OPTIMAL CONTROL OF MARKOV JUMP PROCESSES : ASYMPTOTIC ANALYSIS, ALGORITHMS, AND APPLICATION TO MODELLING OF CHEMICAL REACTION SYSTEMS

WEI ZHANG[1 2] , CARSTEN HARTMANN[1 2] , AND MAX VON KLEIST[1 2]

Abstract. Markov jump processes are used to model various phenomena. In the context of biological or chemical applications one typically refers to the chemical master equation (CME). The CME models the evolution of probability mass of any copy-number combination of interacting particles. When many interacting particles ("species") are considered, the complexity of the CME quickly increases, making direct numerical simulations impossible. This is even more problematic when one is interested in controlling the dynamics defined by the CME. Currently, to the authors knowledge, no method capable of solving the aforementioned task exists. While limit theorems allow for the approximation of the uncontrolled Markov jump process by ordinary differential equations in the large copy number regime (e.g. abundance of interacting particles is high), similar results are missing for the corresponding optimal control problem. In this study, we address this topic for both open loop and feedback control problems. Based on Kurtz's limit theorems, we prove the convergence of the respective control value functions of the underlying Markov decision problem as the copy numbers of the interacting species goes to infinity. In the case of a finite time horizon, a hybrid control policy and -algorithm are proposed to overcome the difficulties due to the curse of dimensionality when the copy number of the involved species is large. Numerical examples are studied to demonstrate the analysis and algorithms.

Key words. Markov jump process, optimal control problem, large number limit, feedback control policy, hybrid control policy.

AMS subject classifications.

1. Introduction.

In the past decades, discrete-state Markov jump processes have been a major research topic in probability theory receiving much attention in applications like economics, physics, biology and chemistry; see e.g. [31, 16, 8, 11, 1, 30]. For example, in the modelling of chemical reactions, a single state is defined as one possible copy-number combination of the distinct interacting chemical species. After a random waiting time, a reaction occurs which changes this copy-number combination. Since the time and order in which chemical reactions occur is random (referred as *instrinsic* noise), the state of the system is a random variable. The Chemical Master Equation (CME) models the probability of all possible outcomes over time, giving rise to an extremely large state space (consisting of all copy-number combinations). Consequently, solving the Chemical Master Equation or approximating its solution computationally is a non-trivial, yet unsolved, task that has been the objective of intense research over the past decades (see e.g. [26] for a summary).

In many real world applications, one does not only aim at propagating or simulating a process forward in time, but aims at controlling and optimizing it. Specifically, the model equations of a controlled system contain extra terms or parameters that can be manipulated by the decision maker according to some control policy. The latter is chosen so that a given cost functional reaches an optimal (e.g. minimum) value. There are two general approaches to an optimal control task, depending on whether the admissible control policies are allowed to depend on the system states (*feedback* or *closed loop* control problem) or not (*open loop* control problem). In the context of a stochastic control, feedback controls are random in the sense that

[1]Institute of Mathematics, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany
[2]Email : wei.zhang@fu-berlin.de, chartman@mi.fu-berlin.de, vkleist@zedat.fu-berlin.de

each realization of the process gives rise to a different control that is adapted according to the random state of the system. Stochastic control problems for Markov jump processes are often termed "Markov decision process" (MDP) [3, 27, 14].

In the case of an open loop control, the control follows a fixed, deterministic policy, which in a stochastic context is reasonable only if one aims at controlling the deterministic evolution of the probability mass of all possible outcomes (the CME), rather than stochastic realizations of the process.

For small or moderately sized systems, both the feedback and open loop control problem can be solved computationally via dynamic programming (e.g. [32, 6]). However, for large systems, as e.g. defined by the Chemical Master Equation (CME), the corresponding problems can usually not be solved by dynamic programming or related (e.g. dual) methods without suitable approximations or remodelling steps. One such remodelling step that has been extensively exploited by control engineers is to replace the CME by a lower dimensional ODE system that ignores the intrinsic noise (e.g., see [18, 23]). These continuous deterministic reaction rate equations model the concentrations of the interacting chemical species by one ordinary differential equation per species. It has been shown in Kurtz's seminal work [20, 19, 22, 21] that the expected particle numbers per unit volume of the original Markov jump processes without control can be approximated by the classical reaction rate equations in the large copy-number regime (parameterized by either the total number of particles N or the reaction volume V). However, the relationship between the optimal control problem of the original Markov jump process and that of the limiting ODE system has not been rigorously studied.

In this article we aim at filling this apparent gap : We study the convergence of the optimal control and its associated costs for the controlled original Markov jump process and the controlled limiting ODE system as $N \to \infty$, both in the open loop and the feedback case. We will confine our analysis to the situation where the control can only be changed at fixed points in times (called *control stages*). Finally, based on the convergence results, we propose a hybrid (deterministic–open–loop, stochastic–feedback) control policy to overcome the curse of dimension when the state space of the CME is large.

The paper is organized as follows: In Section 2, we introduce the mathematical problem along with notations used throughout this paper. Section 3 is devoted to the extension of Kurtz's limit theorem for Markov jump processes to optimal control problems and contains the main results of this paper. Based on this analysis, a hybrid control algorithm is proposed and discussed in Section 4. We present several numerical examples in Section 5, a technical lemma is exemplified in Appendix A.

2. Mathematical Setup.

In this section, we will first introduce our problem, the notations used, and finally sketch two concrete situations in which the problem is relevant.

2.1. Controlled Markov jump processes.

Let $\mathbb{X}$ be a discrete lattice in $\mathbb{R}^n$ and consider the Markov jump process $x(t)$ on it. Suppose that at time $t \geq 0$ and given $x(t) = x \in \mathbb{X}$, the probability for making a transition from x to $x+l$ within the infinitesimal time interval $[t, t+ds)$ is $f(x,l)\,ds$, $l \in \mathbb{X}$. Letting τ denote the waiting time

$$\tau = \inf_{s>t}\{s - t\ ;\ x(s) \neq x(t)\}\,, \tag{2.1}$$

it is known that τ follows an exponential distribution with rate $\lambda(x) = \sum_{l\in\mathbb{X}} f(x,l)$, i.e. $\tau \sim \text{Exp}(\lambda(x))$.

Jump rates. In this work, we suppose that the jump process $x(t)$ depends on both a parameter $N \gg 1$ and control $\nu \in \mathcal{A}$, where $\mathcal{A}$ is the *control set*. In applications, N may be related to system's volume or the magnitude of particle numbers, while ν may affect the jump rates f. To indicate these dependencies, we denote the jump process as $x^{\nu,N}$ and also introduce the normalized process $z^{\nu,N}(t) = N^{-1}x^{\nu,N}(t)$. It is convenient to think of the normalized variable z as a particle density, which is why we will sometimes refer to $z^{\nu,N}(t)$ as the *density process*. Notice that $z^{\nu,N}$ is a Markov jump process on the scaled lattice and, due to its importance in our analysis, we use the notation $\mathbb{X}_N$ and functions $f^{\nu,N} : \mathbb{X}_N \times \mathbb{X}_N \to \mathbb{R}^+$ for its state space and jump rates, respectively. $\mathbb{X}$ and f^ν will be reserved for the corresponding notations related to the original process $x^{\nu,N}$ (their dependence on N is omitted). Specifically, we have $\mathbb{X}_N = \{\frac{x}{N} \,|\, x \in \mathbb{X}\}$ and $f^{\nu,N}(z,l) = f^\nu(Nz, NL)$ for $z, l \in \mathbb{X}_N$, where $\mathbb{R}^+$ is the set consisting of non-negative real numbers.

Controls. We will discuss the control policies and the controlled Markov jump process in detail. For the sake of simplicity, we will refer to the normalized process $z^{\nu,N}$ only, stressing that all considerations are transferrable to $x^{\nu,N}$. Suppose on time interval $[0,T]$, $K+1$ time points $0 = t_0 < t_1 < \cdots < t_j < t_{j+1} < \cdots < t_K = T$ are given and fixed. At each time t_j, $0 \le j < K$, called *control stage*, we are allowed to select some control $\nu_j \in \mathcal{A}$ and apply it to the jump process in order to influence its jump rates. Once a control ν_j is selected at time t_j, it will persistently take effect during the time interval $[t_j, t_{j+1})$. When the selection of controls ν_j is allowed to depend on the system's states, the control policy is called feedback control policy and otherwise is called open loop control policy. More generally, we introduce the sets of open loop and feedback control policies on time $[t_k, T]$ for $0 \le k < K$:

$$\begin{aligned}\mathcal{U}_{o,k} =& \{(\nu_k, \nu_{k+1}, \cdots, \nu_{K-1}) \mid \nu_j \in \mathcal{A},\ k \le j < K\},\\ \mathcal{U}_{f,k} =& \{(\nu_k, \nu_{k+1}, \cdots, \nu_{K-1}) \mid \nu_j : \mathbb{X}_N \to \mathcal{A},\ k \le j < K\}.\end{aligned} \tag{2.2}$$

Notice in the feedback case, while each policy ν_j is a function of state, the same notation will be used to denote its value (i.e. the control selected at t_j) when no ambiguities can arise. For further simplification, let σ denote either 'o' or 'f' and we will write $\mathcal{U}_{\sigma,k}$ to refer to either open loop or feedback control policy set.

For all $\nu \in \mathcal{A}$ and $t \in [0,T]$, the variable $z^N(t,\nu)$ or $z^{\nu,N}(t)$ denotes the controlled process and indicates that control ν is applied at time $t \ge 0$. Given a control policy $u \in \mathcal{U}_{\sigma,k}$, we express the corresponding controlled process in the time interval $[t_k, T]$ as $z^{u,N}(t)$, i.e. $z^{u,N}(t) = z^N(t, \nu_j)$ when $t \in [t_j, t_{j+1})$, $k \le j < K$. The notation $z^{u,N}(t\,;\, z)$ will be used to emphasize that the process starts from a fixed initial state $z \in \mathbb{X}_N$ (the starting time may be nonzero). Introducing the notation $j(t) := k$, if $t \in [t_k, t_{k+1})$, we have $z^{u,N}(t) = z^N(t, \nu_{j(t)})$. Now, let $u = (\nu_0, \nu_1, \nu_2, \cdots, \nu_{K-1}) \in \mathcal{U}_{\sigma,0}$ be a fixed control policy, where application of the control changes the jump rates of the Markov jump process. Specifically, $z^{u,N}(t), t \ge 0$ is a Markov jump process with the property that the probability for system's state to jump from $z^{u,N}(t) = z$ to $z+l$ within the infinitesimal time interval $[t, t+ds)$ at $t \in [t_j, t_{j+1})$, is $f^{\nu_j,N}(z,l)\,ds$ for $l \in \mathbb{X}_N$.

Cost functional. For a control policy $u = (\nu_0, \nu_1, \cdots, \nu_{K-1}) \in \mathcal{U}_{\sigma,0}$, and a process $z^{u,N}$, we define the cost functional

$$J_N(z,u) = \mathbf{E}_z^u\Big[\sum_{j=0}^{K-1}\Big(r\big(z^{u,N}(t_j),\nu_j\big) + \int_{t_j}^{t_{j+1}} \phi\big(z^{u,N}(s),\nu_j\big)\,ds\Big) + \psi\big(z^{u,N}(T)\big)\Big] \tag{2.3}$$

where $\mathbf{E}_z^u$ denotes the expectation over all realizations of $z^{u,N}$ starting at $z^{u,N}(0) = z$ and evolving under the control policy u. The functions $r, \phi : \mathbb{R}^n \times \mathcal{A} \to \mathbb{R}^+$ and $\psi : \mathbb{R}^n \to \mathbb{R}^+$ correspond to the costs at each control stage t_j, the running cost and the terminal cost.

2.2. Limit process and underlying assumptions. Our analysis in the course of the paper is based on Kurtz's limit theorems for jump processes [20, 19, 22, 21] which suggest that, for $u \in \mathcal{U}_{o,0}$, the normalized density proccess $z^{u,N}$ converges to a deterministic limiting process $\tilde{z}^u$ and is governed by the ordinary differential equation (ODE)

$$\frac{d\tilde{z}^u(t)}{dt} = F^{\nu_{j(t)}}(\tilde{z}^u(t))\,, \tag{2.4}$$

or, in integral form,

$$\tilde{z}^u(t) = \tilde{z}^u(0) + \int_0^t F^{\nu_{j(s)}}(\tilde{z}^u(s))\,ds\,. \tag{2.5}$$

Here the vector field F^ν is defined as the limit of

$$F^{\nu,N}(z) = \sum_{l\in\mathbb{X}_N} l\, f^{\nu,N}(z,l)\,, \quad z \in \mathbb{X}_N\,, \tag{2.6}$$

as $N \to \infty$ (see Assumption 2). Convergence of $z^{u,N}$ to $\tilde{z}^u$ will be established below in Theorem 3.1.

Limiting control value. We are interested in approximating the optimal control policy for the jump process by an optimal control $u_{opt,\sigma} \in \mathcal{U}_{\sigma,0}$ of the limiting process, such that

$$J_N(z, u_{opt,\sigma}) \approx U_N(z) \triangleq \inf_{u\in\mathcal{U}_{\sigma,0}} J_N(z,u)\,, \tag{2.7}$$

i.e. the infimum (minimum) is attained, at least approximately, under the limiting policy $u_{opt,\sigma}$.

The function U_N is called the *value function* or *control value* of the underlying stochastic feedback control problem. It is known that an optimal control $u^N_{opt,\sigma} = \operatorname{argmin} J_N(z,u)$ exists when $\mathcal{A}$ is a finite set; see [27] for more details and possible relaxations of the assumptions on the set of admissible controls.

For the related deterministic limit process $\tilde{z}^u$ satisfying (2.4) under some open loop policy $u \in \mathcal{U}_{o,0}$, we define the cost functional by

$$\widetilde{J}(z,u) = \sum_{j=0}^{K-1}\Big[r\big(\tilde{z}^u(t_j),\nu_j\big) + \int_{t_j}^{t_{j+1}} \phi\big(\tilde{z}^u(s),\nu_j\big)\,ds\Big] + \psi\big(\tilde{z}^u(T)\big)\,, \tag{2.8}$$

with the corresponding value function $\widetilde{U}(z) = \inf_{u\in\mathcal{U}_{o,0}} \widetilde{J}(z,u)$. Note that when set $\mathcal{A}$ is finite, the minimizer exists since the number of possible open loop control policies u is finite and equal to $|\mathcal{A}|^K$, i.e. $|\mathcal{U}_{o,0}| = |\mathcal{A}|^K$. Convergence of the value function $U_N \to \widetilde{U}$ will be established in the course of the paper.

