Continuous-Time Mean-Field Markov Decision Models

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Sebastian Höfer, M.Sc.
aus Esslingen am Neckar

Tag der mündlichen Prüfung: 17. Dezember 2025

Referentin: Prof. Dr. Nicole Bäuerle
Korreferent: Prof. Dr. Sören Christensen

Meiner Familie gewidmet

Danksagung

Diese Arbeit ist während meiner Zeit als wissenschaftlicher Mitarbeiter am Institut für Stochastik am Karlsruher Institut für Technologie entstanden. An dieser Stelle möchte ich mich bei den Personen bedanken, die mich auf diesem Weg begleitet und geprägt haben.
Zunächst möchte ich mich ganz herzlich bei meiner Betreuerin Prof. Dr. Nicole Bäuerle bedanken. Sie hat mich über die gesamte Dauer der Promotion fachlich hervorragend unterstützt und sich bei Fragen sofort Zeit genommen. Außerdem hat sie mir in der Lehre die notwendige Freiheit gelassen, um auch den ein oder anderen persönlichen Akzent zu setzen. Ganz besonders geschätzt habe ich ihre Geduld und Unaufgeregtheit, die sie mit mir manchmal vielleicht auch gebraucht hat.
Auch meinem Zweitgutachter Prof. Dr. Sören Christensen möchte ich an dieser Stelle danken. Gerade in der Anfangszeit meiner Promotion war der von ihm mitorganisierte „Workshop on Stochastic Models and Control“ in Travemünde ein super Orientierungspunkt.
Ganz besonders danken möchte ich meiner Familie: meinen Eltern, meinen Brüdern, meiner Tante und meiner Oma, auf die man sich immer zu 100% verlassen kann und die mich immer uneingeschränkt unterstützt haben. Weiter möchte ich mich bei Laura bedanken, die gerade gegen Ende dieser Arbeit die doch teilweise langen Arbeitstage aufgewertet hat.
Neben meiner leiblichen Familie möchte ich mich bei meiner akademischen Familie bedanken. Allen voran bei meiner großen akademischen Schwester Tamara, die meine Zeit am Institut fachlich und menschlich unheimlich bereichert hat. Gleiches gilt ohne Abstriche auch für alle weiteren aktuellen und ehemaligen Stochastik-Buddies, sowie für meine Imkerpartnerin Tatjana. Den familiären und unkomplizierten Umgang miteinander werde ich sehr vermissen und würde mir wünschen, dass dieser in Zukunft für die kommenden „Generationen“ von Doktorandinnen und Doktoranden erhalten bleibt.
Abschließend möchte ich auch all den Menschen, die auf die ein oder andere Weise ihren Anteil an der Promotion hatten, aber nicht alle namentlich genannt werden können, von ganzem Herzen danken.

Prior Publications

Parts of Chapters 3 and 4 are direct quotes from the prior publication

Bäuerle and Höfer (2024). Continuous-time mean field markov decision models, *Applied Mathematics and Optimization* **90**: 12.

Contents

CHAPTER 1

Introduction

1.1. The idea of mean-field Markov decision models

Imagine, one day, for whatever reason, destiny puts you in the following situation: You are navigating on a set of states S and must choose actions from a set of actions A that influence your transition to the next state. If this transition depends solely on the current state and the chosen action, it is called Markovian. To give this process a purpose or a goal, you receive a reward that depends on your current state and the chosen action. As a rational agent, you will try to maximize your expected rewards over a given time horizon. In doing so, you must consider the following trade-off: optimizing your immediate reward while keeping yourself in favorable states to ensure high rewards in the future. In this scenario, you assume the role of the decision-maker in a so-called *Markov decision model*.

Now, suppose that you are no longer acting alone in the system. Instead, a total of N agents are operating simultaneously on the state space S. The agents are statistically equal in the sense that they are all subject to the same transition and reward mechanisms. However, they are not independent, since the transition intensities and rewards may additionally depend on the empirical distribution of the agents among the states, which we assume to be public information, acknowledged by all the agents. Although several optimization criteria exist for such systems, we will focus on a fully cooperative scenario. Here, the agents work in unison to maximize a common objective: the mean of their expected cumulative rewards over the planning horizon. Consequently, the collective goal is to maximize the overall social welfare, rather than individual gains. An equivalent perspective is that of a central controller that aggregates information and assigns actions to the agents.

This setup describes a cooperative multi-agent Markov decision model with mean-field interactions, or, for short, a *mean-field Markov decision model*. A common approach to dealing with such problems is to let the number of agents tend to infinity. In the continuous-time setting with finite state space, we establish the convergence to a deterministic optimization problem, where the state process is described by an ordinary differential equation. This resulting limit problem can then be solved using methods from deterministic optimal control. One of the main objectives of this monograph will be to examine the connections between the mean-field Markov decision model and the corresponding deterministic limit model.

As one might already anticipate, the theory of continuous-time mean-field Markov decision models integrates three distinct fields of mathematics: Markov decision processes, mean-field theory, and deterministic optimal control.

1.2. HISTORICAL REMARKS

This section reviews the key developments of Markov decision processes and mean-field theory. Some historical aspects of deterministic optimal control are discussed in Chapter 4.

MARKOV DECISION PROCESSES

The origins of Markov chains trace back to a scientific debate between Andrey Markov (1856-1922) and Pavel Nekrasov (1853-1924) on the question of whether the independence of a sequence of random variables is a necessary condition for the validity of the law of large numbers. In his work "Extension of the limit theorems of probability theory to a sum of variables connected in a chain", Markov (1907) introduced the concept of a sequence of random variables in which the distribution of one variable depends on the value of the immediately preceding random variable. For this sequence of dependent random variables, he established a version of the law of large numbers. Later, similar concepts of dependent random variables were named Markov chains in his honor.

In the 1950s, Richard Bellman developed a new method for solving sequential decision problems, which he named *dynamic programming*. The core of dynamic programming is the principle of optimality, see Bellman (1954):

"*An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.*"

According to the principle of optimality, once the *current* state is specified, the remaining decisions must be optimal with respect to that state alone, regardless of the preceding history that led to it. This suggested the implementation of a (controlled) Markov chain as the state process. In his groundbreaking book *Dynamic Programming*, Bellman (1957) subsequently coined the term *Markovian decision process*.

In continuous-time, the *value function*, defined as the maximum expected cumulative reward given an initial state, is characterized by a partial differential equation, the so-called *Hamilton-Jacobi-Bellman equation (HJB)*. The corresponding difference equation in discrete-time is commonly referred to as the *Bellman equation*.

Following the pioneering work of Bellman, the field of Markov decision processes (MDPs) has experienced rapid evolution. In this early phase, two fundamental solving algorithms for Markov decision problems emerged. The first, *Value Iteration*, is a direct application of Bellman's work. It iteratively refines the value function for each state, generating a sequence that converges to the optimal value function. Shortly thereafter, Howard (1960) introduced *Policy Iteration*, an efficient algorithm to calculate an optimal policy that consists of two alternating steps: the value-determination operation and the policy-improvement routine. Blackwell (1962, 1965) established a solid mathematical framework for MDPs. In particular, Blackwell studied models with a discount factor β, which is why policies that are optimal for all discount factors sufficiently close to 1 are named *Blackwell-optimal* in his honor.

In recent decades, research on Markov decision processes has diversified significantly, leading to the development of a rich family of models derived from the original. Constrained MDPs (CMDPs) focus on maximizing the reward while adhering to certain side constraints; see Altman (1999). Risk-sensitive MDPs, which incorporate risk aversion into the decision-making process, originated with the foundational work of Howard and Matheson (1972). Multi-Objective MDPs (MOMDPs) were introduced to address applications with multiple, possibly conflicting goals. For instance, a taxi driver should minimize travel time, maximize passenger comfort, minimize fuel consumption, and minimize accident risk. For an overview of the field, see Roijers et al. (2013). Partially-observable MDPs (POMDPs) were introduced for scenarios where the agent's knowledge of the current state is incomplete or uncertain, such as a robot with noisy sensors that is uncertain of its exact location but must still decide which way to turn to reach its goal. Monahan (1982) provides a detailed survey of the topic. Mean-Field MDPs (MFMDPs) represent another branch in this family of Markov decision models, a category to which this monograph can be assigned. A review of the literature and state of the art is given in Section 1.3.

Mean-field theory

The origins of *mean-field theory* lie in physics, more specifically in statistical mechanics. It is typically used to model large systems of interacting particles. A classic example is the Ising model, which describes magnetism in solids. On a crystal lattice with a large number of atoms, each of these atoms possesses a spin that can point either up or down. In the antiferromagnetic case, the orientation of the spins depends on two competing forces: an external magnetic field, which affects all spins equally and favors alignment of all spins in one direction, whereas spin-spin interaction favors antiparallel alignment of neighboring spins. For a large number of spins, the interactions among them are difficult to handle. To simplify the calculation, the complex pairwise interactions exerted by any atom on every

other atom are replaced by an average, effective field (the *mean field*). This effective field is then treated as an additional external field.

The method of mean-field approximation was subsequently applied in other disciplines where large particle interactions occur. A notable field to mention is epidemiology, where mean-field theory is used to model the spread of infectious diseases in a population with a large number of individuals. An early influential example is the so-called SIR model by Kermack and McKendrick (1927) (S = susceptible, I = infected, R = removed/recovered) and models derived from it, such as the SIRD model (D = dead) or the SIRV model (V = vaccinated). In these standard models, the key parameters that determine the course of the epidemic, such as the infection rate, recovery rate, etc., are assumed to be constant for every individual in the population. Individual differences, like age or pre-existing conditions, are averaged out under the mean-field approximation.

In neuroscience, mean-field theory is a common approach to describe the collective dynamics of large populations of neurons. Instead of tracking the state and interaction of every single neuron in a brain region, the theory approximates the total synaptic input to a representative neuron as an average field. This "mean field" represents the aggregate activity of the entire network. For example, in the widely recognized model of Cowan and Wilson (1972), all cells receive the same average excitation level. In this context, mean-field theory allows modeling macroscopic phenomena, such as oscillations and synchronous firing patterns (so-called "brain waves"), that emerge from the complex interplay of thousands of individual neurons, linking microscopic behavior to observable brain function. For an overview of mean-field models in computational neuroscience, see Deco et al. (2008).

Naturally, it was only a small step from the application of mean-field theory in biological neural models to its use in artificial neural networks and machine learning. In very wide networks, where the number of neurons per layer is large, the complex interactions between individual neurons can be effectively analyzed using mean-field approximations. Sompolinski et al. (1988) study random neural networks with a large number N of neurons and show that in the mean-field limit $N \to \infty$ there occurs a sharp transition from a stationary phase to a chaotic phase at a critical value of the gain parameter. Poole et al. (2016) combine these findings with Riemannian geometry to examine the signal propagation in generic, deep neural networks with random weights. An even more direct connection between neural networks and their origins in statistical mechanics is provided by so-called Boltzmann machines, also known as the stochastic Ising model. A Boltzmann machine is a type of stochastic recurrent neural network that was proposed by Hinton and Sejnowski, see Ackley et al. (1985). Neuron activations are updated stochastically according to the Boltzmann distribution, which is the origin of the model's name. To improve the practical efficiency, Peterson and Anderson (1987) proposed a mean-field learning algorithm. Their method approximates neuronal correlations using deterministic mean-field equations, which avoids the slow, stochastic measurement of correlations between the units. In 2024, Geoffrey Hinton was awarded the Nobel Prize in Physics for his foundational discoveries in artificial

neural networks, including his co-invention of the Boltzmann machine.

As a recent development in mathematics, *mean-field games* adapt the core ideas of the established physical theory to model the strategic behavior of a very large number of competitive, interacting agents. The development of mean-field game theory was initiated independently by the work of Lasry and Lions (2007) and Huang et al. (2006). Both recognized that the analysis of systems with a finite number of agents can often be simplified by considering the behavior of systems with an infinite number of agents. The complexity arising from the interaction between the individuals is replaced by considering the interaction of a single representative agent with the *mean field*: the aggregated behavior of the entire population. A core assumption is that the influence of any individual agent on the overall system is negligible, analogous to the infinitesimal impact of a single spin on the macroscopic magnetic field. The optimal strategy of the individual depends on the distribution of the total population, and the evolution of the total population is determined by the sum of the individual optimal strategies. The solution to the mean-field game yields a so-called *mean-field equilibrium*, a term coined by Lasry and Lions (2007), which is an analogue of the Nash equilibrium in classical game theory for an infinite number of players. A rigorous mathematical framework of mean-field game theory is provided in Carmona and Delarue (2018a,b). Since their introduction in 2006, mean-field games have found wide-ranging applications in fields such as economics, finance, engineering, and sociology. To name just a few, Bäuerle and Göll (2023) analyze a financial market with competitive, relative investors. Liu et al. (2018) propose a mean-field game approach to model swarms of robots and Bauso et al. (2016) investigate the propagation of opinions in social networks.

1.3. Related work

Whereas mean-field games typically consider a large number of *competitive*, interacting agents, leading to concepts like (mean-field) Nash equilibria, the focus of this work is on *cooperative* agents who aim to maximize the social welfare, leading to a Pareto-optimal outcome. A related perspective is that of a central controller or social planner, who can observe the agents' states and assign actions to them.

Optimal control of McKean-Vlasov processes

The setting most extensively studied in this context is a system of N agents whose state dynamics are governed by a stochastic differential equation

$$\begin{aligned} dX_t^i &= b(t, X_t^i, \mu_t^N, a_t^i) + \sigma(t, X_t^i, \mu_t^N, a_t^i) dW_t^i, \\ \mu_t^N &= \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i}, \end{aligned}$$

where interaction occurs through the empirical distribution μ_t^N in the drift b and the diffusion σ. Here, $W^1, \ldots, W^N$ are independent Brownian motions, δ denotes the Dirac measure and $a^1, \ldots, a^N$ are controls chosen by the central controller. The aim of the controller is to maximize the average expected cumulative reward of the system given by

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\int_0^T r(t, X_t^i, \mu_t^N, a_t^i)dt + g(X_T^i, \mu_T^N)\right],$$

where r denotes the continuous reward rate and g is a terminal reward.

As the number of agents tends to infinity, one might anticipate that the empirical distribution μ_t^N converges to a probability measure that describes the distribution of a single representative agent. This convergence allows for a significant simplification: the complex, high-dimensional N-agent system can be approximated by the dynamics of the representative agent, whose evolution is described by a stochastic differential equation of McKean-Vlasov type:

$$dX_t = b(t, X_t, \mathbb{P}^{X_t}, a_t) + \sigma(t, X_t, \mathbb{P}^{X_t}, a_t)dW_t,$$

where W is a Brownian motion and $\mathbb{P}^{X_t}$ denotes the distribution of X_t. The mean-field control problem is then to maximize the expected cumulative reward

$$\mathbb{E}\left[\int_0^T r(t, X_t, \mathbb{P}^{X_t}, a_t)dt + g(X_T, \mathbb{P}^{X_t})\right].$$

Because of this relationship, the terms *mean-field optimal control* and *optimal control of McKean-Vlasov processes* are frequently used synonymously in the literature. A foundational work in this context is Lacker (2017), which rigorously connects the optimal control of McKean-Vlasov processes with the control of large interacting particle systems. The informal problem description above is also based on Lacker (2017). In the literature, such models appear in numerous variations. For instance, an additional source of common noise, W^0, is often included in the model, affecting the dynamics of each agent; see, for example, Pham and Wei (2017). In the literature, two main approaches to solving McKean-Vlasov control problems have rapidly emerged. The *stochastic Pontryagin principle* is considered, among others, in Carmona and Delarue (2015), Andersson and Djehiche (2011) and Buckdahn et al. (2011). The term stochastic maximum principle is commonly used as well. The second method is the previously mentioned *dynamic programming*. In the present context, it is applied in Pham and Wei (2017), Pham and Wei (2018), Laurière and Pironneau (2014), Bayraktar et al. (2018), Djete et al. (2022). For a comparison of these two approaches, see Chapter 6 in Carmona and Delarue (2018a) or Carmona et al. (2013). The duality between Pontryagin's maximum principle and dynamic programming will be encountered again in the context of deterministic optimization in Chapter 4.

The case of discrete-time McKean-Vlasov optimal control problems has been less studied in the literature. Pham and Wei (2016) provide a dynamic programming approach, while Dong et al. (2022) and Ahmadova and Mahmudov (2023) address the optimization problem using the discrete-time stochastic maximum principle. A particularly significant and tractable subclass of discrete-time mean-field optimal control is the Linear-Quadratic (LQ) case, where the dynamics are linear in the state and control, and the cost functional is quadratic. This specific structure often allows for explicit solutions. For the finite time-horizon, Elliott et al. (2013) establish necessary and sufficient conditions for solvability and derive optimal controls in closed form. The authors extend these results to the infinite time-horizon setting in Ni, Elliott and Li (2015). Further contributions to the finite-horizon LQ-problem include Song and Liu (2021), Ni, Zhang and Li (2015), and Zhang and Qi (2016), while the infinite-horizon case is addressed in Song and Liu (2020), Ni et al. (2016), and Zhang et al. (2019).

Discrete-time mean-field Markov decision processes

When the state dynamics of the agents are described by a (Markovian) transition kernel rather than a McKean-Vlasov differential equation, the model is commonly referred to as a *mean-field Markov decision problem (MF-MDP)*. The continuous-time version of this framework constitutes the precise setting of this monograph. However, we first review the literature on MF-MDPs in discrete time, which is considerably more extensive. The presumably earliest work in this context is Bordenave and Anantharam (2007). They consider a system of N interacting particles on a finite state space, managed by a central controller selecting actions from a finite set of actions. The authors analyze the system's performance under three criteria: discounted reward, finite average reward, and ergodic (infinite-horizon) average reward. To do this, they employ a dynamic programming approach, formulating Bellman equations for both the finite N-particle system and its deterministic mean-field limit. They prove the convergence of the value functions of the N-agent problems to the value function of the limiting mean-field problem as $N \to \infty$. Furthermore, they show that any limit point of a sequence of optimal policies from the N-agent system is indeed an optimal policy for the mean-field model. They also propose the continuous-time model as a possible extension of their work and stress the technical issue of proving the tightness of the optimal control process. In this monograph, we solve this issue by introducing a suitable topology.

Gast and Gaujal (2011) subsequently extended the results of Bordenave and Anantharam (2007) for MF-MDPs in discrete time. The authors consider a finite state space and a compact action space. Under Lipschitz continuity assumptions on the model parameters, they establish a rate of convergence of order $1/\sqrt{N}$ for the value functions. Further, they prove that an optimal policy of the deterministic mean-field limit model is asymptotically optimal when applied to the N-agent model.

Motivated by Gast and Gaujal (2011), Higuera-Chan et al. (2016) considered a similar framework with the difference that the state transition additionally depends on a random disturbance density which is unknown for the central controller. For the discounted infinite-horizon optimality criterion, they introduce the notion of "eventually optimal" policies for the deterministic mean-field limit model. This weaker notion of optimality is necessary because the controller must learn the unknown disturbance density over time, meaning the policy only converges to the true optimum as time progresses and the statistical estimates improve. Further, the authors prove that eventually optimal policies of the mean-field limit model are "eventually asymptotically optimal" for the N-agent model. In the subsequent paper Higuera-Chan et al. (2017), the same authors address the MF-MDP from a different perspective by framing it as a game against nature. This approach is designed for situations where an unknown parameter (for instance, the distribution of a random disturbance as above) is not only unknown but also unobservable and potentially changing at every time step. Since statistical estimation is not feasible in this setting, the problem is modeled as a game between the central controller and an adversary, referred to as "nature". Nature's goal is to maximize the system's cost by choosing the worst possible value for the unknown parameter at each stage from a given set. Consequently, the controller's objective shifts to finding a minimax policy, which minimizes the cost under this worst-case scenario. The authors show that the minimax policy derived for the deterministic mean-field limit model is asymptotically minimax for the original N-agent system.

In Motte and Pham (2022), the authors analyze discrete-time MF-MDPs where the system is influenced by both idiosyncratic noise (specific to each agent) and a common noise (a single random factor affecting all agents simultaneously). The framework considers compact state and action spaces and focuses on the discounted infinite-horizon optimality criterion. An exceptional property of the model is that both the reward and the transition may depend on the empirical distribution of actions. Another central aspect is the restriction to open-loop controls. An open-loop control is a policy that is determined at the beginning of the process and does not adapt to the system's evolving state. This is in contrast to a feedback (or closed-loop) control, which reacts to the current state of the system at each time step. In Motte and Pham (2022), an open-loop control is modeled as a strategy that depends on time, the agent's initial state, and the realizations of both the common noise and the agent's own idiosyncratic noise, but explicitly not on the agent's state. The open-loop controls are symmetric in the sense that the central controller assigns the same control function to each agent. In this setting, the authors establish that as the number of agents N tends to infinity, the N-agent system converges to a limiting mean-field model. Due to the presence of common noise, this limit is not a deterministic process (as in the previously discussed papers) but a stochastic process which they call "conditional McKean-Vlasov MDP". Under Lipschitz assumptions, the authors derive bounds on the rate of convergence for the value functions and establish the asymptotic optimality of optimal policies of the mean-field limit when applied in the finite N-agent system. Furthermore, the authors show

that their conditional McKean-Vlasov MDP is equivalent to a measure-valued MDP. This equivalence is crucial, as it allows them to apply a dynamic programming approach and formulate a Bellman equation for the limit model. The analysis of this equation proves the existence of optimal policies, which are shown to be randomized feedback policies. The authors emphasize that this need for randomization is a key departure from classical MDP theory, where deterministic policies are sufficient. In their companion paper Motte and Pham (2023), the same authors derive similar results with the main difference being a significant generalization of the admissible open-loop controls. While the first paper considers a symmetric framework where all agents are assigned the same control function, this work allows for fully asymmetric policies, in the sense that each agent can follow a completely different open-loop strategy. Furthermore, the information available to each control is expanded. A policy for a specific agent can now depend on the idiosyncratic noises of the entire population, not just its own. This approach provides additional flexibility and generality, which in turn requires different proof techniques than in Motte and Pham (2022).

Bäuerle (2023) operates in a similar setting, considering compact state and action spaces and common noise. The transition probabilities and rewards of an agent depend not only on its own state and action, but additionally on the empirical distribution of the entire population. The primary focus is to explicitly frame the problem within classical MDP theory. A key distinction from Motte and Pham (2022) is that this is achieved before taking the limit: the paper first establishes that the original N-agent problem is equivalent to a classical MDP on the space of empirical measures. As $N \to \infty$, a mean-field limit MDP is obtained, which remains stochastic due to the presence of the common noise. For the discounted reward criterion, the paper proves the convergence of the value functions and the asymptotic optimality of optimal policies derived from the mean-field limit MDP. Furthermore, the author shows that corresponding results hold for the average reward optimality criterion.

Solving the mean-field limit problem remains a central challenge for the practical application of the mean-field approximation in discrete-time MF-MDPs. To address this challenge, several authors have proposed deep reinforcement learning (RL) methods. A rigorous theoretical treatment of this approach is presented in Carmona et al. (2023). The authors study the discounted infinite-horizon problem where both the state dynamics and the actions are subject to both idiosyncratic and common noise. To solve the resulting measure-valued mean-field limit MDP, they develop a Q-learning algorithm based on a discretization of the limit MDP as well as a deep RL method that avoids discretization.

Papers that focus on the implementation of RL methods to real-world applications are Zhu et al. (2021) and Cui et al. (2021). In the first paper, the authors study a ride-sourcing market in which a platform (taxi company) aims to incentivize a large number of drivers to optimize a combination of corporate profit and system-wide service rate. In the second paper, the authors investigate a queueing system consisting of a large number of schedulers

that allocate jobs to a static number of parallel servers, each equipped with an individual finite queue. The aim of the schedulers is to assign jobs to the servers in such a way that the number of jobs lost due to overload in certain servers is minimized.

An intermediate step toward continuous-time MF-MDPs is the paper by Gast et al. (2012). The authors consider a discrete-time MF-MDP with finite state space and compact action space. The main difference that distinguishes this paper from the previous literature is an innovative time-scaling approach. They transform the discrete-time measure-valued state process into a continuous-time process by affine interpolation, where time is rescaled by an intensity function $I(N)$ that tends to zero as the number of agents N tends to infinity. In the mean-field limit, they obtain a continuous-time deterministic optimization problem. This continuous-time problem is solved using a dynamic programming approach via the corresponding HJB equation. The authors then show that the optimal control derived from this HJB equation, when rescaled to discrete time, provides an asymptotically optimal open-loop policy for the original N-agent system.

Instead of scaling the time axis of discrete-time MF-MDPs to obtain a continuous-time limit, it is natural to consider continuous-time MF-MDPs directly, which is precisely the framework considered in this monograph.

CONTINUOUS-TIME MEAN-FIELD MARKOV DECISION PROCESSES

In comparison to the discrete-time case, the literature on continuous-time MF-MDPs is very sparse. Presumably the most important work in this context is Cecchin (2021). The author investigates a discounted finite-horizon reward model with finite state and compact action space. As Bäuerle (2023) does in the discrete-time case, Cecchin (2021) shows the equivalence between the original high-dimensional N-agent formulation on the joint state space and a more tractable reformulation on the space of empirical measures. In the mean-field limit, a deterministic optimal control problem is obtained, where the value function, under Lipschitz continuity assumptions on the model parameters, is characterized as the unique viscosity solution to the corresponding HJB equation. Under the same assumptions, the author establishes a convergence rate of order $1/\sqrt{N}$ for the value functions. The proof is based on the characterization of the value function as the viscosity solution. Under additional convexity assumptions, convergence rates for the optimal trajectories of the state processes are shown.

Building on the theoretical work of Cecchin (2021), Hofgard et al. (2024) focus on numerical methods for the solution of the corresponding mean-field limit problems. Based on the characterization of the value function as the unique viscosity solution of an HJB equation, they establish both an existence and convergence result for the deep Galerkin method, a deep learning-based algorithm well suited for the numerical solution of high-dimensional nonlinear PDEs.

1.4. Outline and contributions

The aim of this monograph is to contribute to a more profound understanding of continuous-time mean-field Markov decision models and to present the utility of the theory in real-world applications. In Chapter 3, we provide a rigorous theoretical treatment of continuous-time MF-MDPs and establish the central results on convergence and optimality. We consider a large number N of cooperative, interacting agents whose states take values in a finite state space S, collectively aiming to maximize their expected social reward. Under weaker continuity assumptions than in Cecchin (2021), we prove the relative compactness of the sequence of state-action processes for an increasing number of agents, which implies the existence of a weakly converging subsequence. For the corresponding limit state process, which turns out to be deterministic, we derive a characterization by an ordinary differential equation. To address the difficulty of establishing the tightness of the action process, as outlined by Bordenave and Anantharam (2007), an appropriate topological framework is applied, namely the Young topology.

In the mean-field limit, we obtain a deterministic optimization problem for which we provide conditions ensuring that an optimal control of the deterministic limit problem constitutes an asymptotically optimal action process for the (stochastic) system with N agents. Under Lipschitz conditions, we derive a convergence rate of order $1/\sqrt{N}$ for the sequence of value functions. In contrast to Cecchin (2021), the proof is more direct and requires less technical overhead, and we address both the finite- and infinite-horizon case.

An important aspect that has received little attention in the literature on continuous-time MF-MDPs is the translation of the theory into real-world applications. We address this in greater detail in Chapter 4. To solve the corresponding deterministic limit problems, we rely on Pontryagin's maximum principle, in contrast to the majority of the MF-MDP literature, which pursues a dynamic programming approach for the mean-field limit problems. A unique aspect is that we demonstrate, by example, how to model situations in which the number of agents N is not constant but instead follows a birth-and-death process. Such a scenario has not been considered in any of the works cited so far. The specific application in Section 4.3 examines the optimal control of a black bear population.

For both Chapters 3 and 4, a more detailed outline is presented at the beginning. Before turning to the mean-field framework, we first provide the theoretical background on classical continuous-time Markov decision models in Chapter 2.

CHAPTER 2

Continuous-time Markov decision models

The aim of this chapter is to provide the theoretical background on Markov decision models in continuous time. We consider a continuous-time stochastic process on a set of states S. A decision maker can choose actions from a set of actions A to influence the random (Markovian) transition of the process to the next state. In addition, the decision maker receives a continuous reward rate depending on the current state of the process and the currently chosen action. The aim of the decision maker is to find a process of optimal actions to maximize the expected reward over the considered time horizon. To do so, two effects must be balanced: Maximizing the immediate reward as well as keeping the process in favorable states to maximize the future rewards.

Continuous-time Markov decision models are a well-studied topic in the theory of stochastic processes. They form the foundation for continuous-time mean-field Markov decision models. In the following, we give a brief overview of the notation, the construction of the underlying process, and the solution theory. We primarily rely on the monograph by Piunovskiy and Zhang (2020).

Notation. The term *Borel space* in the following definition refers to a Borel subset of a complete separable metric space. For a Borel space M, we denote the Borel-σ-algebra on M by $\mathcal{B}(M)$, the power set by $\mathcal{P}(M)$, and the set of all probability distributions on M by $\mathbb{P}(M)$.

Definition 2.1 (Continuous-time Markov decision model)**.** A continuous-time Markov decision model can be characterized by

$$(S, A, q, r),$$

where

- S is the set of *states*, which we refer to as the *state space*. We assume S to be a nonempty Borel space with the corresponding σ-algebra $\mathcal{B}(S)$.
- A is the *action space* which we assume to be a nonempty Borel space with Borel-σ-algebra $\mathcal{B}(A)$. The elements of A are called *actions*.
- q is a signed kernel that models the transition intensities from one state to the next, i.e.,
$$q : S \times A \times \mathcal{B}(S) \to \mathbb{R}, \quad (x, a, \Gamma) \mapsto q(\Gamma|x, a) = \int_\Gamma q(dy|x, a).$$
Note that the transition intensity depends only on the current state and chosen action, which, among other properties, makes the model Markovian. We make the following assumptions on q:

  ($\widetilde{Q}1$) $q(\Gamma \setminus \{x\}|x, a) \geq 0$ for all $x \in S,\ \Gamma \in \mathcal{B}(S),\ a \in A$.

  ($\widetilde{Q}2$) $q(S|x, a) = 0$ for all $x \in S,\ a \in A$.

  ($\widetilde{Q}3$) $\sup_{a \in A} |q(\{x\}|x, a)| < \infty$ for all $x \in S$.
- r is the reward function. In every state-action combination, the decision maker receives a reward rate according to a measurable function
$$r : S \times A \to [-\infty, \infty), \quad (x, a) \mapsto r(x, a).$$

Remark 2.2. An additional parameter of the model that is frequently considered is the set of all feasible state-action combinations $D \subset S \times A$. The set of all admissible actions in state $x \in S$ is then given by

$$D(x) := \{a \in A \mid (x, a) \in D\}.$$

In the examples considered later, a state-dependent restriction of the action space is not necessary. Generally, in most applications, a non-admissible action $a \in A \setminus D(x)$ can be associated with an unfavorable reward $r(x, a)$ which prevents the decision maker from choosing action a in state x. We therefore refrain from restricting the admissible state-action combinations in the following.

The aim of the controller is to find a process of somewhat "optimal" actions. As it turns out, in the optimal policy the choice of an action depends only on the current state, not on the entire history of states. In addition, the applied action is updated only in the event of

a change of state. In particular, varying the action continuously between two state changes does not increase the value of the control problem. Accordingly, we restrict ourselves to the class of *Markov* policies for the control of the model, since they are sufficient for the topics to come. For an overview of more general concepts of policies, see, e.g., Definitions 1.1.2 and 1.1.3 in Piunovskiy and Zhang (2020).

In the definition below, we allow for relaxed strategies, that is, in contrast to a classical Markov policy, the controller does not have to choose a specific action from the action space but can decide in favor of a probability distribution on A. We call such a policy *relaxed* Markov policy.

Definition 2.3 (Markov policies)**.**

a) A *relaxed Markov policy* is given by a sequence $(\pi_n)_{n\in\mathbb{N}}$ of stochastic kernels

$$\pi_n : S \times \mathcal{B}(A) \to [0,1], \quad (x, \mathcal{A}) \mapsto \pi_n(\mathcal{A} \mid x)$$

such that

i) $x \mapsto \pi_n(\mathcal{A} \mid x)$ is $\mathcal{B}(S)$-measurable for all $n \in \mathbb{N}$ and $\mathcal{A} \in \mathcal{B}(A)$.

ii) $\mathcal{A} \mapsto \pi_n(\mathcal{A} \mid x)$ is a probability measure on $\mathcal{B}(A)$ for all $x \in S$ and $n \in \mathbb{N}$.

b) A relaxed Markov policy $(\pi_n)_{n\in\mathbb{N}}$ is called *relaxed stationary* if $\pi_n \equiv \pi$, i.e., the policy is independent of time.

c) A *deterministic stationary Markov policy* is given by a measurable function

$$\pi : S \to A, \qquad x \mapsto \pi(x) \in A.$$

Let Π be the set of all relaxed Markov policies, Π^s the set of all relaxed stationary Markov policies and Π^d the set of all deterministic stationary Markov policies. Then we have

$$\Pi^d \subset \Pi^s \subset \Pi.$$

The interpretation of a relaxed Markov policy to control a specific problem is the following: After the n-th change of state, the controller instantly applies a new action *distribution* according to π_{n+1}, depending on the current state. This action distribution is continuously applied until the next change of state.

The Markov policies introduced above can be categorized as *feedback policies* or *closed-loop policies*, since the chosen action of the controller depends on the current state of the system. In contrast, an *open-loop policy* determines actions without knowledge of the actual state. Thus, an open-loop policy is simply a function $\pi^{\mathrm{ol}} : [0,\infty) \to A$ that specifies an action for every point in time $t \in [0,\infty)$.

The above transition intensity is defined in terms of a fixed action $a \in A$. Considering the transition intensity for a relaxed Markov policy $(\pi_n)_{n\in\mathbb{N}}$ after the n-th jump, we set for $\Gamma \in \mathcal{B}(S)$

$$q(\Gamma|x, \pi_{n+1}) := \int_A q(\Gamma|x, a)\pi_{n+1}(da|x).$$

In a similar way, the rate of reward that the controller receives during the time interval between the n-th state change and the $(n+1)$-th state change while being in state $x \in S$ is given by

$$r(x, \pi_{n+1}) = \int_A r(x, a)\pi_{n+1}(da|x).$$

2.1. CONSTRUCTION OF THE UNDERLYING PROBABILITY SPACE

Before turning to the solution theory of continuous-time Markov decision processes, we first give a rigorous description of the underlying probability space together with the state and action processes defined on it. In the following, capital letters denote random variables, while lower-case letters denote the corresponding realizations.

The *state process* $(X_t)_{t\geq 0}$ is a stochastic process, i.e., a family of random variables

$$X_t : \Omega \to S, \qquad t \in [0, \infty)$$

on the measurable space

$$(\Omega, \mathcal{F}) := \Big((S \times [0, \infty))^\infty, \mathcal{B}((S \times [0, \infty))^\infty)\Big).$$

We denote an element of Ω by

$$\omega = (x_0, t_1, x_1, t_2, x_2, t_3, \ldots).$$

For the elements of Ω we define the canonical projections

$$\begin{aligned} \tilde{X}_n &: \Omega \to S, & \tilde{X}_n(\omega) &= x_n, \quad n \in \mathbb{N}_0, \\ \tau_n &: \Omega \to [0, \infty), & \tau_n(\omega) &= t_n, \quad n \in \mathbb{N}. \end{aligned}$$

Further, we set

$$T_n := \sum_{j=1}^{n} \tau_j, \quad n \in \mathbb{N}, \qquad T_0 := 0.$$

Then the (controlled) *state process* is given by

$$X_t := \sum_{n\in\mathbb{N}_0} \mathbb{1}_{\{T_n \leq t < T_{n+1}\}} \tilde{X}_n, \qquad t \in [0, \infty).$$

The interpretation of the construction is as follows: The random variables τ_n describe the sojourn times in the states $\tilde{X}_{n-1}$. Based on the sojourn times, T_n describes the time of the

n-th jump of the process and $\tilde{X}_n$ the state of the process on the interval $[T_n, T_{n+1})$. One can easily observe that by construction the embedded discrete-time process of $(X_t)_{t\geq 0}$ is just $(\tilde{X}_n)_{n\in\mathbb{N}_0}$, and that the paths of the state process are piecewise constant and càdlàg (see Definition B.15).

The n-term history of the process is defined by

$$\begin{aligned} H_0 &:= X_0, \\ H_n &:= (X_0, \tau_1, X_1, \tau_2, \ldots, \tau_n, X_n), \qquad n \in \mathbb{N}. \end{aligned}$$

We denote the set of all n-term histories by

$$\mathbf{H}_n := \{(x_0, \tau_1, x_1, \ldots, \tau_n, x_n) \mid (x_0, \tau_1, x_1, \ldots) \in \Omega\}.$$

Since we consider relaxed Markov policies, the chosen action depends only on the current state of the process. Thus, the *action process* $(\pi_t)_{t\geq 0}$ corresponding to a policy $(\pi_n)_{n\in\mathbb{N}} \in \Pi$ is defined by

$$\pi_t := \sum_{n\in\mathbb{N}_0} \mathbb{1}_{\{T_n < t \leq T_{n+1}\}} \pi_{n+1}(\,\cdot\,|\tilde{X}_n) + \mathbb{1}_{\{t=0\}} \pi_1(\,\cdot\,|\tilde{X}_0), \qquad t \in [0,\infty).$$

In contrast to the state process, the action process has piecewise constant càglàd paths. This means that a new decision can only be made after a change of state has already occurred.

To complete the probability space, it remains to construct the probability measure $\mathbb{P}^\pi_\gamma(d\omega)$ which depends on the initial state distribution $\gamma \in \mathbb{P}(S)$ and the chosen policy $\pi = (\pi_n)_{n\in\mathbb{N}} \in \Pi$ and is therefore referred to as the *strategic measure*. We construct $\mathbb{P}^\pi_\gamma(d\omega)$ on Ω recursively on the spaces of histories $\mathbf{H}_0, \mathbf{H}_1, \ldots$ using the Ionescu-Tulcea theorem.
The initial distribution $\gamma(dx_0)$ is the distribution of $H_0 = X_0$. To apply the theorem, define for any $n \in \mathbb{N}$ the stochastic kernels G_n on $\mathcal{B}([0,\infty) \times S)$ given $h_{n-1} \in \mathbf{H}_{n-1}$ by

$$G_n : \mathbf{H}_{n-1} \times \mathcal{B}([0,\infty) \times S) \to [0,1],$$

$$G_n(\Gamma_\tau \times \Gamma_X | h_{n-1}) = \int_{\Gamma_\tau} q(\Gamma_X \setminus \{x_{n-1}\} | x_{n-1}, \pi_n) e^{q(\{x_{n-1}\}|x_{n-1},\pi_n)\cdot s} ds$$

for all $\Gamma_\tau \in \mathcal{B}([0,\infty))$ and $\Gamma_X \in \mathcal{B}(S)$. Strictly speaking, the stochastic kernels depend only on the current state x_{n-1} and not on the entire history h_{n-1} since we consider Markov policies. Nevertheless, this notation is convenient for the application of version A.1 of Ionescu-Tulcea's theorem.

Applying the Ionescu-Tulcea theorem, we obtain the corresponding unique probability measure $\mathbb{P}^\pi_\gamma$ on $(\Omega, \mathcal{F})$. Let $\Gamma_X \in \mathcal{B}(S)$, $\Gamma_\tau \in \mathcal{B}([0,\infty))$ and $\Gamma_H \in \mathcal{B}(\mathbf{H}_{n-1})$ be such that $\Gamma_H = \Gamma_{H_0} \times \Gamma_{H_1} \times \ldots \times \Gamma_{H_{n-1}}$ where $\Gamma_{H_0} \in \mathcal{B}(S)$ and $\Gamma_{H_1}, \ldots, \Gamma_{H_{n-1}} \in \mathcal{B}([0,\infty) \times S)$.

Then the strategic measure $\mathbb{P}_\gamma^\pi$ satisfies

$$\mathbb{P}_\gamma^\pi(H_0 \in \Gamma_X) = \mathbb{P}_\gamma^\pi(X_0 \in \Gamma_X) = \gamma(\Gamma_X)$$

and

$$\begin{aligned}
&\mathbb{P}_\gamma^\pi(\tau_n \in \Gamma_\tau, X_n \in \Gamma_X, H_{n-1} \in \Gamma_H) \\
&= \int_{\Gamma_H} G_n(\Gamma_\tau \times \Gamma_X | h_{n-1})\ \mathbb{P}_\gamma^\pi(H_{n-1} \in dh_{n-1}) \\
&= \int_{\Gamma_{H_{n-1}}} \int_{\Gamma_H} G_n(\Gamma_\tau \times \Gamma_X | h_{n-1})\ G_{n-1}(dh_{n-1}|h_{n-2})\ \mathbb{P}_\gamma^\pi(H_{n-2} \in dh_{n-2}) \\
&= \dots \\
&= \int_{\Gamma_0} \int_{\Gamma_1} \dots \int_{\Gamma_{H_{n-1}}} \int_{\Gamma_H} G_n(\Gamma_\tau \times \Gamma_X | h_{n-1})\ G_{n-1}(dh_{n-1}|h_{n-2}) \dots G_1(dh_1|h_0)\ \gamma(dx_0).
\end{aligned}$$

Remark 2.4. A state $x \in S$ is called absorbing if $q(\{x\}|x,a) = 0$. In an absorbing state, the sojourn time is infinite, which complicates the notation of the probability space; see, e.g., the discussion in Section 1.1.3.1 in Piunovskiy and Zhang (2020). There, the authors introduce an artificial isolated point $x_\infty \notin S$ and allow ∞ as a sojourn time. That is why, in order to keep the notation simple, we tacitly assume here that no absorbing states exist. If so, the issue can be avoided by dividing the absorbing state x_{abs} into two states x_{abs}^1 and x_{abs}^2, which continuously alternate with a certain intensity.

Figure 2.1.: $q(\{x_{abs}\}|x_{abs},\cdot) = 0$ and $q(\{x_{abs}^2\}|x_{abs}^1,\cdot)$, $q(\{x_{abs}^1\}|x_{abs}^2,\cdot) > 0$.

2.2. Formulation and solution of the optimal control problem

To compare different policies and ultimately determine optimal policies, we consider the expected discounted reward over the time horizon as the optimization criterion. Thus, let $\beta > 0$ be the fixed discount rate.

Definition 2.5 (Value of a policy)**.**

a) Let $\pi = (\pi_n)_{n\in\mathbb{N}} \in \Pi$ be an arbitrary relaxed Markov policy with associated action process $(\pi_t)_{t\geq 0}$. The *value of a policy* $(\pi_n)_{n\in\mathbb{N}}$ given some initial state $x \in S$ is

defined by

$$V_\pi(x) = \mathbb{E}_x^\pi \left[\int_0^\infty e^{-\beta t} r(X_t, \pi_t) dt \right] = \mathbb{E}_x^\pi \left[\int_0^\infty e^{-\beta t} \int_A r(X_t, a) \pi_t(da) dt \right].$$

b) The *value function* of the discounted continuous-time Markov decision problem (CTMDP) is given by

$$V(x) := \sup_{\pi=(\pi_n)\in\Pi} V_\pi(x), \qquad x \in S.$$

c) A policy $\pi = (\pi_n) \in \Pi$ is called *(uniformly) optimal* for the discounted CTMDP if

$$V_\pi(x) = V(x) \qquad \forall\ x \in S.$$

Based on the definitions above, the main optimization problem of the discounted continuous-time Markov decision problem can be stated as follows:

(CTMDP) Find a policy $\pi^* = (\pi_n^*) \in \Pi$, such that for all $x \in S$

$$V_{\pi^*}(x) = \sup_{\pi=(\pi_n)\in\Pi} \mathbb{E}_x^\pi \left[\int_0^\infty e^{-\beta t} \int_A r(X_t, a) \pi_t(da) dt \right] = V(x).$$

In order to obtain a well-defined optimization problem, that is, to ensure the existence of a (unique) solution, we have to impose several conditions on the parameters of the model.

Condition (A).
There exist a measurable function $w : S \to [0, \infty)$ and constants $\rho \in \mathbb{R},\ L, b \in [0, \infty)$ such that:

a) $\bar{q}_x := \sup_{a\in A} -q(\{x\}|x, a) \le Lw(x)$ for each $x \in S$.

b) For each $x \in S$ and $a \in A$

$$\int_S w(y) q(dy|x, a) \le \rho w(x) + b.$$

Condition (B).
Condition (A) is satisfied, and there exist a measurable function $w' : S \to [1, \infty)$ and constants $L', b' \ge 0$ and $\rho' \in \mathbb{R}$ such that:

a) $(\bar{q}_x + 1)w'(x) \le L'w(x)$ for each $x \in S$. Here, $w : S \to [0, \infty)$ is the same function as in Condition (A).

b) For each $x \in S$ and $a \in A$

$$\int_S w'(y) q(dy|x, a) \le \rho' w'(x) + b'.$$

c) $\beta > \rho'$.

d) There exists a constant $M' \geq 0$ satisfying

$$|\sup_{a \in A} r(x,a)| \leq M' w'(x), \qquad \forall\, x \in S. \tag{2.1}$$

Condition (C).

a) There exists a continuous function $w' : S \to [1, \infty)$ such that parts b), c), and d) of Condition (B) are satisfied.

b) For each bounded continuous function u on S, the function

$$S \times A \ni (x,a) \mapsto \int_{S \setminus \{x\}} u(y) w'(y) q(dy|x,a)$$

is continuous.

c) The function $(x,a) \mapsto r(x,a)$ is upper semicontinuous (see Definition A.2), where $(x,a) \in S \times A$.

d) The action space is compact.

If the model consists of only a finite number of states, the conditions are considerably simplified.

Lemma 2.6. *In the case of a finite state space, it holds that*

i) Condition (A) is satisfied with $w(x) \equiv 1$,

ii) Condition (B) is satisfied with $w'(x) \equiv 1$.

Proof. Since the number of states is finite, we have

$$\bar{q}_x \leq \max_{x \in S} \bar{q}_x =: L, \qquad \text{resp.} \qquad \bar{q}_x + 1 \leq L + 1 =: L' \qquad \forall\, x \in S,$$

which concludes part a) of Conditions (A) and (B).
With $w(x) \equiv w'(x) \equiv 1$ part b) is trivial in both conditions since

$$\int_S q(dy|x,a) = 0,$$

choose, for example, $\rho = \rho' = 0$ and $b, b' \in [0, \infty)$ arbitrarily. By assumption, the discount rate is positive, that is, $\beta > 0$, which concludes part c) of Condition (B). Part d) of Condition (B) follows by setting $M' := \max_{x \in S} |\sup_{a \in A} r(x,a)|$. □

As a first result, we state that the value of an arbitrary relaxed Markov policy is bounded. Thus, the optimization problem (CTMDP) is well-defined in the sense that the value function $V(x)$ is finite for every $x \in S$.

Theorem 2.7. *Suppose Conditions (A), (B), and (C) are satisfied (for the same function w'). Then, for each policy $\pi = (\pi_n) \in \Pi$ we have*

$$V_\pi(x) \leq \frac{M'(\beta w'(x) + b')}{\beta(\beta - \rho')} < \infty.$$

Based on the function w' in Condition (C), we define the set of all w'-bounded measurable functions on S:

$$\mathbf{B}_{w'}(S) := \left\{ u : S \to \mathbb{R} \mid u \text{ measurable and } \sup_{x \in S} \frac{|u(x)|}{w'(x)} < \infty \right\}.$$

The space $\mathbf{B}_{w'}(S)$ is complete and is often referred to as the *weighted normed space*.

The following theorem suggests a candidate u^* for the optimal value V that is the unique solution of the so-called *optimality equation* or *Bellman equation*. The second part of the theorem states that there exists a measurable mapping $\varphi^* : S \to A$ that maximizes the optimality equation and thus is a candidate for the optimal control of the system.

