Mean field for Markov Decision Processes: from Discrete to Continuous Optimization

Nicolas Gast[a,b], Bruno Gaujal[a,c], Jean-Yves Le Boudec[d]

[a] *Laboratoire Informatique de Grenoble, UMR 5217, 110 av. de la Chimie, 38041 Grenoble, France*
[b] *Grenoble Universités, 38041 Grenoble, France*
[c] *INRIA Grenoble - Rhône-Alpes, 655 avenue de l'Europe, 38 334 Saint Ismier Cedex, France*
[d] *EPFL IC-LCA2 - Lausanne, Switzerland*

Abstract

We study the convergence of Markov Decision Processes made of a large number of objects to optimization problems on ordinary differential equations (ODE). We show that the optimal reward of such a Markov Decision Process, satisfying a Bellman equation, converges to the solution of a continuous Hamilton-Jacobi-Bellman (HJB) equation based on the mean field approximation of the Markov Decision Process. We give bounds on the difference of the rewards, and a constructive algorithm for deriving an approximating solution to the Markov Decision Process from a solution of the HJB equations. We illustrate the method on three examples pertaining respectively to investment strategies, population dynamics control and scheduling in queues are developed. They are used to illustrate and justify the construction of the controlled ODE and to show the gain obtained by solving a continuous HJB equation rather than a large discrete Bellman equation.

1. Introduction

The purpose of this paper is to study optimization problems on Markovian systems composed of a large number of interacting objects.

Consider a system of N objects evolving in a common environment. At each time step, objects change their state randomly according to some probability kernel Γ^N. This kernel depends on the number of objects in each state and also on the decisions of a centralized controller. The goal of this paper is to study the behavior of the controlled system when N becomes large.

Several papers investigate the asymptotic behavior of such systems, but without controllers. For example, in [3, 15], the authors showed that under mild conditions the system converges to a deterministic limit when N grows. The limiting system can be of two types, depending on the intensity $I(N)$ (the intensity is the probability than an object changes its state between two time steps). If $I(N) = O_{N\to\infty}(1)$, the system converges to a dynamical system in discrete time [15]. If $I(N)$ goes to 0 as N grows, the limiting system is a continuous time dynamical system and can be described by ordinary differential equations (ODE).

In [8], the authors consider the controlled case when the intensity is $O(1)$. In that case, the optimization problem of the system of size N converges to a deterministic optimization problem in discrete time. Solving the deterministic system allows one to compute policies that are asymptotically optimal as the number of objects grows.

In this paper, we focus on the $o(1)$ case, which is substantially different from the discrete time case. We consider a Markov decision process where at each time step, a central controller chooses an action from a predefined set that will modify the dynamics of the system. For this, its gets a reward depending on the current state of the system and on the action. The goal of the controller is to maximize the expected reward over a finite time horizon. We show that when N goes to infinity, this problem converges to an optimization problem on an ordinary differential equation.

More precisely, we show that when the Markov decision process is such that its empirical density measure is Markovian, then its optimal reward converges to the optimal reward of its mean field approximation, given by the solution of an HJB equation. Furthermore, the optimal policy of

the limit continuous system is also asymptotically optimal for the original discrete system. Our method relies on bounding techniques used in stochastic approximation and learning [4, 1]. We also introduce an original coupling method, where, to each sample path of the Markov decision process, we associate a random trajectory, obtained as a solution of the ODE, i.e. the mean field limit, controlled by random actions.

This convergence result has an algorithmic counterpart. Basically, when confronted with a large Markov Decision problem, one can first solve the HJB equation for the associated mean field limit and then build a decision policy for the initial system that is asymptotically optimal.

Few papers in the literature are concerned with the problem of mixing the limiting behavior of a large number of objects with optimization. In [6], the value function of the Markov decision process is approximated by a linearly parametrized class of functions, and a fluid approximation of the MDP is used. It is shown that a solution of the HJB equation is a value function for a modification of the original MDP problem. In [20, 7], the curse of dimensionality of dynamic programming is circumvented by approximating the value function by linear regression. In [16], a continuous optimization problem is seen as a fluid limit of a discrete Markov chain. The chain is ad-hoc and is constructed in the purpose of solving the optimizing the fluid limit of a queuing system. Our approach, which uses a mean field approximation, is more structural. In particular it allows one to consider the case where the intensity of actions is only bounded in expectation. Also, our method is different as we explicitly account for the rate of convergence of the original model to its mean field limit, and we obtain explicit bounds.

These results have two main implications. The first one is to justify the construction of controlled ODEs as good approximations of large discrete controlled systems. This construction is often done without rigorous proofs. A discussion based on the vaccination example is given in Section 4.2.

The second implication concerns the effective computation of an optimal control policy. In the discrete case, this is usually done by using dynamic programming for the finite horizon case or by computing a fixed point of the Bellman equation in the discounted case. Both approaches suffer from the curse of dimensionality that makes them impractical when the state space is too large. In our context, the size of the state space is exponential in N, making the problem even more acute. In practice, modern supercomputers only allow one to tackle optimal control problems where N is no larger than a few tens.

The mean field approach offers an alternative to brute force computations. By letting N go to infinity, the discrete problem is replaced by a limit Hamilton-Jacobi-Bellman equation that is deterministic and where the dimensionality of the original system has been hidden in the density measure. Solving the HJB equation numerically is sometimes rather easy, as in the examples in Sections 4.1 and 4.2. It provides a deterministic optimal policy whose reward with a finite (but large) number of objects is remarkably close to the optimal reward.

In this paper we focus on the finite horizon case, though the technique applies mutatis mutandis to infinite horizon with discount.

The rest of the paper is structured as follows. Section 2 contains definitions, some notation, and hypotheses. Section 3 gives the theoretical results and the resulting algorithms. Section 4 illustrates the application of our method on a few examples. Section 5 contains proofs.

2. Notations and Definitions

2.1. System with N Objects

We consider a system composed by N *objects*. Each object has a state in the finite set $\mathcal{S} = \{1 \dots S\}$. Time is discrete and the state of the object n at step $k \in \mathbb{N}$ is denoted $X_n^N(k)$. The state of the system at time k is $X^N(k) \stackrel{\text{def}}{=} \left(X_1^N(k) \dots X_N^N(k)\right)$. For all $i \in \mathcal{S}$, we denote by $M^N(k)$ the empirical measure of the objects $\left(X_1^N(k) \dots X_N^N(k)\right)$ at time k:

$$M^N(k) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \delta_{X_n^N(k)},$$

where δ_x denotes the Dirac measure in x. $M^N(k)$ is a probability measure on $\mathcal{S}$ and $M_i^N(k)$ denotes the proportions of objects in state i at time k (also called the density): $M^N(k)[i] = \sum_{n=1}^{N} \mathbf{1}_{X_n^N(k)=i}$.

The system $\left(X^N(k)\right)_{k\in\mathbb{N}}$ is a Markov process once the sequence of the actions taken by the controller is fixed. This means that there exists a kernel $\Gamma^N(i_1 \dots i_N, j_1 \dots j_N, a)$ such that if the controller takes the action $A^N(k)$ at time t and the system is in state $X^N(k)$, then:

$$\mathcal{P}\left(X^N(k+1) = j_1 \dots j_N | X^N(k) = i_1 \dots i_N, A^N(k) = a\right) = \Gamma^N\left(i_1 \dots i_N, j_1 \dots j_N, a\right) \quad (1)$$

The main assumption on the kernel Γ^N is that it is invariant by any permutation of the objects. This implies in particular that the objects are only distinguishable through their state. Moreover, this means that the process $M^N(k)$ is also Markovian once the sequence of actions is given. In the following, we will focus on the process of density of the system, $\left(M^N(k)\right)_{k\in\mathbb{N}}$, whose kernel is denoted by Γ^N.

2.2. Action, Reward and Policy

At every time k, a centralized controller chooses an action $A^N(k) \in \mathcal{A}$ where $\mathcal{A}$ is called the action set. $(\mathcal{A}, d)$ is a compact metric space for some distance d. The purpose of Markov Decision is to compute optimal *policies*. A policy $\pi = (\pi_0, \pi_1, \dots, \pi_k, \dots)$ is a sequence of decision rules that specifies which actions should be used at time k. In general Markov decision processes, π_k may depend on all the history of the Process $M^N(0) \dots M^N(k)$. However, it can be shown that when the state space is finite, deterministic Markovian policies are dominant [17], therefore, we will only focus on them. For each k, π_k is a function $\mathcal{P}(\mathcal{S}) \to \mathcal{A}$. $M_\pi^N(k)$ denotes the density of the system at time k when the controller applies policy π.