Standing assumptions. The subsequent analysis rests on the following assumptions :

ASSUMPTION 1. *For some fixed* $1 < \alpha \leq 2$, *we assume that*

$$M_N := \sup_{\nu \in \mathcal{A}} \sup_{z \in \mathbb{X}_N} \Big(\sum_{l \in \mathbb{X}_N} |l|^\alpha f^{\nu,N}(z,l) \Big) < \infty \,, \tag{2.9}$$

and satisfies

$$\lim_{N \to \infty} M_N = 0 \,.$$

ASSUMPTION 2. *There exists a function* $F^\nu : \mathbb{R}^n \to \mathbb{R}^n$, *such that*

$$\omega_N = \sup_{z \in \mathbb{X}_N \,,\, \nu \in \mathcal{A}} \big| F^{\nu,N}(z) - F^\nu(z) \big| \tag{2.10}$$

satisfies

$$\lim_{N \to \infty} \omega_N = 0 \,.$$

ASSUMPTION 3. *There exists a constant* $L_F \geq 0$, *such that*

$$|F^\nu(z') - F^\nu(z)| \leq L_F |z' - z| \,, \quad \forall z, z' \in \mathbb{R}^n \,,\ \nu \in \mathcal{A} \,.$$

ASSUMPTION 4. *There exist* $L_r, L_\phi, L_\psi, M_r, M_\phi, M_\psi \geq 0$, *such that*

$$\begin{aligned} &|r(z_1,\nu) - r(z_2,\nu)| \leq L_r |z_1 - z_2| \,, \quad |\phi(z_1,\nu) - \phi(z_2,\nu)| \leq L_\phi |z_1 - z_2|, \\ &|\psi(z_1) - \psi(z_2)| \leq L_\psi |z_1 - z_2| \,, \quad \forall z_1, z_2 \in \mathbb{R}^n \,,\ \nu \in \mathcal{A} \,. \end{aligned}$$

Moreover, $|r(0,\nu)| \leq M_r$, $|\phi(0,\nu)| \leq M_\phi$, $\forall \nu \in \mathcal{A}$, *as well as* $|\psi(0)| \leq M_\psi$.

Assumption 2 states that $F^{\nu,N}(z)$ converges to $F^\nu(z)$ uniformly for all $\nu \in \mathcal{A}$, while Assumption 3 states that the family of limiting vector fields $F^\nu(z)$ are Lipschitz functions of $z \in \mathbb{R}^n$ with Lipschitz constant L_F, uniformly for $\nu \in \mathcal{A}$.

2.3. Applications. Here we consider two prototypical examples of Markov jump processes, which appear relevant in the context of optimal control and to which our results can be applied.

Density dependent Markov chain. The first example is what is called *density dependent Markov chain* [20]. One feature is that jump rates depend on the density of the system's states. Concrete models include the predator-prey model, elementary chemical reactions such as $\mathrm{B} + \mathrm{C} \rightleftharpoons \mathrm{D}$ or epidemic models [20, 19]. Following the notations of Subsection 2.1 and denoting the density dependent Markov chain as $x^{\nu,N}(\cdot)$, it holds that the rate of jumping from x to $x + l$ under the control $\nu \in \mathcal{A}$, is given by $N f^\nu(x/N, l)$ for $l \in \mathbb{X}$. As a consequence, $f^{\nu,N}(z, l/N) = N f^\nu(z,l)$ is the rate at which the normalized process $z^{u,N}(\cdot) = N^{-1} x^{u,N}(\cdot)$ jumps from $z = x/N$ to $z + l/N = (x+l)/N$.

Notice that Assumption 1 holds, if we assume that

$$M_\alpha = \sup_{\nu \in \mathcal{A}} \sup_{z \in \mathbb{R}^n} \Big(\sum_{l \in \mathbb{X}} |l|^\alpha f^\nu(z,l) \Big) < \infty \,, \tag{2.11}$$

since $M_N = N^{1-\alpha} M_\alpha$ with $\alpha > 1$. Furthermore, if we define

$$F^\nu(z) = \sum_{l \in \mathbb{X}} l f^\nu(z, l), \quad \forall z \in \mathbb{R}^n, \tag{2.12}$$

then (2.6) becomes

$$F^{\nu,N}(z) = \sum_{l \in \mathbb{X}_N} l f^{\nu,N}(z, l) = \sum_{l \in \mathbb{X}} \frac{l}{N} \cdot N f^\nu(z, l) = F^\nu(z), \quad z \in \mathbb{X}_N,$$

which implies that Assumption 2 trivially holds with $\omega_N \equiv 0$.

Chemical reactions. As a second example, we mention systems of chemical reactions. Consider a reaction network consisting of n chemical species that can undergo m different chemical reactions:

$$\sum_{i=1}^n v_{ki} S_i \xrightarrow{\kappa'_k} \sum_{i=1}^n v'_{ki} S_i, \quad k = 1, \cdots, m. \tag{2.13}$$

Here the S_i are the different chemical species, κ'_k is the rate constant of the k-th reaction, v_{ki}, v'_{ki} are the molecule numbers of species S_i consumed or generated when the k-th reaction fires. Now let $x_i(t)$ be the number of molecules of species S_i at time t and define

$$x(t) = (x_1(t), x_2(t), \cdots, x_n(t))^T \in \mathbb{N}^n \tag{2.14}$$

to be the state of the chemical system at time t. When the k-th reaction fires at time $t > 0$, the system's state jumps from $x(t)$ to $x(t) + (v'_k - v_k)$ where

$$v_k = (v_{k1}, v_{k2}, \cdots, v_{kn})^T \in \mathbb{N}^n, \quad v'_k = (v'_{k1}, v'_{k2}, \cdots, v'_{kn})^T \in \mathbb{N}^n. \tag{2.15}$$

In order to fully describe the system as a jump process, we still need to specify the Poisson intensity of each reaction (propensity function). Let λ denote a generic propensity function. We will restrict ourself to elementary reactions, which involve at most two molecules:

1. $\emptyset \xrightarrow{\kappa} S_i, \quad \lambda = \kappa N$
2. $S_i \xrightarrow{\kappa} S_j, \quad \lambda = \kappa x_i$
3. $2 S_i \xrightarrow{\kappa} S_j, \quad \lambda = \frac{\kappa}{N} x_i(x_i - 1)$
4. $S_i + S_j \xrightarrow{\kappa} S_k, \quad \lambda = \frac{\kappa}{N} x_i x_j,$

where N in the last two reactions is a constant related to the system volume (e.g., the total number of molecules or a test tube volume). Note that, in general, the propensity $\lambda_k = \lambda_k(x)$ is a function of the system state. The dynamics of $x(t)$ can then be written as

$$x(t) = x(0) + \sum_{k=1}^m (v'_k - v_k) Y_k \left(\int_0^t \lambda_k(x(s))\, ds \right) \tag{2.16}$$

where the $Y_k(\cdot), 1 \le k \le m$ are independent Poisson processes with unit intensity. For a system of controlled chemical reactions, we use the notation $\lambda_k^{\nu,N}(x)$ to indicate that the propensities not only depend on N, but also on the control $\nu \in \mathcal{A}$ via the rate constants κ. It follows from the definition of the reactive events that transition rates and propensity functions are related by

$$f^\nu(x, l) = \begin{cases} \lambda_k^{\nu,N}(x) & \text{if } l = v'_k - v_k \text{ for some } 1 \le k \le m, \\ 0 & \text{otherwise}. \end{cases}$$

Notice that if only reactions of type 1,2 or 4 are involved, the process defined by f^ν is an instance of the aforementioned density dependent Markov chain; when reactions of type 3 are involved, then the limiting vector field F^ν can be computed from $F^{\nu,N}$ by exploiting that $f^{\nu,N}(z,l) = N\kappa z_i(z_i - N^{-1})$ if $l = v_k' - v_k$ for some $1 \le k \le m$ and 0 otherwise

3. Asymptotic analysis of the optimal control problem. In this section, we study optimal control problems in the large number regime based on Kurtz's limiting theorem [20, 19, 22, 21]. As a first step, we confine our attention to the open loop control problem that turns out to be a direct application of Kurtz's theorem, given that Assumptions 3 and 4 apply. Specifically, we show that $J_N(z,u) \to \widetilde{J}(z,u)$ for $\forall z \in \mathbb{X}_N, u \in \mathcal{U}_{o,0}$ as $N \to \infty$ (Theorem 3.2). Then, as a second step, we consider the feedback control problem and prove that $U_N(z) \to \widetilde{U}(z)$, and, especially, if $u \in \mathcal{U}_{o,0}$ and $\widetilde{J}(z,u) = \widetilde{U}(z)$, then $|J_N(z,u) - U_N(z)| \to 0$ as $N \to \infty$ (Theorem 3.3). As we will discuss in detail, an important consequence of Theorem 3.3 is that the optimal (open loop) control policy for the limiting ODE system is almost optimal for the Markov jump process if $N \gg 1$, i.e., it is asymptotically optimal among all feedback control policies in $\mathcal{U}_{f,0}$. Finally, we extend the analysis of the finite time-horizon case to discounted optimal control problems on an infinite time-horizon (Theorem 3.4).

3.1. ODE approximation of the normalized Markov jump process. Let u be some open loop control policy, and recall that $z^{u,N}(t) = N^{-1}x^{u,N}(t)$ denotes the normalized Markov jump process. The limit process as $N \to \infty$ is described by the following theorem.

THEOREM 3.1. *Let $z^{u,N}(t)$ be the normalized jump process under the policy $u \in \mathcal{U}_{o,0}$. Further let $\tilde{z}^u(t)$ be the solution of (2.4) on $t \in [0,T]$ for some $T > 0$. If Assumption 3 holds, then*

$$\mathbf{E}^u\left[\sup_{0\le s\le t}\left|z^{u,N}(s) - \tilde{z}^u(s)\right|\right] \le \left\{\mathbf{E}\left[\left|z^{u,N}(0) - \tilde{z}^u(0)\right|\right] + C_{T,N}\right\} e^{L_F t},$$

for $0 \le t \le T$, with the constant

$$C_{T,N} = T\omega_N + \frac{\alpha}{2(\alpha-1)}\Big(\frac{4TM_N}{\alpha-1}\Big)^{\frac{1}{\alpha}}, \tag{3.1}$$

and ω_N and M_N as defined in (2.10) and (2.9). If $z^{u,N}(0) = \tilde{z}^u(0) = z_0$ and Assumptions 1 and 2 are met, we have for any control policy $u \in \mathcal{U}_{o,0}$:

$$\lim_{N\to\infty} \mathbf{E}^u_{z_0}\left[\sup_{0\le s\le t}|z^{u,N}(s) - \tilde{z}^u(s)|\right] = 0\,. \tag{3.2}$$

Proof. Let $w^{u,N}$ be the martingale (see [20])

$$w^{u,N}(t) = z^{u,N}(t) - z^{u,N}(0) - \int_0^t F^{\nu_{j(s)},N}(z^{u,N}(s))\,ds\,, \tag{3.3}$$

and consider the coupled Markov process $(z^{u,N}(t), w^{u,N}(t))$. For some differentiable function φ of w, Dynkin's formula [7, 25] entails

$$\begin{aligned}
&\mathbf{E}^u\left[\varphi\big(w^{u,N}(t)\big)\right] - \mathbf{E}^u\left[\varphi\big(w^{u,N}(0)\big)\right] \\
&= \int_0^t \mathbf{E}^u\Big[\sum_{l\in\mathbb{X}_N}\Big(\varphi\big(l + w^{u,N}(s)\big) - \varphi\big(w^{u,N}(s)\big) - l\cdot\nabla\varphi\big(w^{u,N}(s)\big)\Big) f^{\nu_{j(s)},N}(z^{u,N}(s), l)\Big]\,ds\,.
\end{aligned}$$

In particular, setting $\varphi(z) = |z|^\alpha$ and using Lemma A.1 from Appendix A, we obtain

$$\mathbf{E}^u\left[|w^{u,N}(t)|^\alpha\right] \leq \frac{4t}{2^\alpha(\alpha-1)} \sup_{\nu\in\mathcal{A}} \sup_{z\in\mathbb{X}_N} \Big(\sum_{l\in\mathbb{X}_N} |l|^\alpha f^{\nu,N}(z,l)\Big) = \frac{4tM_N}{2^\alpha(\alpha-1)}\,,$$

which, by Hölder's inequality and Doob's maximal inequality, implies that

$$\begin{aligned}\mathbf{E}^u\left[\sup_{0\leq s\leq t}|w^{u,N}(s)|\right] &\leq \mathbf{E}^u\left[\sup_{0\leq s\leq t}|w^{u,N}(s)|^\alpha\right]^{\frac{1}{\alpha}} \\ &\leq \frac{\alpha}{\alpha-1}\mathbf{E}^u\left[|w^{u,N}(t)|^\alpha\right]^{\frac{1}{\alpha}} \\ &\leq \frac{\alpha}{2(\alpha-1)}\left(\frac{4tM_N}{\alpha-1}\right)^{1/\alpha}.\end{aligned} \tag{3.4}$$

Combining (3.3) and (2.5) and taking Assumption 3 into consideration, it follows that

$$\begin{aligned}\left|z^{u,N}(t)-\tilde{z}^u(t)\right| \leq& \left|z^{u,N}(0)-\tilde{z}^u(0)\right| + L_F\int_0^t \left|z^{u,N}(s)-\tilde{z}^u(s)\right|ds \\ &+\int_0^t \left|F^{\nu_{j(s)},N}\left(z^{u,N}(s)\right)-F^{\nu_{j(s)}}\left(z^{u,N}(s)\right)\right|ds + \left|w^{u,N}(t)\right| \\ \leq& \left|z^{u,N}(0)-\tilde{z}^u(0)\right| \\ &+L_F\int_0^t\left|z^{u,N}(s)-\tilde{z}^u(s)\right|ds + t\,\omega_N + \left|w^{u,N}(t)\right|.\end{aligned}$$

Now let $y^{u,N}(t) = \sup_{0\leq s\leq t}\left|z^{u,N}(s)-\tilde{z}^u(s)\right|$. Then

$$y^{u,N}(t) \leq y^{u,N}(0) + L_F\int_0^t y^{u,N}(s)\,ds + T\omega_N + \sup_{0\leq s\leq T}|w^{u,N}(s)|\,.$$

Taking expectations and applying Gronwall's inequality and (3.4) yields

$$\mathbf{E}^u\left[y^{u,N}(t)\right] \leq \left\{\mathbf{E}^u\left[y^{u,N}(0)\right] + T\omega_N + \frac{\alpha}{2(\alpha-1)}\left(\frac{4TM_N}{\alpha-1}\right)^{1/\alpha}\right\}e^{L_F t}\,.$$

The second part of the assertions follows upon taking the limit $N\to\infty$ with $z^{u,N}(0) = \tilde{z}^u(0)$. □

Before we proceed to the closed loop case, a few remarks:

Remark 1. *If $z^{u,N,i}, i = 1,2$ denote two jump processes that have the same jump rates, but start from different initial values at time $t = 0$, then, for any $t\in[0,T]$, $T > 0$, the difference between the two solutions is uniformly bounded by*

$$\mathbf{E}^u\left[\sup_{0\leq s\leq t}\left|z^{u,N,2}(s)-z^{u,N,1}(s)\right|\right] \leq \left\{\mathbf{E}\left[\left|z^{u,N,2}(0)-z^{u,N,1}(0)\right|\right] + 2C_{T,N}\right\}e^{L_F t}\,.$$

Remark 2. *For the density dependent Markov chain of Subsection 2.3, it holds that $\omega_N = 0$ and $M_N = N^{1-\alpha}M_\alpha$, where M_α is given in (2.11). And thus*

$$C_{T,N} = \mathcal{O}\left(N^{1/\alpha-1}\right). \tag{3.5}$$

Here, the simplest case is when $z^{u,N}$ is one-dimensional, in which case $\mathcal{A}$ is a singleton and therefore the control u will be omitted for the remainder of this paragraph. Suppose that $f(z,1)=1$ and $f(z,l)=0$ for $l\neq 1$, $z\geq 0$, which implies that $F(z)=1$, which is Lipschitz continuous with Lipschitz constant $L_F=0$. For the initial value $z_0=0$, equation (2.5) yields $\tilde{z}(t)=t$ and $z^N(t)=N^{-1}P(Nt)$, where $P(\cdot)$ is a Poisson process with unit intensity. Further note that Assumption 1 holds with $\alpha=2$ and $M_\alpha=1$, so that Theorem 3.1 entails

$$\mathbf{E}^u_{z_0}\left[\sup_{0\leq s\leq T}\left|\frac{P(Ns)}{N}-s\right|\right]\leq\left(\frac{4T}{N}\right)^{1/2}.$$

3.2. Optimal control on finite time horizon. In this subsection we extend the previous considerations to the case of both open and closed loop optimal control on an finite time-horizon.