Theorem 2.8 (Optimality equation, measurable selection and verification theorem)**.** *Suppose Conditions (A), (B) and (C) are satisfied. Then the optimality equation*

$$\beta u(x) = \sup_{a \in A} \left\{ r(x,a) + \int_S u(y) q(dy|x,a) \right\}, \qquad \forall\ x \in S \tag{2.2}$$

admits a unique upper semicontinuous solution $u^* \in \mathbf{B}_{w'}(S)$.
Further there exists a measurable mapping $\varphi^* : S \to A$ which provides the supremum in (2.2):

$$\beta u^*(x) = \sup_{a \in A} \left\{ r(x,a) + \int_S u^*(y) q(dy|x,a) \right\} \tag{2.3}$$

$$= r(x, \varphi^*(x)) + \int_S u^*(y) q(dy|x, \varphi^*(x)), \qquad \forall\ x \in S. \tag{2.4}$$

Each measurable mapping $\varphi^* : S \to A$ that satisfies (2.4) defines a deterministic stationary strategy, also denoted by φ^*, which is uniformly optimal for the discounted CTMDP:

$$V(x) = V_{\varphi^*}(x) = u^*(x), \qquad \forall\ x \in S.$$

Theorems 2.7 and 2.8, together with the proofs, can be found in Piunovskiy and Zhang (2020) (Lemma 3.1.1, Theorem 3.1.1, and Theorem 3.1.2). Note that in the monograph they consider Markov Decision theory from the perspective of a minimization problem, which is of course equivalent to the maximization problem presented here. Nevertheless, this leads to some minor modifications of the theorems and the corresponding conditions: Obviously, instead of infima we consider suprema in (2.1), (2.2), (2.3). Further, in part c) of Condition (C), we require the reward rate r to be upper semicontinuous. As a result,

the solution u^* of the optimality equation (2.2) is also upper semicontinuous. Despite these modifications, the character of the proofs remains the same and the adaptations can be implemented in a straightforward manner. Therefore, we limit ourselves to referring to the proofs in Piunovskiy and Zhang (2020) here.

The very core of the solution theory is the fact that, under the imposed conditions, we always find an optimal deterministic stationary policy, implying that no relaxation of the actions is needed. Consequently, we can restrict ourselves to the set Π^d in the search for an optimal policy.

CHAPTER 3

Continuous-time mean-field Markov decision models

The aim of this chapter is to provide a rigorous theoretical treatment of continuous-time mean-field Markov decision models and to establish the central results on convergence and optimality. We consider a large number N of cooperative, interacting agents whose states take values in a finite state space S, collectively aiming to maximize their expected social reward.

An intuitive approach to embedding this scenario in an MDP framework is to consider the joint state space S^N and to capture the states of all agents in an N-dimensional state process $(X^1, \dots, X^N)$ as presented in Section 3.1. Due to the nature of its construction, we refer to this formulation as the *multi-agent model*. A drawback of this approach is that the state space grows exponentially with the number of agents, resulting in a computationally intractable optimization problem.

Therefore, in Section 3.2, we adopt a different approach. Instead of tracking the individual states of all agents, we only observe the empirical distribution μ of the population over the states. The new state process is thus given by the vector $\mu = (\mu^1, \dots, \mu^{|S|})$. Although this simplification means losing information about the exact state of each agent, it does not decrease the value of the problem due to the symmetry of the model parameters. In fact, we show the equivalence between both the multi-agent and the measure-valued approach in the sense that both formulations generate the same optimal expected reward. This method of reformulating the problem is well-established in the MF-MDP literature and has been employed in similar frameworks, for example, by Bäuerle (2023) in the discrete-time case and by Cecchin (2021) in the continuous-time case. In contrast to Cecchin (2021), we can

state our solution theorem for the N-agent model (in the form of a Bellman equation) under weaker continuity conditions.

In Section 3.3, we study the mean-field limit that is obtained by letting the number of agents N tend to infinity. We prove that the sequence of state-action processes for an increasing number of agents is relatively compact, which implies the existence of a weakly converging subsequence. We resolve the difficulty pointed out by Bordenave and Anantharam (2007) of establishing the tightness of the action process by endowing the corresponding space with the Young topology. Additionally, we prove that the limit state process is deterministic and characterized by an ordinary differential equation. Mere continuity assumptions are sufficient for these results.

The corresponding mean-field limit problem turns out to be a deterministic optimization problem, which is analyzed in Section 3.4. We show that the optimal value of the limit model serves as an asymptotic upper bound for the N-agent value functions. In addition, we provide conditions under which an optimal control of the limit problem constitutes an asymptotically optimal action process for the (stochastic) system with N agents.

While the preceding sections focus on infinite-horizon models, Section 3.4.1 extends the results to the finite-horizon case. Under Lipschitz conditions on the transition intensity and the reward, we establish a convergence rate of order $1/\sqrt{N}$ for the value functions towards the optimal value of the limit model. In contrast to Cecchin (2021), we show this in a straightforward manner without the use of viscosity solutions. To conclude the chapter, we demonstrate that the previously derived results still hold if resource constraints are imposed on the actions available to the central controller.

Similar deterministic optimization problems to the mean-field limit problem in Section 3.4 arise in the context of *fluid problems* within queueing theory. In the field of queueing networks, *fluid scaling* is a mathematical method used to convert a complex, stochastic system into a simplified, deterministic fluid model; see, among others, Bäuerle (2000, 2002) and Chen (1995). In the stochastic network, jobs arrive randomly according to a certain arrival rate and are processed by one or more servers, each operating at a specific service rate. The corresponding state process represents the number of jobs in each queue. When analyzing this complex system using fluid scaling, two transformations are applied to the process:

i) Time Scaling: The process is accelerated by scaling time by a factor γ.

ii) Space Scaling: The state space is scaled down by the same factor γ, which effectively reduces the magnitude of individual jumps.

As $\gamma \to \infty$, the stochastic state process converges to a deterministic limit process, the *fluid limit*, which is characterized by an ordinary differential equation with a structure similar to the one we establish for the mean-field limit state process in Section 3.3. A notable property of MF-MDPs is that fluid scaling arises "naturally" in the following sense: In

the N-agent system, a state change by a single agent induces a jump of size $1/N$ in the empirical distribution, which corresponds to the space scaling in fluid models as the number of agents grows. On the other hand, the principle of time scaling is also reflected in the N-agent system. As the number of agents N increases, the collective rate of state changes scales proportionally with N. Consequently, the average time between any two consecutive jumps in the empirical distribution is inversely proportional to N, which is equivalent to the time acceleration in the fluid model. We present an application that investigates a stochastic queueing network through the lens of continuous-time MF-MDPs in Section 4.5.

Having outlined the structure of this chapter, we now proceed with the formal description of the basic multi-agent model.

3.1. The multi-agent continuous-time Markov decision process

We consider a finite number N of agents, each moving on a *finite* set of states S according to a continuous-time Markov decision process. We refer to the entirety of the agents as the system (of agents) in reference to particle systems in physics. The vector $\mathbf{x}_t = (x_t^1, \dots, x_t^N) \in S^N$ describes the state of the system at time $t \in [0, \infty)$, where x_t^k is the state of the agent $k \in \{1, \dots, N\}$. The action space is the same for each agent and is given as a compact Borel space A. The action space of the system is accordingly A^N. We denote an action of the system by $\mathbf{a} = (a^1, \dots, a^N) \in A^N$, where a^k is the action chosen by the agent $k \in \{1, \dots, N\}$. We assume that there are no restrictions on the admissible actions depending on the current state; that is, the entire action space A is available in every state. Note that states, actions, etc. of the system are marked in bold, in contrast to states and actions of individuals.

We impose the following independence assumption on the agents:

(I) Given the current state of the system $\mathbf{x} \in S^N$, the next transition of an agent occurs independently of the other agents.

In addition, the agents are *statistically equal* in the sense that the numbering of the agents does not affect the statistical properties of the particle system. More precisely, the parameters of the system, such as state and action space, one-agent transition intensity, one-agent reward of the underlying Markov decision process, do not depend on the number $k = 1, \dots, N$ of the agent; see the upcoming definitions.

In the description of the model, we roughly follow the order of Chapter 2. For the construction of the system state process, we follow the notation of Section 2.1 as well as Piunovskiy and Zhang (2020). The state process of the system is defined on the measurable space $(\Omega, \mathcal{F}) := \big((S^N \times [0, \infty))^\infty, \mathcal{B}((S^N \times [0, \infty))^\infty\big)$.

We denote an element of Ω by $\omega = (\mathbf{x}_0, t_1, \mathbf{x}_1, t_2, \ldots)$. Now define

$$\begin{aligned}
\tilde{\mathbf{X}}_n &: \Omega \to S^N, \quad \tilde{\mathbf{X}}_n(\omega) = \mathbf{x}_n, \quad n \in \mathbb{N}_0, \\
\tau_n &: \Omega \to [0, \infty), \quad \tau_n(\omega) = t_n, \quad n \in \mathbb{N}, \\
T_n &:= \sum_{k=1}^{n} \tau_k, \quad n \in \mathbb{N}, \qquad T_0 := 0.
\end{aligned}$$

The controlled state process of the system is then given by

$$\mathbf{X}_t := \sum_{n \in \mathbb{N}_0} \mathbb{1}_{\{T_n \le t < T_{n+1}\}} \tilde{\mathbf{X}}_n, \qquad t \in [0, \infty).$$

The construction of the process can be interpreted as follows: The random variable τ_n describes the sojourn time of the system in state $\tilde{\mathbf{X}}_{n-1}$. Based on the sojourn times, T_n describes the time of the n-th jump of the process and $\tilde{\mathbf{X}}_n$ the state of the process on the interval $[T_n, T_{n+1})$. By construction, the continuous-time state process $(\mathbf{X}_t)$ has piecewise-constant càdlàg paths, and the embedded discrete-time process is $(\tilde{\mathbf{X}}_n)$.

The system is controlled by policies, following the notation of Definition 2.3. The general theory on continuous-time Markov Decision Processes states that the optimal policy can be found among the deterministic stationary Markov policies (see the discussion in Chapter 2). Nevertheless, in view of the sections to come, we allow randomization (relaxation) for the control of the system, that is, we operate in the class of relaxed stationary policies in the sense of Definition 2.3 b). Hence, a policy for the system is given by a collection of N stochastic kernels $\pi(d\mathbf{a} \mid \mathbf{x}) = (\pi^k(da \mid \mathbf{x}))_{k=1,\ldots,N}$, where

$$\pi^k : S^N \times \mathcal{B}(A) \to [0, 1], \quad (\mathbf{x}, \mathcal{A}) \mapsto \pi^k(\mathcal{A} \mid \mathbf{x}) \qquad \text{(kernel for agent } k\text{)}.$$

Given the state $\mathbf{x}$ of the system, $\pi^k(\cdot \mid \mathbf{x})$ is the probability distribution on the action space A chosen by the agent k. The action process is thus defined by

$$\pi_t := \sum_{n \in \mathbb{N}_0} \mathbb{1}_{\{T_n < t \le T_{n+1}\}} \pi(\,\cdot \mid \tilde{\mathbf{X}}_n) + \mathbb{1}_{\{t=0\}} \pi(\,\cdot \mid \tilde{\mathbf{X}}_0), \qquad t \in [0, \infty).$$

At this point, it should be emphasized once again that, in contrast to the state process, the action process has piecewise constant càglàd paths, which reflects that a new decision can only be taken after a change of state has occurred.

Notation. To denote arbitrary states that do not describe a specific state of an agent, we use i or j in the following to underline the finiteness of the state space. States that refer to an actual state of an agent are still denoted by x^k.

To prepare the description of the transition mechanism in our model, we define the empirical distribution of the agents' states, i.e.,

$$\mu[\mathbf{x}] := \frac{1}{N}\sum_{k=1}^{N}\delta_{x^k},$$

where δ_{x^k} is the Dirac measure at the point x^k. Since S is finite, δ_{x^k} can be interpreted as a vector with $|S|$ components, such that $\delta_{x^k}(i) = \mathbb{1}_{\{x^k=i\}}$ for a state $i \in S$. Accordingly, we denote the proportion of agents in the state $i \in S$ by $\mu(i)$.

The transition intensities for one agent are given by a signed kernel

$$q : S \times A \times \mathbb{P}(S) \times \mathcal{P}(S) \to \mathbb{R}, \quad (i, a, \mu, \Gamma) \mapsto q(\Gamma \mid i, a, \mu) = \sum_{j\in\Gamma} q(\{j\} \mid i, a, \mu).$$

Note that the transition of an agent depends not only on its own state and action but additionally on the empirical distribution of the agents' states.

We make the following assumptions on q:

(Q1) $q(\{j\}|i, a, \mu) \geq 0$ for all $i, j \in S,\ j \neq i,\ a \in A, \mu \in \mathbb{P}(S)$.

(Q2) $\sum_{j\in S} q(\{j\}|i, a, \mu) = 0$ for all $i \in S,\ a \in A,\ \mu \in \mathbb{P}(S)$.

(Q3) $\sup_{i,a,j,\mu} |q(\{j\}|i, a, \mu)| =: q_{max} < \infty$.

(Q4) $\mu \mapsto q(\{j\}|i, a, \mu)$ is continuous w.r.t. weak convergence for all $i, j \in S,\ a \in A$.

(Q5) $a \mapsto q(\{j\}|i, a, \mu)$ is continuous for all $i, j \in S,\ \mu \in \mathbb{P}(S)$.

Remark 3.1. Note that (Q3) follows from (Q4) and (Q5), since the state space A as well as the space $\mathbb{P}(S)$ are compact (see Lemma A.4) and therefore the continuous functions in (Q4) and (Q5) attain their maximum. Together with the finiteness of the set of states S, the claim follows. However, we list (Q3) to emphasize its significance.

Based on the transition intensities for one agent, the transition intensities of the system are given by

$$q(\{(x^1, \dots, x^{k-1}, j, x^{k+1}, \dots x^N)\}|\mathbf{x}, \mathbf{a}) := q(\{j\}|x^k, a^k, \mu[\mathbf{x}]) \tag{3.1}$$

for all $(\mathbf{x}, \mathbf{a}) \in S^N \times A^N, j \in S, j \neq x^k$, and

$$q(\{\mathbf{x}\}|\mathbf{x}, \mathbf{a}) := \sum_{k=1}^{N} q(\{x^k\}|x^k, a^k, \mu[\mathbf{x}]).$$

All other intensities are zero. Further, we set for a relaxed stationary Markov policy $\pi = (\pi^k(da|\mathbf{x}))_{k=1,\dots,N}$

$$q(\{(x^1, \dots, x^{k-1}, j, x^{k+1}, \dots, x^N)\}|\mathbf{x}, \pi) = \int_A q(\{j\}|x^k, a, \mu[\mathbf{x}])\pi^k(da|\mathbf{x}).$$

The intensity in (3.1) describes the transition of agent k from state $x^k \in S$ to state $j \in S$, while all other agents remain in their current state. This definition is sufficient to describe the transition mechanism of the system, since only one agent can change its state at a time. The following remark gives a more detailed explanation of this observation.

Remark 3.2. Let $\mathbf{x}_n$ be the state of the system after the n-th change of state and $\pi(d\mathbf{a} \mid \cdot)$ the fixed policy of the system. It is a classical result from the theory of continuous-time Markov chains that the sojourn time τ_{n+1} in state $\mathbf{x}_n$ is exponentially distributed with the parameter $-q(\{\mathbf{x}_n\}|\mathbf{x}_n, \pi)$, see, e.g., Theorem 13.3.4 in Brémaud (2020). Now let σ^k be the sojourn time of agent k in state x_n^k for $k \in \{1, \dots, N\}$. Given $\mathbf{x}_n$, due to the independence assumption (I), the random variables $\sigma^1, \dots, \sigma^N$ are independent with distribution

$$\sigma^k \sim \mathrm{Exp}\bigg(- \int_A \underbrace{q(\{x_n^k\} \mid x_n^k, a, \mu[\mathbf{x}_n])}_{\leq 0} \pi^k(da \mid \mathbf{x}_n) \bigg).$$

The sojourn time of the system in state $\mathbf{x}_n$ is then given by $\tau_{n+1} = \min\{\sigma^1, \dots, \sigma^N\}$, which is again exponentially distributed with parameter

$$- \sum_{k=1}^{N} \int_A q(\{x_n^k\} \mid x_n^k, a, \mu[\mathbf{x}_n]) \pi^k(da \mid \mathbf{x}_n) = -q(\{\mathbf{x}_n\}|\mathbf{x}_n, \pi). \tag{3.2}$$

In particular, the probability that two sojourn times coincide (i.e., $\sigma^k = \sigma^l$ for $k, l \in \{1, \dots, N\}$ with $l \neq k$) is zero, which means that no two agents can change their state simultaneously with positive probability. In (3.2) we used the minimum property of the exponential distribution, which states that for two independent random variables $X \sim \mathrm{Exp}(\lambda_1),\ Y \sim \mathrm{Exp}(\lambda_2)$ we have $\min(X, Y) \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$, see, e.g., Example 2.25 in Klenke (2020).

Note that in a certain sense, there is a slight abuse of notation in the definition of the intensities, since we use the letter q for both the individual transition intensity and the system transition intensity. It should always be clear from the context which one is present. Further, one can easily observe that the properties (Q1)-(Q5) of the one-agent-intensities directly imply properties ($\widetilde{Q}1$)-($\widetilde{Q}3$) (stated in Section 3.1) for the intensities of the system.

The probability measure of the N-agent process is now given by the following transition kernels

$$\mathbb{P}^\pi(\tau_n \leq t, \tilde{\mathbf{X}}_n \in \Gamma | \tilde{\mathbf{X}}_{n-1}) = \int_0^t q(\Gamma \setminus \{\tilde{\mathbf{X}}_{n-1}\} | \tilde{\mathbf{X}}_{n-1}, \pi) e^{s \cdot q(\{\tilde{\mathbf{X}}_{n-1}\} | \tilde{\mathbf{X}}_{n-1}, \pi)} ds$$

for all $t \geq 0$ and $\Gamma \in \mathcal{P}(S^N)$. In particular, the sojourn times τ_n are exponentially distributed with parameter $-q(\{\tilde{\mathbf{X}}_{n-1}\}|\tilde{\mathbf{X}}_{n-1}, \pi)$. Note that by using this construction, the probability measure depends on the chosen policy, which justifies the name strategic measure.

Returning to the model's control mechanism, keep in mind that the policy of an agent $\pi^k(da \mid \mathbf{x})$ is allowed to depend on the state of the whole system, i.e., we assume that each agent has information about the position of all other agents. Therefore, we can interpret our model as a centralized control problem, where all information is collected and shared by a central controller.

The goal of the central controller is to maximize the expected social reward of the system. In order to implement this, we introduce the (stationary) reward function for one agent as

$$r : S \times A \times \mathbb{P}(S) \to \mathbb{R}, \qquad (i, a, \mu) \mapsto r(i, a, \mu),$$

which depends not only on the state and action of the single agent, but also on the empirical distribution of the system. We make the following assumptions on the reward function:

(R1) For all $(i, a) \in S \times A$ the function $\mu \mapsto \mu(i)r(i, a, \mu)$ is continuous w.r.t. weak convergence.

(R2) For all $i \in S$ and $\mu \in \mathbb{P}(S)$, the function $a \mapsto r(i, a, \mu)$ is continuous.

Since the action space A is assumed to be compact, (R1) and (R2) imply (using the same arguments as in Remark 3.1) that the following expression is bounded:

$$\sup_{(i,a)\in S\times A,\ \mu\in\mathbb{P}(S)} |\mu(i)r(i, a, \mu)| =: (\mu r)_{max} < \infty \tag{3.3}$$

Then we define the (social) reward of the system as the average of the agents' rewards, i.e.,

$$r(\mathbf{x}, \mathbf{a}) := \frac{1}{N} \sum_{k=1}^{N} r(x^k, a^k, \mu[\mathbf{x}]). \tag{3.4}$$

The continuity assumptions (R1) and (R2), together with the finiteness of the state space, directly imply the measurability of the social reward function $(\mathbf{x}, \mathbf{a}) \mapsto r(\mathbf{x}, \mathbf{a})$.

For a relaxed stationary Markov policy $\pi = (\pi^k(da \mid \mathbf{x}))_{k=1,\dots,N}$ we set

$$r(\mathbf{x}, \pi) := \frac{1}{N} \sum_{k=1}^{N} \int_A r(x^k, a, \mu[\mathbf{x}])\pi^k(da \mid \mathbf{x}).$$

The aim of the central controller is to find the *social optimum*, i.e., to maximize the joint expected discounted reward of the system over an infinite time horizon. For a policy $\pi \in \Pi^s$, a discount rate $\beta > 0$, and an initial configuration $\mathbf{x} \in S^N$, define the value function

$$\begin{aligned} V_\pi(\mathbf{x}) &= \mathbb{E}^\pi_{\mathbf{x}}\Big[\int_0^\infty e^{-\beta t} r(\mathbf{X}_t, \pi_t) dt \Big] \\ V(\mathbf{x}) &= \sup_\pi V_\pi(\mathbf{x}). \end{aligned} \tag{3.5}$$

We do not discuss solution procedures for this optimization problem at this point, since we rephrase the problem in the following section and present asymptotically optimal solution methods in Section 3.4.

3.2. THE MEASURE-VALUED CONTINUOUS-TIME MARKOV DECISION PROCESS

As N increases, so does the state space S^N, which increases the complexity of the model and makes it computationally more expensive to solve. Therefore, we seek some simplifications. A reasonable approach is to exploit the symmetry of the system by capturing not the state of every single agent, but the relative or empirical distribution of the agents over the $|S|$ states.

Thus, let $\mu_t^N := \mu[\mathbf{X}_t]$ and define as the new state space the set of all empirical measures of N atoms in S, i.e.,

$$\mathbb{P}_N(S) := \{\mu \in \mathbb{P}(S) \mid \mu = \mu[\mathbf{x}], \text{ for some } \mathbf{x} \in S^N\}.$$

Then the new state process μ_t^N is simply given by

$$\mu_t^N = \sum_{n \in \mathbb{N}_0} \mathbb{1}_{\{T_n \leq t < T_{n+1}\}} \mu[\tilde{\mathbf{X}}_n], \qquad t \in [0, \infty). \tag{3.6}$$

As action space, we take the $|S|$-fold Cartesian product $\mathbb{P}(A)^{|S|}$ of $\mathbb{P}(A)$, the set of all probability distributions on the action space. Hence, an action is given by $|S|$ probability measures $\alpha(d\mathbf{a}) = (\alpha^i(da))_{i \in S}$, where the i-th component indicates the distribution of the agents' actions in state $i \in S$.

For the policies we again restrict to relaxed stationary Markov policies given by a collection of $|S|$ stochastic kernels $\hat{\pi}(d\mathbf{a}|\mu) = (\hat{\pi}^i(da|\mu))_{i \in S}$, where

$$\hat{\pi}^i : \mathbb{P}_N(S) \times \mathcal{B}(A) \to [0,1], \quad (\mu, \mathcal{A}) \mapsto \hat{\pi}^i(\mathcal{A} \mid \mu) \qquad \text{(kernel for state } i\text{)}.$$

In what follows, we denote $\tilde{\mu}_n^N := \mu[\tilde{\mathbf{X}}_n]$. Then we can express the action process by setting

$$\hat{\pi}_t := \sum_{n \in \mathbb{N}_0} \mathbb{1}_{\{T_n < t \leq T_{n+1}\}} \hat{\pi}(\,\cdot \mid \tilde{\mu}_n^N) + \mathbb{1}_{\{t=0\}} \hat{\pi}(\,\cdot \mid \tilde{\mu}_0^N), \qquad t \in [0, \infty). \tag{3.7}$$

The transition intensities of the process $(\mu_t^N)_{t \geq 0}$ are given by

$$q(\{\mu^{i \to j}\}|\mu, \alpha) = N\mu(i) \int_A q(\{j\}|i, a, \mu)\alpha^i(da), \quad \mu \in \mathbb{P}_N(S), \alpha \in \mathbb{P}(A)^{|S|}, \tag{3.8}$$

with $\mu^{i \to j} := \mu - \frac{1}{N}\delta_i + \frac{1}{N}\delta_j$ for all $i, j \in S, i \neq j$, if $\mu(i) > 0$.

The natural interpretation of (3.8) is as follows:

- $q(\{\mu^{i\to j}\}|\mu,\alpha)$: Intensity for the transition of one arbitrary agent in state $i \in S$ to state $j \in S$, while all other agents remain in their current state.
- $N\mu(i)$: Number of agents in state i.
- $\int_A q(\{j\}|i,a,\mu)\alpha^i(da)$: Intensity for the transition of an individual agent in state $i \in S$ to state $j \in S$ while performing the action $\alpha^i \in \mathbb{P}(A)$.

Further, we set for all $\mu \in \mathbb{P}_N(S)$ and $\alpha \in \mathbb{P}(A)^{|S|}$

$$q(\{\mu\}|\mu,\alpha) := -\sum_{i,\mu(i)>0}\sum_{j\neq i} q(\{\mu^{i\to j}\}|\mu,\alpha).$$

All other intensities are zero, since again only one agent can change its state at a time.

The probability distribution of the measure-valued process under a fixed policy $\hat{\pi}$ is now given by the following transition kernels

$$\mathbb{P}^{\hat{\pi}}(\tau_n \le t, \tilde{\mu}_n^N \in B|\tilde{\mu}_{n-1}^N) = \int_0^t q(B|\tilde{\mu}_{n-1}^N,\hat{\pi})e^{s\cdot q(\{\tilde{\mu}_{n-1}^N\}|\tilde{\mu}_{n-1}^N,\hat{\pi})}ds$$

for all $t \ge 0$ and $B \subset \mathbb{P}_N(S)$ measurable, where the random variables (τ_n) are the same as before.

The reward function of the system is derived from the reward for one agent:

$$r(\mu,\alpha) := \sum_{i\in S}\int_A \mu(i)r(i,a,\mu)\alpha^i(da). \tag{3.9}$$

In view of (3.3), $r(\mu,\alpha)$ is bounded.

Lemma 3.3. *Assumptions (R1) and (R2) imply that for every $\alpha \in \mathbb{P}(A)^{|S|}$, the function $\mu \mapsto r(\mu,\alpha)$ is continuous w.r.t. weak convergence.*

Proof. Since S is finite, the weak convergence $\mu_n \Rightarrow \mu$ of a sequence of probability measures $(\mu_n)_{n\in\mathbb{N}} \subset \mathbb{P}(S)$ to a limit measure $\mu \in \mathbb{P}(S)$ corresponds to the componentwise convergence $\mu_n(i) \to \mu(i)$ for every $i \in S$. We obtain

$$\begin{aligned}\lim_{n\to\infty} r(\mu_n,\alpha) &= \lim_{n\to\infty}\sum_{i\in S}\int_A \mu_n(i)r(i,a,\mu_n)\alpha^i(da)\\ &= \sum_{i\in S}\int_A \mu(i)r(i,a,\mu)\alpha^i(da) = r(\mu,\alpha),\end{aligned}$$

where we use the fact that the integrand is bounded (see (3.3)) and dominated convergence to exchange the limit and the integral. Together with (R1), the claim follows. □

The aim in this model is again to maximize the joint expected discounted reward of the system over an infinite time horizon. For a policy $\hat{\pi}$, a discount rate $\beta > 0$, and an initial configuration $\mu \in \mathbb{P}_N(S)$, define the value function

$$\begin{aligned} V^N_{\hat{\pi}}(\mu) &= \mathbb{E}^{\hat{\pi}}_{\mu}\Big[\int_0^\infty e^{-\beta t} r(\mu^N_t, \hat{\pi}_t) dt\Big], \\ V^N(\mu) &= \sup_{\hat{\pi}} V^N_{\hat{\pi}}(\mu). \end{aligned} \tag{3.10}$$

We can now show that both formulations (3.5) and (3.10) are equivalent in the sense that the optimal values are the same.

Theorem 3.4. *We have $V(\mathbf{x}) = V^N(\mu)$ for $\mu = \mu[\mathbf{x}]$ for all $\mathbf{x} \in S^N$.*

Proof. First of all, observe that the reward function r in (3.4) in the multi-agent problem is symmetric, that is, $r(\mathbf{x}, \mathbf{a}) = r(s(\mathbf{x}), s(\mathbf{a}))$ for any permutation $s(\cdot)$ of the vectors.
Now for a stationary policy π for the multi-agent problem, define for all states $i \in S$ with $\mu(i) > 0$:

$$\hat{\pi}^i(da|\mu) := \frac{1}{N\mu(i)} \sum_{k=1}^N \pi^k(da|\mathbf{x}) \mathbb{1}_{\{x^k = i\}}$$

where $\mu = \mu[\mathbf{x}]$. Practically, we consider all agents in state i and take a convex combination of their action distributions as the action distribution in state i. By this construction, we transfer a stationary policy π in the multi-agent setting to a stationary policy in the measure-valued setting.

Choosing $\hat{\pi}$ in the measure-valued MDP yields the reward (where again $\mu = \mu[\mathbf{x}]$)

$$\begin{aligned} r(\mu, \hat{\pi}) &= \sum_{i \in S} \int_A r(i, a, \mu) \mu(i) \left(\frac{1}{N\mu(i)} \sum_{k=1}^N \pi^k(da|\mathbf{x}) \mathbb{1}_{\{x^k = i\}} \right) \\ &= \frac{1}{N} \sum_{k=1}^N \sum_{i \in S} \mathbb{1}_{\{x^k = i\}} \int_A r(i, a, \mu) \pi^k(da|\mathbf{x}) = r(\mathbf{x}, \pi). \end{aligned}$$

Thus, the reward in both formulations is the same. Finally, the transition intensity in the multi-agent model of one agent changing its state from i to j is given by (again $\mu = \mu[\mathbf{x}]$)

$$\begin{aligned} &\sum_{k=1}^N \mathbb{1}_{\{x^k = i\}} \int_A q(\{j\}|i, a, \mu) \pi^k(da|\mathbf{x}) \\ &= N\mu(i) \int_A q(\{j\}|i, a, \mu) \frac{1}{N\mu(i)} \sum_{k=1}^N \pi^k(da|\mathbf{x}) \mathbb{1}_{\{x^k = i\}} \\ &= N\mu(i) \int_A q(\{j\}|i, a, \mu) \hat{\pi}^i(da|\mu) = q(\{\mu^{i \to j}\}|\mu, \hat{\pi}). \end{aligned}$$

Thus, the empirical measure process of the multi-agent problem is statistically equal to the measure-valued MDP process and produces the same expected reward under measure-dependent policies. Now let π^* be an optimal policy for the multi-agent problem and $\hat{\pi}^*$

the corresponding measure-valued policy. We obtain

$$V(\mathbf{x}) = V_{\pi^*}(\mathbf{x}) = V^N_{\hat{\pi}^*}(\mu) \leq V^N(\mu). \tag{3.11}$$

Conversely, let $\hat{\psi}$ be an arbitrary policy for the measure-valued MDP. We construct a policy for the multi-agent problem by assigning each agent $k = 1, \dots, N$ the decision rule

$$\psi^k(da|\mathbf{x}) = \sum_{i \in S} \hat{\psi}^i(da|\mu) \mathbb{1}_{\{x^k = i\}}. \tag{3.12}$$

Again, both policies deliver the same reward for arbitrary $\mu = \mu[\mathbf{x}]$:

$$\begin{aligned} r(\mathbf{x}, \psi) &= \frac{1}{N} \sum_{k=1}^{N} \int_A r(x^k, a, \mu) \psi^k(da|\mathbf{x}) \\ &= \frac{1}{N} \sum_{k=1}^{N} \sum_{i \in S} \mathbb{1}_{\{x^k = i\}} \int_A r(x^k, a, \mu) \psi^k(da|\mathbf{x}) = r(\mu, \hat{\psi}). \end{aligned}$$

The transition intensity for the change of an agent from state i to j in both models using policy ψ resp. $\hat{\psi}$ is given by

$$\begin{aligned} &\sum_{k=1}^{N} \mathbb{1}_{\{x^k = i\}} \int_A q(\{j\}|i, a, \mu) \psi^k(da|\mathbf{x}) \\ &= N\mu(i) \int_A q(\{j\}|i, a, \mu) \hat{\psi}^i(da|\mu) = q(\{\mu^{i \to j}\}|\mu, \hat{\psi}). \end{aligned}$$

To conclude the proof, let $\hat{\psi}^*$ be an optimal policy for the measure-valued problem and ψ^* the corresponding multi-agent policy. Again, we obtain

$$V^N(\mu) = V^N_{\hat{\psi}^*}(\mu) = V_{\psi^*}(\mathbf{x}) \leq V(\mathbf{x}).$$

Together with the converse inequality (3.11), the result follows. □

Remark 3.5. It is possible to extend the previous result to a situation where reward and transition intensity additionally depend on the empirical distribution of actions, see, e.g., Motte and Pham (2022). However, due to the definition of the Young topology, which we use later, it is not possible to transfer the convergence results to this setting.

The problem introduced previously is a classical continuous-time Markov decision process and can be solved with the theory established in Chapter 2.

Lemma 3.6. *The measure-valued model satisfies Conditions (A), (B), and (C).*

Proof. Since the state space $\mathbb{P}_N(S)$ is finite, Conditions (A) and (B) follow directly from Lemma 2.6 with $w(\mu) \equiv w'(\mu) \equiv 1$.

In order to verify Condition (C), consider the functions

$$q : \mathbb{P}_N(S) \times \mathbb{P}(A)^{|S|} \to \mathbb{R}, \qquad (\mu, \alpha) \mapsto q(\{\nu\}|\mu, \alpha), \quad \nu \in \mathbb{P}_N(S),$$
$$r : \mathbb{P}_N(S) \times \mathbb{P}(A)^{|S|} \to \mathbb{R}, \qquad (\mu, \alpha) \mapsto r(\mu, \alpha).$$

Both functions are continuous in μ by construction, since the domain is finite. Regarding the continuity in α, recall that

$$q(\{\mu^{i\to j}\}|\mu, \alpha) = N\mu(i) \int_A q(\{j\}|i, a, \mu)\alpha^i(da),$$
$$r(\mu, \alpha) = \sum_{i\in S} \int_A \mu(i) r(i, a, \mu)\alpha^i(da).$$

In this representation, it is clear that the functions are continuous in α w.r.t. weak convergence. Note that the integrands are continuous in a and bounded due to (Q3) and (Q5) resp. (R2) and (3.3). Note that the continuity of $q(\{\mu\}|\mu, \alpha)$ follows from the continuity of $q(\{\mu^{i\to j}\}|\mu, \alpha)$. For other $\nu \in \mathbb{P}_N(S)$ the intensity is constantly zero in α.

Now suppose that u is a bounded continuous function on $\mathbb{P}_N(S)$. Then the function

$$(\mu, \alpha) \mapsto \sum_{\nu\in\mathbb{P}_N(S)\setminus\{\mu\}} u(\nu) \cdot 1 \cdot q(\{\nu\}|\mu, \alpha)$$

is continuous as a composition of continuous functions.

The compactness of $\mathbb{P}(A)^{|S|}$ follows directly from the compactness of A together with Lemma A.4 and Theorem A.3. Thus, Condition (C) is satisfied, which completes the proof. □

Theorem 3.7. *There exists a unique continuous function $v : \mathbb{P}_N(S) \to \mathbb{R}$ satisfying the optimality equation*

$$\beta v(\mu) = \sup_{\alpha\in\mathbb{P}(A)^{|S|}} \left\{ r(\mu, \alpha) + \sum_{\nu\in\mathbb{P}_N(S)} v(\nu) q(\{\nu\}|\mu, \alpha) \right\} \tag{3.13}$$

for all $\mu \in \mathbb{P}_N(S)$. Moreover, there exists a measurable mapping $\hat{\pi} : \mathbb{P}_N(S) \to \mathbb{P}(A)^{|S|}$ which provides the supremum in (3.13)*. The mapping $\hat{\pi}$ defines a deterministic stationary Markov policy $\hat{\pi}(\cdot|\mu)$ which is optimal for the discounted measure-valued CTMDP:*

$$V^N(\mu) = V^N_{\hat{\pi}}(\mu) = v(\mu), \qquad \forall\ \mu \in \mathbb{P}_N(S).$$

The result follows directly from the solution theory in Chapter 2, in particular from Theorem 2.8. Note that the optimal Markov policy $\hat{\pi}$ in the previous theorem is deterministic as a measurable selector into the action space $\mathbb{P}(A)^{|S|}$; the components $\hat{\pi}^i(\cdot|\mu)$ are indeed probability distributions on A.

Of course, an optimal policy in the measure-valued setting can be directly implemented in the original problem; see the construction (3.12) in the proof of Theorem 3.4. The advantage of the measure-valued formulation is the reduction of the cardinality of the state space. Suppose, e.g., that $S = \{0, 1\}$, that is, all agents are either in state 0 or state 1. Then $|S^N| = 2^N$ in the multi-agent model, whereas $|\mathbb{P}_N(S)| = N + 1$ in the measure-valued case. However, even in this simplified setting, the computation may be inefficient if N is large. For a finite number of actions, the computational effort in the measure-valued model is of the order $\mathcal{O}((N+1)^2 \cdot |A|^2) \sim \mathcal{O}(N^2)$, implying a quadratic dependence on the number of agents. On the other hand, in the original model we have $\mathcal{O}(|S^N|^2 \cdot |A|^2) \sim \mathcal{O}(2^{2N})$, i.e., there is an exponential growth of the effort with the number of agents.

Nevertheless, even the measure-valued model may be intractable to solve by conventional methods for a large number of agents. Naturally, the question arises of how the model behaves when the number of agents tends to infinity.

3.3. Convergence of the state-action process

In this section, we discuss the behavior of the system when the number of individuals tends to infinity. Roughly, the approach is to use the generator of the underlying Markov chain of the system to construct a martingale, for which weak convergence to the zero process is shown. A similar procedure has been carried out by Bäuerle (2000) in the context of fluid models.

We start by specifying the underlying topological spaces in which the state and action processes are embedded, in order to investigate the asymptotic behavior. In what follows, the state process $(\mu_t^N)_{t\geq 0}$ is considered a stochastic element of $D_{\mathbb{P}(S)}[0, \infty)$, the space of càdlàg paths with values in $\mathbb{P}(S)$ equipped with the Skorokhod J_1-topology and the metric d_{J_1}. On $\mathbb{P}(S)$, we choose the total variation metric. For an overview, see Section B.3.

Furthermore, for an action process $(\hat{\pi}_t)_{t\geq 0}$, where $\hat{\pi}_t = (\hat{\pi}_t^i)_{i\in S}$, see (3.7), we consider $(\hat{\pi}_t^i)_{t\geq 0}$ as a stochastic element of $\mathcal{R} := \{\rho : [0, \infty) \to \mathbb{P}(A) \mid \rho \text{ measurable}\}$, the space of measurable functions that assign for each point in time $t \geq 0$ a probability distribution on the action space. The space $\mathcal{R}$ is endowed with the Young topology (cf. Davis (1993)). It is possible to show that $\mathcal{R}$ is compact and metrizable. For a derivation of the notion of convergence used on $\mathcal{R}$, see Section B.2.

In the following, we denote a relaxed stationary Markov policy in the measure-valued model by $\hat{\pi}^N = (\hat{\pi}^{N,i})_{i\in S}$, where the superscript N emphasizes that the model with a fixed number N of agents is considered. We begin with the observation that the state process is well-behaved in the following sense:

Lemma 3.8. *For an arbitrary relaxed stationary Markov policy $\hat{\pi}^N$, the empirical state process $(\mu_t^N)_{t\geq 0}$ is non-explosive, i.e., the number of jumps in any finite time interval is almost surely finite.*

The statement follows since the intensities are bounded and the state space is finite. The result can be found in Piunovskiy and Zhang (2020), Theorem 2.2.4.

As a next step, we recapitulate that for $N \in \mathbb{N}$ and a fixed relaxed stationary Markov policy $\hat{\pi}^N$, the empirical state process $(\mu_t^N)_{t\geq 0}$ describes a time-continuous (homogeneous) Markov chain on a finite state space $\mathbb{P}_N(S)$. The strong generator of this Markov chain is given by

$$\mathcal{A}f(\mu) = \sum_{\nu \in \mathbb{P}_N(S)} (f(\nu) - f(\mu)) q(\{\nu\}|\mu, \hat{\pi}^N),$$

where the arbitrary function $f : \mathbb{P}_N(S) \to \mathbb{R}$ is naturally measurable and bounded due to the finiteness of the domain $\mathbb{P}_N(S)$. For a reference for this representation of the generator, see, e.g., Ethier and Kurtz (1986) Chapter 4, Section 2, Eq. (2.1). Now consider the function f defined as the projection on the j-th component of $\mu \in \mathbb{P}(S)$, i.e.,

$$f : \mathbb{P}_N(S) \to \mathbb{R}, \qquad f(\mu) = \mu(j), \qquad j \in S,$$

which describes the proportion of agents in state j. Define for arbitrary $j \in S$, the one-dimensional process

$$\begin{aligned} M_t^N(j) &:= \mu_t^N(j) - \mu_0^N(j) - \int_0^t \mathcal{A}f(\mu_s) ds \\ &= \mu_t^N(j) - \mu_0^N(j) - \int_0^t \sum_{\nu \in \mathbb{P}_N(S)} (\nu(j) - \mu_s^N(j)) q(\{\nu\}|\mu_s^N, \hat{\pi}_s^N) ds. \end{aligned}$$

Then, for each $j \in S$, the process $M^N(j) := (M_t^N(j))_{t\geq 0}$ is a martingale w.r.t. the filtration $\mathcal{F}_t^N = \sigma(\mu_s^N, s \leq t)$. This follows from the Dynkin formula, see, e.g., Davis (1993), Proposition 14.13. Next, we derive a refined representation of $M^N(j)$. Note that the difference $\nu(j) - \mu_s^N(j)$ can either be $-1/N$ if an individual changes from state j to a state $k \neq j$ or it could be $1/N$ if an individual changes from state $i \neq j$ to state j. In all other cases the difference is 0. Since by (Q2)

$$\sum_{k\neq j} \int_A q(\{k\}|j, a, \mu_s^N) \hat{\pi}_s^{N,j}(da) = -\int_A q(\{j\}|j, a, \mu_s^N) \hat{\pi}_s^{N,j}(da) \tag{3.14}$$

we obtain by inserting the intensity (3.8) and by using (3.14)

$$\begin{aligned} M_t^N(j) =& \mu_t^N(j) - \mu_0^N(j) - \int_0^t \sum_{\nu \in \mathbb{P}_N(S)} (\nu(j) - \mu_s^N(j)) q(\{\nu\}|\mu_s^N, \hat{\pi}_s^N) ds \\ \overset{(3.8)}{=} & \mu_t^N(j) - \mu_0^N(j) - \int_0^t \sum_{k\neq j} \left(-\frac{1}{N}\right) N \mu_s^N(j) \int_A q(\{k\}|j, a, \mu_s^N) \hat{\pi}_s^{N,j}(da) ds \\ & - \int_0^t \sum_{i\neq j} \left(\frac{1}{N}\right) N \mu_s^N(i) \int_A q(\{j\}|i, a, \mu_s^N) \hat{\pi}_s^{N,i}(da) ds \\ \overset{(3.14)}{=} & \mu_t^N(j) - \mu_0^N(j) - \int_0^t \sum_{i\in S} \mu_s^N(i) \int_A q(\{j\}|i, a, \mu_s^N) \hat{\pi}_s^{N,i}(da) ds. \end{aligned} \tag{3.15}$$

With this representation, we prove that the sequence of stochastic processes $(M^N(j))_{N\in\mathbb{N}}$ converges weakly (denoted by $\Rightarrow$) to the zero process in the Skorokhod J_1-topology.

Lemma 3.9. *We have for all $j \in S$ that*

$$(M_t^N(j))_{t\geq 0} \Rightarrow 0, \quad N \to \infty,$$

where $0 \in D_{\mathbb{R}}[0,\infty)$ represents the zero process.

Proof. The idea of the proof is to apply Theorem B.20. In a first step, we have to verify that the finite dimensional distributions of $(M_t^N(j))_{t\geq 0}$ converge weakly to zero, i.e., for any finite collection of points in time $\{t_1,\dots,t_n\} \subset [0,\infty)$ it must hold that

$$(M_{t_1}^N(j),\dots,M_{t_n}^N(j)) \Rightarrow (0,\dots,0), \tag{3.16}$$

where $\Rightarrow$ is the usual weak convergence of random vectors in $\mathbb{R}^n$.

We start by showing that $M_t^N(j)$ is bounded for any fixed $t \in [0,\infty)$:

$$\begin{aligned}
|M_t^N(j)| &= \left|\mu_t^N(j) - \mu_0^N(j) - \int_0^t \sum_{i\in S} \mu_s^N(i) \int_A q(\{j\}|i,a,\mu_s^N)\hat{\pi}_s^{N,i}(da)ds\right| \\
&\leq |\mu_t^N(j) - \mu_0^N(j)| + \int_0^t \sum_{i\in S} \mu_s^N(i) \int_A \underbrace{|q(\{j\}|i,a,\mu_s^N)|}_{\leq q_{max}} \hat{\pi}_s^{N,i}(da)ds \\
&\leq 1 + q_{max}\cdot t < \infty
\end{aligned} \tag{3.17}$$

Therefore $(M_t^N(j))_{t\geq 0}$ are square-integrable martingales in the sense that $\mathbb{E}[(M_t^N(j))^2] < \infty$ for every point in time $t \in [0,\infty)$.

Next, we state some further properties of $(M_t^N(j))_{t\geq 0}$. Notice that, regarding the evolution in time, the integrand in (3.15), and in particular the action process $(\hat{\pi}_t^N)_{t\geq 0}$, depends only on the behavior of the state process $(\mu_t^N)_{t\geq 0}$, since we consider a stationary policy. Consequently, the process $(M_t^N(j))_{t\geq 0}$ is $(\mathcal{F}_t^N)_{t\geq 0}$-adapted and the integrand is constant in time as long as no jump occurs in the state process. Therefore, the integral is a continuous, piecewise linear process in time with a maximum slope of q_{max}. In order to obtain $(M_t^N(j))_{t\geq 0}$, the process $(\mu_t^N(j))_{t\geq 0}$ is added to the integral and the constant $\mu_0^N(j)$ is subtracted. Thus, the process $(M_t^N(j))_{t\geq 0}$ is piecewise linear as well, exhibiting jumps of height $1/N$ at the same points in time as $(\mu_t^N(j))_{t\geq 0}$. Lemma A.13 now states that the paths of $(M_t^N(j))_{t\geq 0}$ have finite variation on compact intervals. Note that Lemma 3.8 ensures that the number of jumps is finite in any interval $[a,b] \subset [0,\infty)$. Theorem A.14 then implies that $(M_t^N(j))_{t\geq 0}$ is a quadratic pure jump process with quadratic variation

$$\begin{aligned}
[M^N(j), M^N(j)]_t &= \sum_{0<s\leq t} (M_s^N(j) - M_{s-}^N(j))^2 \\
&= \frac{1}{N^2} \cdot \left|\{s \in (0,t] \mid \mu_s^N(j) \neq \mu_{s-}^N(j)\}\right|,
\end{aligned}$$

which in fact is the number of jumps of $(\mu_t^N(j))_{t\geq 0}$ in $[0,t]$ multiplied by the squared magnitude of a jump $1/N^2$.

From Remark 3.2 it follows that the expected sojourn time of the *system* in an arbitrary system state $\mu \in \mathbb{P}_N(S)$ is at least $\frac{1}{N\cdot q_{max}}$, which in turn implies that the expected number of jumps of the system state process $(\mu_t^N)_{t\geq 0}$ in a time interval $[0,t]$ is at most $N \cdot q_{max} \cdot t$.

Putting things together, we obtain

$$\begin{aligned}
\mathbb{E}[(M_t^N(j))^2] = \mathbb{E}[[M^N(j), M^N(j)]_t] &= \frac{1}{N^2} \cdot \mathbb{E}\left[\left|\{s \in (0,t] \mid \mu_s^N(j) \neq \mu_{s-}^N(j)\}\right|\right] \\
&\leq \frac{1}{N^2} \cdot \mathbb{E}\left[\left|\{s \in (0,t] \mid \mu_s^N \neq \mu_{s-}^N\}\right|\right] \\
&\leq \frac{1}{N^2} \cdot N \cdot q_{max} \cdot t \qquad (3.18) \\
&= \frac{1}{N} \cdot q_{max} \cdot t \quad \overset{N\to\infty}{\longrightarrow} \quad 0. \qquad (3.19)
\end{aligned}$$

Note that the first equality follows from part b) of Theorem A.14.

