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

Nicolas Gast, Bruno Gaujal and Jean-Yves Le Boudec

Abstract—We study the convergence of Markov decision processes, composed of a large number of objects, to optimization problems on ordinary differential equations. We show that the optimal reward of such a Markov decision process, which satisfies 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 an 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. They are used to illustrate and justify the construction of the controlled ODE and to show the advantage of solving a continuous HJB equation rather than a large discrete Bellman equation.

I. INTRODUCTION

In this paper we study dynamic optimization problems on Markov decision processes 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 $P^N$. This kernel depends on the number of objects in each state, as well as on the decisions of a centralized controller. Our goal 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 [2], [19], the authors show that under mild conditions, as $N$ grows, the system converges to a deterministic limit. The limiting system can be of two types, depending on the intensity $I(N)$ (the intensity is the probability that 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 [19]. 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 (ODEs).

Contributions

Here, 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 the controller receives 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 $N$ becomes large this problem converges to an optimization problem on an ordinary differential equation.

More precisely, we focus on the case where the Markov decision process is such that its empirical occupancy measure is also Markovian; this occurs when the system consists of many interacting objects, the objects can be observed only through their state and the system evolution depends only on the collection of all states. We show that the optimal reward converges to the optimal reward of the mean field approximation of the system, which is given by the solution of an HJB equation. Furthermore, the optimal policy of the mean field approximation is also asymptotically optimal in $N$, 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 that is obtained as a solution of the ODE, i.e. the mean field limit, controlled by random actions.

This convergence result has an algorithmic by-product. Roughly speaking, when confronted with a large Markov decision process, we 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 in $N$.

Our 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. In Section IV-C2 we illustrate this point with an example of malware infection in computer systems.

The second one 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, which 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 computers only allow us to tackle such control problems when $N$ is no larger than a few tens [20].

The mean field approach offers an alternative to brute force computations. By letting $N$ go to infinity, the discrete problem is replaced by an Hamilton-Jacobi-Bellman equation where the dimensionality of the original system has been hidden in the occupancy measure. Solving the HJB equation numerically is sometimes rather easy, as in the examples in Sections IV-C1 and IV-C2. It provides a deterministic optimal policy whose
reward with a finite (but large) number of objects is remarkably close to the optimal reward.

Related Work

Several 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 [25], [8], the curse of dimensionality of dynamic programming is circumvented by approximating the value function by linear regression. Here, we use instead a mean field limit approximation and prove asymptotic optimality in $N$ of limit policy.

In [10], the authors also consider Markov decision processes with a growing number of objects, but when the intensity is $O(1)$. In their case, the optimization problem of the system of size $N$ converges to a deterministic optimization problem in discrete time. In this paper however, we focus on the $o(1)$ case, which is substantially different from the discrete time case because the limiting system does not evolve in discrete time anymore.

Actually, most of the papers dealing with mean field limits of optimization problems over large systems are set in a game theory framework, leading to the concept of mean field games introduced in [18]. The objects composing the system are seen as $N$ players of a game with distributed information, cost and control; their actions lead to a Nash equilibrium. To the best of our knowledge, the classic case with global information and centralized control has not yet been considered. Our work focuses precisely on classic Markov decision problems, where a central controller (our objects are passive), aims at minimizing a global cost function.

For example, a series of papers by M. Huang, P.E. Caines and P. Malhame such as [12], [13], [14] investigate the behavior of systems made of a large number of objects under distributed control. They mostly investigate Linear-Quadratic-Gaussian (LQG) dynamics and use the fact that, here, the solution can be given in closed form as a Riccati equation to show that the limit satisfies a Nash fixed point equation. Their more general approach uses the Nash Equivalence Certainty principle introduced in [12]. The limit equilibrium could or could not be a global optimal. Here, we consider the general case where the dynamics and the cost may be arbitrary (we do not assume LQG Dynamics) so that the optimal policy is not given in closed form. The main difference with their approach comes from the fact that we focus instead on centralized control to achieve a global optimum. The techniques to prove convergence are rather different. Our proofs are more in line with classic mean field arguments and use stochastic approximation techniques.

Another example is the work of Tembine and others [23], [24], on the limits of games with many players. The authors provide conditions under which the limit when the number of players grows to infinity commutes with the fixed point equation satisfied by a Nash equilibrium. Again, our investigation solves a different problem and focuses on the centralized case. In addition, our approach is more algorithmic; we construct two intermediate systems: one with a finite number of objects controlled by a limit policy and one with a limit system controlled by a stochastic policy induced by the finite system.

Structure of the paper

The rest of the paper is structured as follows. In Section II we give definitions, some notation and hypotheses. In Section III we describe our main theoretical contributions. In Section IV we describe our resulting algorithm and illustrate the application of our method with a few examples. The details of all proofs are in Section V and Section VI concludes the paper.

II. NOTATIONS AND DEFINITIONS

A. System with $N$ Objects

We consider a system composed of $N$ objects. Each object has a state from the finite set $S = \{1 \ldots S\}$. 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 $X^N(k) \overset{\text{def}}{=} (X^N_1(k) \ldots X^N_N(k))$. For all $i \in S$, we denote by $M^N(k)$ the empirical measure of the objects $(X^N_1(k) \ldots X^N_N(k))$ at time $k$:

$$M^N(k) \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \delta_{X^N_i(k)}.$$  \hfill (1)

where $\delta_{x}$ denotes the Dirac measure in $x$. $M^N(k)$ is a probability measure on $\mathcal{S}$ and its $i$th component $M^N(k)[i]$ denotes the proportions of objects in state $i$ at time $k$ (also called the occupancy measure): $M^N(k)[i] = \frac{1}{N} \sum_{n=1}^{N} 1_{X^N_n(k)=i}$.

The system $(X^N(k))_{k \in \mathbb{N}}$ is a Markov process once the sequence of the actions taken by the controller is fixed. Let $\Gamma^N$ be the transition kernel, namely $\Gamma^N$ is a mapping $\mathcal{S}^N \times \mathcal{S}^N \times \mathcal{A} \to [0, 1]$, where $\mathcal{A}$ is the set of possible actions, such that for every $x \in \mathcal{S}^N$ and $a \in \mathcal{A}$, $\Gamma^N(x, a)$ is a probability distribution on $\mathcal{S}^N$ and further, if the controller takes the action $A^N(k)$ at time $t$ and the system is in state $X^N(k)$, then:

$$\mathcal{P}\{X^N(k+1)=y_1 \ldots y_N | X^N(k)=x_1 \ldots x_N, A^N(k)=a\} = \Gamma^N(x_1 \ldots x_N, y_1 \ldots y_N, a)$$  \hfill (2)

We assume that

(A0) Objects are observable only through their states in particular, the controller can observe the collection of all states $X^N_1, X^N_2, \ldots$, but not the identities $n = 1, 2, \ldots$. This assumption is required for mean field convergence to occur. In practice, it means that we need to put into the object state any information that is relevant to the description of the system.

Assumption (A0) translates into the requirement that the kernel be invariant by object re-labeling. Formally, let $\mathcal{S}^N$ be the set of permutations of $\{1, 2, \ldots, N\}$. By a slight abuse of notation, for $\sigma \in \mathcal{S}^N$ and $x \in \mathcal{S}^N$ we also denote with $\sigma(x)$ the collection of object states after the permutation, i.e. $\sigma(x) \overset{\text{def}}{=} (x_{\sigma^{-1}(1)}, \ldots x_{\sigma^{-1}(N)})$. The requirement is that

$$\Gamma^N(\sigma(x), \sigma(y), a) = \Gamma^N(x, y, a)$$  \hfill (3)
for all \(x, y \in \mathcal{S}^N, \sigma \in \mathcal{S}^N\) and \(a \in \mathcal{A}\). A direct consequence, shown in Section V, is:

**Theorem 1.** For any given sequence of actions, the process \(M^N(t)\) is a Markov chain.

**B. 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. \((A, d)\) is a compact metric space for some distance \(d\). The purpose of Markov decision control is to compute optimal policies. A policy \(\pi = (\pi_0, \pi_1, \ldots, \pi_k, \ldots)\) is a sequence of decision rules that specify the action at every time instant. The policy \(\pi_k\) might depend on the sequence of past and present states of the process \(X^N\), however, it is known that when the state space is finite, the action set compact and the kernel and the reward are continuous, there exists a deterministic Markovian policy which is optimal (see Theorem 4.4.3 in [21]). This implies that we can limit ourselves to policies that depend only on the current state \(X^N(k)\).