Open loop control. As a straight consequence of Theorem 3.1 and Assumptions 3–4, we obtain the following result for the open loop control problem.

THEOREM 3.2. *Suppose that Assumptions 3 and 4 hold true. Further let $u\in\mathcal{U}_{o,0}$ be any control policy of the form $u=(\nu_0,\nu_1,\cdots,\nu_{K-1})$. Then*

$$\left|J_N(z,u)-\widetilde{J}(z,u)\right|\leq\left\{L_\phi\frac{e^{L_FT}-1}{L_F}+(KL_r+L_\psi)e^{L_FT}\right\}C_{T,N}\,. \tag{3.6}$$

with the the constant $C_{T,N}$ as defined in (3.1) and the Lipschitz constants L_ϕ, L_F, L_r, L_ψ as given in Assumptions 3–4, with the convention

$$\frac{e^{L_FT}-1}{L_F}=0\quad\text{for}\quad L_F=0.$$

Furthermore, under Assumptions 1 and 2, we have

$$\lim_{N\to\infty}|J_N(z,u)-\widetilde{J}(z,u)|=0\,, \tag{3.7}$$

uniformly for all control policies $u\in\mathcal{U}_{o,0}$.

Proof. Exploiting the Lipschitz continuity of the cost functions ϕ, r and ψ (Assumption 4), together with Theorem 3.1, we can conclude that

$$\begin{aligned}\left|J_N(z,u)-\widetilde{J}(z,u)\right| &= \Bigg|\mathbf{E}^u_z\Bigg[\sum_{j=0}^{K-1}\Big(r(z^{u,N}(t_j),\nu_j)-r(\tilde{z}^u(t_j),\nu_j)\\ &\quad+\int_{t_j}^{t_{j+1}}\big(\phi(z^{u,N}(s),\nu_j)-\phi(\tilde{z}^u(s),\nu_j)\big)\,ds\Big)+\psi(z^{u,N}(T))-\psi\big(\tilde{z}^u(T)\big)\Bigg]\Bigg|\\ &\leq\sum_{j=0}^{K-1}\left\{L_r\mathbf{E}^u_z\big[\big|z^{u,N}(t_j)-\tilde{z}^u(t_j)\big|\big]+L_\phi\int_{t_j}^{t_{j+1}}\mathbf{E}^u_z\big[\big|z^{u,N}(s)-\tilde{z}^u(s)\big|\big]\,ds\right\}\\ &\quad+L_\psi\mathbf{E}^u_z\big[\big|z^{u,N}(T)-\tilde{z}^u(T)\big|\big]\\ &\leq\sum_{j=0}^{K-1}\left\{L_\phi\int_{t_j}^{t_{j+1}}C_{T,N}\,e^{L_Fs}ds+L_rC_{T,N}\,e^{L_Ft_j}\right\}+L_\psi C_{T,N}\,e^{L_FT}\\ &\leq\left\{L_\phi\frac{e^{L_FT}-1}{L_F}+(KL_r+L_\psi)e^{L_FT}\right\}C_{T,N}\,.\end{aligned}$$

This proves the assertion. ☐

Closed loop control. Now we consider the case of a feedback control problem. In accordance with (2.3), we define the conditional cost functional for $u \in \mathcal{U}_{f,k}$, $0 \le k < K$ and the corresponding value function as

$$\begin{aligned} J_N(z,u,k) &= \mathbf{E}^u_{t_k,z}\left[\sum_{j=k}^{K-1}\Big(r(z^{u,N}(t_j),\nu_j) + \int_{t_j}^{t_{j+1}} \phi(z^{u,N}(s),\nu_j)\,ds\Big) + \psi\big(z^{u,N}(T)\big)\right], \\ U_N(z,k) &= \inf_{u\in\mathcal{U}_{f,k}} J_N(z,u,k)\,, \end{aligned} \tag{3.8}$$

with the shorthand $\mathbf{E}^u_{t_k,z}[\cdot] = \mathbf{E}^u[\,\cdot\,|\, z^{u,N}(t_k) = z]$ for the conditional expectation over all realizations of the controlled process starting at $z^{u,N}(t_k) = z$. By definition, the value function, also called *optimal cost-to-go*, is the minimum control value from t_k to T as a function of the initial data (z, t_k). In particular, it holds that $U_N(z,K) = \psi(z)$.

Then in complete analogy with the above definitions, we define

$$\begin{aligned} \widetilde{J}(z,u,k) &= \sum_{j=k}^{K-1}\Big(r(\tilde{z}^u(t_j),\nu_j) + \int_{t_j}^{t_{j+1}} \phi(\tilde{z}^u(s),\nu_j)\,ds\Big) + \psi\big(\tilde{z}^u(T)\big)\,, \quad u \in \mathcal{U}_{o,k}\,, \\ \widetilde{U}(z,k) &= \inf_{u\in\mathcal{U}_{o,k}} \widetilde{J}(z,u,k)\,. \end{aligned}$$

to be the cost functional and the value function of the deterministic limit process. In what follows, we will omit the dependence of J_N, $\widetilde{J}$ and U_N on k when $k = 0$ so that the notations are consistent with (2.3) and (2.8).

By the dynamic programming principle [27], the necessary conditions for optimality are given in terms of Bellman's equations for the two value functions :

$$\begin{aligned} U_N(z,k) &= \inf_{\nu\in\mathcal{A}} \mathbf{E}^\nu\left[r(z,\nu) + \int_{t_k}^{t_{k+1}} \phi(z^{\nu,N}(s),\nu)\,ds + U_N(z^{\nu,N}(t_{k+1}),k+1)\right], \\ \widetilde{U}(z,k) &= \inf_{\nu\in\mathcal{A}} \left\{r(z,\nu) + \int_{t_k}^{t_{k+1}} \phi(\tilde{z}^\nu(s),\nu)\,ds + \widetilde{U}(\tilde{z}^\nu(t_{k+1}),k+1) \colon \tilde{z}^\nu(t_k) = z\right\}, \end{aligned} \tag{3.9}$$

with $0 \le k \le K-1$ and the terminal conditions

$$U_N(z,K) = \widetilde{U}(z,K) = \psi(z)\,. \tag{3.10}$$

Notice that in (3.9), we have used the notation $\mathbf{E}^\nu = \mathbf{E}^\nu_{t_k,z}$ for the conditional expectation and $z^{\nu,N}(t)$, $\tilde{z}^\nu(t)$ for the processes, since the involved quantities only depend on the control ν selected at t_k rather than the whole control policy.

Before we proceed, we shall first introduce some constants in order to simplify the notations later on. Let $h = \max\{|t_{j+1} - t_j| \colon 0 \le j \le K-1\}$. In accordance with (3.1), we set

$$C_{h,N} = h\,\omega_N + \frac{\alpha}{2(\alpha-1)}\left(\frac{4h\,M_N}{\alpha-1}\right)^{1/\alpha}\,. \tag{3.11}$$

We also introduce the sequences of numbers a_k, b_k, $0 \le k \le K$, satisfying the recursive relations

$$\begin{aligned} a_k &= L_r + L_\phi e^{L_F h} h + a_{k+1} e^{L_F h}\,, \\ b_k &= L_\phi C_{h,N} e^{L_F h}(t_{k+1} - t_k) + a_{k+1} C_{h,N} e^{L_F h} + b_{k+1}\,, \end{aligned} \tag{3.12}$$

for $0 \le k \le K-1$ and $a_K = L_\psi$, $b_K = 0$. The last two expressions can be made more explicit :

$$
\begin{aligned}
a_k &= \left(L_r + L_\phi e^{L_F h} h\right) \frac{e^{L_F h(K-k)} - 1}{e^{L_F h} - 1} + L_\psi e^{(K-k)L_F h}\,, \\
b_k &= C_{h,N} e^{L_F h} \left\{ L_\phi (T - t_k) + \left[\frac{L_r + L_\phi e^{L_F h} h}{e^{L_F h} - 1} \left(\frac{e^{L_F h(K-k)} - 1}{e^{L_F h} - 1} - (K-k) \right) \right.\right. \\
&\quad \left.\left. +\; L_\psi \frac{e^{L_F h(K-k)} - 1}{e^{L_F h} - 1} \right] \right\}
\end{aligned}
\tag{3.13}
$$

for $0 \le k \le K$. Notice that under Assumptions 1 and 2, both $C_{h,N}$ and b_k go to zero as $N \to \infty$. We have the following approximation property of the limiting value function.

THEOREM 3.3. *Suppose that Assumptions 3 and 4 hold. Let $z_0 \in \mathbb{X}_N$ be fixed and $z \in \mathbb{X}_N$ be random with $\mathbf{E}[|z - z_0|] < \infty$. Then, for $0 \le k \le K$, we have*

$$
\mathbf{E}[|U_N(z,k) - \widetilde{U}(z_0,k)|] \le a_k \mathbf{E}[|z - z_0|] + b_k\,. \tag{3.14}
$$

with a_k, b_k as given by (3.12) or (3.13). Further suppose that $u \in \mathcal{U}_{o,0}$ is the optimal (open loop) control policy for the process $\tilde{z}^u$, i.e. $\widetilde{J}(z_0,u) = \widetilde{U}(z_0,0)$. Then

$$
|J_N(z_0,u) - U_N(z_0,0)| \le b_0 + \left\{ L_\phi \frac{e^{L_F T} - 1}{L_F} + \left(K L_r + L_\psi\right) e^{L_F T} \right\} C_{T,N}\,. \tag{3.15}
$$

Especially, if Assumptions 1–2 are met, it holds that

$$
\lim_{N\to\infty} |J_N(z_0,u) - U_N(z_0,0)| = 0.
$$

Proof. We first prove (3.14) by backward induction from $k = K$ to $k = 0$. To this end, let $\mathbf{E}[\cdot]$ denote the expectation with respect to the random variable $z \in \mathbb{X}_N$. For $k = K$, the terminal condition (3.10) and the Lipschitz continuity of the terminal cost ψ imply that

$$
\mathbf{E}[|U_N(z,K) - \widetilde{U}(z_0,K)|] = \mathbf{E}[|\psi(z) - \psi(z_0)|] \le L_\psi \mathbf{E}[|z - z_0|]\,,
$$

therefore (3.14) holds with $a_K = L_\psi$, $b_K = 0$. Now suppose that (3.14) is true for $k+1 \le K$. Then, using the Bellman equation (3.9) for the value function together with the Lipschitz continuity of r, l (cf. Assumption 4), it follows that

$$
\begin{aligned}
\mathbf{E}[|U_N(z,k) - \widetilde{U}(z_0,k)|] \le\; &\mathbf{E}\Bigg[\sup_{\nu\in\mathcal{A}} \Bigg\{ \left|r(z,\nu) - r(z_0,\nu)\right| \\
&+ \mathbf{E}^\nu \left[\left| \int_{t_k}^{t_{k+1}} \left(\phi(\tilde{z}^\nu(s),\nu) - \phi(z^{\nu,N}(s),\nu)\right) ds \right| \right] \\
&+ \mathbf{E}^\nu \left[\left| U_N(z^{\nu,N}(t_{k+1}),k+1) - \widetilde{U}(\tilde{z}^\nu(t_{k+1}),k+1) \right| \right] \Bigg\} \Bigg] \\
\le\; &\mathbf{E}\Bigg[L_r |z - z_0| + L_\phi \sup_{\nu\in\mathcal{A}} \int_{t_k}^{t_{k+1}} \mathbf{E}^\nu |\tilde{z}^\nu(s) - z^{\nu,N}(s)|\, ds \\
&+ \sup_{\nu\in\mathcal{A}} \mathbf{E}^\nu \left[\left| U_N(z^{\nu,N}(t_{k+1}),k+1) - \widetilde{U}(\tilde{z}^\nu(t_{k+1}),k+1) \right| \right] \Bigg]\,.
\end{aligned}
$$

Using Theorem 3.1 and the inductive step $k+1 \mapsto k$, we conclude

$$\begin{aligned}\mathbf{E}|U_N(z,k) - \widetilde{U}(z_0,k)| \leq & \mathbf{E}\Big[L_r|z-z_0| + L_\phi\big(|z-z_0| + C_{h,N}\big)e^{L_F h}(t_{k+1}-t_k) \\ & + \sup_{\nu\in\mathcal{A}}\big\{a_{k+1}\mathbf{E}^\nu|z^{\nu,N}(t_{k+1}) - \tilde{z}^\nu(t_{k+1})| + b_{k+1}\big\}\Big] \\ \leq & \big(L_r + L_\phi e^{L_F h}h + a_{k+1}e^{L_F h}\big)\,\mathbf{E}[|z-z_0|] \\ & + L_\phi C_{h,N}e^{L_F h}(t_{k+1}-t_k) + a_{k+1}C_{h,N}e^{L_F h} + b_{k+1} \\ = & a_k\mathbf{E}[|z-z_0|] + b_k\,,\end{aligned}$$

where the recursive relation (3.12) has been used in the last equation. This proves (3.14) for all $k \leq K$. Equation (3.15) now follows from (3.14) and Theorem 3.2, using the triangle inequality :

$$\begin{aligned}0 \leq & J_N(z_0,u) - U_N(z_0,0) \\ \leq & |J_N(z_0,u) - \widetilde{J}(z_0,u)| + |\widetilde{U}(z_0,0) - U_N(z_0,0)| \\ \leq & b_0 + \left(L_\phi\frac{e^{L_F T}-1}{L_F} + \big(KL_r + L_\psi\big)e^{L_F T}\right)C_{T,N}\,.\end{aligned}$$

