memory loss (see [13, 25]).

Thus, when given empirical observation of power-law decay, the possibilities enumerated above must be considered. Regarding the question specifically of Ab decay, it is always possible to model the asymptote  $\nu$  in Eq. (2) in order to study the remaining component  $h_s(t)$  (see [1]). Regarding aggregation, that effect would explain power-law decay, but not specifically reciprocal-time decay implied by  $k = -1$ . However, ultimately, to accept power-law decay of the Ab response, a plausible model must be proposed.

### 3 Autonomous *versus* Non-autonomous Dynamic Systems

We take an autonomous dynamic system to be one in which the dynamic law is unchanging in time. This characterizes models of biochemical systems with static decay or interaction rates. This type of model is commonly used to model the immune response.

For example, in [23] the effect of variable or dynamic vaccine dose administration on Ab response was studied experimentally, and compared to computational model predictions. The model includes as state variables  $C_{Ag}$ ,  $C_{IgG}$ ,  $C_{IgM}$ ,  $C_{IC}$  and  $C_{PC}$ , representing concentrations of antigen, immunoglobulin G

(IgG) antibodies, immunoglobulin M (IgM) antibodies, immune complexes (IC) and plasma cells (PC), respectively. The dynamic laws were defined by a system of five ordinary differential equations. Apart from an endogenous input  $F(t)$  of antigen, representing a designed vaccine schedule, the model is time homogeneous, and defined by static decay and interaction rates among the system variables.

As another example, we consider the model proposed in [16]. It contains only two state variables:  $T$ , the concentration of T-cells; and  $C$ , the concentration of peptide-MHCs (pMHC), which transport antigen fragments for recognition by T-cell receptors. This is required for stimulation of the adaptive immune response, specifically, growth of the T-cell population requires interaction with pMHC. The model equations are

$$\begin{aligned}\frac{dT}{dt} &= \alpha \frac{TC}{K + T + C} - \delta T, \\ \frac{dC}{dt} &= -\mu C,\end{aligned}$$

where  $K$  is a constant, and  $\alpha$ ,  $\delta$  and  $\mu$  are positive system parameters. We assume  $\alpha > \delta$ . Initially, the environment is saturated with pMHC, so that  $C \gg K + T$ , and  $T \propto e^{(\alpha-\delta)t}$ . Then  $C$  decays exponentially at rate  $-\mu$ , independently of the other components of the system. Eventually,  $C \ll K + T$ , so that T-cells decay exponentially at rate  $-\delta$ . The model predicts an interesting relationship at the time  $t^*$  of peak T-cell concentration, in particular,

$$\frac{T(t^*)}{T(0)} = \left( \frac{C(0)}{T(0)} \right)^{(\alpha-\delta)/(\alpha-\delta+\mu)}.$$

The prediction that the peak fold increase of T-cells is positively related to the initial input of antigen, but inversely related to the initial T-cell concentration was observed experimentally in [18].

However, if Ab decay is truly power-law, then it is difficult to see how it can be driven by an autonomous dynamic system (we have noted that [9] models power-law decay as a mixture of autonomous decay models, but this degree of heterogeneity does not appear to be compatible with known immune response function).

Cellular processes are believed to possess, in general, robustness properties which ensure uniformity of outcomes under varying conditions (see, for example, [2]). Robustness can take various forms, for example, insensitivity to model parameter values, as defined in [20]. It seems reasonable, therefore, to model the adaptive immune response as a control system relying on biologically plausible control mechanisms. In fact, if Ab decay is not only consistently power-law, but power-law specifically with rate  $\propto 1/t$ , then the argument for this approach is strengthened all the more.

## 4 Control Model for FDC Decay

Follicular dendritic cells (FDC) are found in the B-cell follicles of secondary lymph nodes, the primary site of the adaptive immune response. Their function is to capture and retain antigen in immunogenic form, and to induce Ab response by supporting germinal centers (GC), the sites of B-cell maturation. They are nonmigratory, and form a reticula network which defines a microenvironment. Under the conventional model of the adaptive immune response, it would be reasonable to conjecture that humoral Ab levels are proportional to GC concentration, which is in turn proportional to FDC concentration (for example, the ratio of FDC antigen retaining reticula and GCs was reported to be 1:1 in mouse lymph tissue in [24]). Therefore, it is plausible that a control model for FDC concentration can explain reciprocal-time decay of humoral Ab levels (see [1]).

### 4.1 Model Definition

A non-autonomous dynamic model which predicts reciprocal-time decay is quite easy to construct. For example, if  $F = tC_t$  is a balance equation for concentration  $C_t$ , and  $F$  is held constant, then  $C_t = F/t$ . The question, of course, is whether or not such a balance equation has any relevance to the problem at hand. In fact, we will argue that the quite unique functionality of the FDC makes this equation very relevant.

Suppose there exists a population of activated FDCs, the initial size being a positive real number  $C_0 = N \in \mathbb{R}$ . The model system  $\mathcal{S}$  is partitioned into a reservoir  $\mathcal{R}$  and an FDC population  $\mathcal{F}$ . Flow through  $\mathcal{S}$  is given by:

$$\text{External antigen source} \rightarrow \mathcal{R} \rightarrow \mathcal{F} \rightarrow \text{Antigen clearance}.$$

Antigen transport pathways exist in  $\mathcal{R}$ , while antigen retained in FDCs exists in  $\mathcal{F}$ .

Let  $C_t$ ,  $F_t$  be the population size of still active FDCs and the total amount of antigen in  $\mathcal{F}$  at time  $t \in [0, \infty)$ , respectively. We take  $C_t \in [0, N]$ ,  $F_t \in [0, \infty)$  to be real valued, with initial values  $C_0 = N$ ,  $F_0 = 0$ .

Define the following rules:

- (A1) As long as a unit FDC remains active it ingests antigen at a rate of  $\mu$  per unit time.
- (A2) A unit FDC may be deactivated at any time, at which point its total ingested antigen is released.
- (A3) No FDC can be created or reactivated.

Under rules (A1)–(A3) the balance equation

$$F_t = \mu t C_t, \quad t \geq 0 \tag{6}$$

must hold. Differentiating (6) then gives

$$\frac{dF_t}{dt} = \mu \left[ C_t + t \frac{dC_t}{dt} \right]. \tag{7}$$

The terms of Eq. (7) have an intuitive interpretation. Antigen is ingested at a rate of  $\mu$  per unit cell, giving the term  $\mu C_t$ . At time  $t$  a unit FDC has ingested  $\mu t$  units of antigen, therefore a decay rate of  $dC_t/dt < 0$  forces release of antigen from  $\mathcal{F}$  at the rate  $-\mu t dC_t/dt$ . Thus, the system steady state  $dF_t/dt = 0$  is characterized by both constant antigen retention  $F_t = F_\infty$  and reciprocal-time decay of the FDC population  $C_t = F_\infty/\mu t$ .

## 4.2 A Homeostatic Control Model

The next problem is to introduce a control effector into (7). Define the double-logarithmic derivative

$$k_t = \frac{d \log C_t}{d \log t} = \frac{C_t^{-1} dC_t}{t^{-1} dt}.$$

The solution to  $k_t \equiv k$  yields the power-law decay of Eq. (1). We may then rewrite (7) as

$$\frac{dF_t}{dt} = \mu \cdot C_t [1 + k_t], \quad (8)$$

from which a simple control effector emerges. Maintaining  $k_t \equiv -1$  forces  $dF_t/dt = 0$ , and  $k_t > -1$  or  $k_t < -1$  forces increase or decrease in  $F_t$ , respectively. Thus, feedback control of  $k_t$ , which determines the decay rate of  $C_t$ , provides a mechanism for homeostatic maintenance of the system steady state. Interestingly, the steady state is mathematically equivalent to reciprocal-time decay of  $C_t$ , and therefore of humoral Ab levels. This would predict the universal observation of reciprocal-time decay reported in [1].

Of course, the problem remains of proposing a biologically plausible control law for  $k_t$  with the system steady state as an attractor. It would be reasonable to assume that control is effected at the individual cell level, taking the form

$$\frac{dC_t}{dt} = -\bar{\lambda}(F_t, C_t, t)C_t \quad (9)$$

for some unit cell decay control function  $\bar{\lambda} \geq 0$ . We can substitute the balance Eq. eqrefeq.balance into (9) to obtain a first-order ordinary differential equation (ODE):

$$\frac{dC_t}{dt} = -\bar{\lambda}(\mu t C_t, C_t, t)C_t. \quad (10)$$

In this form,  $\bar{\lambda}$  could be interpreted as a stochastic FDC failure (deactivation) rate.

## 4.3 Exponential Decay Cannot Yield Homeostatic Control

Suppose  $\mathcal{R}$  always contains sufficient antigen for FDC ingestion, and the unit cell decay rate is constant at  $\bar{\lambda}(F_t, C_t, t) \equiv \rho > 0$ , resulting in exponential population decay. The solution to (9) is  $C_t = C_0 \exp(-\rho t)$ , in which case  $F_t = \mu t C_0 \exp(-\rho t)$ . This function possesses a global maximum at  $t = 1/\rho$ . Therefore,  $F_t$  increases to peak level  $F_{max} = (\mu/\rho)C_0 \exp(-1)$  then converges to zero. Thus a statistic decay rate for  $C_t$  cannot yield homeostatic control of the steady state.

#### 4.4 Balance Equations for Steady State Antigen Flow Through System $\mathcal{S}$

To construct a plausible homeostatic control we will expand the definition of the system. We define the amount of antigen  $E_t \in [0, \infty)$  contained in  $\mathcal{R}$ . This is the antigen available for FDC ingestion. The initial reservoir level is then  $E_0 = R > 0$ . Let  $A_t$  be the total amount of additional antigen entering  $\mathcal{R}$  by time  $t$ . Then let  $B_t$  be the total antigen released by deactivated FDCs by time  $t$ . We must have

$$\frac{dB_t}{dt} = -\mu t \frac{dC_t}{dt}. \quad (11)$$

Assuming  $B_t$  is lost to the system, the balance equation may be expanded to

$$\begin{aligned} F_t &= \mu t C_t, \\ R + \Delta_t &= E_t + F_t, \quad t \geq 0 \text{ where } \Delta_t = A_t - B_t. \end{aligned} \quad (12)$$

If the net flow of antigen through  $\mathcal{S}$  is zero, then the additional balance condition

$$\Delta_t = 0 \quad (13)$$

holds. Accepting (12) and (13), convergence to the steady state,  $F_t \rightarrow F_\infty$  can be then expressed as:

$$\lim_{t \rightarrow \infty} E_t = E_\infty < R. \quad (14)$$

In other words, under the system steady state  $\mathcal{R}$  is indefinitely depleted in part or in full. In this case  $F_\infty = R - E_\infty$ , forcing invariant reciprocal-time decay  $C_t = (R - E_\infty)/\mu t$ .

The control function  $\bar{\lambda}$  may depend on any of the quantities in (12), assuming they satisfy the balance conditions, and so the system remains governed by the control equation

$$\frac{dC_t}{dt} = -\bar{\lambda}(A_t, B_t, C_t, E_t, F_t, t)C_t. \quad (15)$$

#### 4.5 Control Based on Allocation of Available Antigen

A reasonable conjecture is that FDC deactivation is upregulated by antigen scarcity, similar to the model proposed in [16] (Sect. 3). Suppose antigen is made available to a single FDC by a Poisson arrival process of rate  $\gamma$ . A failure occurs when an interarrival time exceeds some threshold  $\kappa$ . This failure results in the deactivation of the FDC, and the release of its retained antigen. Since this failure rate will depend on both antigen availability ( $\propto E_t$ ) and competition for antigen ( $\propto C_t$ ) this becomes a potential control effector.

A Poisson process is well approximated by a discrete time arrival process. Independent binary random variables  $X_i$ ,  $i = 1, 2, \dots$  with mean  $q_\delta$  are observed at times  $\delta i$ ,  $i = 1, 2, \dots$ . An arrival occurs at time  $\delta i$  if  $X_i = 1$ . The constraint  $\gamma\delta = q_\delta$  forces an arrival rate of  $\gamma$ . A failure is initiated at time  $\delta i$  if  $X_i = 1$  and  $X_{i+1} = \dots = X_{i+n_\delta} = 0$ , where  $\kappa = n_\delta\delta$  (we lose no generality in choosing  $\delta$

so that  $n_\delta$  is an integer). The expected number of failure initiations  $N_F$  in time interval  $[0, N\delta]$  is

$$E[N_F] = \sum_{i=1}^N P(X_i = 1, X_{i+1} = \dots = X_{i+n_\delta} = 0) = Nq_\delta(1 - q_\delta)^{n_\delta}.$$

The rate of failure initiation is therefore

$$\rho_\delta = \frac{Nq_\delta(1 - q_\delta)^{n_\delta}}{N\delta} = \frac{N\gamma\delta(1 - \gamma\delta)^{\kappa/\delta}}{N\delta} = \gamma(1 - \gamma\delta)^{\kappa/\delta}.$$

Finally, refining the discrete approximation gives failure rate

$$\rho = \lim_{\delta \rightarrow 0} \rho_\delta = \gamma \exp(-\gamma\kappa).$$

The argument is completed by noting that under general conditions the aggregation of  $m$  arrival processes approaches in distribution a Poisson process as  $m \rightarrow \infty$ , so that the model will be reasonably robust with respect to assumptions (see [8]). Under the proposed model the antigen arrival rate per FDC is proportional to  $E_t/C_t$ , therefore the FDC failure rate would be

$$\bar{\lambda}(C_t, E_t) = \begin{cases} \gamma E_t/C_t \exp(-\gamma\kappa E_t/C_t) & ; E_t > 0 \\ \infty & ; E_t = 0 \end{cases}, \quad (16)$$

noting that the population is extinguished essentially instantaneously when  $E_t = 0$  (i.e. when  $\mathcal{R}$  is depleted).

To remain active, the aggregate antigen arrival rate  $\gamma^*$  for an individual FDC must be larger than  $\mu$ . Under these conditions, the neighborhood of an FDC is essentially saturated with available antigen, and therefore able to maintain the maximum ingestion rate  $\mu$ . As antigen is depleted the quantity  $E_t/C_t$  decreases, forcing  $\gamma^*$  to approach  $\mu$ , making an ingestion failure event increasingly likely. Thus, this failure model predicts property (A1). Convergence to reciprocal-time decay under this control law when net antigen flow is zero is verified in the following theorem (see [1] for proof).

**Theorem 1.** *Suppose the control function  $\bar{\lambda}$  of Eq. (15) is given by Eq. (16). Suppose balance Eqs. (12) and (13) hold. Then there exists a constant  $t^*$ , dependent only on parameters  $(\mu, \gamma, \kappa)$ , for which the following statements hold:*

- (i) *For any initial state  $(t, C_t) = (t_0, C_{t_0})$  for which  $t_0 > t^*$  there exists a positive constant  $r^*$  such that for all large enough  $R^*$  we have:*

$$0 < C_t < \frac{R^*}{\mu t + r^*},$$

*and therefore  $E_t/C_t > r^*$ ,  $t \geq t_0$ , where  $R^* = E_0$  is taken to be the initial reservoir quantity.*

- (ii) *Given the initial conditions of statement (i), if  $E_0 = R^*$  then  $\lim_{t \rightarrow \infty} \mu t C_t = R^*$ .*

Thus, under the conditions of Theorem 1,  $C_t$  possesses reciprocal-time decay in the limit, and steady state antigen retention  $\lim_{t \rightarrow \infty} F_t = R$ , with complete reservoir depletion  $\lim_{t \rightarrow \infty} E_t = 0$ . The steady state retention level  $F_\infty$  therefore depends on  $R$  but not on the model parameters  $(\mu, \gamma, \kappa)$ .

## 5 Computer Simulations

We next demonstrate the model using computer simulations reported in [1]. We can observe the convergence of  $C_t$  to reciprocal-time decay, as the initial reservoir  $\mathcal{R}$  of antigen is depleted and retained in the FDC population  $\mathcal{F}$ .

Balance Eqs. (12) and (13) are assumed to hold, and we use the control model with failure rate  $\bar{\lambda}$  given by Eq. (16). We take time interval to be  $t \in [0, 1000]$ , with initial FDC population  $C_0 = 10^3$ . The initial antigen level is varied by setting  $R/C_0 = 25000, 5000, 1000$ . The antigen ingestion rate is set to  $\mu = 10^3$ . To determine the parameters for  $\bar{\lambda}$  consider the case  $R/C_0 = 1000$ . This gives an antigen arrival rate per FDC at  $t = 0$  of  $\gamma E_0/C_0 = \gamma R/C_0 = \gamma 1000$ . Equating this to  $\mu$  gives  $\gamma = 1$ . Given ingestion rate  $\mu$  it would be reasonable to set  $\kappa$  to be some factor of  $\mu^{-1}$ , so we set  $\kappa = \mu^{-1} = 1/1000$ . The model was discretized by time intervals  $\Delta t = 10^{-4}$ .

Figure 1 shows model pathways for varying initial resource  $R/C_0 = 25000, 5000, 1000$  (columns 1–3). In row 1 plots of  $C_t$  and  $E_t$  are shown with a vertical log scale. Row 2 shows  $C_t$  and  $E_t$  on a log-log scale. Grid lines parallel to  $t^{-1}$  are superimposed. For display  $E_0, C_0$  are both normalized to equal 100% in rows 1–2. Row 3 gives the double logarithmic decay rate  $k_t$  as a function of time. Row 4 gives the relative concentration of retained antigen  $F_t/R$ .

The behavior for each set of initial conditions is unvarying, and conforms to the model's prediction. Each example begins with a short period of decay at  $k_t$  close to 0, then approaches  $k_t = -1$  by times ranging from  $t \approx 100 - 250$  (rows 2–3).  $F_t$  quickly reaches its predicted steady state level  $R$  (row 4).

We next examine the robustness of the model to perturbation. Figure 2 is based on the same model used for Fig. 1 ( $R/C_0 = 25000$ ) but with various forms of stochastic noise introduced (columns 1–3). For the “random resource spikes” model the reservoir  $\mathcal{R}$  was supplemented by bulk arrivals of 500 antigen units according to a Poisson process of rate 0.04. For the remaining models multiplicative noise was incorporated by multiplying  $dC_t/dt$  by a log-normal random variable at each computation point (the exponentiated normal random variates had mean  $\mu = 0$  and standard deviations  $\sigma = 0.1, 1$ ).