Further, we assume that the controller can only observe object states. Therefore she cannot make a difference between states that result from object relabeling, i.e. the policy depends on \(X^N(k)\) in a way that is invariant by permutation. By Lemma 2 in Section V-B, it depends on \(M^N(k)\) only. Thus, we may assume that, for every \(k\), \(\pi_k\) is a function \(\mathcal{P}(\mathcal{S}) \to \mathcal{A}\). Let \(M^N(k)\) denotes the occupancy measure of the system at time \(k\) when the controller applies policy \(\pi\).

The controller focuses on a finite-time horizon \([0; H^N]\). If the system has an occupancy measure \(M^N(k)\) at time step \(k \in [0; H^N]\) and if the controller chooses the action \(A^N(k)\), she gets an instantaneous reward \(r^N(M^N(k), A^N(k))\). At time \(H^N\), she gets a final reward \(r_f(M^N(H^N))\). The value of a policy \(\pi\) is the expected gain over the horizon \([0; H^N]\) starting from \(m_0\) when applying the policy \(\pi\). It is defined by

\[
V^N_\pi(m) \overset{\text{def}}{=} \mathbb{E}\left( \sum_{k=0}^{H^N-1} r^N(M^N_k), \pi(M^N_k) \right) + r_f(M^N(H^N)) | M^N(0) = m) \tag{4}
\]

The goal of the controller is to find an optimal policy that maximizes the value. We denote by \(V^*_\pi(m)\) the optimal value when starting from \(m\):

\[
V^*_\pi(m) = \sup_{\pi} V^N_\pi(m) \tag{5}
\]

**C. Scaling Assumptions**

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

\[
F^N(m, a) \overset{\text{def}}{=} \mathbb{E}(M^N(k+1) - M^N(k) | M^N(k) = m, A^N(k) = a) \tag{6}
\]

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 \(I(N) \in (0; 1), N = 1, 2, 3, \ldots\), called the intensity of the model with \(\lim_{N \to \infty} I(N) = 0\) and a sequence \(I_0(N), N = 1, 2, 3, \ldots\), also with \(\lim_{N \to \infty} I_0(N) = 0\) such that for all \(m \in \mathcal{P}(\mathcal{S})\) and \(a \in \mathcal{A}\): \(|F^N(m, a)/I(N) - f(m, a)| \leq I_0(N)\). 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 change 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 \((M^N(t))_{t \in \mathbb{R}^+}\) as the affine interpolation of \(M^N(k)\), rescaled by the intensity function, i.e. \(M^N\) is affine on the intervals \([kI(N), (k+1)I(N)]\), \(k \in \mathbb{N}\)

\[
\hat{M}^N(kI(N)) = M^N(k) \tag{7}
\]

Similarly, \(\hat{M}^N_\pi\) denotes the affine interpolation of the occupancy measure under policy \(\pi\). 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

\[
H^N = \left\lfloor \frac{T}{I(N)} \right\rfloor \tag{8}
\]

\[
r^N(m, a) = I(N)r_f(m, a) \tag{9}
\]

for every \(m \in \mathcal{P}(\mathcal{S})\) and \(a \in \mathcal{A}\) (where \([x]\) denotes the largest integer \(\leq x\)). The final reward \(r_f\) does not depend on \(N\).

**D. Limiting System (Mean Field Limit)**

We will see in Section III that as \(N\) grows, the stochastic system \(M^N_\pi\) converges to a deterministic limit \(m_\pi\), the mean field limit. For more clarity, all the stochastic variables (i.e., when \(N\) is finite) are in uppercase and 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 action functions do not take into account the state to determine the next action. For an action function \(\alpha\) and an initial condition \(m_0\), we consider the following ordinary integral equation for \(m(t), t \in \mathbb{R}^+\):

\[
m(t) - m(0) = \int_0^t r_f(m(s), \alpha(s))ds. \tag{7}
\]

This equation is equivalent to an ODE, but is easier to manipulate in integral form. In the rest of the paper, we make a slight abuse of language and refer to it as an ODE. Under the foregoing assumptions on \(f\) and \(\alpha\), this equation satisfies the Cauchy Lipschitz condition and has a unique solution once the initial condition \(m(0) = m_0\) is fixed. We call \(\phi_t, t \in \mathbb{R}^+,\) the corresponding semi-flow: the unique solution of Eq.(7) is

\[
m(t) = \phi_t(m_0, \alpha) \tag{8}
\]

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

\[
v_\alpha(m_0) = \int_0^T r_f(m_0, \alpha(s), \alpha(s))ds + r_f(\phi_T(m_0, \alpha)). \tag{9}
\]
This equation looks similar to the stochastic case (4) although there are two main differences. The first 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 value of the deterministic limit \( v_* (m_0) \):

\[
v_* (m_0) = \sup_{\alpha} v_{\alpha} (m_0),
\]

(10)

where the supremum is taken over all action functions from \([0; T] \to \mathcal{A}\).

E. Table of Notations

We remind here a list of the main notations used throughout the paper.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M^N_k )</td>
<td>Empirical measure of the system with ( N ) objects, under ( \pi ), at time ( k ). (Section II-B)</td>
</tr>
<tr>
<td>( F^N (m, a) )</td>
<td>Drift of the system with ( N ) objects when the state is ( m ) and the action is ( a ), Eq.(6)</td>
</tr>
<tr>
<td>( f (m, a) )</td>
<td>Drift of the limiting system (limit of rescaled ( F^N (m, a) ) as ( N \to \infty )), Eq.(11)</td>
</tr>
<tr>
<td>( \Phi_t (m_0, \alpha) )</td>
<td>State of the limiting system: ( \Phi_t (m_0, \alpha) = m_0 + \int_0^t f (\Phi_s (m_0, \alpha), s) ds ), Eq.(8)</td>
</tr>
<tr>
<td>( \pi^N )</td>
<td>Policy for the system with ( N ) objects: associates an action ( a \in \mathcal{A} ) to each ( k, M^N (k) )</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Action function for the limiting system: associates an action to each ( \alpha : [0; T] \to \mathcal{A} )</td>
</tr>
<tr>
<td>( \pi_* )</td>
<td>Optimal policy for the system with ( N ) objects</td>
</tr>
<tr>
<td>( \alpha_* )</td>
<td>Optimal action function for the limiting system (if it exists)</td>
</tr>
<tr>
<td>( V^N_{\pi} (m) )</td>
<td>Value of policy ( \pi ) (= expected gain) for the system with ( N ) objects starting from ( m ) under policy ( \pi ), Eq.(4)</td>
</tr>
<tr>
<td>( V^N_{\alpha} (m) )</td>
<td>Optimal value for the system ( N ): ( V^N_{\alpha} (m) = \sup_{\pi} V^N_{\pi} (m) = V^N (m) ), Eq.(5)</td>
</tr>
<tr>
<td>( V^N_0 (m) )</td>
<td>Value for the system ( N ) when applying the action function ( \alpha ), Eq.(12)</td>
</tr>
<tr>
<td>( v_\alpha (m) )</td>
<td>Value of the limiting system starting from ( m ) under action function ( \alpha ), Eq.(9)</td>
</tr>
<tr>
<td>( v_* (m) )</td>
<td>Optimal value of the limiting system: ( v_* (m) = \sup_{\alpha} v_{\alpha} (m) = v_{\alpha_*} (m) ), Eq.(10)</td>
</tr>
</tbody>
</table>

F. Summary of Assumptions

In Section III 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 \( \mathbb{R}^S \) and for \( m, m' \in \mathcal{P} (\mathcal{S}) \) we define \( \| m \| \) as the \( L^2 \)-norm of \( m \) and \( \langle m, m' \rangle = \sum_{s=1}^S m_s m'_s \) as the usual inner product.

a) (A1) (Transition probabilities): Objects can be observed only through their state, i.e., the transition probability matrix (or transition kernel) \( \Gamma^N \), defined by Eq.(2), is invariant under permutations of \( 1 \ldots N \).