Convergence $|J_N(z_0,u) - U_N(z_0,0)| \to 0$ as $N \to \infty$ readily follows from Assumptions 1–2. □

Remark 3. *As discussed in Remark 2, we have* $C_{T,N} = \mathcal{O}(N^{1/\alpha-1})$ *and thus* $b_0 = \mathcal{O}(N^{1/\alpha-1})$ *for the density dependent Markov chain of Subsection 2.3. As a consequence, we can explicitly compute the order of convergence in Theorems 3.2 and 3.3, viz.*

$$|J_N(z,u) - \widetilde{J}(z,u)| \leq CN^{1/\alpha-1}\,, \quad u \in \mathcal{U}_{o,0}\,,$$

and

$$|J_N(z_0,u^*) - U_N(z_0,0)| \leq CN^{1/\alpha-1}\,,$$

with $C>0$ *being a generic constant,* u^* *being the optimal open loop policy for the limiting process* $\tilde{z}^u$*, and* U_N *being the value function of the stochastic feedback optimal control problem.*

3.3. Optimal control on an infinite time-horizon with discounted cost. As a final step of our analysis, we consider optimal control problems with cost functional

$$J_N(z,u) = \mathbf{E}_z^u\left[\sum_{j=0}^{\infty} e^{-\beta t_j}\left(r(z^u(t_j),\nu_j) + \int_{t_j}^{t_{j+1}}\phi(z^u(s),\nu_j)\,ds\right)\right]\,, \tag{3.16}$$

where $\beta > 0$ is a discount factor. For simplicity, we assume that the control set $\mathcal{A}$ is finite and that the time stages at which the controls can be changed are uniformly disctributed, i.e. $t_j = jh$ for some $h > 0$.

The analysis of the (discounted) open loop control problem on an infinite time-horizon is similar to the previous situation (see Theorem 3.2), which is why it is omitted. Instead we shall focus on the feedback control problem. Let

$$\mathcal{U}_f = \{(\nu_0,\nu_1,\cdots) \mid \nu_j : \mathbb{X}_N \to \mathcal{A}\,, \quad 0 \leq j < \infty\}\,. \tag{3.17}$$

It is known (e.g. [27]) that the value function $U_N(z) = \inf_{u\in\mathcal{U}_f} J_N(z,u)$ solves the Bellman equation

$$U_N(z) = \min_{\nu\in\mathcal{A}} \mathbf{E}_z^\nu \left[r(z,\nu) + \int_0^h \phi(z^{\nu,N}(s),\nu)ds + \lambda U_N(z^{\nu,N}(h)) \right] , \tag{3.18}$$

where $\lambda = e^{-\beta h} < 1$. It is moreover known [27] that there is a map $\pi_N \colon \mathbb{X}_N \to \mathcal{A}$, such that $u_{opt} = (\pi_N(\cdot), \pi_N(\cdot), \cdots) \in \mathcal{U}_f$ is an optimal feedback policy that satisfies $U_N(z) = J_N(z, u_{opt})$ and can be determined by the dynamic programming (i.e. Bellman) equation via

$$\pi_N(z) \in \operatorname*{argmin}_{\nu\in\mathcal{A}} \left\{ r(z,\nu) + \mathbf{E}_z^\nu \left[\int_0^h \phi(z^{\nu,N}(s),\nu)\, ds + \lambda U_N(z^{\nu,N}(h)) \right] \right\} , \quad z \in \mathbb{X}_N \, .$$

In correspondence with the stochastic control problem, the natural candidate for the cost functional of the deterministic limit dynamics $\tilde{z}^u(\cdot)$ reads

$$\widetilde{J}(z,u) = \sum_{j=0}^{\infty} e^{-\beta j h} \left(r(\tilde{z}^u(jh),\nu_j) + \int_{jh}^{(j+1)h} \phi(\tilde{z}^u(s),\nu_j)\, ds \right) , \tag{3.19}$$

where $u \in \mathcal{U}_o := \{\, (\nu_0, \nu_1, \cdots) \mid \nu_j \in \mathcal{A} \, , \; 0 \le j < \infty \}$. By the dynamic programming principle, the corresponding value function $\widetilde{U}(z) = \inf_{u\in\mathcal{U}_0} \widetilde{J}(z,u)$ is the solution of

$$\widetilde{U}(z) = \min_{\nu\in\mathcal{A}} \left\{ r(z,\nu) + \int_0^h \phi(\tilde{z}^\nu(s),\nu)\, ds + \lambda \widetilde{U}(\tilde{z}^\nu(h)) \right\} , \tag{3.20}$$

where $\tilde{z}^\nu(0) = z$ (note that the value functions of our infinite time-horizon control problems do not explicitly depend on time).

Additional assumptions. Now call $\mathbb{X}_\infty \subset \mathbb{R}^n$ the state space of $\tilde{z}^\nu$ and assume that a map $\pi_\infty : \mathbb{X}_\infty \to \mathcal{A}$ exists such that

$$\pi_\infty(z) \in \operatorname*{argmin}_{\nu\in\mathcal{A}} \left\{ r(z,\nu) + \int_0^h \phi(\tilde{z}^\nu(s),\nu)\, ds + \lambda \widetilde{U}(\tilde{z}^\nu(h)) \right\} , \quad z \in \mathbb{X}_\infty \, , \tag{3.21}$$

where, again, $\tilde{z}^\nu(0) = z$ and $u_{z,opt} = (\nu_0, \nu_1, \cdots) \in \mathcal{U}_o$, with $\nu_j = \pi_\infty(\tilde{z}^u(jh;\, z))$, being the optimal open loop control policy. For the following analysis, Assumption 4 is replaced by the stricter condition that the functions ϕ, r are not only Lipschitz but also bounded :

Assumption $4'$. *There exist constants $L_r, L_\phi, M_r, M_\phi > 0$, such that for all $\nu \in \mathcal{A}$,*

$$|r(z,\nu) - r(z',\nu)| \le L_r |z - z'| \, , \quad |\phi(z,\nu) - \phi(z',\nu)| \le L_\phi |z - z'| \, , \quad z, z' \in \mathbb{R}^n \, .$$

Moreover $|r(z,\nu)| \le M_r$, $|\phi(z,\nu)| \le M_\phi$, for all $\nu \in \mathcal{A}$ and $z \in \mathbb{R}^n$.

The assumption implies that

$$\widetilde{J}(z,u) \le \sum_{j=0}^{\infty} e^{-\beta j h} \left(M_r + \int_{jh}^{(j+1)h} M_\phi \, ds \right) = \frac{M_r + M_\phi h}{1 - e^{-\beta h}} =: M_J \, . \tag{3.22}$$

Similarly, $J_N(z,u) \le M_J$ and therefore the same upper bound applies to $\widetilde{U}(z)$ and $U_N(z)$. The next theorem states that the value function of the stochastic control problem converges to the value function of the corresponding limiting ODE.

Theorem 3.4. *Suppose that Assumptions 3 and $4'$ hold. Then*

1. For every $\epsilon > 0$, there exists $C_\epsilon > 0$, such that

$$\sup_{|z-z'|\le R} |\widetilde{U}(z) - \widetilde{U}(z')| \le C_\epsilon R + \epsilon\,, \quad R \ge 0\,.$$

In particular, $\widetilde{U}$ is uniformly continuous on $\mathbb{X}_\infty$.

2. Further suppose that Assumptions 1 and 2 are met. Then, for all $\epsilon > 0$, there exists $N' \in \mathbb{N}$, such that

$$\Delta U_N := \sup_{z\in\mathbb{X}_N} |U_N(z) - \widetilde{U}(z)| \le \epsilon \quad \forall N \ge N'\,.$$

3. Under Assumptions 1 and 2 and given an ϵ'-optimal policy $u_z = (\nu_0, \nu_1, \cdots) \in \mathcal{U}_o$ with $\nu_j = \pi(\tilde{z}^{u_z}(jh\,;\,z))$, $\pi : \mathbb{X}_\infty \to \mathcal{A}$ for the (3.19) that satisfies

$$\widetilde{U}(z) \le \widetilde{J}(z, u_z) \le \widetilde{U}(z) + \epsilon' \quad \forall z \in \mathbb{X}_\infty\,,$$

the following holds true : for all $\epsilon > \epsilon'/(1-\lambda)$ there exists a constant $N' \in \mathbb{N}$, such that

$$J_N(z, u_z) \le U_N(z) + \epsilon \quad \forall N \ge N'\,,\ z \in \mathbb{X}_N\,.$$

That is, u_z is an ϵ-optimal control policy for the feedback optimal control problem (3.16).

Proof.

1. Given $z, z' \in \mathbb{X}_\infty$, let $\nu = \pi_\infty(z)$. By the Lipschitz continuity of the cost function, Assumption 4′, and (3.20)–(3.21), we have

$$\begin{aligned}\widetilde{U}(z') - \widetilde{U}(z) \le & r(z',\nu) + \int_0^h \phi(\tilde{z}^\nu(s;z'),\nu)\,ds + \lambda\widetilde{U}(\tilde{z}^\nu(h;z')) \\ & - r(z,\nu) - \int_0^h \phi(\tilde{z}^\nu(s;z),\nu)\,ds - \lambda\widetilde{U}(\tilde{z}^\nu(h;z)) \\ \le & L_r|z-z'| + \int_0^h L_\phi\big|\tilde{z}^\nu(s;z') - \tilde{z}^\nu(s;z)\big|\,ds \\ & + \lambda\big|\widetilde{U}(\tilde{z}^\nu(h;z)) - \widetilde{U}(\tilde{z}^\nu(h;z'))\big|\,,\end{aligned} \tag{3.23}$$

where we use the notation $\tilde{z}^\nu(s;z)$ to denote the solution of the ODE (2.5) with initial condition $\tilde{z}^\nu(0;z) = z$ and $\tilde{z}^\nu(s;z')$ for the initial value z'. By Assumption 3,

$$\big|\tilde{z}^\nu(t;z) - \tilde{z}^\nu(t;z')\big| \le e^{L_F t}\big|z - z'\big|\,, \quad t \ge 0. \tag{3.24}$$

Now for all $R \ge 0$, we define the function

$$G(R) = \sup_{|z-z'|\le R} \big|\widetilde{U}(z) - \widetilde{U}(z')\big|\,. \tag{3.25}$$

From (3.22), it readily follows that $G(R) \le M_J$ since $\widetilde{U}$ is non-negative. By combining (3.23) and (3.24), we find

$$G(R) \le \left(L_r + \frac{L_\phi\big(e^{L_F h} - 1\big)}{L_F}\right) R + \lambda G\big(e^{L_F h}R\big)\,,$$

which, upon iterating the above inequality k times, leads to

$$G(R) \le \left(L_r + \frac{L_\phi(e^{L_F h} - 1)}{L_F}\right) \frac{1-\lambda^k e^{L_F kh}}{1-\lambda e^{L_F h}} R + \lambda^k M_J\,. \tag{3.26}$$

The first conclusion follows by noticing that $\lambda < 1$.

2. We proceed with our comparison of the value functions $U_N(z)$ and $\widetilde{U}(z)$. From (3.22) and the non-negativity of $\widetilde{U}$ and U_N, we conclude that $\Delta U_N \le M_J$. Now let $z \in \mathbb{X}_N \subset \mathbb{X}_\infty$ with $\pi_\infty(z) = \nu \in \mathcal{A}$. Using the dynamic programming equations (3.18) and (3.20), Theorem 3.1, the definition (3.11) of the constant $C_{h,N}$, we conclude that

$$
\begin{aligned}
U_N(z) - \widetilde{U}(z) \le& \mathbf{E}^\nu\left[\int_0^h \phi(z^{\nu,N}(s),\nu)\,ds + \lambda U_N(z^{\nu,N}(h))\right] \\
&- \int_0^h \phi(\tilde{z}^\nu(s),\nu)\,ds - \lambda \widetilde{U}(\tilde{z}^\nu(h)) \\
\le& \mathbf{E}^\nu\left[\int_0^h L_\phi\big|z^{\nu,N}(s) - \tilde{z}^\nu(s)\big|\,ds + \lambda U_N(z^{\nu,N}(h)) - \lambda\widetilde{U}(z^{\nu,N}(h))\right] \\
&+ \lambda\mathbf{E}^\nu\Big[\big|\widetilde{U}(z^{\nu,N}(h)) - \widetilde{U}(\tilde{z}^\nu(h))\big|\Big] \\
\le& \frac{L_\phi}{L_F} C_{h,N}\big(e^{L_F h} - 1\big) + \lambda\Delta U_N + \lambda\mathbf{E}^\nu G\big(|z^{\nu,N}(h) - \tilde{z}^\nu(h)|\big)\,.
\end{aligned}
$$

Since the same upper bound for $\Delta U_N := \widetilde{U}(z) - U_N(z)$ is obtained by taking the supremum over $z \in \mathbb{X}_N$ and using (3.26), it follows that

$$
\begin{aligned}
\Delta U_N \le& \frac{1}{1-\lambda}\left\{C_{h,N}\left[\frac{L_\phi}{L_F}\big(e^{L_F h} - 1\big) + \lambda\left(L_r + \frac{L_\phi(e^{L_F h} - 1)}{L_F}\right)\right.\right. \\
&\left.\left.\times\ \frac{1-\lambda^k e^{L_F k h}}{1-\lambda e^{L_F h}} e^{L_F h}\right] + \lambda^{k+1} M_J\right\}
\end{aligned}
$$

holds for all $k \in \mathbb{N}$. Assumptions 1 and 2 moreover entail that $C_{h,N} \to 0$, as $N \to \infty$. Hence the upper limit for ΔU_N can be made arbitrarily small as follows: Let $\epsilon > 0$ and choose some $k \in \mathbb{N}$, such that $\lambda^{k+1} M_J \le (1-\lambda)\epsilon/2$. Then pick $N' \in \mathbb{N}$, such that the first term in the sum inside the curly brackets is bounded by $(1-\lambda)\epsilon/2$ whenever $N \ge N'$; as a consequence, $\Delta U_N \le \epsilon$ for all $N \ge N'$.