If the system has density $M^N(k)$ at time k and if the controller chooses the action $A^N(k)$, she gets an *instantaneous reward* $r^N(M^N(k), A^N(k))$. The expected average reward over a finite-time horizon $[0; H^N]$ starting from m_0 when applying the policy π is defined by

$$V_\pi^N(m) \stackrel{\text{def}}{=} \mathbb{E}\left(\sum_{k=0}^{\lfloor H^N \rfloor} r^N\left(M_\pi^N(k), \pi(M_\pi^N(k))\right) \middle| M_\pi^N(0) = m \right) \quad (2)$$

The goal of the controller is to find a optimal policy that maximizes the expected reward. We denote by $V_*^N(m)$ the optimal reward when starting from m:

$$V_*^N(m) = \sup_\pi V_\pi^N(m).$$

2.3. Scaling Assumptions

If at some time k, the system has density $M^N(k) = m$ and the controller chooses action $A^N(k) = a$, the system goes into state $M^N(k+1)$ with probabilities given by the kernel $\Gamma^N(M^N(k), A^N(k))$. The expectation of the difference between $M^N(k+1)$ and $M^N(k)$ is called the *drift* and is denoted by $F^N(m, a)$:

$$F^N\left(m, a\right) \stackrel{\text{def}}{=} \mathbb{E}\left[M^N(k+1) - M^N(k) | M^N(k) = m, A^N(k) = a\right].$$

In order to study the limit with N, we assume that F^N goes to 0 at speed $I(N)$ when N goes to infinity and that $F^N/I(N)$ converges to a Lipschitz continuous function f. More precisely, we assume that there exists a sequence $I(N) \in (0; 1)$, $N = 1, 2, 3...$, called the *intensity* of the model with $\lim_{N\to\infty} I(N) = 0$ and a sequence $I_0(N)$, $N = 1, 2, 3...$, also with $\lim_{N\to\infty} I_0(N) = 0$ such that for all $m \in \mathcal{P}(\mathcal{S})$ and $a \in \mathcal{A}$: $\left|\frac{1}{I(N)} F^N(m, a) - f(m, a)\right| \leq I_0(N)$. In a sense, $I(N)$ represents the order of magnitude of the number of objects that change their state within one unit of time.

The changes of $M^N(k)$ during a time step is of order $I(N)$. This suggests a rescaling of time by $I(N)$ to obtain an asymptotic result. We define the continuous time process $\left(\hat{M}^N(t)\right)_{t\in\mathbb{R}^+}$ as the

affine interpolation of $M^N(k)$, rescaled by the intensity function, i.e. $\hat{M}^N$ is affine on the intervals $[kI(N), (k+1)I(N)]$, $k \in \mathbb{N}$ and

$$\hat{M}^N(kI(N)) = M^N(k).$$

Similarly, $\hat{M}^N_\pi$ denotes the affine interpolation of the density under policy π. Thus, $I(N)$ can also be interpreted as the duration of the time slot for the system with N objects.

We assume that the time horizon and the reward per time slot scale accordingly, i.e. we impose

$$\begin{aligned} H^N &= \left\lfloor \frac{T}{I(N)} \right\rfloor \\ r^N(m,a) &= I(N) r(m,a) \end{aligned}$$

for every $m \in \mathcal{P}(\mathcal{S})$ and $a \in \mathcal{A}$ (where $\lfloor x \rfloor$ denotes the largest integer $\leq x$).

2.4. Limiting System (Mean Field Limit)

We will see in Section 3 that as N grows, the stochastic system $\hat{M}^N_\pi$ converges to a deterministic limit m_π, the mean field limit. For more clarity, all the stochastic variables (*i.e.*, when N is finite) are in uppercase while their limiting deterministic values are in lowercase.

An action function $\alpha : [0;T] \to \mathcal{A}$ is a piecewise Lipschitz continuous function that associates to each time t an action $\alpha(t)$. Note that action functions and policies are different in the sense that actions functions do not take into account the state to define the next action. For an action function α and an initial condition m_0, we consider the following ordinary differential equation for $m(t)$, $t \in \mathbb{R}^+$:

$$m(t) - m(0) = \int_0^t f(m(s), \alpha(s)) ds. \tag{3}$$

Under the foregoing assumptions on f and α, this ODE satisfies the Cauchy Lipschitz condition and therefore has a unique solution once the initial condition $m(0) = m_0$ is fixed. We call ϕ_t, $t \in \mathbb{R}^+$, the corresponding semi-flow, i.e.

$$m(t) = \phi_t(m_0, \alpha)$$

is the unique solution of Eq.(3).

As for the system with N objects, we define $v_\alpha(m_0)$ as the reward of the limiting system over a finite horizon $[0;T]$ when applying the action function α and starting from $m(0) = m_0$:

$$v_\alpha(m_0) \stackrel{\text{def}}{=} \int_0^T r\left(\phi_s(m_0, \alpha), \alpha(s)\right) ds. \tag{4}$$

This equation looks similar to the stochastic case (2) although there are two main differences. The first one is that the system is deterministic. The second is that it is defined for action functions and not for policies. We also define the optimal reward of the deterministic limit $v_*(m_0)$:

$$v_*(m_0) = \sup_\alpha v_\alpha(m_0),$$

where the supremum is taken over all possible actions functions from $[0;T] \to \mathcal{A}$.

2.5. Summary of Assumptions

In Section 3 we establish theorems for the convergence of the discrete stochastic optimization problem to a continuous deterministic one. These theorems are based on several technical assumptions, which are given next. Since $\mathcal{S}$ is finite, the set $\mathcal{P}(\mathcal{S})$ is the simplex in $\mathbb{R}^{\mathcal{S}}$ and for $m, m' \in \mathcal{P}(\mathcal{S})$ we define $\|m\|$ as the ℓ^2-norm of m and $\langle m, m' \rangle = \sum_{i=1}^S m_i m'_i$ as the usual inner product.

(A1) (Transition Kernel). Objects can be observed only through their state, *i.e.*, the transition kernel Γ^N, defined by Eq.(1), is invariant by permutations of $1 \dots N$.

There exist some non random functions $I_1(N)$ and $I_2(N)$ such that $\lim_{N\to\infty} I_1(N) = \lim_{N\to\infty} I_2(N) = 0$ and such that for all m and any policy π, the number of objects that perform a transition between time slot k and $k+1$ per time slot $\Delta_\pi^N(k)$ satisfies

$$\begin{aligned} \mathbb{E}\left(\Delta_\pi^N(k)\middle| M_\pi^N(k) = m\right) &\leq N I_1(N) \\ \mathbb{E}\left(\Delta_\pi^N(k)^2\middle| M_\pi^N(k) = m\right) &\leq N^2 I(N) I_2(N) \end{aligned}$$

where $I(N)$ is the intensity function of the model, defined in the following assumption A2.

(A2) (Convergence of the Drift). There exist some non random functions $I(N)$ and $I_0(N)$ and a function $f(m,a)$ such that $\lim_{N\to\infty} I(N) = \lim_{N\to\infty} I_0(N) = 0$ and

$$\left\| \frac{1}{I(N)} F^N(m,a) - f(m,a) \right\| \leq I_0(N)$$

f is defined on $\mathcal{P}(\mathcal{S}) \times \mathcal{A}$ and there exists L_2 such that $|f(m,a)| \leq L_2$.

(A3) (Lipschitz Continuity). There exist constants L_1, K and K_r such that for all $m, m' \in \mathcal{P}(\mathcal{S})$, $a, a' \in \mathcal{A}$:

$$\begin{aligned} \left\|F^N(m,a) - F^N(m',a)\right\| &\leq L_1 \left\|m - m'\right\| I(N) \\ \left\|f(m,a) - f(m',a')\right\| &\leq K(\left\|m - m'\right\| + d(a,a')) \\ \left|r(m,a) - r(m',a)\right| &\leq K_r \left\|m - m'\right\| \end{aligned}$$

We also assume that the reward is bounded: $\sup_{m,a\in\mathcal{A}} |r(m,a)| \stackrel{\text{def}}{=} \|r\|_\infty < \infty$.

To make things more concrete, here is a simple but useful case where all assumptions are true.

- There are constants c_1 and c_2 such that the expectation of the number of objects that perform a transition in one time slot is $\leq c_1$ and its standard deviation is $\leq c_2$,
- and $F^N(m,a)$ can be written under the form $\frac{1}{N}\varphi(m,a,1/N)$ where φ is a continuous function on $\Delta_S \times \mathcal{A} \times [0,\epsilon)$ for some neighborhood Δ_S of $\mathcal{P}(\mathcal{S})$ and some $\epsilon > 0$, continuously differentiable with respect to m.