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 \( \pi \), the number of objects that perform a transition between time slot \( k \) and \( k + 1 \) per time slot \( \Delta^N (k) \) satisfies

\[
\mathbb{E} \left( \Delta^N (k) | M^N (k) = m \right) \leq N I_1 (N)
\]

\[
\mathbb{E} \left( \Delta^N (k)^2 | M^N (k) = m \right) \leq N^2 I_1 (N) I_2 (N)
\]

(11)

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

b) (A2) (Convergence of the Drift): There exist some non-random functions \( f (m, a) \) 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

\[
\| f (m, a) - f (m', a) \| \leq K (\| m - m' \| + d (a, a'))
\]

\[
| r (m, a) - r (m', a) | \leq K_r | m - m' |
\]

\[
| r_f (m, a) - r_f (m', a) | \leq K_r | m - m' |
\]

We also assume that the reward is bounded: \( \sup_{m, a \in \mathcal{A}} \max (| r (m, a) |, | r_f (m) |) \leq M \). 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 \( \varphi \) is a continuous function on \( \Delta_S \times \mathcal{A} \times [0, \epsilon) \) for some neighborhood \( \Delta_S \) of \( \mathcal{P} (\mathcal{S}) \) and some \( \epsilon > 0 \), continuously differentiable with respect to \( m \).

In this case we 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 \( I_2 (N) = (c_1^2 + c_2^2)/N \).

III. MEAN FIELD CONVERGENCE

In Section III-A we establish the main results, then, in Section III-B, we provide the details of the method used to derive them.

A. Main Results

The first result establishes convergence of the optimization problem for the system with \( N \) objects to the optimization problem of the mean field limit:

**Theorem 2 (Optimal System Convergence).** Assume (A0) to (A3). If \( \lim_{N \to \infty} M^N (0) = m_0 \) almost surely [resp. in probability] then:

\[
\lim_{N \to \infty} V^N_* (M^N (0)) = v_* (m_0)
\]

almost surely [resp. in probability], where \( V^N_* \) and \( v_* \) are the optimal values for the system with \( N \) objects and the mean field limit, defined in Section II.
The proof is given in Section V-F.

The second result states that an optimal action function for the mean field limit provides an asymptotically optimal strategy for the system with \( N \) objects. We need, at this point, to introduce a first auxiliary system, which is a system with \( N \) objects controlled by an action function borrowed from the mean field limit. More precisely, let \( \alpha \) be an action function that specifies the action to be taken at time \( t \). Although \( \alpha \) has been defined for the limiting system, it can also be used in the system with \( N \) objects. In this case, the action function \( \alpha \) can be seen as a policy that does not depend on the state of the system. At step \( k \), the controller applies action \( \alpha(kI(\omega)) \). By abuse of notation, we denote by \( M_{N}^{\alpha} \), the state of the system when applying the action function \( \alpha \) (it will be clear from the notation whether the subscript is an action function or a policy). The value for this system is defined by

\[
V_{\alpha}^{N}(m_{0}) \overset{\text{def}}{=} \mathbb{E}\left( \sum_{k=0}^{H_{N}} r(M_{N}^{\alpha}(k), \alpha(kI(\omega))) + r_{f}(M_{N}^{\alpha}(H_{N})) \right| M_{N}^{\alpha}(0) = m_{0} \). \tag{12}
\]

Our next result is the convergence of \( M_{N}^{\alpha} \) and of the value:

**Theorem 3.** Assume (A0) to (A3); \( \alpha \) is a piecewise Lipschitz continuous action function on \([0; T]\), of constant \( K_{\alpha} \), and with at most \( p \) discontinuity points. Let \( M_{N}^{\alpha}(t) \) be the linear interpolation of the discrete time process \( M_{N}^{\alpha} \). Then for all \( \epsilon > 0 \):

\[
\mathbb{P}\left\{ \sup_{0 \leq t \leq T} \left| \tilde{M}_{N}^{\alpha}(t) - \phi_{\alpha}(m_{0}, \alpha) \right| > \left| M_{N}(0) - m_{0} \right| \right. \\
\left. + I_{0}(N, \alpha)T + \epsilon \left| e^{L_{1}T} \right| \right\} \leq \frac{J(N, T)}{\epsilon^{2}} \tag{13}
\]

and

\[
\left| V_{\alpha}^{N}(M_{N}(0)) - v_{\alpha}(m_{0}) \right| \leq B^{J}(N, \left| M_{N}(0) - m_{0} \right|) \tag{14}
\]

where \( J, I_{0} \) and \( B^{J} \) are defined in Section V-A and satisfy \( \lim_{N \to \infty} I_{0}(N, \alpha) = \lim_{N \to \infty} J(N, T) = 0 \) and \( \lim_{N \to \infty, \delta \to 0} B^{J}(N, \delta) = 0 \).

In particular, if \( \lim_{N \to \infty} \tilde{M}_{N}^{\alpha}(0) = m_{0} \) almost surely [resp. in probability] then \( \lim_{N \to \infty} V_{\alpha}^{N}(M_{N}(0)) = v_{\alpha}(m_{0}) \) almost surely [resp. in probability].

The proof is given in Section V-E.

As the reward function \( r(m, a) \) is bounded and the time-horizon \([0; T]\) is finite, the set of values of \( m \) starting from the initial condition \( m \), \( \{v_{\alpha}(m) : \alpha \text{ action function}\} \), is bounded. This set is not necessarily compact because the set of action functions may not be 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 \( \alpha^{\epsilon} \) such that \( v_{\alpha}(m) = \sup_{\alpha} v_{\alpha}(m) \leq v_{\alpha^{\epsilon}} + \epsilon \). Combining Theorem 2 and Theorem 3, we have that \( \lim_{N \to \infty} V_{\alpha^{\epsilon}}^{N} = v_{\alpha^{\epsilon}} \geq v_{\alpha} - \epsilon = \lim_{N \to \infty} V_{\alpha}^{N} - \epsilon \). This shows that \( \alpha^{\epsilon} \) is optimal up to \( 2\epsilon \) for \( N \) large enough. This shows the following corollary:

**Corollary 4 (Asymptotically Optimal Policy).** If \( \alpha_{*} \) is an optimal action function for the limiting system and if \( \lim_{N \to \infty} M_{N}^{\alpha}(0) = m_{0} \) almost surely [resp. in probability], then we have:

\[
\lim_{N \to \infty} \left| V_{\alpha_{*}}^{N} - V_{\alpha}^{N} \right| = 0,
\]

almost surely [resp. in probability].

In other words, an optimal action function for the limiting system is asymptotically optimal for the system with \( N \) objects.

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 for very small values of \( N \). In Section IV-B, we present two algorithms that can be used to build static or adaptive policies that are asymptotically optimal for the system with \( N \) objects.

**B. Derivation of Main Results**

1) Second Auxiliary System: The method of proof uses a second auxiliary system, the process \( \phi_{\alpha}(m_{0}, A_{\pi}^{N}) \) defined below. It is a limiting system controlled by an action function derived from the policy of the original system with \( N \) objects.

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

Recall that for any \( m_{0} \in \mathcal{P}(\mathcal{S}) \) and any action function \( \alpha \), \( \phi_{\alpha}(m_{0}, \alpha) \) is the solution of the ODE (7). For every \( \omega \), \( \phi_{\alpha}(m_{0}, A_{\pi}^{N}(\omega)) \) is the solution of the limiting system with action function \( A_{\pi}^{N}(\omega) \), i.e.

\[
\phi_{\alpha}(m_{0}, A_{\pi}^{N}(\omega)) = m_{0} + \int_{0}^{t} f(\phi_{\alpha}(m_{0}, A_{\pi}^{N}(\omega)), A_{\pi}^{N}(\omega)(s)) \, ds.
\]

When \( \omega \) is fixed, \( \phi_{\alpha}(m_{0}, A_{\pi}^{N}(\omega)) \) is a continuous time deterministic process corresponding to one trajectory \( M_{\pi}^{N}(\omega) \). When considering all possible realizations of \( M_{\pi}^{N} \), \( \phi_{\alpha}(m_{0}, A_{\pi}^{N}) \) is a random, continuous time function "coupled" to \( M_{\pi}^{N} \). Its randomness comes only from the action term \( A_{\pi}^{N} \), in the ODE. In the following, we omit to write the dependence in \( \omega \). \( A_{\pi}^{N} \) and \( M_{\pi}^{N} \) will always designate the processes corresponding to the same \( \omega \).