The convergence (3.19) now implies that for an arbitrary but fixed point in time $t \in [0,\infty)$, the sequence $(M_t^N(j))_{N\in\mathbb{N}}$ satisfies

$$M_t^N(j) \quad \overset{L^2}{\longrightarrow} \quad 0. \qquad (3.20)$$

Now, let $\{t_1, \ldots, t_k\} \subset [0,\infty)$ be an arbitrary finite set of points in time. From (3.20) it follows that the marginal distributions $M_{t_k}^N(j)$ of the random vector $(M_{t_1}^N(j), \ldots, M_{t_n}^N(j))$ converge in L^2 to zero, and in particular they converge in probability. The convergence in probability of the marginal distributions is equivalent to the convergence in probability of the corresponding random vector, see Proposition B.4, i.e., we have

$$(M_{t_1}^N(j), \ldots, M_{t_n}^N(j)) \quad \overset{\mathbb{P}}{\longrightarrow} \quad (0, \ldots, 0), \qquad (3.21)$$

which implies the desired weak convergence (3.16).

To apply Theorem B.20, it remains to check the relative compactness, which we show using Theorem B.19. First note that $(M^N(j))_{N\in\mathbb{N}}$ is a sequence of stochastic processes with sample paths in $D_{\mathbb{R}}[0,\infty)$, where $\mathbb{R}$ is separable and $(\mathbb{R}, |\cdot|)$ is complete. Due to boundedness (3.17), choose for every $\varepsilon > 0$ and rational $t \geq 0$ the compact set $\Gamma_{t,\varepsilon} = [-1 - q_{max} \cdot t, 1 + q_{max} \cdot t]$. Then we obtain

$$\mathbb{P}(M_t^N(j) \in \Gamma_{t,\varepsilon}) = 1.$$

In what follows, let σ be an arbitrary $(\mathcal{F}_t^N)$-stopping time with $\sigma \leq T$ almost surely.

For every $T > 0$ it holds that

$$\begin{aligned}
&\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\sigma}\mathbb{E}\left[\min\left\{1,\left|M_\sigma^N(j)-M_{\sigma+\delta}^N(j)\right|\right\}\right]\\
\leq&\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\sigma}\mathbb{E}\left[\left|M_\sigma^N(j)-M_{\sigma+\delta}^N(j)\right|\right]\\
\leq&\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\sigma}\left(\mathbb{E}\left[\left|\{s\in(\sigma,\sigma+\delta]\mid \mu_s^N(j)\neq\mu_{s-}^N(j)\}\right|\right]\cdot\frac{1}{N}+\delta\cdot q_{max}\right)\\
\leq&\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\sigma}\left(\mathbb{E}\left[\left|\{s\in(\sigma,\sigma+\delta]\mid \mu_s^N\neq\mu_{s-}^N\}\right|\right]\cdot\frac{1}{N}+\delta\cdot q_{max}\right)\\
\leq&\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\sigma}\left(N\cdot q_{max}\cdot\delta\cdot\frac{1}{N}+\delta\cdot q_{max}\right)=0.
\end{aligned}$$

For the second inequality, we use the fact that the process $(M_t^N(j))_{t\geq 0}$ is piecewise linear with a maximum slope of q_{max} and has jumps of height $1/N$ at the same points in time as $(\mu_t^N(j))_{t\geq 0}$, see the discussion above. The last inequality follows as in (3.18) since the expected number of jumps of the system in the interval $(\sigma,\sigma+\delta]$ is just $N\cdot q_{max}\cdot\delta$.

Theorem B.19 now suggests that the sequence $(M^N(j))_{N\in\mathbb{N}}$ is relatively compact. Together with (3.21), the conditions of Theorem B.20 are fulfilled, and the sequence of processes $(M_t^N(j))_{t\geq 0}$ converges weakly on $[0,\infty)$ towards the zero-process in the sense of the Skorokhod d_{J_1}-metric for $N\to\infty$. □

Furthermore, Lemma 3.9 directly implies the stochastic convergence of the sequence of martingale processes $(M_t^N)_{t\geq 0}\subset D_{\mathbb{R}^{|S|}}[0,\infty)$.

Corollary 3.10. *We have that*

$$(M_t^N)_{t\geq 0}\xrightarrow{\mathbb{P}}(0,\dots,0),\quad N\to\infty,$$

where $(0,\dots,0)\in D_{\mathbb{R}^{|S|}}[0,\infty)$ *is the vector of zero-processes.*

Proof. Part c) of Proposition B.4 implies for all $j\in S$ the stochastic convergence

$$(M_t^N(j))_{t\geq 0}\xrightarrow{\mathbb{P}}0,\quad N\to\infty.$$

Part b) of Proposition B.4 then implies the stochastic convergence of the entire martingale process. □

The next Theorem marks the main convergence result in this chapter. It states that the sequence of state-action processes for an increasing number of agents is relatively compact, hence there exists a weakly converging subsequence. It turns out that the limit state process has almost surely continuous paths, which is intuitively reasonable since, as discussed earlier, the height of the jumps of the state process is $1/N$ and therefore converges to zero as the number of agents tends to infinity. However, the more interesting property of the limit state process is that it is deterministic and characterized by an ordinary differential equation. In the theorem we abbreviate $\mu^N := (\mu_t^N)_{t\geq 0}$ resp. $\hat{\pi}^N := (\hat{\pi}_t^N)_{t\geq 0}$.

Theorem 3.11. *A sequence of arbitrary state-action processes $(\mu^N, \hat{\pi}^N)_{N\in\mathbb{N}}$ is relatively compact. Thus, there exists a subsequence (N_k) which converges weakly*

$$(\mu^{N_k}, \hat{\pi}^{N_k}) \Rightarrow (\mu^*, \hat{\pi}^*), \text{ as } k \to \infty.$$

Moreover, the limit $(\mu^*, \hat{\pi}^*)$ satisfies

a) (μ_t^*) has almost surely continuous paths,

b) and for each state $j \in S$ we have

$$\mu_t^*(j) = \mu_0^*(j) + \int_0^t \sum_{i\in S} \mu_s^*(i) \int_A q(\{j\}|i, a, \mu_s^*)\hat{\pi}_s^{*,i}(da)ds. \tag{3.22}$$

Proof. We start by showing the relative compactness of a sequence $(\mu^N)_{N\in\mathbb{N}}$, using Theorem B.19. The sequence $(\mu^N)_{N\in\mathbb{N}}$ has paths in $D_{\mathbb{P}(S)}[0,\infty)$, where $\mathbb{P}(S)$ is complete and separable with respect to the total variation distance; see Lemma A.10.
In what follows, let again σ be an arbitrary $(\mathcal{F}_t^N)$-stopping time with $\sigma \le T$ almost surely. For every $\varepsilon > 0$ and rational $t \ge 0$ choose the compact set $\Gamma_{t,\varepsilon} \equiv \mathbb{P}(S)$. Then we obtain by construction of the model

$$\mathbb{P}(\mu_t^N \in \Gamma_{t,\varepsilon}) = 1.$$

Moreover, for every $T > 0$ it holds that

$$\begin{aligned}
&\lim_{\delta\to 0} \limsup_{N\to\infty} \sup_{\sigma} \mathbb{E}[\min\{1, \|\mu_\sigma^N - \mu_{\sigma+\delta}^N\|_{TV}\}] \\
\le &\lim_{\delta\to 0} \limsup_{N\to\infty} \sup_{\sigma} \mathbb{E}[\|\mu_\sigma^N - \mu_{\sigma+\delta}^N\|_{TV}] \\
\le &\lim_{\delta\to 0} \limsup_{N\to\infty} \sup_{\sigma} \mathbb{E}\left[\left|\{s \in (\sigma, \sigma+\delta] \mid \mu_s^N \ne \mu_{s-}^N\}\right|\right] \cdot \frac{1}{N} \\
\le &\lim_{\delta\to 0} \limsup_{N\to\infty} \sup_{\sigma} N \cdot q_{max} \cdot \delta \cdot \frac{1}{N} = 0.
\end{aligned}$$

The second inequality holds because $\|\mu_s^N - \mu_t^N\|_{TV} = \frac{1}{N}$, provided that in $[s,t]$ only one state change occurs, i.e., one agent changes its state. This follows directly from the characterization of the total variation distance in Lemma A.9. The third inequality holds in the same way as the inequality (3.18) in the proof of Lemma 3.9. Theorem B.19 then states that $(\mu^N)_N$ is relatively compact.

Since $\mathcal{R}$ is compact, so is $\mathcal{R}^{|S|}$ and we directly obtain the relative compactness of $(\hat{\pi}^N)_N$. The relative compactness of the sequence of state-action processes $(\mu^N, \hat{\pi}^N)_N$ follows from Proposition 3.2.4 in Ethier and Kurtz (1986). Thus, a converging subsequence exists. To simplify the notation, we still denote it by (N).

To prove the continuity of the limit state process, define for arbitrary $\mu \in D_{\mathbb{P}(S)}[0,\infty)$

$$J(\mu,u) = \sup_{0\le t\le u} \|\mu_t - \mu_{t-}\|_{TV}.$$
$$J(\mu) = \int_0^\infty e^{-u} J(\mu,u)du.$$

For the sequence of state processes $(\mu^N)_N$ we obtain

$$\lim_{N\to\infty} J(\mu^N) = \lim_{N\to\infty} \int_0^\infty e^{-u} \sup_{0\le t\le u} \|\mu_t^N - \mu_{t-}^N\|_{TV}\, du \le \lim_{N\to\infty} \frac{1}{N} = 0.$$

We exploit the fact that there can only be jumps of height $\frac{1}{N}$ in the state process with N agents. Theorem B.21 then implies the a.s. continuity of the limit state process $(\mu_t^*)_{t\ge 0}$.

It remains to justify the characterization of the limit state process in part b). Due to Skorokhod's representation theorem, see Theorem B.6, we find a probability space $(\tilde\Omega, \tilde{\mathcal{A}}, \tilde{\mathbb{P}})$ with $D_{\mathbb{P}(S)}[0,\infty)$-valued random variables $\tilde\mu^*, (\tilde\mu^N)_{N\in\mathbb{N}}$ and $\mathcal{R}^{|S|}$-valued random variables $\tilde\pi^*, (\tilde\pi^N)_{N\in\mathbb{N}}$ such that

$$(\mu^N, \hat\pi^N) \overset{\mathcal{D}}{=} (\tilde\mu^N, \tilde\pi^N) \qquad (\mu^*, \hat\pi^*) \overset{\mathcal{D}}{=} (\tilde\mu^*, \tilde\pi^*), \quad \text{and} \quad (\tilde\mu^N, \tilde\pi^N) \xrightarrow{a.s.} (\tilde\mu^*, \tilde\pi^*).$$

Define the function

$$M : D_{\mathbb{P}(S)}[0,\infty) \times \mathcal{R}^{|S|} \to D_{\mathbb{R}^{|S|}}[0,\infty), \tag{3.23}$$
$$M_t^{\mu,\hat\pi}(j) := \mu_t(j) - \mu_0(j) - \int_0^t \sum_{i\in S} \mu_s(i) \int_A q(\{j\}|i,a,\mu_s)\hat\pi_s^i(da)ds, \qquad j \in S.$$

We examine the asymptotic behavior of M for the sequences $(\mu^N, \hat\pi^N)_{N\in\mathbb{N}}$ resp. $(\tilde\mu^N, \tilde\pi^N)_{N\in\mathbb{N}}$. Corollary 3.10 provides immediately

$$M^{\mu^N,\hat\pi^N} \xrightarrow{\mathbb{P}} (0,\dots,0), \quad N\to\infty, \tag{3.24}$$

where $(0,\dots,0) \in D_{\mathbb{R}^{|S|}}[0,\infty)$ is the vector of zero-processes.

Now consider the sequence of stochastic processes $(M^{\tilde\mu^N,\tilde\pi^N})_{N\in\mathbb{N}}$. Since the limit process μ^* is almost surely continuous, so is $\tilde\mu^*$ and Theorem B.22 implies uniform convergence on compact intervals, i.e., componentwise for almost all $\omega \in \tilde\Omega$ we obtain

$$\lim_{N\to\infty} \sup_{0\le s\le t} \|\tilde\mu_s^N(\omega) - \tilde\mu_s^*(\omega)\|_{TV} = 0$$

for every $t \in [0,\infty)$. This implies the pointwise convergence

$$\tilde\mu_t^N(\omega) \to \tilde\mu_t^*(\omega), \quad t \in [0,\infty) \tag{3.25}$$

for almost all $\omega \in \tilde\Omega$.

Further, for the integral in the definition of $M^{\tilde{\mu}^N,\tilde{\pi}^N}$ we obtain element-wise for almost all $\omega \in \tilde{\Omega}$

$$\begin{aligned}
&\left|\int_0^t \sum_{i\in S} \tilde{\mu}_s^N(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^N)\tilde{\pi}_s^{N,i}(da)ds - \int_0^t \sum_{i\in S} \tilde{\mu}_s^*(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^*)\tilde{\pi}_s^{*,i}(da)ds\right| \\
\leq &\left|\int_0^t \sum_{i\in S} \tilde{\mu}_s^N(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^N)\tilde{\pi}_s^{N,i}(da)ds - \int_0^t \sum_{i\in S} \tilde{\mu}_s^*(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^*)\tilde{\pi}_s^{N,i}(da)ds\right| \\
+ &\left|\int_0^t \sum_{i\in S} \tilde{\mu}_s^*(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^*)\tilde{\pi}_s^{N,i}(da)ds - \int_0^t \sum_{i\in S} \tilde{\mu}_s^*(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^*)\tilde{\pi}_s^{*,i}(da)ds\right|.
\end{aligned}$$

The second expression tends to 0 for $N \to \infty$ due to the definition of the Young topology and the fact that $a \mapsto q(\{j\}|i,a,\tilde{\mu}_s^*)$ is continuous by assumption. The first expression can be bounded by

$$\begin{aligned}
&\int_0^t \sum_{i\in S} \int_A \left|\tilde{\mu}_s^N(i) q(\{j\}|i,a,\tilde{\mu}_s^N) - \tilde{\mu}_s^*(i) q(\{j\}|i,a,\tilde{\mu}_s^*)\right| \tilde{\pi}_s^{N,i}(da)ds \\
\leq &\int_0^t \sum_{i\in S} \sup_{a\in A} \left|\tilde{\mu}_s^N(i) q(\{j\}|i,a,\tilde{\mu}_s^N) - \tilde{\mu}_s^*(i) q(\{j\}|i,a,\tilde{\mu}_s^*)\right| ds
\end{aligned}$$

which also tends to zero due to dominated convergence, (Q4), (Q5) and Lemma A.5. Together with (3.25) we obtain the almost sure convergence

$$M^{\tilde{\mu}^N,\tilde{\pi}^N} \xrightarrow{a.s.} M^{\tilde{\mu}^*,\tilde{\pi}^*}, \quad N \to \infty,$$

where

$$M_t^{\tilde{\mu}^*,\tilde{\pi}^*}(j) = \tilde{\mu}_t^*(j) - \tilde{\mu}_0^*(j) - \int_0^t \sum_{i\in S} \tilde{\mu}_s^*(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^*)\tilde{\pi}_s^{*,i}(da)ds, \qquad j \in S. \quad (3.26)$$

Now, observe that the function M in (3.23) is measurable. This follows from the measurability of the projections $(\mu,\hat{\pi}) \mapsto M_t^{\mu,\hat{\pi}}$ for each $t \geq 0$ and Proposition 3.7.1 in Ethier and Kurtz (1986). This provides the implication

$$(\mu^N,\hat{\pi}^N) \overset{\mathcal{D}}{=} (\tilde{\mu}^N,\tilde{\pi}^N) \quad \Longrightarrow \quad M^{\mu^N,\hat{\pi}^N} \overset{\mathcal{D}}{=} M^{\tilde{\mu}^N,\tilde{\pi}^N}.$$

Regarding the limit $N \to \infty$, from $M^{\mu^N,\hat{\pi}^N} \Rightarrow (0,\dots,0)$ and the uniqueness of the weak limit, see, e.g., Remark 13.13 in Klenke (2020), we obtain $M^{\tilde{\mu}^*,\tilde{\pi}^*} \overset{\mathcal{D}}{=} (0,\dots,0)$ and thus $M^{\tilde{\mu}^*,\tilde{\pi}^*}(\omega) = (0,\dots,0)$ for almost all $\omega \in \tilde{\Omega}$. Equation (3.26) implies that the limit $\tilde{\mu}^*$ almost surely satisfies the differential equation

$$\tilde{\mu}_t^*(j) = \tilde{\mu}_0^*(j) + \int_0^t \sum_{i\in S} \tilde{\mu}_s^*(i) \int_A q(\{j\}|i,a,\tilde{\mu}_s^*)\tilde{\pi}_s^{*,i}(da)ds, \qquad j \in S. \quad (3.27)$$

In a last step, we use the distributional equality $\mu^* \overset{\mathcal{D}}{=} \tilde{\mu}^*$ resp. $\hat{\pi}^* \overset{\mathcal{D}}{=} \tilde{\pi}^*$ to conclude that μ^* almost surely satisfies the desired differential equation

$$\mu_t^*(j) = \mu_0^*(j) + \int_0^t \sum_{i \in S} \mu_s^*(i) \int_A q(\{j\}|i, a, \mu_s^*) \hat{\pi}_s^{*,i}(da) ds, \qquad j \in S.$$

which concludes the proof. □

In the following, we refer to the differential equation in part b) of the theorem as the *state equation*.

3.3.1. Solvability of the state equation

First, note that the question of the existence of a solution to the state equation does not arise, as Theorem 3.11 ensures the existence of a limit process $(\mu_t^*)_{t \geq 0}$ satisfying the state equation. We therefore proceed directly to the derivation of an intuitive yet important property of the solutions.

Theorem 3.12 (Solutions to the state equation are distributions on S)**.**
Let $(\mu_t)_{t \geq 0}$ be a solution to the state equation (3.22) with initial distribution $\mu_0 \in \mathbb{P}(S)$ and action process $\hat{\pi} = (\hat{\pi}_t)_{t \geq 0}$. Then μ_t constitutes a probability distribution on the state space S for every point in time $t \geq 0$.

Proof. We start with the claim that $(\mu_t)_{t \geq 0}$ is nonnegative. Since the state process starts in the initial distribution μ_0, it is clear that $\mu_0(j) \geq 0$. Let $t_0 \geq 0$ be the first point in time where a component of the process reaches zero, that is,

$$t_0 = \inf\{t \geq 0 \mid \exists j \in S : \ \mu_t(j) = 0\}.$$

The differential form of the state equation is given by

$$\mu_t'(j) = \sum_{i \in S} \mu_t(i) \int_A q(\{j\}|i, a, \mu_t) \hat{\pi}_t^i(da), \quad j \in S, \tag{3.28}$$

with initial value μ_0. Now, let $j_0 \in S$ be a state such that $\mu_{t_0}(j_0) = 0$. Then we obtain for the derivative

$$\begin{aligned} \mu_{t_0}'(j_0) &= \sum_{i \in S \setminus \{j_0\}} \mu_{t_0}(i) \int_A q(\{j_0\}|i, a, \mu_{t_0}) \hat{\pi}_{t_0}^i(da) + \underbrace{\mu_{t_0}(j_0)}_{=0} \int_A q(\{j_0\}|j_0, a, \mu_{t_0}) \hat{\pi}_{t_0}^{j_0}(da) \\ &= \sum_{i \in S \setminus \{j_0\}} \mu_{t_0}(i) \int_A \underbrace{q(\{j_0\}|i, a, \mu_{t_0})}_{\geq 0} \hat{\pi}_{t_0}^i(da) \geq 0, \end{aligned}$$

where (Q1) ensures $q(\{j_0\}|i, a, \mu_{t_0}) \geq 0$. The positivity of the derivative at points of zero, together with the continuity of the paths of the state process, see part a) of Theorem 3.11, implies the nonnegativity of $(\mu_t)_{t \geq 0}$.

It remains to verify the normalization condition $\sum_{j\in S}\mu_t(j)=1$ for every time point $t\geq 0$. We have for $t\geq 0$

$$\begin{aligned}\sum_{j\in S}\mu_t(j) &= \sum_{j\in S}\left(\mu_0(j)+\int_0^t\sum_{i\in S}\mu_s(i)\int_A q(\{j\}|i,a,\mu_s)\hat{\pi}_s^i(da)ds\right)\\ &= \sum_{j\in S}\mu_0(j)+\sum_{j\in S}\int_0^t\sum_{i\in S}\mu_s(i)\int_A q(\{j\}|i,a,\mu_s)\hat{\pi}_s^i(da)ds\\ &= 1+\int_0^t\sum_{i\in S}\mu_s(i)\int_A\underbrace{\sum_{j\in S}q(\{j\}|i,a,\mu_s)}_{=0}\hat{\pi}_s^i(da)ds\\ &= 1,\end{aligned}$$

where (Q2) is applied to establish the last equality. Nonnegativity and normalization then imply the statement. □

Since the limiting state process constitutes a distribution on the state space for every point in time, one of the $|S|$ differential equations in (3.22) can be omitted.

Having established this structural property of solutions to the state equation, we now turn to the question of uniqueness. First of all, recall the following assumption, see 3.1, imposed on the transition intensity q.

(Q4) $\mu\mapsto q(\{j\}|i,a,\mu)$ is continuous w.r.t. weak convergence for all $i,j\in S,\ a\in A$.

In order to guarantee the unique solvability of the state equation, it is sufficient to assume Lipschitz continuity for $\mu\mapsto q(\{j\}|i,a,\mu)$. More precisely, instead of (Q4) we have to assume (Q4'):

(Q4') For all $(i,a)\in S\times A$ and $j\in S$ there exists a uniform constant $L_2>0$ such that

$$|q(\{j\}|i,a,\mu)-q(\{j\}|i,a,\nu)|\leq L_2\|\mu-\nu\|_{TV}$$

for all $\mu,\nu\in\mathbb{P}(S)$.

Under assumption (Q4'), we can derive a Lipschitz property for the integrand of the state equation.

Lemma 3.13. *For an action process $\hat{\pi}=(\hat{\pi}_t)_{t\geq 0}\in\mathcal{R}^{|S|}$ define the function*

$$f:[0,\infty)\times\mathbb{P}(S)\to\mathbb{R}^{|S|},\qquad f_j(t,\mu)=\sum_{i\in S}\mu(i)\int_A q(\{j\}|i,a,\mu)\,\hat{\pi}_t^i(da),\qquad j\in S.$$

Suppose that assumption (Q4') is satisfied. Then there exists a constant $K>0$ such that

$$\|f(t,\nu)-f(t,\mu)\|_1\leq K\cdot\|\nu-\mu\|_{TV}$$

for all $(t,\mu),(t,\nu)\in[0,\infty)\times\mathbb{P}(S)$.

Proof. For arbitrary $\mu, \nu \in \mathbb{P}(S)$ we have in every component $j \in S$

$$\begin{aligned}|f_j(t,\mu) - f_j(t,\nu)| &= \left|\sum_{i\in S}\int_A \Big(\mu(i)q(\{j\}|i,a,\mu) - \nu(i)q(\{j\}|i,a,\nu)\Big)\hat{\pi}_t^i(da)\right| \\ &\leq \sum_{i\in S}\int_A \Big|\mu(i)q(\{j\}|i,a,\mu) - \nu(i)q(\{j\}|i,a,\nu)\Big|\hat{\pi}_t^i(da).\end{aligned}$$

For the absolute value in the integral, it holds that

$$\begin{aligned}&\Big|\mu(i)q(\{j\}|i,a,\mu) - \nu(i)q(\{j\}|i,a,\nu)\Big| \\ \leq\ &\Big|\mu(i)q(\{j\}|i,a,\mu) - \mu(i)q(\{j\}|i,a,\nu) + \mu(i)q(\{j\}|i,a,\nu) - \nu(i)q(\{j\}|i,a,\nu)\Big| \\ =\ &\Big|\mu(i)\Big(q(\{j\}|i,a,\mu) - q(\{j\}|i,a,\nu)\Big) + q(\{j\}|i,a,\nu)\Big(\mu(i) - \nu(i)\Big)\Big| \\ \leq\ &\mu(i)\Big|q(\{j\}|i,a,\mu) - q(\{j\}|i,a,\nu)\Big| + |q(\{j\}|i,a,\nu)|\Big|\mu(i) - \nu(i)\Big| \\ \leq\ &L_2\|\mu-\nu\|_{TV} + 2q_{max}\|\mu-\nu\|_{TV},\end{aligned}$$

where L_2 is the Lipschitz-constant given by (Q4') and q_{max} is the bound for the intensities given by (Q3). Above we obtain

$$\begin{aligned}|f_j(t,\mu) - f_j(t,\nu)| &\leq \sum_{i\in S}\int_A \Big|\mu(i)q(\{j\}|i,a,\mu) - \nu(i)q(\{j\}|i,a,\nu)\Big|\hat{\pi}_t^i(da) \\ &\leq \sum_{i\in S}\int_A (L_2 + 2q_{max})\|\mu-\nu\|_{TV}\ \hat{\pi}_t^i(da) \\ &= |S|(L_2 + 2q_{max})\|\mu-\nu\|_{TV}.\end{aligned}$$

This implies the statement since

$$\|f(t,\nu) - f(t,\mu)\|_1 = \sum_{j\in S}|f_j(t,\mu) - f_j(t,\nu)| \leq \underbrace{|S|^2(L_2 + 2q_{max})}_{=:K}\cdot\|\mu-\nu\|_{TV}.$$

□

The Lipschitz property of the integrand f of the state equation enables us to establish the main uniqueness result of this section.

Theorem 3.14 (Existence and uniqueness of solutions of the state equation)**.**
Let $\mu_0 \in \mathbb{P}(S)$ be an initial distribution and $\hat{\pi} = (\hat{\pi}_t)_{t\geq 0} \in \mathcal{R}^{|S|}$ be an arbitrary action process. Suppose that, in addition to assumptions (Q1)-(Q5), the transition intensities satisfy assumption (Q4').
*Then there exists a **unique** solution $\mu^* = (\mu_t^*)_{t\geq 0}$ that satisfies the state equation*

$$\mu_t(j) = \mu_0(j) + \int_0^t \sum_{i\in S}\mu_s(i)\int_A q(\{j\}|i,a,\mu_s)\hat{\pi}_s^i(da)ds, \qquad j \in S, \tag{3.29}$$

globally on $[0,\infty)$.

Further, let $(\mu_0^N) \subset \mathbb{P}_N(S)$ *be such that* $\mu_0^N \Rightarrow \mu_0$*. Then, for every sequence* $(\hat{\pi}^N) \subset \mathcal{R}^{|S|}$ *with* $\hat{\pi}^N \Rightarrow \hat{\pi}$ *we have*

$$(\mu^N, \hat{\pi}^N) \Rightarrow (\mu^*, \hat{\pi}), \text{ as } N \to \infty,$$

where for each N*,* $\mu^N \in D_{\mathbb{P}(S)}[0, \infty)$ *denotes the state process with initial distribution* μ_0^N *resulting from applying the action process* $\hat{\pi}^N$ *in the* N*-agent problem* (3.10)*.*

Proof. For $N \in \mathbb{N}$, let $\mu_0^N \in \mathbb{P}_N(S)$ be initial distributions such that $\mu_0^N \Rightarrow \mu_0$. Further, let $(\mu_t^N)_{t\geq 0}$ be the N-agent state process with initial distribution μ_0^N resulting from applying the action process $\hat{\pi}$. Then, Theorem 3.11 guarantees the existence of a weakly converging subsequence $\mu^{N_k} \Rightarrow \mu^*$ such that the limit state process μ^* satisfies the state equation (3.29). It remains to show that under assumption (Q4'), the solution μ^* is unique.

Let $(\mu_t)_{t\geq 0}$ and $(\nu_t)_{t\geq 0}$ be two solutions of the state equation with the same initial distribution $\mu_0 \in \mathbb{P}(S)$ and the same action process $\hat{\pi}$. According to Theorem 3.12, we know that $\mu_t \in \mathbb{P}(S)$ and $\nu_t \in \mathbb{P}(S)$ for all $t \geq 0$. For each component $j \in S$, both solutions satisfy the differential equation:

$$\mu_t(j) = \mu_0(j) + \int_0^t f_j(s, \mu_s)\, ds,$$
$$\nu_t(j) = \mu_0(j) + \int_0^t f_j(s, \nu_s)\, ds,$$

where the function f is defined as in Lemma 3.13.

Now consider the difference of the two solutions in terms of the total variation distance. For arbitrary $t \geq 0$ we have

$$\begin{aligned}
\|\mu_t - \nu_t\|_{TV} &= \frac{1}{2} \sum_{j\in S} |\mu_t(j) - \nu_t(j)| \\
&= \frac{1}{2} \sum_{j\in S} \left| \int_0^t (f_j(s, \mu_s) - f_j(s, \nu_s)) ds \right| \\
&\leq \frac{1}{2} \sum_{j\in S} \int_0^t |f_j(s, \mu_s) - f_j(s, \nu_s)| ds \\
&= \int_0^t \left(\frac{1}{2} \sum_{j\in S} |f_j(s, \mu_s) - f_j(s, \nu_s)| \right) ds \\
&= \int_0^t \frac{1}{2} \cdot \|f(s, \mu_s) - f(s, \nu_s)\|_1 \, ds \\
&\leq \int_0^t \frac{1}{2} \cdot K \cdot \|\mu_s - \nu_s\|_{TV} \, ds.
\end{aligned}$$

Gronwall's inequality A.8 applied to $d(t) := \|\mu_t - \nu_t\|_{TV}$ implies that $d(t) = \|\mu_t - \nu_t\|_{TV} \equiv 0$ for all $t \geq 0$. Thus, we conclude that $\mu_t \equiv \nu_t$.

The second part of the theorem regarding the weak convergence of the sequence of state-action processes $(\mu^N, \hat{\pi}^N)$ results from the fact that every weakly convergent subsequence (N_k) has the same limit process $(\mu^*, \hat{\pi})$. Observe that the existence of a convergent subsequence is guaranteed by Theorem 3.11 and that the limit process $(\mu^*, \hat{\pi})$ of every convergent subsequence is characterized by the identical state equation with a unique solution. □

Note that Theorem 3.11 only yields the existence of weakly convergent subsequences, while Theorem 3.29 states conditions for the weak convergence of the full sequence of state-action processes $(\mu^N, \hat{\pi}^N)$ itself.

We conclude the section with an example that indeed demonstrates that the solution to the state equation is not necessarily unique.

Example 3.15. Suppose that the state space is $S = \{1, 2\}$ and the system is uncontrolled. State 1 is absorbing, that is, $q(\{1\}|1, \mu) = q(\{2\}|1, \mu) = 0$ (since the system is uncontrolled, we skip the action from the notation). So agents can only change from state 2 to 1. The intensity of such a change is

$$q(\{1\}|2, \mu) = \begin{cases} \frac{\mu(1)^{\frac{1}{3}}}{1-\mu(1)}, & \text{if } \mu(1) \leq 0.99 \\ \frac{0.99^{\frac{1}{3}}}{0.01} & \text{if } \mu(1) \geq 0.99. \end{cases}$$

Set $q(\{2\}|2, \mu) = -q(\{1\}|2, \mu)$ to satisfy (Q2). Intensities are bounded and continuous. Since we have the relation $\mu_t(1) + \mu_t(2) = 1$, it is sufficient to investigate $\mu_t(1)$. The state equation for $(\mu_t(1))_{t\geq 0}$ in differential form is given by

$$\begin{aligned} \mu_t'(1) &= \mu_t(1)q(\{1\}|1, \mu_t) + (1 - \mu_t(1))q(\{1\}|2, \mu_t) \\ &= \begin{cases} (\mu_t(1))^{\frac{1}{3}}, & \mu_t(1) \leq 0.99, \\ \frac{0.99^{\frac{1}{3}}}{0.01}(1 - \mu_t(1)), & \mu_t(1) \geq 0.99. \end{cases} \end{aligned} \tag{3.30}$$

As the initial condition, we assume that all agents start in state 2, that is, $\mu_0(1) = 0$. Under these conditions, the initial value problem has at least two solutions:

$$\tilde{\mu}_t(1) \equiv 0, \qquad \text{and} \qquad \bar{\mu}_t(1) = \begin{cases} (\frac{2}{3}t)^{\frac{3}{2}}, & t < \frac{3}{2} \cdot 0.99^{\frac{2}{3}}, \\ 1 - C \cdot e^{-99.6655 \cdot t}, & t \geq \frac{3}{2} \cdot 0.99^{\frac{2}{3}}, \end{cases}$$

for a (very large) constant C, such that $1 - C \cdot e^{-99.6655 \cdot \frac{3}{2} \cdot 0.99^{\frac{2}{3}}} \stackrel{!}{=} 0.99$. Note that $t^* = \frac{3}{2} \cdot 0.99^{\frac{2}{3}}$ is the "critical" value in the sense that $(\frac{2}{3}t^*)^{\frac{3}{2}} = 0.99$.

Now consider the following sequence of initial distributions:

$$\mu_0^N := \begin{cases} (0, 1), & N \text{ even}, \quad \text{(all } N \text{ agents start in state 2)} \\ \left(\frac{1}{N}, \frac{N-1}{N}\right), & N \text{ odd}. \quad \text{(exactly one agent starts in state 1)} \end{cases}$$

Obviously $(\mu_0^N)_{N\in\mathbb{N}} \Rightarrow (0,1)$. However, the limit process of the even subsequence is $\tilde{\mu}_t^N(1) \equiv 0$ since the intensity to change from state 2 to 1 remains zero. On the other hand, the odd subsequence converges to the second solution $(\bar{\mu}_t(1))_{t\geq 0}$.

The background of this example is the following: The function $\mu \mapsto q(\{1\}|2,\mu)$ is continuous on $[0,1]$, but not Lipschitz continuous at the point $\mu = 0$. This is easily seen, as the slope of the intensity becomes unbounded as μ approaches zero; see Figure 3.2. Consequently, assumption (Q4) is satisfied, but (Q4') is in fact not. Consequently, solutions to the state equation exist, but uniqueness does not hold.

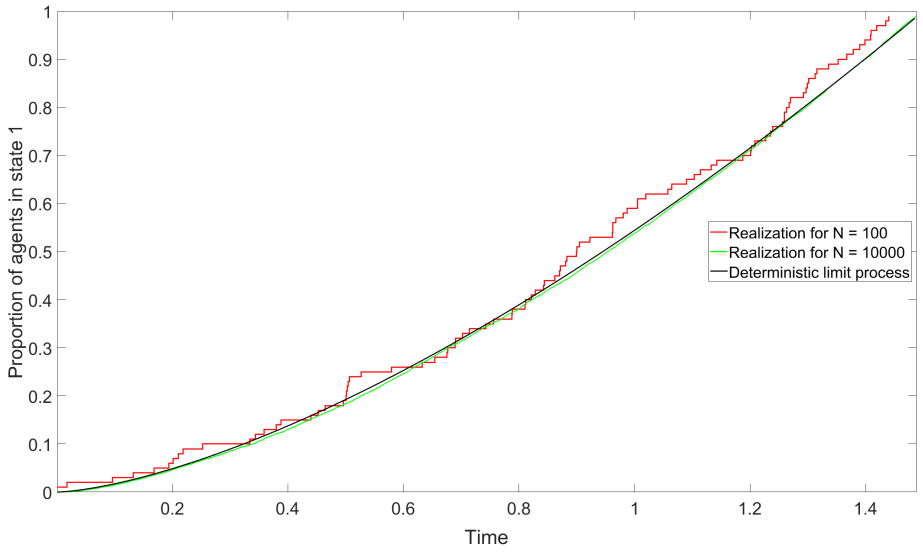

Figure 3.1.: Colourful lines: State trajectories $\mu_t^N(1)$ for $N = 100$ (red) and $N = 10000$ (green) agents in Example 3.15 when one agent starts in state 1.
Black line: Deterministic limit process $(\bar{\mu}_t(1))_{t\geq 0}$.

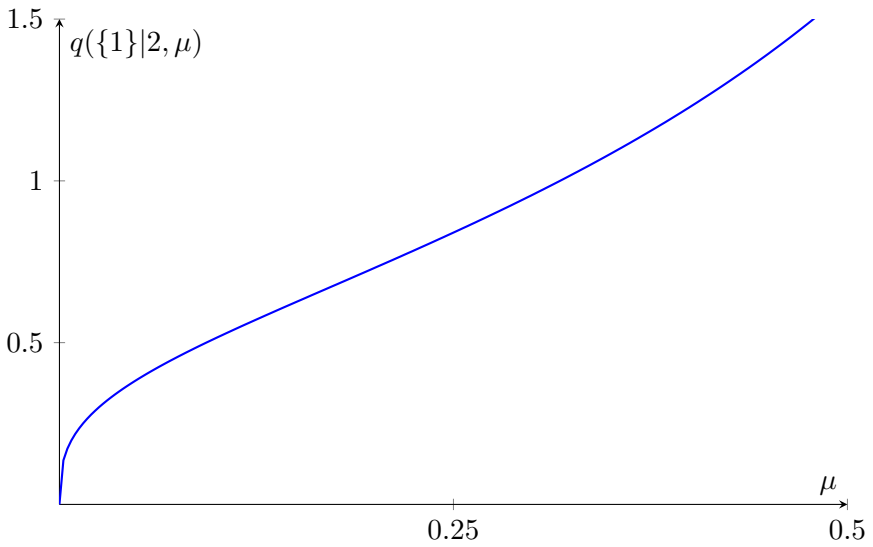

Figure 3.2.: Intensities $q(\{1\}|2,\mu)$ as a function of $\mu \in [0, 0.5]$.

3.4. The deterministic limit model

The key message of the previous section is the result that in the asymptotic case, when the number of agents tends to infinity, the limit state process is almost surely deterministic and characterized by an ordinary differential equation. This property leads to a deterministic limit control problem that serves as an asymptotic upper bound for the optimization problem with N agents. Moreover, an optimal control of the limit model can be used to establish a sequence of asymptotically optimal policies for the N-agent model.

Consider the following deterministic optimization problem:

$$
\begin{aligned}
(F) \qquad & \sup_{\hat{\pi}} \int_0^\infty e^{-\beta t} r(\mu_t, \hat{\pi}_t) dt, \\
& s.t.\ \mu_0 \in \mathbb{P}(S),\ \hat{\pi}_t^i \in \mathbb{P}(A), \\
& \mu_t(j) = \mu_0(j) + \int_0^t \sum_{i \in S} \mu_s(i) \int_A q(\{j\}|i, a, \mu_s) \hat{\pi}_s^i(da) ds, \quad \forall\ t \geq 0,\ j = 1, \ldots, |S|.
\end{aligned} \tag{3.31}
$$

We denote the maximum value of this problem by $V^F(\mu_0)$. The label (F) of the optimization problem is motivated by its similarity to the deterministic fluid problem considered in Bäuerle (2000). In a first step, we show that this value provides an asymptotic upper bound to the value of problem (3.10).

Theorem 3.16 (Asymptotic upper bound)**.**
For all $(\mu_0^N) \subset \mathbb{P}_N(S),\ \mu_0^* \in \mathbb{P}(S)$ *with* $\mu_0^N \Rightarrow \mu_0^*$ *and for all sequences of policies* $(\hat{\pi}^N)_{N \in \mathbb{N}} \subset \mathbb{P}(A)^{|S|}$ *we have*

$$\limsup_{N \to \infty} V_{\hat{\pi}^N}^N(\mu_0^N) \leq V^F(\mu_0^*).$$

Proof. Pick a subsequence (N_k) along which the lim sup is attained, i.e.,

$$\lim_{k \to \infty} V_{\hat{\pi}^{N_k}}^{N_k}(\mu_0^{N_k}) = \limsup_{N \to \infty} V_{\hat{\pi}^N}^N(\mu_0^N).$$

According to Theorem 3.11 there exists a further subsequence (N_{k_l}) of corresponding state and action processes such that

$$(\mu^{N_{k_l}}, \hat{\pi}^{N_{k_l}}) \Rightarrow (\mu^*, \hat{\pi}^*),\ \text{as } l \to \infty.$$

For convenience we still denote this subsequence by (N). As in the proof of Theorem 3.11, due to Skorokhod's representation theorem, see Theorem B.6, we find a probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbb{P}})$ with $D_{\mathbb{P}(S)}[0, \infty)$-valued random variables $\tilde{\mu}^*, (\tilde{\mu}^N)_{N \in \mathbb{N}}$, resp. $\mathcal{R}^{|S|}$-valued random variables $\tilde{\pi}^*, (\tilde{\pi}^N)_{N \in \mathbb{N}}$ such that

$$(\mu^N, \hat{\pi}^N) \overset{\mathcal{D}}{=} (\tilde{\mu}^N, \tilde{\pi}^N) \qquad (\mu^*, \hat{\pi}^*) \overset{\mathcal{D}}{=} (\tilde{\mu}^*, \tilde{\pi}^*), \quad \text{and} \quad (\tilde{\mu}^N, \tilde{\pi}^N) \xrightarrow{a.s.} (\tilde{\mu}^*, \tilde{\pi}^*).$$

For the sequence of value functions along the relabeled subsequence (N), we obtain

$$\begin{aligned}
\lim_{N\to\infty} V^N_{\hat{\pi}^N}(\mu_0^N) &= \lim_{N\to\infty} \mathbb{E}\Big[\int_0^\infty e^{-\beta t} r(\mu_t^N, \hat{\pi}_t^N)dt\Big] \\
&= \lim_{N\to\infty} \mathbb{E}\Big[\int_0^\infty e^{-\beta t} r(\tilde{\mu}_t^N, \tilde{\pi}_t^N)dt\Big] \\
&= \lim_{N\to\infty} \mathbb{E}\Big[\int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^N(i) r(i,a,\tilde{\mu}_t^N)\tilde{\pi}_t^{N,i}(da)dt\Big] \\
&= \mathbb{E}\Big[\lim_{N\to\infty}\int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^N(i) r(i,a,\tilde{\mu}_t^N)\tilde{\pi}_t^{N,i}(da)dt\Big] \\
&= \mathbb{E}\Big[\int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^*(i) r(i,a,\tilde{\mu}_t^*)\tilde{\pi}_t^{*,i}(da)dt\Big] \\
&= \mathbb{E}\Big[\int_0^\infty e^{-\beta t} r(\tilde{\mu}_t^*, \tilde{\pi}_t^*)dt\Big] \\
&\le V^F(\mu_0^*). \qquad (3.32)
\end{aligned}$$

The first equality is just the definition of the value function. The distributional equality $(\mu^N, \hat{\pi}^N) \overset{\mathcal{D}}{=} (\tilde{\mu}^N, \tilde{\pi}^N)$ implies the second equality. The third and sixth equalities follow from the representation (3.9) of the social reward. The fourth equality is obtained by dominated convergence, since μr is bounded; see (3.3). The last inequality is true due to the fact that by Theorem 3.11 the limit process $(\tilde{\mu}^*, \tilde{\pi}^*)$ satisfies the constraints of problem (F). In particular, the limit process $\tilde{\mu}^*$ almost surely satisfies the required differential equation; see (3.27). It remains to show the fifth equality. Componentwise, for almost all $\omega \in \tilde{\Omega}$ we have

$$\begin{aligned}
&\left|\int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^N(i) r(i,a,\tilde{\mu}_t^N)\tilde{\pi}_t^{N,i}(da)dt - \int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^*(i) r(i,a,\tilde{\mu}_t^*)\tilde{\pi}_t^{*,i}(da)dt\right| \\
\le &\left|\int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^N(i) r(i,a,\tilde{\mu}_t^N)\tilde{\pi}_t^{N,i}(da)dt - \int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^*(i) r(i,a,\tilde{\mu}_t^*)\tilde{\pi}_t^{N,i}(da)dt\right| \\
+ &\left|\int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^*(i) r(i,a,\tilde{\mu}_t^*)\tilde{\pi}_t^{N,i}(da)dt - \int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \tilde{\mu}_t^*(i) r(i,a,\tilde{\mu}_t^*)\tilde{\pi}_t^{*,i}(da)dt\right|.
\end{aligned}$$

The second expression tends to zero for $N \to \infty$ due to the definition of the Young topology and the fact that $a \mapsto r(i,a,\mu)$ is continuous by (R2). The first expression can be bounded from above by

$$\begin{aligned}
&\int_0^\infty e^{-\beta t} \sum_{i\in S}\int_A \left|\tilde{\mu}_t^N(i) r(i,a,\tilde{\mu}_t^N) - \tilde{\mu}_t^*(i) r(i,a,\tilde{\mu}_t^*)\right| \tilde{\pi}_t^{N,i}(da)dt \\
\le &\int_0^\infty e^{-\beta t} \sum_{i\in S} \sup_{a\in A}\left|\tilde{\mu}_t^N(i) r(i,a,\tilde{\mu}_t^N) - \tilde{\mu}_t^*(i) r(i,a,\tilde{\mu}_t^*)\right| dt
\end{aligned}$$

which also tends to zero for $N \to \infty$ due to (R1), (R2), Lemma A.5 and dominated convergence. Thus, the statement follows. □

If the state equation in (F) admits a unique solution, we can construct a strategy that is asymptotically optimal in the sense that the upper bound in the previous theorem is attained in the limit. Conditions for unique solvability are provided in Section 3.3.1. Suppose that $(\mu^*, \hat{\pi}^*)$ is an optimal state-action trajectory for problem (F). Then, for the N-agent problem, we consider the strategy

$$\hat{\pi}_t^{N,i} := \hat{\pi}_t^{*,i}, \qquad i \in S, \tag{3.33}$$

which applies the kernel $\hat{\pi}_t^{*,i}$ at each time $t \geq 0$, regardless of the current state μ_t^N of the process. More precisely, the strategy considered is now a deterministic (*open-loop*) policy rather than a (*closed-loop*) feedback policy.

Theorem 3.17. *Suppose $\hat{\pi}^*$ is an optimal control for (F) where the corresponding differential equation in (F) has a unique solution μ^* and let $(\mu_0^N) \subset \mathbb{P}_N(S)$ be such that $\mu_0^N \Rightarrow \mu_0^* \in \mathbb{P}(S)$. Then if we use the strategy $\hat{\pi}^*$ for the problem (3.10) for any N we obtain*

$$\lim_{N\to\infty} V_{\hat{\pi}^*}^N(\mu_0^N) = V^F(\mu_0^*).$$

Thus, we call $\hat{\pi}^*$ asymptotically optimal.

Proof. First, note that $\hat{\pi}^*$ is admissible for each N in the sense of (3.33). In addition, let $(\mu_t^N)_{t\geq 0}$ be the corresponding state process when N agents are present. From Theorem 3.14 we directly obtain that the sequence of state processes itself is weakly convergent to the limit process μ^*, i.e.,

$$\mu^N \Rightarrow \mu^*, \text{ as } N \to \infty.$$

The almost sure pointwise convergence of the Skorokhod representation in (3.25) implies the weak convergence $\mu_t^N \Rightarrow \mu_t^*$ for all $t \geq 0$.

The continuous mapping theorem B.5, together with Lemma 3.3 then ensures for all $t \geq 0$ the weak convergence of the system rewards

$$r(\mu_t^N, \hat{\pi}_t^*) \Rightarrow r(\mu_t^*, \hat{\pi}_t^*), \text{ as } N \to \infty. \tag{3.34}$$

Now, we obtain for the sequence of value functions

$$\begin{aligned}
\lim_{N\to\infty} V_{\hat{\pi}^*}^N(\mu_0^N) &= \lim_{N\to\infty} \mathbb{E}\Big[\int_0^\infty e^{-\beta t} r(\mu_t^N, \hat{\pi}_t^*) dt\Big] \\
&= \lim_{N\to\infty} \int_0^\infty e^{-\beta t}\, \mathbb{E}\big[r(\mu_t^N, \hat{\pi}_t^*)\big] dt \\
&= \int_0^\infty e^{-\beta t} \lim_{N\to\infty} \mathbb{E}\big[r(\mu_t^N, \hat{\pi}_t^*)\big] dt \\
&= \int_0^\infty e^{-\beta t} r(\mu_t^*, \hat{\pi}_t^*) dt \\
&= V^F(\mu_0^*).
\end{aligned}$$

Recall that the rewards are bounded, see (3.3). For the second equality, we interchange expectation and integration by applying Fubini's theorem. The interchange of the limit in the third equality is due to dominated convergence. The fourth equality results from the weak convergence $r(\mu_t^N, \hat{\pi}_t^*) \Rightarrow r(\mu_t^*, \hat{\pi}_t^*)$ together with Lemma B.3. Note that since the limit state process is almost surely deterministic, so is $r(\mu_t^*, \hat{\pi}_t^*)$. The last equality is true due to the fact that $\hat{\pi}^*$ is an optimal control for problem (F) with state trajectory μ^*. □

In the proof of Theorem 3.17 we use the fact that the limit state trajectory is unique. If we drop the assumption of a unique solution in Theorem 3.17 we only obtain

$$\limsup_{N\to\infty} V_{\hat{\pi}^*}^N(\mu_0^N) \leq V^F(\mu_0^*),$$

see Theorem 3.16. For an illustration of a system with multiple solutions of the state equation, see Example 3.15.