3. We consider an ϵ'-optimal policy $\pi\colon \mathbb{X}_\infty \to \mathcal{A}$ and define $u_z = (\nu_1, \nu_2, \cdots) \in \mathcal{U}_o$ with $\nu_j = \pi(\tilde{z}^u(jh; z))$, such that $\widetilde{U}(z) \le \widetilde{J}(z, u_z) \le \widetilde{U}(z) + \epsilon'$ for all $z \in \mathbb{X}_\infty$. Now let $\nu = \pi(z)$, $z \in \mathbb{X}_N \subset \mathbb{X}_\infty$. Then, using the upper bound for ΔU_N,

$$
\begin{aligned}
&r(z,\nu) + \mathbf{E}^\nu\left[\int_0^h \phi\big(z^{\nu,N}(s),\nu\big)\,ds + \lambda U_N(z^{\nu,N}(h))\right] \\
=& r(z,\nu) + \int_0^h \phi(\tilde{z}^\nu(s),\nu)\,ds + \lambda\widetilde{U}\big(\tilde{z}^\nu(h)\big) \\
&+ \mathbf{E}^\nu\left[\int_0^h \big(\phi\big(z^{\nu,N}(s),\nu\big) - \phi(\tilde{z}^\nu(s),\nu)\big)\,ds + \lambda U_N(z^{\nu,N}(h)) - \lambda\widetilde{U}(\tilde{z}^\nu(h))\right] \\
\le& \widetilde{J}(z,u_z) + \frac{L_\phi}{L_F} C_{h,N}\big(e^{L_F h} - 1\big) + \lambda\Delta U_N + \lambda\mathbf{E}^\nu\big[G\big(|\tilde{z}^\nu(h) - z^{\nu,N}(h)|\big)\big] \\
\le& \widetilde{U}(z) + \epsilon' + \delta' \\
\le& U_N(z) + \epsilon' + 2\delta'\,.
\end{aligned}
$$

Together with the Bellman equation (3.18), the above inequality yields

$$
J_N(z,u_z) \le U_N(z) + \frac{\epsilon' + 2\delta'}{1-\lambda}\,, \quad z \in \mathbb{X}_N\ ,
$$

by which the assertion is proved since $\delta' > 0$ can be made arbitrarily small.

□

4. Algorithms. In this section we discuss some numerical aspects of the control problems studied in this paper. In contrast to the previous sections, this part involves some heuristics, and we confine ourselves to the optimal control problem for a jump process on a finite time-horizon $[0, T]$ with a finite control set $\mathcal{A}$. To this end, we assume that the parameter N is large, but finite, and we remind the reader again that $x^{u,N}$ denotes the original Markov jump process with a control policy u and $z^{u,N} = N^{-1}x^{u,N}$ stands for the normalized process. The state spaces on which $x^{u,N}$ and $z^{u,N}$ live are denoted by $\mathbb{X}$ and $\mathbb{X}_N$.

4.1. Tau-leaping method. In order to compute the optimal control policy, it is necessary to simulate trajectories of the underlying Markov jump process and to estimate the corresponding cost. The stochastic simulation algorithm (SSA) [9, 10, 13] is a typical Monte Carlo method: At each time step, it determines the waiting time in (2.1) as well as the next state according to the jump rates between the current state and the next possible states. When N is large, however, the system becomes numerically stiff because a large number of jump events occur within a short time interval. Since SSA traces every single jump event of the system, the effective step size of the method decreases rapidly, which renders the SSA inefficient.

As a remedy to this problem, the tau-leaping method [12, 28, 5, 15, 13] aims at increasing the effective step size by updating the state vector according to the transitions that may occur within a given time interval. Roughly speaking, instead of computing the waiting time and the next jump, the idea of the tau-leaping method is to ask "how many times will each jump occur within a given time interval" and then update the state vector accordingly. With a proper and carefully chosen step size [5], the tau-leaping method can approximate the SSA quite well and meanwhile reduce the simulation time up to 1 or 2 orders of magnitude.

4.2. State space truncation. While our previous analysis suggested that the optimal open loop control of the limiting ODE system may be a reasonable approximation whenever N is sufficiently large, there may be situations where this criterium is not met. Consequently, it may be necessary to compute a feedback control, at least for parts of the state space.

The computational complexity for solving the feedback optimal control problem is proportional to the number of states $\mathbb{X}$ considered (which grows of the order N^n, with n being the number of species). Therefore, truncating the state space $\mathbb{X}$ is necessary before numerically solving for the optimal feedback control. One such approach to truncate the state space is to consider only states $x = (x^1, x^2, \cdots, x^n) \in \mathbb{X}$ that lie with a hypercube defined by $x^i \in [c_i N, c'_i N]$, $1 \le i \le n$, where $0 \le c_i < c'_i$ are minimum and maximum average densities per species. The cut-off values c_i, c'_i could, e.g., be determined by launching a couple of independent simulations of the jump process controlled by candidate open loop control policies.

Once a truncated state space $\mathbb{X}_{cut}$ has been constructed, then a simple algorithm (Algorithm 1) to compute the optimal feedback control policy can be based on the necessary optimality condition (3.9) with terminal condition $U_N(\cdot, K) = \psi$ where the expectation value in (3.9) is estimated by a Monte Carlo average. The algorithm is summarized below. If T is the total simulation time, $\Delta t > 0$ is the average time step size used to generate trajectories (e.g. by SSA or tau-leaping) and we use M independent realizations for each starting state to approximate the expectation value, the overall computational cost of Algorithm 1 is $\mathcal{O}\big(M \cdot |\mathcal{A}| \cdot |\mathbb{X}_{cut}| \cdot \lceil T/\Delta t \rceil\big)$.

Algorithm 1 Compute the optimal feedback control policy on truncated state space

Set $U_N(\cdot, K) = \psi$.
for $k \leftarrow K-1$ to 0 **do**
  for each $x \in \mathbb{X}_{cut}$ **do**
    for each $\nu \in \mathcal{A}$ **do**
      Set $u = (\nu, \nu_{k+1}, \cdots, \nu_{K-1})$.
      Starting from x at time t_k, generate M trajectories $x_i^{u,N}$ till time t_{k+1}, such that $x_i^{u,N}(t_{k+1}) \in \mathbb{X}_{cut}$ (generate new realization if $x_i^{u,N}(t_{k+1}) \notin \mathbb{X}_{cut}$).
      Let $z = x/N$, $z_i^{u,N} = x_i^{u,N}/N$, compute

$$J(\nu) = \frac{1}{M}\sum_{i=1}^{M}\left(r(z,\nu) + \int_{t_k}^{t_{k+1}} \phi(z_i^{u,N}(s),\nu)\,ds + U_N(z_i^{u,N}(t_{k+1}), k+1)\right).$$

    end for
    Set $\nu_k(z) = \operatorname*{argmin}_{\nu\in\mathcal{A}} J(\nu)$ and $U_N(z,k) = \min_{\nu\in\mathcal{A}} J(\nu)$.
  end for
end for

4.3. Hybrid control. Solving the feedback control problem may be computationally infeasible even after truncation of the state space. To overcome this limitation one may utilize an adaptive state space truncation strategy that exploits information from open loop control policies that are considered reasonable approximations of the optimal feedback control. Based on the analysis in the previous sections, the key idea is to assume that the typical states visited when an optimal open loop policy is applied are also important for computing a sufficiently accurate feedback control policy. To this end, at each stage t_j, the algorithm generates states whose density is scattered around the values of the system controlled by reasonable open loop approximations.

Adaptive truncation strategy. Let $\mathcal{S}_j \subset \mathbb{X}$ denote the finite state set after truncation at the j-th control stage, $0 \le j < K$. We construct $\mathcal{S}_j$ as follows:

1. Compute good (open loop) candidate policies for the Markov jump process. A control policy $u_k \in \mathcal{U}_{o,0}$ is called "good" if $k \le n_{ol}$ and $J(u_k) \le (1+\epsilon_{ol})J(u_0)$ for appropriately chosen $n_{ol} \in \mathbb{N}$, $n_{ol} \ge 1$ and $\epsilon_{ol} \ge 0$. Sort all "good" control policies $u_k \in \mathcal{U}_{o,0}$ by their costs in non-decreasing order.

2. Compute the statistics of the resulting candidate jump processes. For each policy u_k, record the average density $z_{k,j} \in \mathbb{R}^n$ and the standard deviation $\sigma_{k,j} \in \mathbb{R}^n$ at each stage j, $0 \le j < K$.

3. Compute trust regions $\mathcal{S}_j$. For every "good" open loop policy u_k, generate a large number M_{ol} of trajectories and add each trajectory's state $x \in \mathbb{X}$ at stage j to the set $\mathcal{S}_j$ if

$$x^i/N \in \left[z^i_{k,j} - r\sigma^i_{k,j}, z^i_{k,j} + r\sigma^i_{k,j}\right], \quad \forall i \in \{1,\dots,n\} \tag{4.1}$$

where the parameter $r > 0$ is a pre-selected constant of order 1, and $x^i, z^i_{k,j}, \sigma^i_{k,j}$ are the ith components of x, $z_{k,j}$, $\sigma_{k,j} \in \mathbb{R}^n$.

REMARK 4. *A few remarks about the above algorithm.*

1. *In the case that the jump process starts from a fixed initial value x, $\mathcal{S}_0 = \{x\}$ is a singleton containing only the initial state.*
2. *Step 1 can be accomplished by enumerating all possible $u_k \in \mathcal{U}_{o,0}$ and computing the cost $J(u_k)$ by simulating trajectories using SSA or the tau-leaping method. By the central limit theorem (see [21]), the state distribution of the jump process under "good" open loop policy is approximately Gaussian whenever N is not too small, hence the standard statistical estimators for first and second moments computed in Step 2 can capture the distribution to a very good approximation (however algorithmic adaptations can be implemented to account for more complex distributions).*
3. *Ideally, for every "good" control policy u_k and every control stage j, we would like to record all possible (i.e. reachable) states that satisfy (4.1). However, this set may become very large (i.e. densely filled), especially when n is large. Therefore, Step 3 involves the tunable parameter M_{ol} to determine the number of states in $\mathcal{S}_j$ where trajectories generated in Steps 1 or 2 can be reused. The drawback is that important states may be missing when they are not visited by the first M_{ol} trajectories (see below for a patch.)*

Hybrid control policy. Having the state sets $\mathcal{S}_j$ at hand, the task of computing a feedback control policy is to determine maps $\nu_j : \mathcal{S}_j \to \mathcal{A}$, $0 \le j < K$ according to a modification of Algorithm 1. Keeping in mind that the trust regions $\mathcal{S}_j$ may be only poorly sampled, it is quite possible that, at some control stage j, the system fails to reach $\mathcal{S}_j$ under the action ν_{j-1}. To remedy this defect, we propose the following strategy : Let the best available open loop policy be denoted by $u_0 = (\nu_0^0, \nu_1^0, \cdots, \nu_{K-1}^0)$, and let us consider the j-th control stage, $0 \le j < K$ where we suppose that the system has ended up in a state $x \notin \mathcal{S}_j$. Further let x' be one of the nearest states to x among all states in $\mathcal{S}_j$, i.e. $x' \in \operatorname{argmin}_{x' \in \mathcal{S}_j} |x - x'|$. Then we apply the control $\nu_j(x')$ if $|x - x'|/N \le \epsilon_{near}$, otherwise we use ν_j^0. In other words, we replace the original candidate control by the modified control policy $u = (\bar{\nu}_0, \bar{\nu}_1, \cdots, \bar{\nu}_{K-1}) \in \mathcal{U}_{f,0}$ with

$$\bar{\nu}_j : \mathbb{X} \to \mathcal{A}, \quad \bar{\nu}_j(x) = \begin{cases} \nu_j(x'), & \text{if } |x' - x|/N < \epsilon_{near} \\ \nu_j^0, & \textit{otherwise} \end{cases} \tag{4.2}$$

In the following we keep using ν_j instead of $\bar{\nu}_j$ when no ambigularity will arise. This strategy can prevent problems that arise when the feedback ν_j at stage j cannot be computed because some rare, but important states are missing due to insufficient sampling of the trust region $\mathcal{S}_j$. Notice that the algorithmic modification can be easily switched off by setting $\epsilon_{near} = 0$. In this case, the feedback policy is applied only when states belong to $\mathcal{S}_j$, while open loop policies are applied otherwise. In agreement with the notation used in Sections 1–3, we define

$$\mathcal{U}_{h,k} = \{(\bar{\nu}_k, \bar{\nu}_{k+1}, \cdots, \bar{\nu}_{K-1}) \mid \nu_j : \mathcal{S}_j \to \mathcal{A}, \, k \le j < K\}, \quad 0 \le k < K, \tag{4.3}$$

as the set of all hybrid control policies, where $\bar{\nu}_j$ is defined as in (4.2). The algorithmic task now boils down to find the optimal hybrid control policy $u \in \mathcal{U}_{h,0}$. For solving this task, we consider the cost function $J_N(z, u, k)$ as in (3.8) and define a modified value function as

$$U_N(z, k) = \inf_{u \in \mathcal{U}_{h,k}} J_N(z, u, k), \quad Nz \in \mathcal{S}_k. \tag{4.4}$$

By definition, the value function satisfies the terminal condition $U_N(z, K) = \psi(z)$ and a modified Bellman equation as a necessary optimality condition :

$$\begin{aligned} U_N(z,k) = &\min_{\nu\in\mathcal{A}} \mathbf{E}^{\nu}\Bigg[\sum_{j=k}^{\tau-1}\Big(r\big(z^{u,N}(t_j),\nu_j(z^{u,N}(t_j))\big) \\ &+ \int_{t_j}^{t_{j+1}} \phi\big(z^{u,N}(s),\nu_j(z^{u,N}(t_j))\big)\,ds\Big) + U_N\big(z^{u,N}(t_\tau),\tau\big)\Bigg], \quad Nz\in\mathcal{S}_k\,. \end{aligned} \tag{4.5}$$

where $z^{u,N}(t_k) = z$, $u = (\nu_k, \nu_{k+1}, \cdots, \nu_{K-1})$ with $\nu_k = \nu$ and $(\nu_{k+1}, \cdots, \nu_{K-1}) \in \mathcal{U}_{h,k+1}$ is the optimal hybrid control policy starting from stage $k+1$. The terminal index τ is a stopping time, depending on the particular realization, and is either the smallest stage index such that $k < \tau < K$ and $Nz^{u,N}(t_\tau) \in \mathcal{S}_\tau$, or $\tau = K$ otherwise. Notice that only states z such that $Nz \in \mathcal{S}_k$ contribute to (4.5). Based on it, we can compute the optimal hybrid control policy by backward iteration in Algorithm 2 below.