2) Convergence of Controlled System: 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 \( \pi \).

**Theorem 5.** Under Assumptions (A0) to (A3), for any \( \epsilon > 0 \) and any policy \( \pi \):

\[
\mathbb{P}\left\{ \sup_{0 \leq t \leq T} \left| \tilde{M}_{\pi}^{N}(t) - \phi_{\pi}(m_{0}, A_{\pi}^{N}) \right| > \left| M_{N}^{\alpha}(0) - m_{0} \right| + I_{0}(N,T) + \epsilon \left| e^{L_{1}T} \right| \right\} \leq \frac{J(N,T)}{\epsilon^{2}} \tag{15}
\]
where $\hat{M}_N$ is the linear interpolation of the discrete time system with $N$ objects) and $J$ is defined in Section V-A.

Recall that $I_0(N)$ and $J(N,T)$ for a fixed $T$ go to 0 as $N \to \infty$. The proof is given in Section V-C.

3) Convergence of Value: Let $\pi$ be a policy and $A_\pi$ the sequence of actions corresponding to a trajectory $M_\pi$ as we just defined. Eq.(9) defines the value for the deterministic limit when applying a sequence of actions. This defines a random variable $v_{A_\pi}(m_0)$ that corresponds to the value of the limit system when using $A_\pi$ as action function. The random part comes from $A_\pi$, $\mathbb{E}[v_{A_\pi}(m_0)]$ designates the expectation of this value over all possible $A_\pi$. A first consequence of Theorem 5 is the convergence of $V_\pi^N(M_N(0))$ to $\mathbb{E}[v_{A_\pi}(m_0)]$ with an error that can be uniformly bounded.

**Theorem 6** (Uniform convergence of the value). Let $A_\pi$ be the random action function associated with $M_\pi$, as defined earlier. Under Assumptions (A0) to (A3),

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

where $B$ is defined in Section V-A.

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

The proof is given in Section V-D.

4) Putting Things Together: The proof of the main result uses the two auxiliary systems. The first auxiliary system provides a strategy for the system with $N$ objects derived from an action function of the mean field limit; it cannot do better than the optimal value for the system with $N$ objects, and is close to the optimal value of the mean field limit. Therefore, the optimal value for the system with $N$ objects is lower bounded by the optimal value for the mean field limit. The second auxiliary system is used in the opposite direction, which shows that, roughly speaking, for large $N$ the two optimal values are the same. We give the details of the derivation in Section V-F.

IV. APPLICATIONS

A. Hamilton-Jacobi-Bellman Equation and Dynamic Programming

Let us now consider the finite time optimization problem for the stochastic system and its limit from a constructive point of view. As the state space is finite, we can compute the optimal value by using a dynamic programming algorithm. If $U^N(m,t)$ denotes the optimal value for the stochastic system starting from $m$ at time $t/I(N)$, then $U^N(m,t) = \sup_\pi \mathbb{E}\left[\sum_{k=t/I(N)}^{T/I(N)} r^N(M^N(k)) : M^N(t) = m\right]$. The optimal value can be computed by a discrete dynamic programming algorithm [21] by setting $U^N(m,T) = r_f(m)$ and

$$U^N(m,t) = \sup_{a \in A} \mathbb{E}\left[r^N(m,a) + U^N(M^N(t+1/I(N)),t+1/I(N)) : M^N(t) = m, A^N(t) = a\right].$$

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

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:

$$\frac{\partial u(m,t)}{\partial t} + \max_a \{\nabla u(m,t).f(m,a) + r(m,a)\} = 0. \quad (16)$$

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

B. Algorithms

Corollary 4 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. Corollary 4 shows that the value $V_\pi^N$ of the policy $\pi$ obtained by Algorithm 1 is asymptotically optimal:

$$\lim_{N \to \infty} V_\pi^N(M_N(0)) = \lim_{N \to \infty} \inf V_\pi^N(M_N(0)).$$

**Algorithm 1**: Algorithm constructing a static policy for the system with $N$ objects, over the finite horizon $H$.

begin
From the original system with $N$ objects, construct the occupancy measure $M_N$ and its kernel $\Gamma^N$ and let $M_N(0)$ be the initial occupancy measure; Compute the limit $f$ of the drift of $\Gamma^N$; Solve the HJB equation (16) on $[0,HI(N)]$. This provides an optimal control function $\alpha^\pi(M^N_0, t)$; Construct a discrete control $\pi$ for the discrete system: the action to be taken under state $M_N(k)$ at step $k$ is

$$\pi(M_N(k), k) \equiv \alpha^\pi(\phi_{k/I(N)}(M_N(0), \alpha)).$$

return $\pi$;

The policy $\pi$ 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 $\pi'$ from 0 to the final horizon $H$ is given in Algorithm 2.

In practice, the total value of the adaptive policy $\pi'$ should be larger than the value of the static policy $\pi$ because it uses on-line corrections at each step, before taking a new action. However Theorem 2 does not provide a proof of its asymptotic optimality.

C. Examples

In this section, we develop three examples. The first one can be seen as a simple illustration of optimal mean field. The
An equivalent model is that at every time step (which size an HJB equation must be solved over a much smaller
size) the utility loses customers, which eventually reduces the gain.

The purpose of this last example is to show that a discrete
Markovian system made of
utility fixes their price
and
utility
decreases as
the
probability of a new subscription, and
the
values
of
the
utility
is
a
proportion
of
subscribers
at
time
The
Hamilton-Jacobi-
Bellman
equation
is
\[ \frac{\partial}{\partial t} u_*(t, x) + H \left( x, \frac{\partial}{\partial x} u_*(t, x) \right) = 0 \]
with
\[ H(x, p) = \max_{\alpha \in [0, 1]} \left[ p(s(\alpha) - x(s(\alpha) + a(\alpha)) + \alpha x) \right]. \]
\[ \frac{dx}{dt} = -x(t)\alpha(t) + (1 - x(t))(1 - \alpha(t)) = 1 - x(t) - \alpha(t), \]
and
\[ H(x, p) = \max (x(1 - p), (1 - x)p). \]
The solution of the HJB equation can be given in closed form. The optimal policy is to choose action \( \alpha = 1 \) if \( x > 1/2 \) or \( x > 1 - \exp(-(T-t)) \), and 0 otherwise. Figure 1 shows the evolution of the proportion of subscribers \( x(t) \) when the optimal policy is used. The colored area corresponds to all the points \((t, x)\) where the optimal policy is \( \alpha = 1 \) (fix a high price) and the white area is where the optimal policy is to choose \( \alpha = 0 \) (low price).

To show that this policy is indeed optimal, one has to compute the corresponding value of the benefit \( u(t, x) \) and show that it satisfies the HJB equation. This can be done using a case analysis, by computing explicitly the value of \( u(t, x) \) in the zones \( Z_1, Z_2, Z_3 \) and \( Z_4 \) displayed in Figure 1, and check that \( u(t, x) \) satisfies Eq.(18) in each case.
2) Infection Strategy of a Viral Worm: This second example has two purposes. The first one is to provide a rigorous justification of the use of a continuous optimization approach for this classic problem in population dynamics and to show that the continuous limit provides insights on the structure of the optimal behavior for the discrete system. Here, the optimal action function can be shown to be of the bang-bang type for the limit problem, by using tools from continuous optimization such as the Pontryagin maximum principle. Theorem 2 shows that a bang-bang policy should also be asymptotically optimal in the discrete case.

The second purpose is to compare numerically the performance of the optimal policy of the deterministic limit \( \alpha_* \) and the performance of other policies for the stochastic system for small values of \( N \). We show that \( \alpha_* \) is close to optimal even for \( N = 10 \) and that it outperforms another classic heuristic.