A direct implementation of the policy $\hat{\pi}^*$ in the problem (3.10) eventually requires continuous updating of the policy, which may be unfavorable in practice. This can be avoided by using a modified policy instead. We assume here that $t \mapsto \hat{\pi}_t^*$ is piecewise continuous and left-continuous. The restriction to left-continuous controls is merely for convenience and has no practical significance, since from any piecewise continuous optimal policy one can construct an optimal version with left-continuous paths. This holds since action processes which differ only on a countable set of time points yield the same objective value.

Now, let $(\hat{t}_n)_{n\in\mathbb{N}}$ be the discontinuity points in time of $\hat{\pi}^*$ and define the set

$$\{T_n^N, n \in \mathbb{N}\} \cup \{\hat{t}_n, n \in \mathbb{N}\} =: \{\tilde{T}_1^N < \tilde{T}_2^N < \ldots\}$$

where T_n^N describes the time of the n-th jump of the N-agent state process. Then $(\tilde{T}_n^N)$ is the ordered sequence of the time points in this set. Further, we set $\tilde{T}_0^N := 0$. We impose the following assumption on the sequence $\{\tilde{T}_0^N < \tilde{T}_1^N < \tilde{T}_2^N < \ldots\}$:

(Δ) For an increasing number of agents, the N-agent system satisfies for arbitrary $\vartheta > 0$

$$\Delta_N := \sup_{n\in\mathbb{N}:\ \tilde{T}_n^N \leq \vartheta} (\tilde{T}_n^N - \tilde{T}_{n-1}^N) \overset{N\to\infty}{\longrightarrow} 0, \quad \text{a.s.}$$

In other words, assumption (Δ) ensures that the sojourn times of the N-agent system uniformly converge to zero on compact intervals, as the number of agents tends to infinity. This is an intuitive property in many applications, since the intensities of the N-agent system, and hence the number of jumps, grow linearly with the number of agents; see (3.8). Nevertheless, assumption (Δ) is not necessarily satisfied; take for example the degenerate case in which all intensities are zero.

We now turn to the construction of the modified policy. Define

$$\Theta_N(t) := \sum_{n=0}^{\infty} \tilde{T}_n^N \cdot \mathbb{1}_{(\tilde{T}_n^N, \tilde{T}_{n+1}^N]}(t) + \tilde{T}_0^N \cdot \mathbb{1}_{\{0\}}(t).$$

Based on the set of jump times $\{\tilde{T}_0^N < \tilde{T}_1^N < \tilde{T}_2^N < \dots\}$, we define the action process

$$\hat{\pi}_t^{N,*} := \sum_{n=0}^{\infty} \hat{\pi}^*_{\tilde{T}_n^N} \cdot \mathbb{1}_{(\tilde{T}_n^N, \tilde{T}_{n+1}^N]}(t) + \hat{\pi}^*_{\tilde{T}_0^N} \cdot \mathbb{1}_{\{0\}}(t) = \hat{\pi}^*_{\Theta_N(t)}. \tag{3.35}$$

The idea of the action process $(\hat{\pi}_t^{N,*})_{t\geq 0}$ is to update it to match $\hat{\pi}^*$ only when an agent changes its state or when $\hat{\pi}^*$ exhibits a jump and to keep it constant otherwise. It can be shown that this sequence of policies is also asymptotically optimal under mild assumptions.

Theorem 3.18. *Suppose assumption* (Δ) *holds, let* $\hat{\pi}^*$ *be a piecewise continuous, left-continuous optimal control for* (F)*, where the corresponding differential equation in* (F) *has a unique solution* (μ_t^*)*. Further, let* $(\mu_0^N) \subset \mathbb{P}_N(S)$ *be such that* $\mu_0^N \Rightarrow \mu_0 \in \mathbb{P}(S)$*. For each* N*, apply the policy* $(\hat{\pi}_t^{N,*})_{t\geq 0}$ *from* (3.35) *in the* N*-agent problem* (3.10)*. Then*

$$\lim_{N\to\infty} V^N_{\hat{\pi}^{N,*}}(\mu_0^N) = V^F(\mu_0).$$

Proof. First of all, note that the action process $\hat{\pi}^{N,*}$ is stochastic and depends on the trajectory of the N-agent state process. The main idea of the proof is to show the convergence of the action process $\hat{\pi}^{N,*} \to \hat{\pi}^*$ with respect to the Young topology, componentwise for almost all $\omega \in \Omega$. Having ensured that, the statement follows since

$$\lim_{N\to\infty} V^N_{\hat{\pi}^{N,*}}(\mu_0^N) = \lim_{N\to\infty} \mathbb{E}\Big[\int_0^\infty e^{-\beta t} r(\mu_t^{N,*}, \hat{\pi}_t^{N,*}) dt\Big] = \mathbb{E}\Big[\int_0^\infty e^{-\beta t} r(\mu_t^*, \hat{\pi}_t^*) dt\Big] = V^F(\mu_0). \tag{3.36}$$

Here $(\mu_t^{N,*})$ denotes the corresponding state process to strategy $(\hat{\pi}^{N,*})$. Since the state equation for $\hat{\pi}^*$ admits a unique solution (μ_t^*), Theorem 3.14 implies the weak convergence $\mu^{N,*} \Rightarrow \mu^*$. The arguments used to establish (3.36) are the same as those in the proof of Theorem 3.16; in particular, see the chain of equalities preceding inequality (3.32). Note that inequality (3.32) becomes an equality here since $\hat{\pi}^*$ is optimal for (F).

It remains to justify the convergence $\hat{\pi}^{N,*,i} \to \hat{\pi}^{*,i}$ with respect to the Young topology for every $i \in S$, componentwise for almost all $\omega \in \Omega$. Let $\psi : [0,\infty) \times A \to \mathbb{R}$ be an arbitrary function that is measurable in t, continuous in a and satisfies

$$\int_0^\infty \max_{a\in A} |\psi(t,a)| dt < \infty. \tag{3.37}$$

Then $\hat{\pi}^{N,*,i}(\omega) \to \hat{\pi}^{*,i}$ for an $\omega \in \Omega$ is equivalent to the convergence

$$\begin{aligned} &\left| \int_0^\infty \int_A \psi(t,a)\ \hat{\pi}_t^{N,*,i}(da)dt - \int_0^\infty \int_A \psi(t,a)\ \hat{\pi}_t^{*,i}(da)dt \right| \\ = &\left| \int_0^\infty \underbrace{\int_A \psi(t,a)\ \left(\hat{\pi}^{*,i}_{\Theta_N(t)} - \hat{\pi}_t^{*,i}\right)(da)}_{=:f_N(t)}\ dt \right| \quad = \quad \left| \int_0^\infty f_N(t)\ dt \right| \overset{N\to\infty}{\longrightarrow} 0. \end{aligned}$$

Now, fix an $\omega \in \Omega$ for which assumption (Δ) holds and let $\varepsilon > 0$ be arbitrary. Then (3.37) implies that there exists $\vartheta > 0$ such that, independently of N, we have

$$\int_\vartheta^\infty |f_N(t)|\, dt \le \int_\vartheta^\infty \max_{a\in A} |\psi(t,a)| \cdot \|\hat{\pi}^{*,i}_{\Theta_N(t)} - \hat{\pi}_t^{*,i}\|_{TV}\ dt \le \int_\vartheta^\infty \max_{a\in A} |\psi(t,a)| dt \le \frac{\varepsilon}{3}.$$

It remains to show that $\int_0^\vartheta |f_N(t)|\, dt \le \frac{2\varepsilon}{3}$ for N sufficiently large. Since the N-agent state process is càdlàg and the action process $\hat{\pi}^*$ is piecewise continuous in time, the number of elements in the intersection $[0,\vartheta] \cap \{\tilde{T}_1^N < \tilde{T}_2^N < \ldots\}$ is necessarily finite and we denote the elements by $t_1, \ldots, t_{k_N}$. We assume w.l.o.g. that $t_{k_N} < \vartheta$. Now (3.37) implies that we find a small $\delta_1 > 0$, such that for the open intervals $U_j := (t_j, t_j + \delta_1)$, $j = 1, \ldots, k_N$, we obtain

$$\int_{\bigcup_{j=1}^{k_N} U_j} |f_N(t)|\, dt \le \int_{\bigcup_{j=1}^{k_N} U_j} \max_{a\in A} |\psi(t,a)| dt \le \frac{\varepsilon}{3}.$$

Now, define the compact sets $K_1 := [0, t_1]$, $K_j := [t_{j-1} + \delta_1, t_j]$ for $j = 2, \ldots, k_N$ and $K_{k_N+1} := [t_{k_N} + \delta_1, \vartheta]$. Then we have the identity

$$K := [0,\vartheta] \setminus \bigcup_{j=1}^{k_N} U_j = \bigcup_{j=1}^{k_N+1} K_j.$$

Each set K_j, $j = 1, \ldots, k_N + 1$, is compact and does not contain any jumps of the N-agent state process or the action process. In particular, the mapping $t \mapsto \hat{\pi}_t^*$ is continuous on the compact set K and therefore uniformly continuous on K by the Heine–Cantor theorem. Thus, there exists $\delta_2 > 0$ such that for every $t \in K$ with $|\Theta_N(t) - t| < \delta_2$ we have

$$\|\hat{\pi}^{*,i}_{\Theta_N(t)} - \hat{\pi}_t^{*,i}\|_{TV} \le \frac{\varepsilon}{3C_\vartheta}, \tag{3.38}$$

where $C_\vartheta := \int_0^\vartheta \max_{a\in A} |\psi(t,a)| dt$. If $C_\vartheta = 0$, then $\psi = 0$ a.e. on $[0,\vartheta]$, hence $f_N(t) = 0$ on $[0,\vartheta]$ and $\int_0^\vartheta |f_N(t)|\, dt = 0$. Hence, the claim is trivial. Suppose $C_\vartheta > 0$. Note that by construction, for every $t \in K$, the preceding jump time $\Theta_N(t)$ is also contained in K.

Assumption (Δ) now ensures that $\Delta_N := \sup_{n\in\mathbb{N}:\ \tilde{T}_n^N \le \vartheta} (\tilde{T}_n^N - \tilde{T}_{n-1}^N)$ converges uniformly to zero for almost all $\omega \in \Omega$. Therefore, we find $N_0 \in \mathbb{N}$ such that $|\Theta_N(t) - t| \le \Delta_N \le \delta_2$ for all $t \in K$ and consequently (3.38) holds for all $N \ge N_0$. Thus, for $N \ge N_0$ we obtain

$$\int_K |f_N(t)| dt \le \int_K \max_{a\in A} |\psi(t,a)| \cdot \|\hat{\pi}^{*,i}_{\Theta_N(t)} - \hat{\pi}_t^{*,i}\|_{TV}\ dt \le C_\vartheta \cdot \frac{\varepsilon}{3C_\vartheta} = \frac{\varepsilon}{3}.$$

Putting things together, we finally obtain for $N \geq N_0$

$$\int_0^\infty |f_N(t)|dt = \int_\vartheta^\infty |f_N(t)|dt + \int_{[0,\vartheta]\setminus K} |f_N(t)|dt + \int_K |f_N(t)|dt \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$$

which implies the desired convergence $\hat{\pi}^{N,*,i}(\omega) \to \hat{\pi}^{*,i}$ with respect to the Young topology. Thus, the statement follows. □

Remark 3.19. a) General statements about the existence of optimal controls in (F) can only be made under additional assumptions. A classical result is the Theorem of Filippov-Cesari (see Seierstad (1987) Theorem 8 in Chapter II.8 for the finite time horizon problem and Theorem 15 in Chapter III.7 for the infinite horizon problem). We provide the finite-horizon version in detail in Chapter 4, Theorem 4.2.

b) Let $\hat{\pi}_t(\cdot) = \hat{\pi}(\cdot|\mu_t)$ be an optimal *feedback* policy for problem (F). If $\mu \mapsto \hat{\pi}(\cdot|\mu)$ is continuous and the corresponding state equation admits a unique solution μ^*, then this feedback rule is also asymptotically optimal for the problem (3.10). However, if the mapping is not continuous, the convergence may not hold. A particular class of feedback controls that are, in general, discontinuous in μ, is the class of threshold policies, where the chosen action can change abruptly once a certain threshold in μ is crossed. The statement of convergence in the continuous case is presented in the following Corollary 3.20, while an example where convergence does not hold is provided in Section 4.5.

Corollary 3.20 (Asymptotic optimality of continuous feedback policies)**.**
In the situation of Theorem 3.17, suppose that for the problem (F) we obtain an optimal feedback policy $\hat{\pi}_t(\cdot) = \hat{\pi}(\cdot|\mu_t)$. If $\mu \mapsto \hat{\pi}(\cdot|\mu)$ is continuous w.r.t. weak convergence and the corresponding state equation admits a unique solution, then this feedback rule is also asymptotically optimal for the N-agent problem (3.10)*.*

Proof. Let $(\mu^N)_{N\in\mathbb{N}}$ denote the sequence of state processes that result from applying the feedback rule. Define $\hat{\pi}^N := (\hat{\pi}(\cdot|\mu_t^N))_{t\geq 0}$. Theorem 3.11 then ensures the existence of a weakly converging subsequence

$$(\mu^{N_k}, \hat{\pi}^{N_k}) \Rightarrow (\mu^*, \hat{\pi}^*), \text{ as } k \to \infty.$$

As in the proof of Theorem 3.17, the almost sure pointwise convergence of the Skorokhod representation in 3.25 implies the weak convergence $\mu_t^{N_k} \Rightarrow \mu_t^*$ for all $t \geq 0$.

The continuity of $\mu \mapsto \hat{\pi}(\cdot|\mu)$ together with the continuous mapping theorem B.5 implies $\hat{\pi}(\cdot|\mu_t^{N_k}) \Rightarrow \hat{\pi}(\cdot|\mu_t^*)$ for every $t \in [0,\infty)$ as $k \to \infty$. The uniqueness of the weak limit yields $\hat{\pi}^* = (\hat{\pi}(\cdot|\mu_t^*))_{t\geq 0}$.

Now, since the state equation for the feedback rule $\hat{\pi}(\cdot|\mu)$ admits a unique solution, Theorem 3.14 states that the sequence of state processes itself is weakly convergent, i.e

$$\mu^N \Rightarrow \mu^*, \text{ as } N \to \infty,$$

and, in particular, $\hat{\pi}(\cdot|\mu_t^N) \Rightarrow \hat{\pi}(\cdot|\mu_t^*)$ for every $t \in [0, \infty)$.

Next, we obtain for the difference of rewards

$$\begin{aligned}\left|r(\mu_t^N, \hat{\pi}(\cdot|\mu_t^N)) - r(\mu_t^*, \hat{\pi}(\cdot|\mu_t^*))\right| \leq & \underbrace{\left|r(\mu_t^N, \hat{\pi}(\cdot|\mu_t^N)) - r(\mu_t^*, \hat{\pi}(\cdot|\mu_t^N))\right|}_{=:R_1} \\ & + \underbrace{\left|r(\mu_t^*, \hat{\pi}(\cdot|\mu_t^N)) - r(\mu_t^*, \hat{\pi}(\cdot|\mu_t^*))\right|}_{=:R_2}. \end{aligned} \tag{3.39}$$

Using implicitly a Skorokhod representation argument as in the proof of Theorem 3.16, bounding the first term yields

$$\begin{aligned} R_1 &\leq \sum_{i \in S} \int_A \left|\mu_t^N(i) r(i, a, \mu_t^N) - \mu_t^*(i) r(i, a, \mu_t^*)\right| \hat{\pi}^i(da|\mu_t^N) \\ &\leq \sum_{i \in S} \sup_{a \in A} \left|\mu_t^N(i) r(i, a, \mu_t^N) - \mu_t^*(i) r(i, a, \mu_t^*)\right| \end{aligned}$$

which tends to zero for $N \to \infty$ due to (R1), (R2) and Lemma A.5, which implies $R_1 \Rightarrow 0$.

For the second term R_2, we obtain

$$R_2 = \left| \sum_{i \in S} \left(\int_A \mu_t^*(i) r(i, a, \mu_t^*) \hat{\pi}^i(da|\mu_t^N) - \int_A \mu_t^*(i) r(i, a, \mu_t^*) \hat{\pi}^i(da|\mu_t^*) \right) \right|.$$

Assumption (R2) together with $\hat{\pi}(\cdot|\mu_t^N) \Rightarrow \hat{\pi}(\cdot|\mu_t^*)$ yields $R_2 \Rightarrow 0$ as $N \to \infty$. Part c) of Proposition B.4 ensures that R_1 and R_2 converge in probability to zero.

In total, we obtain the convergence in probability of the right-hand side in 3.39 and conclude

$$r(\mu_t^N, \hat{\pi}(\cdot|\mu_t^N)) \Rightarrow r(\mu_t^*, \hat{\pi}(\cdot|\mu_t^*)), \text{ as } N \to \infty. \tag{3.40}$$

From there, the convergence of the sequence of value functions

$$\lim_{N \to \infty} V_{\hat{\pi}}^N(\mu_0^N) = V^F(\mu_0^*)$$

follows from the analogous chain of equations subsequent to (3.34) in the proof of Theorem 3.17. □

To conclude the section, we establish a short result that will be helpful in applying Pontryagin's maximum principle in certain applications. In the case where the transition intensity is affine in the action and the reward is concave in the action, relaxation of

control is unnecessary to attain the optimal value of the deterministic limit problem (F). We show the statement in the case of a one-dimensional control variable. The statement remains valid for higher-dimensional control variables. However, for the sake of clarity in the notation, we restrict ourselves to the one-dimensional case.

Lemma 3.21. *Assume that the action space A is given by an interval $A = [a_{\min}, a_{\max}] \subset \mathbb{R}$. In addition, assume that for each $i, j \in S$ and $\mu \in \mathbb{P}(S)$*

- *$a \mapsto q(\{j\} \mid i, a, \mu)$ is affine, and*
- *$a \mapsto r(i, a, \mu)$ is concave.*

Let $\hat{\pi}^*$ be an optimal (relaxed) control for (F), where the corresponding state equation with initial distribution $\mu_0 \in \mathbb{P}(S)$ admits a unique solution μ^*. Then there exists an unrelaxed control $(a_t)_{t\geq 0}$ taking values in $A^{|S|}$ that is also optimal.

Proof. The idea is to define a_t^i as the expectation of $\hat{\pi}_t^{*,i}$, that is, we have for every $i \in S$

$$a_t^i := \int_A a \; \hat{\pi}_t^{*,i}(da).$$

The convexity of the action space $A = [a_{\min}, a_{\max}]$ implies $(a_t)_{t\geq 0} \subset A^{|S|}$. Since the transition intensity is affine and the reward is concave in the action, due to Jensen's inequality, see Theorem A.6 a) and c), we obtain for every $i, j \in S$ and every $\mu \in \mathbb{P}(S)$

$$r(i, a_t^i, \mu) \geq \int_A r(i, a, \mu)\hat{\pi}_t^{*,i}(da), \qquad q(\{j\} \mid i, a_t^i, \mu) = \int_A q(\{j\} \mid i, a, \mu)\hat{\pi}_t^{*,i}(da).$$

Consequently, $(a_t)_{t\geq 0}$ induces the same (unique) state process as $(\hat{\pi}_t^*)_{t\geq 0}$, since the state equations of both controls coincide. In addition, regarding the rewards, we have along the corresponding optimal state trajectory

$$\begin{aligned} V^F(\mu_0) = \int_0^\infty e^{-\beta t} r(\mu_t^*, \hat{\pi}_t^*)dt &= \int_0^\infty e^{-\beta t} \sum_{i\in S} \int_A r(i, a, \mu_t^*)\hat{\pi}_t^{*,i}(da)\mu_t^*(i)dt \\ &\leq \int_0^\infty e^{-\beta t} \sum_{i\in S} r(i, a_t^i, \mu_t^*)\mu_t^*(i)dt \\ &= \int_0^\infty e^{-\beta t} r(\mu_t^*, a_t)dt \\ &\leq V^F(\mu_0). \end{aligned}$$

We conclude that $(\hat{\pi}_t^*)_{t\geq 0}$ and $(a_t)_{t\geq 0}$ are equivalent in the sense that both generate the same value, which implies the statement. □

The proof for an m-dimensional control variable where the action space is given by a hyperrectangle in $\mathbb{R}^m$ is essentially the same; the unrelaxed action process $(a_1(t), \ldots, a_m(t))$ is obtained by taking the expectation for each component $j = 1, \ldots, m$ of the relaxed action process.

3.4.1. RATE OF CONVERGENCE IN THE FINITE HORIZON PROBLEM

In the previous section, we analyzed the asymptotic behavior of the value functions of the N-agent model in comparison with the optimal value V^F of the limit model (F) in the case of an *infinite* time horizon. We now turn to problems with a *finite* time horizon. In this setting, it is possible not only to study convergence but also to derive results on the rate of convergence. Instead of (3.10), we consider the following N-agent problem:

For a control $\hat{\pi}$, a discount rate $\beta > 0$ and an initial configuration $\mu_0 \in \mathbb{P}_N(S)$, define the finite-horizon value function

$$\begin{aligned} V_{\hat{\pi}}^{N,T}(\mu_0) &= \mathbb{E}_{\mu_0}^{\hat{\pi}}\Big[\int_0^T e^{-\beta t} r(\mu_t^N, \hat{\pi}_t) dt + g(\mu_T^N)\Big], \\ V^{N,T}(\mu_0) &= \sup_{\hat{\pi}} V_{\hat{\pi}}^{N,T}(\mu_0). \end{aligned} \tag{3.41}$$

with terminal time $T > 0$ and terminal reward $g : \mathbb{P}(S) \to \mathbb{R}$ for the final state.

The corresponding finite-horizon limit problem based on (F) is given by

$$\begin{aligned} (F_T) \quad & \sup_{\hat{\pi}} \int_0^T e^{-\beta t} r(\mu_t, \hat{\pi}_t) dt + g(\mu_T) \\ & s.t.\ \mu_0 \in \mathbb{P}(S),\ \hat{\pi}_t^i \in \mathbb{P}(A), \\ & \mu_t(j) = \mu_0(j) + \int_0^t \sum_{i \in S} \mu_s(i) \int_A q(\{j\}|i, a, \mu_s) \hat{\pi}_s^i(da) ds, \quad t \in [0,T],\ j \in S. \end{aligned} \tag{3.42}$$

We denote the optimal value by $V^{F,T}$.

Remark 3.22 (Restriction of Theorem 3.11 to a finite time horizon)**.**
The results of Theorem 3.11 directly transfer from the infinite-horizon setting to any finite time interval $[0,T]$. To illustrate this, suppose $(\mu^N)_{N\in\mathbb{N}} \subset D_{\mathbb{P}(S)}[0,\infty)$ is a weakly converging sequence of state processes with $\mu^N \Rightarrow \mu^*$. Denote the restriction of the processes to $[0,T]$ by $\mu_{|[0,T]}$. Since the limit process μ^* has almost surely continuous paths, we directly obtain the weak convergence $\mu^N_{|[0,T]} \Rightarrow \mu^*_{|[0,T]}$. This is true since convergence in $D_{\mathbb{P}(S)}[0,\infty)$ is equivalent to convergence in $D_{\mathbb{P}(S)}[0,T]$ for each continuity point T of the limit process, see Theorem 16.2 in Billingsley (1999).
A similar restriction argument applies to the convergence of action processes with respect to the Young topology, e.g., by testing against continuous extensions of admissible test functions $\psi : [0,T] \times A \to \mathbb{R}$.

We now aim to transfer Theorem 3.17 to the finite-horizon setting. Under stricter assumptions, it is possible to prove that the rate of convergence of the value functions in the finite-horizon problem (3.41) is $1/\sqrt{N}$. In order to obtain this rate, we impose Lipschitz conditions on the reward function and the intensity functions. Thus, assume that (Q4') is satisfied (see Section 3.3.1). Additionally, we impose *one* of the following two Lipschitz conditions on the reward function:

(R1') For all $(i,a) \in S \times A$ there exists a uniform constant $L_1 > 0$ s.t.

$$\begin{aligned} |\mu(i)r(i,a,\mu) - \nu(i)r(i,a,\nu)| &\leq L_1 \|\mu - \nu\|_{TV}, \\ |g(\mu) - g(\nu)| &\leq L_1 \|\mu - \nu\|_{TV} \end{aligned}$$

for all $\mu, \nu \in \mathbb{P}(S)$.

(R1") For all $(i,a) \in S \times A$ there exist uniform constants $L_1 > 0$ and $r_{max} > 0$ s.t.

$$\begin{aligned} |r(i,a,\mu) - r(i,a,\nu)| &\leq L_1 \|\mu - \nu\|_{TV}, \\ |g(\mu) - g(\nu)| &\leq L_1 \|\mu - \nu\|_{TV} \\ \sup_{(i,a)\in S\times A,\ \mu\in\mathbb{P}(S)} |r(i,a,\mu)| &\leq r_{max} \end{aligned}$$

for all $\mu, \nu \in \mathbb{P}(S)$.

We start with the observation that the Lipschitz continuity of the single-agent reward function extends directly to the system's reward function.

Lemma 3.23. *Let $\alpha \in \mathbb{P}(A)^{|S|}$ be an action of the system of agents, and $\mu, \nu \in \mathbb{P}(S)$ two empirical distributions of the agents on S. Then, under either (R1') or (R1"), there exists a constant $L_3 > 0$ such that the reward rate of the system is bounded by*

$$|r(\mu,\alpha) - r(\nu,\alpha)| \leq L_3 \|\mu - \nu\|_{TV}.$$

Proof. We start by proving the claim under (R1'):

$$\begin{aligned} |r(\mu,\alpha) - r(\nu,\alpha)| &= \Big| \sum_{i\in S} \int_A \mu(i) r(i,a,\mu) \alpha^i(da) - \sum_{i\in S} \int_A \nu(i) r(i,a,\nu) \alpha^i(da) \Big| \\ &\leq \sum_{i\in S} \int_A \big|\mu(i) r(i,a,\mu) - \nu(i) r(i,a,\nu)\big| \alpha^i(da) \\ &\leq \sum_{i\in S} \int_A L_1 \|\mu - \nu\|_{TV} \alpha^i(da) \\ &= \underbrace{|S| \cdot L_1}_{=:L_3} \ \|\mu - \nu\|_{TV} \end{aligned}$$

Note that the second inequality follows from assumption (R1').

Under assumption (R1"), we obtain

$$\begin{aligned} |r(\mu,\alpha) - r(\nu,\alpha)| &= \Big| \sum_{i\in S} \int_A \mu(i) r(i,a,\mu) \alpha^i(da) - \sum_{i\in S} \int_A \nu(i) r(i,a,\nu) \alpha^i(da) \Big| \\ &\leq \sum_{i\in S} \int_A \big|\mu(i) r(i,a,\mu) - \nu(i) r(i,a,\nu)\big| \alpha^i(da) \\ &= \sum_{i\in S} \int_A \big|\mu(i) r(i,a,\mu) - \nu(i) r(i,a,\mu) + \nu(i) r(i,a,\mu) - \nu(i) r(i,a,\nu)\big| \alpha^i(da) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in S}\int_A |r(i,a,\mu)(\mu(i)-\nu(i)) + \nu(i)(r(i,a,\mu)-r(i,a,\nu))|\alpha^i(da)\\
&\le \sum_{i\in S}\int_A |r(i,a,\mu)||\mu(i)-\nu(i)| + \nu(i)|r(i,a,\mu)-r(i,a,\nu)|\alpha^i(da)\\
&\overset{(1)}{\le} \sum_{i\in S}\int_A r_{max}|\mu(i)-\nu(i)| + |r(i,a,\mu)-r(i,a,\nu)|\alpha^i(da)\\
&\overset{(2)}{\le} \sum_{i\in S}\int_A 2r_{max}\|\mu-\nu\|_{TV} + L_1\|\mu-\nu\|_{TV}\ \alpha^i(da)\\
&= \underbrace{|S|(2r_{max}+L_1)}_{=:L_3}\ \|\mu-\nu\|_{TV}.
\end{aligned}
$$

Inequality (1) follows from the boundedness of r. We obtain (2) by applying (R1") and the representation $\|\mu_t^N - \mu_t^*\|_{TV} = \frac{1}{2}\sum_{i\in S}|\mu_t^N(i)-\mu_t^*(i)|$, see Lemma A.9. □

Having established the Lipschitz continuity of the rewards of the system, we now derive the rate of convergence for the value functions.

Theorem 3.24 (Rate of convergence of the value functions)**.**
In the finite-horizon setting, under assumptions (Q1)–(Q5), with (Q4) replaced by (Q4') and (R1) replaced by either (R1') or (R1") (along with (R2)), suppose that $\mathbb{E}\left[\|\mu_0^N - \mu_0\|_{TV}\right] \le \frac{L_0}{\sqrt{N}}$ *for a constant* $L_0 > 0$*. Furthermore, let* $\hat{\pi}^*$ *be an optimal control of* (F_T)*. Then*

$$\left|V_{\hat{\pi}^*}^{N,T}(\mu_0^N) - V^{F,T}(\mu_0)\right| \le \frac{\tilde{L}}{\sqrt{N}}$$

for a constant $\tilde{L} > 0$ *that is independent of* N *but depends on* T*.*

Proof. Let $(\mu_t^*)_{t\in[0,T]}$ be the unique solution of the finite-horizon state equation

$$\mu_t(j) = \mu_0(j) + \int_0^t \sum_{i\in S}\mu_s(i)\int_A q(\{j\}|i,a,\mu_s)\hat{\pi}_s^{*,i}(da)ds, \quad \forall\ t\in[0,T],\ j\in S.$$

Note that such a unique solution exists; see Theorem 3.14. Additionally, let $(\mu_t^N)_{t\in[0,T]}$ be the corresponding state process in the N-agent model when each agent is executing policy $\hat{\pi}^*$; see (3.33). The object of interest is the difference

$$
\begin{aligned}
\left|V_{\hat{\pi}^*}^{N,T}(\mu_0^N) - V^{F,T}(\mu_0)\right| &= \left|\mathbb{E}_{\mu_0^N}^{\hat{\pi}^*}\Big[\int_0^T e^{-\beta t}r(\mu_t^N,\hat{\pi}_t^*)dt + g(\mu_T^N)\Big] - \int_0^T e^{-\beta t}r(\mu_t^*,\hat{\pi}_t^*)dt - g(\mu_T^*)\right|\\
&\le \mathbb{E}_{\mu_0^N}^{\hat{\pi}^*}\Big[\int_0^T e^{-\beta t}|r(\mu_t^N,\hat{\pi}_t^*) - r(\mu_t^*,\hat{\pi}_t^*)|dt + |g(\mu_T^N) - g(\mu_T^*)|\Big]
\end{aligned}
\tag{3.43}
$$

We show that the difference in rewards is bounded. Assumptions (R1') or (R1") imply immediately

$$|g(\mu_T^N) - g(\mu_T^*)| \le L_1\|\mu_T^N - \mu_T^*\|_{TV}.$$

Furthermore, Lemma 3.23 implies the Lipschitz continuity of $\mu \mapsto r(\mu, \pi^*)$ with a constant $L_3 > 0$:

$$|r(\mu_t^N, \hat{\pi}_t^*) - r(\mu_t^*, \hat{\pi}_t^*)| \leq L_3 \|\mu_t^N - \mu_t^*\|_{TV}$$

Inserting the bounds for the reward in 3.43 results in the inequality

$$\left|V_{\hat{\pi}^*}^{N,T}(\mu_0^N) - V^{F,T}(\mu_0)\right| \leq \mathbb{E}_{\mu_0^N}^{\hat{\pi}^*}\Big[\int_0^T e^{-\beta t} L_3 \|\mu_t^N - \mu_t^*\|_{TV} dt + L_1 \|\mu_T^N - \mu_T^*\|_{TV}\Big]. \quad (3.44)$$

Next, we aim to derive a bound for $\mathbb{E}_{\mu_0^N}^{\hat{\pi}^*}[\|\mu_t^N - \mu_t^*\|_{TV}]$. Recall from (3.15) the representation

$$\mu_t^N(j) = \mu_0^N(j) + \int_0^t \sum_{i \in S} \mu_s^N(i) \int_A q(\{j\}|i, a, \mu_s^N) \hat{\pi}_s^{*,i}(da) ds + M_t^N(j),$$

for all $t \in [0, T]$ and $j \in S$. Together with the state equation, we obtain for every $j \in S$

$$\begin{aligned}
|\mu_t^N(j) - \mu_t^*(j)| = & \Big|\mu_0^N(j) + \int_0^t \sum_{i \in S} \mu_s^N(i) \int_A q(\{j\}|i, a, \mu_s^N) \hat{\pi}_s^{*,i}(da) ds + M_t^N(j) \\
& - \mu_0(j) - \int_0^t \sum_{i \in S} \mu_s^*(i) \int_A q(\{j\}|i, a, \mu_s^*) \hat{\pi}_s^{*,i}(da) ds\Big| \\
= & \Big|\mu_0^N(j) + \int_0^t \sum_{i \in S} \mu_s^N(i) \int_A q(\{j\}|i, a, \mu_s^N) \hat{\pi}_s^{*,i}(da) ds + M_t^N(j) \\
& - \int_0^t \sum_{i \in S} \mu_s^N(i) \int_A q(\{j\}|i, a, \mu_s^*) \hat{\pi}_s^{*,i}(da) ds + \int_0^t \sum_{i \in S} \mu_s^N(i) \int_A q(\{j\}|i, a, \mu_s^*) \hat{\pi}_s^{*,i}(da) ds \\
& - \mu_0(j) - \int_0^t \sum_{i \in S} \mu_s^*(i) \int_A q(\{j\}|i, a, \mu_s^*) \hat{\pi}_s^{*,i}(da) ds\Big| \\
\leq & |\mu_0^N(j) - \mu_0(j)| + \int_0^t \sum_{i \in S} \mu_s^N(i) \int_A |q(\{j\}|i, a, \mu_s^N) - q(\{j\}|i, a, \mu_s^*)| \hat{\pi}_s^{*,i}(da) ds \\
& + \int_0^t \sum_{i \in S} |\mu_s^N(i) - \mu_s^*(i)| \int_A q(\{j\}|i, a, \mu_s^*) \hat{\pi}_s^{*,i}(da) | ds + |M_t^N(j)| \\
\overset{(1)}{\leq} & |\mu_0^N(j) - \mu_0(j)| + \int_0^t L_2 \|\mu_s^N - \mu_s^*\|_{TV} ds + q_{max} \int_0^t \sum_{i \in S} |\mu_s^N(i) - \mu_s^*(i)| ds + |M_t^N(j)| \\
\overset{(2)}{\leq} & |\mu_0^N(j) - \mu_0(j)| + (L_2 + 2q_{max}) \int_0^t \|\mu_s^N - \mu_s^*\|_{TV} ds + |M_t^N(j)|.
\end{aligned}$$

To obtain (1) apply (Q3) and (Q4'). The last inequality (2) follows from the representation of the total variation distance in Lemma A.9. Using Jensen's inequality (see Theorem A.6), we obtain

$$\mathbb{E}_{\mu_0^N}^{\hat{\pi}^*}[|M_t^N(j)|] \overset{\text{Jensen}}{\leq} \sqrt{\mathbb{E}_{\mu_0^N}^{\hat{\pi}^*}[(M_t^N(j))^2]} \leq \frac{\sqrt{q_{max} t}}{\sqrt{N}},$$

where the second inequality is already shown in the proof of Lemma 3.9, see (3.18). Note that the root function is strictly concave on $[0, \infty)$.

Now define the constants

$$L_4 := \frac{1}{2} \cdot |S| \cdot (L_2 + 2q_{max}), \qquad L_5 := \frac{1}{2} \cdot |S| \cdot \sqrt{q_{max}}.$$

Then, for every $t \in [0, T]$, we can establish the desired bound

$$\begin{aligned}
\mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}[\|\mu_t^N - \mu_t^*\|_{TV}] &= \mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\left[\frac{1}{2}\sum_{j\in S}|\mu_t^N(j) - \mu_t^*(j)|\right] \\
&\leq \frac{1}{2}\sum_{j\in S}\mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\left[|\mu_0^N(j) - \mu_0(j)| + (L_2 + 2q_{max})\int_0^t \|\mu_s^N - \mu_s^*\|_{TV} ds + |M_t^N(j)|\right] \\
&\leq \mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}[\|\mu_0^N - \mu_0\|_{TV}] + L_4\int_0^t \mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}[\|\mu_s^N - \mu_s^*\|_{TV}]ds + \frac{L_5}{\sqrt{N}}\sqrt{t} \\
&\leq \frac{L_0}{\sqrt{N}} + \frac{L_5}{\sqrt{N}}\sqrt{t} + L_4\int_0^t \mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}[\|\mu_s^N - \mu_s^*\|_{TV}]ds.
\end{aligned}$$

Gronwall's inequality, see Lemma A.8, implies that for all $t \in [0, T]$

$$\mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}[\|\mu_t^N - \mu_t^*\|_{TV}] \leq \frac{1}{\sqrt{N}}(L_0 + L_5\sqrt{t})e^{L_4 t}.$$

Finally, in (3.44) we obtain

$$\begin{aligned}
\left|V^{N,T}_{\hat{\pi}^*}(\mu_0^N) - V^{F,T}(\mu_0)\right| &\leq \mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\Big[\int_0^T e^{-\beta t}L_3\|\mu_t^N - \mu_t^*\|_{TV}dt + L_1\|\mu_T^N - \mu_T^*\|_{TV}\Big] \\
&= L_3\int_0^T e^{-\beta t}\,\mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\Big[\|\mu_t^N - \mu_t^*\|_{TV}\Big]dt + L_1\,\mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\Big[\|\mu_T^N - \mu_T^*\|_{TV}\Big] \\
&\leq L_3\int_0^T e^{-\beta t}\,\frac{1}{\sqrt{N}}(L_0 + L_5\sqrt{t})\,e^{L_4 t}\,dt + L_1\frac{1}{\sqrt{N}}\,(L_0 + L_5\sqrt{T})\,e^{L_4 T} \\
&= \frac{1}{\sqrt{N}}\left(L_3\int_0^T (L_0 + L_5\sqrt{t})\,e^{-(\beta - L_4)t}\,dt + L_1(L_0 + L_5\sqrt{T})\,e^{L_4 T}\right) \\
&=: \frac{\tilde{L}}{\sqrt{N}},
\end{aligned}$$

which in turn implies the statement. □

The statement about the convergence rate can be extended to the infinite horizon problem when the discount rate is large enough.

Corollary 3.25 (Rate of convergence for infinite time horizon)**.**
In the infinite-horizon setting, under assumptions (Q1)–(Q5), with (Q4) replaced by (Q4') and (R1) replaced by either (R1') or (R1") (along with (R2)), suppose that $\mathbb{E}\left[\|\mu_0^N - \mu_0\|_{TV}\right] \leq \frac{L_0}{\sqrt{N}}$ *for a constant* $L_0 > 0$*. Furthermore, let* $\hat{\pi}^*$ *be an optimal control of* (F)*.*

If the discount rate satisfies $\beta > L_4$, then

$$\left|V^N_{\hat{\pi}^*}(\mu_0^N) - V^F(\mu_0)\right| \leq \frac{\tilde{L}}{\sqrt{N}}$$

for a constant $\tilde{L} > 0$ that is independent of N.

Proof. First of all, note that the terminal reward is absent in the infinite-horizon setting. Using the identical arguments as in the proof of Theorem 3.24, we obtain for the difference of value functions

$$\begin{aligned}
\left|V^N_{\hat{\pi}^*}(\mu_0^N) - V^F(\mu_0)\right| &= \left|\mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\Big[\int_0^\infty e^{-\beta t} r(\mu_t^N, \hat{\pi}_t^*)dt\Big] - \int_0^\infty e^{-\beta t} r(\mu_t^*, \hat{\pi}_t^*)dt\right| \\
&\leq \mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\Big[\int_0^\infty e^{-\beta t} |r(\mu_t^N, \hat{\pi}_t^*) - r(\mu_t^*, \hat{\pi}_t^*)|dt\Big] \\
&\leq \mathbb{E}^{\hat{\pi}^*}_{\mu_0^N}\Big[\int_0^\infty e^{-\beta t} L_3 \|\mu_t^N - \mu_t^*\|_{TV} dt\Big] \\
&\leq L_3 \int_0^\infty e^{-\beta t} \frac{1}{\sqrt{N}}(L_0 + L_5\sqrt{t})\, e^{L_4 t} dt \\
&= \frac{1}{\sqrt{N}}\left(L_3 \int_0^\infty (L_0 + L_5\sqrt{t})\, e^{-(\beta - L_4)t} dt\right) \\
&\overset{\beta > L_4}{=} \frac{1}{\sqrt{N}} L_3 \left(\frac{L_0}{\beta - L_4} + \frac{L_5\sqrt{\pi}}{2(\beta - L_4)^{3/2}}\right) \\
&=: \frac{\tilde{L}}{\sqrt{N}}.
\end{aligned}$$

□

Remark 3.26. Note that the assumption $\mathbb{E}\left[\|\mu_0^N - \mu_0\|_{TV}\right] \leq \frac{L_0}{\sqrt{N}}$ in Theorem 3.24 is satisfied if the initial states of the N agents are sampled i.i.d. from a distribution $\mu_0 \in \mathbb{P}(S)$.

Proof. We denote the initial state of the agent $k \in \{1, \ldots, N\}$ by $X_0^k \sim \mu_0$. Since agents receive their initial state independently according to μ_0, the number of agents with initial state $i \in S$ is binomially distributed: For an arbitrary state $i \in S$ we have

$$\sum_{k=1}^N \mathbb{1}_{\{X_0^k = i\}} \sim \mathrm{Bin}(N, \mu_0(i))$$

To simplify the notation, we define $p := \mu_0(i)$. Furthermore, we identify μ_0^N with the relative distribution of the N agents over the states, that is

$$\mu_0^N(i) := \frac{1}{N}\sum_{k=1}^N \delta_{\{X_0^k = i\}}, \qquad i \in S.$$

We obtain

$$\mathrm{Var}\left(\mu_0^N(i)\right) = \mathrm{Var}\left(\frac{1}{N}\sum_{k=1}^N \delta_{\{X_0^k = i\}}\right) = \frac{p(1-p)}{N}.$$

Using Chebyshev's inequality, see Theorem A.7, we get for every $i \in S$

$$\begin{aligned}
\mathbb{E}\left[|\mu_0^N(i) - \mu_0(i)|\right] &= \mathbb{E}\left[\left|\mu_0^N(i) - p\right|\right] \\
&= \int_0^\infty \mathbb{P}\Big(\left|\mu_0^N(i) - p\right| \geq \gamma\Big) d\gamma \\
&= \int_0^{\sqrt{\frac{p(1-p)}{N}}} \mathbb{P}\Big(\left|\mu_0^N(i) - p\right| \geq \gamma\Big) d\gamma + \int_{\sqrt{\frac{p(1-p)}{N}}}^\infty \mathbb{P}\Big(\left|\mu_0^N(i) - p\right| \geq \gamma\Big) d\gamma \\
&\leq \sqrt{\frac{p(1-p)}{N}} + \int_{\sqrt{\frac{p(1-p)}{N}}}^\infty \frac{p(1-p)}{N\gamma^2} d\gamma \\
&= 2\sqrt{\frac{p(1-p)}{N}}.
\end{aligned}$$

This concludes the proof since

$$\begin{aligned}
\mathbb{E}\left[\|\mu_0^N - \mu_0\|_{TV}\right] &= \mathbb{E}\left[\frac{1}{2}\sum_{i\in S} |\mu_0^N(i) - \mu_0(i)|\right] \\
&= \frac{1}{2}\sum_{i\in S} \mathbb{E}\left[|\mu_0^N(i) - \mu_0(i)|\right] \leq \frac{1}{\sqrt{N}} \underbrace{\sum_{i\in S} \sqrt{\mu_0(i)(1-\mu_0(i))}}_{=:L_0}
\end{aligned}$$

□

3.4.2. RESOURCE CONSTRAINTS FOR THE ACTION PROCESS

Recall that the set of action processes is given by $\mathcal{R}^{|S|}$, where

$$\mathcal{R} := \{\rho : [0,\infty) \to \mathbb{P}(A) \mid \rho \text{ measurable}\}.$$

The only relevant property used for the relative compactness of a sequence of action processes in $\mathcal{R}^{|S|}$ and for the results derived therefrom is the compactness of the set $\mathcal{R}^{|S|}$. Consequently, all results remain valid even if we restrict the action processes to compact subsets of $\mathcal{R}^{|S|}$.

An important special case is given by so-called resource constraints. In this setting, the central controller does not have unrestricted access to every action in every state. More precisely, there exists an action $\tilde{a} \in A$, a subset of states $\tilde{S} \subset S$, and a constant $c < |\tilde{S}|$ such that for almost all $t \geq 0$, the action process $\hat{\pi} \in \mathcal{R}^{|S|}$ has to satisfy

$$\sum_{i\in\tilde{S}} \hat{\pi}_t^i(\{\tilde{a}\}) \leq c < |\tilde{S}|.$$

Note that modifying the action process at time points of Lebesgue measure zero does not affect its value. We therefore only require the resource constraint to be satisfied for almost all $t \geq 0$.

In the case of a finite action space A, the following theorem ensures that the above restriction induces a compact subset of $\mathcal{R}^{|S|}$. An example of a model with a resource constraint in the action is provided in Section 4.5.

Theorem 3.27 (Resource constraints for finite action space)**.**
Let $\tilde{a} \in A$, $\tilde{S} \subset S$, and $c < |\tilde{S}|$. Suppose that the set of actions A is finite. Then the restricted set of action processes

$$\mathcal{R}_{RC} := \left\{ \rho \in \mathcal{R}^{|S|} \;\middle|\; \sum_{i \in \tilde{S}} \rho_t^i(\{\tilde{a}\}) \leq c \;\; \textit{for almost all } t \geq 0 \right\} \subset \mathcal{R}^{|S|}$$

is compact.

Proof. Since closed subsets of compact sets are compact as well, see Theorem (11.3) in Davis (1993), it is sufficient to show that $\mathcal{R}_{RC}$ is closed. Thus, let $(\rho^n)_{n\in\mathbb{N}} \subset \mathcal{R}_{RC}$ be a sequence of restricted action processes that converges to $\rho \in \mathcal{R}^{|S|}$ with respect to the Young topology, that is,

$$\int_0^\infty \int_A \psi(t,a)\rho_t^{i,n}(da)dt \;\longrightarrow\; \int_0^\infty \int_A \psi(t,a)\rho_t^{i}(da)dt \tag{3.45}$$

for every $i \in S$ and for all $\psi : [0,\infty) \times A \to \mathbb{R}$ such that ψ is measurable in t and $\int_0^\infty \max_{a\in A} |\psi(t,a)|dt < \infty$. Due to the finiteness of A, the continuity of ψ in a is trivial.