In each case the models exhibit the same limiting behavior seen in Fig. 1, despite persistent random perturbations. For the random resource spikes model the assumption of constant system resource  $R$  is violated, but without apparent effect on the approach to the predicted system steady state (Fig. 2 column 1).

For the multiplicative noise model with  $\sigma = 0.1$  (Fig. 2, column 2), the behavior differs little from the corresponding noiseless model (Fig. 1, column 1). What is of some interest is the stable fluctuation of  $k_t$  about the steady state value  $k = -1$ , suggesting an efficient negative feedback control able to maintain

reciprocal-time decay. Setting  $\sigma = 1$  results in considerably more noise (Fig. 2, column 3). The decay rate  $k_t$  no longer fluctuates about  $k = -1$  in a stable manner, but instead subjects the system to frequent and extremely large decay rates. In this case, fluctuation of  $E_t$  is more evident (rows 1, 2). Despite this, the system steady state is maintained.

## 6 Discussion

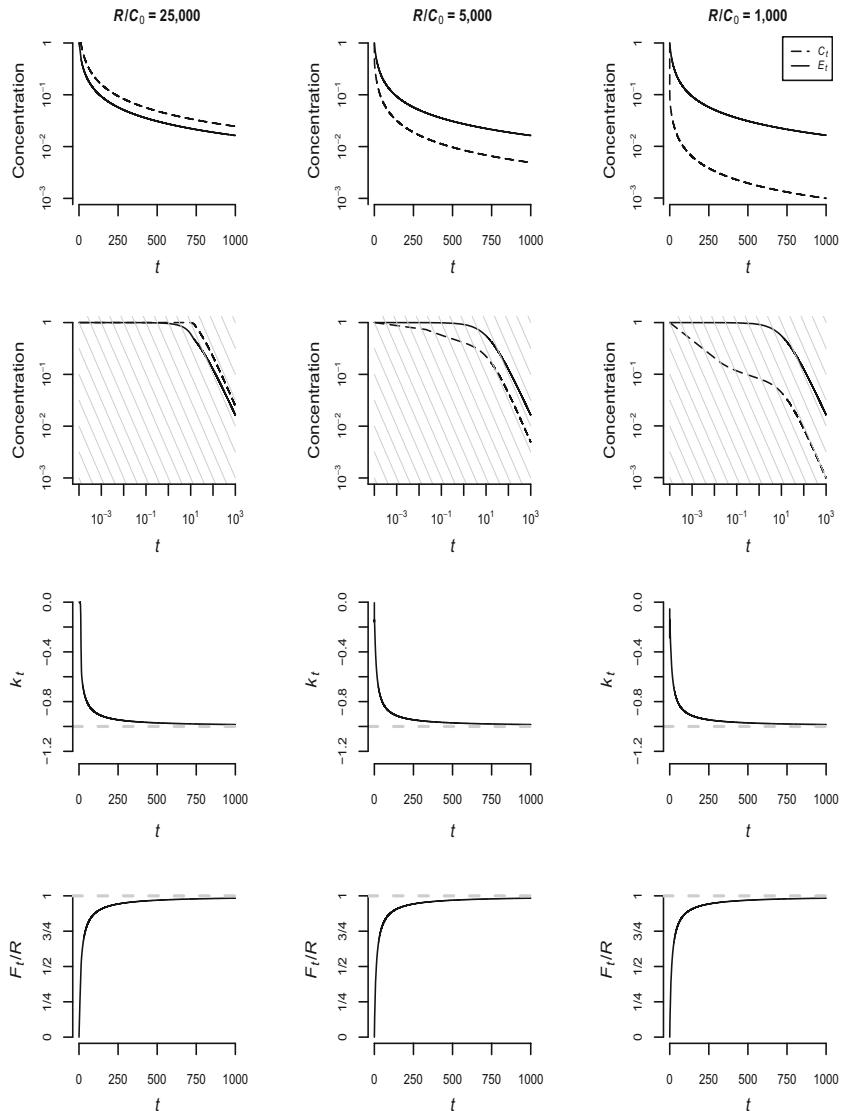
The model proposed in [1] achieves a number of objectives. First, it predicts the universal reciprocal-time decay that has been widely reported in the literature. Furthermore, reciprocal-time decay was demonstrated to be a stable attractor. Remarkably, the model conforms to the robustness principle of insensitivity to model parameter values (see [20]) in the sense that the long-term behavior does not depend on any model parameters other than the initial antigen level  $R$ , provided this value is large enough (Theorem 1).

The remaining questions have to do with the biological plausibility of the model. In fact, there is a striking concordance between cell properties required by the model and those widely reported of FDCs, which are generally unique to this cell type.

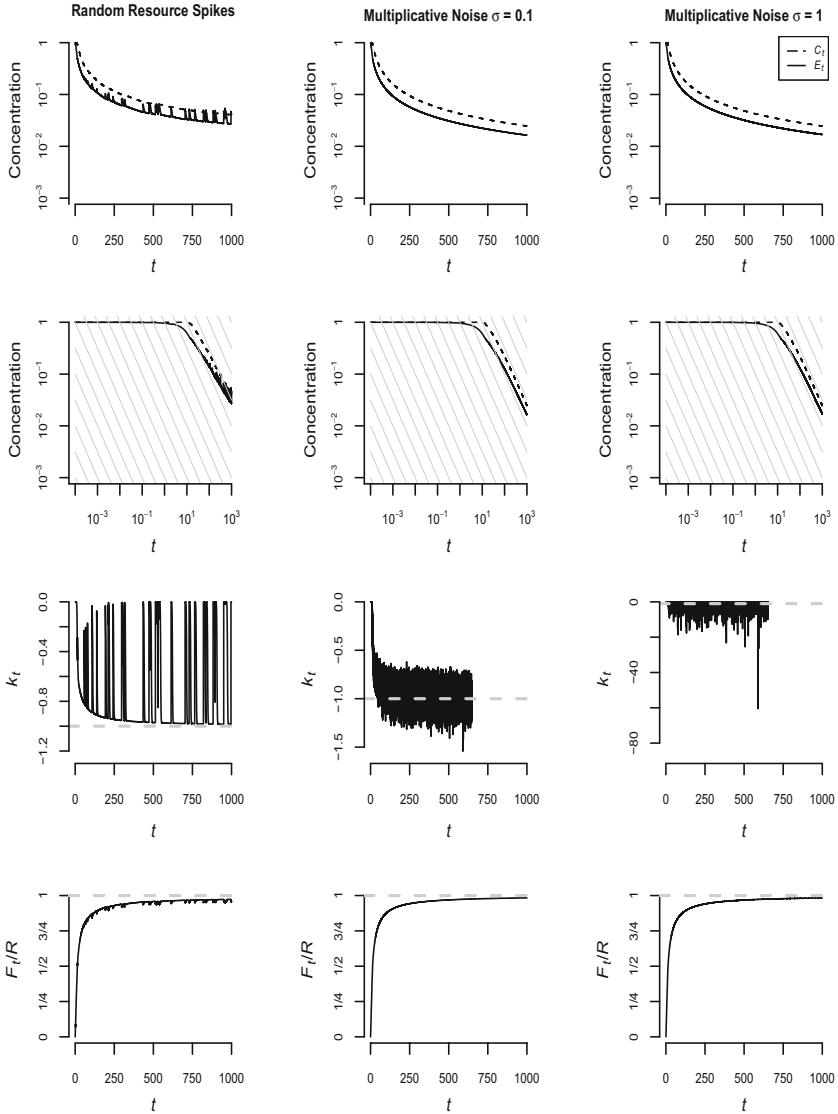
The ability of FDCs to retain intact antigen for extended periods has been consistently reported. This property is frequently conjectured to be related to long term persistence of Ab concentrations (see [11]).

Regarding properties **(A1)**–**(A2)**, it was reported in [14] that maintenance of FDC functionality requires continual lymphotoxin  $\alpha/\beta$  (LT) signalling. Inhibition of LT signalling not only prevents FDC ingestion of antigen, but eliminates previously ingested antigen. The authors write that “[a] surprising observation is that the maintenance of pre-existing FDCs in a differentiated state requires continual interaction with B lymphocytes expressing LT $\alpha\beta$

Thus, the model of [1] is able to unify disparate observations of FDC function, providing a simple regulatory principle which predicts a robust, universal reciprocal-time decay rate for any adaptive immune response. Remarkably, under this principle no feedback is required to terminate the immune response. Rather, at the highest level the control is open-loop, with the FDC population functioning collectively as an immune response timer.



**Fig. 1.** Plots show model pathways for varying total resource  $R/C_0 = 25000, 5000, 1000$  (columns 1–3). See Sect. 5 for descriptions. In row 1 plots of  $C_t$  and  $E_t$  are shown with a vertical log-scale. Row 2 shows  $C_t$  and  $E_t$  on a log-log scale. Grid lines parallel to  $t^{-1}$  are superimposed. For display purposes  $E_0, C_0$  are both normalized to equal 100% in rows 1–2. Row 3 gives the double-logarithmic decay rate  $k_t$  as a function of time. A horizontal reference line is included at  $k = -1$ . Row 4 gives the relative concentration of  $F_t/R$ . A horizontal reference line is included at  $F_t/R = 1$ .



**Fig. 2.** Plots show model used for Fig. 1 with  $R/C_0 = 25000$  incorporating various forms of stochastic noise (columns 1–3). See Sect. 5 for descriptions. In row 1 plots of  $C_t$  and  $E_t$  are shown with a vertical log-scale. Row 2 shows  $C_t$  and  $E_t$  on a log-log scale. Grid lines parallel to  $t^{-1}$  are superimposed. For display purposes  $E_0, C_0$  are both normalized to equal 100% in rows 1–2. Row 3 gives the double-logarithmic decay rate  $k_t$  as a function of time. A horizontal reference line is included at  $k = -1$ . Row 4 gives the relative concentration of  $F_t/R$ . A horizontal reference line is included at  $F_t/R = 1$ .

## References

1. Almudevar, A.: A model for the regulation of follicular dendritic cells predicts invariant reciprocal-time decay of post-vaccine antibody response. *Immunol. Cell Biol.* **95**(9), 832–842 (2017)
2. Alon, U.: An Introduction to Systems Biology: Design Principles of Biological Circuits. CRC Press, Boca Raton (2019)
3. Ansel, K.M., Ngo, V.N., Hyman, P.L., Luther, S.A., Förster, R., Sedgwick, J.D., Browning, J.L., Lipp, M., Cyster, J.G.: A chemokine-driven positive feedback loop organizes lymphoid follicles. *Nature* **406**(6793), 309–314 (2000)
4. Borrow, R., Andrews, N., Findlow, H., Waight, P., Southern, J., Crowley-Luke, A., Stapley, L., England, A., Findlow, J., Miller, E.: Kinetics of antibody persistence following administration of a combination meningococcal serogroup C and haemophilus influenzae type B conjugate vaccine in healthy infants in the United Kingdom primed with a monovalent meningococcal serogroup C vaccine. *Clin. Vaccine Immunol.* **17**(1), 154–159 (2010)
5. Chen, S., Zhou, Z., Wei, F.-X., Huang, S.-J., Tan, Z., Fang, Y., Zhu, F.-C., Wu, T., Zhang, J., Xia, N.-S.: Modeling the long-term antibody response of a hepatitis E vaccine. *Vaccine* **33**(33), 4124–4129 (2015)
6. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. *SIAM Rev.* **51**(4), 661–703 (2009)
7. David, M.-P., Van Herck, K., Hardt, K., Tibaldi, F., Dubin, G., Descamps, D., Van Damme, P.: Long-term persistence of anti-HPV-16 and -18 antibodies induced by vaccination with the as04-adjuvanted cervical cancer vaccine: Modeling of sustained antibody responses. *Gynecol. Oncol.* **115**(3), S1–S6 (2009)
8. Feller, W.: Probability Theory and Its Applications, 2nd edn., vol. 2. Wiley, New York (1971)
9. Fraser, C., Tomassini, J.E., Xi, L., Golm, G., Watson, M., Giuliano, A.R., Barr, E., Ault, K.A.: Modeling the long-term antibody response of a human papillomavirus (HPV) virus-like particle (VLP) type 16 prophylactic vaccine. *Vaccine* **25**(21), 4324–4333 (2007)
10. Gesemann, M., Scheiermann, N.: Quantification of hepatitis B vaccine-induced antibodies as a predictor of anti-HBs persistence. *Vaccine* **13**(5), 443–447 (1995)
11. Heesters, B.A., Myers, R.C., Carroll, M.C.: Follicular dendritic cells: dynamic antigen libraries. *Nat. Rev. Immunol.* **14**(7), 495–504 (2014)
12. Honorati, M., Palareti, A., Dolzani, P., Busachi, C., Rizzoli, R., Facchini, A.: A mathematical model predicting anti-hepatitis B virus surface antigen (HBs) decay after vaccination against hepatitis B. *Clin. Exp. Immunol.* **116**(1), 121–126 (1999)
13. Jost, A.: Die Assoziationsfestigkeit in ihrer Abhängigkeit von der Verteilung der Wiederholungen. Leopold Voss, Leipzig (1897)
14. Mackay, F., Browning, J.L.: Turning off follicular dendritic cells. *Nature* **395**(6697), 26–27 (1998)
15. Marrack, P., Scott-Browne, J., MacLeod, M.K.: Terminating the immune response. *Immunol. Rev.* **236**, 5–10 (2010)
16. Mayer, A., Zhang, Y., Perelson, A.S., Wingreen, N.S.: Regulation of T cell expansion by antigen presentation dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **116**(13), 5914–919 (2019)
17. Myung, I.J., Kim, C., Pitt, M.A.: Toward an explanation of the power law artifact: insights from response surface analysis. *Mem. Cognit.* **28**(5), 832–840 (2000)

18. Quiel, J., Caucheteux, S., Laurence, A., Singh, N.J., Bocharov, G., Ben-Sasson, S.Z., Grossman, Z., Paul, W.E.: Antigen-stimulated CD4 T-cell expansion is inversely and log-linearly related to precursor number. *Proc. Natl. Acad. Sci. U.S.A.* **108**(8), 3312–3317 (2011)
19. Reed, W.J., Hughes, B.D.: From gene families and genera to incomes and internet file sizes: why power laws are so common in nature. *Phys. Rev. E* **66**(6), 067103 (2002). 4 pp.
20. Savageau, M.A.: Parameter sensitivity as a criterion for evaluating and comparing the performance of biochemical systems. *Nature* **229**(5286), 542–544 (1971)
21. Southern, J., McVernon, J., Gelb, D., Andrews, N., Morris, R., Crowley-Luke, A., Goldblatt, D., Miller, E.: Immunogenicity of a fourth dose of Haemophilus Influenzae type B (Hib) conjugate vaccine and antibody persistence in young children from the United Kingdom who were primed with acellular or whole-cell pertussis component-containing Hib combinations in infancy. *Clin. Vaccine Immunol.* **14**(10), 1328–1333 (2007)
22. Swart, E., van Gageldonk, P., de Melker, H., van der Klis, F., Berbers, G., Mollema, L.: Long-term protection against diphtheria in the Netherlands after 50 years of vaccination: Results from a seroepidemiological study. *PloS One* **11**(2), e0148605 (2016)
23. Tam, H.H., Melo, M.B., Kang, M., Pelet, J.M., Ruda, V.M., Foley, M.H., Hu, J.K., Kumari, S., Crampton, J., Baldeon, A.D., Sanders, R.W., Moore, J.P., Crotty, S., Langer, R., Anderson, D.G., Chakraborty, A.K., Irvine, D.J.: Sustained antigen availability during germinal center initiation enhances antibody responses to vaccination. *Proc. Natl. Acad. Sci. U.S.A.* **113**(43), E6639–E6648 (2016)
24. Tew, J.G., Kosco, M.H., Burton, G.F., Szakal, A.K.: Follicular dendritic cells as accessory cells. *Immunol. Rev.* **117**(1), 185–211 (1990)
25. Wickelgren, W.A.: Trace resistance and the decay of long-term memory. *J. Math. Psychol.* **9**(4), 418–455 (1972)
26. Wixted, J.T.: On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychol. Rev.* **111**(4), 864–879 (2004)



# Swarm Intelligence and Swarm Robotics in the Path Planning Problem

Quoc Bao Diep<sup>(✉)</sup>, Thanh Cong Truong, and Ivan Zelinka

Faculty of Electrical Engineering and Computer Science, VSB - Technical University of Ostrava, 17. Listopadu 15, Ostrava, Czech Republic  
diepquocbao@gmail.com, {cong.thanh.truong.st,ivan.zelinka}@vsb.cz

**Abstract.** In this chapter, we introduce the basic characteristics of swarm intelligence, the path planning problem for robots, and how to apply the self-organizing migrating algorithm, a representative of swarm intelligence to solve that real-world problem. We set up simulations in the Matlab environment with four common possible scenarios to demonstrate the effectiveness of the solution.

**Keywords:** Self-organizing migrating algorithm · SOMA · Path planning · Swarm intelligence

**AMS(2020) Subject Classification:** Primary 68T40 · Secondary 93C85

## 1 Introduction

Along with the development of artificial intelligence, swarm intelligence (SI) increasingly prove its important role, participating in most of the real-world technical problems. SI is derived from the observation of the intelligent behavior of creatures to form algorithms that solve complex problems with simple rules. The popular SI algorithms can be mentioned as particle swarm optimization [9], artificial bee colony [8], firefly algorithm [12], ant colony optimization [5], especially the self-organizing migrating algorithm [11, 13] that will be focused on in this chapter.

These algorithms have been applied to solve complex problems in many fields, such as analysis of the performance of the fish school search algorithm running in graphic processing units [10], training the radial basis function network for data classification and disease diagnosis [7], adaptive routing in telecommunications networks [6], resource allocation scheme for 5G C-RAN [1] and task scheduling in cloud-based internet of things applications [3].

But what kind of problems can apply SI algorithm to solve? And how to solve them? Most problems arise in practice that requires optimal solutions, which are minimum or maximum values, or solutions that satisfy some constraint. Accordingly, the optimization problems are the objects to be solved by the SI

algorithms, and one of the most important things to do is modeling the given problem into a mathematical model described by equations.