In this case one can choose $I(N) = 1/N$, $I_0(N) = c_0/N$ (where c_0 is an upper bound to the norm of the differential $\frac{\partial\varphi}{\partial m}$), $I_1(N) = c_1/N$ and $I_2(N) = (c_1^2 + c_2^2)/N$.

3. Mean Field Convergence

In this section we establish the main result, Theorem 4, which states the convergence of the optimization problem for the system with N objects to the optimization problem of the mean field limit. To this end we introduce two auxiliary systems. The former is the process $\phi_t(m_0, A_\pi^N)$ defined below, which is a random continuous time system, coupled to the original system with N objects. The latter is M_α^N, also defined below, which is a discrete time system with N objects under a deterministic action function.

3.1. First Auxiliary System

Consider the system with N objects under policy π. The process M_π^N is defined on some probability space Ω. To each $\omega \in \Omega$ corresponds a trajectory $M_\pi^N(\omega)$. For each $\omega \in \Omega$, we define an action function $A_\pi^N(\omega)$. This random function is piecewise constant on each interval $[kI(N), (k+1)I(N))$ $(k \in \mathbb{N})$ and is such that $A_\pi^N(\omega)(kI(N)) \stackrel{\text{def}}{=} \pi_k(M^N(k))$ is the action taken by the controller of the system with N objects at time slot k, under policy π.

Recall that for any $m_0 \in \mathcal{P}(\mathcal{S})$ and action function α, $\phi_t(m_0, \alpha)$ is the solution of the ODE (3). For every ω, $\phi_t(m_0, A_\pi^N(\omega))$ is the solution of the limiting system with action function $A_\pi^N(\omega)$, i.e.

$$\phi_t(m_0, A_\pi^N(\omega)) - m_0 = \int_0^t f(\phi_s(m_0, A_\pi^N(\omega)), A_\pi^N(\omega)(s))ds.$$

When ω is fixed, $\phi_t(m_0, A_\pi^N(\omega))$ is a continuous time deterministic process corresponding to one trajectory $M_\pi^N(\omega)$. When considering all possible realizations of M_π^N, $\phi_t(m_0, A_\pi^N)$ is a random, continuous time function "coupled" to M_π^N. Its randomness comes only from the action term A_π^N, in the ODE. In the following, we omit to write the dependence in ω. A_π^N and M_π^N will always designate the processes corresponding to the same ω.

The following result is the main technical result; it shows the convergence of the controlled system in probability, with explicit bounds. Notice that it does not require any regularity assumption on the policy π (recall that $\hat{M}_\pi^N$ is the linear interpolation of the discrete time system with N objects).

Theorem 1. *Under Assumption (A1,A2,A3), for any $\epsilon > 0$, $N \geq 1$ and any policy π:*

$$\mathbb{P}\left\{\sup_{0\leq t\leq T}\left\|\hat{M}_\pi^N(t) - \phi_t(m_0, A_\pi^N)\right\| > \left[\left\|M^N(0) - m_0\right\| + I_0(N)T + \epsilon\right]e^{L_1 T}\right\} \leq \frac{J(N,T)}{\epsilon^2} \tag{5}$$

with

$$J(N,T) = 8T\left\{L_1^2\left[I_2(N)I(N)^2 + I_1(N)^2\left(T + I(N)\right)\right] + S^2\left[2I_2(N) + I(N)\left(I_0(N) + L_2\right)^2\right]\right\}$$

Note that $I_0(N)$ and $J(N,T)$ for a fixed T go to 0 as $N \to \infty$. The proof is given in Appendix 5.1.

Let π is a policy and A_π^N the sequence of actions corresponding to a trajectory M_π^N as we just defined. Eq.(4) defines the reward for the deterministic limit when applying a sequence of action. This defines a random variable $v_{A_\pi^N}(m_0)$ which corresponds to the reward of System ∞ when applying A_π^N. The random part comes from A_π^N. $\mathbb{E}\left[v_{A_\pi^N}(m_0)\right]$ designates the expectation of this reward over all possible A_π^N. A first consequence of Theorem 1 is the convergence of $V_\pi^N\left(M^N(0)\right)$ to $\mathbb{E}\left[v_{A_\pi^N}(m_0)\right]$ with an error that can be uniformly bounded.

Theorem 2 (Uniform convergence of the reward)**.** *Let A_π^N be the random action function associated with M_π^N, as defined earlier. Under Assumptions (A1,A2,A3),*

$$\left|V_\pi^N\left(M^N(0)\right) - \mathbb{E}\left[v_{A_\pi^N}(m_0)\right]\right| \leq B\left(N, \left\|M^N(0) - m_0\right\|\right)$$

with

$$\begin{aligned} B(N,\delta) \quad &\stackrel{\text{def}}{=} \quad I(N)\left\|r\right\|_\infty + K_r\left(\delta + I_0(N)T\right)\frac{e^{L_1 T} - 1}{L_1} \\ &+ \quad \frac{3}{2^{\frac{1}{3}}}\left[\frac{K_r}{L_1}\left(e^{L_1 T} - 1 + \frac{I(N)}{2}\right)\right]^{\frac{2}{3}}\left\|r\right\|_\infty^{\frac{1}{3}} J(N,T)^{\frac{1}{3}} \end{aligned} \tag{6}$$

Note that $\lim_{N\to\infty,\delta\to 0} B(N,\delta) = 0$; in particular, if $\lim_{N\to\infty} M_\pi^N(0) = m_0$ almost surely [resp. in probability] then $\left|V_\pi^N\left(M^N(0)\right) - \mathbb{E}\left[v_{A_\pi^N}(m_0)\right]\right| \to 0$ almost surely [resp. in probability].

The proof is given in appendix 5.2.

3.2. Second Auxiliary System

We now introduce the second auxiliary system. Let α be an action function that specifies the action to be taken at time t. Although α has been defined for the limiting system, it can also be used in the system with N objects. In that case, the action function α can be seen as a policy that does not depend on the state of the system. At step k, the controller applies action $\alpha(kI(N))$. By abuse of notation, we denote by M_α^N, the state of the system when applying the action function α

(it will be clear from the notation whether the subscript is an action function or a policy). Similarly we define

$$V_\alpha^N(m_0) \stackrel{\text{def}}{=} \mathbb{E}\left(\sum_{k=0}^{H^N} r\left(M_\alpha^N(k), \alpha(kI(N))\right) \middle| M_\alpha^N(0) = m_0\right)$$

A second consequence of Theorem 1 is the convergence of M_α^N and of the reward:

Theorem 3. *Assume (A1,A2,A3); α is a piecewise Lipschitz continuous action function on $[0;T]$, of constant K_α, and with at most p discontinuity points. Let $\hat{M}_\alpha^N(t)$ be the linear interpolation of the discrete time process M_α^N. Then*

$$\mathbb{P}\left\{\sup_{0\le t\le T}\left\|\hat{M}_\alpha^N(t) - \phi_t(m_0,\alpha)\right\| > \left[\left\|M^N(0) - m_0\right\| + I_0'(N,\alpha)T + \epsilon\right] e^{L_1 T}\right\} \le \frac{J(N,T)}{\epsilon^2} \tag{7}$$

with

$$I_0'(N,\alpha) = I_0(N) + I(N)Ke^{(K-L_1)T}\left(\frac{K_\alpha}{2} + 2\left(1+\min(1/I(N),p)\right)\|\alpha\|_\infty\right)$$

Further,

$$\left|V_\alpha^N\left(M^N(0)\right) - v_\alpha(m_0)\right| \le B'\left(N, \left\|M^N(0) - m_0\right\|\right) \tag{8}$$

with $B'(N,\delta)$ as in Eq.(6) but with $I_0(N)$ replaced by $I_0'(N,\alpha)$.

Note that $\lim_{N\to\infty,\delta\to 0} B'(N,\delta) = 0$; in particular, if $\lim_{N\to\infty} M_\pi^N(0) = m_0$ almost surely [resp. in probability] then $\lim_{N\to\infty} V_\alpha^N\left(M^N(0)\right) = v_\alpha(m_0)$ almost surely [resp. in probability].

The proof is given in appendix 5.3.

3.3. Convergence of Optimization Problems

Theorem 4 (Optimal System Convergence)**.** *Assume (A1,A2,A3). If $\lim_{N\to\infty} M^N(0) = m_0$ almost surely [resp. in probability] then:*

$$\lim_{N\to\infty} V_*^N\left(M^N(0)\right) = v_*\left(m_0\right)$$

almost surely [resp. in probability].