We must show that $\rho \in \mathcal{R}_{RC}$. In order to prove this, suppose that $\rho \notin \mathcal{R}_{RC}$. Then there exists $\varepsilon > 0$ such that the set

$$\mathcal{T} := \left\{ t \geq 0 \;\middle|\; \sum_{i \in \tilde{S}} \rho_t^i(\{\tilde{a}\}) \geq c + \varepsilon \right\}$$

has positive Lebesgue measure $\lambda(\mathcal{T}) > 0$. Thus, there exists $m > 0$ such that $\lambda(\mathcal{T} \cap [0,m]) > 0$. Now define $\mathcal{T}_m := \mathcal{T} \cap [0,m]$ and the test function

$$\psi(t,a) = \mathbb{1}_{\mathcal{T}_m}(t) \cdot \mathbb{1}_{\{a=\tilde{a}\}}(a).$$

The function is measurable in t and it holds that

$$\int_0^\infty \max_{a\in A} |\psi(t,a)|dt = \int_0^\infty \psi(t,\tilde{a})dt = \lambda(\mathcal{T}_m) < \infty.$$

Inserting ψ in the notion of convergence (3.45) and adding up over the states in $\tilde{S}$, we obtain

$$\lim_{n\to\infty} \sum_{i\in\tilde{S}} \int_0^\infty \int_A \psi(t,a)\rho_t^{i,n}(da)dt \;=\; \sum_{i\in\tilde{S}} \int_0^\infty \int_A \psi(t,a)\rho_t^{i}(da)dt. \tag{3.46}$$

The left-hand side can be bounded by

$$\begin{aligned}\lim_{n\to\infty}\sum_{i\in\tilde{S}}\int_0^\infty\int_A \psi(t,a)\rho_t^{i,n}(da)dt &= \lim_{n\to\infty}\sum_{i\in\tilde{S}}\int_{\mathcal{T}_m}\rho_t^{i,n}(\{\tilde{a}\})dt \\ &= \lim_{n\to\infty}\int_{\mathcal{T}_m}\underbrace{\sum_{i\in\tilde{S}}\rho_t^{i,n}(\{\tilde{a}\})}_{\leq c \quad a.e.}\,dt \;\leq\; c\cdot\lambda(\mathcal{T}_m), \qquad (3.47)\end{aligned}$$

where we use that $\rho^n \in \mathcal{R}_{RC}$ for every $n \in \mathbb{N}$. Conversely, for the right-hand side in (3.46) we obtain

$$\sum_{i\in\tilde{S}}\int_0^\infty\int_A \psi(t,a)\rho_t^{i}(da)dt = \sum_{i\in\tilde{S}}\int_{\mathcal{T}_m}\rho_t^{i}(\{\tilde{a}\})dt = \int_{\mathcal{T}_m}\underbrace{\sum_{i\in\tilde{S}}\rho_t^{i}(\{\tilde{a}\})}_{\geq c+\varepsilon}\,dt \geq (c+\varepsilon)\cdot\lambda(\mathcal{T}_m). \quad (3.48)$$

Now inequalities (3.47) and (3.48) create a contradiction to the identity (3.46) and therefore to $\rho \notin \mathcal{R}_{RC}$. This implies that the set $\mathcal{R}_{RC}$ is closed and thus, the statement follows. □

CHAPTER 4

Applications

In this chapter, we apply the previously derived theory to several applications. Theorem 3.17 establishes a method for solving mean-field MDPs with a large number of interacting agents, at least asymptotically: By letting the number of agents tend to infinity, we obtain the deterministic limit problem (F) resp. (F_T) for a finite time horizon. If we find an optimal control for the limit problem, it is asymptotically optimal for the corresponding N-agent model. For a finite time horizon, Theorem 3.24 even provides a statement on the rate of convergence, allowing us to quantify the approximation between the value functions $V^{N,T}$ and $V^{F,T}$. The benefit of the theory thus depends in particular on the solvability of the mentioned deterministic optimization problems. Since the subsequent applications primarily concern the finite-horizon case, recall the corresponding deterministic limit problem (F_T):

$$
\begin{aligned}
(F_T) \quad & \sup_{\hat{\pi}} \int_0^T e^{-\beta t} r(\mu_t, \hat{\pi}_t) dt + g(\mu_T) \\
& s.t.\ \mu_0 \in \mathbb{P}(S),\ \hat{\pi}_t^i \in \mathbb{P}(A), \\
& \mu_t(j) = \mu_0(j) + \int_0^t \sum_{i \in S} \mu_s(i) \int_A q(\{j\}|i, a, \mu_s) \hat{\pi}_s^i(da) ds, \quad t \in [0, T],\ j \in S.
\end{aligned}
$$

As will be shown in Section 4.1, if either the action space is finite, or relaxation of the control is unnecessary, which is the case in most applications, the problem can be reformulated as a standard problem in deterministic optimal control theory.

$$
\begin{aligned}
(OC) \qquad & \sup_{u=(u_t)} \int_0^T e^{-\beta t} \tilde{r}(x_t, u_t) dt + \tilde{g}(x_T), \\
& s.t.\ u_t = (u_t^1, u_t^2, \ldots, u_t^m) \in \mathbb{R}^m, \\
& \quad x_t = (x_t^1, x_t^2, \ldots, x_t^n) \in \mathbb{R}^n, \qquad x_0 \in \mathbb{R}^n, \\
& \quad x_t' = \tilde{f}(x_t, u_t), \qquad t \in [0, T].
\end{aligned} \tag{4.1}
$$

Such optimal control problems originate from the calculus of variations, whose foundations were established in the 18th century by Euler and Lagrange. Modern optimal control theory, which specifically addresses optimization problems like (OC), was substantially developed during the 1950s. In the process, two different philosophies for solving deterministic control problems emerged.

The first is the previously mentioned *dynamic programming*, developed by Richard Bellman. In this method, given $t \in [0, T]$ and $x(t) = \xi$, the so-called Bellman function is defined by

$$
B(\xi, t) = \max_{(u_t)} \int_t^T e^{-\beta s} \tilde{r}(x_s, u_s) ds + \tilde{g}(x_T).
$$

The core of dynamic programming is the following result: Under certain conditions, a state-control pair $(x_t, u_t)_{t \in [0,T]}$ is optimal if and only if the Bellman function $B(\xi, t)$ satisfies the Hamilton-Jacobi-Bellman (HJB) equation:

$$
-B_t(\xi, t) = \max_u \left\{ e^{-\beta t} \tilde{r}(\xi, u) + B_x(\xi, t) \tilde{f}(\xi, u) \right\}
$$

for all $(\xi, t) \in \mathbb{R}^n \times [0, T]$ for which (u^*) is continuous. For reference, see, e.g., Section 3.4 of Grass et al. (2008). Thus, determining the optimal control requires solving a partial differential equation on the entire state-time domain. Cecchin (2021), who considers a continuous-time mean-field MDP similar to ours, chooses the dynamic programming approach to solve his deterministic limit model. Under certain Lipschitz assumptions on the model parameters, he shows that the corresponding value function is characterized as the unique viscosity solution of the HJB.

To solve our limit problem (F_T), we rely on the second main approach to deterministic control problems, which is known in the literature as *Pontryagin's maximum principle* and originates from the work of Pontryagin et al. (1962). In contrast to dynamic programming, only ordinary differential equations need to be considered. Further, unlike the HJB, which must be satisfied for every point ξ in the state space, Pontryagin's maximum principle provides conditions that only need to be satisfied along a single, optimal trajectory, making it potentially more computationally efficient. A downside of the maximum principle is that it offers only necessary conditions for optimality, which, however, will prove to be adequate for our applications. Furthermore, in contrast to the approach of Cecchin (2021), our deterministic limit problem does not require Lipschitz conditions on the model parameters, but merely continuity.

The chapter is organized as follows: In Section 4.1, we present the central elements of optimal control theory and adapt Pontryagin's maximum principle to our framework.
The first application, discussed in Section 4.2, is a mean-field MDP that models the optimal spread of a computer virus in a mobile wireless network. The objective of the central controller, who assumes the role of the attacker, is to optimally adjust the lethality of the virus. In order to solve the problem, we explicitly derive the structure of the optimal control.
The second application, presented in Section 4.3, addresses the optimal control of a black bear population. We demonstrate how to model scenarios where the number of agents (bears) N is not constant but instead follows a birth-and-death process. Based on Pontryagin's maximum principle, we describe an approach for numerically solving the limit problem.
In Section 4.4, we examine a production site consisting of a large number of statistically equal machines. In this application, we consider an infinite time horizon. The limit problem is solved using a concept related to Pontryagin's maximum principle, known in the literature as the *most rapid approach path (MRAP)*. The key to the optimal control is to identify an optimal stationary state and to reach that state as fast as possible.
Finally, in Section 4.5, we consider a queueing network that includes a resource constraint, as discussed in Section 3.4.2. The purpose of the application is to illustrate that optimal feedback controls of the deterministic limit model with a fixed initial distribution μ_0 are not necessarily optimal when implemented in the N-agent problem.

4.1. Pontryagin's maximum principle

Recall that a control $(\hat{\pi}_t)_{t\geq 0}$ of the deterministic limit model (F) specifies a probability distribution on the action space for each state $i \in S$ and each point in time t, that is, $\hat{\pi}_t^i \in \mathbb{P}(A)$. In order to apply Pontryagin's maximum principle, we have to ensure that the control region of (F) is a subset of $\mathbb{R}^m$ as required in the standard control problem (OC), see (4.1). Therefore, throughout this chapter, we assume that one of the following assumptions is satisfied.
Assumptions:

i) The action space A is finite OR

ii) The action space is given by a hyperrectangle in $\mathbb{R}^m$, that is, $A = [a_{\min}, a_{\max}]$, where $a_{\min}, a_{\max} \in \mathbb{R}^m$. In addition, assume that the one-agent transition intensity q is affine in the action and that the one-agent reward r is concave in the action.

Note that assumption ii) implies that relaxation of the control is unnecessary; see Lemma 3.21. In this case, the control region of the deterministic limit problem is $A^{|S|}$. Otherwise, if relaxation is necessary, the control region is $\mathbb{P}(A)^{|S|}$. In the case of a finite action space, $\mathbb{P}(A)$ corresponds to the standard probability simplex in $\mathbb{R}^{|A|}$. If there is no need to specify

an action in a state $i \in S$ because neither the reward $r(i, \cdot, \mu)$ nor the transition intensity $q(\{j\} \mid i, \cdot, \mu)$ depend on the action, in this specific state we refrain from specifying a distribution $\hat{\pi}^i \in \mathbb{P}(A)$ resp. an action $a^i \in A$ and do not take the state $i \in S$ into account when determining a control. In the upcoming applications, this means that not all dimensions of $\mathbb{P}(A)^{|S|}$ resp. $A^{|S|}$ are needed to control the system. In the following, we denote the control region of the system resulting from the considerations above by A_{Det}.

Furthermore, in contrast to problem (3.42), we set the discount rate β to zero in the following, as it does not play a role in the finite-horizon applications to come. The (finite-horizon) deterministic limit problem (F_T) is then given by

$$
\begin{aligned}
(F_T) \quad & \sup_{a=(a_t)} \int_0^T r(\mu_t, a_t)dt + g(\mu_T) =: V_a^{F,T} \\
& s.t.\ \mu_0 \in \mathbb{P}(S),\ a_t \in A_{Det}, \\
& \mu_t(j) = \mu_0(j) + \int_0^t \sum_{i \in S} \mu_s(i) \int_A q(\{j\}|i, a, \mu_s) a_s^i(da)ds, \quad \forall\ t \in [0,T],\ j \in S. \quad (4.2)
\end{aligned}
$$

In the remainder of the section, we provide the basic definitions and results that are necessary to apply Pontryagin's maximum principle to problem (F_T). For this, we primarily rely on the approach presented in Chapters 3.2 and 3.3 in Grass et al. (2008). In what follows, the partial derivative of an arbitrary function f with respect to a (state) variable x is denoted by f_x, while its derivative with respect to time is denoted by f'.

Definition 4.1 (Optimal State-Control Trajectory)**.**

a) The state-control trajectory $(\mu_t, a_t)_{t \in [0,T]} \subset \mathbb{P}(S) \times A_{Det}$ is called *admissible* for (F_T) if (a_t) is piecewise continuous, for which the left- and right-hand limits exist and (μ_t) is continuous and piecewise continuously differentiable, which satisfies (4.2) for all points in time where (a_t) is continuous.

b) Let $(\mu_t^*, a_t^*)_{t \in [0,T]}$ be an admissible state-control trajectory for (F_T). If

$$V_{a^*}^{F,T} \geq V_a^{F,T}, \qquad \text{for all admissible controls } (a_t),$$

then (μ_t^*, a_t^*) is called an *optimal solution* for (F_T).

We now turn to the question of when we can guarantee the existence of an optimal solution, based on the parameters of the model. For notational convenience, as in Section 3.3.1, we write $(f_1, \ldots, f_{|S|})$ for the right-hand side of the state equation. For the following existence theorem, define the set

$$N(\mu, A_{Det}) := \left\{ \left(r(\mu, a) + \gamma, f_1(\mu, a), \ldots, f_{|S|}(\mu, a)\right) \mid \gamma \leq 0, a \in A_{Det} \right\}.$$

Theorem 4.2 (Existence of an optimal solution, Filippov-Cesari)**.**
Consider the optimal control problem (F_T). Suppose that an admissible pair (μ_t, a_t) exists and assume that the set $N(\mu, A_{Det})$ is convex for every $\mu \in \mathbb{P}(S)$. Then there exists an optimal pair (μ_t^*, a_t^*).

Proof. The Filippov-Cesari existence theorem states four assumptions that together guarantee the existence of an optimal pair (μ_t^*, a_t^*):

i) There exists an admissible pair (μ_t, a_t),

ii) $N(\mu, A_{Det})$ is convex for every $\mu \in \mathbb{P}(S)$,

iii) A_{Det} is closed and bounded,

iv) There exists a number b such that $\|\mu_t\|_{TV} \leq b$ for all $t \in [0, T]$ and all admissible pairs (μ_t, a_t).

Since A is assumed to be a compact Borel space, Lemma A.4 implies assumption iii). Furthermore, μ_t is a distribution on S for every point in time; therefore, we have $\|\mu_t\|_{TV} \leq 1$ for every $t \in [0, T]$. Thus, to show existence, it remains to check assumptions i) and ii), which together imply the statement. The cited version of Filippov-Cesari is Theorem II.8 in Seierstad (1987), see also Theorem 5.1.ii in Cesari (1983). □

The following theorem marks the central result in this section and provides necessary conditions for an optimal control.

Theorem 4.3 (Pontryagin's Maximum Principle)**.**
Let $(\mu_t^*, a_t^*)_{t\in[0,T]}$ be an optimal solution to (F_T). Then there exists a continuous and piecewise continuously differentiable function $p_t = (p_1(t), \ldots, p_{|S|}(t))$ with $p_t \in \mathbb{R}^{|S|}$ satisfying for all $t \in [0, T]$

$$H(\mu_t^*, a_t^*, p_t) = \max_{a \in A_{Det}} H(\mu_t^*, a, p_t) \tag{4.3}$$

and at every point t where (a_t^*) is continuous

$$p_t' = -H_\mu(\mu_t^*, a_t^*, p_t).$$

Furthermore, the transversality condition

$$p_T = g_\mu(\mu_T^*)$$

holds, where

$$\begin{aligned} H(\mu_t, a_t, p_t) &= r(\mu_t, a_t) + \sum_{j\in S} p_j(t)\Big(\sum_{i\in S} \mu_t(i) \int_A q(\{j\}|i, a, \mu_t) a_t^i(da)\Big) \\ &= r(\mu_t, a_t) + \sum_{j=1}^{|S|} p_j(t) f_j(\mu_t, a_t) \end{aligned}$$

is called the Hamiltonian of (F_T), while p is called the adjoint function.

For a proof of Pontryagin's maximum principle in a general framework, see Section 2.4 in Seierstad (1987) or Section 3.4.2 in Grass et al. (2008).

Remark 4.4. The state equation is autonomous in the sense that the right-hand side does not explicitly depend on t. Since the reward function $r(\mu, a)$ does not explicitly depend on t either, we call the whole optimization problem (F_T) autonomous. For such an autonomous control problem with an admissible state-control trajectory $(\mu_t, a_t)_{t\in[0,T]}$ satisfying the necessary conditions of Pontryagin's maximum principle, it holds that the Hamiltonian is constant when evaluated on $(\mu_t, a_t)_{t\in[0,T]}$, that is

$$H(\mu_t, a_t, p_t) \equiv C, \qquad \forall\, t \in [0, T],$$

for a constant $C \in \mathbb{R}$, where p is the corresponding adjoint function. For reference, see Section 5.3, Equation (5.3-40) in Kirk (1970).

Remark 4.5. A control $(a_t)_{t\in[0,T]}$ is referred to as *bang-bang* if for every $t \in [0, T]$, the control takes values only on the boundary of the control region. If, for example, the control region is given by $A_{Det} = [a_{min}, a_{max}]$, we have $a_t \in \{a_{min}, a_{max}\}$. Controls of a bang-bang type frequently occur if the Hamiltonian depends linearly on the control variable. For a reference, see, e.g., Chapter 17 in Lenhart and Workman (2007). In this linear one-dimensional case, the maximization condition (4.3) reveals the following representation of the optimal control:

$$a_t^* := \begin{cases} a_{min}, & \text{if} \quad H_a(\mu_t^*, a, p(t)) < 0, \\ ?, & \text{if} \quad H_a(\mu_t^*, a, p(t)) = 0, \\ a_{max}, & \text{if} \quad H_a(\mu_t^*, a, p(t)) > 0. \end{cases}$$

In the literature, $\varphi(t) := H_a(\mu_t^*, a_t^*, p(t))$ is referred to as the *switching function*. If $\varphi \equiv 0$ on an interval of nonzero length, the maximization condition (4.3) does not provide additional information about the optimal control. This is known as *singular* optimal control. An application in which the optimal control (a_t^*) includes a singular segment is presented in Section 4.4.

4.2. Spreading malware

This example is based on the deterministic control model considered in Khouzani et al. (2012) and deals with the propagation of a virus in a mobile wireless network. The network consists of a large number of mobile devices that are vulnerable to a computer virus controlled by an attacker. The attacker acts as the central controller, aiming to maximize the damage dealt to the network by determining the lethality of the virus. This scenario is embedded in an epidemiological SIRD model (S = susceptible, I = infected, R = recovered, D = dead), which is an extension of the classical SIR model (R = removed) introduced

by Kermack and McKendrick (1927). SIR models as well as compartmental models such as the SIRD model gained increased attention in the literature during the COVID-19 pandemic; see, e.g., Calafiore et al. (2020), Fernandez-Villaverde and Jones (2022), among many others.

Khouzani et al. (2012) proposed a deterministic optimization problem and derived, using Pontryagin's maximum principle, that the optimal control is bang-bang. Essentially, the idea is to initially set the lethality to zero, allowing the virus to spread through the network. After a critical point in time, the lethality is set to its maximum until the end of the time horizon, in order to cause as much damage as possible.

Because of the multi-device nature of the problem, it can be interpreted as the mean-field limit model of an N-agent Markov decision problem; see, e.g., Gast et al. (2012). There, the authors consider a discrete-time N-agent model that they transform into a continuous-time process by the affine interpolation of the agents' discrete-time empirical distribution, where time is rescaled by an intensity function I. Letting the number N of devices tend to infinity, they obtain a deterministic continuous-time optimization problem, which is a variant of the one considered by Khouzani et al. (2012). It turns out that the optimal bang-bang control derived in Khouzani et al. (2012) is asymptotically optimal for the discrete-time N-agent model of Gast et al. (2012).

We derive similar relationships between our model and the deterministic model considered by Khouzani et al. (2012). We begin by describing the basic N-device problem in our framework.

4.2.1. Description of the N-agent model

Suppose that there are N devices in the network. A device can be in one of the following states:

- *Susceptible (S)*: The device is not yet contaminated, but susceptible to infection.
- *Infected (I)*: The device is contaminated by the virus.
- *Dead (D)*: The virus has destroyed the software of the device.
- *Recovered (R)*: The device has a security patch that makes it immune to the virus.

The states D and R are absorbing. The joint process $\mu_t^N = (S_t^N, I_t^N, D_t^N, R_t^N)$ is a controlled continuous-time Markov chain where $S_t^N,\ I_t^N,\ D_t^N,\ R_t^N$ represent the fractions of devices in the states S, I, D, R, respectively. In this model, we have $S_t^N + I_t^N + D_t^N + R_t^N = 1$ and $S_t^N, I_t^N, D_t^N, R_t^N \geq 0$. The initial state of the network is $\mu_0^N = (S_0^N, I_0^N, D_0^N, R_0^N) = (1 - I_0^N, I_0^N, 0, 0)$ with $0 < I_0^N < 1$ such that $(I_0^N)_{N\in\mathbb{N}} \Rightarrow I_0 \in (0, 1)$.
The action of the attacker is to determine the lethality of the virus for infected devices. At each point in time, the attacker chooses the rate $a_t \in [0, \bar{a}]$, at which the infected devices are destroyed. Thus, the action space is given by $A = [0, \bar{a}]$, where a lethality rate of zero

means that infected devices are not destroyed by the virus at that time.
The transition rates of one device are as follows: A susceptible device gets infected at rate $\lambda_{SI} I_t^N$, where $\lambda_{SI} > 0$. The rate is proportional to the fraction of infected devices, which reflects that the risk of infection increases with the number of infected devices. This constitutes an interaction between an individual agent and the empirical distribution of the system. In addition, a susceptible device recovers at rate $\lambda_{SR} > 0$, which is the rate at which the security patch is distributed. An infected device is destroyed by the virus at rate $a_t \in [0, \bar{a}]$ chosen by the attacker and recovers at rate $\lambda_{IR} > 0$. The rates are shown in the following figure:

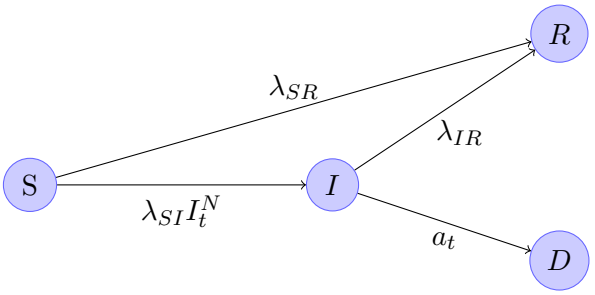

Figure 4.1.: Transition intensities of a device between its possible states.

The intensities of one device at time t can be summarized as

$$q(\{I\}|S, \cdot, \mu_t^N) = \lambda_{SI} I_t^N, \quad q(\{R\}|S, \cdot, \mu_t^N) = \lambda_{SR},$$
$$q(\{D\}|I, a_t, \mu_t^N) = a_t, \quad q(\{R\}|I, \cdot, \mu_t^N) = \lambda_{IR}.$$

Thus, the diagonal elements of the intensity matrix are given by

$$q(\{S\}|S, \cdot, \mu_t^N) = -\lambda_{SI} I_t^N - \lambda_{SR}, \qquad q(\{I\}|I, a_t, \mu_t^N) = -a_t - \lambda_{IR},$$

$$q(\{D\}|D, \cdot, \mu_t^N) = q(\{R\}|R, \cdot, \mu_t^N) = 0.$$

All other intensities are zero. Note that (Q1)-(Q5) and, in particular, (Q4') are satisfied. The aim of the attacker is to maximize its expected reward over the finite time interval $[0, T]$. Every infected device generates a reward for the attacker, e.g., by collecting information and transferring data. More precisely, the reward function of a device is given by

$$r(i, \cdot, \mu_t^N) = \frac{1}{T} \mathbb{1}_{\{i=I\}} I_t^N.$$

Note that the reward is proportional to the fraction of infected devices, which means that the collected information is more valuable when more devices are infected. Conditions (R1) and (R2) are satisfied. Furthermore, we allow for a terminal reward for each destroyed device at the end of the planning horizon T, i.e.,

$$g(\mu_T^N) = D_T^N.$$

Since the intensities are affine in a and the (terminal) reward does not depend on the action, there is no need for a relaxed control; see Lemma 3.21. Consequently, the control region of the deterministic limit model is given by $A_{Det} = [0, \bar{a}]$. Note that, in particular, the reward rate and the terminal reward satisfy conditions (R1') and (R1").

4.2.2. Deterministic limit model

By letting the number of devices tend to infinity, we obtain the finite-horizon deterministic limit optimization problem (F_T); see (3.42). In the following, we denote the limit state process by $\mu_t = (S_t, I_t, D_t, R_t)$. Since the state process forms a distribution for each point in time $t \in [0, T]$, we use the property $R_t = 1 - S_t - I_t - D_t$ and therefore omit the state R in the upcoming considerations.

$$
\begin{aligned}
(SM) \qquad & \sup_a D_T + \frac{1}{T}\int_0^T I_t^2 dt, \\
& s.t.\ a_t \in [0, \bar{a}],\ I_0 \in (0,1),\ \text{and for all } t \in [0,T] \\
& \qquad S_t = 1 - I_0 + \int_0^t -\lambda_{SI} I_s S_s - \lambda_{SR} S_s ds, && (4.4) \\
& \qquad I_t = I_0 + \int_0^t \lambda_{SI} I_s S_s - \lambda_{IR} I_s - a_s I_s ds, && (4.5) \\
& \qquad D_t = \int_0^t a_s I_s ds. && (4.6)
\end{aligned}
$$

We start the analysis of the problem by verifying the assumptions of the Filippov-Cesari existence theorem.

Theorem 4.6. *The optimal control problem (SM) has an optimal pair $(\mu_t^*, a_t^*)_{t\in[0,T]}$.*

Proof. For $a_t \equiv 0$, the corresponding state-control trajectory (μ_t, a_t) is admissible, which implies assumption i) of the Filippov-Cesari existence theorem 4.2. It remains to check the convexity of the set

$$
N(\mu, [0,\bar{a}]) = \left\{ \left(\frac{1}{T} I^2 + \gamma,\ -\lambda_{SI} IS - \lambda_{SR} S,\ \lambda_{SI} IS - \lambda_{IR} I - aI,\ aI \right) \,\middle|\, \gamma \le 0,\ a \in [0,\bar{a}] \right\}
$$

for every $\mu \in \mathbb{P}(S)$. Now, let $n_1, n_2 \in N(\mu, [0,\bar{a}])$ with $\gamma_1, \gamma_2 \le 0$ and $a_1, a_2 \in [0,\bar{a}]$. For $\lambda \in [0,1]$ we obtain that $n_3 := \lambda n_1 + (1-\lambda) n_2$ is again in $N(\mu, [0,\bar{a}])$ with $\gamma_3 = \lambda\gamma_1 + (1-\lambda)\gamma_2 \le 0$ and $a_3 = \lambda a_1 + (1-\lambda) a_2 \in [0,\bar{a}]$, which implies the convexity of $N(\mu, [0,\bar{a}])$ for every $\mu \in \mathbb{P}(S)$ and thus the statement. □

The deterministic optimal control problem (SM) was studied by Khouzani et al. (2012). They show that the optimal control has the following structure: Set the lethality to zero at the beginning and let the virus spread in the network to profit from the infected devices until a critical point in time $t^* \in [0,T]$. From then on, set the lethality to the maximum $\bar{a}$ until the end of the planning horizon to destroy as many devices as possible to maximize the terminal reward. Thus, the structure of the optimal control is bang-bang. For

completeness, we provide a detailed proof of the bang-bang structure using Pontryagin's maximum principle.

Theorem 4.7 (Optimal control in the Spreading Malware Model (SM))**.**
There exists a point in time $t^* \in [0,T)$ such that the optimal control is given by

$$a_t^* = \begin{cases} 0, & t \in [0,t^*), \\ \bar{a}, & t \in (t^*,T]. \end{cases}$$

Proof. We start with the observation that the fraction of infected devices is positive throughout the time horizon. Even in the case $a_t \equiv \bar{a}$ for all $t \in [0,T]$, the fraction of infected devices decreases at most exponentially at rate $\lambda_{IR} + \bar{a}$, but never reaches zero, i.e., $I_t \geq I_0 e^{-(\lambda_{IR}+\bar{a})t} > 0$.

The Hamiltonian function for (SM) is given by

$$\begin{aligned} H(S,I,D,a,p_S,p_I,p_D) = \frac{1}{T}I^2 &+ p_S(-\lambda_{SI}IS - \lambda_{SR}S) \\ &+ p_I(\lambda_{SI}IS - \lambda_{IR}I - aI) \\ &+ p_D(aI) \end{aligned}$$

where p_S, p_I and p_D are the adjoint functions to the corresponding differential equations. Pontryagin's maximum principle yields the following necessary conditions for an optimal control $(a_t^*)_{t\in[0,T]}$:

Necessary Conditions:
Let $(\mu_t^*, a_t^*)_{t\in[0,T]}$ be an optimal solution of (SM). Then there exist continuous and piecewise continuously differentiable adjoint functions $p_S, p_I, p_D : [0,T] \to \mathbb{R}$ satisfying for all $t \in [0,T]$

$$H(\mu_t^*, a_t^*, p_S(t), p_I(t), p_D(t)) = \max_{a\in[0,\bar{a}]} H(\mu_t^*, a, p_S(t), p_I(t), p_D(t)), \tag{4.7}$$

and at every point t where (a_t^*) is continuous

$$\begin{aligned} p_S'(t) &= -H_S(\mu_t^*, a_t^*, p_S(t), p_I(t), p_D(t)) \\ &= \lambda_{SI} I_t^* (p_S(t) - p_I(t)) + \lambda_{SR}\, p_S(t) \end{aligned} \tag{4.8}$$

$$\begin{aligned} p_I'(t) &= -H_I(\mu_t^*, a_t^*, p_S(t), p_I(t), p_D(t)) \\ &= -\frac{2}{T} I_t^* + \lambda_{SI} S_t^* (p_S(t) - p_I(t)) + a_t^* (p_I(t) - p_D(t)) + \lambda_{IR}\, p_I(t) \end{aligned} \tag{4.9}$$

$$p_D'(t) = -H_D(\mu_t^*, a_t^*, p_S(t), p_I(t), p_D(t)) = 0. \tag{4.10}$$

Furthermore, the transversality condition holds:

$$p_S(T) = g_S(\mu_T^*) = 0, \qquad p_I(T) = g_I(\mu_T^*) = 0, \qquad p_D(T) = g_D(\mu_T^*) = 1. \tag{4.11}$$

The transversality condition, together with (4.10) directly implies that $p_D \equiv 1$. In order to investigate the maximization condition (4.7), observe that the Hamiltonian is linear in a, where the term depending on a is given by

$$I_t(p_D(t) - p_I(t)) = I_t(1 - p_I(t)). \tag{4.12}$$

Condition (4.7) together with (4.12) gives the following property of the optimal solution:

$$a_t^* = \begin{cases} 0, & \text{if} \quad I_t(1 - p_I(t)) < 0, \\ \bar{a}, & \text{if} \quad I_t(1 - p_I(t)) > 0. \end{cases} \tag{4.13}$$

With the transversality condition (4.11), we obtain $a_T^* = \bar{a}$, which means that at the end of the planning horizon, it is optimal to set the lethality rate to the maximum. Define the switching function $\varphi(t) := I_t(1 - p_I(t))$. Note that φ is continuous and piecewise continuously differentiable with $\varphi(T) = I_T > 0$. Therefore, there exists an interval (t, T) where $a_t = \bar{a}$.

In order to prove the statement, we show that $\varphi(t)$ has at most one sign change at a unique point in time $t^* \in [0, T)$ over the planning horizon and that there cannot exist other points in time where $\varphi(t) = 0$. We have the following representation of the derivative in time of φ, wherever $(a_t^*)_{t\in[0,T]}$ is continuous:

$$\begin{aligned}
\varphi' &= I'(1 - p_I) - I p_I' \\
&= \frac{I'\varphi}{I} - I\Big(- \frac{2}{T} I + \lambda_{SI} S(p_S - p_I) + a(p_I - 1) + \lambda_{IR}\, p_I \Big) \\
&= \frac{I'\varphi}{I} + \frac{2}{T} I^2 + \lambda_{SI} IS(p_I - p_S) + aI(1 - p_I) - \lambda_{IR}\, p_I I \\
&= \frac{I'\varphi}{I} + \frac{2}{T} I^2 + \lambda_{SI} IS(p_I - p_S) + aI(1 - p_I) - \lambda_{IR}\, p_I I \\
&\quad + \Big(H - \frac{1}{T} I^2 - \lambda_{SI} IS(p_I - p_S) + \lambda_{SR}\, p_S S + \lambda_{IR}\, p_I I - aI(1 - p_I) \Big) \\
&= \frac{I'\varphi}{I} + H + \lambda_{SR}\, p_S S + \frac{1}{T} I^2.
\end{aligned} \tag{4.14}$$

We now state four facts that will be proved later. For all $t \in (0, T)$, we have

i) $H \equiv c$ for a constant $c > 0$ when evaluated for an admissible pair $(\mu_t, a_t)_{t\in[0,T]}$ that satisfies the necessary conditions.

ii) $p_S(t) > 0$.

In addition, we have two basic analytical results. Let $f : [0, T] \to \mathbb{R}$ be continuous and piecewise continuously differentiable.

iii) Suppose that $f(T) > 0$ and that $t_0 := \sup\{t \in [0,T] \mid f(t) = 0\}$ exists. Then we have $f'(t_0^+) \geq 0$.

iv) Let $t_0, t_1 \in [0,T]$ be two consecutive zeros of f, i.e., $f(t_0) = f(t_1) = 0$ and $f(t) \neq 0$ for $t \in (t_0, t_1)$. Then if $f'(t_0^+) \neq 0$ and $f'(t_1^-) \neq 0$, it holds that $f'(t_0^+)$ and $f'(t_1^-)$ must have opposite signs.

First, we conclude that φ cannot be equal to zero over an interval of nonzero length. To see this, assume that such an interval with $\varphi = 0$ exists. Since $(a_t^*)_{t\in[0,T]}$ is piecewise continuous, there exists a subinterval with $\varphi = 0$ and $\varphi' = 0$ where $(a_t^*)_{t\in[0,T]}$ is continuous, and thus

$$0 = \varphi' \overset{(4.14)}{=} \frac{I'\varphi}{I} + H + \lambda_{SR}\, p_S S + \frac{1}{T} I^2 > 0,$$

a contradiction. The positivity of the right-hand side follows because $\varphi = 0$, $H > 0$ by i) and $p_S > 0$ by ii).

Let τ be a point in time where $\varphi(\tau) = 0$. Having justified that φ is not equal to zero over an interval of nonzero length, we know that in a neighborhood of τ it holds that $\varphi(t) \neq 0$, and $(a_t^*)_{t\in[0,T]}$ is continuous in the left and right neighborhood of τ, respectively. We get, using the continuity of φ, $\varphi(\tau) = 0$, i) and ii),

$$\varphi'(\tau^-) = \varphi'(\tau^+) = H + \lambda_{SR}\, p_S S + \frac{1}{T} I^2 > 0. \tag{4.15}$$

Now, iv) together with (4.15) implies that there exists at most one point in time t^* with $\varphi = 0$ in the interval $[0,T)$, and φ exhibits a sign change at t^*. In light of (4.13), if φ indeed has a zero t^* in $[0,T)$, the continuity of φ together with $\varphi(T) > 0$ implies $a_t^* = \bar{a}$ on $(t^*, T]$ and $a_t^* = 0$ on $[0, t^*)$. If φ has no zero in $[0,T)$ and is therefore strictly positive on $[0,T]$, the optimal control is given by $a_t^* \equiv \bar{a}$, which is covered by setting $t^* = 0$ in the formulation of the theorem. Hence, the statement follows. It remains to check the facts i)-iv).

Proof of i):
From Remark 4.4, it is immediately clear that the Hamiltonian is constant. The positivity of the constant can be verified using the transversality condition (4.11):

$$H(\mu_T, a_T, p(T)) = \frac{1}{T} I_T^2 + a_T I_T > 0.$$

Proof of ii):
In fact, we show $p_S(t) > 0$ and $p_I(t) - p_S(t) > 0$ for all $t \in (0,T)$. We start by investigating the behavior at the end of the planning horizon. We have

$$\begin{aligned}
&p_S'(T) = 0,\\
&p_S''(T) = -\lambda_{SI} I_T \Big(-\frac{2}{T} I_T - a_T \Big) > 0,\\
&p_I'(T) = p_I'(T) - p_S'(T) = -\frac{2}{T} I_T - a_T < 0.
\end{aligned}$$

Recall that since $a_t = \bar{a}$ on an interval before the end of the planning horizon, the derivatives p_S' and p_I' are continuous. Thus, there exists an interval $(T-\varepsilon, T)$ such that $p_S(t) > 0$ and $p_I(t) - p_S(t) > 0$ on $(T-\varepsilon, T)$. Define

$$t_0 := \inf\{t \in [0,T] \mid p_S(s) > 0,\ p_I(s) - p_S(s) > 0 \quad \forall\ s \in (t,T)\}.$$

Suppose $t_0 > 0$. We distinguish three cases:

1. $p_S(t_0) = 0$ and $p_I(t_0) - p_S(t_0) > 0$:
   Then we have
   $$p_S'(t_0^+) = -\lambda_{SI} I_{t_0} p_I(t_0) < 0,$$
   which is a contradiction to fact iii). Therefore, this case is not possible.

2. $p_S(t_0) > 0$ and $p_I(t_0) - p_S(t_0) = 0$:
   It is clear that $p_S(t_0) = p_I(t_0)$. We obtain
   $$\begin{aligned} p_I'(t_0^+) - p_S'(t_0^+) &= -\frac{2}{T} I_{t_0} - a_{t_0}(1 - p_I(t_0)) + \lambda_{IR}\ p_I(t_0) - \lambda_{SR}\ p_S(t_0) \\ &= -\frac{2}{T} I_{t_0} - a_{t_0}(1 - p_I(t_0)) + \lambda_{IR}\ p_I(t_0) - \lambda_{SR}\ p_S(t_0) \\ &\quad - \frac{H}{I_{t_0}} + \frac{1}{T} I_{t_0} - \lambda_{SR}\ p_S(t_0)\frac{S_{t_0}}{I_{t_0}} - \lambda_{IR}\ p_I(t_0) + a_{t_0}(1 - p_I(t_0)) \\ &= -\frac{1}{T} I_{t_0} - \lambda_{SR}\ p_S(t_0) - \frac{H}{I_{t_0}} - \lambda_{SR}\ p_S(t_0)\frac{S_{t_0}}{I_{t_0}} < 0. \end{aligned}$$
   Again, this is a contradiction to fact iii). Thus, this case is not possible as well.

3. $p_S(t_0) = 0$ and $p_I(t_0) - p_S(t_0) = 0$:
   Here the second case gets reduced to
   $$p_I'(t_0^+) - p_S'(t_0^+) = -\frac{1}{T} I_{t_0} - \frac{H}{I_{t_0}} < 0,$$
   which is again not possible.

Since none of the three cases can occur, we conclude that $t_0 > 0$ is impossible, which means that $p_S(t) > 0$ and $p_I(t) - p_S(t) > 0$ for all $t \in (0,T)$.

Proof of iii):
Since we assumed $f(T) > 0$, and since t_0 is the latest time at which the continuous function f is zero, it is clear that $f(t) > 0$ on $(t_0, T]$. Because f is piecewise continuously differentiable, the right-hand derivative $f'(t_0^+)$ exists. We obtain

$$f'(t_0^+) = \lim_{h \to 0^+} \frac{f(t_0+h) - f(t_0)}{h} = \lim_{h \to 0^+} \underbrace{\frac{f(t_0+h)}{h}}_{\geq 0} \geq 0.$$

Proof of iv):
Since f is continuous and t_0 and t_1 are two consecutive zeros of f, we have $f(t) > 0$ or $f(t) < 0$ on (t_0, t_1). First, assume that $f(t) > 0$ on (t_0, t_1). We have

$$f'(t_0^+) = \lim_{h \to 0^+} \frac{f(t_0 + h) - f(t_0)}{h} = \lim_{h \to 0^+} \underbrace{\frac{f(t_0 + h)}{h}}_{>0} \geq 0.$$

The case $f'(t_0^+) = 0$ is excluded in the statement, therefore we obtain $f'(t_0^+) > 0$. On the other hand, we have

$$f'(t_1^-) = \lim_{h \to 0^+} \frac{f(t_1) - f(t_1 - h)}{h} = \lim_{h \to 0^+} \underbrace{\frac{-f(t_1 - h)}{h}}_{<0} \leq 0.$$

Again, by assumption $f'(t_1^-) \neq 0$ and thus $f'(t_1^-) < 0$. The case where $f(t) < 0$ on (t_0, t_1) follows analogously, with all signs reversed.

□

Remark 4.8. As discussed, Pontryagin's maximum principle provides only necessary conditions for an optimal pair. So why are the necessary conditions sufficient here for optimality? Theorem 4.6 ensures the existence of an optimal pair. Furthermore, the bang-bang control $(a_t^*)_{t \in [0,T]}$ of Theorem 4.7 is optimal since it is the unique candidate to satisfy the maximization condition (4.7). A control deviating from the switching property (4.13) would result in a violation of the maximum property (4.7) and thus cannot be optimal.

4.2.3. ILLUSTRATION

Having shown that the structure of the optimal control is bang-bang, as a next step, we point out how to determine the critical point in time for a specific choice of parameters. For this, we view the value function as a function of the critical point in time:

$$V^{SM}(t) := D_T + \frac{1}{T} \underbrace{\int_0^t I_s^2 ds}_{\text{Regime } a=0} + \frac{1}{T} \underbrace{\int_t^T I_s^2 ds}_{\text{Regime } a=\bar{a}},$$

where the system of differential equations for the state process (4.4)-(4.6) is evaluated under the bang-bang control $(a_s)_{s \in [0,T]}$ with $a_s = \bar{a}\ \mathbb{1}_{\{s>t\}}$. The function $V^{SM}(t)$ is then maximized (numerically). For illustration, we choose the intensities $\lambda_{SI} = 0.6$, $\lambda_{SR} = \lambda_{IR} = 0.1$, the maximum lethality rate $\bar{a} = 1$, the initial proportion of infected devices $I_0 = 0.1$, and a time horizon of $T = 10$. Figure 4.2 shows the graph of $V^{SM}(t)$ for the specified set of parameters. We conclude that in this case, the critical point in time is approximately $t^* = 6.26$ with an optimal value of $V^{SM}(t^*) = 0.3832$. The terminal fraction of dead devices is $D_{10} = 0.3353$, as shown in Figure 4.3. This means that the

impact of the terminal value D_{10} on the value $V^{SM}(t^*)$ is significantly greater than that of the continuous reward associated with the proportion of infected devices.

A simulation of the system with N devices can be found in Figure 4.3. Shown are the trajectories of the optimal state distribution in (SM) and the simulated paths of $N = 1000$ devices for the same set of parameters as in Figure 4.2. The simulated paths are almost indistinguishable from the deterministic trajectories. Recall that Theorem 3.17 ensures that the control $(a_t)_{t\in[0,T]}$ with $a_t = \bar{a}\ \mathbb{1}_{\{t>6.26\}}$ is asymptotically optimal for the system with N devices. In particular, since we consider a finite time horizon and the conditions (Q4'), (R1'), and (R1") are satisfied, Theorem 3.24 implies that the convergence rate of the value functions is of order $1/\sqrt{N}$.

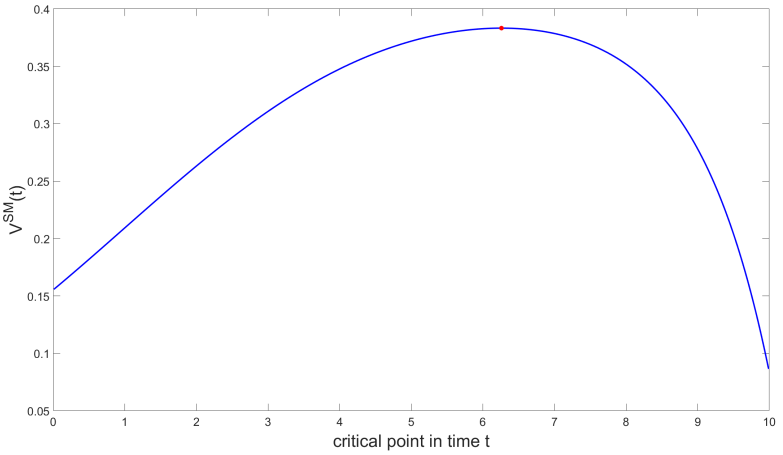

Figure 4.2.: Graph of $V^{SM}(t)$ for the specified set of parameters.

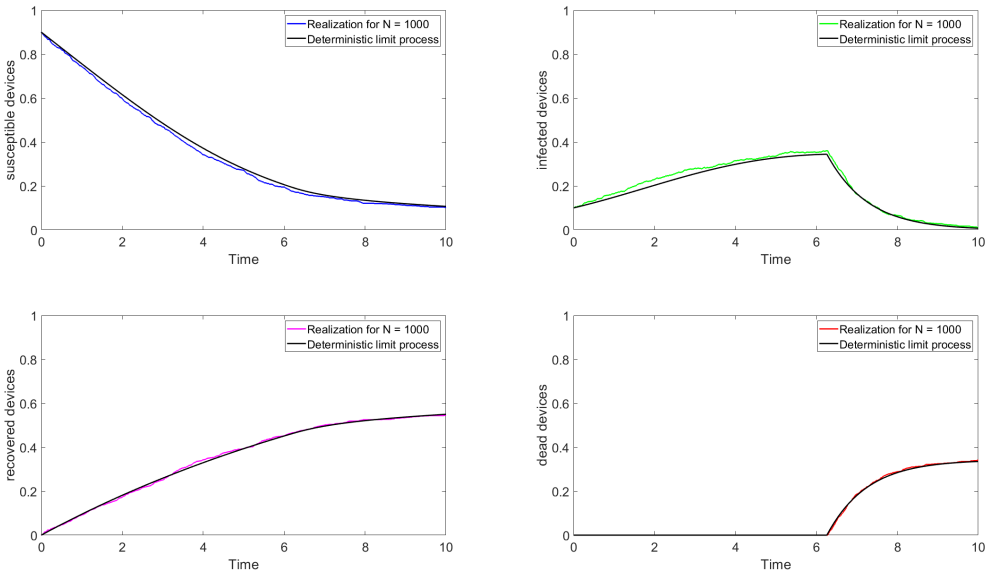

Figure 4.3.: State trajectories for $N = 1000$ devices under optimal control.

4.3. Black bears

Until now, we have considered problems where the number of agents N is fixed in the sense that no agent enters or leaves the system during the optimization period. But we can also apply our theory to control a population following a birth-and-death process, i.e., agents are born or die during the time horizon. The central controller can vary the birth rate and/or death rate in order to bring the population into optimal shape. To illustrate this, we study the control of a metapopulation harvesting model for black bears, introduced by Gross et al. (2005). They present a deterministic optimal control problem of the following form: Three state variables represent the black bear populations in three different habitats (national park, forest, urban area). The control variables are the hunting rates in the national park and in the forest. The controller's aim is to minimize the number of bears in the urban area through (costly) preventive hunting in the park and the forest. To derive an optimal solution, they formulate the optimality system given by the necessary conditions of Pontryagin's minimum principle, which they solve numerically.

The central question is whether we can interpret this deterministic optimal control problem as a limit model of an N-agent mean-field MDP. We present an illustrative construction that relies on the introduction of an additional state in which the "inactive" agents are collected.

4.3.1. Description of the N-agent model

Consider a population of black bears, where the population is divided into three different habitats: A national park (P), the surrounding common and unpreserved forest (F) and urban areas populated by humans, which we refer to as the *outside* (O). Naturally, the aim of the central controller is to concentrate the population in the national park and to minimize the number of bears in urban areas, i.e., the outside. We call the collection of national park, forest, and outside the *ecosystem*.

To cast the birth-and-death process into our classic N-agent framework, we assume that the number of bears in the ecosystem is limited over the planning horizon. The true number of bears in the ecosystem is less than the number N of potentially available bears in the model. Typically, the number of bears alive in the ecosystem will be significantly lower than N. But how do we model the birth and death of bears if the number of potentially available bears in the system should remain constant N?

To address this problem, we introduce an additional state which we call *transcendence* ($\mathcal{T}$). The natural choice of symbol for the transcendence would be T. Unfortunately, T is already reserved for the time horizon. Instead, we use the symbol $\mathcal{T}$. In the transcendent state, we collect all bears (out of the total N) that are currently not "alive" and waiting to be born. When a bear is born, it switches from the transcendent state to the ecosystem. And if a bear dies, it leaves the ecosystem and moves to the transcendent state. Note that it is perfectly possible here that one bear may undergo multiple cycles of life and death.