This example is taken from [15] and considers the propagation of infection by a viral worm. Actually, similar epidemic models have been validated through experiments, as well as simulations as a realistic representation of the spread of a virus in mobile wireless networks (see [7, 22]). A susceptible node is a mobile wireless device, not contaminated by the worm but prone to infection. A node is infective if it is contaminated by the worm. An infective node spreads the worm to a susceptible node whenever they meet, with probability \( \beta \). The worm can also choose to kill an infective node, i.e., render it completely dysfunctional - such nodes are denoted dead. A functional node that is immune to the worm is referred to as recovered. The goal of the worm is to maximize the damages done to the network by choosing the rate \( \alpha(t) \) at which it kills node at time \( t \). Let the total number of nodes in the network be \( N \). Let the proportion of susceptible, infective, recovered and dead nodes at time \( t \) be denoted by \( S(t) \), \( I(t) \), \( R(t) \) and \( D(t) \), respectively. Under a uniform mobility model, the probability that a susceptible node becomes infected is \( \beta I/N \). The immunization of susceptibles (resp. infectives) happens at a fixed rate \( q \) (resp. \( b \)). This means that a susceptible (resp. infective) node is immunized with probability \( q/N \) (resp. \( b/N \)) at every time step. The probability for a node to be killed by the worm is \( \alpha/N \) where \( \alpha \in [0; \alpha_{\text{max}}] \) is the action taken by the worm.

The damages are function of the number of dead nodes at the end of the time-horizon \( D(T) \) and of instantaneous damages during the time horizon that depend on the proportion of infected nodes, \( f(I(t)) \). The controller wants to maximize the expectation of the sum of these two damage functions:

\[
\mathbb{E}\left(D_s(T) + \frac{1}{NT} \sum_{k=1}^{NT} f(I_s(k))\right).
\]

At this point, authors of [15] invoke the classic results of Kurtz [17] to show that the dynamics of this population process converges to the solution of the following differential equations,

\[
\begin{align*}
\frac{dS}{dt} &= -\beta IS - qS \\
\frac{dI}{dt} &= \beta IS - bI - \alpha(t)I \\
\frac{dR}{dt} &= \alpha(t)I \\
\frac{dD}{dt} &= bI + qS,
\end{align*}
\]

where \( \alpha(t) \) is the action taken by the worm at time \( t \).

This system actually satisfies assumptions \((A_1, A_2, A_3)\), which allows us not only to obtain the mean field limit, but also to say more about the optimization problem. In the continuous control problem, the objective of the worm is to find an action function \( \alpha \) such that the damage function \( D(T) + \frac{1}{T} \int_0^T f(I(t))dt \) is maximized under the constraint \( 0 \leq \alpha(t) \leq \alpha_{\text{max}} \) (where \( f \) is convex). In [15], this problem is shown to have a solution and the Pontryagin maximum principle is used to show that the optimal action function \( \alpha_* \) is of bang-bang type: there exists \( t_1 \in [0 \ldots T] \) s.t.

\[
\alpha_*(t) = \begin{cases} 
0 & \text{for } 0 < t < t_1 \\
\alpha_{\text{max}} & \text{for } t_1 < t < T
\end{cases}
\]  

(21)

Theorem 2 makes the formal link between the optimization of the model on an individual level and the previous resolution of the optimization problem on the differential equations, done in [15]. It allows us to formally claim that the policy \( \alpha_* \) of the worm is indeed asymptotically optimal when the number of objects goes to infinity.

We investigated numerically the performance of \( \alpha_* \) against various infection policies for small values of the number of nodes in the system \( N \). These results are reported on Figure 2, where we compare four values:

- \( v_s \) - the optimal value of the limiting system;
- \( V_*^N \) - the optimal expected damage for the system with \( N \) objects (MDP problem);
- \( V_{\alpha_*}^N \) - the expected value of the system with \( N \) objects when applying the action function \( \alpha_* \) that is optimal for the limiting system;
- the performance of a heuristic where, instead of choosing a threshold as suggested by the limiting system (21), the killing probability \( \alpha(t) \) is fixed to some \( \nu \) for the whole time. The curve on the figure is drawn for the optimal \( \nu \) (recomputed for each parameter \( N \)).

We implemented a simulator that follows strictly the model of infection described earlier in this part. We chose parameters similar to those used in [15]: the parameter for the evolution of the system are \( \beta = 0.6 \), \( q = .1 \), \( b = .1 \), \( \alpha_{\text{max}} = 1 \) and the damage function to be optimized is \( D(T) + \frac{1}{T} \int_0^T I^2(t)dt \) with \( T = 10 \). The initial state is \( S(0) = .7 \), \( I(0) = .3 \), \( D(0) = R(0) = 0 \). However, it should be noted that the choice of these parameters does not qualitatively influence the results. Thanks to the relatively small size of the system, these four quantities can be computed numerically using a backward induction. The optimal policies for the deterministic limit consists in not killing machines until \( t_1 = 4.9 \) and in killing machines at a maximum rate after that time: \( \alpha_*(t) = 1_{\{t \geq 4.9\}} \).

Theorem 2 shows that \( \alpha_* \) is asymptotically optimal (\( \lim_{N \to \infty} V_*^N = \lim_{N \to \infty} V_{\alpha_*}^N = v_* \)), but Figure 2(a) shows that, already for low values of \( N \), these three quantities are very close. A classic heuristic for this maximal infection problem is to kill a node with a constant probability \( \alpha(t) = \nu \) regardless of the time horizon. Our numerical study shows that \( \alpha_* \) outperforms this heuristic by more than 20%. The performance of this heuristic does not increase with the size of the system \( N \).

In order to illustrate the convergence of the values \( V_*^N \) and \( V_{\alpha_*}^N \) to \( v_* \), Figure 2(b) is a detailed view of Figure 2(a) where...
When they do not use their computer, it is available for the volunteer computing system such as BOINC http://boinc.berkeley.edu/. Volunteer computing means that people make their personal computer available for a computing system. When they do not use their computer, it is available for the computing system. However, as soon as they start using their computer, it becomes unavailable for the computing system. These systems are becoming more and more popular and provide large computing power at a very low cost [16].

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

The model is a discrete time model of a queueing system. Actually, a more natural continuous-time Markov model could also be handled similarly, by using uniformization.

There are $U^N$ users. Each user has a state $x \in \{\text{on, off}\}$. At each time step, an active user sends one job with probability $p^N_1$ and becomes inactive with probability $p^N_i$. An inactive user sends no jobs to the system and becomes active with probability $p^N_a$.

There are $C$ clusters in the system. Each cluster $c$ contains $Q^N_c$ computing resources. Each resource has a buffer of bounded size $j_c$. A resource can either be valid or broken. If it is valid and if it has one or more job in its queue, it completes one job with probability $\mu_c/N$ at this time slot. A resource gets broken with probability $p^N_b/N$. In that case, it discards all the packets of its buffer. A broken resource becomes valid with probability $p^N_v/N$.

At each time step, the broker takes an action $\alpha \in \mathcal{A}$ and sends the packets it received to the clusters according to the distribution $\alpha$. A packet sent to cluster $c$ joins the queue of one of the resources, $\nu_c$; according to a local rule (for example chosen uniformly among the $Q^N_c$ resources composing the cluster). If the queue of resource $\nu$ is full, the packet is lost. The goal of the broker is to minimize the number of losses plus the total size of the queues over a finite horizon (and hence the response time of accepted packets).

This model is represented in Figure 3.

Figure 2. Damage caused by the worm for various infection policies as a function of the size of the system $N$. The goal of the worm is to maximize the damage (higher means better). Panel (a) shows the optimal value $v_*$ for the limiting system (mean field limit), the optimal value $v^N_*$ for the system with $N$ objects, the value $v^N_*$ of the asymptotically optimal policy given in Corollary 4 and the value of a classic heuristic. Panel (b) zooms the $y$-axis around the values of the optimal policies.

Figure 3. The brokering problem in a desktop grid system, such as Boinc
The system has an intensity $I(N) \overset{\text{def}}{=} 1/N$. The number $C$ of clusters is fixed and does not depend on $N$, as well as the sizes $J_c$ of the buffers. However, both the number of users $U^N$, and the number of resources in the clusters $Q^N_c$, are linear in $N$. Finally, by construction, all the state changes occur with probabilities that scale with $1/N$.