This chapter presents how to model the path planning problem for robots avoiding detected obstacles towards the target by applying the self-organizing migrating algorithm (SOMA), a representative of the SI. Section 2 presents the main concept of the SI and introduces the SOMA algorithm. Section 3 deals with the problem of path planning for robots. Details of the simulated settings in this research are presented in Sect. 4. Section 5 shows the simulations that prove the correctness of the solution. Finally, we conclude in Sect. 6.

## 2 Swarm Intelligence

### 2.1 General Concept

Swarm intelligence (SI) is a common name referring to the algorithms that operate on the mechanism simulating the collective intelligence behaviors of the creatures. It works on a population (or some sub-populations) of many individuals that interact with each other (both competing and cooperating) or with the environment (migration and survival) to solve specific problems such as foraging, protecting the nest or moving safely in the natural habitat.

Slightly different from the evolutionary algorithms, which operate on Darwin's theory of evolution, individuals in the SI population do not inherit the genetic properties from generation to generation, but rather will share the knowledge with each other in the same generation under loops. This sharing of information is the key for the SI algorithm to find the global optimal solution to the given problems.

A flock of birds, for example, is searching for food in space, and one individual alerts the remaining members to fly towards its cry (sharing information) when it finds the food. On the way the others move to that food source, they can find a more abundant source than the previous signal, they will share it again. And that process is repeated until the whole flock meets together on where the most food source is.

Inspired by those observations, the SI algorithms are designed to mimic these behaviors. The next subsection will present the SOMA algorithm, a representative of the SI algorithm.

### 2.2 Self-Organizing Migrating Algorithm

The SOMA algorithm was first introduced in [14]. It bore all the characteristics of the SI that we will analyze below.

The first operation of SOMA is to create an initial population containing a predetermined number of individuals in a given search space, representing the natural habitat in the foraging bird example above. These individuals are the solutions to the optimization problem that have been encoded.

The population is then evaluated by the cost function (will be presented in the next section), and the fitness value represents the amount of food as in the

above example. The best individual is selected, and the remaining individuals move towards that member. They will likely find better positions on the path they move. At the end of each such migration, a new best individual is selected again and the process continues until the algorithm has found a solution that satisfies the given requirement [11, 13].

In the problem to be addressed in this chapter, SOMA plays the role of generating a dynamic set of next stops for the robot in real-time. At a specific time, based on the necessary information such as the current position of the robot, the position of the target, and the obstacles detected by sensors, SOMA calculates the next best position that the robot should move to. These positions are generated in real-time and become the moving path for the robot.

To execute that description, the algorithm first initializes a random group of individuals around the current position of the robot (in the searching range) based on Eq. (1). Then the best position is selected after evaluating all individuals (named Leader), and the algorithm goes into the first migration loop.

$$Po_{individual_{ith}} = Po_{actual} + rnd_{-1 \rightarrow 1} ro_{range}, \quad (1)$$

where:

- $Po_{individual_{ith}}$ : position of the  $i^{th}$  individual,
- $Po_{actual}$ : actual position of robot,
- $ro_{range}$ : maximum moving range of robot,
- $rnd_{-1 \rightarrow 1}$ : uniformly distributed random number from  $-1$  to  $1$ .

During this loop, the remaining members will one-by-one move towards the Leader using the rule given in Eq. 2.

$$Po_{new} = Po_{current} + (Po_{leader} - Po_{current}) n PRTVector_j \quad (2)$$

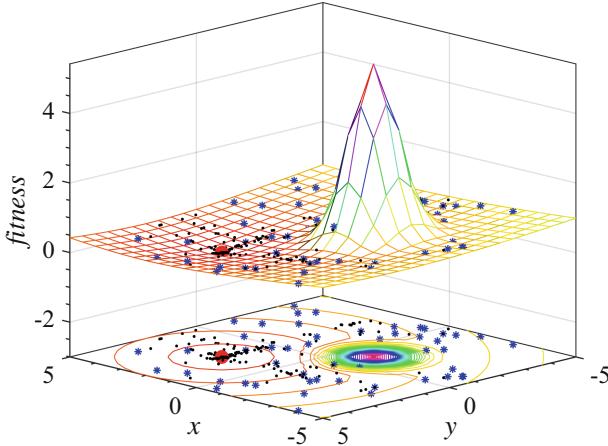
where:

- $Po_{new}$ : the new position of the current individual,
- $Po_{current}$ : the current position of the current individual,
- $Po_{leader}$ : the Leader position in this migration loop,
- $PRTVector_j$ : the perturbatively factor, created by Eq. 3,
- $n$ : moving step, from 0 by  $Step$  to  $PathLength$ .

$$\text{if } rnd_{0 \rightarrow 1} < PRT; PRTVector_j = 1; \text{ else, } PRTVector_j = 0. \quad (3)$$

After each individual completes its move, the best position on its path is selected to be compared with the initial. It will replace the initial position if it has a better fitness value. When the last member completes its job, a new best individual throughout whole the population is then selected again to replace the old Leader and a new migration loop begins.

Those processes are terminated when the entire population has achieved a given number of migrations. And the final Leader is the position where the robot



**Fig. 1.** The migration of individuals in a population of SOMA when the robot detects obstacles.

will move. Figure 1 depicts the principle of the SOMA algorithm, where the hill represents an obstacle, the blue points represent the initial individuals, the black points represent the locations after the migration of the initial population, the red point is the globally optimal location where the robot will move to.

The next section describes how to build the fitness function.

### 3 The Path Planning Problem of Swarm Robots

For any optimization problem, the fitness function is an important component, is the object to solve. In some situations, the fitness function is already given. But in some cases, we have to build the fitness function by modeling that problem. In this section, we present how to turn the robot path planning problem into a fitness function.

Starting with a simple rule, the robot is as close to the target and as far away from the obstacles as possible. Equation 4 generally describes the elements  $X$  stated, where  $n$  is the number of the target and obstacles detected by the sensors. The goal is to minimize the function  $f(X)$ .

$$f(X) = \sum_{i=1}^n X_i \quad (4)$$

For the principle as close to the target as possible, we see that the value of the function  $f(X)$  should be proportional to the distance from the robot to the target. Equation 5 constructs the first element of the fitness function in detail, where  $(x_{robot}, y_{robot})$  and  $(x_{target}, y_{target})$  are the current positions of the robot and target respectively, and  $a_1$  is the equilibrium coefficient.

$$X_1 = a_1 \sqrt{(x_{target} - x_{robot})^2 + (y_{target} - y_{robot})^2} \quad (5)$$

Similarly, with the rule that as far as possible from obstacles, the value of the  $f_{(X)}$  function should be inversely proportional to the distance from the robot to detected obstacles. Equation 6 describes this in detail.

$$X_i = a_i \sum_0^{n_{obstacle}} e^{-(c - r_{obstacle}) dis_{obstacle}} \quad (6)$$

where:

$$dis_{obstacle} = \sqrt{(x_{obstacle} - x_{robot})^2 + (y_{obstacle} - y_{robot})^2}$$

- $X_i$ : the obstacle elements of the  $f_{(X)}$ ,
- $a_i$ : the equilibrium coefficient,
- $n_{obstacle}$ : the number of detected obstacles,
- $c$ : the influential coefficient of obstacles,
- $r_{obstacle}$ : the radius of detected obstacles,
- $dis_{obstacle}$ : the distance from the robot to detected obstacles.

In the framework of this chapter, we do not go into details about robot kinematics and dynamics. We assume that the robot can move smoothly from point A to a nearby point without any problem, and the SOMA will generate the dynamic set of that points [2].

Due to the robot's physical limitations, the maximum distance between the two points mentioned is limited, named  $d_{limit}$ . However, no matter how big the  $d_{limit}$  is, the algorithm quality is completely independent of this distance.

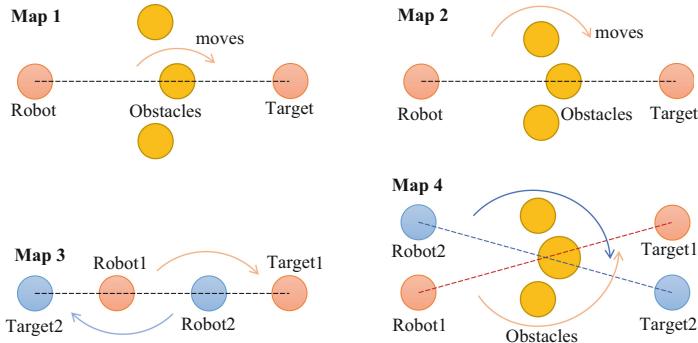
## 4 Experiment Setup

To rigorously evaluate the feasibility of the proposed solution, we built 4 selective scenarios, covering most of the basic situations that can occur in the real-world.

### 4.1 Selective Scenarios

The first scenario is the simplest one, having a robot (with a respective target) and three static obstacles. The location of the obstacles is intentionally arranged so that they are symmetrical and centered on the line connecting the robot to the target. The gap between the three obstacles is calculated wide enough for the robot to move through them. This scenario is set up to test the ability of the robot to pass through sufficient gaps between obstacles (see Map 1 of Fig. 2).

The second scenario is similar to the first but the distance between the obstacles has been changed so that they are smaller than the physical size of the robot (it cannot move through those gaps). The aim is to trap the robot into the local minima and observe how to escape from the trapped area of the robot (Map 2 of Fig. 2).



**Fig. 2.** Selective scenarios to test the operability of robots.

In the third scenario, two robots with two respective targets were set. There is no obstacle in this map, but robots will be obstacles to each other. All of them were intentionally put in a straight line so that the robots will move in opposite directions. This situation tests the possibility of mutual avoidance between robots (Map 3 of Fig. 2).

The last one is the most complex scenario. Two robots, two respective targets, and three obstacles were used. The robots are on the same side of the obstacles, and the targets are on the opposite side but diagonally. The obstacles located in the middle are not only to prevent the movement of the robots, but also trap the robot to the local minimum associated with the other robot. This scenario tests the generality of the proposed algorithm (Map 4 of Fig. 2).

**Table 1.** Locations of obstacles and robots in Cartesian coordinates - Map 1 and 2 (in decimeter)

The object	Obstacle 1	Obstacle 2	Obstacle 3	Robot	Target
$x_{map_1}$	04	11	13	01	20
$y_{map_1}$	11	04	13	01	20
$r_{map_1}$	03	03	03	-	-
$x_{map_2}$	06	13	11	01	20
$y_{map_2}$	13	06	11	01	20
$r_{map_2}$	02	02	02	-	-

The detailed locations of robots, obstacles, and targets are shown in Tables 1, 2, and 3.

**Table 2.** Locations of robots in Cartesian coordinates - Map 3 (in *decimeter*)

The object	Robot 1	Robot 2	Target 1	Target 2
$x_{map_3}$	06	17	20	03
$y_{map_3}$	06	17	20	03

**Table 3.** Locations of obstacles and robots in Cartesian coordinates - Map 4 (in *decimeter*)

The object	Obs 1	Obs 2	Obs 3	Ro 1	Ro 2	Tar 1	Tar 2
$x_{map_4}$	07	14	11	05	02	17	22
$y_{map_4}$	14	07	11	02	05	22	17
$r_{map_4}$	02	02	03	–	–	–	–

## 4.2 Control Parameters

The objects were drawn using Matlab software R2020b version in Windows 10 Pro Edition 20H2 Version. The SOMA for each robot is also programmed using Matlab. The control parameters of the algorithm are given in Table 4.

Popsize = 40; Migration = 20: for 2-Dimensional problem and the objective function to solve is not too complicated, those values are suitable.

PRT = 0.1, Step = 0.11, Pathlength = 3: These options are common to the SOMA algorithm, and it is selected based on the recommendation from the original paper [11, 13].

All robots used in simulations have a radius  $r_{robot} = 0.8$  dm. The sensors have a radius of active range  $r_{sensor} = r_{robot} + 2.8$  dm. The maximum step of the robots is  $d_{limit} = 0.4$  dm.

## 5 Simulation Results

The simulation results are presented in the form of selected figures captured from the robot's movement in form of 2D and 3D.

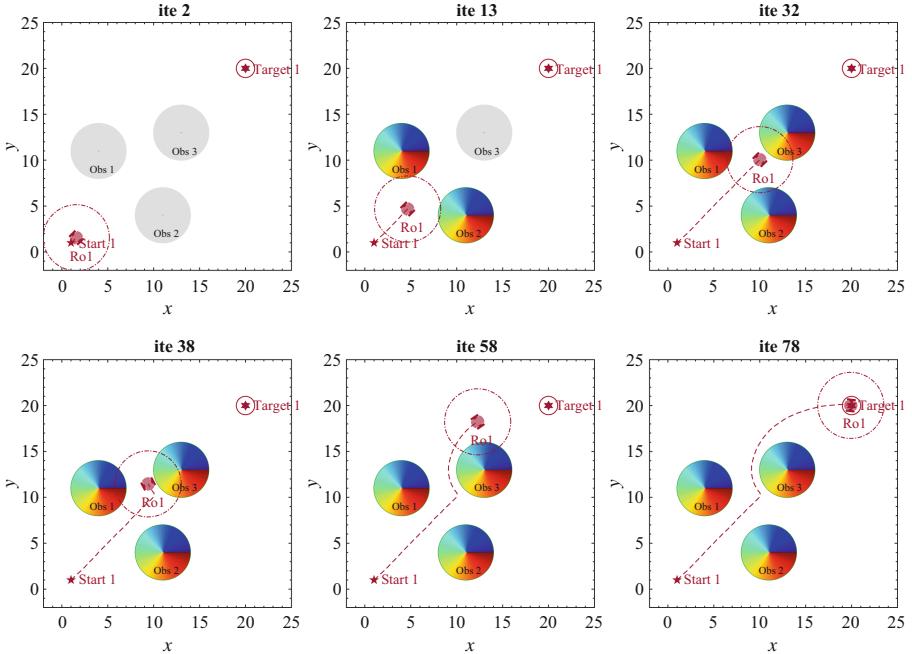
In those 2D figures, robots are plotted on Cartesian coordinates with obstacles and targets also. The circle around the robot represents the working range of the sensors. Obstacles are drawn in a dark color circle. As the robot moves, obstacles detected in the sensor's active area will be represented by bright colors. These obstacles will turn bright colors when they are detected by sensors located on the robot.

**Table 4.** The control parameter values of SOMA.

Migration	PopSize	Step	PRT	PathLength
20	40	0.11	0.1	3.0

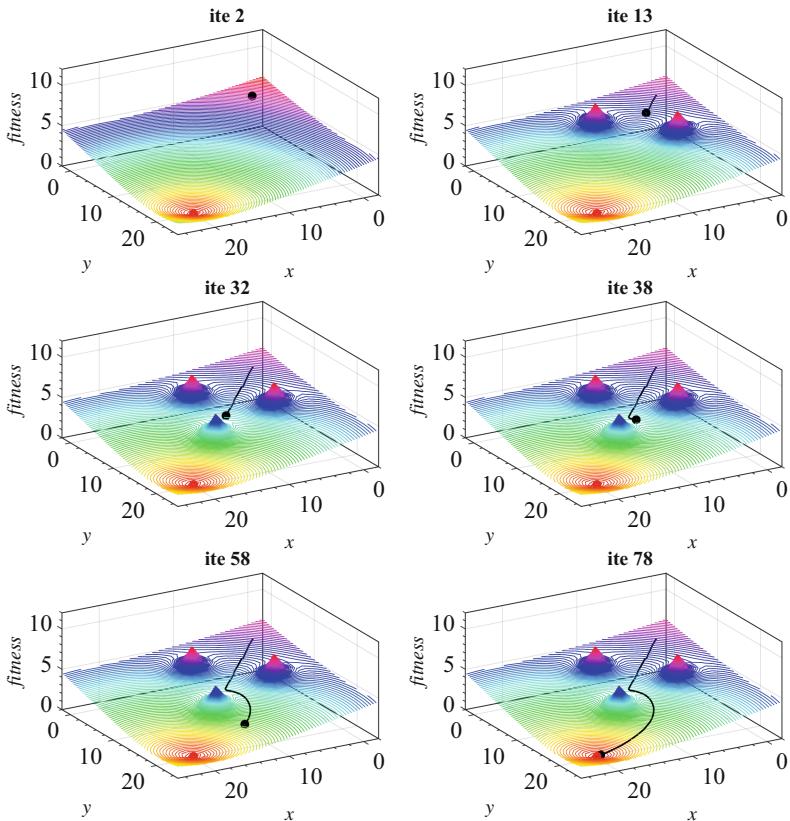
In 3D figures, the robot is represented as a big black dot, and the robot's path is represented by small black dots. Contour lines represent the surrounding environment, they can change depending on the distance from the robot to the target and the obstacles.

### 5.1 Results for Map 1



**Fig. 3.** The movement process of the robot in Map 1: move through the gaps between obstacles to hit the target.

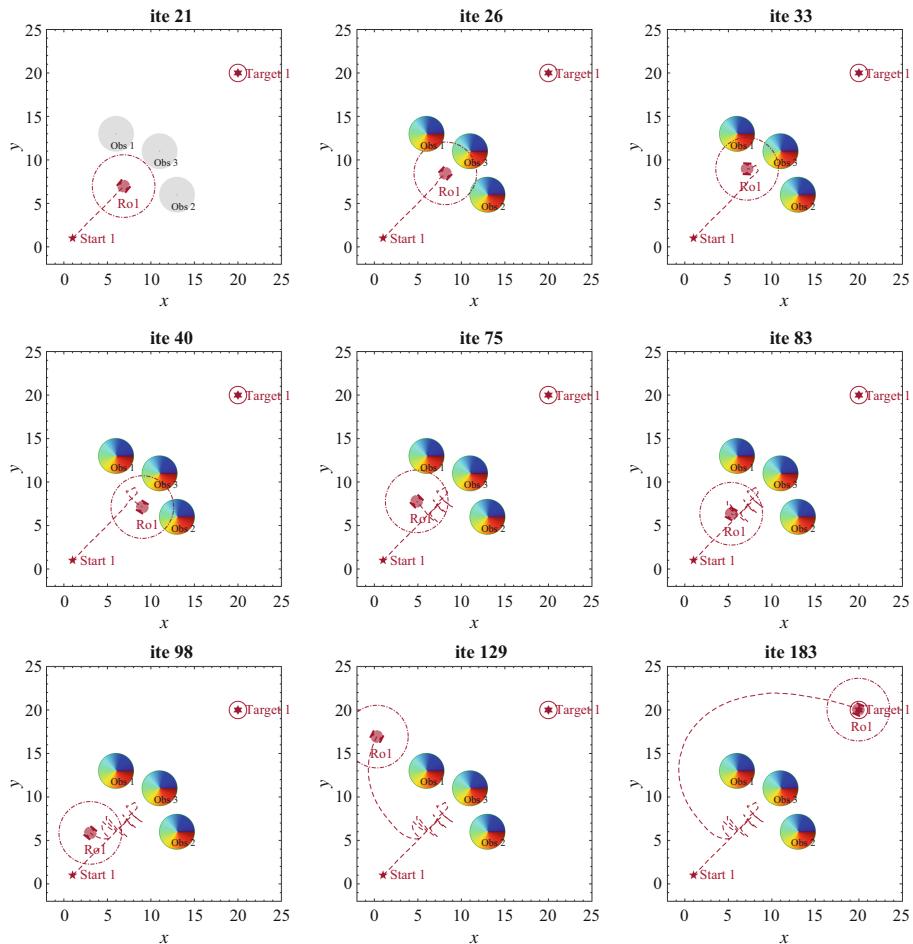
Figures 3 and 4 show the robot's movement in 2D and 3D, respectively. They were captured at the step of  $2^{nd}$ ,  $13^{th}$ ,  $32^{nd}$ ,  $38^{th}$ ,  $58^{th}$ , and  $78^{th}$ . At the  $2^{nd}$  step in Fig. 3, the robot has not detected the obstacles yet so they are in a dark color, and the robot tends to move straight towards the target. At this moment, the contours on the 3D map of Fig. 4 are also “flat” (without hills).