Algorithm 2 Compute the optimal hybrid control policy

Set $U_N(\cdot, K) = \psi$.
for $k \leftarrow K-1$ to 0 **do**
  for each x such that $x \in \mathcal{S}_k$ **do**
    for each $\nu \in \mathcal{A}$ **do**
      Set $u = (\nu, \nu_{k+1}, \cdots, \nu_{K-1})$.
      Generate M trajectories $x_i^{u,N}$ from time t_k to t_{τ_i} where $k < \tau_i$ and t_{τ_i} is either the first time when $x_i^{u,N}(t_{\tau_i}) \in \mathcal{S}_{\tau_i}$ or $\tau_i = K$, $1 \le i \le M$.
      Let $z_j^{u,N} = x_j^{u,N}/N$, compute

$$\begin{aligned} J(\nu) = &\frac{1}{M}\sum_{i=1}^{M}\Big\{\sum_{j=k}^{\tau_i-1}\Big[r(z_j^{u,N},\nu_j(z_i^{u,N}(t_j))) + \int_{t_j}^{t_{j+1}}\phi\big(z_i^{u,N}(s),\nu_j(z_i^{u,N}(t_j))\big)\,ds\Big] \\ &+ U_N(z_i^{u,N}(t_{\tau_i}),\tau_i)\Big\}. \end{aligned}$$

    end for
    Set $\nu_k(z) = \operatorname*{argmin}_{\nu\in\mathcal{A}} J(\nu)$ and $U_N(z,k) = \min_{\nu\in\mathcal{A}} J(\nu)$.
  end for
end for

A computational bottleneck in computing the hybrid control policy for $\epsilon_{near} > 0$ is the solution of the minimization problem $\min_{x'\in\mathcal{S}_j} |x - x'|$, i.e. to find the nearest neighbor of x in $\mathcal{S}_j$. The computational complexity of an exact minimization based on a pairwise comparison is $\mathcal{O}\big(|\mathcal{S}_j|\big)$, which would increase the computational cost of Algorithm 2 to $\mathcal{O}\big(M\cdot|\mathcal{A}|\cdot|\mathcal{S}_j|^2\cdot\lceil T/\Delta t\rfloor\big)$ (assuming $\tau = k+1$ and $|\mathcal{S}_j|$ are constant for simplicity). However, by employing a so-called *k-d* tree data structure [4] to store the states in $\mathcal{S}_j$, the computational complexity of finding the nearest neighbor can be reduced to $\mathcal{O}\big(\ln|\mathcal{S}_j|\big)$, by which the total computational cost is of the order

$$\mathcal{O}\big(M\cdot|\mathcal{A}|\cdot|\mathcal{S}_j|\cdot\ln|\mathcal{S}_j|\cdot\lceil T/\Delta t\rfloor\big)\,.$$

In the numerical examples in Section 5 below, our implementation uses the ANN (Approximate

Nearest Neighbor) library [24], which provides operations on k-d trees and efficient algorithms for finding the first k-th ($k = 1$ in our case) nearest neighbors.

5. Numerical examples. In this section, we consider several numerical examples in order to demonstrate the analysis and the algorithms discussed in the previous sections.

5.1. Birth-death process. First we consider the one-dimensional birth-death process which can be described as

$$x - 1 \xleftarrow{\kappa_-} x \xrightarrow{\kappa_+} x + 1\,, \tag{5.1}$$

where $x \in \mathbb{N}^+$. We suppose that the process has a density dependent birth rate which is $x \cdot \kappa_+$ when the current state is x and, similarly, $x \cdot \kappa_-$ for the death rate. We fix $T = 3.0$ and $K = 3$, i.e. the control can be switched at time $t = 0.0, 1.0, 2.0$. Two control/parameterizaton sets $\mathcal{A}_1$, $\mathcal{A}_2$ shown in Table 5.1 are considered. Each set contains two controls ν_0, ν_1 that affect the jump rates κ_- and κ_+. For the optimal control problem, let x_t^u be system's state at $t \in [0, T]$ with control $u \in \mathcal{U}_{\sigma,0}$ and set $r(z, \nu) = \psi(z) = 0$, $\phi(z, \nu) = |z - 1.0|$ for $\nu \in \mathcal{A}_i$, $i = 1, 2$, giving the cost function

$$J_N(z_0, u) = \mathbf{E}^u_{z_0}\left[\int_0^3 |z_t^{u,N} - 1.0|\, dt\right]\,, \quad u \in \mathcal{U}_{\sigma,0}\,, \tag{5.2}$$

with $z_t^{u,N} = x_t^{u,N}/N$ and $z_0^{u,N} = z_0$. Fixing $z_0 = 1.2$ and picking one of two control sets $\mathcal{A}_1$, $\mathcal{A}_2$, we shall compare optimal open loop and feedback control policies for the jump process as N increases, as well as the optimal (open loop) control policy for the related deterministic ODE

$$\frac{d\tilde{z}_t^u}{dt} = (\kappa_+ - \kappa_-)\tilde{z}_t^u, \qquad \tilde{z}_0^u = 1.2\,. \tag{5.3}$$

Open loop control. In the case of open loop control, there are $|\mathcal{U}_{o,0}| = 2^3 = 8$ different control policies in total for both the jump process (5.1) and the deterministic ODE (5.3), regardless of the value N, since one of the two controls ν_0, ν_1 can be selected at any of the three control stages. The optimal control is obtained by simply comparing the costs of all 8 possible policies. In Figure 5.1, the evolutions of the mean and the standard deviation of density $z_t^{u,N}$ are shown for different N. For both sets $\mathcal{A}_1$, $\mathcal{A}_2$, it is observed that the standard deviations decrease and the means get closer to that of ODE with the optimal control policy as N grows larger. For the control set $\mathcal{A}_2$, we observe that the suboptimal policy $u_2 = (1, 1, 0)$ is almost as good as the optimal policy $u_1 = (1, 0, 1)$ of the ODE system. (For the ease of notation, we use the index of the control action to denote the policy, e.g. $(1, 0, 1)$ means (ν_1, ν_0, ν_1).) For the jump processes with $N = 40$ or $N = 100$, u_2 is even better than u_1; cf. Figure 5.3.

Feedback control. Now we turn to the feedback control problem, in which case the optimal control policy can be obtained by iterating the dynamic programming equations (3.9)–(3.10) by backward iteration. As the state space $\mathbb{X} = \mathbb{N}^+$ is infinite, finite state truncation is necessary for Algorithm 1 be functional. The form of the cost functional (5.2) and the initial condition $z_0 = 1.2$ suggests to project the dynamics onto the finite subspace $\mathbb{X}_{cut} = \{N/2, N/2 + 1, \cdots, 2N\} \subset \mathbb{N}^+$ with reflecting boundary conditions at $x \in \{N/2,\, 2N\}$ (assuming that $N/2$ is an integer). Note that for large values of N, computing the optimal policy is infeasible, even for the truncated system (cf. Subsection 4.2), which calls for the adaptive truncation strategy.

No.	control	$\mathcal{A}_1$		$\mathcal{A}_2$	
		κ_-	κ_+	κ_-	κ_+
0	ν_0	$\underline{0.6}$	1.0	$\underline{0.8}$	1.0
1	ν_1	1.0	0.8	1.0	0.8

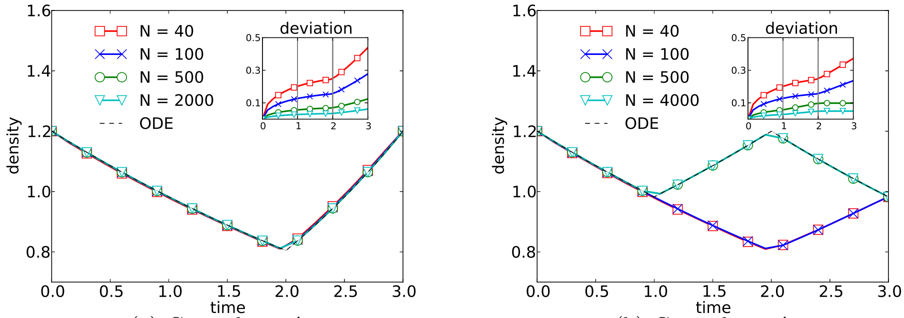
Table 5.1: Two different control sets $\mathcal{A}_1$, $\mathcal{A}_2$ for the birth-death jump process. Each set contains two controls where the underlined entries indicate different control actions in $\mathcal{A}_1$ and $\mathcal{A}_2$.

Figure 5.2 shows the mean and the standard deviation of $z_t^{u,N}$ under the two optimal feedback control policies as a function of time and for increasing N. Generally, for both control set $\mathcal{A}_1$ and $\mathcal{A}_2$, the optimal feedback control policies lead to smaller costs as compared to the optimal open loop controls. Specifically, we observe in Figure 5.2(a) that for $\mathcal{A}_1$ the standard deviations decrease and means converge to the density of the optimally controlled ODE system (by u_2) as N increases. For $\mathcal{A}_2$, due to the existence of another competing policy, u_2, some states with density close to $z = 1.0$ may select the control ν_1 while others select ν_0 (see Table 5.1), which leads to a significant rise in the standard deviation between the control stages $t = 1.0$ and $t = 2.0$ (see Figure 5.2(b)); we moreover notice that the convergence of the first moment of the controlled jump process at time $t = 2.0$ to the ODE solution is slower than in case of the control set $\mathcal{A}_1$ as N increases. The last observation is in agreement with Figure 5.4(a) that shows the bimodal probability density function of the optimally controlled process at time $t = 2.0$ that becomes even more pronounced for larger values of N. Nevertheless, Figure 5.3(a) clearly shows the convergence of the cost values of both open loop and feedback control policies as N increases, inline with the theoretical prediction. As a final demonstration, Figures 5.3(a) and 5.4(b) show a comparison of SSA and tau-leaping, with the clear indication that the results of tau-leaping are close to the SSA prediction, but at much lower computational cost.

Hybrid control. Finally we consider the hybrid control policy following the procedure discussed in Subsection 4.3 where we confine our attention to the control set $\mathcal{A}_2$. To assess the approximation quality of the hybrid control algorithm, we compute the cost under the open loop control policies for various values of N and with 5000 trajectories for each policy. As "good" control policies we define suboptimal controls with $n_{ol} = 2$ and $\epsilon_{ol} = 0.05$ (see page 17). The trust regions $\mathcal{S}_j$ are computed from $M_{ol} = 5000$ realizations for each "good" open loop policy according to (4.1) with $r = 2.5$. As Figure 5.4(c) illustrated, the cardinality of the sets $\mathcal{S}_1$ and $\mathcal{S}_2$ is much smaller than the cardinality of $\mathbb{X}_{cut}$ used in the feedback control case, which can lead to a tremendeous reduction of the computational effort as compared to Algorithm 1 at almost no loss of numerical accuracy (see Figure 5.3).

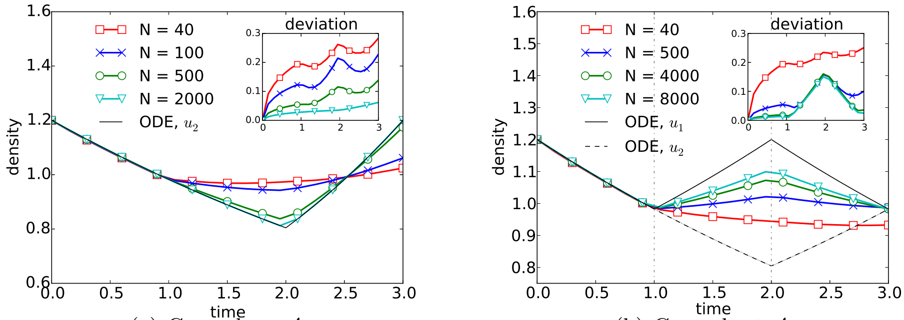
5.2. Predator-prey model. In this section, we consider a two dimensional predator-prey model on the state space $\mathbb{X} = \mathbb{N}^+ \times \mathbb{N}^+$. We call A and B the prey and predator species, and let $x = (x^1, x^2) \in \mathbb{X}$ denote the numbers of species A and B. We suppose that both prey and predator reproduce or decease naturally, with the predator eating the prey in order to reproduce. Recalling the notation explained in Subsection 2.3, the dynamics of A, B species can be modelled as a jump process on $\mathbb{X}$ according to the rules (see [19])

1. $\mathrm{A} \xrightarrow{\lambda_1} 2\,\mathrm{A}$, $\quad \mathrm{A} \xrightarrow{\mu_1} \emptyset$



(a) Control set $\mathcal{A}_1$ (b) Control set $\mathcal{A}_2$

Fig. 5.1: Evolution of empirical mean and standard deviation (inset plot) of the density process $z^{u,N}$ under the optimal open loop control policies in comparison with the ODE solutions. Here N is the scaling number and controls are switched at times $t = 0.0, 1.0, 2.0$. (a) Control Set 1. The optimal policy is $u_2 = (1, 1, 0)$ for the jump process for all N and for the ODE system. (b) Control Set 2. The optimal policy is $u_2 = (1, 1, 0)$ for the jump process with $N = 40, 100$, but it is $u_1 = (1, 0, 1)$ for $N = 500, 4000$ and the ODE system.

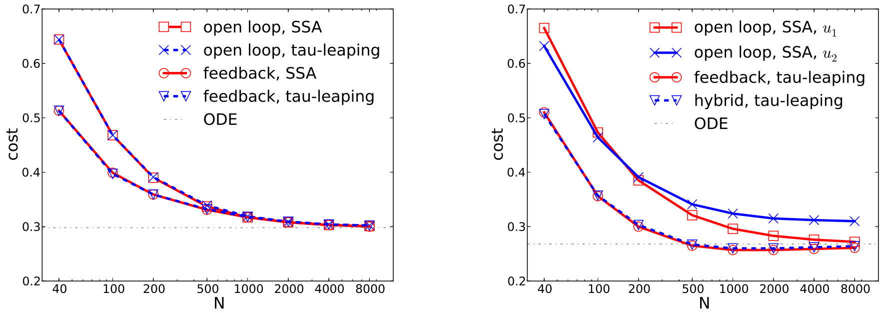


(a) Control set $\mathcal{A}_1$ (b) Control set $\mathcal{A}_2$

Fig. 5.2: Evolution of empirical mean and standard deviation (inset plot) of density the density process $z^{u,N}$ under the optimal feedback control policies in comparison with the ODE solutions. N is the scaling number and controls are switched at times $t = 0.0, 1.0, 2.0$. (a) Control set $\mathcal{A}_1$: as N increases, the standard deviations decrease and the average gets closer to the ODE solution under the optimal policy $u_2 = (1, 1, 0)$. (b) Control set $\mathcal{A}_2$: the policies $u_1 = (1, 0, 1)$ and $u_2 = (1, 1, 0)$ are the dominant (sub)optimal control policies for the ODE system.