The limiting system is described by the variable $m_c(t)$, which represents the fraction of users who are on, and the variables $q_{c,j}(t)$ and $b_c(t)$ that, respectively, represent the fraction of resources in cluster $c$ having $j$ jobs in their buffer and the fraction of resources in cluster $c$ that are broken. For an action function $\alpha(\cdot)$, we denote by $\alpha_c(\cdot)$ the fraction of packets sent to cluster $c$. Finally, $m_c$ denotes the fraction of users (both active or inactive) and $q_{c,J}$ the fraction of processors in cluster $c$. These fractions are constant (independent of time) and satisfy $m + q_1 + \cdots + q_{c,J} = 1$. We get the following equations:

\begin{align}
\frac{dm_c}{dt} &= -p_a m_c(t) + p_a (m - m_c) \\
\frac{dq_{c,j}}{dt} &= p_b b_c - \alpha_c(t) p_a m_c q_{c,j} + \mu_c q_{c,j-1} - p_b q_{c,j} \\
\frac{dq_{c,J}}{dt} &= \alpha_c(t) p_a m_c q_{c,J-1} - \mu_c q_{c,J} - p_b q_{c,J} \\
\frac{db_c}{dt} &= -p_a b_c(t) + p_a \sum_{j=0}^{J_c} q_{c,j}. \tag{26}
\end{align}

where (23) and (25) hold for each cluster $c$ and (24) holds for each cluster $c$ and for all $j \leq J_c$. The cost associated to the action function $\alpha$: \(I(0,\alpha) \overset{\text{def}}{=}-T \left[ \sum_{c=1}^{C} \sum_{j=1}^{J_c} j q_{c,j}(t) + \gamma (\sum_{c=1}^{C} \sum_{j=1}^{J_c} q_{c,j-1}(t)) \right. \)

\begin{align}
\left. + \sum_{c=1}^{C} \alpha_c(t) p_a m_c(q_{c,J-1}(t) + b_c(t))) \right] \right] dt \tag{27}
\end{align}

The first part of (27) represents the cost induced by the number of jobs in the system. The second part of (27) represents the cost induced by the losses. The parameter $\gamma$ gives weight on the cost induced by the losses.

The HJB problem becomes minimizing (27) subject to the variables $u_a, q_{k,i}, b_k$ satisfying Equations (22) to (26). This system is made of $(C + 2)J_c$ ODEs. Solving the HJB equation numerically in this case can be challenging but remains more tractable than solving the original Bellman equation over $J^N$ states. The curse of dimensionality is so acute for the discrete system that it cannot be solved numerically with more than 10 processors [5].

### V. PROOFS

#### A. Details of Scaling Constants

$I_0(N,\alpha) \overset{\text{def}}{=} I_0(N) + I(N)K e^{(K-L_t)T}$

\begin{align}
&\left( \frac{K_T}{2} + 2 (1 + \min(1/I(N), p)) \|\alpha\|_\infty \right)
\end{align}

\begin{align}
J(N,T) \overset{\text{def}}{=} 8T \left[ L_1 \left[ I_2(N) I(N)^2 + I_1(N)^2 (T + I(N)) \right] \right. \\
&+ S^2 \left[ 2I_2(N) + I(N) (I_0(N) + L_2)^2 \right] \left. \right] \\
B(N,\delta) \overset{\text{def}}{=} I(N) \|\alpha\|_\infty \\
&+ K_\epsilon \sqrt{2} I_1(N) \\
&+ K_\epsilon \left( \delta + I_0(N)T \right) \left( e^{L_t T} + \frac{e^{L_t T} - 1}{L_t} \right) \\
&+ \frac{3 S^2}{2 T} \left[ e^{L_t T} + \frac{e^{L_t T} - 1 + (I(N)/2)^2}{L_t} \right] \\
&\cdot K_\epsilon^2 \left( \|\alpha\|_\infty \right) J(N,T) \left( (T + 1)^2 \right)
\end{align}

$B'(N,\delta)$ has the same expression as $B(N,\delta)$ replacing $I_0(N)$ by $I_0(N,\alpha)$ in the second line.

#### B. Proof of Theorem 1

We begin with a few general statements. Let $P$ be the set of probabilities on $S$ and $\mu^N : S^N \to P$ defined by $\mu^N(x) = \sum_{c=1}^{C} \sum_{j=1}^{J_c} x_{c,j}$ for all $x \in S$. Also let $P^N$ be the image set of $\mu^N$, i.e. the set of all occupancy measures that are possible when the number of objects is $N$. The following establishes that if two global states have the same occupancy measure, then they differ by a permutation.

Lemma 1. For all $x, x' \in S^N$, if $\mu^N(x) = \mu^N(x')$ there exists some $\sigma \in S^N$ such that $x' = \sigma(x)$.

**Proof:** By induction on $N$. Its is obvious for $N = 1$. Assume the lemma holds for $N - 1$ and let $x, x' \in S^N$, with $\mu^N(x) = \mu^N(x')$. There is at least one coordinate, say $i$, such that $x'_i = x_i$, because there is the same number of occurrences of $s = x_1$ in both $x$ and $x'$. Let $y = x_2 \ldots x_N$ and $y' = x'_2 \ldots x'_{i-1} x_{i+1} \ldots x_N$. Then $\mu^{N-1}(y) = \mu^{N-1}(y')$, therefore there exists some $\tau \in S^{N-1}$ such that $y' = \tau(y)$. Define $\sigma$ by $\sigma(1) = i, \sigma(j) = \tau(j) + 1$ for $j \geq 2$, so that $x' = \sigma(x)$. Clearly $\sigma$ is a permutation of $\{1, \ldots, N\}$. Let $f : S^N \to E$ where $E$ is some arbitrary set. We say that $f$ is invariant under $\mathcal{S}^N$ if $f \circ \sigma = f$ for all $\sigma \in \mathcal{S}^N$. The following results states that if a function of the global state is invariant under permutations, it is a function of the occupancy measure.

Lemma 2. If $f : S^N \to E$ is invariant under $\mathcal{S}$ then there exists $\widetilde{f} : P^N \to E$ such that $\widetilde{f} \circ \mu^N = f$.

**Proof:** Define $\widetilde{f}$ as follows. For every $m \in P^N$ pick some arbitrary $x_0 \in (\mu^N)^{-1}(m)$ and let $\tilde{f}(m) = f(x_0)$. Now let $x$, perhaps different from $x_0$, such that $\mu^N(x) = m$. By Lemma 1, there exists some $\sigma \in S^N$ such that $x = \sigma(x_0)$ therefore $f(x) = f(x_0) = \tilde{f}(\mu^N(x))$. This is true for every $m \in P^N$ thus $f(x) = \tilde{f}(\mu^N(x))$ for every $x \in S^N$.

The sequence of actions $a_k$ is given and $N$ is fixed. We are thus given a time-inhomogeneous Markov chain $X^N$ on $S^N$, with transition kernel $G_k, k \in \mathbb{N}$, given by $G_k(x,y) = \Gamma^N(x,y,a_k)$, such that for any permutation $\sigma \in S^N$ and any states $x, y$ we have

\begin{align}
G_k(\sigma(x),\sigma(y)) = G_k(x,y)
\end{align}

Let $\mathcal{F}(k)$ be the $\sigma -$ field generated by $X^N(s)$ for $s \leq k$ and $\mathcal{G}(k)$ be the $\sigma-$ field generated by $M^N(s)$ for $s \leq k$. Note that because $M^N = \mu^N \circ X^N, \mathcal{G}(k) \subset \mathcal{F}(k)$. 

\begin{align}
G_k(\sigma(x),\sigma(y)) = G_k(x,y)
\end{align}

Let $\mathcal{F}(k)$ be the $\sigma -$ field generated by $X^N(s)$ for $s \leq k$ and $\mathcal{G}(k)$ be the $\sigma-$ field generated by $M^N(s)$ for $s \leq k$. Note that because $M^N = \mu^N \circ X^N, \mathcal{G}(k) \subset \mathcal{F}(k)$. 

\begin{align}
G_k(\sigma(x),\sigma(y)) = G_k(x,y)
\end{align}

Let $\mathcal{F}(k)$ be the $\sigma -$ field generated by $X^N(s)$ for $s \leq k$ and $\mathcal{G}(k)$ be the $\sigma-$ field generated by $M^N(s)$ for $s \leq k$. Note that because $M^N = \mu^N \circ X^N, \mathcal{G}(k) \subset \mathcal{F}(k)$.
Pick some arbitrary test function \( \psi : S^N \to \mathbb{R} \) and fix some time \( k \geq 1 \); we will now compute \( \mathbb{E} \left( \psi(M^N_{k(I)}(k)) | F(k-1) \right) \). Because \( M^N_{k} \) is a function of \( X^N \) and \( X^N \) is a Markov chain, \( \mathbb{E} \left( \psi(M^N_{k}) | F(k-1) \right) \) is a function, say \( \psi \), of \( X^N(k-1) \). We have, for any fixed \( \bar{M} \) (Lemma 4) and the second part \( \psi \) integral formula.