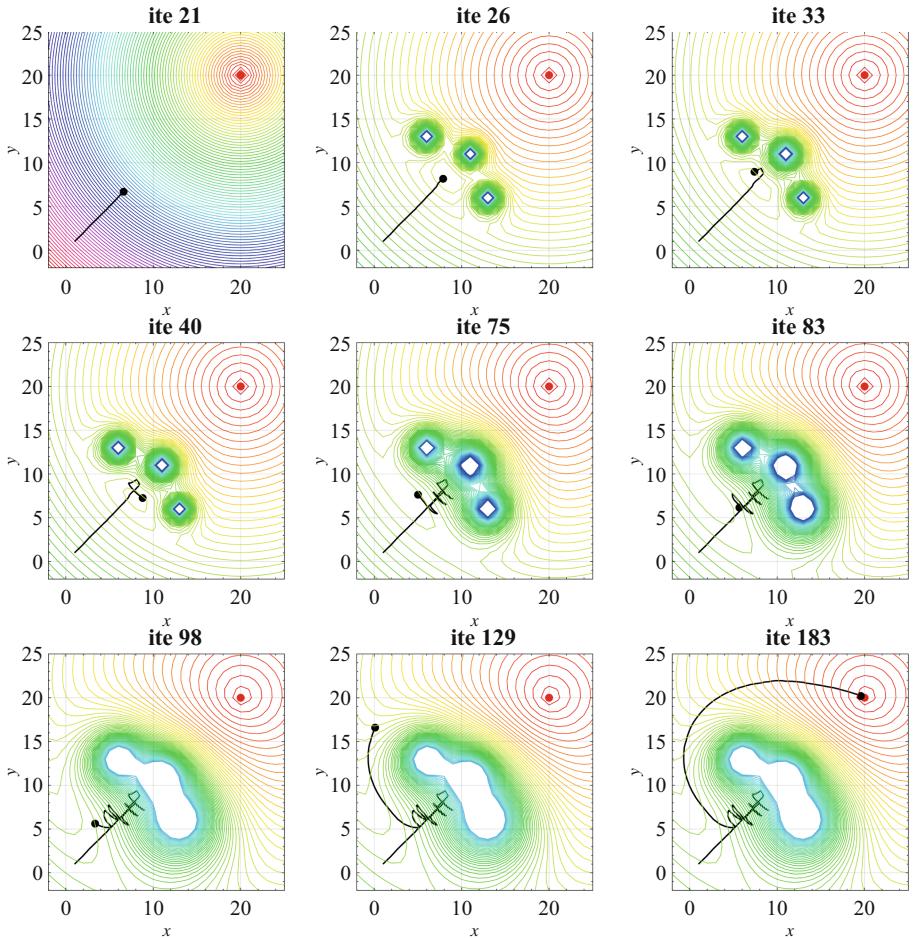
**Fig. 4.** The movement process of the robot in Map 1 presented in 3D.

However, in steps 13<sup>th</sup> and 32<sup>nd</sup>, the obstacles are detected, and they have changed color, hills appear respectively in the 3D contour maps. The robot will move along these contour lines from high to low and avoid colliding on the rising hills (which are obstacles).

In the simple situation of Map 1, the distance between obstacles is large enough for the robot to pass, so the robot has no trapped between three obstacles. The robot takes 78 steps to hit its target on this Map.



**Fig. 5.** The movement process of the robot in Map 2: cannot move through the gaps, escape from the trap to hit the target.



**Fig. 6.** The movement process of the robot in Map 2 presented in contour map.

## 5.2 Results for Map 2

Scenario 2 is intentionally arranged so that the distance between obstacles is not enough for the robot to move through. In this situation, the robot will be trapped between obstacles and will not be able to move out of the trap zone. To solve this problem, the equilibrium coefficient will change the value leading to a change the width of the hill accordingly, thereby escaping the robot from the trapped area [4].

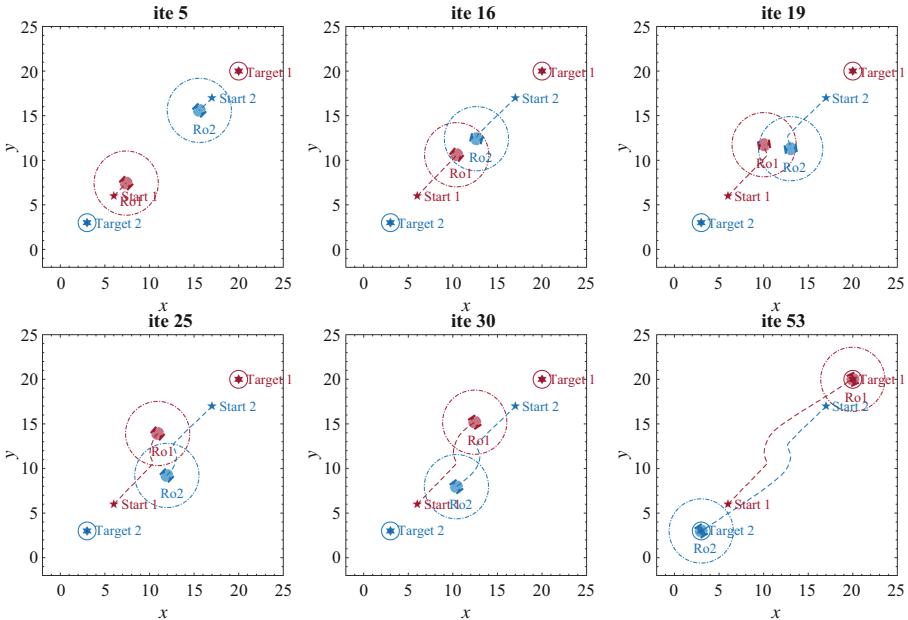
Figures 5 and 6 show the entire operation of the robot, captured at steps 21<sup>st</sup>, 26<sup>th</sup>, 33<sup>rd</sup>, 40<sup>th</sup>, 75<sup>th</sup>, 83<sup>rd</sup>, 98<sup>th</sup>, 129<sup>th</sup>, and 183<sup>rd</sup>.

In step 21<sup>st</sup>, similar to map 1, the robot has not detected any obstacles so it moves straight to the target. However, in step 26<sup>th</sup>, all three obstacles were detected, at which time the robot was trapped in the contour as shown in steps

$26^{th}$  and  $33^{rd}$ . As mentioned above, the equilibrium coefficients start to change, resulting in the size of the hills growing up, the contour changing continuously. The robot follows these contour lines, shown in steps  $40^{th}$  to  $98^{th}$ , and exits the trap towards the target, shown in steps  $129^{th}$ , and  $183^{rd}$ .

The robot took 183 steps in this scenario to escape the trap and hit the target.

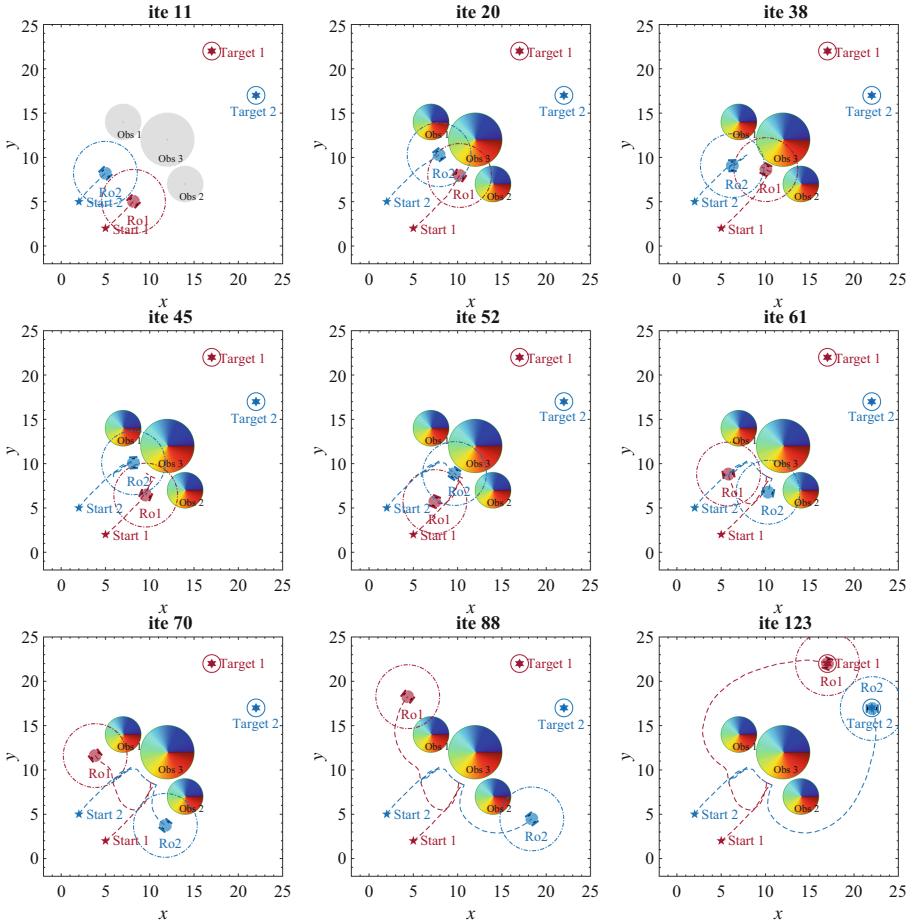
### 5.3 Results for Map 3



**Fig. 7.** The movement process of the robot in Map 3: face-to-face between two robots.

Different from map 1 and map 2, map 3 has two robots and two targets respectively. There are no obstacles on this map. Instead, the positions of the robots and the targets are intentionally arranged so that they are each other's obstacles.

Figure 7 shows the movement of two robots, captured at steps  $5^{th}$ ,  $16^{th}$ ,  $19^{th}$ ,  $25^{th}$ ,  $30^{th}$ , and  $53^{rd}$ . At step  $5^{th}$ , the robots have not detected the other robot yet so they move straight to their target. But in step  $16^{th}$ , both robots are in the detection range of sensors, and they avoid each other as shown in steps  $19^{th}$  to  $30^{th}$ . Finally, they finish their work on step  $53^{rd}$ .



**Fig. 8.** The movement process of the robot in Map 4: a complex combination in a single scenario.

#### 5.4 Results for Map 4

The last scenario is the most complex one to test the general operability of robots. Two robots, two respective targets, and three obstacles are present on this map. They are arranged so that robots will be stuck between obstacles and the remaining robot will be another obstacle, moving around, preventing each other's path.

Figure 8 reveals this operation process, captured at steps of 11<sup>th</sup>, 20<sup>th</sup>, 38<sup>th</sup>, 45<sup>th</sup>, 52<sup>nd</sup>, 61<sup>st</sup>, 70<sup>th</sup>, 88<sup>th</sup>, and 123<sup>rd</sup>.

At step 11<sup>th</sup>, three obstacles have not been detected and two robots are not in each other's path so they move towards the targets. However, in step 20<sup>th</sup>, three obstacles are blocking the way of both robots. Furthermore, the remaining

robot now becomes the fourth obstacle preventing the other robot's path. In step 38<sup>th</sup>, Robot 2 turned its head, moved backward to find a way out of the trap zone, and Robot 1 moved along obstacles.

In steps 45<sup>th</sup> to 61<sup>st</sup>, the robots move around in the trap to find a way to escape, and they start out of the trap in step 70<sup>th</sup>. Once out of the trap zone, there are no obstacles left, so the robots approach their target without any problem, as shown in step 88<sup>th</sup>. They hit the final target at step 123<sup>rd</sup>.

## 6 Conclusions

In this chapter, we have introduced a common practical application that is using the self-organizing migrating algorithm to plan the path for the robot in real-time. For this problem, SOMA plays a role in generating a dynamic set of moving points from the starting position to the target that the robot must pass through. When obstacles are detected by sensors, the inversely proportional components appear in the fitness function and the algorithm will search a next point that satisfies both the criteria of avoiding obstacles and towards the target. The limitations of the solution such as the parameters in the model are fine-tuned by experience will be overcome in the next studies.

**Acknowledgement.** The following grants are acknowledged for the financial support provided for this research: Grant of SGS No. SP2020/78, VSB-Technical University of Ostrava.

## References

1. Ari, A.A.A., Gueroui, A., Titouna, C., Thiare, O., Aliouat, Z.: Resource allocation scheme for 5G C-RAN: a swarm intelligence based approach. *Comp. Netw.* **165**, 106957 (2019)
2. Bao, D.Q., Zelinka, I.: Obstacle avoidance for swarm robot based on self-organizing migrating algorithm. *Procedia Comput. Sci.* **150**, 425–432 (2019)
3. Boveiri, H.R., Khayami, R., Elhoseny, M., Gunasekaran, M.: An efficient swarm-intelligence approach for task scheduling in cloud-based internet of things applications. *J. Ambient Intell. Humaniz. Comput.* **10**(9), 3469–3479 (2019)
4. Diep, Q.B., Zelinka, I., Senkerik, R.: An algorithm for swarm robot to avoid multiple dynamic obstacles and to catch the moving target. In: International Conference on Artificial Intelligence and Soft Computing, pp. 666–675. Springer, Cham (2019)
5. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. *EEE Comput. Intell. Mag.* **1**(4), 28–39 (2006)
6. Ducatelle, F., Di Caro, G.A., Gambardella, L.M.: Principles and applications of swarm intelligence for adaptive routing in telecommunications networks. *Swarm Intell.* **4**(3), 173–198 (2010)
7. Horng, M.H., Lee, Y.X., Lee, M.C., Liou, R.J.: Firefly metaheuristic algorithm for training the radial basis function network for data classification and disease diagnosis. *Theory New Appl. Swarm Intell.* **4**(7), 115–132 (2012)
8. Karaboga, D., Basturk, B.: On the performance of artificial bee colony (ABC) algorithm. *Appl. Soft Comput.* **8**(1), 687–697 (2008)

9. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN 1995 International Conference on Neural Networks, vol. 4, pp. 1942–1948. IEEE (1995)
10. Lins, A., Bastos-Filho, C.J., Nascimento, D.N., Junior, M.A.O., de Lima-Neto, F.B.: Analysis of the performance of the fish school search algorithm running in graphic processing units. In: Parpinelli, R., Lopes, H.S. (eds.) Theory and New Applications of Swarm Intelligence, pp. 17–32. IntechOpen (2012). <https://doi.org/10.5772/30360>
11. Pluhacek, M., Zelinka, I., Senkerik, R., Davendra, D.: Inspired in SOMA: perturbation vector embedded into the chaotic PSO algorithm driven by Lozi chaotic map. In: Davendra, D., Zelinka, I. (eds.) New Optimization Techniques in Engineering. Studies in Computational Intelligence. Springer (2016)
12. Yang, X.S.: Firefly algorithms for multimodal optimization. In: International Symposium on Stochastic Algorithms, pp. 169–178. Springer, Heidelberg (2009)
13. Zelinka, I.: SOMA - self-organizing migrating algorithm. In: Onwubolu, G.C., Babu, B.V. (eds.) New Optimization Techniques in Engineering, pp. 167–217. Springer, Heidelberg (2004)
14. Zelinka, I., Lampinen, J.: SOMA – self-organizing migrating algorithm. In: MENDEL – 6th International Conference on Soft Computing, Brno, Czech Republic (2000)



# Utilizing Differential Evolution into Optimizing Targeted Cancer Treatments

Michail-Antisthenis Tsompanas<sup>1(✉)</sup>, Larry Bull<sup>1</sup>, Andrew Adamatzky<sup>1</sup>, and Igor Balaz<sup>2</sup>

<sup>1</sup> Unconventional Computing Laboratory, Department of Computer Science and Creative Technologies, University of the West of England, Bristol, UK  
`{antisthenis.tsompanas,larry.bull,andy.adamatzky}@uwe.ac.uk`

<sup>2</sup> Laboratory for Meteorology, Physics and Biophysics, Faculty of Agriculture, University of Novi Sad, Novi Sad, Serbia  
`igor.balaz@df.uns.ac.rs`

**Abstract.** Working towards the development of an evolvable cancer treatment simulator, the investigation of including evolutionary optimization methods was considered. Namely, Differential Evolution (DE) is studied here, motivated by the high efficiency of variations of this technique in real-valued problems. A basic DE algorithm, namely “DE/rand/1” was used to optimize *in silico* the design of a targeted drug delivery system (DDS) for tumor treatment on PhysiCell simulator. The suggested approach proved to be more efficient than a standard Genetic Algorithm (GA), which was not able to escape local minima after a predefined number of generations. The key attribute of DE that enables it to outperform standard GAs, is the fact that it keeps the diversity of the population high, throughout all the generations. This work will be incorporated with ongoing research in a more wide applicability platform that will design, develop and evaluate targeted DDSs aiming cancer tumours.