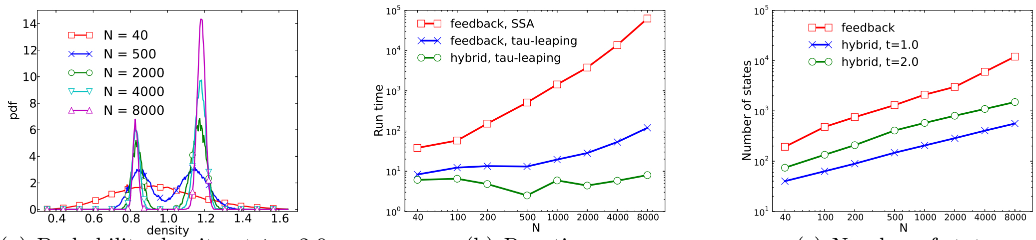
2. $\mathrm{B} \xrightarrow{\lambda_2} 2\,\mathrm{B}$, $\mathrm{B} \xrightarrow{\mu_2} \emptyset$
3. $\mathrm{A} + \mathrm{B} \xrightarrow{b} \mathrm{B}$, $\mathrm{A} + \mathrm{B} \xrightarrow{c} \mathrm{A} + 2\,\mathrm{B}$.

A control corresponds to a vector $\nu = (\lambda_1, \mu_1, \lambda_2, \mu_2, b, c)$, where each parameter assumes only positive real values. Now we define the jump vectors $l_1 = (1, 0)$, $l_2 = (0, 1)$ and consider the normalized state vector $z = (z_1, z_2) = x/N \in \mathbb{R}^2$ for a large, but fixed scaling parameter $N \gg 1$.



(a) Control set $\mathcal{A}_1$ (b) Control set $\mathcal{A}_2$

Fig. 5.3: Cost values for jump process with different scaling number N. Both SSA and tau-leaping methods are used to sample trajectories. For control set $\mathcal{A}_2$, $u_1 = (1, 0, 1)$, $u_2 = (1, 1, 0)$ are the two most optimal open loop policies.



(a) Probability density at $t = 2.0$ (b) Run time (c) Number of states

Fig. 5.4: Dynamics under the control set $\mathcal{A}_2$. (a) shows the probability distribution of states of the jump process at time $t = 2.0$ under the optimal feedback control. (b) displays the CPU run time for different values of N, where the algorithm is run in parallel with 10 processors in each case. (c) gives the number of states in $\mathbb{X}_{cut}$ (feedback control) and $\mathcal{S}_1$, $\mathcal{S}_2$ (hybrid control). Notice that the curve corresponding to the feedback controls is not a straight line because, in our simulation, the size $\mathbb{X}_{cut}$ does not scale linearly with N.

The normalized jump rates are then given by

$$\begin{aligned} f^\nu(z, l_1) &= \lambda_1 z_1 \,, \qquad f^\nu(z, -l_1) = (\mu_1 + b z_2) z_1 \,, \\ f^\nu(z, l_2) &= (\lambda_2 + c z_1) z_2 \,, \qquad f^\nu(z, -l_2) = \mu_2 z_2 \,, \end{aligned} \tag{5.4}$$

giving rise to the limit vector field as $N \to \infty$:

$$F^\nu(z) = \big((\lambda_1 - \mu_1) z_1 - b z_1 z_2 \,,\, c z_1 z_2 - (\mu_2 - \lambda_2) z_2\big). \tag{5.5}$$

Our aim is to study the optimal control problem on a finite time-horizon $[0, T]$, with terminal time $T = 5.0$ and $K = 5$ control stages at times $t = j \times 1.0$, $0 \leq j \leq 4$. We define the cost functional as

$$J_N(z_0, u) = \mathbf{E}^u_{z_0}\left[\int_0^{5.0} \Big(|z_t^{1,u,N} - 2 z_t^{2,u,N}| + |z_t^{1,u,N} - 1.5|\Big)\, dt\right], \quad u \in \mathcal{U}_{\sigma,0}\,, \tag{5.6}$$

No.	control	λ_1	μ_1	λ_2	μ_2	b	c
0	ν_0	2.5	0.2	0.2	2.0	2.0	2.0
1	ν_1	$\underline{2.7}$	0.2	0.2	$\underline{1.5}$	2.0	2.0
2	ν_2	2.5	0.2	0.2	$\underline{2.5}$	2.0	2.0

Table 5.2: The control set $\mathcal{A}$ contains three different controls to modify the rates in the predator-prey model. The major effects are indicated by underlined rates.

N	50	100	200	500	1000	2000	4000
$\Delta_s t$	1.8×10^{-3}	9.0×10^{-4}	4.5×10^{-4}	1.8×10^{-4}	9.0×10^{-5}	4.5×10^{-5}	2.2×10^{-5}
$\Delta_\tau t$	1.8×10^{-3}	9.0×10^{-4}	4.5×10^{-4}	3.9×10^{-4}	1.1×10^{-3}	2.7×10^{-3}	3.2×10^{-3}

Table 5.3: Average time step when SSA (row with label $\Delta_s t$) or tau-leaping (row with label $\Delta_\tau t$) are used to generate realization of the predator-prey model.

where $z_t^{u,N} = (z_t^{1,u,N}, z_t^{2,u,N}) = N^{-1} x_t^{u,N}$ is the normalized jump process with initial condition $z_0^{u,N} = z_0$. In our numerical experiment, we set $z_0 = (1.0, 0.4)$ and choose $N = 50$, 100, 200, 500, 1000, 2000, 4000.

The particular choice of cost functional J_N is motivated by the observation that the stable equilibrium of the prey species is at $z^1 = 1.5$ with roughly about two times more prey than predator. The control set $\mathcal{A}$ contains three different controls and is shown in Table 5.2: Observe that, in comparison with ν_0, the prey reproduces faster under the control ν_1 and predators decease more slowly, while ν_2 has the reverse effect.

Open loop control. We do a brute-force calculation of the optimal open loop control policy based on ordering all possible $3^5 = 243$ policies in $\mathcal{U}_{o,0}$ according to their cost. In each case, 50000 trajectories are sampled using both SSA and tau-leaping. From Table 5.3, we conclude that for large N (≥ 500), tau-leaping outperforms the SSA, as is indicated by the large increment of the effective time step. Except for system with $N = 50$ whose optimal open loop control policy is $u_1 = (0, 2, 1, 0, 2)$ with cost 11.26, the optimal policies for other larger N are all $u_2 = (0, 2, 1, 2, 2)$, which is also the optimal policy for the limiting ODE system (for $N = 50$, u_2 is the second best policy with cost 11.30). See Figure 5.7. The empirical means of the normalized process $z^{u,N}$ and the standard deviations are shown in Figure 5.5 for various values of N. As can be expected from the theoretical predictions, we observe that the mean values approach the solution of the limiting ODE, with the standard deviations decreasing as N increases. Convergence of the cost values to the cost value of the limit ODE system is also observed in Figure 5.7.

Hybrid control. We continue to study the hybrid control policy introduced in Subsection 4.3. Firstly, all 243 possible open loop control policies are ordered by their costs, among which we identify all "good" policies for $n_{ol} = 3$, $\epsilon_{ol} = 0.05$. Then, secondly, we estimate the first two moments of the process under all "good" policies based on 5000 independent realizations of the process. Thirdly, for each "good" policy, we generate M_{ol} trajectories once again and collect the accessed states at time t_j in $\mathcal{S}_j$, $1 \leq j < M$ according to the membership criterion (4.1) for $r = 3.0$. (Note that $\mathcal{S}_0$ contains only a single element). The minimum and maximum

N	50	100	200	500	1000	2000	4000
N_g	5	5	3	3	3	3	3
M_{ol}	5000	10000	10000	10000	20000	20000	30000
$\min_{1\leq j\leq 4} \|S_j\|$	4090	8738	12024	11545	23120	26060	40463
$\max_{1\leq j\leq 4} \|S_j\|$	11420	30572	25784	14587	29369	29597	44513
$9N^2$	22500	90000	360000	2250000	9000000	36000000	144000000

Table 5.4: Predator-prey model with hybrid control. The row "$9N^2$" shows the estimated state space cardinality after truncation if a simple cut-off criterion is used. The row "N_g" shows the number of "good" open control policies, and "M_{ol}" denotes the number of trajectories generated for each "good" open policy in the calculation of the sets $\mathcal{S}_j$. The other two rows contain the minimum and maximum numbers of states in the sets $\mathcal{S}_j$.

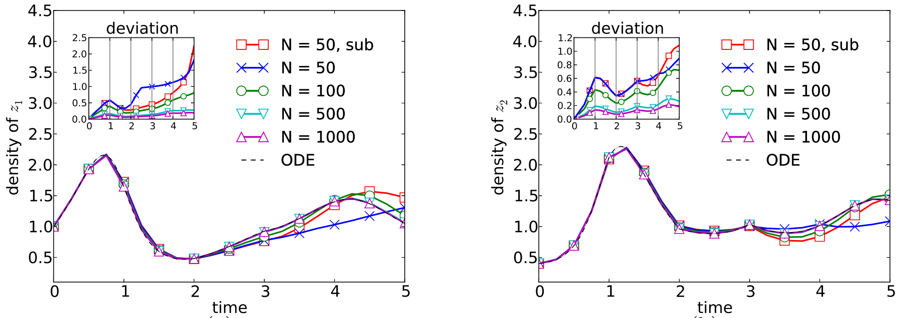
ϵ_{near}	N	50	100	200	500	1000	2000	4000
0.0	r_{ol}	13.6%	13.6%	38.1%	66.1%	66.6%	73.2%	74.7
	time	1.0h	5.3h	5.6h	7.1h	5.0h	5.0h	8.2h
	cost	10.72	9.88	9.58	9.27	9.18	9.13	9.11
0.02	r_{ol}	3.3%	1.1%	0.9%	0.6%	0.3%	0.4%	0.3%
	r_{near}	10.2%	12.0%	36.4%	65.5%	66.3%	72.9%	74.3%
	time	1.1h	5.5h	5.5h	7.0h	5.7h	5.5h	7.2h
	cost	10.60	9.81	9.47	9.25	9.18	9.13	9.11

Table 5.5: Predator-prey model with hybrid control. The rows "r_{ol}" and "r_{near}" record the relative frequencies of using an open loop policy or a feedback policy of a nearest neighbor when the hybrid control policy is applied (see Subsection 4.3). The row "time" shows the CPU run time (in hours) needed to compute the optimal hybrid control policy with 20 processors running in parallel for each N.

cardinalities $\min_{1\leq j\leq M-1} |\mathcal{S}_j|$ and $\max_{1\leq j\leq M-1} |\mathcal{S}_j|$ of the trust regions $\mathcal{S}_j$ are shown in Table 5.4.

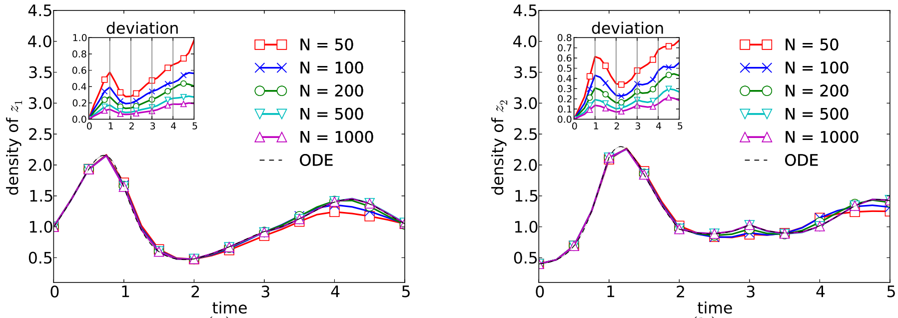
The reader should bear in mind that, if we wanted to compute the optimal feedback control policy on a globally truncated state space (see Subsection 4.2) then it would be necessary to include states whose normalized components are within $[0, 3.0] \times [0, 3.0]$ as suggested by the empirical mean and standard deviation of the process (see Figure 5.5), which would result in $9N^2$ states in total; even for moderate predator-prey populations, computing the optimal feedback policy on $\mathbb{X}_{cut}$ is therefore extremely costly. Compared to this approach, the adaptive state truncation that gives rise to the $\mathcal{S}_j$ is much more efficient in that the overall number of states involved in the computation of the optimal policy is much smaller; see Table 5.4 and Figure 5.8.

Finally, we compute the optimal hybrid policy using Algorithm 2 and apply it to the predator-prey model in the way explained in Subsection 4.3. The resulting cost values that were estimated based on 50000 independent realizations are shown in Table 5.5 and Figure 5.7 and clearly demonstrate the superiority of the hybrid controls over the optimal open loop control



(a) (b)

Fig. 5.5: Evolution of empirical mean and standard deviation (inset plot) of the normalized predator and prey states under the optimal open loop policy. The curve labeled by "$N = 50$, sub" corresponds to a system of size $N = 50$ that is controlled by the suboptimal policy u_2, which becomes the optimal policy for larger N. "ODE" corresponds to the limiting ODE under the optimal policy u_2.



(a) (b)

Fig. 5.6: Evolution of mean and standard deviation (inset plot) of the normalized predator-prey system under by the hybrid control policy.

policies (in particular, see Table 5.5 for $N = 50, 100, 200$.) To explain the observed gain in the numerical speed-up, Table 5.5 also records the relative frequencies r_{ol} of switching to an open loop policy: For $\epsilon_{near} = 0.0$, we observe that the hybrid control frequently switches to the optimal open loop policy, which is an indicator that the rust regions $\mathcal{S}_j$ are too small as the dynamics often hits an "unknown" state outside $\mathcal{S}_j$. Yet, for $\epsilon_{near} = 0.02$, we find that r_{ol} decreases significantly which suggests that the $\mathcal{S}_j$ contain almost all states that are close to the accessible states under the given control policy. Note, moreover, that the resulting cost value for $\epsilon_{near} = 0.02$ is slightly improved over the choice $\epsilon_{near} = 0.0$.

Before we conclude, we would like to stress the important observation that the standard deviation of the process is smaller under the hybrid control policy (similarly for the feedback policy) than that under the optimal open loop policy. This effect can be revealed by comparing

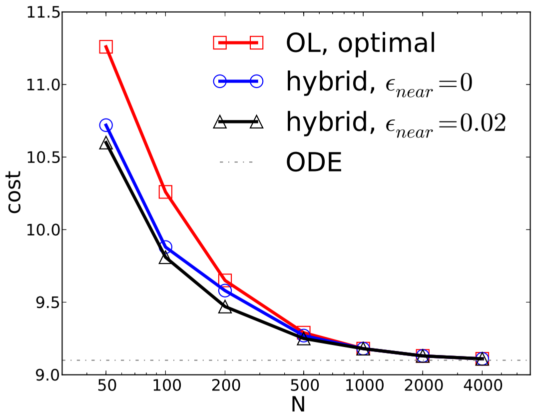


Fig. 5.7: Cost values of the predator-prey model under the optimal open loop control policy, hybrid control policies with $\epsilon_{near} = 0$ and 0.02, for various values of N. The dotted horizontal line is the optimal cost for the limiting ODE system.