Let us now mathematically embed the birth and death process in the N-agent model. The obvious choice for the state space is $S = \{P, F, O, \mathcal{T}\}$. The joint process $\mu_t^N = (P_t^N, F_t^N, O_t^N, \mathcal{T}_t^N)$ is a controlled continuous-time Markov chain where the components $P_t^N,\ F_t^N,\ O_t^N,\ \mathcal{T}_t^N$ represent the fraction of black bears in states $P, F, O, \mathcal{T}$, respectively. The action of the central controller is to determine the bear hunting rates in the national park and in the forest, respectively. The action space is given by $A := \{(a_p, a_f) \mid a_p \in [0,1], a_f \in [0,1]\}$, where $a_p \in [0,1]$ denotes the hunting rate in the park, and $a_f \in [0,1]$ the hunting rate in the forest. The transition intensities for an individual bear can be summarized as follows:

i) Birth intensity:

$$q(\{P\}|\mathcal{T}, \cdot, \mu_t^N) = g \cdot \frac{P_t^N}{\mathcal{T}_t^N} \cdot \mathbb{1}_{\{\mathcal{T}_t^N \geq \varepsilon\}} + g \cdot \frac{P_t^N}{\varepsilon} \cdot \mathbb{1}_{\{\mathcal{T}_t^N < \varepsilon\}},$$
$$q(\{F\}|\mathcal{T}, \cdot, \mu_t^N) = g \cdot \frac{F_t^N}{\mathcal{T}_t^N} \cdot \mathbb{1}_{\{\mathcal{T}_t^N \geq \varepsilon\}} + g \cdot \frac{F_t^N}{\varepsilon} \cdot \mathbb{1}_{\{\mathcal{T}_t^N < \varepsilon\}}.$$

Here, $g > 0$ denotes the population growth rate, and $\varepsilon > 0$ is fixed and small. In order to guarantee a birth rate independent of the number of transcendent bears for the ecosystem in the deterministic limit model, the birth intensity for one individual bear in the transcendent state, at least as long as $\mathcal{T}_t^N$ is greater than or equal to ε. For lower values of $\mathcal{T}_t^N$, it is replaced by ε to prevent the intensity from exploding. Further, the birth intensity of the park/forest population is proportional to the number of bears in the park/forest, i.e., the more bears in the area, the more baby bears.

Reproduction only occurs in the national park and the forest. Therefore, the birth rate in the outside area is zero:

$$q(\{O\}|\mathcal{T}, \cdot, \mu_t^N) = 0.$$

ii) Bear hunting:

$$q(\{\mathcal{T}\}|P, a_t, \mu_t^N) = a_p(t), \qquad q(\{\mathcal{T}\}|F, a_t, \mu_t^N) = a_f(t).$$

The central controller determines the hunting rate for both the national park and the forest, respectively. We abbreviate $a_t = (a_p(t), a_f(t))$. In the outside, no hunting occurs:

$$q(\{\mathcal{T}\}|O, a_t, \mu_t^N) = 0.$$

iii) Emigration from the park:

$$q(\{F\}|P, \cdot, \mu_t^N) = \frac{m_p \cdot g}{K} \cdot P_t^N \cdot \left(1 - \frac{F_t^N}{K}\right) \cdot \mathbb{1}_{\{F_t^N \leq K\}}.$$

Here, $m_p \in [0,1]$ is the fraction of the park boundary adjacent to the forest. The parameter $K \in (0, 0.5)$ denotes the maximum capacity of bears in the park and in the forest, respectively. Intuitively, this means that the maximum proportion of the population that can occupy the park and the forest is the same. The emigration intensity increases proportionally with the density of bears in the park P_t^N, and decreases with the density of bears in the forest F_t^N. Additionally, a higher growth rate g also increases the migration pressure.

The emigration intensity from the park to the outside area consists of two terms:

$$q(\{O\}|P,\cdot,\mu_t^N) = \frac{(1-m_p)\cdot g}{K}\cdot P_t^N + \frac{m_p\cdot g}{K}\cdot P_t^N \cdot \min\Big\{\frac{F_t^N}{K},1\Big\}.$$

The first term describes direct emigration to the outside. The second term represents bears that would migrate to the forest but eventually end up in the outside area due to overpopulation of the forest.

iv) Emigration from the forest:

$$\begin{aligned} q(\{P\}|F,\cdot,\mu_t^N) &= \frac{m_f\cdot g}{K}\cdot F_t^N \cdot \Big(1-\frac{P_t^N}{K}\Big)\cdot \mathbb{1}_{\{P_t^N\le K\}},\\ q(\{O\}|F,\cdot,\mu_t^N) &= \frac{(1-m_f)\cdot g}{K}\cdot F_t^N + \frac{m_f\cdot g}{K}\cdot F_t^N \cdot \min\Big\{\frac{P_t^N}{K},1\Big\}. \end{aligned}$$

Here, $m_f \in [0,1]$ is the fraction of the forest boundary adjacent to the park. The intensities are analogous to the emigration from the park.

v) Intensities of stay:

$$\begin{aligned} q(\{\mathcal{T}\}|\mathcal{T},\cdot,\mu_t^N) &= -g\cdot\frac{P_t^N+F_t^N}{\mathcal{T}_t^N}\cdot\mathbb{1}_{\{\mathcal{T}_t^N\ge\varepsilon\}} - g\cdot\frac{P_t^N+F_t^N}{\varepsilon}\cdot\mathbb{1}_{\{\mathcal{T}_t^N<\varepsilon\}},\\ q(\{P\}|P,a_t,\mu_t^N) &= -\frac{m_p\cdot g}{K}\cdot P_t^N\cdot\Big(1-\frac{F_t^N}{K}\Big)\cdot\mathbb{1}_{\{F_t^N\le K\}}\\ &\quad -\frac{(1-m_p)\cdot g}{K}\cdot P_t^N - \frac{m_p\cdot g}{K}\cdot P_t^N\cdot\min\Big\{\frac{F_t^N}{K},1\Big\}\\ &\quad - a_p(t)\\ &= -\frac{g}{K}\cdot P_t^N - a_p(t),\\ q(\{F\}|F,a_t,\mu_t^N) &= -\frac{g}{K}\cdot F_t^N - a_f(t),\\ q(\{O\}|O,\cdot,\mu_t^N) &= 0. \end{aligned}$$

All other intensities are zero. We can observe two key characteristics of the model: First, bear mortality occurs exclusively through hunting; there is no natural death rate causing a transition back to the transcendent state. Second, note that the outside area constitutes an

absorbing state. Black bears do not migrate back to the forest or the park once they enter the outside area, nor can they be hunted there. With respect to these two characteristics, we follow Gross et al. (2005) and adopt these two assumptions without assessing their plausibility.

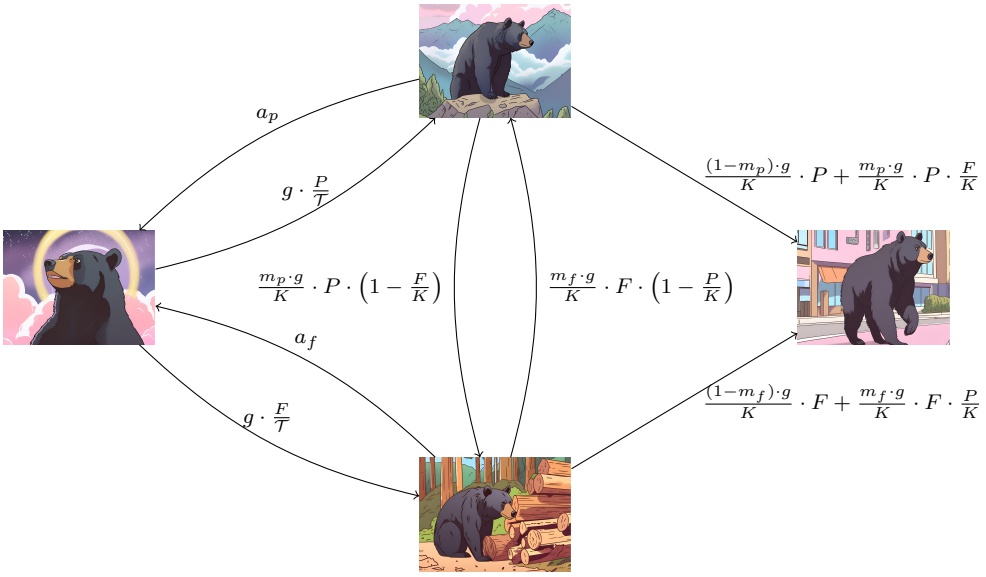

Figure 4.4.: Transition intensities of a bear in the case $\mathcal{T}_t^N \geq \varepsilon$, $F_t^N \leq K$ and $P_t^N \leq K$. Left: $\mathcal{T}$, Right: O, Top: P, Bottom: F.

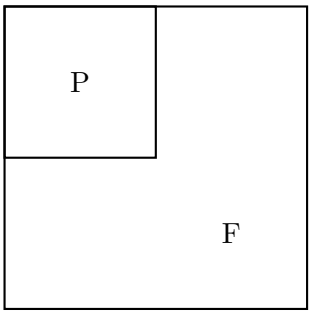

Figure 4.5.: Possible arrangement of national park and forest for the parameters $m_p = 1/2$ and $m_f = 1/4$.

The aim of the central controller is to minimize the proportion of bears in the outside urban areas by hunting bears in the park and the forest. Hunting a bear incurs a cost c_p (park) or c_f (forest). Usually, the price of hunting a bear in a national park is much higher than in an unpreserved forest, hence typically $c_p \gg c_f$. This leads to the following reward (cost) function for an individual bear:

$$r(i, a_t, \mu_t^N) = -\mathbb{1}_{\{i=O\}} - c_p \cdot a_p(t)^2 - c_f \cdot a_f(t)^2.$$

This application does not include a terminal reward or cost. Note that the intensities are affine in a and the reward is concave with respect to the control variables. By Lemma 3.21, no relaxed controls are required. Hence, the control region of the deterministic limit model is $A_{Det} = [0,1]^2$.

Remark 4.9. Note that conditions (R1) and (R2) are satisfied, since the reward function $r(i, a_t, \mu_t^N)$ depends quadratically on a_t and does not depend on μ_t^N. However, we include μ_t^N in the argument for the sake of completeness. The intensities are chosen in such a way that conditions (Q1)-(Q3) are satisfied. In particular, the introduction of the parameter ε ensures that the birth intensities remain bounded. Furthermore, observe that all intensities are Lipschitz continuous in μ_t^N, hence (Q4) and (Q4') hold as well. Finally, (Q5) holds because the bear hunting intensity is linear in the control.

4.3.2. Deterministic limit model

In order to determine the optimal hunting rates to effectively control the population, we consider the corresponding deterministic control problem (F_T), obtained by letting the number of black bears N tend to infinity. In what follows, we denote the limit state process by $\mu_t = (P_t, F_t, O_t, \mathcal{T}_t)$. The initial state of the N-bear system is given by $\mu_0^N = (P_0^N, F_0^N, O_0^N, \mathcal{T}_0^N) = (P_0^N, F_0^N, 0, 1 - P_0^N - F_0^N) \in \mathbb{P}_N(S)$ with $P_0^N \leq K$ and $F_0^N \leq K$. In addition, we assume that the sequence of initial distributions converges weakly, i.e., $(P_0^N, F_0^N, 0, 1 - P_0^N - F_0^N)_{N\in\mathbb{N}} \Rightarrow (P_0, F_0, 0, 1 - P_0 - F_0) \in \mathbb{P}(S)$, where again $P_0 \leq K$ and $F_0 \leq K$. Before turning to the deterministic limit problem, we discuss in more detail the relationship between the population in state $\mathcal{T}$ and the parameter ε.

Remark 4.10. The transcendent state $\mathcal{T}$ is an artificial dummy state that helps embed the application of population control into the theory of Mean-Field Markov Decision Models. The actual fraction of the bear population in the state $\mathcal{T}$ is of minor importance as long as it does not risk running empty. For example, the two configurations $(P, F, O, \mathcal{T}) = (0.25, 0.25, 0, 0.5)$ and $(P, F, O, \mathcal{T}) = (0.1, 0.1, 0, 0.8)$ are equivalent in the sense that the restriction of the population to the actual ecosystem is in both cases given by $(P, F, O) = (0.5, 0.5, 0)$. The outflow from state $\mathcal{T}$ is determined by the intensity

$$q(\{\mathcal{T}\}|\mathcal{T}, \cdot, \mu_t^N) \quad = -g \cdot \frac{P_t^N + F_t^N}{\mathcal{T}_t^N} \cdot \mathbb{1}_{\{\mathcal{T}_t^N \geq \varepsilon\}} - g \cdot \frac{P_t^N + F_t^N}{\varepsilon} \cdot \mathbb{1}_{\{\mathcal{T}_t^N < \varepsilon\}}$$

which is bounded below by $-2g/\varepsilon$. The corresponding state equation for the evolution of $(\mathcal{T}_t)_{t\geq 0}$ given by Theorem 3.11 is

$$\begin{aligned}\mathcal{T}_t &= \mathcal{T}_0 + \int_0^t \underbrace{a_p(s)P_s + a_f(s)F_s}_{\geq 0} + \mathcal{T}_s \cdot \left(-g \cdot \frac{P_s + F_s}{\mathcal{T}_s} \cdot \mathbb{1}_{\{\mathcal{T}_s \geq \varepsilon\}} - g \cdot \frac{P_s + F_s}{\varepsilon} \cdot \mathbb{1}_{\{\mathcal{T}_s < \varepsilon\}}\right) ds \\ &\geq \mathcal{T}_0 + \int_0^t \mathcal{T}_s \cdot \left(-\frac{2g}{\varepsilon}\right) ds.\end{aligned}$$

By inverting the signs we obtain

$$-\mathcal{T}_t \leq -\mathcal{T}_0 - \frac{2g}{\varepsilon}\int_0^t -\mathcal{T}_s ds.$$

Applying Gronwall's inequality (A.8) with $f(t) = -\mathcal{T}_t, A = -\mathcal{T}_0$ and $C = -2g/\varepsilon$ gives

$$-\mathcal{T}_t \leq -\mathcal{T}_0 \cdot e^{-\frac{2g}{\varepsilon}\cdot t},$$

which is again equivalent to

$$\mathcal{T}_t \geq \mathcal{T}_0 \cdot e^{-\frac{2g}{\varepsilon}\cdot t}.$$

The key message of the remark is now the following: If we choose the free variable $\mathcal{T}_0$ sufficiently large, the deterministic limit process $(\mathcal{T}_t)_{t\geq 0}$ does not drop below ε over the planning horizon. Therefore, the process remains in the regime $\mathcal{T}_t \geq \varepsilon$, so we can ignore the case distinction $\mathcal{T}_t \gtrless \varepsilon$. For the remainder of the section, assume $\mathcal{T}_0$ is chosen sufficiently large.

With $P_0, F_0 \leq K$ and the continuity of the limit state process (μ_t) ensured by Theorem 3.11, observe that by the specified intensities, (P_t) and (F_t) cannot exceed the maximum capacity K. Therefore, we may drop the min and the indicator terms in the corresponding intensities. In view of this and Remark 4.10, the deterministic control problem (F_T) is given by

$$\begin{aligned}
(BB) \qquad & \sup_{a=(a_p,a_f)} \int_0^T -O_t - c_p \cdot a_p(t)^2 - c_f \cdot a_f(t)^2 dt, \\
& s.t.\ a_p(t), a_f(t) \in [0,1], \quad P_0, F_0 \geq 0, \quad P_0, F_0 \leq K \text{ and for all } t \in [0,T] \\
& \quad P_t = P_0 + \int_0^t gP_s - \frac{g}{K}P_s^2 + \frac{m_f \cdot g}{K}\Big(1 - \frac{P_s}{K}\Big)F_s^2 - a_p(s)P_s ds, \\
& \quad F_t = F_0 + \int_0^t gF_s - \frac{g}{K}F_s^2 + \frac{m_p \cdot g}{K}\Big(1 - \frac{F_s}{K}\Big)P_s^2 - a_f(s)F_s ds, \\
& \quad O_t = 0 \ + \int_0^t g(1-m_p)\frac{P_s^2}{K} + g(1-m_f)\frac{F_s^2}{K} + \frac{m_f \cdot g}{K}F_s^2\frac{P_s}{K} + \frac{m_p \cdot g}{K}P_s^2\frac{F_s}{K} ds.
\end{aligned}$$

Since the state process (μ_t) provides a distribution on the state space for every $t \in [0,T]$, only $|S| - 1$ differential equations are sufficient to describe the process. Thus, we omit the differential equation for the added transcendent state $\mathcal{T}$ and result in a deterministic control problem equivalent to that in Gross et al. (2005).

We again start the investigation of (BB) by verifying the assumptions of the Filippov-Cesari existence theorem.

Theorem 4.11. *The optimal control problem (BB) has an optimal pair $(\mu_t^*, a_t^*)_{t\in[0,T]}$.*

Proof. For $a_t = (a_p(t), a_f(t)) \equiv (0,0)$, with the corresponding state process (μ_t) the state-control trajectory (μ_t, a_t) is admissible, which satisfies assumption i) of the Filippov-Cesari existence theorem 4.2. The control $a_t \equiv (0,0)$ corresponds to the situation in which there

is no bear hunting during the optimization horizon. It remains to check the convexity of the set

$$
\begin{aligned}
N(\mu,[0,1]^2) = \Bigg\{ \Bigg(& -O - c_p a_p^2 - c_f a_f^2 + \gamma, \\
& gP - \frac{g}{K}P^2 + \frac{m_f \cdot g}{K}\Big(1-\frac{P}{K}\Big)F^2 - a_p P, \\
& gF - \frac{g}{K}F^2 + \frac{m_p \cdot g}{K}\Big(1-\frac{F}{K}\Big)P^2 - a_f F, \\
& g(1-m_p)\frac{P^2}{K} + g(1-m_f)\frac{F^2}{K} + \frac{m_f \cdot g}{K}F^2\frac{P}{K} + \frac{m_p \cdot g}{K}P^2\frac{F}{K}\Bigg) \\
& \Bigg| \ \gamma \le 0,\ (a_p, a_f) \in [0,1]^2 \Bigg\}
\end{aligned}
$$

for every $\mu \in \mathbb{P}(S)$. The last component does not depend on a at all, the second and third depend linearly. Thus, they are not critical with respect to the convexity of $N(\mu,[0,1]^2)$ and we only need to consider the first component in greater detail.

So, let $n_1, n_2 \in N(\mu,[0,1]^2)$ with $\gamma_1, \gamma_2 \le 0$ and $a_1 = (a_{p,1}, a_{f,1}), a_2 = (a_{p,2}, a_{f,2}) \in [0,1]^2$. We must show that the convex combination $n_3 := \lambda n_1 + (1-\lambda)n_2$ is again in $N(\mu,[0,1]^2)$ for any $\lambda \in [0,1]$. The task is to find suitable parameters a_3 and γ_3 for n_3. Define the function $c(a) := -c_p a_p^2 - c_f a_f^2$. It is clear that the function is concave, thus we have

$$
c(\lambda a_1 + (1-\lambda)a_2) \ge \lambda c(a_1) + (1-\lambda)c(a_2). \tag{4.16}
$$

Now set

$$
\begin{aligned}
a_3 &= \lambda a_1 + (1-\lambda)a_2, \\
\gamma_3 &= \lambda\gamma_1 + (1-\lambda)\gamma_2 + \lambda c(a_1) + (1-\lambda)c(a_2) - c(a_3)
\end{aligned}
$$

The inequality (4.16) implies that $\gamma_3 \le 0$. The first component of n_3, denoted by $n_3(1)$, is given by

$$
\begin{aligned}
n_3(1) &= \lambda n_1(1) + (1-\lambda)n_2(1) \\
&= \lambda(-O - c_p a_{p,1}^2 - c_f a_{f,1}^2 + \gamma_1) + (1-\lambda)(-O - c_p a_{p,2}^2 - c_f a_{f,2}^2 + \gamma_2) \\
&= -O + \lambda c(a_1) + \lambda\gamma_1 + (1-\lambda)c(a_2) + (1-\lambda)\gamma_2 \\
&= -O + \underbrace{\lambda c(a_1) + \lambda\gamma_1 + (1-\lambda)c(a_2) + (1-\lambda)\gamma_2 - c(a_3)}_{=\gamma_3} + c(a_3) \\
&= -O + c(a_3) + \gamma_3.
\end{aligned}
$$

We conclude that by the choice of a_3 and γ_3, the convex combination n_3 is again in $N(\mu,[0,1]^2)$, which completes the proof. □

We now turn to the solution of the optimal control problem (BB) and derive a unique representation of the optimal control depending on the state process and the adjoint function given by Pontryagin's maximum principle. Later, the resulting representation of the control is evaluated numerically in Section 4.3.3.

Theorem 4.12. *Let $(\mu_t^*, a_t^*)_{t\in[0,T]}$ be an optimal solution to (BB) with corresponding adjoint function $p = (p_P, p_F, p_O)$. Then the optimal control $(a_t^*)_{t\in[0,T]}$ is of the form*

$$a_p^*(t) = \begin{cases} 0, & -\frac{p_P(t)P_t^*}{2c_p} < 0, \\ -\frac{p_P(t)P_t^*}{2c_p}, & -\frac{p_P(t)P_t^*}{2c_p} \in [0,1], \\ 1, & -\frac{p_P(t)P_t^*}{2c_p} > 1. \end{cases} \qquad a_f^*(t) = \begin{cases} 0, & -\frac{p_F(t)F_t^*}{2c_f} < 0, \\ -\frac{p_F(t)F_t^*}{2c_f}, & -\frac{p_F(t)F_t^*}{2c_f} \in [0,1], \\ 1, & -\frac{p_F(t)F_t^*}{2c_f} > 1. \end{cases}$$

Proof. The Hamiltonian for (BB) is given by

$$\begin{aligned} H(P,F,O,a_p,a_f,p_P,p_F,p_O) =& -O - c_p a_p^2 - c_f a_f^2 \\ &+ p_P\left(gP - \frac{g}{K}P^2 + \frac{m_f\cdot g}{K}\left(1-\frac{P}{K}\right)F^2 - a_pP\right) \\ &+ p_F\left(gF - \frac{g}{K}F^2 + \frac{m_p\cdot g}{K}\left(1-\frac{F}{K}\right)P^2 - a_fF\right) \\ &+ p_O\left((1-m_p)\frac{P^2}{K} + g(1-m_f)\frac{F^2}{K} + \frac{m_f\cdot g}{K^2}PF^2 + \frac{m_p\cdot g}{K^2}P^2F\right) \end{aligned}$$

Pontryagin's maximum principle gives the following necessary conditions for an optimal control $(a_t^*)_{t\in[0,T]}$:

Necessary Conditions:
Let $(\mu_t^*, a_t^*)_{t\in[0,T]}$ be an optimal solution of (BB). Then there exist continuous and piecewise continuously differentiable adjoint functions $p_P, p_F, p_O : [0,T] \to \mathbb{R}$ satisfying for all $t \in [0,T]$

$$H(\mu_t^*, a_p^*(t), a_f^*(t), p_P(t), p_F(t), p_O(t)) = \max_{a_p,a_f\in[0,1]} H(\mu_t^*, a_p, a_f, p_P(t), p_F(t), p_O(t)), \quad (4.17)$$

and at every point t where (a_t^*) is continuous

$$\begin{aligned} p_P'(t) =& -H_P(\mu_t^*, a_p^*(t), a_f^*(t), p_P(t), p_F(t), p_O(t)) \\ =& -p_P(t)\left(g - \frac{2g}{K}P_t^* - \frac{m_f g}{K^2}(F_t^*)^2 - a_p^*(t)\right) \\ &- p_F(t)\cdot\frac{2m_p g}{K}P_t^*\left(1-\frac{F_t^*}{K}\right) \\ &- p_O(t)\left(\frac{2g(1-m_p)}{K}P_t^* + \frac{m_f g}{K^2}(F_t^*)^2 + \frac{2m_p g}{K^2}P_t^*F_t^*\right) \end{aligned} \quad (4.18)$$

$$\begin{aligned} p'_F(t) =& - H_F(\mu_t^*, a_p^*(t), a_f^*(t), p_P(t), p_F(t), p_O(t)) \\ =& - p_P(t) \cdot \frac{2m_f g}{K} F_t^* \left(1 - \frac{P_t^*}{K}\right) \\ & - p_F(t) \left(g - \frac{2g}{K} F_t^* - \frac{m_p g}{K^2} (P_t^*)^2 - a_f^*(t)\right) \\ & - p_O(t) \left(\frac{2g(1-m_f)}{K} F_t^* + \frac{2m_f g}{K^2} P_t^* F_t^* + \frac{m_p g}{K^2} (P_t^*)^2\right) \end{aligned} \tag{4.19}$$

$$\begin{aligned} p'_O(t) =& -H_O(\mu_t^*, a_p^*(t), a_f^*(t), p_P(t), p_F(t), p_O(t)) \\ =& \, 1. \end{aligned} \tag{4.20}$$

Furthermore, the transversality condition holds:

$$p_P(T) = p_F(T) = p_O(T) = 0.$$

The transversality condition, together with (4.20), implies that $p_O(t) = t - T$. In order to investigate the maximization condition (4.17), we take into account the partial derivatives of the Hamiltonian with respect to the control variables a_p and a_f:

$$\begin{aligned} H_{a_p}(\mu_t^*, a_p, a_f, p_P(t), p_F(t), p_O(t)) &= -2c_p a_p - p_P(t) P_t^*, \\ H_{a_f}(\mu_t^*, a_p, a_f, p_P(t), p_F(t), p_O(t)) &= -2c_f a_f - p_F(t) F_t^*, \\ H_{a_p a_p}(\mu_t^*, a_p, a_f, p_P(t), p_F(t), p_O(t)) &= -2c_p < 0, \\ H_{a_f a_f}(\mu_t^*, a_p, a_f, p_P(t), p_F(t), p_O(t)) &= -2c_f < 0. \end{aligned}$$

It follows that the Hamiltonian is strictly concave in the arguments a_p and a_f. The unique maximum of the Hamiltonian related to the control variables is therefore given by

$$a_p^*(t) = \begin{cases} 0, & -\frac{p_P(t)P_t^*}{2c_p} < 0, \\ -\frac{p_P(t)P_t^*}{2c_p}, & -\frac{p_P(t)P_t^*}{2c_p} \in [0,1], \\ 1, & -\frac{p_P(t)P_t^*}{2c_p} > 1. \end{cases} \qquad a_f^*(t) = \begin{cases} 0, & -\frac{p_F(t)F_t^*}{2c_f} < 0, \\ -\frac{p_F(t)F_t^*}{2c_f}, & -\frac{p_F(t)F_t^*}{2c_f} \in [0,1], \\ 1, & -\frac{p_F(t)F_t^*}{2c_f} > 1. \end{cases}$$

□

4.3.3. Numerical solution and the forward-backward sweep method

The aim of this section is to give an idea of how to solve deterministic limit problems numerically. There is a broad spectrum of numerical methods for optimal control problems, including Runge-Kutta schemes, direct and indirect shooting methods, direct and indirect transcription methods, etc. For an overview, see Betts (1998) or Rao (2010).

Here, we follow the approach presented by Lenhart and Workman (2007), Chapter 4. The starting point is to use the maximization condition (4.17) to obtain an expression for the optimal control a^* as a function of the state process μ^* and the adjoint function p. For problem (BB), we have already achieved this through Theorem 4.12. Having established the representation of a^*, the next step is to insert a^* into the differential equations for the state process and the adjoint function. Observe that the differential equation for the state process satisfies an initial condition, while the transversality condition gives a terminal condition for the adjoint. At this point, solving the optimization problem reduces to solving two systems of differential equations. To address these, we present in detail a method that is commonly known as the *Forward-Backward Sweep method (FBSM)*. For a review of the theory and implementation of the FBSM, as well as acceleration techniques, see Sharp et al. (2021).

Outline of the FBSM solving algorithm:

1. Divide the planning horizon $[0, T]$ into an equidistant fine partition $0 = t_0 < t_1 < \ldots < t_n = T$. The aim is to find an approximation $\hat{a} = (\hat{a}_{t_0}, \ldots, \hat{a}_{t_n})$ such that $\hat{a}_{t_j} \approx a^*_{t_j}$. As an initial guess for $\hat{a}$ in (BB) we simply choose $\hat{a} \equiv (0, 0)$.

2. Solve the system of differential equations for the state process with the initial condition μ_0, using the current approximation for the control $\hat{a}$. We obtain an approximation for the state process $\hat{\mu} = (\hat{\mu}_{t_0}, \ldots, \hat{\mu}_{t_n})$. Note that the state equation does not depend on the adjoint function.

3. Solve the system of differential equations for the adjoint function with the terminal condition given by the transversality condition, using the current approximations for the control $\hat{a}$ and the state process $\hat{\mu}$. We obtain an approximation for the adjoint process $\hat{p} = (\hat{p}_{t_0}, \ldots, \hat{p}_{t_n})$.

4. Update the control approximation $\hat{a}^{\text{new}}$ by inserting the approximations $\hat{\mu}$ and $\hat{p}$ into the representation of a^* given by the maximization condition (4.3).

5. If the difference between successive iterates is sufficiently small (e.g., $\|\hat{a}^{\text{new}} - \hat{a}\| < \varepsilon$), terminate the algorithm. If not, return to Step 2.

In order to solve the systems of differential equations in Steps 2 and 3, we follow the approach of Lenhart and Workman (2007) and present a fourth-order Runge-Kutta method. Note that any other standard ODE solver may be adequate. Let h be the step size given by the partition of the planning horizon. Further, denote the systems by

$$\mu_t' = f_1(\mu_t), \qquad p_t' = f_2(p_t).$$

Then the new approximation for the state process is constructed forward in time by the rule

$$\hat{\mu}_{t+h} = \hat{\mu}_t + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

where $t \in \{t_0, t_1, \ldots, t_{n-1}\}$ and

$$k_1 = f_1(\hat{\mu}_t), \qquad k_2 = f_1(\hat{\mu}_t + \frac{h}{2}k_1), \qquad k_3 = f_1(\hat{\mu}_t + \frac{h}{2}k_2), \qquad k_4 = f_1(\hat{\mu}_t + hk_3).$$

The transversality condition gives a terminal value for the adjoint system, therefore we use a backward approach to approximate the trajectory of the adjoint:

$$\hat{p}_{t-h} = \hat{p}_t - \frac{h}{6}(\tilde{k}_1 + 2\tilde{k}_2 + 2\tilde{k}_3 + \tilde{k}_4),$$

where $t \in \{t_n, t_{n-1}, \ldots, t_1\}$ and

$$\tilde{k}_1 = f_2(\hat{p}_t), \qquad \tilde{k}_2 = f_2(\hat{p}_t - \frac{h}{2}\tilde{k}_1), \qquad \tilde{k}_3 = f_2(\hat{p}_t - \frac{h}{2}\tilde{k}_2), \qquad \tilde{k}_4 = f_2(\hat{p}_t - h\tilde{k}_3).$$

The procedure of constructing the state process forward in time and the adjoint backward in time is what gives the Forward-Backward Sweep method its name. McAsey et al. (2012) prove the convergence of the FBSM under certain Lipschitz conditions on the functions of the optimal control problem. In particular, we show that problem (BB) satisfies these conditions.

Convergence of the forward-backward sweep method

Assume that the optimality system can be written in the following form:

$$\begin{aligned}
\mu_t' &= h_0(\mu_t, a_t), \\
p_t' &= h_1(\mu_t, a_t) + p_t \cdot h_2(\mu_t, a_t), \\
a_t^* &= h_3(\mu_t, p_t),
\end{aligned}$$

with the boundary conditions $\mu_0 \in \mathbb{P}(S)$ and $p(T) = 0$, where h_3 is a unique representation of the optimal control depending on the state process μ and the adjoint p, usually obtained by the maximization condition (4.3). In problem (BB), the function h_3 is given by Theorem 4.12. In this section we assume that h_3 and, consequently, a^* are continuous in time. Note that this is satisfied for the optimal control in (BB), since the state processes P^*, F^* and the corresponding adjoint functions p_P and p_F are continuous in time.

In order to state the main convergence result of the section, we impose the following Lipschitz assumptions upon the model functions:

Assumption (L): The functions h_0, h_1, h_2, h_3 are Lipschitz continuous in both arguments, with Lipschitz constants $L_{h_0}, \ldots, L_{h_3}$. Moreover, $\|p\|_\infty < \infty$ and $\|h_2\|_\infty < \infty$.

Theorem 4.13 (McAsey et al. (2012))**.** *Assume that h_3 and, consequently, a^* are continuous in time. In addition, assume **(L)** holds. Now suppose that either the Lipschitz constants are small enough or that the planning horizon T is small enough. Then*

$$\max_{j=0,\dots,n} \{\|\mu_{t_j} - \hat{\mu}^k_{t_j}\| + \|p_{t_j} - \hat{p}^k_{t_j}\| + \|a^*_{t_j} - \hat{a}^k_{t_j}\|\} \longrightarrow 0$$

as $k, n \to \infty$, where k denotes the k-th iteration of the FBSM.

For the proof, see Theorem 3.3 in McAsey et al. (2012).

Proposition 4.14. *The problem (BB) satisfies the assumption **(L)**.*

Proof. The function h_0 corresponds to the right-hand side of the state equations in the optimization problem (BB). The control variables a_p and a_f enter h_0 linearly. Thus, h_0 is Lipschitz continuous with respect to a_p and a_f. The three state variables P, F, and O enter h_0 linearly and quadratically, where we know that $P, F, O \in [0, 1]$. Therefore, the partial derivatives of h_0 with respect to the state variables are bounded, which implies that h_0 is Lipschitz continuous with respect to P, F, and O as well.

The right-hand sides of the differential equations for the adjoint variables p_P, p_F, and p_O (see (4.18), (4.19), and (4.20)) define the functions h_1 and h_2. The function h_1 is simply given by $h_1 \equiv (0, 0, 1)$ and is therefore trivially Lipschitz continuous. The Lipschitz continuity of h_2 follows from the same reasoning as for h_0. Additionally, since $P, F, O \in [0, 1]$, it follows that h_2 itself is bounded, that is, $\|h_2\|_\infty < \infty$. In particular, since the adjoint system $p' = h_1 + p \cdot h_2$ is linear in the adjoint variables with bounded time-dependent coefficients h_1 and h_2 on the finite time horizon $[0, T]$, it follows that the adjoint variables remain bounded. Consequently, we have $\|p\|_\infty < \infty$.

Finally, as already mentioned, h_3 is given by the representation of the optimal control in Theorem 4.12. Since the state variables P and F as well as the adjoint variables p_P and p_F enter linearly, it follows that h_3 is Lipschitz continuous.

□

We refrain from diving deeper into the discussion of when the Lipschitz constants are "small enough" to ensure convergence. Instead, we proceed with the empirical observation that the FBSM performs well in our black bear application and turn, in the following section, to the presentation of the numerical solution for a specific choice of parameters.

4.3.4. Illustration

To compute and illustrate the optimal trajectories, we apply the Forward-Backward Sweep method to problem (BB). We consider a spatial arrangement of park and forest with boundary parameters $m_p = 0.5$ and $m_f = 0.25$; a possible outcome is shown in Figure 4.5. The reproduction rate is set to $g = 0.1$, and the maximum capacity for both park and forest is $K = 0.2$. The hunting costs for bears are $c_p = 10000$ in the park and $c_f = 10$

in the forest. As initial conditions, the bear population is distributed among the states as follows: $P_0 = 0.1,\ F_0 = 0.1,\ O_0 = 0$ and $\mathcal{T}_0 = 0.8$. The specified planning horizon is $T = 25$.

The resulting trajectories in Figure 4.6 reveal distinct dynamics across the regions. The bear concentration in the park increases steadily throughout the planning horizon, approaching the capacity limit of $K = 0.2$. In contrast, the forest concentration initially decreases slightly but then begins to grow after reaching a minimum, also approaching the capacity limit of $K = 0.2$. The concentration outside the designated regions exhibits a continuous and significant increase, indicating the expected steady outflow of the population from the park and forest to the surrounding areas. Given that the urban areas in our model represent an absorbing state, the trajectory's course is not particularly surprising.

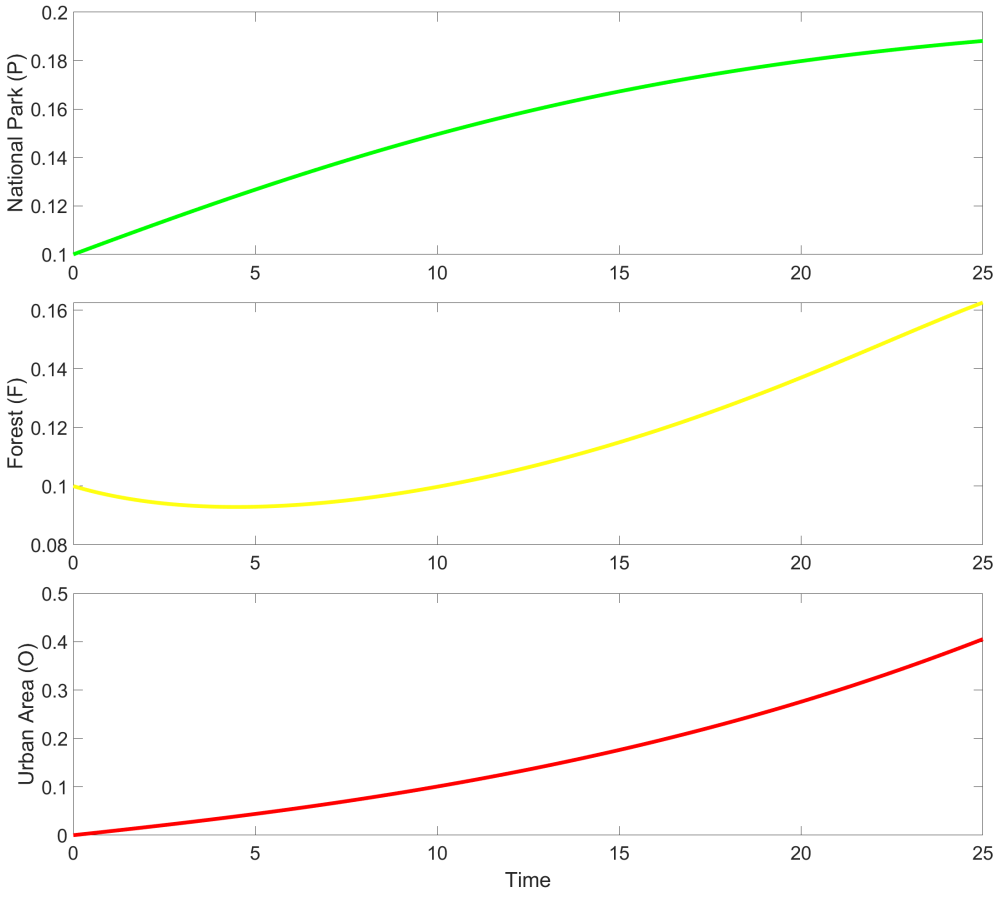

Figure 4.6.: Optimal state trajectories (P_t^*), (F_t^*), and (O_t^*).

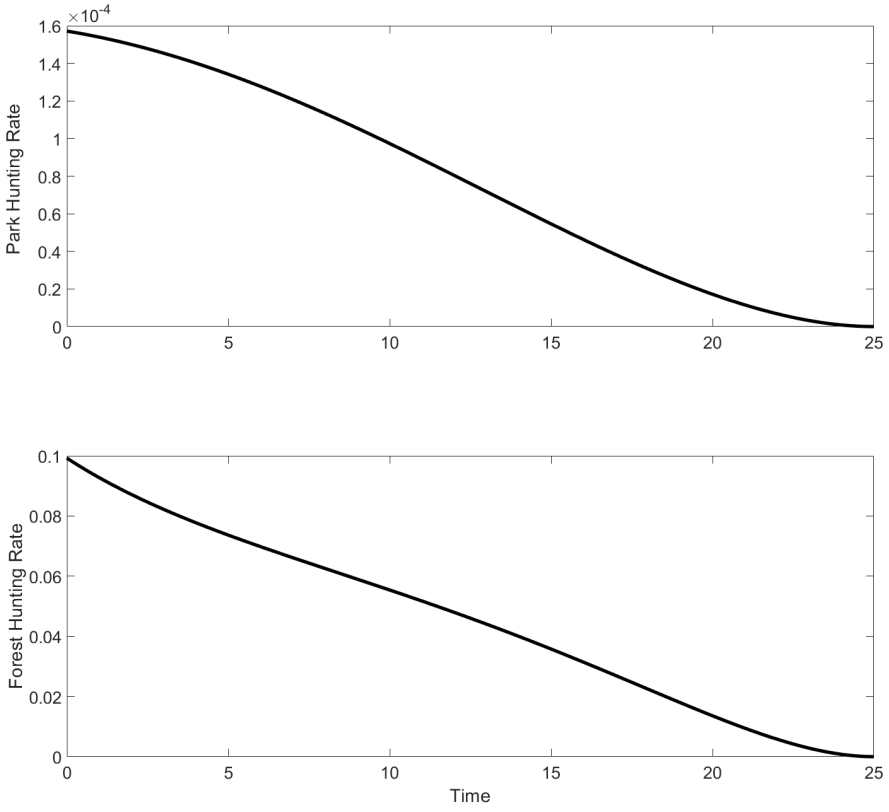

Figure 4.7.: Optimal hunting rates $(a_t^*) = (a_p^*(t), a_f^*(t))$.

Regarding the hunting rates in Figure 4.7, for both the park and the forest, we observe a clear decline. The hunting rate in the forest starts at a relatively high level and then declines steadily to nearly zero. This initial high hunting pressure explains the temporary decrease in the forest population at the beginning of the planning horizon. In contrast, the hunting rate in the park remains very low throughout the entire period and also diminishes over time. The substantial difference in the magnitude of hunting rates between the two regions can be attributed to the significant cost difference, with $c_f = 10$ in the forest and $c_p = 10000$ in the park.

This application has two key messages. First, it shows that by introducing an artificial transcendent state, it is possible to embed systems in which agents follow a birth-and-death process into the framework of N-agent mean-field MDPs. Second, we have shown that the deterministic limit model presented in Gross et al. (2005) can be interpreted as a limit model of an N-agent mean-field MDP. Theorem 3.17 then ensures that optimal controls of the limit model are asymptotically optimal when applied to the N-agent model. Modeling the population in a discrete framework is well justified and may, in fact, be closer to reality than interpreting the number of bears as a continuous quantity as in the limit model, given that bear counts are inherently discrete and typically modest in size. To get a sense of scale, the current black bear population in the Great Smoky Mountains National Park is about 1,900 individuals, according to the National Park Service (2025).

4.4. Machine replacement

Imagine a company that produces certain goods with the help of a large number of statistically equal machines. A machine can break down and stop producing, requiring costly repairs to restore productivity. We assume that the repair of a batch of machines becomes cheaper as more machines are out of order, which leads to a certain interaction between the machines. The task of the company is to control the maintenance of the production site, balancing the rewards from producing goods and the cost of service in an optimal manner during the planning horizon. Unlike the other applications in this chapter, this model considers an infinite time horizon. In the resulting deterministic limit model, we have the fraction of functional machines as the state variable and the fraction of broken machines that do not receive repair as the control variable. We solve the limit model using a concept known in the literature as the *most rapid approach path (MRAP)*. The optimal control is characterized by two phases: first, reaching an optimal singular state as fast as possible, and second, maintaining that state thereafter.

A related problem dealing with the optimal maintenance of a single machine was proposed by Thompson (1968). He formulates a deterministic optimal control problem, which he solves by applying Pontryagin's maximum principle. The optimal maintenance derived by Thompson turns out to be of bang-bang type.

A related mean-field application is presented in Huang and Ma (2016). They propose a discrete-time mean-field stochastic game with N agents, each operating on the state space $S = [0, 1]$. For each agent, the state space can be interpreted as the risk or distress level. The action space of an agent is a set consisting of two actions $A = \{a_0, a_1\}$, where a_0 suggests doing nothing and a_1 resets the stress level. The agents then interact through their cost functions. Translated to the machine replacement situation, the state space $[0, 1]$ can be interpreted as the condition of a machine and the action a_1 as the call for maintenance. Huang and Ma show that for the derived mean-field limit model the optimal policy of an agent is a pure Markov threshold policy, that is, there exists a parameter $c \in [0, 1]$ such that the agent chooses to wait if its state $x \in [0, 1]$ is below c, and otherwise, in the case $x \geq c$, it takes the action a_1 to reset the risk level. Further, they prove that the system of threshold policies from the mean-field limit provides an ε-Nash equilibrium for the underlying N-agent game.

We start the investigation of our machine replacement model with a description of the N-machine problem.

4.4.1. Description of the N-agent model

Suppose a company has N statistically equal machines. Each machine can be in one of the following states:

- *Working (w)*: The machine is functional and produces a reward for the company,
- *Broken (b)*: The machine is out of service and needs repair in order to produce again,

thus $S = \{w, b\}$. The joint process $\mu_t^N = (w_t^N, b_t^N)$ is a controlled continuous-time Markov chain where w_t^N represents the fraction of working machines and b_t^N the fraction of broken machines. Thus, we have $w_t^N + b_t^N = 1$ and $w_t^N, b_t^N \geq 0$ for all $t \geq 0$. The initial fraction of working machines in the N-machine system is given by $w_0^N \in [0, 1]$, where we assume that $(w_0^N)_{N\in\mathbb{N}} \Rightarrow w_0 \in [0, 1]$.
The action of the company is to determine for each machine whether it receives a repair or not. The basic action space for a machine is therefore given by $A = \{a^r, a^d\}$, where a^r represents the action "repair" and a^d the action "do nothing". Naturally, a working machine does not need repair; this is ensured by appropriate rewards.
The transition rates of one machine are as follows: A working machine breaks down at a fixed rate $\lambda_{wb} > 0$. A broken machine that gets repaired changes to the state "working" with rate $\lambda_{bw} > 0$. Thus, we can summarize the transition rates of one machine by

$$q(\{b\}|w, a^d, \mu_t^N) = \lambda_{wb}, \quad q(\{w\}|b, a, \mu_t^N) = \lambda_{bw}\mathbb{1}_{\{a=a^r\}}.$$

The diagonal elements of the intensity matrix are given by

$$q(\{w\}|w, a^d, \mu_t^N) = -\lambda_{wb}, \quad q(\{b\}|b, a, \mu_t^N) = -\lambda_{bw}\mathbb{1}_{\{a=a^r\}},$$

and all other intensities are zero. Obviously, (Q1)-(Q5) and (Q4') are satisfied.

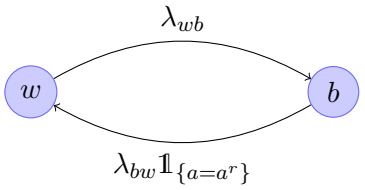

Figure 4.8.: Transition intensities of one machine between the two states.

Each working machine produces a reward rate $R > 0$ whereas the company has to pay a fixed cost rate of $C > 0$ when it has to call the service for repair for a machine, i.e.,

$$r(i, a, \mu_t^N) = R \cdot \mathbb{1}_{\{i=w\}}\mathbb{1}_{\{a=a^d\}} - C \cdot \mathbb{1}_{\{i=b\}}\mathbb{1}_{\{a=a^r\}}\frac{1}{b_t^N}.$$

Note that (R1), (R2) are satisfied. Observe that, in contrast to the transition, an interaction between individuals occurs in the reward. The cost of repair of a broken machine is smaller

the more other machines are broken. This can be explained by economies of scale. For example, the mechanic only has to travel to the site once, and these costs are then shared between all the machines to be repaired.
As mentioned above, there is no incentive for the company to repair a working machine; thus, the action chosen in this case is always a^d. In terms of the action process for the system of machines this means $\hat{\pi}_t^w(\{a^d\}) \equiv 1$. To describe the action process of the system $(\hat{\pi}_t)_{t\in[0,\infty)}$, it is therefore sufficient to denote the proportion of broken machines that are NOT being repaired at a certain point in time t. To keep the notation simple, we denote $a_t^N := \hat{\pi}_t^b(\{a^d\})$. The fraction of broken machines that receive service is then simply given by the process $(1 - a_t^N)_{t\in[0,\infty)}$. Following the discussion in Section 4.1, since the action space is finite, the control region of the deterministic limit model is $A_{Det} = \mathbb{P}(A)^2$, which simplifies to $A_{Det} = [0, 1]$.

The reward rate for the entire system is now given by

$$r(\mu_t^N, a_t^N) = Rw_t^N - C(1 - a_t^N).$$

The aim of the company is to find the balance between the reward that a machine produces and the costs of repair, which decrease for the individual machine the more machines are broken. For the rest of the chapter, we specify the parameters $R = 2$, $C = 1$, $\lambda_{wb} = 1$ and $\lambda_{bw} = 2$.