Proof of Theorem 4. This theorem is a direct consequence of Theorem 3 and Theorem 2. We do the proof for almost sure convergence, the proof for convergence in probability is similar. To prove the theorem we prove

$$\limsup_{N\to\infty} V_*^N(M^N(0)) \le v_*(m_0) \le \liminf_{N\to\infty} V_*^N(M^N(0)) \tag{9}$$

- Let $\epsilon > 0$ and $\alpha(.)$ be an action function such that $v_\alpha(m_0) \ge v_*(m_0) - \epsilon$ (such an action is $\epsilon-$optimal). Theorem 3 shows that $\lim_{N\to\infty} V_\alpha^N(M^N(0)) = v_\alpha(m_0) \ge v_*(m_0) - \epsilon$ a.s. This shows that $\liminf_{N\to\infty} V_*^N(M^N(0)) \ge \lim_{N\to\infty} V_\alpha^N(M^N(0)) \ge v_*(m_0) - \epsilon$; this holds for every $\epsilon > 0$ thus $\liminf_{N\to\infty} V_*^N(M^N(0)) \ge v_*(m_0)$ a.s., which establishes the second inequality in Eq.(9), on a set of probability 1.
- Let $B(N,\delta)$ be as in Theorem 2, $\epsilon > 0$ and π^N such that $V_*^N(M^N(0)) \le V_{\pi^N}^N(M^N(0)) + \epsilon$. By Theorem 2, $V_{\pi^N}^N(M^N(0)) \le \mathbb{E}\left(v_{A_{\pi^N}^N}(m_0)\right) + B(N,\delta^N) \le v_*(m_0) + B(N,\delta^N)$ where $\delta^N \stackrel{\text{def}}{=} \left\|M^N(0) - m_0\right\|$. Thus $V_*^N(M^N(0)) \le v_*(m_0) + B(N,\delta^N) + \epsilon$. If further $\delta^N \to 0$ a.s. it follows that $\limsup_{N\to\infty} V_*^N(M^N(0)) \le v_*(m_0) + \epsilon$ a.s. for every $\epsilon > 0$, thus $\limsup_{N\to\infty} V_*^N(M^N(0)) \le v_*(m_0)$ a.s.

□

In particular, this theorem, along with Theorem 3 shows that an optimal policy for the limiting system is asymptotically optimal for the system with N objects as N goes to infinity. Since the reward function $r(m,a)$ is bounded and the time-horizon $[0;T]$ is finite, the set of reward when starting from the point m, $\{v_\alpha(m) : \alpha \text{ action function}\}$, is bounded. This set is not necessarily compact since the set of action function is not closed (a limit of Lipschitz continuous functions is not necessarily Lipschitz continuous). However, as it is bounded, for all $\epsilon > 0$, there exists an action function α^ϵ such that $v_*(m) = \sup_\alpha v_\alpha(m) \leq v_{\alpha^\epsilon} + \epsilon$. Theorem 4 shows that α^ϵ is optimal up to a term 2ϵ for N big enough.

In particular, this shows that as N grows, policies that do not take into account the state of the system (*i.e.*, action functions) are asymptotically as good as adaptive policies. In practice however, adaptive policies might perform better, especially from small values of N. However, it is in general impossible to prove convergence for the adaptive policy.

In fact, in many cases, the optimal policies π used for the control of stochastic systems are not continuous and exhibits thresholds. In those cases M_π^N does not necessarily converge and obtaining asymptotics can be difficult. In some particular case, like for the best response dynamics studied in [9], limit theorems can be obtain but at the cost of a greater complexity. This is beyond the scope of the present paper.

3.4. Hamilton-Jacobi-Bellman equation and dynamic programming

Let us consider the finite time optimization problem for the stochastic system and its limit on a constructive point of view. Since the state space is finite, one can compute the optimal reward by using a dynamic programming algorithm. If $U^N(m,t)$ denotes the optimal reward for the stochastic system starting from m at time $t/I(N)$, then $U^N(m,t) = \sup_\pi \mathbb{E}\left[\sum_{k=t/I(N)}^{T/I(N)} r^N(M_\pi^N(k)) : M^N(t) = m\right]$. $U^N(x,t)$. The optimal reward can be computed by a discrete dynamic programming algorithm [17] by setting $U^N(m,T) = r^N(m)$ and

$$U^N(m,t) = \sup_{a\in\mathcal{A}} \mathbb{E}\left(r^N(m,a) + U^N(M^N(t+I(N)), t+I(N))\middle|\, \bar{M}^N(t) = m, A^N(t) = a\right). \quad (10)$$

Then, the optimal cost over horizon $[0;T/I(N)]$ is $V_*^N(m) = U(m,0)$.

Similarly, if we denote by $u(m,t)$ the optimal cost over horizon $[t;T]$ for the limiting system, $u(m,t)$ satisfies the classical Hamilton-Jacobi-Bellman equation:

$$\dot{u}(m,t) + \max_a \{\nabla u(m,t).f(m,a) + r(m,a)\} = 0. \quad (11)$$

This provides a way to compute the optimal reward as well as the optimal policy by solving the partial differential equation above.

3.5. Algorithmic Construction

Theorem 4 above can be used to design an effective construction of an asymptotically optimal policy for the system with N objects over the horizon $[0,H]$ by using the procedure described in Algorithm 1.

Theorem 4 says that under policy π, the total reward V_π^N is asymptotically optimal:

$$\lim_{N\to\infty} V_\pi^N(M^N(0)) = \liminf_{N\to\infty} V_*^N(M^N(0)).$$

The policy π constructed by Algorithm 1 is static in the sense that it does not depend on the state $M^N(k)$ but only on the initial state $M^N(0)$, and the deterministic estimation of $M^N(k)$ provided by the differential equation. One can construct a more adaptive policy by updating the starting point of the differential equation at each step. This new procedure, constructing an adaptive policy π' from 0 to the final horizon H is given in Algorithm 2.

In practice, the total reward of the adaptive policy π' is larger than the reward of the static policy π because it uses on-line corrections at each step, before taking a new action. However Theorem 4 does not provide a proof of its asymptotic optimality.

Algorithm 1: Static algorithm constructing a policy for the system with N objects, over the finite horizon.

begin

From the original system with N objects, construct the density measure M^N and its kernel Γ^N and let $M^N(0)$ be the initial density;

Compute the limit of the drift of Γ^N, namely the function f;

Solve the HJB equation (11) on the interval $[0, HI(N)]$. This provides an optimal control function $\alpha(M_0^N, t)$;

Construct a discrete control $\pi(M^N(k), k)$ for the discrete system, that gives the action to be taken under state $M^N(k)$ at step k:

$$\pi(M^N(k), k) \stackrel{\text{def}}{=} \alpha(\phi_{kI(N)}(M^N(0), \alpha)).$$

return π;

Algorithm 2: Adaptive algorithm constructing a policy π' for the system with N objects, over the finite horizon H.

begin

$M := M^N(0)$; $k := 0$;

repeat

$\alpha_k(M, \cdot) :=$ solution of (11) over $[kI(N), HI(N)]$ starting in M;

$\pi'(M, k) := \alpha_k(\phi_{kI(N)}(M, \alpha_k))$;

M is changed by applying kernel $\Gamma^N_{\pi'}$;

k:= k+1;

until k=H;

return π';

4. Application examples

The goal of our approach is twofold. First, the goal is to provide a justification for the study of deterministic optimization problem as a good approximation of a stochastic problem. The second goal is to provide new methods for the resolution of some problems. The first one is illustrated by the first two examples while the last example provides a problem that should be developed in future work.

4.1. Harvesting

We begin by a first simple example that can be seen as a simplified discrete Merton's problem. This first example shows a case where the optimization problem in the infinite system can be solved in closed form. This can be seen as an ideal case for the mean field approach: while the original system is difficult to solve even numerically when N is large, taking the limit when N goes to infinity make it simple to solve, even analytically.

We consider a system made of N objects that can be either in state G (Good) or B (Bad). The controller can take two actions. The first one leads objects to go from B to G. The second one leads objects to go from G to B. The controller earn $1/N$ dollars per object that goes from G to B. This problem is often referred as a harvesting problem since the controller has the choice between harvest (action 1) and let it produce (action 0).

Within our framework, we represent this problem with a system of N interacting objects. If the action is 1, at each time step each object in state B goes to state G with probability $1/N$ while if the action taken is 0, each object in state G goes to state B. The intensity of the model is $I(N) = 1/N$. If $x(t)$ is the fraction of objects in state G and $\alpha(t) \in \{0; 1\}$ the action taken by the controller, the mean field limit of the system is:

$$\frac{\partial x}{\partial t} = 1 - x(t) - \alpha(t), \tag{12}$$

and the quantity to maximize is $\int_0^T x(t)\alpha(t)dt$.