Figure 5.5 with Figure 5.6 for the same value of N, and it suggests that besides providing smaller costs, both hybrid and feedback control policies have a positive effect on the speed of convergence towards the deterministic limit dynamics.

6. Conclusions and future directions. Due to their wide applicability, Markov Decision Processes have been the subject of intensive research. While the theory is quite developed, algorithms for numerically computing optimal controls are restricted to small or moderately sized systems.

The aim of this paper was to analyze optimal control problems for Markov jump process in the large number regime (parameterized by the "particle" number $N \gg 1$), i.e. when the state space is too large to compute optimal feedback controls using standard algorithms. Based on Kurtz's limit theorems, we have established convergence results for the value functions of the optimal control problems on finite and infinite time-horizons as $N \to \infty$. Our results suggest that the optimal open loop control policy for the limiting deterministic system is a good substitute for the controlled Markov jump process, for which the optimal feedback policy may not be computable. Nonetheless, for a given jump process with a possibly large, but finite N, the approximation error induced by replacing the optimal stochastic control by the limiting deterministic control is difficult to assess; even for large values of N the stochastic dynamics controlled by a deterministic open loop control policy is not robust under the intrinsic random perturbations, and may hence deviate considerably from the optimal regime. To account for this lack of robustness, we proposed an algorithmic strategy to compute a *hybrid* control policy that is based on a combination of deterministic (open loop) and stochastic (closed loop) controls. The key idea is to truncate the state space adaptively in time, exploiting data gathered from stochastic simulations under near-optimal open loop policies, and then to apply the optimal feedback control policy for all times, in which the stochastic realizations resides inside the truncated state space (for all other states, the suboptimal open loop policy is applied). Note that the proposed algorithmic scheme has some conceptional similarity to reinforcement learning

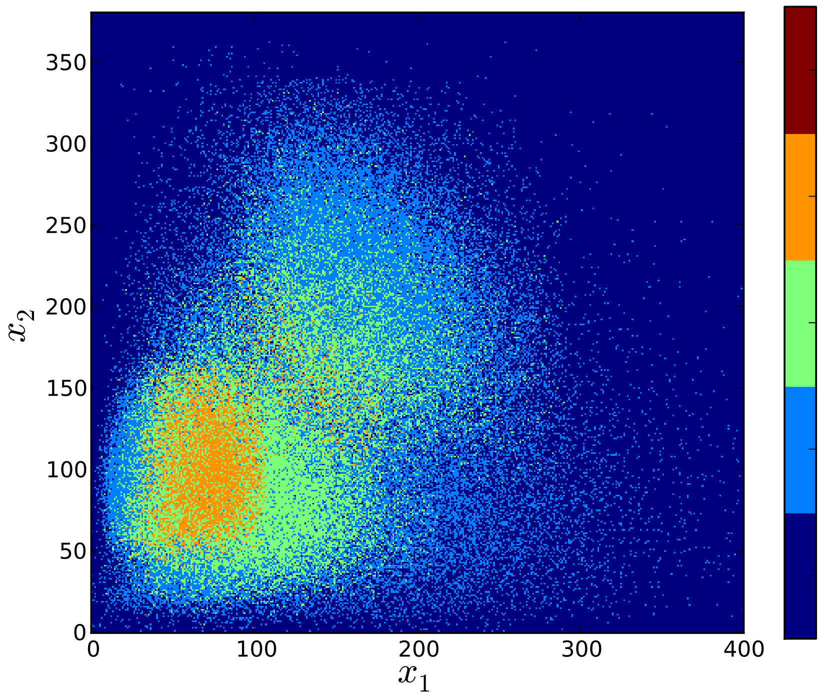


Fig. 5.8: States selected to construct the hybrid control policy in the predator-prey model for $N = 100$. The value at each grid point $x = (x_1, x_2)$ counts how many sets $\mathcal{S}_j$ contain the state x, i.e. the value at $x \in \mathbb{X}$ is equal to $\sum_{j=1}^{4} \mathbf{1}_{\mathcal{S}_j}(x)$.

procedures [29]. Both the accuracy and the practicability of the proposed *hybrid* algorithm have been demonstrated numerically with two simple birth-death and predator-prey models.

Throughout the article, we have assumed that the cost can be expressed as a function of the normalized process $z^N(t) = N^{-1}x^N(t)$, which in many cases is the natural variable scaling. In some cases, however, such as complex chemical reaction networks, it might be necessary to consider a more general scaling of the form $z_i^N(t) = N^{-\alpha_i}x_i(t)$, $\alpha_i \geq 0$, in which each chemical species comes with its own scaling order. Then, in the limit $N \to \infty$ it may happen that the limit of $z_i^N(t)$ can be deterministic, stochastic or even hybrid when some of the α_i are equal to zero and others are positive. We emphasize that the convergence analysis is much more involved in these cases, not to mention determining the correct scaling of the variables (see [2, 17]), and we refer to future work that will address these issues.

Appendix A. A technical lemma. The following inequality has been used in the proof of Theorem 3.1.

LEMMA A.1. *Let $\varphi(z) = |z|^\alpha$, for $z \in \mathbb{R}^n$ and $1 < \alpha \leq 2$. Then*

$$0 \leq \varphi(z + w) - \varphi(w) - z \cdot \nabla\varphi(w) \leq \frac{4}{\alpha - 1}\varphi\left(\frac{z}{2}\right), \quad \forall z, w \in \mathbb{R}^n. \tag{A.1}$$

Proof. The case $w = 0$ can be readily verified. Now assume $w \neq 0$ and consider $z = (z_1, 0, 0, \cdots, 0)^T$, $w = (w_1, w')^T$ where $z_1, w_1 \in \mathbb{R}$, $w' \in \mathbb{R}^{n-1}$. With $g(r) = r^\alpha$, $r > 0$, it follows that

$$\begin{aligned}
&\varphi(z+w) - \varphi(w) - z \cdot \nabla\varphi(w) \\
=&g\Big(\sqrt{(z_1+w_1)^2 + |w'|^2}\Big) - g\Big(\sqrt{w_1^2 + |w'|^2}\Big) - g'\Big(\sqrt{w_1^2 + |w'|^2}\Big)\frac{w_1 z_1}{\sqrt{w_1^2 + |w'|^2}} \\
=&\int_0^{z_1}\int_0^r \Bigg[g''\Big(\sqrt{(s+w_1)^2 + |w'|^2}\Big)\frac{(s+w_1)^2}{(s+w_1)^2 + |w'|^2} \\
&+ g'\left(\sqrt{(s+w_1)^2 + |w'|^2}\right)\left(\frac{1}{\sqrt{(s+w_1)^2 + |w'|^2}} - \frac{(s+w_1)^2}{\left((s+w_1)^2 + |w'|^2\right)^{\frac{3}{2}}}\right)\Bigg] dsdr \\
=&\int_0^{z_1}\int_0^r \left[g''\left(\sqrt{(s+w_1)^2 + |w'|^2}\right)\frac{(s+w_1)^2}{(s+w_1)^2 + |w'|^2} + \frac{|w'|^2 g'\left(\sqrt{(s+w_1)^2 + |w'|^2}\right)}{\left((s+w_1)^2 + |w'|^2\right)^{\frac{3}{2}}}\right] dsdr\,.
\end{aligned}$$

Since $1 < \alpha \leq 2$, we conclude that $g', g'' \geq 0$, hence $\frac{g'(r)}{r} = \frac{g''(r)}{\alpha-1} = \alpha r^{\alpha-2}$ is non-increasing for $r > 0$, and $\frac{a + \frac{b}{\alpha-1}}{a+b} \leq \frac{1}{\alpha-1}$, $\forall a, b > 0$. Therefore

$$\begin{aligned}
0 \leq&\varphi(z+w) - \varphi(w) - z \cdot \nabla\varphi(w) \\
=&\int_0^{z_1}\int_0^r g''\Big(\sqrt{(s+w_1)^2 + |w'|^2}\Big)\frac{(s+w_1)^2 + \frac{1}{\alpha-1}|w'|^2}{(s+w_1)^2 + |w'|^2}\, ds\, dr \\
\leq&\frac{1}{\alpha-1}\int_0^{z_1}\int_0^r g''\Big(\sqrt{(s+w_1)^2 + |w'|^2}\Big)\, ds\, dr \\
\leq&\frac{1}{\alpha-1}\int_0^{z_1}\int_0^r g''\big(|s+w_1|\big)\, ds\, dr \\
\leq&\frac{2}{\alpha-1}\int_0^{|z_1|}\int_0^{\frac{r}{2}} g''(s)\, ds\, dr \leq \frac{4}{\alpha-1} g\left(\frac{|z_1|}{2}\right) = \frac{4}{\alpha-1} g\left(\frac{|z|}{2}\right)\,.
\end{aligned}$$

For the general case, let $A \in \mathcal{O}(n)$ be a rotation, such that $Az = (z_1, 0, 0, \cdots, 0)^T$, $z_1 \in \mathbb{R}$. Then

$$\begin{aligned}
&\varphi(z+w) - \varphi(w) - z \cdot \nabla\varphi(w) \\
=&g\big(|z+w|\big) - g\big(|w|\big) - g'\big(|w|\big)\frac{w}{|w|} \cdot z \\
=&g\big(|Az+Aw|\big) - g\big(|Aw|\big) - g'\big(|Aw|\big)\frac{Aw}{|Aw|} \cdot Az \\
=&\varphi(Az+Aw) - \varphi(Aw) - Az \cdot \nabla\varphi(Aw) \\
\leq&\frac{4}{\alpha-1} g\left(\frac{|Az|}{2}\right) = \frac{4}{\alpha-1} g\left(\frac{|z|}{2}\right),
\end{aligned}$$

therefore the conclusion also holds for general $z \in \mathbb{R}^n$. ∎

REFERENCES

[1] Linda J.S. Allen, *An Introducttion to stochastic Processes with Application to Biology*, CRS Press, 2011.

[2] Karen. Ball, Thomas. G. Kurtz, Lea Popovic, and Greg Rempala, *Asymptotic Analysis of Multiscale Approximations to Reaction Networks*, Ann. Appl. Probab., 16 (2006), pp. 1925–1961.

[3] Richard E. Bellman, *Dynamic Programming*, Dover Publications, Incorporated, 2003.

[4] Jon Louis Bentley, *Multidimensional binary search trees used for associative searching*, Commun. ACM, 18 (1975), pp. 509–517.

[5] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold, *Efficient step size selection for the tau-leaping simulation method*, J. Chem. Phys., 124 (2006), p. 044109.

[6] Sulav Duwal, Stefanie Winkelmann, Christof Schütte, and Max von Kleist, *Optimal treatment strategies in the context of 'treatment for prevention' against HIV-1 in resource-poor settings.*, PLoS Comput. Biol., 11 (2015), p. e1004200.

[7] Eugene B. Dynkin, *Markov processes, vol. 1*, Grundlehren der mathematischen Wissenschaften, Academic press Berlin, New York, 1965.

[8] Crispin Gardiner, *Stochastic Methods*, Springer, 2010.

[9] Daniel T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, J. Comput. Phys., 22 (1976), pp. 403 – 434.

[10] ———, *Exact stochastic simulation of coupled chemical reactions*, J. Phys. Chem., 81 (1977), pp. 2340–2361.

[11] ———, *Markov Processes: An Introduction for Physical Scientists*, Academic Press, 1991.

[12] ———, *Approximate accelerated stochastic simulation of chemically reacting systems*, J. Chem. Phys., 115 (2001), pp. 1716–1733.

[13] ———, *Stochastic simulation of chemical kinetics*, Annu. Rev. Phys. Chem., 58 (2007), pp. 35–55.

[14] Ronald A. Howard, *Dynamic Programming and Markov Processes*, The MIT Press, 1th ed., 1960.

[15] Yucheng Hu and Tiejun Li, *Highly accurate tau-leaping methods with random corrections*, J. Chem. Phys., 130 (2009), p. 124109.

[16] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier, 3 ed., 2007.

[17] Hye-Won Kang and Thomas G. Kurtz, *Separation of time-scales and model reduction for stochastic reaction networks*, Ann. Appl. Probab., 23 (2013), pp. 529–583.

[18] Donald E. Kirk, *Optimal Control Theory: An Introduction*, Dover Books on Electrical Engineering Series, Dover Publications, 2004.

[19] Thomas G. Kurtz, *Solutions of ordinary differential equations as limits of pure jump markov processes*, J. Appl. Prob., 7 (1970), pp. 49–58.

[20] ———, *Limit theorems for sequences of jump Markov processes approximating ordinary differential processes*, J. Appl. Prob., 8 (1971), pp. 344–356.

[21] ———, *Limit theorems and diffusion approximations for density dependent markov chains*, in Stochastic Systems: Modeling, Identification and Optimization, I, vol. 5 of Mathematical Programming Studies, 1976, pp. 67–78.

[22] ———, *Strong approximation theorems for density dependent markov chains*, Stoch. Proc. Appl., 6 (1978), pp. 223 – 240.

[23] Suzanne Lenhart and John T. Workman, *Optimal Control Applied to biological models*, Chapman & Hall/CRC Mathematical and Computational Biology, Chapman & Hall/CRC, 2007.

[24] David M. Mount and Sunil Arya, *ANN Programming Manual*. `http://www.cs.umd.edu/~mount/ANN/`, 2010.

[25] Bernt Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer, 5th ed., 2000.

[26] Jürgen Pahle, *Biochemical simulations: stochastic, approximate stochastic and hybrid approaches.*, Brief Bioinform, 10 (2009), pp. 53–64.

[27] Martin L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc., New York, NY, USA, 1st ed., 1994.

[28] Muruhan Rathinam, Linda R. Petzold, Yang Cao, and Daniel T. Gillespie, *Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method*, J. Chem. Phys., 119 (2003), pp. 12784–12794.

[29] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[30] Jörgen W. Weibull, *Evolutionary Game Theory*, MIT Press, 1997.

[31] Darren J. Wilkinson, *Stochastic Modelling for Systems Biology*, Chapman & Hall/CRC, 2006.

[32] Stefanie Winkelmann, Christof Schütte, and Max von Kleist, *Markov control processes with rare state observation: Theory and application to treatment scheduling in HIV-1*, Commun. Math. Sci., 12 (2014), pp. 859–877.