4.4.2. DETERMINISTIC LIMIT MODEL

To obtain the deterministic limit optimization problem (F), see (3.31), let the number of machines N tend to infinity. The limit state process is denoted by $\mu_t = (w_t, b_t) = (w_t, 1-w_t)$ and the control by $(a_t)_{t\in[0,\infty)}$, where a_t represents the fraction of broken machines that do not receive repair. Since we consider an infinite-horizon model, let $\beta > 0$ be the discount rate. Then, the limit problem (F), denoted in this setting by (MR), is given by

$$\begin{aligned}(MR) \qquad & \sup_a \int_0^\infty e^{-\beta t} \cdot (2w_t - (1 - a_t))dt, \\ & s.t. \ \text{ for all } t \in [0,\infty): \ w_0 \in [0,1], \ a_t \in [0,1] \text{ and} \\ & w_t = w_0 + \int_0^t 2(1 - w_s)(1 - a_s) - w_s ds. \end{aligned} \tag{4.21}$$

Since the control variable enters the problem linearly, potential candidates for the optimal control are either bang-bang or singular controls (cf. Remark 4.5). As it turns out, in this scenario, the structure of the optimal control is indeed to reach a singular state as fast as possible.

Theorem 4.15 (Optimal control in the Machine Replacement Model (MR))**.**

a) Let $\beta \in (0,3]$. Then there exists a singular state $\hat{w}(\beta) \in [0, \frac{1}{2})$ such that the optimal control for (MR) is given by

$$a_t^* = \begin{cases} 1, & w_t > \hat{w}(\beta), \\ 1 - \frac{\hat{w}(\beta)}{2(1-\hat{w}(\beta))}, & w_t = \hat{w}(\beta), \\ 0, & w_t < \hat{w}(\beta). \end{cases}$$

The singular state $\hat{w}(\beta)$ can be determined as the unique solution in the interval $[0,1]$ of the quadratic equation

$$4w^2 + (\beta - 8)w + 3 - \beta = 0, \tag{4.22}$$

which exists for all $\beta \in (0,3]$.

b) Let $\beta > 3$. Then the optimal control to (MR) is given by $a_t^* \equiv 1$.

Intuitively, the optimal control in the case $\beta \in (0,3]$ can be described as follows: If the initial state is $w_0 \in (\hat{w}(\beta), 1]$, intentionally let the machines break down until the singular state is reached. Conversely, if the initial state is $w_0 \in [0, \hat{w}(\beta))$, repair the broken machines until the singular state is reached. From that point on, the optimal control is to maintain the singular level, as it represents the optimal long-term balance between profit and repair costs. For $\beta > 3$, future rewards are discounted so heavily that they no longer outweigh the repair costs. Consequently, the optimal control is to never repair.

Lemma 4.16. *Let $\beta \in (0,3]$ and suppose that the system under optimal control has reached the singular state at time $\hat{t}$, that is, $w_{\hat{t}} = \hat{w}(\beta)$ and $a_{\hat{t}}^* = 1 - \frac{\hat{w}(\beta)}{2(1-\hat{w}(\beta))} =: \hat{a}$. Then the system is indeed singular in the sense of Remark 4.5, satisfying*

$$\varphi(t) = H_a(w_t^*, a_t^*, p_t) \overset{!}{=} 0, \qquad \forall\, t \in [\hat{t}, \infty).$$

Proof. First, observe that once the system enters the state $\hat{w}(\beta)$, it remains there under the optimal control (a_t^*). This follows from the fact that the derivative of the state process vanishes:

$$w'_{\hat{t}} = 2(1 - \hat{w}(\beta))(1 - \hat{a}) - \hat{w}(\beta) = 2(1 - \hat{w}(\beta)) \left(1 - \left(1 - \frac{\hat{w}(\beta)}{2(1 - \hat{w}(\beta))}\right)\right) - \hat{w}(\beta) = 0.$$

The Hamiltonian for problem (MR) is given by

$$H(w, a, p) = 2w - (1 - a) + p\big(2(1 - w)(1 - a) - w\big),$$

where (p_t) is the corresponding adjoint function.

Differentiating the Hamiltonian with respect to the control a, we obtain the switching function

$$\varphi(t) = H_a(w_t, a_t, p_t) = 1 - 2p_t(1 - w_t).$$

For the system to be singular on $[\hat{t}, \infty)$, we must show that $\varphi(t) = 0$ for all $t \geq \hat{t}$. Since $w_t = \hat{w}(\beta)$ is constant on this path, this requires that the adjoint p_t also takes a constant value $\hat{p}$, which satisfies

$$1 - 2\hat{p}(1 - \hat{w}(\beta)) = 0 \quad \Longrightarrow \quad \hat{p} = \frac{1}{2(1 - \hat{w}(\beta))}.$$

The adjoint function (p_t) in the discounted case is given by the differential equation

$$\begin{aligned} p_t' &= \beta p_t - H_w(w_t^*, a_t^*, p_t) \\ &= \beta p_t - (2 - p_t(3 - 2a_t^*)) \\ &= p_t(\beta + 3 - 2a_t^*) - 2. \end{aligned}$$

For the adjoint to be constant, its time derivative must be zero. Inserting the expressions for $\hat{p}$ and $\hat{a}$ in terms of $\hat{w} = \hat{w}(\beta)$ yields:

$$\begin{aligned} & & 0 &= \hat{p}(\beta + 3 - 2\hat{a}) - 2 \\ \iff & & 0 &= \frac{1}{2(1-\hat{w})}\left(\beta + 3 - 2\left(1 - \frac{\hat{w}}{2(1-\hat{w})}\right)\right) - 2 \\ \iff & & 0 &= 4\hat{w}^2 + (\beta - 8)\hat{w} + (3 - \beta). \end{aligned}$$

This is precisely the defining quadratic equation for the singular state $\hat{w}(\beta)$. Consequently, the triple $(\hat{w}, \hat{a}, \hat{p})$ constitutes a stationary state satisfying $w_t' = 0$, $a_t' = 0$, $p_t' = 0$, and, in particular, $\varphi(t) = 0$ for all $t \in [\hat{t}, \infty)$, which implies the statement. □

The control given in part a) of Theorem 4.15, in which a singular state should be reached as fast as possible, constitutes a so-called *most rapid approach path* to the singular state.

Definition 4.17 (Most Rapid Approach Path (MRAP))**.**
A most rapid approach path (MRAP) $(w_t^*)_{t\in[0,\infty)}$ to a given path $(\tilde{w}_t)_{t\in[0,\infty)}$ has the property

$$|w_t^* - \tilde{w}_t| \leq |w_t - \tilde{w}_t|, \qquad \text{for all} \quad t \in [0, \infty)$$

for all admissible state trajectories $(w_t)_{t\in[0,\infty)}$.

To prove Theorem 4.15 and establish the optimality of the MRAP approach, we adapt the method from Section 3.3 of Feichtinger and Hartl (1986), see also Section 3.5.1 of Grass et al. (2008). First, define the functions

$$A(w) = 2w - 1, \qquad B(w) \equiv 1, \qquad a(w) = 2 - 3w, \qquad b(w) = 2w - 2, \qquad \Psi(a) = a.$$

Then, problem (MR) can be written in the form

$$\begin{aligned}(MR) \qquad & \sup_a \int_0^\infty e^{-\beta t} \cdot g(w_t, a_t) dt, \\ & s.t. \text{ for all } t \in [0,\infty): \ w_0 \in [0,1], \ a_t \in [0,1] \text{ and} \\ & w_t = w_0 + \int_0^t f(w_s, a_s) ds,\end{aligned}$$

with

$$g(w,a) = A(w) + B(w)\Psi(a), \qquad f(w,a) = a(w) + b(w)\Psi(a).$$

Using this representation of the optimization problem, the following proposition ensures the sufficient optimality of the MRAP method.

Proposition 4.18 (MRAP Theorem, Feichtinger and Hartl (1986))**.**
Based on the model parameters, define the functions

$$\begin{aligned}M(w) &= A(w) - \frac{a(w)B(w)}{b(w)} = \frac{4w^2 - 3w}{2w - 2}, \\ N(w) &= \frac{B(w)}{b(w)} = \frac{1}{2w-2}\end{aligned}$$

and consider the equation

$$I(w) := \beta N(w) + M_w(w) \stackrel{!}{=} 0. \tag{4.23}$$

Assume that (4.23) *has a unique solution* $\hat{w}$ *in* $[0,1)$ *and that*

$$I(w) \begin{cases} > 0, & 0 \le w < \hat{w}, \\ < 0, & \hat{w} < w < 1. \end{cases}$$

Assume further that for any admissible trajectory $(w_t)_{t\in[0,\infty)}$ *it holds that*

$$\lim_{t\to\infty} e^{-\beta t} \int_{w_t}^{\hat{w}} N(\xi) d\xi \ge 0. \tag{4.24}$$

If for any $w_0 \in [0,1]$ *an MRAP to* $\hat{w}$ *exists, then it is the optimal solution.*

If there exists no admissible solution $\hat{w}$ *of* (4.23) *and if* $I(w) < 0$ *for all* $w \in [0,1)$*, then the MRAP to the state* $w = 0$ *is optimal.*

For a proof, see Satz 3.2 in Feichtinger and Hartl (1986). The proof of Theorem 4.15 now relies on verifying the conditions given in the proposition.

Proof of Theorem 4.15. We show that the conditions from Proposition 4.18 result in precisely the optimal control claimed above. We begin with the investigation of equation (4.23). We obtain

$$\begin{aligned} I(w) &= \beta N(w) + M_w(w) \\ &= \frac{\beta}{2w-2} + \frac{4w^2-8w+3}{2(w-1)^2} \\ &= \frac{\beta(w-1)+4w^2-8w+3}{2(w-1)^2} \overset{!}{=} 0. \end{aligned}$$

Setting the numerator to zero yields the quadratic equation

$$4w^2 + (\beta-8)w + 3 - \beta \overset{!}{=} 0. \tag{4.25}$$

We now distinguish between the two cases for the discount rate β.

1. Case: $\beta > 3$
Consider the numerator of $I(w)$ as a function of w:

$$h : [0,1] \to \mathbb{R}, \qquad h(w) = 4w^2 + (\beta-8)w + 3 - \beta.$$

On the closed interval $[0,1]$, the convex parabola h attains its maximum at the endpoints. We have $h(1) = -1 < 0$, independent of β and $h(0) = 3 - \beta < 0$ for $\beta > 3$. Thus, $h(w)$ and, consequently, $I(w)$ are negative throughout the interval $[0,1)$. Proposition 4.18 then implies that the MRAP to $w = 0$ is optimal, which corresponds to the strategy to never repair anything. Thus, we obtain $a_t^* \equiv 1$.

2. Case: $\beta \in (0,3]$
Consider again the numerator h of $I(w)$ that takes the form of a convex parabola. We have $h(0) = 3 - \beta \geq 0$ for $\beta \in (0,3]$ and $h(1) = -1$ independently of β. Hence, we conclude that h has a unique zero $\hat{w}$ in $[0,1]$. In particular, since $h(\frac{1}{2}) = -\frac{1}{2}\beta$, this zero is located in $[0, \frac{1}{2})$. Moreover, $I(w)$ is positive to the left of $\hat{w}$ and negative to the right of $\hat{w}$. Note that the exact location of $\hat{w}$ depends on the discount rate β. We therefore use the notation $\hat{w}(\beta)$ in the formulation of the theorem.
It remains to show that the "transversality condition" (4.24) is satisfied. Thus, let $(w_t)_{t\in[0,\infty)}$ be an arbitrary admissible state trajectory. As $t \to \infty$, we have that w_t remains bounded within the interval $[0, \frac{2}{3} + \varepsilon]$ for arbitrarily small $\varepsilon > 0$. This is because even if we always repair at full capacity (i.e., $a_t \equiv 0$), the state dynamics are given by

$$w_t' = 2(1 - w_t) - w_t = 2 - 3w_t.$$

The solution to this ODE is $w_t = \frac{2}{3} + \left(w_0 - \frac{2}{3}\right) e^{-3t}$, which converges to $\frac{2}{3}$ as $t \to \infty$.

Now consider the integral

$$\begin{aligned}\int_{w_t}^{\hat{w}} N(\xi)d\xi &= \int_{w_t}^{\hat{w}} \frac{1}{2\xi - 2} d\xi = \left[\frac{1}{2}\ln(1-\xi)\right]_{w_t}^{\hat{w}} \\ &= \frac{1}{2}\left(\ln(1-\hat{w}) - \ln(1-w_t)\right) \\ &= \frac{1}{2}\ln\left(\frac{1-\hat{w}}{1-w_t}\right).\end{aligned}$$

Since w_t is bounded away from 1 for t sufficiently large, the integral $\int_{w_t}^{\hat{w}} N(\xi)d\xi$ is also bounded as $t \to \infty$. Additionally, the exponential discount term tends to zero. Hence, it follows that

$$\lim_{t\to\infty} e^{-\beta t} \int_{w_t}^{\hat{w}} N(\xi)d\xi = 0.$$

Thus, the transversality condition (4.24) is satisfied.

Proposition 4.18 now ensures that an MRAP to the unique solution $\hat{w}$ of (4.25) in $[0,1)$ is optimal. In the case $w_t > \hat{w}$, the fastest possible control to reach $\hat{w}$ is to set $a_t = 1$ and let the machines break down. Conversely, if $w_t < \hat{w}$, choose $a_t = 0$ to increase the fraction of working machines as fast as possible. If the singular state $\hat{w}$ is reached, it is optimal to maintain this level of working machines. To achieve this, the action a_t must be chosen such that the derivative of the state process vanishes, that is

$$0 \overset{!}{=} w_t' = 2(1-w_t)(1-a_t) - w_t \quad \Longleftrightarrow \quad a_t = 1 - \frac{w_t}{2(1-w_t)}$$

Inserting $\hat{w}$ completes the representation of the optimal control claimed in the theorem. Finally, note that since $\hat{w} \in [0, \frac{1}{2})$, we have $a_t^* \in [0,1]$ for all $t \geq 0$, and thus the optimal control is admissible. □

4.4.3. Illustration

For illustration, we consider (MR) for the parameters $\beta = 2$ and $w_0 = 1$. As shown in Section 4.4.2, the control problem, together with the corresponding state differential equation, can be solved explicitly. For this specific choice of parameters, the optimal singular state is determined as the solution in $[0,1]$ of the quadratic equation

$$4w^2 - 6w + 1 = 0.$$

We obtain $\hat{w}(2) = \frac{1}{4}(3 - \sqrt{5}) \approx 0.191$. According to Theorem 4.15, the optimal control is given by

$$a_t^* = \begin{cases} 1, & w_t > \hat{w}(\beta), \\ 1 - \frac{\hat{w}(\beta)}{2(1-\hat{w}(\beta))} \approx 0.882, & w_t = \hat{w}(\beta). \end{cases}$$

The optimal state-control trajectory evolves as follows: Starting from the initial state $w_0 = 1$, it is optimal to set $a_t^* = 1$, allowing the machines to decay. In this first phase,

solving the state equation (4.21) yields $w_t^* = e^{-t}$. The singular state $\hat{w}$ is reached at $\hat{t} = -\ln\left(\frac{1}{4}(3-\sqrt{5})\right) \approx 1.656$. Thereafter, it is optimal to maintain the singular level by choosing $a_t^* \approx 0.882$. Recall that Theorem 3.17 ensures that $(a_t^*)_{t\in[0,\infty)}$ is asymptotically optimal for the N-machine problem.

Figure 4.9 visualizes the optimal deterministic state trajectory (w_t^*) and compares it to simulated paths for various numbers of machines N under an open-loop implementation of the optimal control (a_t^*). The convergence of the state processes can be clearly anticipated. Several simulations for $N = 1000$ machines are shown in Figure 4.10.

Note that the representation of the optimal control in Theorem 4.15 can, in principle, also be implemented as a feedback control. However, the induced mapping $w \mapsto a^*(\cdot|w)$ has a jump discontinuity at $\hat{w}(\beta)$, so Corollary 3.20 does not apply. Thus, we use an open-loop implementation here, since we have shown its asymptotic optimality in Theorem 3.17. A detailed discussion on why an open-loop implementation of the optimal control may be preferable to a feedback implementation is presented in Section 4.5.

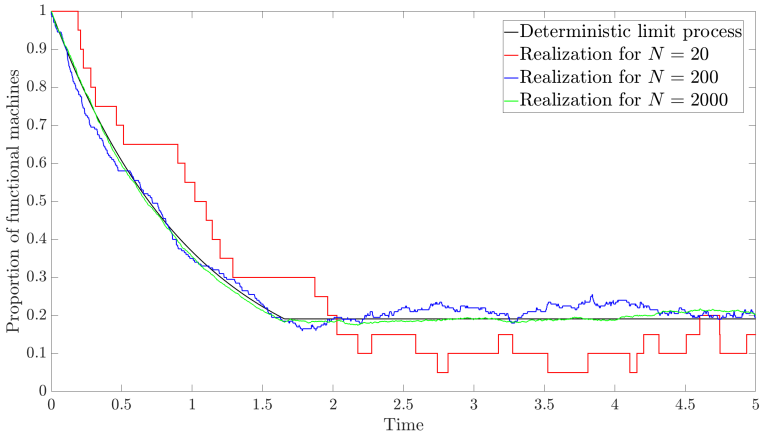

Figure 4.9.: Deterministic state process under optimal control and realizations of the state process for $N = 20,\ 200,\ 2000$ under open-loop implementation of $(a_t^*)_{t\in[0,\infty)}$.

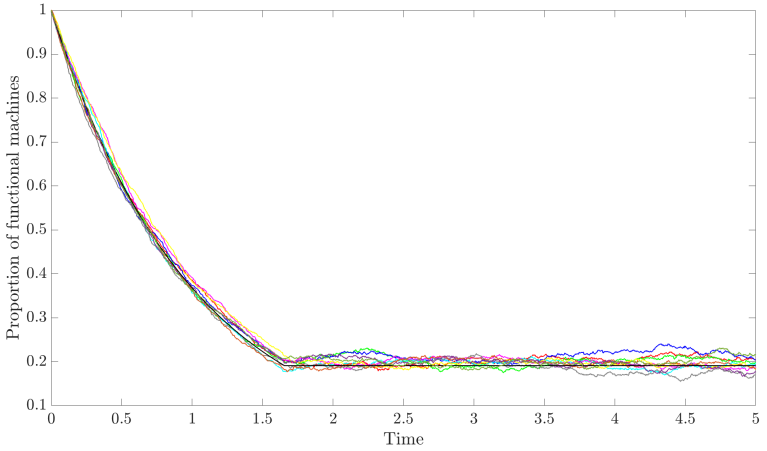

Figure 4.10.: Deterministic state process under optimal control and ten realizations of the state process for $N = 1000$ under open-loop implementation of $(a_t^*)_{t\in[0,\infty)}$.

4.5. Resource competition

In Theorem 3.17 it was stated that an optimal control $(\hat{\pi}_t^*)$ of the deterministic limit model is asymptotically optimal for the N-agent problem when applied in an open-loop manner. More precisely, for the N-agent model, we considered the control

$$\hat{\pi}_t^{N,i} := \hat{\pi}_t^{*,i}, \qquad i \in S,$$

which applies the kernel $\hat{\pi}_t^{*,i}$ at each time $t \geq 0$, regardless of the current state μ_t^N of the N-agent process. Corollary 3.20 ensures the asymptotic optimality of feedback policies in the case where $\mu \mapsto \hat{\pi}^*(\cdot|\mu)$ is continuous with respect to weak convergence.

The following example shows that, in general, optimal feedback policies $\hat{\pi}(\cdot|\mu)$ of the deterministic limit model (F) with fixed initial distribution μ_0 are not necessarily asymptotically optimal when implemented in the N-agent problem. In particular, convergence of the N-agent state processes $(\mu^N)_{N\in\mathbb{N}}$ to the optimal deterministic state process μ^* as well as convergence of the value functions $V^N_{\hat{\pi}(\cdot|\mu)}$ to the optimal value V^F of the limit problem is not guaranteed.

The example is an adaptation of the queueing network considered in Kumar and Seidman (1989) and Rybko and Stolyar (1992) to our setting. The Rybko-Stolyar network is an example of an unstable subcritical network, in the sense that the number of jobs in the network tends to infinity as $t \to \infty$, although the arrival rates of jobs at each service station are strictly less than the service rates. For a detailed discussion of instability in subcritical queueing networks, see Chapter 3 in Bramson (2008).

4.5.1. Description of the N-agent model

Suppose we have a network of eight queues $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$, consisting of two lines representing queues 1 to 4 (upper line) and queues 5 to 8 (lower line). The set of queues S corresponds to the basic state space in our mean-field MDP context. The aim of the network is to process a total of N jobs, where jobs of type 1 pass through the upper line and jobs of type 2 pass through the lower line. Queues 2 and 7 share server A, while queues 3 and 6 share server B. The action of the central controller is to determine which queue the servers A and B work on. Thus, the basic action space is $A = \{0, 1\}$, where for server A the action $a = 1$ symbolizes processing jobs of type 1 in the upper line. In this case, the transition intensity at queue 2 is fully activated, while $a = 0$ activates the intensity at queue 7. For queues 3 and 6, the control mechanism is the same for server B.

The path of a job of type 1 in the upper line is now as follows: The intensity for leaving the initial queue 1 is $\lambda_1 = 1$, the full intensity for leaving queue 2 is $\lambda_2 = 1$ when processed by server A, and finally, the full intensity for leaving queue 3 is $\lambda_3 = 1$ when processed by server B. The dynamics of jobs in the lower line are analogous, with $\lambda_5 = \lambda_6 = \lambda_7 = 1$. Queues 4 and 8 are absorbing and collect the fully processed jobs.

The joint process $\mu_t^N = (\mu_t^N(1), \dots, \mu_t^N(8))$ is a controlled continuous-time Markov chain, where $\mu_t^N(i)$ represents the fraction of jobs in queue $i \in S$. A natural choice for an initial configuration is $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$.

In this model, we assume that servers A and B can split their capacity between queues 2 and 7, resp. 3 and 6. There is a resource constraint in the sense that $\hat{\pi}^2(\{1\} \mid \mu_t^N) + \hat{\pi}^7(\{1\} \mid \mu_t^N) \leq 1$ for server A and $\hat{\pi}^3(\{1\} \mid \mu_t^N) + \hat{\pi}^6(\{1\} \mid \mu_t^N) \leq 1$ for server B, meaning that in the N-agent mean-field network, the sum of the activation probabilities in queues 2 and 7 as well as the sum of the activation probabilities in queues 3 and 6 is constrained by 1, see Section 3.4.2. W.l.o.g., we impose $\hat{\pi}^2(\{1\} \mid \mu_t^N) + \hat{\pi}^7(\{1\} \mid \mu_t^N) = 1$ and $\hat{\pi}^3(\{1\} \mid \mu_t^N) + \hat{\pi}^6(\{1\} \mid \mu_t^N) = 1$, since any inequality would amount to idling, which cannot increase the value. Thus, we are able to introduce the following simplifying notation

$$\begin{aligned} \hat{\pi}^2(\{1\} \mid \mu_t^N) &= a_t, & \hat{\pi}^3(\{1\} \mid \mu_t^N) &= 1 - b_t, \\ \hat{\pi}^7(\{1\} \mid \mu_t^N) &= 1 - a_t, & \hat{\pi}^6(\{1\} \mid \mu_t^N) &= b_t, \end{aligned}$$

where we have $a_t, b_t \in [0, 1]$ for every $t \in [0, T]$. The control region of the deterministic limit model is accordingly $A_{Det} = [0, 1]^2$.

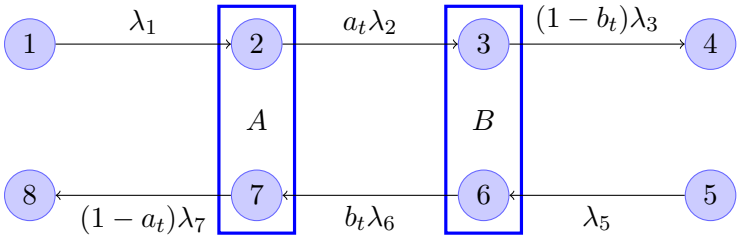

Figure 4.11.: Transition intensities of one agent in the N-agent mean-field network.

To summarize, intensities are given by

$$\begin{aligned} q(\{2\}|1, \cdot\,, \mu_t^N) &= 1, & q(\{6\}|5, \cdot\,, \mu_t^N) &= 1, \\ q(\{3\}|2, a_t, \mu_t^N) &= a_t, & q(\{4\}|3, b_t, \mu_t^N) &= 1 - b_t, \\ q(\{7\}|6, b_t, \mu_t^N) &= b_t, & q(\{8\}|7, a_t, \mu_t^N) &= 1 - a_t. \end{aligned}$$

All other intensities (besides the intensities of stay) are zero. Note that (Q1)-(Q5) and (Q4') are satisfied.

Assume that the objective of the central controller is to process as many jobs as possible into the absorbing final queues 4 and 8. This intuition results in the following terminal reward function:

$$g(\mu_T^N) = \min\{\mu_T^N(4), 0.4\} + \min\{\mu_T^N(8), 0.4\}.$$

To ease the analysis of optimality, the terminal reward is bounded by the fraction of 0.4 in every line of the network. Every control that manages to process a sufficient proportion of

jobs into the absorbing queues 4 and 8 generates the maximum terminal reward.

Further, assume that jobs in queues 3 and 7 generate costs whenever at least one percent of the jobs are present in these queues. The reward function for one job is therefore given by

$$r(3,\cdot,\mu_t^N) = -\mathbb{1}_{\{\mu_t^N(3)>0.01\}} \cdot (\mu_t^N(3) - 0.01),$$
$$r(7,\cdot,\mu_t^N) = -\mathbb{1}_{\{\mu_t^N(7)>0.01\}} \cdot (\mu_t^N(7) - 0.01).$$

In every other queue, no reward or cost is generated. Note that the reward functions are chosen in such a way that they are continuous in $\mu(3)$ resp. $\mu(7)$. Thus, (R1) and (R2) are satisfied.

4.5.2. Deterministic limit model

We now let the number of jobs in the network N tend to infinity, and obtain the finite-horizon deterministic limit problem (F_T), see (3.42). In the following, we denote the limit state process by $\mu_t = (\mu_t(1), \dots, \mu_t(8))$ and the control by $(a_t, b_t)_{t\in[0,T]}$. For the sequence of initial states, we require $\mu_0^N \Rightarrow \mu_0$ for a distribution $\mu_0 \in \mathbb{P}(S)$. As the time horizon, we choose $T = 45$. The result is the following deterministic optimization problem:

$$\begin{aligned}
(RC) \qquad & \sup_{(a,b)} g(\mu_T) + \int_0^{45} \mu_s(3)\cdot r(3,\cdot,\mu_s) + \mu_s(7)\cdot r(7,\cdot,\mu_s)ds, \\
& s.t.\ a_t, b_t \in [0,1], \quad \mu_0 \in \mathbb{P}(S) \text{ and for all } t \in [0,45] \\
& \qquad \mu_t(1) = \mu_0(1) + \int_0^t -\mu_s(1)ds \\
& \qquad \mu_t(2) = \mu_0(2) + \int_0^t \mu_s(1) - a_s\mu_s(2)ds \\
& \qquad \mu_t(3) = \mu_0(3) + \int_0^t a_s\mu_s(2) - (1-b_s)\mu_s(3)ds \\
& \qquad \mu_t(4) = \mu_0(4) + \int_0^t (1-b_s)\mu_s(3)ds \\
& \qquad \mu_t(5) = \mu_0(5) + \int_0^t -\mu_s(5)ds \\
& \qquad \mu_t(6) = \mu_0(6) + \int_0^t \mu_s(5) - b_s\mu_s(6)ds \\
& \qquad \mu_t(7) = \mu_0(7) + \int_0^t b_s\mu_s(6) - (1-a_s)\mu_s(7)ds \\
& \qquad \mu_t(8) = \mu_0(8) + \int_0^t (1-a_s)\mu_s(7)ds
\end{aligned}$$

Since the state process forms a distribution for each point in time $t \in [0,45]$, one of the state equations above can be omitted.

Now suppose that the initial distribution of jobs is given by the natural choice $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$. We consider the following heuristic feedback control, in which the servers give priority to the exit queues 3 and 7 whenever the fractions are above the

threshold of 0.01. Otherwise, queues 2 and 6 receive priority. If the threshold is reached exactly, a special rule is applied.

$$a_t^{pr} := \begin{cases} 0, & \mu_t(7) > 0.01 \\ c_t & \mu_t(7) = 0.01 \\ 1, & \mu_t(7) < 0.01 \end{cases} \qquad b_t^{pr} := \begin{cases} 0, & \mu_t(3) > 0.01 \\ c_t & \mu_t(3) = 0.01 \\ 1, & \mu_t(3) < 0.01 \end{cases}$$

Here, the parameter c_t is chosen in such a way that the inflow in queues 3 and 7 corresponds to the outflow of these queues. This allows us to maintain the fraction of jobs exactly at the threshold of 0.01. Since the evolution of the state process in the deterministic limit model is symmetric, we do not need to differentiate the parameter c_t for the two different servers.

Figure 4.12 displays the control (a_t^{pr}, b_t^{pr}) resulting from applying the heuristic priority-rule feedback control in the deterministic limit model for the initial distribution $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$. As already mentioned, the upper and the lower line are symmetric, and consequently, we have $a_t^{pr} = b_t^{pr}$ for every $t \in [0, 45]$. Note that for a short period of time in the beginning ($t \in [0, 0.2147)$), it is optimal to perform $a_t^{pr} = b_t^{pr} = 1$ until the thresholds of 0.01 in queues 3 and 7 are reached. From then on, a_t^{pr} and b_t^{pr} are chosen such that the threshold level is preserved in queues 3 and 7.

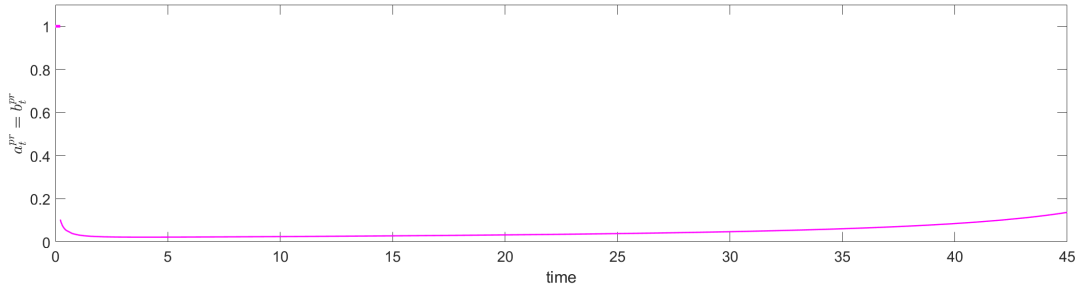

Figure 4.12.: "Priority-rule" control $(a_t^{pr}, b_t^{pr})_{t\in[0,45]}$ of the deterministic limit model.

Why is this priority rule optimal for the deterministic limit model with initial distribution $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$? Clearly, the maximum possible reward that can be achieved is 0.8. This is the case when the fraction of jobs in queues 3 and 7 is kept below the threshold throughout the planning horizon and additionally, at the terminal time $T = 45$ the fraction of completely processed jobs in queues 4 and 8 exceeds the reward limit of 0.4. The priority rule enforces these properties of the state trajectories and thus yields the maximum reward (see Figure 4.13).

Theorem 3.17 now establishes the asymptotic optimality of the control $(a_t^{pr}, b_t^{pr})_{t\in[0,45]}$ when applied in an open-loop manner to the N-agent model with initial distribution $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$. Figure 4.13 shows the state process of the deterministic limit model under the optimal control $(a_t^{pr}, b_t^{pr})_{t\in[0,45]}$, as well as a realization of the corresponding

N-agent state process for $N = 1000$ jobs. The convergence of the state processes can already be clearly anticipated.

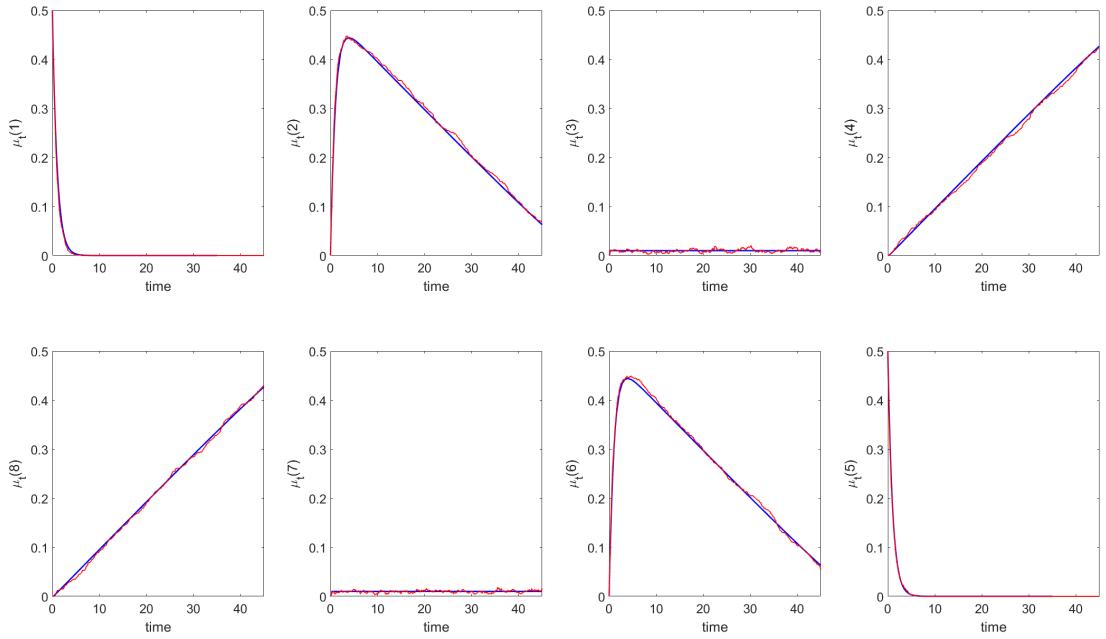

Figure 4.13.: Deterministic state process under optimal control (blue), Realization of the state process for $N = 1000$ jobs under open-loop implementation of the priority rule $(a_t^{pr}, b_t^{pr})_{t\in[0,45]}$ (red).

4.5.3. Non-optimality of the priority-rule feedback control in the N-agent model

The priority-rule feedback control can easily be implemented in the N-agent model. A major observation is that the threshold of 0.01 cannot be attained exactly for $\mu_t^N(3)$ and $\mu_t^N(7)$ if the number of jobs N is not divisible by 100. Consequently, the parameter c_t is irrelevant for implementing the control for such N, and we have $a_t^{pr,N}, b_t^{pr,N} \in \{0, 1\}$ for all $t \in [0, T]$ in the N-agent model.

As it turns out, the nature of the state processes in the N-agent model is fundamentally different from that of the deterministic limit model. Figure 4.14 depicts a realization of the state process for $N = 3001$ jobs. The evolution over time is as follows: For a short period of time in the beginning, we have $\mu_t^N(3) < 0.01$ and $\mu_t^N(7) < 0.01$. Thus, $a_t^{pr,N} = b_t^{pr,N} = 1$ is executed, which results in an increase in the fractions of jobs in queues 3 and 7. When the fraction in one of the two lines exceeds the threshold of 0.01, the priority of the respective server switches, while the other server continues to operate as before. W.l.o.g. assume that this occurs in the upper line (as realized in Figure 4.14), i.e., $a_t^{pr,N} = 1$ and $b_t^{pr,N} = 0$. Then the inflow of jobs in queue 3 is still greater than the outflow, causing congestion of jobs in queue 3 that generates cost. The distribution of jobs in the lower line remains constant.

Only when the majority of jobs in the upper line are processed to queue 4 and the fraction of jobs in queue 3 falls below the threshold, the priority of the system is reversed and the jobs in the lower line are processed, causing again congestion of jobs in queue 7 that generates cost, until finally, the majority of jobs in the lower line are processed to queue 8.

Surprisingly, when applied to the N-agent model, the priority-rule processes jobs to the absorbing queues more quickly than the deterministic limit model. This effect arises because the outflow of queues 3 and 7 is proportional to the fraction of jobs in these queues, which is typically higher in the N-agent model due to the congestion. However, this faster processing is not reflected in a higher reward.

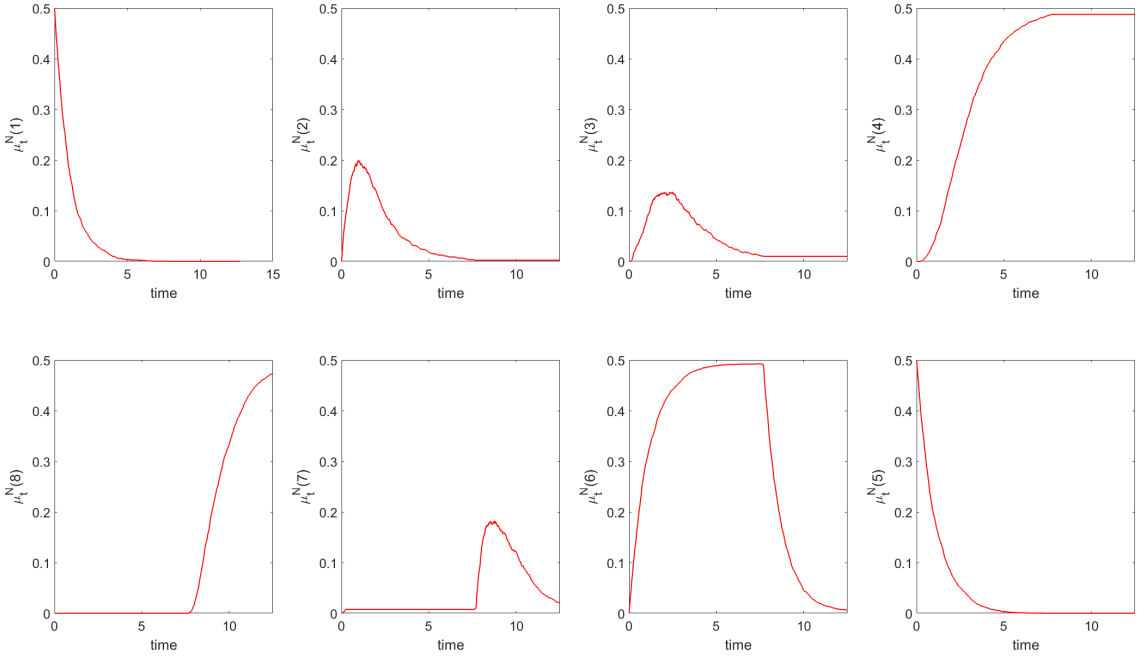

Figure 4.14.: Realization of the state process for $N = 3001$ jobs under the priority-rule feedback control.

Note that the evolution of the N-agent system strongly depends on which threshold is exceeded first — that of the upper line in queue 3 or the lower line in queue 7. This determines which line is emptied first. Figure 4.15 shows eight different realizations of the state process for $N = 2001$ jobs obtained by applying the priority rule. In four realizations, the upper line is emptied first, while in the other four, the lower line is emptied first.

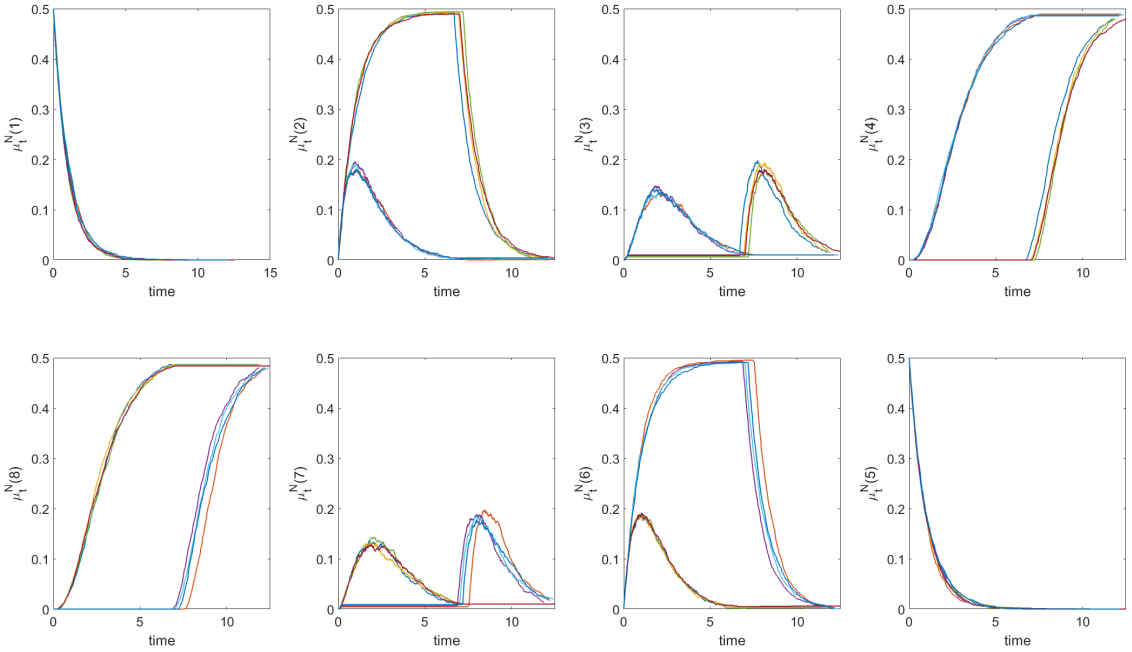

Figure 4.15.: Eight realizations of the state process for $N = 2001$ jobs under the priority-rule feedback control.

The implementation of the priority-rule feedback control in the N-agent model reveals several interesting results. First, we find that, although the priority is optimal for problem (RC) at least for the initial distribution $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$ as it attains the maximum possible reward, the priority rule is not asymptotically optimal for the N-agent model, since it creates congestion in queues 3 and 7 that generates cost. We also conclude that convergence of the state processes does not generally occur for arbitrary feedback rules. The intuition behind this observation is quite obvious: the N-agent model may be found in states that cannot be reached in the deterministic limit model. In this example, these are queues where $\mu_t^N(3) > 0.01$ or $\mu_t^N(7) > 0.01$. In such states that are unreachable for the limit model, the feedback control may trigger a behavior of the N-agent state process fundamentally different from the deterministic limit state process.

4.5.4. An asymptotically optimal feedback control for the N-agent model

The aim of this section is to refine the priority-rule feedback control such that it becomes asymptotically optimal when applied to the N-agent model. The main idea is to interrupt the inflow of jobs in queues 3 and 7 whenever the fraction of jobs in these queues is above the threshold. This leads to the following structure for the refined feedback control:

i) Case $\mu_t(3) < 0.01$ and $\mu_t(7) < 0.01$: Build up the fraction of jobs in queues 3 and 7 up to the threshold of 0.01.

$$a_t = 1, \qquad b_t = 1$$

ii) Case $\mu_t(3) > 0.01$ and $\mu_t(7) > 0.01$: Reduce the fraction of jobs in queues 3 and 7 down to the threshold of 0.01.

$$a_t = 0, \qquad b_t = 0$$

iii) Case $\mu_t(3) > 0.01$ and $\mu_t(7) \leq 0.01$: Reduce the fraction of jobs in queue 3 down to the threshold of 0.01. During this process the fraction of jobs in queue 7 is reduced as well.

$$a_t = 0, \qquad b_t = 0$$

iv) Case $\mu_t(3) \leq 0.01$ and $\mu_t(7) > 0.01$: Reduce the fraction of jobs in queue 7 down to the threshold of 0.01. The fraction of jobs in queue 3 is reduced as well.

$$a_t = 0, \qquad b_t = 0$$

v) Case $\mu_t(3) = 0.01$ and $\mu_t(7) = 0.01$: Adjust the control parameters in such a way that the inflow in queues 3 and 7 corresponds to the outflow of these queues. This allows us to maintain the fraction of jobs exactly at the threshold of 0.01.

$$a_t = c_t, \qquad b_t = c_t$$

vi) The two cases in which one of the queues is at the threshold $\mu_t(\cdot) = 0.01$ and the other queue is below the threshold are combinations of cases i) and v). These cases are of minor practical relevance.

Why is this feedback control optimal for the deterministic limit model? For the initial state $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$, the refined control is equivalent to the priority-rule feedback control in the sense that it induces the same state (and action) process and therefore generates the maximum reward. Note that $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$ is the least favorable initial configuration for processing as many jobs as possible to queues 4 and 8 until the end of the planning horizon. If the control provides $\mu_T(4) > 0.4$ and $\mu_T(8) > 0.4$ in this least favorable case, then this holds for every other initial configuration as well. Additionally, for any initial configuration with $\mu_0(3) > 0.01$ or $\mu_0(7) > 0.01$, which corresponds to the cases ii), iii) or iv), the feedback control applies $a_t = 0$ and $b_t = 0$, which is the fastest strategy to reduce the fraction of jobs in queues 3 and 7 below the threshold of 0.01 and therefore minimizing the costs. Thus, we conclude that the presented feedback control is optimal for every initial configuration for the deterministic limit model.

The refined feedback control described above can easily be implemented in the N-agent model. As for the priority-rule feedback control, the threshold of 0.01 cannot be attained exactly for $\mu_t^N(3)$ and $\mu_t^N(7)$ if the number of jobs N is not divisible by 100. Thus, only the cases i)-iv) are relevant to implement the control for such N. The main difference between the two feedback policies is the behavior when the threshold in queues 3 or 7

is exceeded. The refined control then cuts the inflow in these queues instantly until the surplus has been processed. Consequently, the threshold is exceeded by at most one job, and congestion in queues 3 and 7 is avoided; see Figure 4.16.

By the structure of the reward function, exceeding the threshold by one job at most results in costs that tend to zero as the number of jobs tends to infinity. Additionally, as suggested in Figure 4.16, for N sufficiently large, the fraction of jobs in queues 4 and 8 reliably surpasses the reward limit of 0.4 at $T = 45$. We conclude that the refined feedback control is asymptotically optimal for the N-agent model, generating the maximum possible reward of 0.8 for the initial distribution $\mu_0 = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0)$.

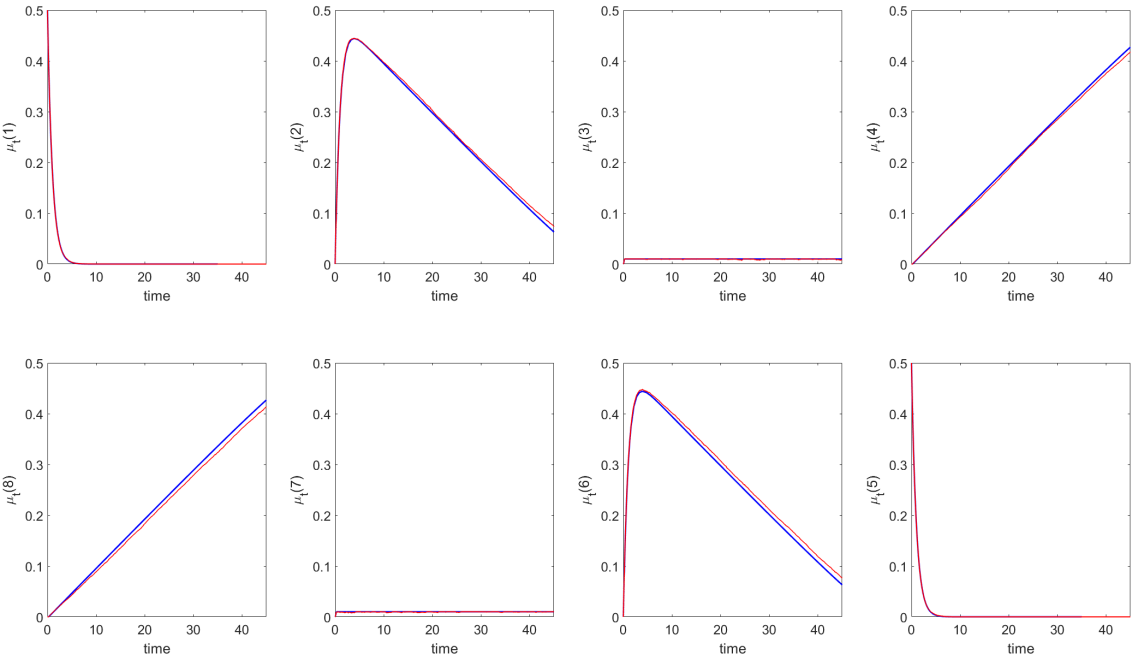

Figure 4.16.: Deterministic state process under optimal feedback control (blue), Realization of the state process for $N = 20001$ jobs under the refined feedback control (red).