**Keywords:** Differential evolution · Cancer treatment · Evolutionary algorithm · PhysiCell simulator · Optimization

**AMS(2020) subject classification:** Primary 68W50 · Secondary 92-08 · 90C27

## 1 Introduction

The vast diversity of cell types discovered in cancerous tumours [1, 22] and their ability to resist conventional treatment due the existence of subclonal populations [2, 15], is motivating more complex treatment options. First steps into using multitarget, multistage and multicomponent nanoparticles are promising [17] and need to be further investigated as these techniques may hold the key to effective cancer treatments. Consequently, building computational tools that could discover the optimum design parameters of a treatment through efficiently

explore and exploit large parameter search spaces, are of paramount importance. In accordance with this concept, the evolutionary *in silico* optimization of a targeted DDS was investigated utilizing a robust evolutionary algorithm (EA).

DE gained popularity over other well-established EAs, as it follows similar algorithmic steps with standard EAs, but was able to surpass them in terms of efficiency [16]. There have been several proposals on how to enhance its performance [8, 18] and these alternative algorithmic approaches, building on the initial methodology, were tested in real problems, as well as numerical benchmark problems [5, 6, 14, 23].

DE was initially proposed in [16], as an effective methodology of optimization over continuous spaces of nonlinear and non differentiable functions. The method was introduced as an alternative to previous direct search approaches, aiming at three main objectives: the ability to find the global optimum, the fast convergence to this optimum and the need of a small amount of control parameters for the procedure. Moreover, it was developed to be easily implemented in parallel computing platforms. The methodology is population based, where for every generation the new individuals are produced after applying the scaled difference (hence the name differential evolution) of some predefined individuals to another predefined base individual.

The ability to easily extract attributes of the population, such as the distance of its members and their directions, through the aforementioned methodology, is what makes DE so powerful. This characteristic is defined as self-referential mutation [13]. Given the aforementioned advantages of DE compared with other EAs, it was chosen to investigate the optimization of a targeted DDS on a cancer tumour, when simulated with PhysiCell [10].

PhysiCell [10] is a multicellular, agent-based simulator that was designed to extend the BioFVM [9] framework, to form a virtual laboratory. PhysiCell is open source and offers several sample projects, one of which is the one studied in this study. More specifically, sample project “anti-cancer biorobots” [10] was developed as a possible tool to investigate the targeted cancer treatment, i.e. with drugs transported by specialized nanoparticles that would target specific cells of the cancer tumours.

Previously, PhysiCell was deployed as a virtual laboratory in the optimization process of the design of nanoparticle carriers of cancer treating compounds [12, 19–21] and the process of mapping immunotherapies [11]. More specifically, the use of PhysiCell led the training of surrogate-assisted evolutionary algorithms exploring the efficient solutions of the design of nanoparticle-based drug delivery systems for cancer [12]. A similar application of PhysiCell was examined in [19], where the optimization was achieved by a novel memetic algorithm inspired by the fundamental haploid-diploid lifecycle of eukaryotic organisms [3, 4]. Moreover, active learning and genetic algorithms incorporated with the PhysiCell simulator, enabled the efficient exploration of biological and clinical constraints for cancer immunotherapy [11].

## 2 Differential Evolution

The methodology of DE comprises of the following steps: initialization, mutation, recombination (or crossover) and selection. At first, a population of individuals (possible solutions) is formed by randomly picking values on the D-dimensional search space of the problem to be optimized. The random function used is mainly uniformly distributed to cover sufficiently the search space, which could be limited with lower and upper boundaries depending on the definition of the problem.

Then, the mutation operator is employed on the initial population. More specifically, for every individual in the population a new individual is generated (named the mutant individual) by a mathematical expression of the parameters of randomly chosen or predefined individuals. For example, the mutant individual ( $v_i$ ) can be the linear combination of three randomly selected individuals as defined by the following equation, where  $x_{r1}$ ,  $x_{r2}$  and  $x_{r3}$  are randomly selected individuals from the population and  $F$  is a scaling factor, which is a positive number.

$$v_i = x_{r1} + F \cdot (x_{r2} - x_{r3})$$

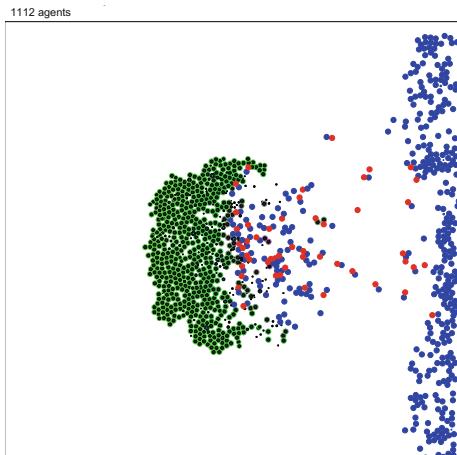
Note here, that the variations of DE methodology are defined with a notation in the form of “DE/base/num”. The “base” part of the notation refers to the technique of choosing the individual that will be used as the base individual to which the scaled difference will be added. The “num” part indicates the amount of pairs of individuals that will produce a scaled difference each, to be added on the base individual. Thus, the aforementioned DE variation is denoted as “DE/rand/1”.

The produced mutant individuals, then, undergo the crossover operator, in order to be recombined with individuals from the initial population. Each individual chosen from the initial population is denoted as the target individual and the produced individual after the crossover as the trial individual. Two crossover strategies are mainly used, namely exponential and binomial. However, binomial is dominant in the literature, where every parameter in the D-dimensional solution is treated separately to the others, and chosen from the mutant or the target individual based on a probability defined as the  $CR$  parameter. This parameter is known as the crossover rate.

The final step of the algorithm is then evaluation, by the use of the selection operator. Similar to standard GAs, the trial individuals produced from the previous step, are evaluated with the fitness function. If their fitness is better than the one of their target individual, they replace the target individual in the next generation. Otherwise the target individual is retained. When all trial individuals are tested and the appropriate selection is made to form the population of the next generation, the algorithm runs again the population through the mutation, crossover and selection operators, until the termination criteria are met (i.e. the computation budget).

### 3 Methodology

The sample project from PhysiCell [10] framework that was investigated, is the “anti-cancer biorobots”. In this simulation all entities are simulated as agents. Namely, there are three different types of agents with different functionalities, the cancer cells, the chemical compound (defined as cargo agents) and the functionalized nanoparticles (defined as worker agents). The outputs of PhysiCell include graphical representation of the agents in the simulated areas. An example after 10 days of simulated growth and treatment of a tumour is depicted in Fig. 1. The cancer cells are illustrated as green, the chemical compound as blue, while the nanoparticles as red agents.



**Fig. 1.** The graphical representation output of PhysiCell [10] after 10 days of simulated growth and treatment of a cancer tumour.

The optimization problem here was defined as the discovery of the design of nanoparticle agents in PhysiCell simulator (v.1.4.1) that will result in lower number of remaining cancer cells. Each possible solution is mapped in a 6-dimensional space where the parameters studied and their boundaries are the attached worker migration bias [0,1], the unattached worker migration bias [0,1], worker relative adhesion [0,10], worker relative repulsion [0,10], worker motility persistence time (min) [0,10] and the cargo release  $O_2$  threshold (mmHg) [0,20]. These parameters are selected as they dictate the way the simulated nanoparticle agents behave, namely the specifics of worker agent migration (deterministic for 1 and Brownian for 0) are determined by the parameters attached and unattached worker migration bias. The simulated time that each worker agent moves in a direction before changing to a new one is defined by the parameter worker motility persistence time. The behaviour of the worker agents in accordance with the cargo agents is described by the rest of the aforementioned parameters. For more details the reader can refer to [10].

Due to the stochastic nature of the simulator, each possible solution is evaluated after extracting the average value of the remaining cancer cells of 5 runs. Each run executes 7 days of growing an initial 200 micron radius tumor and 3 days of applying the treatment. The execution of the simulation on an Intel® Xeon® CPU E5-2650 at 2.20 GHz (using 8 of the 48 cores) and 64GB RAM was completed after 5 min.

For comparison reasons, a generic GA was applied to optimize the aforementioned problem. The parameters of the GA were chosen as population size  $P = 20$ , tournament size  $T = 2$  for selection and replacement operations, uniform crossover with probability  $X = 80\%$  and per allele mutation rate of  $\mu = 20\%$  with random step size of  $s = [-5, 5]\%$ .

On the other hand, the “DE/rand/1” strategy was implemented and tested for the optimization of a targeted DDS on a cancer tumour, by simulating this procedure with PhysiCell. The parameters of the DE algorithm were chosen as population size  $P = 20$ , scaling factor  $F = 0.5$  and crossover rate  $CR = 0.9$ . Note that these parameters are not fine-tuned to enhance the performance of the algorithm, but are most commonly used throughout the literature.

The rest of the parameters used to define the simulation of the tumour environment by PhysiCell were retained unchanged and assigned the same values as from the developers of the simulator [10]. More specifically these parameters are listed in Table 1.

Note here that the computational budget for one test of each evolutionary methodology was set to 1000 evaluations of the simulator. That translates to an evaluation of 200 possible solutions, because of the 5 run average used to partially alleviate the stochasticity effects of the simulator. Thus, having populations of 20 individuals, it results to 10 generations. Moreover, note that the total time required for an evolutionary test (1000 evaluations) reaches 5000 min or c. 3.5 days. Consequently, the comparison of the minuscule possible overhead of the DE, when compared to a GA, is not analysed in this application.

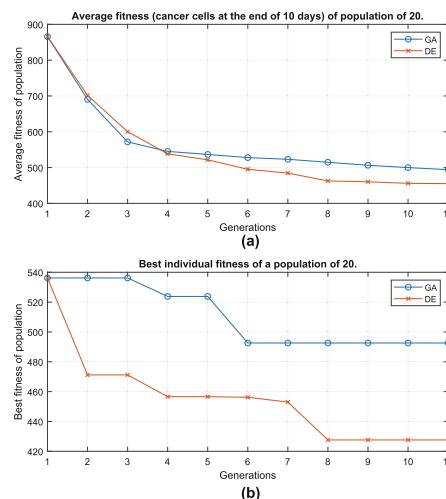
## 4 Results

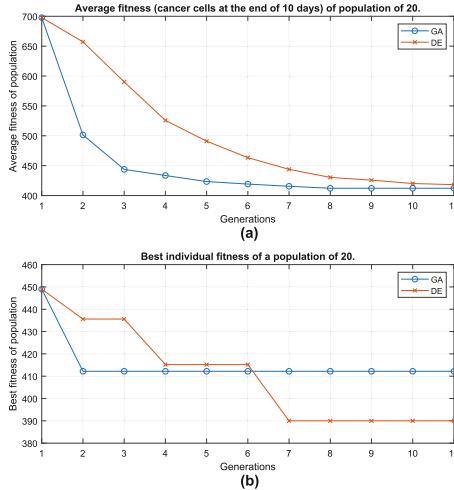
Three comparison tests were run, each using the same initial population for the GA and DE optimization. The results of the average fitness (remaining cancer cells after 10 days of simulation) of all individuals of the population through out the generations are depicted in Figs. 2, 3 and 4(a). Furthermore, the fitness of the best individual for both GA and DE throughout the generations was illustrated in Figs. 2, 3 and 4(b).

In Fig. 2(a) it is apparent that both approaches force the population to converge towards a lower fitness. However, in Fig. 2(b) the fact that DE outperforms the GA in finding better solutions from the first few generations is shown. Moreover, the GA seems to be stuck from the sixth generation in a local minimum.

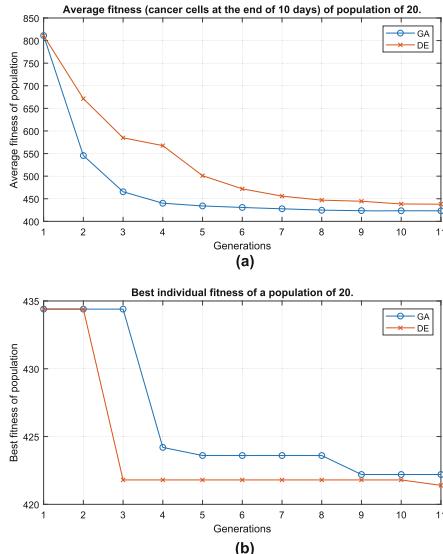
**Table 1.** Unaltered parameters of PhysiCell simulator.

Parameter	Value
Damage rate	0.03333 min <sup>-1</sup>
Repair rate	0.004167 min <sup>-1</sup>
Drug death rate	0.004167 min <sup>-1</sup>
Elastic coefficient	0.05 min <sup>-1</sup>
Cargo $O_2$ relative uptake	0.1 min <sup>-1</sup>
Cargo apoptosis rate	4.065e-5 min <sup>-1</sup>
Cargo relative adhesion	0
Cargo relative repulsion	5
Maximum relative cell adhesion distance	1.25
Maximum elastic displacement	50 μm
Maximum attachment distance	18 μm
Minimum attachment distance	14 μm
Motility shutdown detection threshold	0.001
Attachment receptor threshold	0.1
Worker migration speed	2 μm/min
Worker apoptosis rate	0 min <sup>-1</sup>
Worker $O_2$ relative uptake	0.1 min <sup>-1</sup>

**Fig. 2.** Average (a) and best fitness (b) of the populations evolved with GA and DE in the first comparison run.

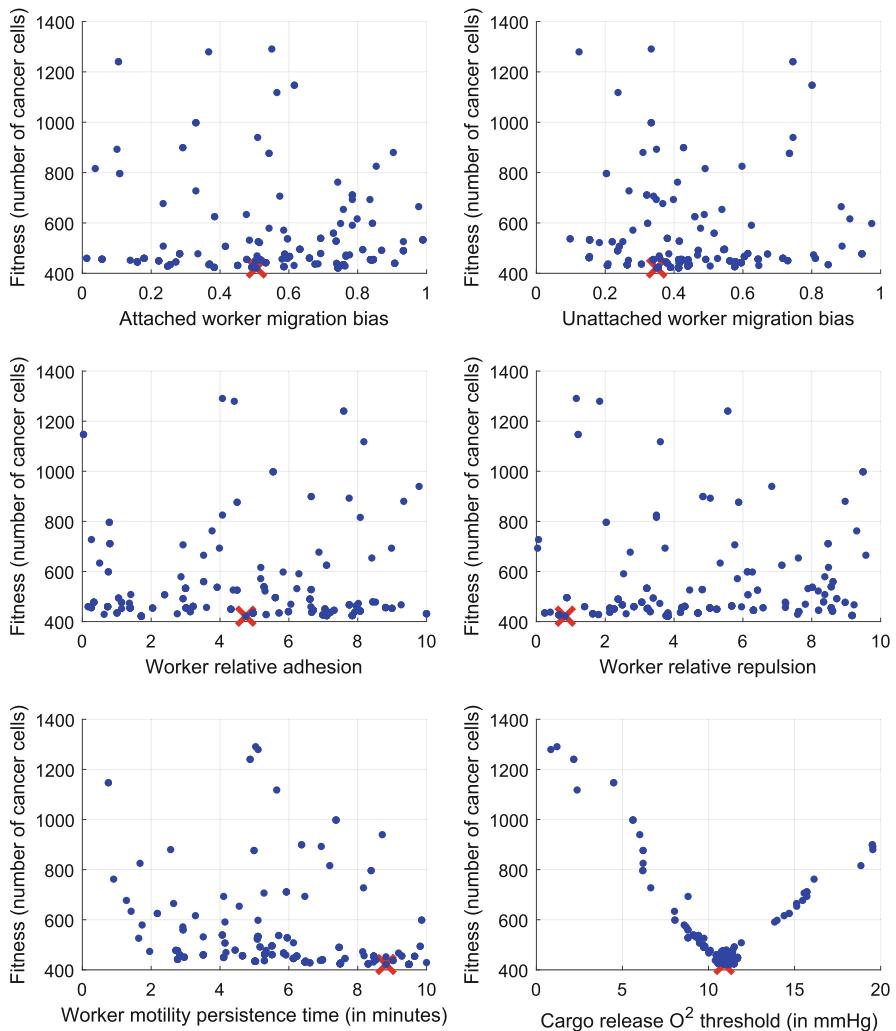


**Fig. 3.** Average (a) and best fitness (b) of the populations evolved with GA and DE in the second comparison run.



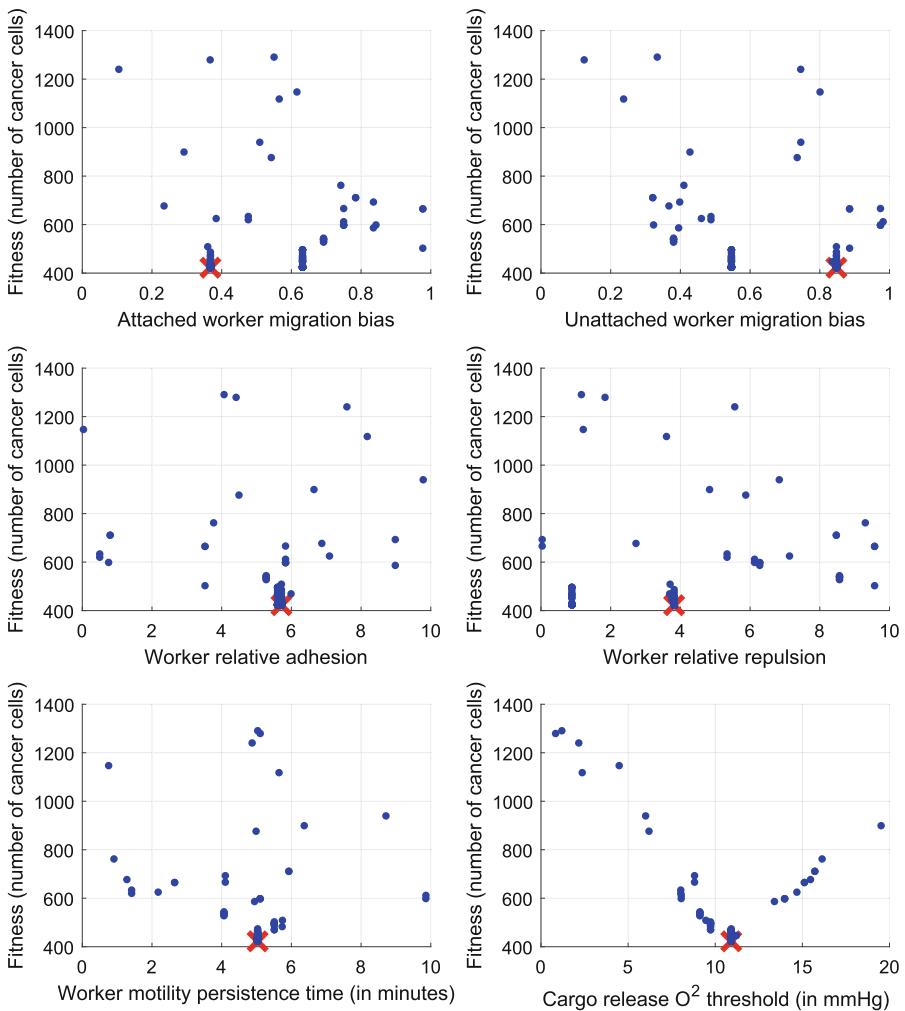
**Fig. 4.** Average (a) and best fitness (b) of the populations evolved with GA and DE in the third comparison run.