Let us call $u(x, t)$ the reward when the remaining time is t and there is a proportion x of users in state G. The classical Hamilton-Jaccobi-Bellman equation is

$$\frac{\partial}{\partial t}u(t, x) = \max\left(x\left(1 - \frac{\partial}{\partial x}u(t, x)\right), (1 - x)\frac{\partial}{\partial x}u(t, x)\right).$$

The optimal solution of this HJB equation can be given in closed form. The optimal action is to chose action 1 if $x > 1/2$ or $x > 1 - \exp(-t)$, and 0 otherwise.

4.2. Optimal vaccination of a population

The second example has two purposes. The first objective is to provide a justification of a classical problem in population dynamics. The second one is to show that the mean field approach provides numerical insights in this case as well. Even if there is no close form for the solution, the optimal action function can be shown to be of threshold type for the limit problem. This can be used to make numerical computations much easier.

We consider the propagation of a disease in a population. The classical epidemiological model is the SIR model of [12]. An individual is either susceptible (S), infected (I) or recover (R). A susceptible individual become infected by contact of an infected individual or of an external source (such as plants or animals). Infected individual can recover with a constant rate. A recovered individual has already have the disease and is immune for a time. It can be susceptible again after some time. There exist many variations of this model. The reader is referred to [10] for a more complete description.

This problem can be viewed as an optimization problem. To each infected individual is associated a cost, representing for example the fact that it cannot work. A cost is also associated to the vaccination program that can represent the cost of the vaccination campaign or the side-effects on the population. The focus here is to study vaccination strategies of a population. A controller

can choose to spend some amount of money to vaccinate people: it increases the probability of a susceptible person to go in the recovered state without having been infected.

The model is as follows. By abuse of notation, we denote S, I, R the proportion of individuals each state. Each time step, a susceptible person becomes infected with a probability $\beta I/N$ (probability of meeting some who is infected). An infected person becomes recovered with probability γ/N and a recovered person becomes susceptible with probability μ/N. The vaccination intensity is modeled by a value $u \in [0; 1]$. A susceptible person becomes recovered with probability u/N. The immediate cost at each step is proportional to the number of infected people plus a cost for the vaccination. The cost of vaccination is typically chosen to be proportional to the square of the vaccination rate. Therefore the vaccination cost is chosen to be $I + \tau u^2$.

This system satisfies assumptions (A_1, A_2, A_3) and we can compute its mean field limit. The infinite system is described by the following system of differential equations:

$$\begin{array}{rcl} \frac{\partial S}{\partial t} & = & \gamma R - uS - \beta IS \\ \frac{\partial I}{\partial t} & = & \beta IS - \mu I \\ \frac{\partial R}{\partial t} & = & \mu I + uS - \gamma R. \end{array} \tag{13}$$

The objective is to find $u(t)$ such that $\int_0^T I + \tau u^2$ is minimized.

This deterministic optimization problem as well as several variations has been studied in the literature. In [21], the authors show the existence and uniqueness of an optimal control and provide a numerical method to compute the optimal solution. The same technique has also been applied to HIV treatment, see [11, 13] for example. This model has also been studied when the state space is restricted to a two states S and I [19]. In this case, the vaccination is replaced by a treatment that makes people go from I to S. It can be shown that there exists a threshold I^* such that if we are far enough from the end of the time horizon, the strategy is to fully vaccinate (choose $u = 1$) when $I > I^*$ and no vaccination if $I < I^*$.

Theorem 2 makes the formal link between the description of the model on an individual level and the resolution of an optimization problem on differential equations, that is often left loose. It shows that most of these models, based on differential equations and are in fact mean field approximation of stochastic behaviors. Using our framework, we can show that the optimal control computed by solving the ODEs are indeed asymptotically optimal for the discrete model at the individual level.

4.3. Brokering problem

We consider a model of a model of a volunteer computing system like BOINC `http://boinc.berkeley.edu/`. Volunteer computing means that people make available their personal computer for a computing system. When they do not use their computer, their computer is available for the computing system. However, as soon as they start using their computer, it becomes unavailable for the computing system. These systems become more and more popular and provide large computing power at a very low cost [14].

The Markovian model with N objects is defined as follows. The N objects represents the number of users that can submit jobs to the system as well as the number of resources that can perform the jobs. The resources are grouped in a small number of clusters and all resources in the same cluster share the same characteristics in term of speed and availability. Users send jobs to a central broker whose job is to balance the load among the clusters.

There are U^N users. Each user has a state $u \in \{\text{active}, \text{inactive}\}$. At each time step, an active user sends a job to the system with a probability p_j^N and becomes inactive with probability p_i^N. An inactive user sends no jobs to the system and becomes active with probability p_a^N.

The are M clusters. The cluster i contains Q_i^N resources. Each resource has a buffer of bounded size J_i. A resource can either be available or busy. If it is available and if it has one or more job in its queue, it completes one job with probability μ_i^N at this time slot. An available resource becomes busy with probability p_b^N. In that case, it discards all the packets of its buffer. A busy resource becomes available with probability p_v^N.

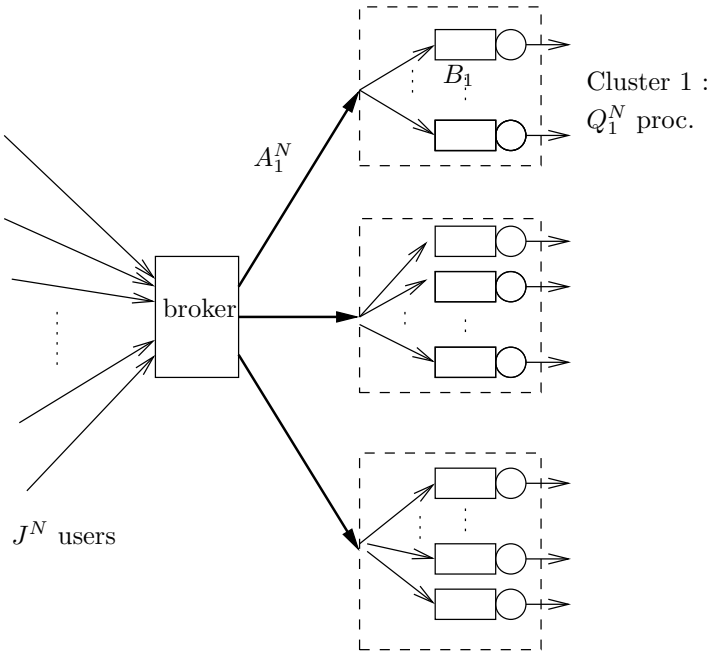

Figure 1: The brokering problem in a desktop grid system, such as Boinc

At each time step, the broker takes an action $a \in \mathcal{P}(\{1 \dots M\})$ and sends the packets it received to the clusters according to the distribution a. A packet sent to cluster i joins the queue of one resource, j according to a local rule (for example the shortest queue among h of them chosen uniformly). If the queue of resource j is full, the packet is lost. The goal of the broker is to minimize the total size of the queues over a finite horizon (and hence the response time of accepted packets) plus the number of losses.

This model is represented by Figure 1.

We consider the model with an intensity $I(N) \stackrel{\text{def}}{=} 1/N$. The number M of clusters is fixed and does not depend on N, as well as the sizes J_i of the buffers. However, both the number of users J^N and the number of resources in the clusters Q^N go to infinity with N. All the probabilities scale with $1/N$. For example the probability for a resource in cluster i to complete a task during one time step is $\mu_i^N \stackrel{\text{def}}{=} \mu_i/N$ for some constant μ_i.