APPENDIX A

Miscellaneous

Theorem A.1 (Ionescu-Tulcea)**.** *Let $\{(X_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ be a sequence of measurable spaces, $Y_n := X_1 \times X_2 \times \ldots \times X_n$ and $Y := X_1 \times X_2 \times \ldots$. Let p be a given probability measure on $(X_1, \mathcal{F}_1)$, and, for $n = 1, 2, \ldots$, let $q_n(dx_{n+1}|y_n)$ be a stochastic kernel on X_{n+1} given $y_n \in Y_n$. Then for each $n = 2, 3, \ldots$, there exists a unique probability measure μ_n on $(Y_n, \prod_{i=1}^n \mathcal{F}_i)$ with $\prod_{i=1}^n \mathcal{F}_i$ being the product σ-algebra of $\{\mathcal{F}_i\}_{i=1}^n$ such that, for all $\Gamma_1 \in \mathcal{F}_1, \ldots, \Gamma_n \in \mathcal{F}_n$,*

$$\mu_n(\Gamma_1 \times \cdots \times \Gamma_n) = \int_{\Gamma_1} \int_{\Gamma_2} \cdots \int_{\Gamma_{n-1}} q_{n-1}(\Gamma_n|x_1, \ldots, x_{n-1})\, q_{n-2}(dx_{n-1}|x_1, \ldots, x_{n-2}) \cdots \\ \cdots q_1(dx_2|x_1)\, p(dx_1).$$

If $f : Y_n \to [0, \infty]$ is a measurable function, then

$$\int_{Y_n} f(y_n) d\mu_n(y_n) = \int_{X_1} \int_{X_2} \cdots \int_{X_n} f(x_1, x_2, \ldots, x_n) q_{n-1}(dx_n|x_1, \ldots, x_{n-1}) \cdots \\ \cdots q_1(dx_2|x_1) p(dx_1).$$

Furthermore, there exists a unique probability measure μ on $(Y, \prod_{i=1}^{\infty} \mathcal{F}_i)$ with $\prod_{i=1}^{\infty} \mathcal{F}_i$ being the product σ-algebra of $\{\mathcal{F}_i\}_{i=1}^{\infty}$ such that, for each n, the marginal of μ on Y_n is μ_n.

This version of the theorem can be found in Proposition B.1.37 in Piunovskiy and Zhang (2020). For a proof, see, e.g., Theorem 2.7.2 in Ash and Doléans-Dade (2000) and Proposition 7.28 in Bertsekas and Shreve (1978).

Definition A.2 (Upper semicontinuous function)**.** Let X be a topological space and c be a $[-\infty, \infty]$-valued function on X. Then c is called *upper semicontinuous* on X if $\{x \in X \mid c(x) \geq \varepsilon\}$ is closed in X for each constant $\varepsilon \in \mathbb{R}$.

Theorem A.3 (Tychonoff)**.** *Let $(M_i, \tau_i)_{i \in I}$ be a family of compact topological spaces. Then, the product $\prod_{i \in I} M_i$, equipped with the product topology, is compact.*

For reference, see, e.g., Theorem 17.8 in Willard (1970)

Lemma A.4. *For a compact Borel space M, the set of all probability distributions $\mathbb{P}(M)$ on M, endowed with the topology of weak convergence, is a compact Polish space.*

For reference, see Proposition A.8 in Lange (2017).

Lemma A.5. *Let X be a separable metric space, Y a compact metric space and $f : X \times Y \to \mathbb{R}$ continuous. Then $x_n \to x$ for $n \to \infty$ implies*

$$\lim_{n \to \infty} \sup_{y \in Y} |f(x_n, y) - f(x, y)| = 0.$$

For a proof, see, e.g., Lemma B.12, Lange (2017).

Theorem A.6 (Jensen)**.** *Let $I \subset \mathbb{R}$ be an interval, $\varphi : I \to \mathbb{R}$, and let X be an I-valued random variable with $\mathbb{E}[|X|] < \infty$.*

a) If φ is concave, then

$$\mathbb{E}[\varphi(X)] \leq \varphi(\mathbb{E}[X]).$$

b) If φ is convex, then

$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]).$$

c) If φ is affine, then

$$\mathbb{E}[\varphi(X)] = \varphi(\mathbb{E}[X]).$$

The affine case follows from a) and b). For a proof of the convex case, see Theorem 7.9 in Klenke (2020).

Theorem A.7 (Chebyshev)**.** *Let X be a real random variable with $\mathbb{E}[X^2] < \infty$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

For a proof, see Theorem 5.11 in Klenke (2020).

Lemma A.8 (Gronwall's inequality)**.** *Let $f : [0, b] \to \mathbb{R}$ be Lebesgue-integrable and $A \geq 0$, $C > 0$ constants such that for all $t \in [0, b]$*

$$0 \leq f(t) \leq A + C \int_0^t f(s)ds.$$

Then the following inequality holds for all $t \in [0, b]$:

$$f(t) \leq Ae^{Ct}.$$

See Gronwall-Lemma in Walz (2017).

Lemma A.9 (Characterization of the total variation distance)**.** *Let μ and ν be probability measures on a finite set S. The total variation distance between μ and ν is given by*

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{i \in S} |\mu(i) - \nu(i)|.$$

For a proof, see, e.g., Proposition 4.2 in Levin and Peres (2017).

Lemma A.10. *Let S be a finite set. Then $\mathbb{P}(S)$ is separable and complete with respect to the total variation distance.*

Proof. For finite S with $|S| = m$, identify $\mathbb{P}(S)$ with the standard probability simplex $\Delta^{m-1} = \{x \in \mathbb{R}^m : x_i \geq 0,\ \sum_{i=1}^m x_i = 1\}$. On a finite state space, $\|\mu - \nu\|_{TV} = \frac{1}{2}\|\mu - \nu\|_1$, so $(\mathbb{P}(S), \|\cdot\|_{TV})$ is (up to a constant factor) the metric subspace $(\Delta^{m-1}, \|\cdot\|_1)$ of $(\mathbb{R}^m, \|\cdot\|_1)$. The simplex Δ^{m-1} is closed and bounded in $\mathbb{R}^m$, hence compact (Heine–Borel); compact metric spaces are complete and separable. □

A.1. Martingales and quadratic variation

In what follows, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space with filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geq 0}$.

Definition A.11 (Martingale)**.** A real-valued stochastic process $(M_t)_{t \geq 0}$ is an *$\mathcal{F}$-martingale* if it is $\mathcal{F}$-adapted, satisfies $\mathbb{E}[|M_t|] < \infty$ for all $t \geq 0$, and for all $0 \leq s \leq t$,

$$\mathbb{E}[M_t \mid \mathcal{F}_s] = M_s, \qquad \mathbb{P}\text{-a.s..}$$

Definition A.12 (Total variation)**.** Let $f : [0, \infty) \to \mathbb{R}$ and $[a, b] \subset [0, \infty)$. The *total variation* of f on $[a, b]$ is

$$\mathrm{TV}_{[a,b]}(f) := \sup_{\mathcal{Z}} \sum_{i=1}^{n} |f(t_i) - f(t_{i-1})|,$$

where the supremum is taken over all finite partitions $\mathcal{Z} = \{a = t_0 < t_1 < \cdots < t_n = b\}$ of $[a, b]$.

Lemma A.13. *Let $f : [0,\infty) \to \mathbb{R}$ be càdlàg with, on every finite interval $[a,b] \subset [0,\infty)$, only finitely many jumps, each of finite height. Assume further that f is linear between the jumps. Then f has finite total variation on $[a,b]$.*

Proof. Fix $[a,b] \subset [0,\infty)$ and assume there is exactly one jump at $t^* \in (a,b)$. If there is more than one jump in $[a,b]$, the reasoning remains the same, but the notation becomes more complicated. Let the slopes on (a,t^*) and (t^*,b) be m^- and m^+, and set $m := \max\{|m^-|,|m^+|\}$. Let $\mathcal{Z} = \{a = t_0 < \ldots < t_n = b\}$ be any partition, and denote by $t_-^* := \max\{t_i \le t^*\}$ and $t_+^* := \min\{t_i \ge t^*\}$ the closest partition points bracketing t^*. With $h := |f(t^*) - \lim_{s\uparrow t^*} f(s)|$ we have

$$\begin{aligned}\sum_{i=1}^{n} |f(t_i) - f(t_{i-1})| &= |m^-| \cdot (t_-^* - a) + |f(t_-^*) - f(t_+^*)| + |m^+| \cdot (b - t_+^*) \\ &\le |m^-| \cdot (t_-^* - a) + m \cdot (t_+^* - t_-^*) + h + |m^+| \cdot (b - t_+^*) \\ &\le m \cdot (b-a) + h\end{aligned}$$

Since the bound is independent of the partition, we obtain

$$\mathrm{TV}_{[a,b]}(f) \le m \cdot (b-a) + h < \infty.$$

□

Theorem A.14 (Characterization of the Quadratic Variation)**.** *Let $(M_t)_{t\ge 0}$ be an $\mathcal{F}$-martingale.*

a) If $(M_t)_{t\ge 0}$ is càdlàg with paths of finite variation on compacts, then its quadratic variation is given by

$$[M,M]_t = M_0^2 + \sum_{0<s\le t} (M_s - M_{s-})^2.$$

In this case $(M_t)_{t\ge 0}$ is called quadratic pure jump.

b) If $\mathbb{E}[[M,M]_t] < \infty$, then

$$\mathbb{E}[M_t^2] = \mathbb{E}[[M,M]_t].$$

For a proof of part a) see Theorem 26 in Protter (2005), Section II.6. For part b) see Corollary 3 in Protter (2005), Section II.6.

APPENDIX B

Convergence of stochastic processes

The aim of this chapter is to give a rigorous characterization of the convergence of the state and action processes, as well as to collect additional results related to the convergence results in Chapter 3.

As mentioned in Section 3.1, the paths of the state process are piecewise constant, right-continuous with existing left limits. Therefore, we consider the state process (3.6) as a stochastic element of $D_{\mathbb{P}(S)}[0,\infty)$, the space of càdlàg functions with values in $\mathbb{P}(S)$. Section B.3 provides the fundamental definitions and results needed to characterize and study convergence in $D_{\mathbb{P}(S)}[0,\infty)$.

In contrast, by the definition of the action process (3.7), its paths are piecewise constant and left-continuous with existing right limits, and thus do not belong to $D_{\mathbb{P}(A)}[0,\infty)$. Instead, we regard the action process $(\hat{\pi}_t)_{t\geq 0}$ as a stochastic element of $\mathcal{R}^{|S|}$, where

$$\mathcal{R} := \{\rho : [0,\infty) \to \mathbb{P}(A) \mid \rho \text{ measurable}\}$$

is the space of measurable functions that assign to each time $t \geq 0$ a probability distribution on the action space. In Section B.2 we introduce a notion of convergence on $\mathcal{R}$ and present additional topological properties.

Section B.1 is devoted to the basic concepts of convergence of random variables and the relations between them.

B.1. CONCEPTS OF CONVERGENCE FOR RANDOM VARIABLES

Definition B.1 (Almost sure convergence, convergence in probability and L^p-convergence)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and (E, d) a separable metric space with associated Borel σ-algebra $\mathcal{B}(E)$. Furthermore, let $X, X_1, X_2, \ldots : \Omega \to E$ be $\mathcal{A}$-$\mathcal{B}(E)$-measurable random variables.

The sequence $(X_n)_{n\in\mathbb{N}}$ converges to X

a) *almost surely*, denoted by $X_n \xrightarrow{a.s.} X$, if there exists a set $N \in \mathcal{A}$ with $\mathbb{P}(N) = 0$ such that
$$d(X(\omega), X_n(\omega)) \overset{n\to\infty}{\longrightarrow} 0, \quad \forall\, \omega \in \Omega \setminus N.$$

b) *in probability*, denoted by $X_n \xrightarrow{\mathbb{P}} X$, if
$$\mathbb{P}(d(X, X_n) > \varepsilon) \overset{n\to\infty}{\longrightarrow} 0, \quad \forall\, \varepsilon > 0.$$

Now, let $(E, d) = (\mathbb{R}, |\cdot|)$ be the Euclidean space and $p \geq 1$ a real number. Further assume that the p-th moments exist, i.e., $\mathbb{E}[|X_n|^p] < \infty$ for all $n \in \mathbb{N}$ and $\mathbb{E}[|X|^p] < \infty$.

The sequence $(X_n)_{n\in\mathbb{N}}$ converges to X

c) *in the p-th mean* or *in* L^p, denoted by $X_n \xrightarrow{L^p} X$, if
$$\mathbb{E}[|X_n - X|^p] \overset{n\to\infty}{\longrightarrow} 0.$$

Definition B.2 (Weak convergence, relative compactness and tightness)**.**
Let (E, d) be a metric space and $\mathbb{P}(E)$ the space of Borel probability measures on E.

a) A sequence $(P_n)_{n\in\mathbb{N}} \subset \mathbb{P}(E)$ converges *weakly* to $P \in \mathbb{P}(E)$, denoted by $P_n \Rightarrow P$, if
$$\int f \, dP_n \overset{n\to\infty}{\longrightarrow} \int f \, dP$$
for all bounded continuous functions $f : E \to \mathbb{R}$.

b) A sequence of E-valued random variables $(X_n)_{n\in\mathbb{N}}$ converges weakly to X ($X_n \Rightarrow X$), if the distributions $(P_{X_n})_{n\in\mathbb{N}}$ converge weakly to P_X.

c) A family of probability measures $\mathcal{P} \subset \mathbb{P}(E)$ is *relatively compact*, if every sequence of elements of $\mathcal{P}$ contains a weakly convergent subsequence.

d) A family of probability measures $\mathcal{P} \subset \mathbb{P}(E)$ is *tight*, if for each $\varepsilon > 0$ there exists a compact set $K \subset E$ such that
$$\inf_{P\in\mathcal{P}} P(K) \geq 1 - \varepsilon.$$

Lemma B.3. *Let $X, X_1, X_2, \dots$ be real random variables. Then the following are equivalent:*

i) $X_n \Rightarrow X$

ii) $\mathbb{E}[f(X_n)] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)]$ *for all continuous and bounded functions* $f : \mathbb{R} \to \mathbb{R}$.

For a proof, see Corollary 13.24 in Klenke (2020).

Proposition B.4 (Properties of the types of convergence)**.**
Let (E, d) be a separable metric space and suppose that $X, X_1, X_2, \dots$ are E-valued random variables defined on the same probability space.

a) The following relations between the types of convergence hold:

$$\begin{aligned} &\textit{i)}\ X_n \xrightarrow{a.s.} X \quad \Longrightarrow \quad X_n \xrightarrow{\mathbb{P}} X, \\ &\textit{ii)}\ X_n \xrightarrow{L^p} X \quad \Longrightarrow \quad X_n \xrightarrow{\mathbb{P}} X, \quad \textit{for any } p \geq 1, \\ &\textit{iii)}\ X_n \xrightarrow{\mathbb{P}} X \quad \Longrightarrow \quad X_n \Rightarrow X. \end{aligned}$$

b) Considering random vectors $X_n := (X_n^{(1)}, \dots, X_n^{(k)})$, $n \in \mathbb{N}$ and $X := (X^{(1)}, \dots, X^{(k)})$, it holds that

$$X_n \xrightarrow{\mathbb{P}} X \quad \Longleftrightarrow \quad X_n^{(j)} \xrightarrow{\mathbb{P}} X^{(j)}, \quad \forall\ j \in \{1, \dots k\}.$$

c) If X is almost surely constant, it holds that

$$X_n \xrightarrow{\mathbb{P}} X \quad \Longleftrightarrow \quad X_n \Rightarrow X.$$

Proofs for part a) can be found in Remark 6.4 in Klenke (2020), Lemma 5.7 and Theorem 5.12 in Kallenberg (2021). For a proof of part b), see Lemma 5.4 in Kallenberg (2021). For part c) see Subsection *Convergence in Probability* in Section 3 in Billingsley (1999).

Theorem B.5 (Continuous mapping theorem)**.**
Let (E_1, d_1) and (E_2, d_2) be metric spaces and let $\varphi : E_1 \to E_2$ be measurable. Denote by U_φ the set of points of discontinuity of φ. If $X, X_1, X_2, \dots$ are E_1-valued random variables with $\mathbb{P}(X \in U_\varphi) = 0$ and $X_n \Rightarrow X$, then $\varphi(X_n) \Rightarrow \varphi(X)$.

For a proof, see Theorem 13.25 in Klenke (2020).

The following well-known representation theorem by Skorokhod establishes a further connection between weak convergence and almost sure convergence. We state the version given in Klenke (2020), Theorem 17.57.

Theorem B.6 (Skorokhod's Representation Theorem)**.**
Let $P, P_1, P_2, \dots$ be probability measures on a Polish space $(E, \mathcal{B}(E))$ with $P_n \Rightarrow P$. Then there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with random variables $X, X_1, X_2, \dots$ with $\mathbb{P}_X = P$ and $\mathbb{P}_{X_n} = P_n$ for every $n \in \mathbb{N}$ such that $X_n \xrightarrow{a.s.} X$.

Theorem B.7 (Prokhorov)**.**
Let (E, d) be a metric space and let $\mathcal{P} \subset \mathbb{P}(E)$ be a family of probability measures. If $\mathcal{P}$ is tight, then $\mathcal{P}$ is relatively compact.

For a proof, see Corollary 3.2.3 in Ethier and Kurtz (1986). If (E, d) is complete and separable, the converse holds as well and tightness and relative compactness are equivalent. This result is known as *Prokhorov's theorem*, see, e.g., Theorem 3.2.2 in Ethier and Kurtz (1986). For our purposes, the above version is sufficient.

Proposition B.8 (Relative compactness in product spaces)**.**
Let (E_k, d_k), $k = 1, 2, \ldots$, be metric spaces, and define the metric space (E, d) by letting

$$E = \prod_{k=1}^{\infty} E_k \quad \text{and} \quad d(x, y) = \sum_{k=1}^{\infty} 2^{-k} \min\{d_k(x_k, y_k), 1\}$$

for all $x, y \in E$. Let $(P_i)_{i \in I} \subset \mathbb{P}(E)$ for some arbitrary index set I. For $k = 1, 2, \ldots$ and each $i \in I$, define $P_i^k \in \mathbb{P}(E_k)$ as the k-th marginal distribution of P_i. Then $(P_i)_{i \in I}$ is tight if and only if $(P_i^k)_{i \in I}$ is tight for $k = 1, 2, \ldots$.

The statement can be found in Ethier and Kurtz (1986), Proposition 3.2.4.

B.2. CONVERGENCE OF THE ACTION PROCESS

The purpose of this section is to characterize convergence in the *space of action processes*

$$\mathcal{R} := \{\rho : [0, \infty) \to \mathbb{P}(A) \mid \rho \text{ measurable}\},$$

the space of measurable functions that assign to each time $t \geq 0$ a probability distribution on the action space A, which we assumed to be a compact Borel space. To specify measurability of an element in $\mathcal{R}$, define for every $B \in \mathcal{B}(A)$ the function

$$m_B : \mathbb{P}(A) \to [0, 1], \qquad \alpha \mapsto \alpha(B).$$

On $\mathbb{P}(A)$, we now consider the initial σ-algebra

$$\sigma_{m_B} := \sigma\left(\bigcup_{B \in \mathcal{B}(A)} m_B^{-1}(\mathcal{B}([0, 1]))\right).$$

In other words, σ_{m_B} is the smallest σ-algebra on $\mathbb{P}(A)$, such that the functions m_B are measurable. Finally, we call $\rho \in \mathcal{R}$ measurable if it is $\mathcal{B}([0, \infty))$-σ_{m_B}- measurable.

To develop a concept of convergence in $\mathcal{R}$ we follow the construction of Davis (1993), Chapter 43. We start with the observation that the space $\mathcal{R}$ is a subset of

$$
\begin{aligned}
X^* &= L^\infty\left([0,\infty), C^*(A)\right) \\
&= \left\{ v : [0,\infty) \to C^*(A) \;\middle|\; \|v\|_* = \operatorname*{ess\,sup}_{t\in[0,\infty)} \|v_t\|_{C^*} < \infty \right\},
\end{aligned}
$$

where

$$
\begin{aligned}
C^*(A) &= \{\nu : \mathcal{B}(A) \to \mathbb{R} \;\mid\; \nu \text{ finite signed measure on } A\}, \\
\|\nu\|_{C^*} &= \nu^+(A) + \nu^-(A).
\end{aligned}
$$

The space X^* is the space of functions that assign to each time $t \geq 0$ a signed measure on the action space. The space of probability measures $\mathbb{P}(A)$ is a subset of the space of signed measures $C^*(A)$. A version of the Markov-Riesz representation theorem, see Theorem VIII.2.26 in Elstrodt (2018), states that $C^*(A)$ is the dual space of the set

$$
C(A) = \{f : A \to \mathbb{R} \mid f \text{ continuous}\}.
$$

Remark B.9. The statement in Elstrodt (2018) is formulated for the class of C^0-functions, the space of functions that *vanish at infinity*. The dual of the space of C^0-functions is the space of finite regular signed measures $\nu : \mathcal{B}(A) \to \mathbb{R}$. In our case, A is a compact Borel space, and thus in particular Polish. Theorem VIII.1.16 together with Conclusion VIII.2.22 in Elstrodt (2018) imply that every signed measure on $\mathcal{B}(A)$ is regular. Additionally, every continuous function with compact domain is necessarily vanishing at infinity. Thus, the version of the Markov-Riesz representation theorem in Elstrodt (2018) is appropriate in our context.

Now, following from the duality of L^p-spaces, $X^* = L^\infty([0,\infty), C^*(A))$ is the dual space of

$$
\begin{aligned}
X &= L^1\left([0,\infty), C(A)\right) \\
&= \left\{ \psi : [0,\infty) \times A \to \mathbb{R} \;\middle|\; \psi \text{ meas. in } t, \text{ cont. in } a \text{ and } \int_0^\infty \max_{a\in A} |\psi(t,a)| dt < \infty \right\},
\end{aligned}
$$

see, e.g., the discussion preceding Proposition 43.3 in Davis (1993).

The *weak* topology on* X^* is defined as the coarsest topology, such that the pairings

$$
\hat{\psi} : X^* \to \mathbb{R}, \quad v \mapsto \langle \psi, v \rangle := \int_0^\infty \int_A \psi(t,a) v_t(da) dt.
$$

are continuous for all $\psi \in X$. The Banach–Alaoglu theorem states that the unit ball

$$
B_1 = \{v \in X^* \mid \|v\|_* \leq 1\} \subset X^*
$$

is compact w.r.t. the weak* topology, see, e.g., Satz 13.9 in Hirzebruch and Scharlau (1991).

Since an element $\rho \in \mathcal{R}$ assigns a probability measure to each point in time, we have $\|\rho\|_* = 1$. Consequently, the space $\mathcal{R}$ is a subset of B_1.

Theorem B.10. *The space of action processes $\mathcal{R}$ is compact.*

Since closed subsets of compact sets are compact as well, see Theorem (11.3) in Davis (1993), it is sufficient to show that $\mathcal{R} \subset B_1$ is closed. For a proof, see Proposition (43.3) in Davis (1993). The compactness of the space $\mathcal{R}^{|S|}$ follows from Theorem A.3.

Having ensured that the limit of a sequence of elements of $\mathcal{R}$ is again an element of $\mathcal{R}$, we now introduce the notion of convergence considered for action processes in Section 3.3.

Theorem B.11 (Topological properties of $\mathcal{R}$)**.**

a) A sequence $(\rho^n)_{n\in\mathbb{N}} \subset \mathcal{R}$ converges to $\rho \in \mathcal{R}$ for $n \to \infty$ in the weak topology if and only if*

$$\int_0^\infty \int_A \psi(t,a)\rho_t^n(da)dt \;\longrightarrow\; \int_0^\infty \int_A \psi(t,a)\rho_t(da)dt$$

for all $\psi \in X$.

b) The space $\mathcal{R}$ is metrizable in the weak topology.*

For a proof of the characterization of convergence in $\mathcal{R}$, see Lemma 13.8 and Lemma 13.2 in Hirzebruch and Scharlau (1991). Note that the proof relies solely on topological arguments and does not require a metric. Nevertheless, we state that $\mathcal{R}$ is metrizable in order to apply Theorem B.8 at some point. For reference, see Lemma B.3 in Forwick (1997).

Remark B.12. The proposed derivation of the convergence of action processes based on the weak* topology is common in control theory. For a broader overview, see, e.g., Chapter 43 in Davis (1993), Chapter III.3.B in Warga (1972), or Appendix B in Forwick (1997). The concept was introduced by Young (1937) in his work on the calculus of variations. Therefore, the corresponding topology bears his name.

Definition B.13 (Young topology)**.**
The *Young topology* on $\mathcal{R}$ is the relative weak* topology of $\mathcal{R}$ considered as a subset of B_1.

Remark B.14.

a) The notion of convergence in Theorem B.11 corresponds to the pointwise convergence of the pairings $\langle \psi, v_n \rangle \to \langle \psi, v \rangle$ for every $\psi \in X$. For this reason, the literature also refers to the weak* topology resp. the Young topology as the *topology of pointwise convergence*.

b) An element $\rho \in \mathcal{R}$ can be interpreted as a stochastic kernel, since

   i) ρ_t is a probability measure on $\mathcal{B}(A)$ for every $t \in [0,\infty)$,

   ii) $t \mapsto \rho_t(B)$ is measurable for every $B \in \mathcal{B}(A)$.

   The second property directly follows from the fact that $t \mapsto \rho_t(B)$ is simply the composition $m_B \circ \rho$, where m_B and ρ are both measurable functions.

B.3. Convergence of the state process

In what follows, let (E, d) be a metric space and $d_{\min} := \min\{d, 1\}$. For the investigation of the state process, we rely on the metric space $(\mathbb{P}(S), \|\cdot\|_{TV})$.

Definition B.15 (The space $D_E[0,\infty)$)**.**

a) A function $f : [0,\infty) \to E$ is called *càglàd (continue à gauche, limite à droite)* if

- f is left-continuous, i.e., $\lim_{s\to t^-} f(s) = f(t)$ for $t \in (0,\infty)$, and
- the right limits exist, i.e., $\lim_{s\to t^+} f(s) =: f(t^+)$ exists for $t \in [0,\infty)$.

b) A function $f : [0,\infty) \to E$ is called *càdlàg (continue à droite, limite à gauche)* if

- f is right-continuous, i.e., $\lim_{s\to t^+} f(s) = f(t)$ for $t \in [0,\infty)$, and
- the left limits exist, i.e., $\lim_{s\to t^-} f(s) =: f(t^-)$ exists for $t \in (0,\infty)$.

c) The set $D_E[0,\infty)$ is the space of all càdlàg functions $f : [0,\infty) \to E$.

In order to study convergence on the space of càdlàg functions, we introduce a metric to measure distances between elements of $D_E[0,\infty)$.

Definition B.16 (The Skorokhod metric d_{J_1})**.**

a) Let

$$\Lambda' := \{\lambda : [0,\infty) \to [0,\infty) \mid \lambda \text{ strictly incr. and cont., } \lambda(0) = 0, \lim_{t\to\infty} \lambda(t) = \infty\},$$

$$\Lambda := \left\{\lambda \in \Lambda' \mid \lambda \text{ Lipschitz and } \gamma(\lambda) := \sup_{s>t\geq 0} \left|\log \frac{\lambda(s)-\lambda(t)}{s-t}\right| < \infty\right\}.$$

For $x, y \in D_E[0,\infty)$ define the *Skorokhod metric* d_{J_1} by

$$d_{J_1}(x,y) := \inf_{\lambda\in\Lambda}\left\{\max\left\{\gamma(\lambda), \int_0^\infty e^{-u} \sup_{t\geq 0}\{d_{\min}(x(t\wedge u), y(\lambda(t)\wedge u))\}du\right\}\right\}.$$

b) The topology induced on $D_E[0,\infty)$ by the metric d_{J_1} is called the *Skorokhod topology* J_1.

Theorem B.17 (Separability and completeness in $D_E[0,\infty)$)**.**
If E is separable, then $D_E[0,\infty)$ is separable. If (E,d) is complete, then $(D_E[0,\infty), d_{J_1})$ is complete.

For a proof, see, e.g., Theorem 3.5.6 in Ethier and Kurtz (1986). Note that, together with Lemma A.10, the theorem implies the separability and completeness of $(D_{\mathbb{P}(S)}[0,\infty), d_{J_1})$.

Remark B.18. Skorokhod proposed four different topologies J_1, J_2, M_1 and M_2 on $D_E[0,1]$ in his work on limit theorems for stochastic processes; see Skorokhod (1956). Since then, the corresponding metric d_{J_1} has become the most popular and is thus frequently referred to as the *standard Skorokhod metric*. The metric introduced in Definition B.16 is an extension of the original Skorokhod metric on $D_E[0,1]$ to the space $D_E[0,\infty)$.

Prokhorov's theorem B.7 links the tightness of a family of probability measures with relative compactness. Applied to a sequence of stochastic processes, it can be shown that a weaker notion of "pointwise tightness" (B.1) in combination with an additional condition is sufficient for the sequence to be relatively compact. The result can be found in Kurtz (1981), Theorem 2.7.

Theorem B.19 (Characterization of relative compactness in $(D_E[0,\infty), d_{J_1})$)**.**
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of stochastic processes with sample paths in $D_E[0,\infty)$, E complete and separable. Let $\mathcal{F}_t^n = \sigma(X_n(s) : s \le t)$ and let M_T^n be the collection of $\mathcal{F}_t^n$ stopping times τ such that $\tau \le T$ a.s. Suppose for every $\varepsilon > 0$ and rational $t \ge 0$ there exists a compact set $\Gamma_{t,\varepsilon} \subset E$ such that

$$\inf_{n\in\mathbb{N}} \mathbb{P}(X_n(t) \in \Gamma_{t,\varepsilon}) > 1 - \varepsilon. \tag{B.1}$$

Then for any $\beta > 0$ either of the following conditions implies $(X_n)_{n\in\mathbb{N}}$ is relatively compact:

a) For $T, \delta > 0$ there exist random variables $\gamma_n^T(\delta) \ge 0$ such that

$$\mathbb{E}[d_{\min}^{\beta}(X_n(t+\delta), X_n(t))|\mathcal{F}_t^n] \le \mathbb{E}[\gamma_n^T(\delta)|\mathcal{F}_t^n], \quad 0 \le t \le T,$$

and $\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{E}[\gamma_n^T(\delta)] = 0$.

b) For every $T > 0$

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \sup_{\tau\in M_T^n} \mathbb{E}[d_{\min}^{\beta}(X_n(\tau+\delta), X_n(\tau))] = 0.$$

The definition of relative compactness B.2 c) implies that a relatively compact sequence of stochastic processes always contains a weakly converging *sub*sequence. If additionally the finite-dimensional distributions of the processes converge weakly, the sequence of stochastic processes itself is weakly convergent. For a proof of the following theorem, see part b) of Theorem 3.7.8 in Ethier and Kurtz (1986).

Theorem B.20 (Characterization of weak convergence in $(D_E[0,\infty), d_{J_1})$)**.**
Let E be separable and let $X, X_1, X_2, \ldots$ be processes with sample paths in $D_E[0,\infty)$.
If $(X_n)_{n\in\mathbb{N}}$ is relatively compact and there exists a dense set $D \subset [0,\infty)$ such that

$$(X_n(t_1), \ldots, X_n(t_k)) \quad \Rightarrow \quad (X(t_1), \ldots, X(t_k))$$

holds for every finite set $\{t_1, \ldots t_k\} \subset D$, then

$$X_n \quad \Rightarrow \quad X.$$

The following theorem gives a criterion under which the limit process of a weakly converging sequence of stochastic processes is continuous. The statement can be found in Ethier and Kurtz (1986), Theorem 3.10.2.

Theorem B.21 (Characterization of continuity for limit processes in $(D_E[0,\infty), d_{J_1})$)**.**
For $x \in D_E[0,\infty)$ define

$$J(x) = \int_0^\infty e^{-u} \min\{J(x,u), 1\} du,$$

where

$$J(x,u) = \sup_{0 \le t \le u} d(x(t), x(t-)).$$

Now let $X, (X_n)_{n\in\mathbb{N}}$ be processes with sample paths in $D_E[0,\infty)$ and suppose that $X_n \Rightarrow X$. Then the following holds:

$$X \text{ is a.s. continuous} \quad \Longleftrightarrow \quad J(X_n) \Rightarrow 0.$$

In the case of a continuous limit process, convergence with respect to d_{J_1} implies uniform convergence on compact intervals.

Theorem B.22 (Uniform convergence on compact intervals)**.**
Let $x \in D_E[0,\infty)$ be continuous and $(x_n)_{n\in\mathbb{N}} \subset D_E[0,\infty)$ such that $\lim_{n\to\infty} d_{J_1}(x_n, x) = 0$. Then

$$\lim_{n\to\infty} \sup_{0 \le t \le u} d(x_n(t), x(t)) = 0.$$

for all $u \ge 0$.

Proof. Since the limit process x is continuous, we have $J(x,u) = 0$ for all $u \ge 0$. Lemma 3.10.1 in Ethier and Kurtz (1986) then implies the statement. □

Bibliography

Ackley, D. H., Hinton, G. E. and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines, *Cognitive science* **9**: 147–169.

Ahmadova, A. and Mahmudov, N. I. (2023). Stochastic maximum principle for discrete time mean-field optimal control problems, *Optimal Control Applications and Methods* **44**: 3361–3378.

Altman, E. (1999). *Constrained Markov Decision Processes*, Chapman & Hall/CRC.

Andersson, D. and Djehiche, B. (2011). A maximum principle for SDEs of mean-field type, *Applied Mathematics & Optimization* **63**: 341–356.

Ash, R. B. and Doléans-Dade, C. (2000). *Probability and measure theory*, San Diego: Academic Press.

Bäuerle, N. (2000). Asymptotic optimality of tracking policies in stochastic networks, *The Annals of Applied Probability* **10**(4): 1065–1083.

Bäuerle, N. (2002). Optimal control of queueing networks: An approach via fluid models, *Advances in Applied Probability* **34**(2): 313–328.

Bäuerle, N. (2023). Mean field Markov decision processes, *Applied Mathematics & Optimization* **88**: 12.

Bäuerle, N. and Göll, T. (2023). Nash equilibria for relative investors via no-arbitrage arguments, *Mathematical Methods of Operations Research* **97**: 1–23.

Bäuerle, N. and Höfer, S. (2024). Continuous-time mean field Markov decision models, *Applied Mathematics and Optimization* **90**: 12.

Bauso, D., Tembine, H. and Basar, T. (2016). Opinion dynamics in social networks through mean-field games, *SIAM Journal on Control and Optimization* **54**: 3225–3257.

Bayraktar, E., Cosso, A. and Pham, H. (2018). Randomized dynamic programming principle and Feynman–Kac representation for optimal control of McKean–Vlasov dynamics, *Transactions of the American Mathematical Society* **370**: 2115–2160.

Bellman, R. (1954). The theory of dynamic programming, *Bulletin of the American Mathematical Society* p. 503–515.

Bellman, R. (1957). *Dynamic programming*, Princeton University Press.

Bertsekas, D. P. and Shreve, S. E. (1978). *Stochastic optimal control*, New York: Academic Press.

Betts, J. T. (1998). Survey of numerical methods for trajectory optimization, *Journal of Guidance, Control, and Dynamics* **21**: 193–207.

Billingsley, P. (1999). *Convergence of probability measures, 2nd ed.*, John Wiley & Sons.

Blackwell, D. (1962). Discrete dynamic programming, *The Annals of Mathematical Statistics* **33**: 719–726.

Blackwell, D. (1965). Discounted dynamic programming, *The Annals of Mathematical Statistics* **36**: 226–235.

Bordenave, C. and Anantharam, V. (2007). Optimal control of interacting particle systems, *Preprint*, HAL Open Archive. https://hal.science/hal-00397327.

Bramson, M. (2008). *Stability of Queueing Networks*, Springer Berlin Heidelberg.

Brémaud, P. (2020). *Markov Chains*, Springer Nature.

Buckdahn, R., Djehiche, B. and Li, J. (2011). A general stochastic maximum principle for SDEs of mean-field type, *Applied Mathematics & Optimization* **64**: 197–216.

Calafiore, G. C., Novara, C. and Possieri, C. (2020). A time-varying SIRD model for the COVID-19 contagion in Italy, *Annual Reviews in Control* **50**: 361–372.

Carmona, R. and Delarue, F. (2015). Forward–backward stochastic differential equations and controlled McKean–Vlasov dynamics, *The Annals of Probability* **43**: 2647–2700.

Carmona, R. and Delarue, F. (2018a). *Probabilistic Theory of Mean Field Games with Applications I*, Springer Cham.

Carmona, R. and Delarue, F. (2018b). *Probabilistic Theory of Mean Field Games with Applications II*, Springer Cham.

Carmona, R., Delarue, F. and Lachapelle, A. (2013). Control of McKean–Vlasov dynamics versus mean field games, *Mathematics and Financial Economics* **7**: 131–166.

Carmona, R., Laurière, M. and Tan, Z. (2023). Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning, *The Annals of Applied Probability* **33**(6B): 5334–5381.

Cecchin, A. (2021). Finite state n-agent and mean field control problems, *ESAIM: Control, Optimisation and Calculus of Variations* **27**: 31.

Cesari, L. (1983). *Optimization - Theory and Applications*, Springer.

Chen, H. (1995). Fluid approximations and stability of multiclass queueing networks: Work-conserving disciplines, *The Annals of Applied Probability* **5**(3): 637–665.

Cowan, J. D. and Wilson, H. R. (1972). Excitatory and inhibitory interactions in localized populations of model neurons, *Biophysical Journal* **12**: 1–24.

Cui, K., Tahir, A., Sinzger, M. and Koeppl, H. (2021). Discrete-time mean field control with environment states, *2021 60th IEEE Conference on Decision and Control (CDC)*, IEEE, pp. 5239–5246.

Davis, M. H. (1993). *Markov models and optimization*, Chapman and Hall.

Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M. and Friston, K. (2008). The dynamic brain: From spiking neurons to neural masses and cortical fields, *PLoS Computational Biology* **4**: 1–35.

Djete, F., Possamai, D. and Tan, X. (2022). McKean–Vlasov optimal control: The dynamic programming principle, *The Annals of Probability* **50**: 791–833.

Dong, B., Nie, T. and Wu, Z. (2022). Maximum principle for discrete-time stochastic control problem of mean-field type, *Automatica* **144**: 110497.

Elliott, R., Li, X. and Ni, Y.-H. (2013). Discrete time mean-field stochastic linear-quadratic optimal control problems, *Automatica* **49**: 3222–3233.

Elstrodt, J. (2018). *Maß- und Integrationstheorie*, Springer Spektrum.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes*, Wiley, New York.

Feichtinger, G. and Hartl, R. F. (1986). *Optimale Kontrolle ökonomischer Prozesse : Anwendungen des Maximumprinzips in den Wirtschaftswissenschaften*, de Gruyter.

Fernandez-Villaverde, J. and Jones, C. I. (2022). Estimating and simulating a SIRD model of COVID-19 for many countries, states, and cities, *Journal of Economic Dynamics and Control* **140**.

Forwick, L. (1997). *Optimale Kontrolle stückweise deterministischer Prozesse*, PhD thesis, Dissertation, Bonn, Universität Bonn.

Gast, N. and Gaujal, B. (2011). A mean field approach for optimization in discrete time, *Discrete Event Dynamic Systems* **21**: 63–101.

Gast, N., Gaujal, B. and Le Boudec, J.-Y. (2012). Mean field for Markov decision processes: from discrete to continuous optimization, *IEEE Transactions on Automatic Control* **57**(9): 2266–2280.

Grass, D., Caulkins, J. P., Feichtinger, G., Tragler, G. and Behrens, D. A. (2008). *Optimal Control of Nonlinear Processes : With Applications in Drugs, Corruption, and Terror*, Springer Berlin Heidelberg.

Gross, L. J., Lenhart, S. and Salinas, R. A. (2005). Control of a metapopulation harvesting model for black bears, *Natural Resource Modeling* **18**(3): 307–321.

Higuera-Chan, C., Jasso-Fuentes, H. and Minjarez-Sosa, A. (2016). Discrete-time control for systems of interacting objects with unknown random disturbance distributions: A mean field approach, *Applied Mathematics & Optimization* **74**(1): 197–227.

Higuera-Chan, C., Jasso-Fuentes, H. and Minjarez-Sosa, A. (2017). Control systems of interacting objects modeled as a game against nature under a mean field approach, *Journal of Dynamics and Games* **4**(1): 59–74.

Hirzebruch, F. and Scharlau, W. (1991). *Einführung in die Funktionalanalysis*, Spektrum Akademischer Verlag.

Hofgard, W., Sun, J. and Cohen, A. (2024). Convergence of the deep Galerkin method for mean field control problems, *arXiv preprint arXiv:2405.13346* .

Howard, R. A. (1960). *Dynamic programming and Markov processes*, Technology Press of Massachusetts Institute of Technology.

Howard, R. A. and Matheson, J. E. (1972). Risk-sensitive Markov decision processes, *Management Science* **18**: 356–369.

Huang, M. and Ma, Y. (2016). Mean field stochastic games: Monotone costs and threshold policies, *2016 IEEE 55th Conference on Decision and Control (CDC)*, IEEE, pp. 7105–7110.

Huang, M., Malhamé, R. P. and Caines, P. E. (2006). Large population stochastic dynamic games: closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle, *Communications in information and systems* **6**: 221–252.

Kallenberg, O. (2021). *Foundations of Modern Probability*, Springer Nature.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society A* **115**: 700–721.

Khouzani, M., Sarkar, S. and Altman, E. (2012). Maximum damage malware attack in mobile wireless networks, *IEEE/ACM Transactions on Networking* **20**(5): 1347–1360.

Kirk, D. E. (1970). *Optimal control theory : an introduction*, Englewood Cliffs, N.J. : Prentice-Hall.

Klenke, A. (2020). *Probability Theory*, Springer Nature.

Kumar, P. and Seidman, T. I. (1989). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems, *Proceedings of the 28th IEEE Conference on Decision and Control*, IEEE, pp. 2028–2031.

Kurtz, T. G. (1981). *Approximation of Population Processes*, CBMS NSF regional conference series in applied mathematics 36, Society for Industrial and Applied Mathematics.

Lacker, D. (2017). Limit theory for controlled McKean–Vlasov dynamics, *SIAM Journal on Control and Optimization* **55**: 1641–1672.

Lange, D. K. (2017). *Cost optimal control of Piecewise Deterministic Markov Processes under partial observation*, PhD thesis, Dissertation, Karlsruhe, Karlsruher Institut für Technologie (KIT), DOI: 10.5445/IR/1000069448.

Lasry, J.-M. and Lions, P.-L. (2007). Mean field games, *Japanese Journal of Mathematics* **2**: 229–260.

Laurière, M. and Pironneau, O. (2014). Dynamic programming for mean-field type control, *Comptes Rendus Mathematique* **352**: 707–713.

Lenhart, S. and Workman, J. T. (2007). *Optimal Control Applied to Biological Models*, Chapman and Hall/CRC.

Levin, D. A. and Peres, Y. (2017). *Markov Chains and Mixing Times*, American Mathematical Society.

Liu, Z., Wu, B. and Lin, H. (2018). A mean field game approach to swarming robots control, *2018 Annual American Control Conference (ACC)* pp. 4293–4298.

Markov, A. A. (1907). Extension of the limit theorems of probability theory to a sum of variables connected in a chain, *in* R. A. Howard (ed.), *Dynamic Probabilistic Systems, Volume I: Markov Chains*, Wiley, New York, pp. 552–576. English translation published in 1971.

McAsey, M., Mou, L. and Han, W. (2012). Convergence of the forward-backward sweep method in optimal control, *Computational Optimization and Applications* **53**: 207–226.

Monahan, G. E. (1982). A survey of partially observable Markov decision processes: Theory, models, and algorithms, *Management Science* **28**: 1–16.

Motte, M. and Pham, H. (2022). Mean-field Markov decision processes with common noise and open-loop controls, *The Annals of Applied Probability* **32**(2): 1421–1458.

Motte, M. and Pham, H. (2023). Quantitative propagation of chaos for mean field Markov decision process with common noise, *Electronic Journal of Probability* **28**: 1–24.

National Park Service (2025). Black bears – great smoky mountains national park, `https://www.nps.gov/grsm/learn/nature/black-bears.htm`. Last updated June 7, 2025. Accessed October 17, 2025.

Ni, Y.-H., Elliott, R. and Li, X. (2015). Discrete-time mean-field stochastic linear–quadratic optimal control problems, ii: Infinite horizon case, *Automatica* **57**: 65–77.

Ni, Y.-H., Li, X. and Zhang, J.-F. (2016). Indefinite mean-field stochastic linear-quadratic optimal control: From finite horizon to infinite horizon, *IEEE Transactions on Automatic Control* **61**(11): 3269–3284.

Ni, Y.-H., Zhang, J.-F. and Li, X. (2015). Indefinite mean-field stochastic linear-quadratic optimal control, *IEEE Transactions on Automatic Control* **60**: 1786–1800.

Peterson, C. and Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks, *Complex Systems* **1**: 995–1019.

Pham, H. and Wei, X. (2016). Discrete time McKean–Vlasov control problem: A dynamic programming approach, *Applied Mathematics & Optimization* **74**: 487–506.

Pham, H. and Wei, X. (2017). Dynamic programming for optimal control of stochastic McKean–Vlasov dynamics, *SIAM Journal on Control and Optimization* **55**: 1069–1101.

Pham, H. and Wei, X. (2018). Bellman equation and viscosity solutions for mean-field stochastic control problem, *ESAIM: Control, Optimisation and Calculus of Variations* **24**: 437–461.

Piunovskiy, A. and Zhang, Y. (2020). Continuous-time Markov decision processes, *Probability Theory and Stochastic Modelling* .

Pontryagin, L., Boltyanskii, V., Gamkrelidze, R. and Mishchenko, E. (1962). *The Mathematical Theory of Optimal Processes*, Wiley, New York. Authorized Translation from the Russian.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J. and Ganguli, S. (2016). Exponential expressivity in deep neural networks through transient chaos, *Advances in Neural Information Processing Systems* **29**.

Protter, P. E. (2005). *Stochastic Integration and Differential Equations*, Springer Berlin Heidelberg New York.

Rao, A. V. (2010). A survey of numerical methods for optimal control, *Advances in the Astronautical Sciences* **135**: 497–528.

Roijers, D. M., Vamplew, P., Whiteson, S. and Dazeley, R. (2013). A survey of multi-objective sequential decision-making, *Journal of Artificial Intelligence Research* **48**: 67–113.

Rybko, A. N. and Stolyar, A. L. (1992). Ergodicity of stochastic processes describing the operation of open queueing networks, *Problems Inform. Transmission* **28**(3): 199–220.

Seierstad, S. (1987). *Optimal Control Theory with Economic Applications*, Elsevier. North-Holland.

Sharp, J. A., Burrage, K. and Simpson, M. J. (2021). Implementation and acceleration of optimal control for systems biology, *Journal of the Royal Society Interface* **18**.

Skorokhod, A. (1956). Limit theorems for stochastic processes, *Theory of probability and its applications* **1**(3): 261–290.

Sompolinski, H., Crisanti, A. and Sommers, H. (1988). Chaos in random neural networks, *Physical Review Letters* **61**: 259–262.

Song, T. and Liu, B. (2020). Solvability and optimal stabilization controls of discrete-time mean-field stochastic system with infinite horizon, *Advances in Difference Equations* (187).

Song, T. and Liu, B. (2021). Discrete-time mean-field stochastic linear-quadratic optimal control problem with finite horizon, *Asian Journal of Control* **23**: 979–989.

Thompson, G. L. (1968). Optimal maintenance policy and sale date of a machine, *Management Science* **14**(9): 543–550.

Walz, G. (2017). *Lexikon der Mathematik, Band 2*, Springer Spektrum.

Warga, J. (1972). *Optimal control of differential and functional equations*, Academic Press.

Willard, S. (1970). *General Topology*, Addison Wesley.

Young, L. C. (1937). Generalized curves and the existence of an attained absolute minimum in the calculus of variations, *Comptes Rendus de la Société des Sciences et des Lettres de Varsovie* **30**: 211–234.

Zhang, H. and Qi, Q. (2016). Optimal control for mean-field system: Discrete-time case, *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 4474–4480.

Zhang, H., Qi, Q. and Fu, M. (2019). Optimal stabilization control for discrete-time mean-field stochastic systems, *IEEE Transactions on Automatic Control* **64**: 1125–1136.

Zhu, Z., Ke, J. and Wang, H. (2021). A mean-field Markov decision process model for spatial-temporal subsidies in ride-sourcing markets, *Transportation Research Part B: Methodological* **150**: 540–565.