In Fig. 3(a) the GA approach seems to converge the average fitness of the population of solutions towards a lower fitness faster than the DE. Also, in Fig. 3(b) the DE approach is outperformed by GA for the first six generations. On the contrary, the GA is again stuck in a local optimum, while the DE is continuously evolving towards better solutions and manages to find a better one at the sixth generation.



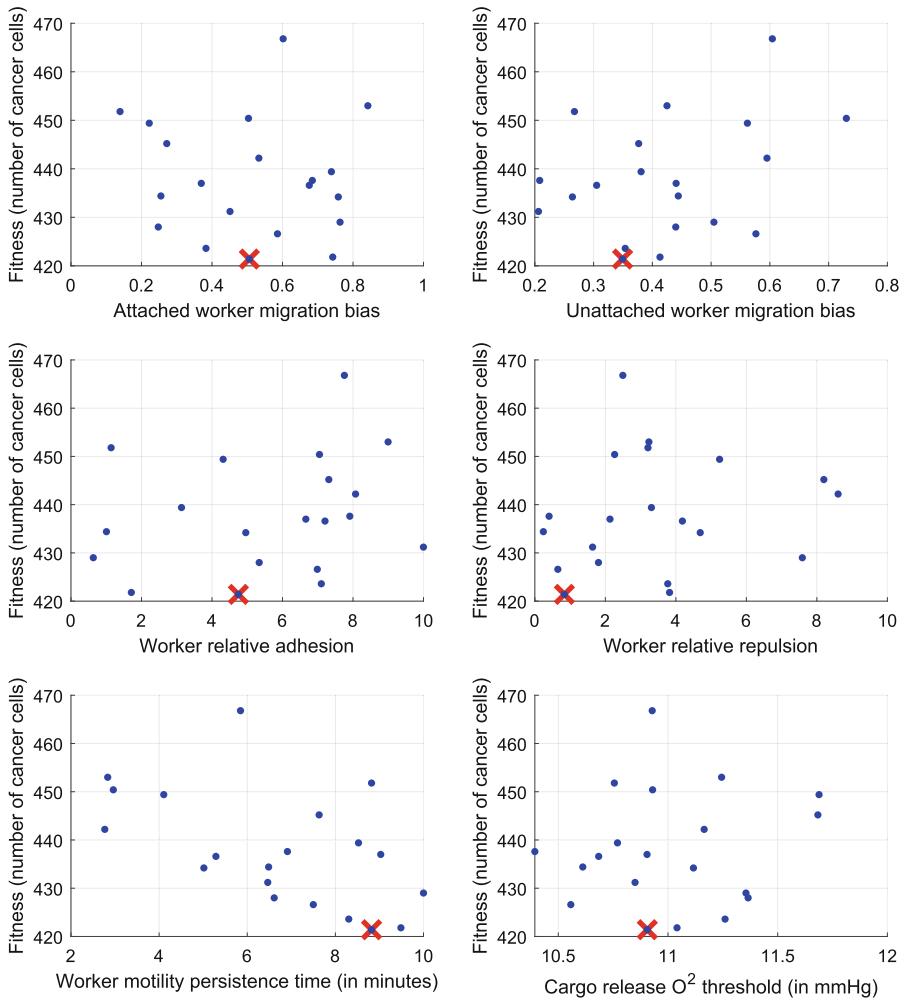
**Fig. 5.** Scatter plot of all individuals tested during the DE approach for the third run. The red “X” mark denotes the best individual found.

Finally, the results from the third comparison run are presented in Fig. 4. As in the previous runs, it can be claimed that DE outperforms the GA. Specifically, in Fig. 4(b) the DE approach reaches a fittest individual than GA at the very first generations. Furthermore, whereas the DE seems to be stuck in a local minimum from the third generation and the GA reaches a solution quite similar to this one, on the last generation the DE manages to escape its minimum and provide an ever better solution.



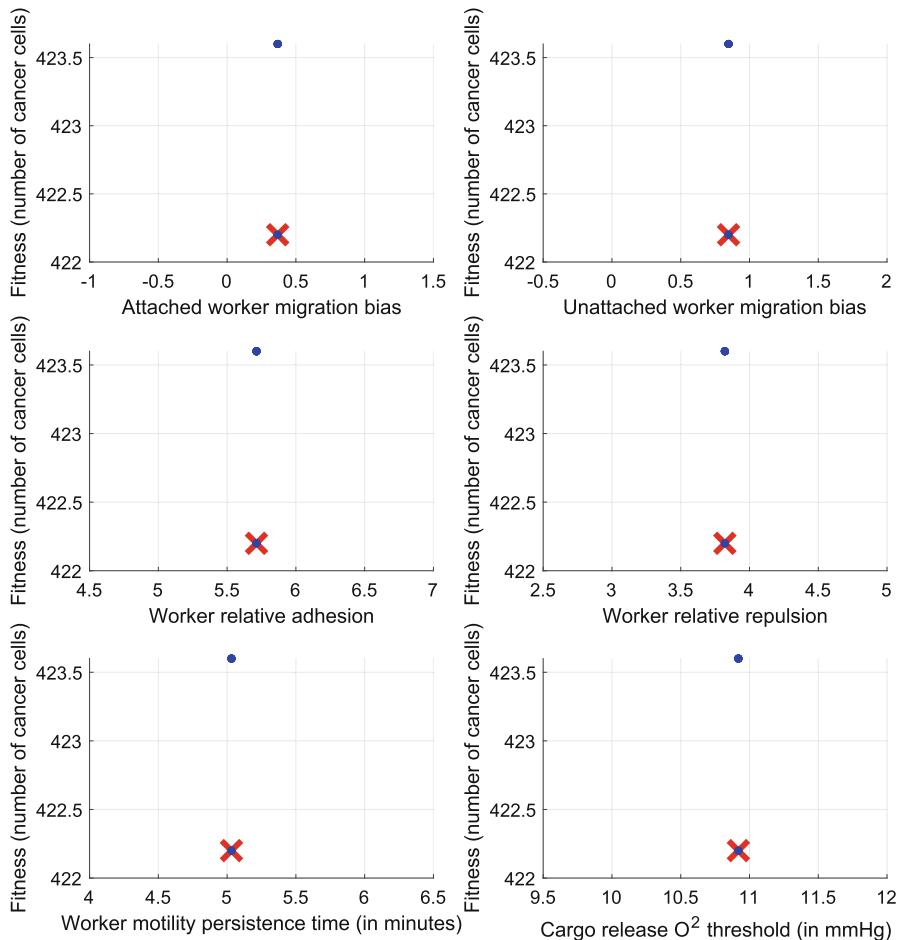
**Fig. 6.** Scatter plot of all individuals tested during the GA approach for the third run. The red “X” mark denotes the best individual found.

The scatter plots of the parameters investigated during the evolution of DE in the third comparison run are outlined in Fig. 5. It is clear that the individuals produced with the DE approach cover the search space better than the ones produced by GA (illustrated in Fig. 6). This is attributed to the fact that DE is designed to tackle models defined in real-values search spaces, whereas GA is not. As a result, the GA is heavily limited by the randomly produced initial population, a disadvantage that is alleviated by the DE methodology.



**Fig. 7.** Scatter plot of the individuals in the final population for the DE approach for the third run. The red “X” mark denotes the best individual found.

In addition, to clearly portray the reason why GA is easily trapped in local optima, while DE manages to escape them, Figs. 7 and 8 are given, that present the final population of the DE and GA approaches for the third run, respectively. The DE approach succeeds in maintaining a high diversity in the final population as shown in Fig. 7, where no duplicate individuals can be spotted. On the other hand, in Fig. 8 the final population of the GA approach is apparently comprised by multiple copies of just two individuals, which are also very close to each other. Consequently, the GA can not escape from these two individuals unless a dramatic mutation happens (not possible as the mutation step size was set here to  $s = [-5, 5]\%$ ).



**Fig. 8.** Scatter plot of the individuals in the final population for the GA approach for the third run. The red “X” mark denotes the best individual found.

## 5 Conclusion

The optimization of the design of targeted DDS on a cancer tumor, simulated by PhysiCell, was studied by utilizing “DE/rand/1” approach. The DE approach was compared with a standard steady state GA with the same computation budget, namely 1000 evaluations. The results derived from the comparison runs of both approaches equipped with the same initial population unveiled that DE is a more robust algorithm to reach a better solution within the same amount of generations. On top of that, DE enables the further exploration of the search space as it maintains a high diversity of the individuals in the final population.

As an aspect of future work, different variations of DE can be investigated with PhysiCell and under alternative ultimate goals, for instance the ability of niching is well documented in variants of DE [7]. Finally, the conclusions driven from this study will be applied on ongoing research towards a more wide applicability platform that will design, develop and evaluate DDSs aiming cancer tumours.

**Acknowledgement.** This project has received funding from the European Union's Horizon 2020 FET Open programme under grant agreement No. 800983.

## References

1. Aktipis, C., Nesse, R.: Evolutionary foundations for cancer biology. *Evol. Appl.* **6**, 144–159 (2013)
2. Bozic, I., Nowak, M.: Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15964–15968 (2014)
3. Bull, L.: On the Baldwin effect. *Artif. Life* **5**, 241–246 (1999)
4. Bull, L.: The evolution of sex through the Baldwin effect. *Artif. Life* **23**, 481–492 (2017)
5. Das, S., Mullick, S.S., Suganthan, P.: Recent advances in differential evolution - an updated survey. *Swarm Evol. Comput.* **27**, 1–30 (2016)
6. Das, S., Suganthan, P.N.: Differential evolution: a survey of the state-of-the-art. *IEEE Trans. Evol. Comput.* **15**, 4–31 (2010)
7. Epitropakis, M.G., Li, X., Burke, E.: A dynamic archive niching differential evolution algorithm for multimodal optimization. In: 2013 IEEE Congress on Evolutionary Computation, pp. 79–86 (2013)
8. Epitropakis, M.G., Tasoulis, D., Pavlidis, N., Plagianakos, V., Vrahatis, M.: Enhancing differential evolution utilizing proximity-based mutation operators. *IEEE Trans. Evol. Comput.* **15**, 99–119 (2011)
9. Ghaffarizadeh, A., Friedman, S., Macklin, P.: BioFVM: an efficient, parallelized diffusive transport solver for 3-D biological simulations. *Bioinformatics* **32**, 1256–1258 (2015)
10. Ghaffarizadeh, A., Heiland, R., Friedman, S., Mumenthaler, S., Macklin, P.: PhysiCell: an open source physics-based cell simulator for 3-D multicellular systems. *PLOS Comput. Biol.* **14**, e1005991 (2018)
11. Ozik, J., Collier, N., Heiland, R., An, G., Macklin, P.: Learning-accelerated discovery of immune-tumour interactions. *Mol. Syst. Des. Eng.* **4**, 747–760 (2019)
12. Preen, R., Bull, L., Adamatzky, A.: Towards an evolvable cancer treatment simulator. *Biosystems* **182**, 1–7 (2019)
13. Price, K.: Differential evolution vs. the functions of the 2<sup>nd</sup>/ICEO. In: Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC 1997), pp. 153–157 (1997)
14. Qin, A., Huang, V., Suganthan, P.N.: Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE Trans. Evol. Comput.* **13**, 398–417 (2008)
15. Schmitt, M., Loeb, L., Salk, J.: The influence of subclonal resistance mutations on targeted cancer therapy. *Nat. Rev. Clin. Oncol.* **13**, 335–347 (2016)

16. Storn, R., Price, K.: Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359 (1997)
17. Sun, J., Luo, C., Wang, Y., He, Z.: The holistic 3M modality of drug delivery nanosystems for cancer therapy. *Nanoscale* **5**, 845–859 (2013)
18. Thomsen, R.: Multimodal optimization using crowding-based differential evolution. In: Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No. 04TH8753) **2**, pp. 1382–1389 (2004)
19. Tsompanas, M.A., Bull, L., Adamatzky, A., Balaz, I.: Haploid-diploid evolution: nature's memetic algorithm. Preprint <https://arxiv.org/abs/1911.07302> (2019)
20. Tsompanas, M.A., Bull, L., Adamatzky, A., Balaz, I.: Novelty search employed into the development of cancer treatment simulations. *Inform. Med. Unlocked* **19**, 100347 (2020)
21. Tsompanas, M.A., Bull, L., Adamatzky, A., Balaz, I.: In silico optimization of cancer therapies with multiple types of nanoparticles applied at different times. *Comput. Methods Programs Biomed.* (2020) <https://doi.org/10.1016/j.cmpb.2020.105886>
22. Waclaw, B., Bozic, I., Pittman, M., Hruban, R., Vogelstein, B., Nowak, M.: A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* **525**, 261–264 (2015)
23. Wang, H., Rahnamayan, S., Sun, H., Omran, M.: Gaussian bare-bones differential evolution. *IEEE Trans. Cybern.* **43**, 634–647 (2013)



# On an Approach to Evaluation of Health Care Programme by Markov Decision Model

Masayuki Horiguchi<sup>(✉)</sup>

Department of Mathematics, Faculty of Science, Kanagawa University,  
2946 Tsuchiya, Hiratsuka, Kanagawa 259-1293, Japan  
[horiguchi@kanagawa-u.ac.jp](mailto:horiguchi@kanagawa-u.ac.jp)

**Abstract.** In this paper, we consider the evaluation of periodic screening programme for woman breast cancer and formulate the model as a partially observable Markov decision process (POMDP). We convert a POMDP with finite state, observation state and action spaces to an equivalent completely observable MDP with continuous state and finite action spaces. By this approach, we have an optimal policy from dynamic programming (DP) equation in an equivalent MDP, but we focus on considering the evaluation in scenarios of periodic screening for participants with silent condition of breast cancer and seeking an answer which programme is better than others for themselves. The aim of this paper is, by using the data sets based on cancer registration and estimated parameters of survival rates and other ratios related to screening and diagnoses in Japan, to evaluate some scenarios of breast cancer screening programme in POMDP.

**Keywords:** Finite MDP · Evaluation of health care programme · POMDP · Practice by MDP

**AMS(2020) Subject Classification:** Primary 90C40 · Secondary 62C10

## 1 Introduction

There are many papers of studying of the cost effectiveness analysis and guideline principles from the standpoint of Markov decision model (e.g., [3, 12, 14, 16, 17], and so on). It is important to consider not only the economic management for medical treatment, surgery or other therapy, etc., but also to minimize risks of harm to the patients and people who participate in the health care programme such as periodic medical examination at work place, quarantine and so on. In Japan mammography screening without clinical breast examination is recommended for woman aged 40–74 years and mammography screening with clinical breast examination is recommended for woman aged 40–64 years in the breast cancer screening programme [5]. The recommendations are concluded by

randomized controlled trial (RCT) and other experiments and comparing the benefits and risks in method of screening and contents of the programme. From the cancer statistics in Japan screening rate in periodic screening examination is about 45% and medical examination rate of participants who recalled after screening is about 80%. The government and municipal governments set numerical targets for the rates to be more than 50% and 90% respectively in the screening programme and examination, and moreover they engage in continuous activities in order to achieve the rates of participants including other cancers, colon, prostate, etc.

In this paper, we consider the evaluation of periodic screening programme for woman breast cancer and formulate the model as a POMDP. We convert a POMDP with finite state, observation state and action spaces to an equivalent completely observable MDP with continuous state and finite action spaces. By this approach, we have an optimal policy from DP equation in an equivalent MDP, but we focus on considering the evaluation in scenarios of periodic screening for participants with silent condition of breast cancer and seeking an answer which programme is better than others. We do not expect to be the main issue in this paper to estimate the parameters of survival rates and other ratios related to screening and diagnoses.

One of the aims of this paper is, by using the data sets based on cancer registration and estimated parameters of survival rates and others in Japan, to evaluate some scenarios of breast cancer screening programme in POMDP. We also search for solution from another point of view to achieve the rate of participants at least 50% coverage and at least 90% coverage those which are mentioned above, we try to show the difference of risk in programme between no screening and screening. From vital statistics [7, 21], screening statistics [10, 19] and estimated value of parameters (KapWeb [9]) we set the estimated values to our model parameters. We hope that our model in this paper will be an help to improve the rates of breast cancer screening in recommended programme by flexibility and benefits of MDP. It is difficult to have data related to untreated and unscreened breast cancer patients, we use the data by public institution like by SEER [8] including patients. Our approach is similar to the method of evaluation of relative mortality risk by Maillert et al. [12]. There are some recent papers related to breast cancer screening modeled as MDP, see [11, 14, 20]. Otten et al. [14] had considered the case to change the intervals of screening dynamically, although we fix intervals of screening by one year. Steimle & Denton [18] and Zhang & Denton [22] had considered the prostate cancer screening in POMDP. Other type of economic evaluation in management of hospital, Sauré & Puterman [16] had discussed on the problem of patient appointment scheduling.

In Sect. 2 we introduce the notation of POMDP and its converted MDP. Section 3 presents some preliminaries for numerical examples in the next section. In Sect. 4 we provides detailed numerical results by comparing the trajectories of cost function and information of state at each step.

It is not our purpose to estimate and/or use the estimated parameters in the middle of observing the states in dynamical system, because of our stand point

of intending to promote participation in the screening, although if we have more detailed data such as the false positive ratio etc. on clinical trials it is possible and important to consider approaches with adaptive control based on preventive intervention and/or avoiding the excessive treatment with regard to the result of screening.