The limiting system is described by the variable $u_a(t)$, that represents the fraction of active users, and the variables $q_{k,i}(t)$ and $b_k(t)$ that represents the fraction of resources in cluster k having i jobs and the fraction of resources in cluster k that are busy. For an action function α, we denote by $\alpha_i(t)$ the fraction of packets send to cluster i. Denoting u the fraction of users (both active or inactive) and q_k the fraction of processors in cluster k, we get the following equations:

$$\frac{\partial u_a(t)}{\partial t} = -p_i u_a(t) + p_a(u - u_a(t)) \tag{14}$$

$$\frac{\partial q_{k,0}(t)}{\partial t} = p_a b_k(t) - \frac{\alpha_k(t) p_j u_a(t)}{q_k} q_{k,0}(t) + \mu_k q_{k,i+1} - p_b q_{k,0}(t) \tag{15}$$

$$\frac{\partial q_{k,i}(t)}{\partial t} = \frac{\alpha_k(t) p_j u_a(t)}{q_k}(q_{k,i-1}(t) - q_{k,i}(t)) + \mu_k(q_{k,i+1} - q_{k,i}) - p_b q_{k,i}(t) \tag{16}$$

$$\frac{\partial q_{k,J_k}(t)}{\partial t} = \frac{\alpha_k(t) p_j u_a(t)}{q_k} q_{k,J_k-1}(t) + -\mu_k q_{k,J_k} - p_b q_{k,J_k}(t) \tag{17}$$

$$\frac{\partial b_k(t)}{\partial t} = -p_v b_k(t) + p_b \sum_{i=0}^{J_k} q_{k,i}(t). \tag{18}$$

where (15) and (17) hold for all cluster k and (16) holds for all cluster k and all $i \leq J_k$. The cost associated to the action function α is:

$$\int_0^T \sum_{k=1}^{M} \sum_{i=1}^{J_k} i q_{k,i}(t) + C \left(\sum_{k=1}^{M} \frac{\alpha_k(t) p_j u_a(t)}{q_k} (q_{k,J_k}(t) + b_k(t)) + \sum_{k=1}^{M} p_b \sum_{i=1}^{J_k} i q_{k,i}(t) \right) dt \qquad (19)$$

The first part of (19) represents the cost induced by the number of jobs in the system. The second part of (19) represents the cost induced by the losses. C is a parameter to give more or less weight on the cost induced by the lost.

The HJB problem becomes minimizing (19) subjects that the variables $u_a, q_{k,i}, b_k$ satisfy Equations (14) to (18). The limit system is made of a system of $(B+2).M$ ODEs. Solving the HJB equation in this case can be challenging but remains more tractable than solving the original Bellman equation over B^{NM} states. The curse of dimensionality is so acute here that the finite system cannot be solved numerically with more than 10 processors in total [5] and one usually settles for heuristics approaches.

References

[1] M. Benaïm. Dynamics of stochastic approximation algorithms. *Séminaire de Probabilités XXXIII. Lecture Notes in Math*, 1709:1–68, 1999.

[2] M. Benaım and J. Weibull. Deterministic approximation of stochastic evolution in games: a generalization. Technical report, mimeo, 2003.

[3] Michel Benaim and Jean-Yves Le Boudec. A Class Of Mean Field Interaction Models for Computer and Communication Systems. *Performance Evaluation*, 65(11-12):823–838, 2008.

[4] A. Benveniste, P. Priouret, and M. Métivier. *Adaptive algorithms and stochastic approximations*. 1990.

[5] Vandy Berten and Bruno Gaujal. Grid brokering for batch allocation using indexes. In Springer, editor, *Network Control and Optimization*, volume 4465 of *LNCS*.

[6] W. Chen, D. Huang, A. Kulkarni, J. Unnikrishnan, Q. Zhu, P. Mehta, S. Meyn, and A. Wierman. Approximate Dynamic Programming using Fluid and Diffusion Approximations with Applications to Power Management. In *Submitted to the 48th IEEE Conference on Decision and Control*, 2009.

[7] DP De Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

[8] Nicolas Gast and Bruno Gaujal. A Mean Field Approach for Optimization in Particles Systems and Applications. Research Report RR-6877, INRIA, 2009.

[9] Z. Gorodeisky. Deterministic approximation of best-response dynamics for the Matching Pennies game. *Games and Economic Behavior*, 66(1):191–201, 2009.

[10] H.W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[11] H.R. Joshi. Optimal control of an HIV immunology model. *Optimal control applications and methods*, 23(4):199–213, 2002.

[12] WO Kermack and AG McKendrick. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal Society of London.*, pages 700–721, 1927.

[13] D. Kirschner, S. Lenhart, and S. Serbin. Optimal control of the chemotherapy of HIV. *Journal of Mathematical Biology*, 35(7):775–792, 1997.

[14] Derrick Kondo, Bahman Javadi, Paul Malecot, Franck Cappello, and David Anderson. Cost-benefit analysis of cloud computing versus desktop grids. In *18th International Heterogeneity in Computing Workshop*, Rome, 2009.

[15] J.Y. Le Boudec, D. McDonald, and J. Mundinger. A generic mean field convergence result for systems of interacting objects. In *Quantitative Evaluation of Systems, 2007. QEST 2007. Fourth International Conference on the*, pages 3–18, 2007.

[16] S. Meyn. *Control techniques for complex networks*. Cambridge University, 2007.

[17] M.L. Puterman. Markov decision processes: discrete stochastic dynamic programming. 1994.

[18] S.M. Ross. *Stochastic Processes, 2nd ed.* John Wiley and Sons, Inc. USA, 1996.

[19] S.P. Sethi and P.W. Staats. Optimal control of some simple deterministic epidemic models. *Journal of the Operational Research Society*, pages 129–136, 1978.

[20] J.N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[21] G. Zaman, Y. Han Kang, and I.H. Jung. Stability analysis and optimal vaccination of an SIR epidemic model. *Biosystems*, 93(3):240–249, 2008.

5. Proofs

5.1. Theorem 1

The proof is inspired by the method in [2]. The main idea of the proof is to write

$$\begin{aligned}\left\|M_\pi^N(k) - \phi_{kI(N)}(m_0, A_\pi^N)\right\| &\leq \left\|M_\pi^N(k) - M^N(0) - \sum_{j=0}^{k-1} f^N(j)\right\| \\ &\quad + \left\|M^N(0) + \sum_{j=0}^{k-1} f^N(j) - \phi_{kI(N)}(m_0, A_\pi^N)\right\|\end{aligned}$$

where $f^N(k) \stackrel{\text{def}}{=} F^N\left(M_\pi^N(k), \pi_k(M_\pi^N(k))\right]$ is the drift at time k if the empirical measure is $M_\pi^N(k)$. The first part is bounded with high probability using a Martingale argument (Lemma 6) and the second part is bounded using an integral formula.

Recall that $\bar{M}_\pi^N(t) \stackrel{\text{def}}{=} M_\pi^N\left(\left\lfloor \frac{t}{I(N)} \right\rfloor\right)$, i.e. $\bar{M}_\pi^N(kI(N)) = M_\pi^N(k)$ for $k \in \mathbb{N}$ and $\bar{M}_\pi^N$ is piecewise constant and right-continuous. Let $\Delta_\pi^N(k)$ be the number of objects that change state between time slots k and $k+1$. Thus,

$$\left\|M_\pi^N(k+1) - M_\pi^N(k)\right\| \leq N^{-1}\sqrt{2}\Delta_\pi^N(k) \tag{20}$$

and thus

$$\left\|\hat{M}_\pi^N(t) - \bar{M}_\pi^N(t)\right\| \leq N^{-1}\sqrt{2}\Delta_\pi^N(k) \tag{21}$$

as well, with $k = \left\lfloor \frac{t}{I(N)} \right\rfloor$. Define

$$Z_\pi^N(k) = M_\pi^N(k) - M^N(0) - \sum_{j=0}^{k-1} F^N\left(M_\pi^N(j), \pi_j(M_\pi^N(j))\right) \tag{22}$$

and let $\hat{Z}_\pi^N(t)$ be the continuous, piecewise linear interpolation such that $\hat{Z}_\pi^N(kI(N)) = Z_\pi^N(k)$ for $k \in \mathbb{N}$. Recall that $A_\pi^N(t) \stackrel{\text{def}}{=} \pi_{\lfloor t/I(N) \rfloor}(M^N(\lfloor t/I(N) \rfloor))$ – $A_\pi^N(t)$ is the action taken by the controller at time $t/I(N)$. It follows from these definitions that:

$$\begin{aligned}\hat{M}_\pi^N(t) &= M_\pi^N(0) + \int_0^t \frac{1}{I(N)} F^N\left(\bar{M}_\pi^N(s), A_\pi^N(s)\right) ds + \hat{Z}_\pi^N(t) \\ &= M_\pi^N(0) + \int_0^t \frac{1}{I(N)} F^N\left(\hat{M}_\pi^N(s), A_\pi^N(s)\right) ds + \hat{Z}_\pi^N(t) \\ &\quad + \int_0^t \frac{1}{I(N)} \left[F^N\left(\bar{M}_\pi^N(s), A_\pi^N(s)\right) - F^N\left(\hat{M}_\pi^N(s), A_\pi^N(s)\right)\right] ds\end{aligned}$$