Because \( [\pi \bar{M}] \) is bounded using an \( N \geq 1 \) and \( \hat{\Psi}_{M_{k}(k)}(\bar{M}(k)) \) is the drift at \( \bar{M}(k) \) and \( f_{\hat{\Psi}_{M_{k}(k)}}(\bar{M}(k)) \) is the integral formula. It follows from these definitions that \( M^N_{k}(t) \) is equal to

\[
M^N_{k}(0) + \int_0^t \frac{ds}{I(N)} F^N \left( \bar{M}^N_{k}(s), A^N_{\bar{M}}(s) \right) + \hat{Z}^N_{M_{k}}(t) = M^N_{k}(0) + \int_0^t \frac{ds}{I(N)} F^N \left( \bar{M}^N_{k}(s), A^N_{\bar{M}}(s) \right) + \hat{Z}^N_{M_{k}}(t) + \int_0^t \frac{ds}{I(N)} \left[ F^N \left( \bar{M}^N_{k}(s), A^N_{\bar{M}}(s) \right) - F^N \left( \bar{M}^N_{k}(s), A^N_{\bar{M}}(s) \right) \right]
\]

Applying Assumption (A2) to the third line, (A3) to the second and fourth lines, and Equation (31) to the fourth line leads to:

\[
\left\| \hat{M}^N_{k}(t) - \phi_{M_{k}}(m_{0}, A_{\bar{M}}) \right\| \leq \left\| M^N_{k}(0) - m_{0} \right\| + \left\| \hat{Z}^N_{M_{k}}(t) \right\| + L_{1} \int_0^t \left\| \hat{M}^N_{k}(s) - \phi_{M_{k}}(m_{0}, A_{\bar{M}}) \right\| ds + \int_0^t \left( I(N) \right)^{\lfloor \frac{t}{I(N)} \rfloor} \sum_{k=0} W_{k+1}
\]

For all \( N, \pi, T, b_1 > 0 \) and \( b_2 > 0 \), define

\[
\Omega_1 = \left\{ \omega \in \Omega : \sup_{0 \leq k \leq T} \sum_{j=0}^{k} W_{k+1} > b_1 \right\},
\]

\[
\Omega_2 = \left\{ \omega \in \Omega : \sup_{0 \leq k \leq T} \left\| \hat{Z}^N_{M_{k}}(k) \right\| > b_2 \right\}.
\]

Let \( (\mathcal{F}_k)_{k \in \mathbb{N}} \) denotes the natural filtration associated with the process. Since \( M(k) \) is Markovian and \( W(k) + 1 \) depends only on \( \bar{M}(k) \) and \( \bar{M}(k+1) \), we have \( \mathbb{E} \left( \bar{W}(k+1) | \mathcal{F}_k \right) = \mathbb{E} \left( \bar{W}(k+1) | \bar{M}(k) \right) \). Moreover, since \( \Delta^N_{\bar{M}}(k) \) is the number of objects that change state during one time step, we have \( \bar{W}(k+1) \leq \sqrt{2N} \Delta^N_{\bar{M}}(k) \). Using Assumption (A1), this shows that \( \mathbb{E} \left( \bar{W}(k+1) | \mathcal{F}_k \right) \leq \sqrt{2I(N)} \) and \( \mathbb{E} \left( \bar{W}(k+1) | \mathcal{F}_k \right) \leq \frac{1}{2}(N)I(N) \). Therefore applying Lemma 3 with \( n = T/I(N) \) shows that for any \( b_1 > 0 \):

\[
\mathbb{P} \left( \Omega_1 \right) \leq \frac{2T}{b_1} \left( I(N) + \frac{I(N)^2}{I(N)} \right) (T + I(N))
\]

Moreover, we show in Lemma 4 that:

\[
\mathbb{P} \left( \Omega_2 \right) \leq \frac{2S}{b_2} \left[ I(N) + I(N) \left( I_0(N) + L_2 \right) \right]^2
\]
Now fix some \( \epsilon > 0 \) and let \( b_1 = \frac{\epsilon}{\varepsilon_1 I(N)}, \) \( b_2 = \epsilon/2 \). For \( \omega \in \Omega \setminus (\Omega_1 \cup \Omega_2) \) and for \( 0 \leq t \leq T \):

\[
\left\| M_{\pi}^N(t) - \phi_t(m_0, A_{\pi}^N) \right\| \leq \left\| M_{\pi}^N(0) - m_0 \right\| + \epsilon + I_0(N)T \\
+ L_1 \int_0^t \left\| M_{\pi}^N(s) - \phi_s(m_0, A_{\pi}^N) \right\| ds
\]

By Grönwall’s lemma: \( \left\| M_{\pi}^N(t) - \phi_t(m_0, A_{\pi}^N) \right\| \) is less than \( \left\| M_{\pi}^N(0) - m_0 \right\| + \epsilon + I_0(N)T \) \( e^{L_1 t} \) \( (35) \)

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.(33) and Eq.(34), concludes the proof.

The proof of Theorem 5 uses the following lemmas.

**Lemma 3.** Let \((W_k)_{k \in \mathbb{N}}\) be a sequence of square integrable, non-negative random variables, adapted to a filtration \((\mathcal{F}_k)_{k \in \mathbb{N}}\), such that \(W_0 = 0\) a.s. and for all \(k \in \mathbb{N}\): \(\mathbb{E}(W_{k+1} | \mathcal{F}_k) \leq \alpha \) and \(\mathbb{E}(W_{k+2} | \mathcal{F}_k) \leq \beta\). Then for all \(n \in \mathbb{N}\) and \(b > 0\):

\[
\mathbb{P}\left( \sup_{0 \leq k \leq n} (W_0 + \ldots + W_k) > b \right) \leq \frac{n\beta + n(n + 1)\alpha^2}{b^2}
\]

**Proof:** Let \(Y_n = \sum_{k=0}^n W_k\). It follows that \(\mathbb{E}(Y_n) \leq n\alpha \) and \(\mathbb{E}(Y_{n+1}) \leq \beta + 2n\alpha^2 + \mathbb{E}(Y_n^2)\) from where we derive that \(\mathbb{E}(Y_n^2) \leq n\beta + n(n + 1)\alpha^2\) \( (36) \)

Now, because \(W_{n+1} \geq 0\):

\[
\mathbb{E}(Y_{n+1}^2 | \mathcal{F}_n) \geq (\mathbb{E}(Y_{n+1} | \mathcal{F}_n))^2 \\
=(Y_n + \mathbb{E}(W_{n+1} | \mathcal{F}_n))^2 \geq Y_n^2
\]

thus \(Y_n^2\) is a non-negative sub-martingale and by Kolmogorov’s inequality:

\[
P\left( \sup_{0 \leq k \leq n} Y_k > b \right) = P \left( \sup_{0 \leq k \leq n} Y_k^2 > b^2 \right) \leq \frac{\mathbb{E}(Y_n^2)}{b^2}
\]

Together with Eq.(36) this concludes the proof.