Before the beginning of each scenario, important elements to describe the MDP, such as transition matrices and cost functions have been calculated by estimated parameters from the population-based data set in advance. We do not consider adaptive control model in this paper. This paper presents the desirable scenarios and principles for woman breast screening by comparing the values of relative mortality risks in every scenario in POMDP.

## 2 Partially Observed Markov Decision Process

In this section we define the sequential decision model to be examined by the similar formulation taken in Monahan [13] (see the generalized and abstract spaces case: refer to Rieder [15], Bäuelre and Rieder [1], Hinderer et al. [6]).

$X = \{0, 1, 2, \dots, N\}$  is the finite state space and  $X_t \in X$  is the state at time  $t = 0, 1, 2, \dots$ . We call the process  $\{X_t\}$  *core process*. Let  $A = \{1, 2, \dots, K\}$  be the finite action space and  $Y = \{1, 2, \dots, M\}$  be the finite observation space. Let  $Y_t \in Y$  denote the observation state at time  $t = 1, 2, \dots$  and we call the process  $\{Y_t\}$  *observation process*. The observation state  $Y_t = y_t$  at time  $t$  represents the information for the decision maker as imprecisely information of unobservable state of core process. In other words, we cannot observe the state  $X_t = x_t$  directly and deduce it by delivering the information of observing state  $Y_t = y_t$  to the decision maker and we have renewed the apriori belief probability on  $X$  at the time  $t - 1$  to the posteriori belief on  $X$  at the present time  $t$  by using the updating operator based on Bayesian mechanism.

We denote by  $\mathcal{P}(X)$  the set of all probability distributions on  $X$ .  $P_{a_t} = (p_{a_t}(j|i))$  denotes the transition law of core process  $\{X_t\}$  and  $p_{a_t}(j|i)$  is the probability of state transition from the present state  $X_t = i \in S$  to the next state  $X_{t+1} = j$  when an action  $a_t \in A$  is taken at time  $t$ . Let  $H_n = A \times Y \times H_{n-1}, n = 1, 2, \dots$  where  $H_0$  is arbitrary.  $h_n = (a_0, y_1, a_1, y_2, \dots, a_{n-1}, y_n) \in H_n$  is the history of action and observed state up to time  $n$ .

$C_a(x, y)$  is the immediate cost function defined on  $X \times Y \times A$  to  $\mathbb{R}$ .  $V_0(i)$  is the terminal cost function defined on  $X$  to  $\mathbb{R}$ . Let  $\pi_i(0) = Pr(X_0 = i), i \in S$  the initial distribution of core process.

Let  $q_a(y|i, j)$  be the transition kernel of observing state  $y$  from  $X \times X \times Y$  to  $[0, \infty)$  for  $a \in A$  when at the present time an action  $a$  is taken and the last state  $i$  and the present state  $j$  are given.

A policy for finite horizon  $T$  is a sequence  $\pi = (f_0, f_1, \dots, f_{T-1})$  of functions  $f_k : H_k \rightarrow A, 0 \leq k \leq T - 1$ . The set of all policies for finite horizon  $T$  is denoted by  $\Pi_T$ .

For each policy  $\pi \in \Pi_T$  and initial state  $X_0 = x \in X$ , a probability measure that describes the stochastic behavior of the partially observed system is defined by usual way (cf. [2, 4]).

Define  $\mu_t = (\mu_t(i)) \in \mathcal{P}(X)$  the information vector at time  $t$ , where we have set

$$\mu_t(i) = \Pr(X_t = i | \mu_0, a_0, y_1, a_1, y_2, \dots, a_{t-1}, y_t) = \Pr(X_t = i | \mu_0, h_t), \quad i \in X.$$

Let  $\mu_0 = (\mu_0(i)) \in \mathcal{P}(X)$  be the apriori distribution of unobservable states on  $X$  at the beginning period  $t = 0$ .

After knowing the information from observation  $Y_t = y_t$  at time  $t = 1, 2, \dots$ , apriori distribution as the information of unobserved state  $x_t$  is updated by the operator  $\Phi_{a_{t-1}}$ , where

$$\tilde{q}_a(j, y | \mu) = \sum_{i \in X} \tilde{q}_a(j, y | i) \mu(i) = \sum_{i \in S} p_a(j | i) q_a(y | i, j) \mu(i)$$

on  $X \times Y$  given  $\mu \in \mathcal{P}(X)$  for each  $a \in A$ :

$$\begin{aligned} \Phi_{a_{t-1}}(\mu_{t-1}, y_t)(j) &= \frac{\tilde{q}_{a_{t-1}}(j, y_t | \mu_{t-1})}{\sum_{j \in X} \tilde{q}_{a_{t-1}}(j, y_t | \mu_{t-1})} \\ &= \frac{\sum_{i \in X} p_{a_{t-1}}(j | i) q_{a_{t-1}}(y_t | i, j) \mu_{t-1}(i)}{\sum_{j \in X} \sum_{i \in X} p_{a_{t-1}}(j | i) q_{a_{t-1}}(y_t | i, j) \mu_{t-1}(i)} \end{aligned} \quad (1)$$

and we set  $\mu_t(j) = \Phi_{a_{t-1}}(\mu_{t-1}, y_t)(j)$ ,  $j \in X$ .

Let  $h_t = (a_0, y_1, a_1, y_2, \dots, a_{t-1}, y_t) \in H_t$ . Then, the posteriori distribution  $\mu_t(h_t)$  is calculated recursively by the updating operator  $\Phi_{a_i}$ ,  $i = 1, \dots, t$  as the following:

$$\mu_i(h_i) = \Phi_{a_{i-1}}(\Phi_{a_{i-2}}(\dots \Phi_{a_1}(\Phi_{a_0}(\mu_0, y_1), y_2), \dots, y_{i-1}), y_i)$$

The following theorem is well-known for finite state and action partially observable model.

**Theorem 1.** *For any fixed policy  $\pi = (a_0, a_1, \dots, a_{T-1})$  the sequence of probability distributions  $\{\mu_i(\cdot)\}$ ,  $i = 1, 2, \dots, T$  is a Markov process, i.e., for any measurable set  $\Gamma \in \mathcal{P}(X)$ ,*

$$\Pr(\mu_i \in \Gamma | \mu_0, a_0, \mu_1, \dots, \mu_{t-1}, a_{t-1}) = \Pr(\mu_i \in \Gamma | \mu_{i-1}, a_{i-1}).$$

We can convert a partially observable Markov decision model to an equivalent Markov decision model with continuous state space  $S = \mathcal{P}(X)$ .

By the requirement for describing the transitions of observation and unobserved state in the system, we define the transition law  $Q_a$ ,  $a \in A$  with stochastic kernel  $q_a(y|x, x')$  from  $X \times X \times Y$  to  $[0, \infty)$  as follows. For  $\mu \in \mathcal{P}(X)$  and  $a \in A$ ,

$$Q_a(\mu; y) = \sum_{x' \in X} \sum_{x \in X} p_a(x'|x) q_a(y|x, x') \mu(x)$$

and

$$Q_a(\mu; x', y) = \sum_{x \in X} p_a(x'|x) q_a(y|x, x') \mu(x).$$

The immediate cost function is defined by

$$C_a(\mu) = \sum_{x \in X} C_a(x) \mu(x), \text{ for } a \in A, \mu \in \mathcal{P}(X).$$

The terminal cost function is defined by

$$V_0(\mu) = \sum_{x \in X} V_0(x) \mu(x), \mu \in \mathcal{P}(X).$$

We denote by  $V_{N,\pi}(\mu)$  as the minimum expected undiscounted total cost function if there are  $N$  steps to go and the initial state is  $\mu$  and policy  $\pi \in \Pi$  is taken.

For history  $h_N = (a_0, y_1, a_1, y_2, \dots, a_{N-1}, y_N)$ , an action  $a_i \in A$  is taken and the information of system  $y_i$  is observed up to time  $N$ . Then we know transition law  $Q_a$  at time  $i$  and states transition of  $(\mu_i, y_i) \in \mathcal{P}(X) \times Y$  occurs according to  $Q_{a_{i-1}}(\mu_i; x, y_i), x \in X$ .

Note that the information vector  $\mu = (\mu_i), i \in X$  is updated by

$$\mu_i(x') = \Phi_{a_{i-1}}(\mu_i, y_i)(x') = \frac{Q_a(\mu; x', y)}{Q_a(\mu; y)}.$$

Then we have the following Dynamic Programming equation:

$$V_n(\mu) = \min_{a \in A} \left\{ C_a(\mu) + \sum_{y \in Y} V_{n-1}(\Phi_a(\mu, y)) Q_a(\mu; y) \right\}, n = 1, 2, \dots, N, \quad (2)$$

where  $V_0$  is terminal cost function.

From the results of finite horizon MDP we have the optimal policy for an equivalent MDP by solving DP equation recursively and backwards, from step  $n = 1$  to  $N$ .

Denote  $a_{N-n}^*$  the minimizer of (2) for each step  $n = 1, 2, \dots, N$  and let  $\pi^* = (a_0^*, a_1^*, \dots, a_{N-1}^*)$ .

**Theorem 2.** *The sequence of minimizer of (2) is an optimal policy for an equivalent MDP, i.e.,  $V_{N,\pi^*}(\mu_0) = \sup_{\pi \in \Pi} V_{N,\pi}(\mu_0)$ .*

### 3 POMDP for Evaluation of Health Care Programmes

We consider the evaluation of health care programme in mass screening of breast cancer as a POMDP. The state transitions of core process is shown in Fig. 1 and it is similar to the transition model of Maillart et al. [12]. In addition, we give subtree of decision making at time  $n$  in Fig. 2.

The state space is  $X = \{0, 1, 2, 3, 4\}$ . Each state  $x \in X$  represents the state of health in which screening participant is. Five states are represented as follows: State 0: no breast cancer, state 1: early breast cancer, state 2: later/advanced breast cancer, state 3: breast cancer induced death and state 4: non-breast cancer induced death. Transition matrix  $P = (p_{ij})$  leads to the state transitions at each time  $t = 1, 2, \dots$ , and satisfies the assumption:

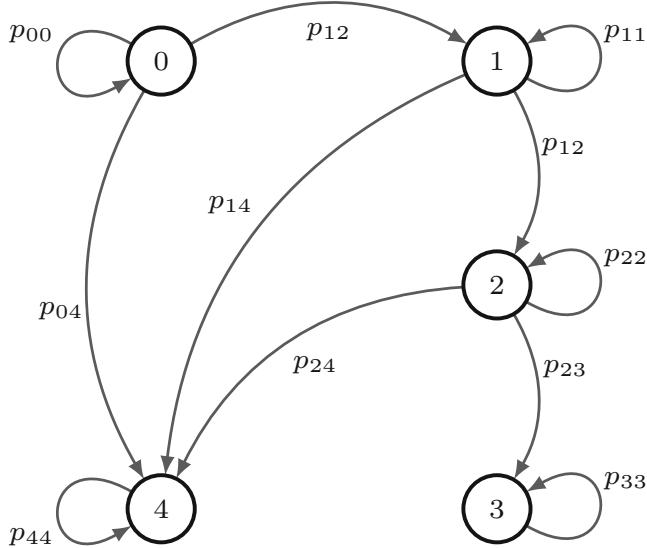
$$\begin{cases} p_{ii} > 0 \text{ for } i = 0, 1, 2, 3, 4, \\ p_{ij} > 0 \text{ if } j = i + 1, i = 0, 1, 2, \\ p_{i4} > 0 \text{ if } i = 0, 1, 2, \\ p_{ij} = 0 \text{ otherwise.} \end{cases}$$

The action space is  $A = \{a_0(\text{no screening}), a_1(\text{screening})\}$ . Every year in the period of screening programme, participants have two alternatives of  $a_0$ : no screening and  $a_1$ : screening, but for simplicity, under the scenario in this paper participants are arranged in advance which alternative is selected in the programme. The observing space is  $Y = \{0, 1\}$  where information (observed) state 0 is negative from the screening test and information state 1 is positive. The parameters  $p_{ij}(\alpha) = p(j|i)(\alpha)$  of transition matrices  $P(\alpha) = (p_{ij}(\alpha))$  are determined by the age  $\alpha$  of participants. Hence the state transition of core process occurs by  $P_a(\alpha), a \in A$  and the Markov process of core state is *non-stationary (non-homogeneous)* (cf. [2, 4]). Let  $I_i = \{n \in \mathbb{N} | 25 + 5(i - 1) \leq n < 25 + 5i\}$ . By abuse of notation, we let  $[\underline{\alpha}_i, \bar{\alpha}_i]$  stand for  $I_i$ .

We set immediate cost functions  $C_a^\alpha$  for participant who aged  $\alpha$  years and  $\alpha \in I_i = [\underline{\alpha}_i, \bar{\alpha}_i]$ .

$$\begin{aligned} C_{a_0}^\alpha(\mu) &= \mu(0)r_0(\alpha) + \mu(1)r_1(\alpha) + \mu(2)r_2(\alpha), \\ C_{a_1}^\alpha(\mu) &= \mu(0)d_0(\alpha)r_0(\alpha) + \mu(1)d_1(\alpha)r_1(\alpha) + \mu(2)d_2(\alpha)r_2(\alpha), \end{aligned}$$

where  $r_0(\alpha)$ : recall rate of screening participants who aged  $\alpha$  years,  $r_1(\alpha)$ : a complementary rate of 10-year relative survival rate  $e_1(\alpha)$  of patients of breast cancer who aged  $\alpha$  years and is in category 1 of breast density (Breast Imaging Reporting and Data System BI-RADS) category. Hence we define  $r_1(\alpha) = 1 - e_1(\alpha)$ . Also, we define  $r_2(\alpha) = 1 - e_2(\alpha)$ , where  $e_2(\alpha)$  is average (aggregation) of the survival rates of category 2 and 3.  $d_0(\alpha)$  denotes predictive value of positive screening results and  $d_1(\alpha)$  is the sensitivity of mammography in association with breast densities in category 1, and  $d_2(\alpha)$  is average of sensitivities of patients who is in category 2 or 3. The above values of parameters are surveyed and estimated by authors as follows:  $r_0, d_0$ : [10, 19],  $r_1, r_2(e_1, e_2)$ : [9],  $d_1, d_2$ : [19].

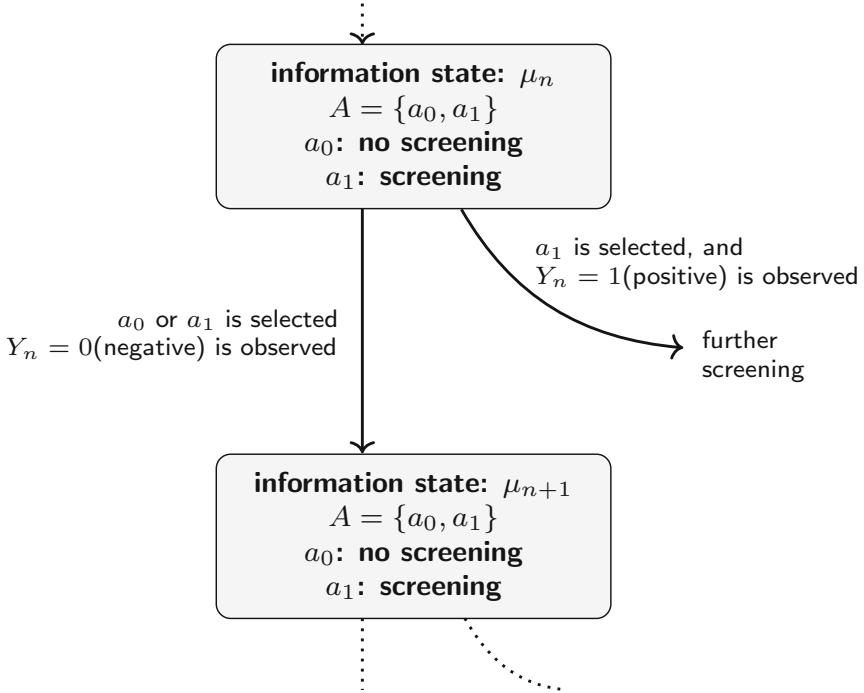


**Fig. 1.** Transition probabilities of core process

In order to update the information  $\mu$  and calculate the value functions  $V_{N,\pi}, V_n$ , we set  $q_{at-1}(y|i,j), t \geq 1$  in Eq. (1) when  $y = 0$  (screening test was negative) by

$$\begin{aligned}
 q_{at-1}(0|0,0) &= q_{at-1}(0|0,1) = q_{at-1}(0|0,4) = 1 - r_0(\alpha_{t-1}), \\
 q_{at-1}(0|1,1) &= q_{at-1}(0|1,2) = q_{at-1}(0|1,4) = 1 - d_1(\alpha_{t-1}), \\
 q_{at-1}(0|2,2) &= q_{at-1}(0|2,3) = q_{at-1}(0|2,4) = 1 - d_2(\alpha_{t-1}), \\
 q_{at-1}(0|i,j) &= 0 \quad \text{otherwise,}
 \end{aligned}$$

where  $\alpha_t$  is age of participant for screening at the  $t$ -th period from the beginning while in screening programme. Moreover, to deduce the transition probabilities  $p_{ij}(\alpha) = p(j|i)(\alpha)$  for screening participants who aged  $\alpha$  years, we follow the estimated cancer relative survival rates  $K_1$  and  $K_2$  of 5 years whose rates are defined similar to  $r_1(\alpha)$  and  $r_2(\alpha)$  respectively at clinical stages by KapWeb [9] in Japan. We also used mortality rate  $a$  of excluding the patients cased breast cancer and estimated incident rate  $b$  of breast cancer from vital and cancer statistics [7, 21] respectively in 2015 and the proportion  $c$  of early and non infiltrating breast cancer found in the past screening programme from Annual report on breast cancer screening in Japan in 2013 [10].