Using the definition of the semi-flow $\phi_t(m_0, A_\pi^N) = m_0 + \int_0^t f(\phi_s(m_0, A_\pi^N), A_\pi^N(s))ds$, we get:

$$\begin{aligned}\hat{M}_\pi^N(t) - \phi_t(m_0, A_\pi^N) &= M_\pi^N(0) - m_0 + \hat{Z}_\pi^N(t) \\ &\quad + \int_0^t \frac{1}{I(N)} \left[F^N\left(\hat{M}_\pi^N(s), A_\pi^N(s)\right) - F^N\left(\phi_s(m_0, A_\pi^N), A_\pi^N(s)\right)\right] ds \\ &\quad + \int_0^t \left[\frac{1}{I(N)} F^N\left(\phi_s(m_0, A_\pi^N), A_\pi^N(s)\right) - f\left(\phi_s(m_0, A_\pi^N), A_\pi^N(s)\right)\right] ds \\ &\quad + \int_0^t \frac{1}{I(N)} \left[F^N\left(\bar{M}_\pi^N(s), A_\pi^N(s)\right) - F^N\left(\hat{M}_\pi^N(s), A_\pi^N(s)\right)\right] ds\end{aligned}$$

Applying Assumption (A2) to the third line, (A3) to the second and fourth lines, and Equation (21) to the fourth line leads to:

$$\begin{aligned}\left\|\hat{M}_\pi^N(t) - \phi_t(m_0, A_\pi^N)\right\| &\leq \left\|M_\pi^N(0) - m_0\right\| + \left\|\hat{Z}_\pi^N(t)\right\| + L_1 \int_0^t \left\|\hat{M}_\pi^N(s) - \phi_s(m_0, A_\pi^N)\right\| ds \\ &\quad + I_0(N)t + \frac{\sqrt{2}L_1 I(N)}{N} \sum_{k=0}^{\lfloor \frac{t}{I(N)} \rfloor} \Delta_\pi^N(k)\end{aligned}$$

For all N, π, T, $b_1 > 0$ and $b_2 > 0$, define

$$\Omega_1 = \left\{ \omega \in \Omega : \sup_{0 \leq k \leq \frac{T}{I(N)}} \sum_{j=0}^{k} \Delta_\pi^N(j) > b_1 \right\},\ \Omega_2 = \left\{ \omega \in \Omega : \sup_{0 \leq k \leq \frac{T}{I(N)}} \left\|Z_\pi^N(k)\right\| > b_2 \right\} \tag{23}$$

Assumption (A1) implies conditions on the first and second order moment of $\Delta_\pi^N(k)$. Therefore by Lemma 5, this shows that for any $b_1 > 0$:

$$\mathbb{P}(\Omega_1) \leq \frac{TN^2}{b_1^2}\left[I_2(N) + \frac{I_1(N)^2}{I(N)^2}(T + I(N))\right] \tag{24}$$

Moreover, we show in Lemma 6 that:

$$\mathbb{P}(\Omega_2) \leq 2S^2 \frac{T}{b_2^2}\left[2I_2(N) + I(N)\left[(I_0(N) + L_2)\right]^2\right] \tag{25}$$

Now fix some $\epsilon > 0$ and let $b_1 = \frac{N\epsilon}{2\sqrt{2}L_1 I(N)}$, $b_2 = \epsilon/2$. For $\omega \in \Omega \setminus (\Omega_1 \cup \Omega_2)$ and for $0 \leq t \leq T$:

$$\begin{aligned}\left\|\hat{M}_\pi^N(t) - \phi_t(m_0, A_\pi^N)\right\| &\leq \left\|M_\pi^N(0) - m_0\right\| + \epsilon + I_0(N)T \\ &\quad + L_1 \int_0^t \left\|\hat{M}_\pi^N(s) - \phi_s(m_0, A_\pi^N)\right\| ds\end{aligned}$$

By Grönwall's lemma:

$$\left\|\hat{M}_\pi^N(t) - \phi_t(m_0, A_\pi^N)\right\| \leq \left[\left\|M_\pi^N(0) - m_0\right\| + \epsilon + I_0(N)T\right] e^{L_1 t} \tag{26}$$

and this is true for all $\omega \in \Omega \setminus (\Omega_1 \cup \Omega_2)$. We apply the union bound $\mathbb{P}(\Omega_1 \cup \Omega_2) \leq \mathbb{P}(\Omega_1) + \mathbb{P}(\Omega_2)$ which, with Eq.(24) and Eq.(25), concludes the proof.

The proof of Theorem 1 uses the following lemmas.

Lemma 5. *Let $(W_k)_{k\in\mathbb{N}}$ be a sequence of square integrable, nonnegative random variables, adapted to a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$, such that $W_0 = 0$ a.s. and for all $k \in \mathbb{N}$:*

$$\begin{aligned}\mathbb{E}(W_{k+1} | \mathcal{F}_k) &\leq \alpha \\ \mathbb{E}\left(W_{k+1}^2 \middle| \mathcal{F}_k\right) &\leq \beta\end{aligned}$$

Then for all $n \in \mathbb{N}$ and $b > 0$:

$$\mathbb{P}\left(\sup_{0 \leq k \leq n} (W_0 + ... + W_k) > b\right) \leq \frac{n\beta + n(n+1)\alpha^2}{b^2} \tag{27}$$

Proof. Let $m_n = \beta n + n(n+1)\alpha^2$ and $Y_n = \sum_{k=0}^n W_k$. It follows that $\mathbb{E}(Y_n) \leq \alpha n$ and

$$\mathbb{E}\left(Y_{n+1}^2 \middle| \mathcal{F}_n\right) \leq \beta + 2n\alpha^2 + Y_n^2$$

so that $Z_n = Y_n^2 - m_n$ is a supermartingale w.r. $\mathcal{F}_n$. Let T_n be the first time $k \leq n$ at which $Y_k > b$ if it exists, otherwise $T_n = n$, so that $Y_{T_n} > b$ if and only if $\sup_{0\leq k\leq n}(Y_k) > b$. By the optional stopping theorem [18, Thm 6.4.1]:

$$\mathbb{E}(Z_{T_n}) \leq \mathbb{E}(Z_0) = 0$$

thus $\mathbb{E}\left(Y_{T_n}^2\right) \leq \mathbb{E}(m_{T_n}) \leq m_n$. Now $\mathbb{E}\left(Y_{T_n}^2\right) \geq \mathbb{E}\left(Y_{T_n}^2 1_{Y_{T_n}>b}\right) \geq b^2\mathbb{P}(Y_{T_n} > b)$ thus $\mathbb{P}(Y_{T_n} > b) \leq \frac{m_n}{b^2}$. □

Lemma 6. *Define Z_π^N as in Eq.(22). For all $N \geq 2$, $b > 0$, $T > 0$ and all policy π:*

$$\mathbb{P}\left(\sup_{0\leq k\leq \lfloor \frac{T}{I(N)} \rfloor} \left\|Z_\pi^N(k)\right\| > b\right) \leq 2S^2\frac{T}{b^2}\left[2I_2(N) + I(N)\left[(I_0(N) + L_2)\right]^2\right]$$

Proof. The proof is inspired by the methods in [1]. For fixed N and $h \in \mathbb{R}^S$, let

$$L_k = \langle h, Z_\pi^N(k)\rangle$$

By the definition of Z^N, L_k is a martingale w.r. to the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ generated by M_π^N. Thus

$$\mathbb{E}\left((L_{k+1} - L_k)^2\middle|\mathcal{F}_k\right) = \mathbb{E}\left(\langle h, M_\pi^N(k+1) - M_\pi^N(k)\rangle^2\middle|\mathcal{F}_k\right) + \langle h, F^N\left(M_\pi^N(k), \pi_k(M_\pi^N(k))\right)\rangle^2$$

By Assumption (A2):

$$\left|\langle h, F^N\left(M_\pi^N(k), \pi(M_\pi^N(k))\right)\rangle\right| \leq (I_0(N) + L_2)\, I(N)\, \|h\|$$

Thus, using Eq.(20) and Assumption (A5):

$$\begin{aligned}\mathbb{E}\left((L_{k+1} - L_k)^2\middle|\mathcal{F}_k\right) &\leq \|h\|^2\left[N^{-2}2\mathbb{E}\left(\Delta_\pi^N(k)^2\middle|\mathcal{F}_k\right) + \left[(I_0(N) + L_2)\, I(N)\right]^2\right] \\ &\leq \|h\|^2\left[2I(N)I_2(N) + \left[(I_0(N) + L_2)\, I(N)\right]^2\right]\end{aligned}$$

We now apply Kolmogorov's inequality for martingales and obtain

$$\mathbb{P}\left(\sup_{0\leq k\leq n} L_k > b\right) \leq \frac{n}{b^2}\|h\|^2\left[2I(N)I_2(N) + \left[(I_0(N) + L_2)\, I(N)\right]^2\right]$$