**Fig. 2.** Subtree of decision making at time  $n$  in POMDP

For the age  $\alpha \in I_i = [\underline{\alpha}_i, \bar{\alpha}_i]$ , the transition matrix  $P(\alpha) = (p_{ij}(\alpha))$  is required to satisfy the following equations: let  $p = p_{11}(\alpha), q = p_{22}(\alpha)$ ,

$$\begin{aligned} p^5 + (1 - p - a)(p^4 + p^3q + p^2q^2 + pq^3 + q^4) &= K_1(\alpha), \\ q^5 &= K_2(\alpha), \\ p_{12} &= bc. \end{aligned}$$

In this paper, we will restrict our attention to compare the scenarios of screening program for participants who has no symptom of breast cancer or never develops symptoms within continued period of the screening program, the value function represents the accumulated mortality risk of people in each scenario. Hence let  $Q_{a_0}(\mu; y) \equiv 1, V_{n-1}(\Phi_{a_1}(\mu; 1)) \equiv 0$  and we do not suit the optimal policy  $\pi^*$  derived from Eq. (2) to the best recommendation for the screening programme. Instead, under POMDP, we show the property of each programme including the recommended one by public institution [5].

## 4 Numerical Analysis

Almost of all statistical data used in this section are based on the aggregate statistics every 5 or 10 years of age of screening participants, the transition probability matrices and parameters for participants in the programme may take different values at every five years old (the beginnings of interval are 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75 and 80 years old).

We show 9 scenarios in Table 1 and consider the accumulative relative mortality risk in each scenario. For example, in scenario 1 the start age of participant is at 40 years old and begins the screening test at the age and continue the screening until at 84 years old. The duration of the scenario is  $N = 45$ . We give below the transition matrices  $P(\alpha)$  of core process for  $\alpha \in I_i, i = 1, 2, \dots, 12$ . The values of parameters are shown in Table 2 and 3. For each scenario, at age of the programme begins for participant we set the information state vector  $\mu'_0 = (1, 0, 0, 0, 0)'$ . The process of scenario 1 begins by transition matrix  $P(\alpha), \alpha \in I_4 = [40, 45]$ , i.e., the first transition of state is occurred by  $P(40)$ . After updating the vector  $\mu_t$  at time  $t$  (as information of participant to the screening programme and in  $t$ -th year of the scenario) until the final period  $N$ , we calculate the values  $V_1, V_2, \dots, V_N$  recursively from DP Eq. (2) by backward induction.

**Table 1.** scenarios of screening programme

Scenario	Age 25–39	Age 40–64	Age 65–84	Duration $N$ (years)
1	–	Screening		45
2	–	Screening	No screening	45
3	–	No screening		45
4	No screening			60
5	No screening	Screening		60
6	No screening	Screening	No screening	60
7	No screening			60
8	–	Screening: aged 45–74 years		35
9	–	Screening	–	25

$$\alpha \in I_1 : P(\alpha) = \begin{pmatrix} 0.9999291 & 0.0000645 & 0 & 0 & 0.0000064 \\ 0 & 0.9441060 & 0.0558876 & 0 & 0.0000064 \\ 0 & 0 & 0.9472137 & 0.0527799 & 0.0000064 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

**Table 2.** Values of the parameters (1)

$\alpha \in I_i$	$I_1$	$I_2, I_3$	$I_4, I_5$	$I_6, I_7$	$I_8, I_9$	$I_{10}, I_{11}$	$I_{12}$
$r_0(\alpha)$	0.04	0.053	0.116	0.095	0.072	0.048	0.048
$r_1(\alpha)$	0.947	0.951	0.971	0.969	0.992	1.000	1.000
$r_2(\alpha)$	0.626	0.728	0.768	0.743	0.758	0.801	0.796
$d_0(\alpha)$	0.007	0.016	0.019	0.025	0.041	0.069	0.105
$d_1(\alpha)$	—	—	1.000	0.875	0.912	—	—
$d_2(\alpha)$	—	—	0.632	0.741	0.794	—	—
$K_1(\alpha)$	0.955	0.981	0.992	0.986	1.000	1.000	1.000
$K_2(\alpha)$	0.7625	0.827	0.8785	0.848	0.866	0.894	0.914

**Table 3.** Values of parameters (2)

$\alpha \in I_i$	$I_1$	$I_2$	$I_3$	$I_4$	$I_5$	$I_6$
$a$	$6.45 \times 10^{-6}$	$9.08 \times 10^{-6}$	$1.41 \times 10^{-5}$	$2.47 \times 10^{-5}$	$3.33 \times 10^{-5}$	$4.65 \times 10^{-5}$
$b$	$8.60 \times 10^{-5}$	$2.46 \times 10^{-4}$	$6.96 \times 10^{-4}$	$1.50 \times 10^{-3}$	$2.31 \times 10^{-3}$	$2.23 \times 10^{-3}$
$c$	0.50 0.25	0.685 0.269		0.726 0.256		0.728 0.173
$\alpha \in I_i$	$I_7$	$I_8$	$I_9$	$I_{10}$	$I_{11}$	$I_{12}$
$a$	$6.41 \times 10^{-5}$	$1.15 \times 10^{-4}$	$1.99 \times 10^{-4}$	$2.86 \times 10^{-4}$	$7.54 \times 10^{-4}$	$1.07 \times 10^{-4}$
$b$	$2.18 \times 10^{-3}$	$2.34 \times 10^{-3}$	$2.31 \times 10^{-3}$	$2.25 \times 10^{-3}$	$2.00 \times 10^{-3}$	$1.70 \times 10^{-3}$
$c$	0.728 0.173	0.764 0.167		0.778 0.173		0.782 0.091

$$\alpha \in I_2 : P(\alpha) = \begin{pmatrix} 0.9997562 & 0.0002347 & 0 & 0 & 0.0000091 \\ 0 & 0.9441060 & 0.0558849 & 0 & 0.0000091 \\ 0 & 0 & 0.9627225 & 0.0372685 & 0.0000091 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_3 : P(\alpha) = \begin{pmatrix} 0.9993219 & 0.0006640 & 0 & 0 & 0.0000141 \\ 0 & 0.944171 & 0.0558149 & 0 & 0.0000141 \\ 0 & 0 & 0.9627225 & 0.0372634 & 0.0000141 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_4 : P(\alpha) = \begin{pmatrix} 0.9985003 & 0.0014750 & 0 & 0 & 0.0000247 \\ 0 & 0.9672830 & 0.0326923 & 0 & 0.0000247 \\ 0 & 0 & 0.9744249 & 0.0255504 & 0.0000247 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_5 : P(\alpha) = \begin{pmatrix} 0.9976973 & 0.0022694 & 0 & 0 & 0.0000333 \\ 0 & 0.9412310 & 0.0587357 & 0 & 0.0000333 \\ 0 & 0 & 0.9744249 & 0.0255418 & 0.0000333 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_6 : P(\alpha) = \begin{pmatrix} 0.9979370 & 0.0020164 & 0 & 0 & 0.0000465 \\ 0 & 0.9539420 & 0.0460115 & 0 & 0.0000465 \\ 0 & 0 & 0.9675628 & 0.0323907 & 0.0000465 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_7 : P(\alpha) = \begin{pmatrix} 0.9979681 & 0.0019678 & 0 & 0 & 0.0000641 \\ 0 & 0.9542050 & 0.0457309 & 0 & 0.0000641 \\ 0 & 0 & 0.9675628 & 0.0323731 & 0.0000641 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

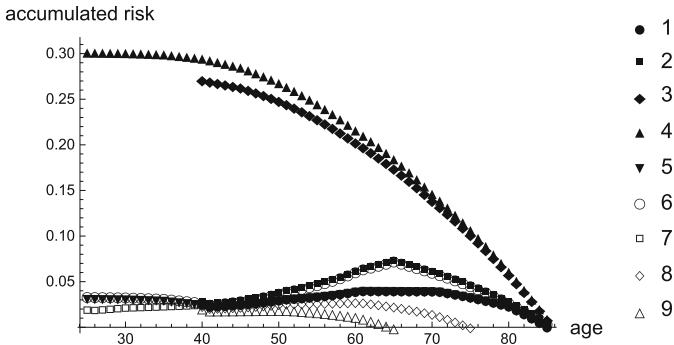
$$\alpha \in I_8 : P(\alpha) = \begin{pmatrix} 0.9977112 & 0.0021739 & 0 & 0 & 0.0001149 \\ 0 & 0.9483570 & 0.0515281 & 0 & 0.0001149 \\ 0 & 0 & 0.9716360 & 0.0282491 & 0.0001149 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_9 : P(\alpha) = \begin{pmatrix} 0.9976481 & 0.0021525 & 0 & 0 & 0.0001995 \\ 0 & 0.9498140 & 0.0499865 & 0 & 0.0001995 \\ 0 & 0 & 0.9716360 & 0.0281646 & 0.0001995 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

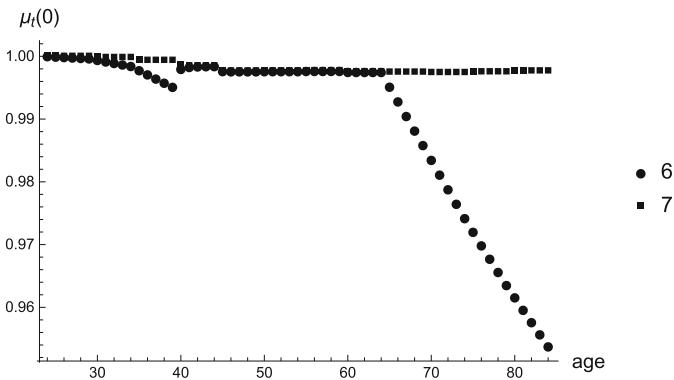
$$\alpha \in I_{10} : P(\alpha) = \begin{pmatrix} 0.9975742 & 0.0021398 & 0 & 0.0002860 & \\ 0 & 0.9370820 & 0.0626320 & 0 & 0.0002860 \\ 0 & 0 & 0.9778393 & 0.0218746 & 0.0002860 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_{11} : P(\alpha) = \begin{pmatrix} 0.9976596 & 0.0019001 & 0 & 0 & 0.0004403 \\ 0 & 0.9405450 & 0.0590147 & 0 & 0.0004403 \\ 0 & 0 & 0.9778393 & 0.0217204 & 0.0004403 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\alpha \in I_{12} : P(\alpha) = \begin{pmatrix} 0.9977629 & 0.0014832 & 0 & 0 & 0.0007538 \\ 0 & 0.9302950 & 0.0689512 & 0 & 0.0007538 \\ 0 & 0 & 0.9830340 & 0.0162122 & 0.0007538 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$



**Fig. 3.** Trajectories of accumulated lifetime mortality risks



**Fig. 4.** Trajectories of information  $\mu_t(0)$

Figure 3 shows the trajectories of value (accumulated lifetime mortality risk) in each scenario. In Fig. 4, the trajectories of  $\{\mu_t(0)\}$  in scenario 6 and 7 are shown. In contrast to choosing the option of no screening at each period decreased the information probability  $\mu_t(0)$  of no breast cancer in scenario 6, continuous participation to screening brought almost flat tendency with few decrease to the information probability  $\mu_t(0)$  with respect to the time  $t$ . Trajectories of  $\{\mu_t(0)\}$  in other scenarios has similar tendency as the main factor of those variations is whether the scenario itself has the duration of no screening or not. So we omit to show trajectories  $\{\mu_t(0)\}$  in the case of other variations.

From the trajectories in Fig. 3, if the longer duration of no screening time is included in the scenario which participants select, the higher accumulated risk  $V_N$  of the scenario is brought. The scenario 8 (mammographic screening without clinical breast examination) and 9 (mammographic screening without clinical breast examination) are both recommended for people living in Japan by National Cancer Center of Japan (NCCJ) (cf. [5]).

**Acknowledgement.** This work was partially supported by MHLW Cancer Control Promotion Program Grant Number JPMH18EA1003.

## References

1. Bäuerle, N., Rieder, U.: *Markov Decision Processes with Applications to Finance*. Springer, Heidelberg (2011)
2. Bertsekas, D.P., Shreve, S.E.: *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, New York (1978)
3. Drummond, M.F., Sculpher, M.J., Claxton, K., Stoddart, G.L., Torrance, G.W.: *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, Oxford (2015)
4. Dynkin, E.B., Yushkevich, A.A.: *Controlled Markov Processes*. Springer, New York (1979)
5. Hamashima, C., et al.: The Japanese guidelines for breast cancer screening. *Jpn. J. Clin.* **46**, 482–492 (2016)
6. Hinderer, K., Rieder, U., Stieglitz, M.: *Dynamic Optimization: Deterministic and Stochastic Models*. Springer, Cham (2016)
7. Hori, M., Matsuda, T., Shibata, A., Katanoda, K., Sobue, T., Nishimoto, H.: Cancer incidence and incidence rates in Japan in 2009: a study of 32 population-based cancer registries for the Monitoring of Cancer Incidence in Japan (MCIJ) project. *Jpn. J. Clin.* **45**, 884–891 (2015)
8. Howlader, N., Noone, A.M., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D.R., Chen, H.S., Feuer, E.J., Cronin, K.A. (eds.) *SEER Cancer Statistics Review, 1975–2017*. National Cancer Institute, Bethesda, MD, based on November 2019 SEER data submission, posted to the SEER web site, April 2020. [https://seer.cancer.gov/csr/1975\\_2017/](https://seer.cancer.gov/csr/1975_2017/)
9. KapWeb: Survival statistics of Japanese association of Clinical Cancer Centers. <https://kapweb.chiba-cancer-registry.org>
10. Kasahara, Y., Tsuji, I., Ohnuki, K., Koibuchi, Y., Ban, K., Furukawa, J., Masuoka, H., Murata, Y., Morita, T., Yoshida, M., Rai, Y.: Annual report 2013 on breast cancer screening in Japan. *J. Jpn. Assoc. Breast Cancer Screen.* **23**, 84–97 (2014)
11. Madadi, M., Zhang, S.: Cost-effectiveness analysis of breast cancer mammography screening policies considering uncertainty in women's adherence. In: Kong, N., Zhang, S. (eds.) *Decision Analytics and Optimization in Disease Prevention and Treatment*, pp. 223–240. Wiley, New York (2018)
12. Maillart, L.M., Ivy, J.S., Ransom, S., Diehl, K.: Assessing dynamic breast cancer screening policies. *Oper. Res.* **56**, 1411–1427 (2008)
13. Monahan, G.E.: A survey of partially observable Markov decision processes: theory, models, and algorithms. *Manag. Sci.* **28**, 1–16 (1982)
14. Otten, J.W.M., Witteveen, A., Vliegen, I.M.H., Siesling, S., Timmer, J.B., IJzerman, M.J.: Stratified breast cancer follow-up using a partially observable Markov decision process. In: Boucherie, R., van Dijk, N. (eds.) *Markov Decision Processes in Practice*, pp. 223–244. Springer, New York (2017)
15. Rieder, U.: Structural results for partially observed control models. *ZOR Methods Models Oper. Res.* **35**, 473–490 (1991)
16. Sauré, A., Puterman, M.L.: Advance patient appointment scheduling. In: Boucherie, R., van Dijk, N. (eds.) *Markov Decision Processes in Practice*, pp. 245–168. Springer, New York (2017)

17. Siebert, U., Alagoz, O., Bayoumi, A.M., Jahn, B., Owens, D.K., Cohen, D.J., Kuntz, K.M.: State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3. *Value Health* **15**, 812–820 (2012)
18. Steimle, L.N., Denton, B.T.: Markov decision processes for screening and treatment of chronic diseases. In: Boucherie, R., van Dijk, N. (eds.) *Markov Decision Processes in Practice*, pp. 189–222. Springer, New York (2017)
19. Suzuki, A., Kuriyama, S., Kawai, M., Amari, M., Takeda, M., Ishida, T., Ohnuki, K., Nishino, Y., Tsuji, I., Shibuya, D., Ohuchi, N.: Age-specific interval breast cancers in Japan: estimation of the proper sensitivity of screening using a population-based cancer registry. *Cancer Sci.* **99**, 2264–2267 (2008)
20. Tunc, S., Alagoz, O., Chhatwal, J., Burnside, E.S.: Using finite-horizon Markov decision processes for optimizing post-mammography diagnostic decisions. In: Kong, N., Zhang, S. (eds.) *Decision Analytics and Optimization in Disease Prevention and Treatment*, pp. 183–200. Wiley, New York (2018)
21. Vital Statistics Japan (Ministry of Health, Labour and Welfare). [https://ganjoho.jp/en/professional/statistics/table\\_download.html](https://ganjoho.jp/en/professional/statistics/table_download.html)
22. Zhang, J., Denton, B.T.: Partially observable Markov decision processes for prostate cancer screening, surveillance, and treatment: a budgeted sampling approximation method. In: Kong, N., Zhang, S. (eds.) *Decision Analytics and Optimization in Disease Prevention and Treatment*, pp. 201–222. Wiley, New York (2018)

# Author Index

## A

- Adamatzky, Andrew, 328  
Almudevar, Anthony, 298  
Anton, Elene, 266  
Avrachenkov, Konstantin E., 192  
Ayesta, Urtzi, 266

## B

- Balaz, Igor, 328  
Bäuerle, Nicole, 108  
Borkar, Vivek S., 192  
Bull, Larry, 328

## D

- de Oliveira, André Marcorin, 87  
Deng, Fan, 221  
Diep, Quoc Bao, 313  
do Valle Costa, Oswaldo Luiz, 87  
Dolhare, Hars P., 192

## F

- Feinberg, Eugene A., 1

## G

- Glauner, Alexander, 108  
González-Sánchez, David, 148  
Guo, Xin, 221

## H

- Horiguchi, Masayuki, 341  
Huo, Haifeng, 19

## J

- Jasso-Fuentes, Héctor, 57  
Jonckheere, Matthieu, 266

## K

- Kara, Ali Devran, 166  
Kasyanov, Pavlo O., 1

## L

- Lipets, Vladimir, 284  
Liu, Li, 248

## M

- Menaldi, Jose-Luis, 57  
Minjárez-Sosa, J. Adolfo, 148

## P

- Patil, Kishor, 192  
Piunovskiy, Alexey B., 38

## R

- Robles-Aguilar, Alan D., 148

## S

- Semenikhin, Konstantin V., 129  
Sonin, Isaac M., 248

## T

- Truong, Thanh Cong, 313  
Tsompanas, Michail-Antisthenis, 328

**V**

- Vásquez-Rojas, Fidel, [57](#)  
Verloop, Ina Maria, [266](#)

**W**

- Wen, Xian, [19](#)

**Y**

- Yüksel, Serdar, [166](#)

**Z**

- Zadorojniy, Alexander, [284](#)  
Zelinka, Ivan, [313](#)  
Zgurovsky, Michael Z., [1](#)  
Zhang, Yi, [221](#)