Let Ξ_h be the set of $\omega \in \Omega$ such that $\sup_{0\leq k\leq n}\langle h, Z_\pi^N(k)\rangle \leq b$ and let $\Xi := \bigcap_{h=\pm\vec{e}_i, i=1\ldots S}\Xi_h$ where $\vec{e}_i$ is the ith vector of the canonical basis of $\mathbb{R}^S$. It follows that, for all $\omega \in \Xi$ and $0 \leq k \leq n$ and $i = 1\ldots S$: $\left|\langle Z_\pi^N(k), \vec{e}_i\rangle\right| \leq b$. This means that for all $\omega \in \Xi$: $\left\|Z_\pi^N(k)\right\| \leq \sqrt{S}b$. By the union bound applied to the complement of Ξ, we have

$$1 - \mathbb{P}(\Xi) \leq 2S\frac{n}{b^2}\left[I(N)I_2(N) + \left[(I_0(N) + L_2)\, I(N)\right]^2\right]$$

Thus we have shown that, for all $b > 0$:

$$\mathbb{P}\left(\sup_{0\leq k\leq n}\left\|Z_\pi^N(k)\right\| > \sqrt{S}b\right) \leq 2S\frac{nI(N)}{b^2}\left[I_2(N) + I(N)\left[(I_0(N) + L_2)\right]^2\right]$$

which, by changing b to $b/\sqrt{S}$, shows the result. □

5.2. Proof of Theorem 2

We use the same notation as in the proof of Theorem 1. By definition of V^N, v and the time horizons:

$$\begin{aligned} V_\pi^N(M^N(0)) - \mathbb{E}\left(v_{A_\pi^N}(m_0)\right) &= \mathbb{E}\left(\int_0^{H^N I(N)} r(\bar{M}_\pi^N(s), A_\pi^N(s)) - r(m_{A_\pi^N}(s), A_\pi^N(s))ds\right) \\ &\quad -\mathbb{E}\left(\int_{H^N I(N)}^T r(m_{A_\pi^N}(s), A_\pi^N(s))ds\right) \end{aligned}$$

The latter term is bounded by $I(N)\left\|r\right\|_\infty$. Let $\epsilon > 0$ and $\Omega_0 = \Omega_1 \cup \Omega_2$ where Ω_1, Ω_2 are as in the proof of Theorem 1. Thus $\mathbb{P}(\Omega_0) \leq \frac{J(N,T)}{\epsilon^2}$ and, using the Lipschitz continuity of r in m (with constant K_r):

$$\left|V_\pi^N(M^N(0)) - \mathbb{E}\left[v_{A_\pi^N}(m_0)\right]\right| \leq I(N)\left\|r\right\|_\infty + \frac{2\left\|r\right\|_\infty J(N,T)}{\epsilon^2} + K_r\mathbb{E}\left[1_{\omega\notin\Omega_0}\int_0^T \left\|\bar{M}_\pi^N(s) - m_{A_\pi^N}(s)\right\| ds\right]$$

For $\omega \notin \Omega_0$ and $s \in [0,T]$: $\int_0^T \left\|\bar{M}_\pi^N(s) - \hat{M}_\pi^N(s)\right\| ds \leq \frac{\epsilon I(N)}{2L_1}$ and, by Eq.(26), $\int_0^T \left\|\hat{M}_\pi^N(s) - m_{A_\pi^N}(s)\right\| ds \leq \left(\left\|M^N(0) - m_0\right\| + I_0(N)T + \epsilon\right)\frac{e^{L_1T}-1}{L_1}$ thus

$$\left|V_\pi^N(M^N(0)) - \mathbb{E}\left[v_{A_\pi^N}(m_0)\right]\right| \leq B_\epsilon(N, \left\|M^N(0) - m_0\right\|) \tag{28}$$

where

$$B_\epsilon(N,\delta) \stackrel{\text{def}}{=} I(N)\left\|r\right\|_\infty + K_r\left(\delta + I_0(N)T + \epsilon\right)\frac{e^{L_1T}-1}{L_1} + \frac{K_rI(N)}{2L_1}\epsilon + \frac{2\left\|r\right\|_\infty J(N,T)}{\epsilon^2}$$

This holds for every $\epsilon > 0$, thus

$$\left|V_\pi^N(M^N(0)) - \mathbb{E}\left[v_{A_\pi^N}(m_0)\right]\right| \leq B(N, \left\|M^N(0) - m_0\right\|) \tag{29}$$

where $B(N,\delta) \stackrel{\text{def}}{=} \inf_{\epsilon>0} B_\epsilon(N,\delta)$. By direct calculus, one finds that $\inf_{\epsilon>0}\left(a\epsilon + b/\epsilon^2\right) = 3/2^{\frac{2}{3}}a^{\frac{2}{3}}b^{\frac{1}{3}}$ for $a > 0, b > 0$, which gives the required formula for $B(N,\delta)$.

5.3. Proof of Theorem 3

Let $\bar{\alpha}^N$ be the right-continuous function constant on the intervals $[kI(N);(k+1)I(N))$ such that $\bar{\alpha}^N(s) = \alpha(s)$. $\bar{\alpha}^N$ can be viewed as a policy independent of m. Therefore, by Theorem 1, on the set $\Omega \setminus (\Omega_1 \cup \Omega_2)$, for every $t \in [0;T]$:

$$\left\|\hat{M}_\alpha(t) - \phi_t(m_0,\alpha)\right\| \leq \left[\left\|M^N(0) - m_0\right\| + I_0(N)T + \epsilon\right]e^{L_1T} + u(t)$$

with $u(t) \stackrel{\text{def}}{=} \left|\phi_t(m_0,\bar{\alpha}^N) - \phi_t(m_0,\alpha)\right|$. We have

$$\begin{aligned} u(t) &\leq \int_0^t \left|f(\phi_s(m_0,\alpha),\alpha(s)) - f(\phi_s(m_0,\bar{\alpha}^N),\bar{\alpha}^N(s))\right| ds \\ &\leq \int_0^t K\left(\left\|\phi_s(m_0,\alpha) - \phi_s(m_0,\bar{\alpha}^N)\right\| + d(\alpha(s),\bar{\alpha}^N(s)\right) ds \\ &\leq K\int_0^t u(s)ds + Kd_1 \end{aligned}$$

where $d_1 \stackrel{\text{def}}{=} \int_0^T \left\|\alpha(t) - \bar{\alpha}^N(t)\right\| dt$. Therefore, using Gronwall's inequality, we have

$$u(t) \leq Kd_1e^{KT}$$

By Lemma 7, this shows Eq.(7). The rest of the proof is as for Theorem 2.

Lemma 7. *If α is a piecewise Lipschitz continuous action function on $[0;T]$, of constant K_α, and with at most p discontinuity points, then*

$$\int_0^T d(\alpha(t), \bar{\alpha}^N(t))dt \leq TI(N)\left(\frac{K_\alpha}{2} + 2\left(1 + \min(1/I(N), p)\right)\|\alpha\|_\infty\right).$$

Proof of lemma 7. Let first assume that $T = kI(N)$. The left handside $d_1 = \int_0^T d(\alpha(t), \bar{\alpha}^N(t))dt$ can be decomposed on all intervals $[iI(N), (i+1)I(N))$:

$$\begin{aligned} d_1 &= \sum_{i=0}^{\lfloor T/I(N)\rfloor} \int_{iI(N)}^{(i+1)I(N)} \left\|\alpha(s) - \bar{\alpha}^N(s)\right\| ds \\ &\leq \sum_{i=0}^{\lfloor T/I(N)\rfloor} \int_{iI(N)}^{(i+1)I(N)} \|\alpha(s) - \alpha(iI(N))\| ds \end{aligned}$$

If α has no discontinuity point on $[iI(N), (i+1)I(N))$, then

$$\int_{iI(N)}^{(i+1)I(N)} d(\alpha(s), \alpha(iI(N)))ds \leq \int_0^{I(N)} K_\alpha s ds \leq K_\alpha 2I(N)^2$$

If α has one or more discontinuity points on $[iI(N), (i+1)I(N))$, then

$$\int_{iI(N)}^{(i+1)I(N)} d(\alpha(s)\alpha(iI(N)))ds \leq \int_{iI(N)}^{(i+1)I(N)} 2\|\alpha\|_\infty ds \leq 2\|\alpha\|_\infty I(N)$$

There are at most $\min(1/I(N), p)$ intervals $[iI(N), (i+1)I(N)]$ that have discontinuity points which shows that

$$d_1 \leq TI(N)(\frac{K_\alpha}{2} + \min(1/I(N), p)2\|\alpha\|_\infty).$$

If $T \neq kI(N)$, then $T = kI(N) + t$ with $0 < t < I(N)$. Therefore, there is an additional term of $\int_{kI(N)}^{kI(N)+t} d(\alpha(s), \bar{\alpha}^N(s))ds \leq 2\|\alpha\|_\infty I(N)$. □