**Lemma 4.** Define \(Z_{\pi}^N\) as in Eq.(32). For all \(N \geq 2\), \(b > 0\), \(T > 0\) and all policy \(\pi\):

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

**Proof:** The proof is inspired by the methods in [1]. For fixed \(N\) and \(h \in \mathbb{R}^2\), let

\[
L_k \overset{\text{def}}{=} \langle h, Z_{\pi}^N(k) \rangle.
\]

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

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

\[+ \langle h, F^N(\pi^N, \pi_k(M_{\pi}^N(k))) \rangle^2 \]

By Assumption (A2):

\[
|\langle h, F^N(\pi^N, \pi_k(M_{\pi}^N(k))) \rangle| \leq (I_0(N) + L_2) I(N) \|h\|
\]

Thus, using that \(\mathbb{E}(\|M_{\pi}^N(k+1) - M_{\pi}^N(k)\|^2 | \mathcal{F}_k)\) is equal to \(\mathbb{E}(\|M_{\pi}^N(k+1) - M_{\pi}^N(k)\|^2 | M_{\pi}^N(k))\) which is less than \(N^{-2}2\mathbb{E}(\Delta^N_{\pi}(k)^2 | M_{\pi}^N(k))\) and Assumption (A1), \(\mathbb{E}(L_{k+1} - L_k) \) \( \mathcal{F}_k \) can be bounded by:

\[
\|h\|^2 \left[ N^{-2}2\mathbb{E}(\Delta^N_{\pi}(k)^2 | M_{\pi}^N(k)) + (I_0(N) + L_2) I(N)^2 \right] \]

\[
\leq b^2 \left[ 2I_2(N) + (I_0(N) + L_2) I(N)^2 \right]
\]

Applying Kolmogorov’s inequality for martingales, we obtain

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

which, by changing \(b \) to \(\sqrt{b} \), shows the result.

**D. Proof of Theorem 6**

We use the same notation as in the proof of Theorem 5. By definition of \(V_{\pi}^N\) and \(v\), (Eq. 4 and 9) and the fact that \(r_{\pi}^N(m, a) = I(N)r(m, a)\), we have:

\[
V_{\pi}^N(M^N(0)) - \mathbb{E}(v_{A^N}(m_0)) = \mathbb{E}\left( \int_0^{H^N} r_{\pi}(s) ds \right)
\]

\[+ \mathbb{E}(r_f(M_{\pi}^N(H^N)) - r_f(m_{A^N}(H^N I(N)))) + r_f(m_{A^N}(H^N I(N))) - r_f(m_{A^N}(T))
\]

\[+ \mathbb{E}\left( \int_0^{H^N} r_{f}(m_{A^N}(s), A_{\pi}^N(s)) ds \right)
\]

Since \(H^N = \lfloor T/N \rfloor\), the latter term is bounded by \(I(N)\|p\|\). Moreover, by assumption (A1) on \(\Delta^N_{\pi}(H^N)\), we have \(\left\| r_f(m_{A^N}(H^N I(N))) - r_f(m_{A^N}(T)) \right\| \leq K_r \left\| m_{A^N}(H^N I(N)) - m_{A^N}(T) \right\| \leq K_r \sqrt{2I_1(N)}\).
Let $\epsilon > 0$ and $\Omega_0 = \Omega_1 \cup \Omega_2$ where $\Omega_1, \Omega_2$ are as in the proof of Theorem 5. Thus $P(\Omega_0) \leq \frac{J(N,T)}{T^{2}}$ and, using, the Lipschitz continuity of $r$ in $m$ (with constant $K_r$):

$$
|V_\pi^N(M^N(0)) - E[v_A^N(m_0)]| \\
\leq I(N) \|r\|_\infty + K_r\sqrt{2}I_1(N) + \frac{2(T+1)\|r\|_\infty J(N,T)}{\epsilon^2} \\
+ K_r E\left(\int_0^T \|M^N_\pi(s) - m_A^N(s)\| ds \\
+ r_f(M^N_\pi(H_N)) - r_f(m(H^N(I(N))))\right).
$$

For $\omega \notin \Omega_0$: $\int_0^T \|M^N_\pi(s) - m_A^N(s)\| ds \leq \frac{\epsilon I(N)}{2L_1}$ and, by Eq. (35), $\int_0^T \|M^N_\pi(s) - m_A^N(s)\| ds$ is less than $(\|M^N(0) - m_0\| + I_0(N)T + \epsilon)\frac{e^{L_1T} - 1}{L_1}$. Thus

$$
|V_\pi^N(M^N(0)) - E[v_A^N(m_0)]| \leq B_\epsilon(N, \|M^N(0) - m_0\|)
$$

where $B_\epsilon(N, \delta) := I(N) \|r\|_\infty + K_r\sqrt{2}I_1(N) + \frac{K_r I(N)}{2L_1} + \frac{2(T+1)\|r\|_\infty J(N,T)}{\epsilon^2} + K_r (\delta + I_0(N)T + \epsilon)\frac{e^{L_1T} - 1}{L_1} + e^{L_1T}$.

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

$$
|V_\pi^N(M^N(0)) - E[v_A^N(m_0)]| \leq B(N, \|M^N(0) - m_0\|)
$$

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

### E. Proof of Theorem 3

Let $\tilde{\alpha}^N$ be the right-continuous function constant on the intervals $[kI(N); (k+1)I(N)]$ such that $\tilde{\alpha}^N(s) = \alpha(s)$. $\tilde{\alpha}^N$ can be viewed as a policy independent of $m$. Therefore, by Theorem 5, on the set $\Omega \setminus (\Omega_1 \cup \Omega_2)$, for every $t \in [0; T]$: 

$$
\|M_\alpha(t) - \phi_\alpha(m_0, \alpha)\| \leq \|\alpha\|_{\infty} + I_0(N)T + \epsilon\|e^{L_1T} + u(t)\| \\
with u(t) := |\phi_\alpha(m_0, \tilde{\alpha}^N) - \phi_\alpha(m_0, \alpha)|. We have
$$

$$
u(t) := \int_0^t |f(\phi_\alpha(m_0, \alpha), \alpha(s)) - f(\phi_\alpha(m_0, \tilde{\alpha}^N), \tilde{\alpha}^N(s))| ds \\
\leq \int_0^t K(\|\phi_\alpha(m_0, \alpha) - \phi_\alpha(m_0, \tilde{\alpha}^N)\| + d(\alpha(s), \tilde{\alpha}^N(s))) ds \\
\leq K\int_0^t u(s) ds + Kd_1
$$

where $d_1 := \int_0^T \|\alpha(t) - \tilde{\alpha}^N(t)\| dt$. Therefore, using Grönwall’s inequality, we have $u(t) \leq Kd_1 e^{Kt}$. By Lemma 5, this shows Eq. (13). The rest of the proof is as for Theorem 6.

### Lemma 5

If $\alpha$ is a piecewise Lipschitz continuous action function on $[0; T]$, of constant $K_\alpha$, and with at most $p$ discontinuity points, then $\int_0^T d(\alpha(t), \tilde{\alpha}^N(t)) dt$ is bounded by

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

### Proof of Lemma 5

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

$$
d_1 = \sum_{i=0}^{[T/I(N)]} \int_{iI(N)}^{(i+1)I(N)} \|\alpha(s) - \tilde{\alpha}^N(s)\| ds \\
\leq \sum_{i=0}^{[T/I(N)]} \int_{iI(N)}^{(i+1)I(N)} \|\alpha(s) - \alpha(iI(N))\| ds.
$$

If $\alpha$ has no discontinuity point on $[iI(N), (i+1)I(N))$, then $\int_{iI(N)}^{(i+1)I(N)} d(\alpha(s), \alpha(iI(N))) ds \leq I(N)\|_{\infty} K_\alpha ds \leq K_\frac{\alpha}{2} I(N)^2$. If $\alpha$ has one or more discontinuity points on $[iI(N), (i+1)I(N))$, then $\int_{iI(N)}^{(i+1)I(N)} d(\alpha(s), \tilde{\alpha}^N(iI(N))) ds \leq 2\|\alpha\|_{\infty} I(N)$.

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 at most $\int_{kI(N)}^{(k+1)I(N)} d(\alpha(s), \tilde{\alpha}^N(s)) ds \leq 2\|\alpha\|_{\infty} I(N)$.

### F. Proof of Theorem 2

This theorem is a direct consequence of Theorem 3 and Theorem 6. We do the proof for almost sure convergence, the proof for convergence in probability is similar. To prove the theorem we prove

$$
\limsup_{N \to \infty} V_\pi^N(M^N(0)) \leq v_\pi(m_0) \leq \liminf_{N \to \infty} V_\pi^N(M^N(0))
$$

for all $N$. Let $\epsilon > 0$ and $\alpha$ be an action function such that $\alpha(m_0) \geq v_\pi(m_0) - \epsilon$ (such an action is called $\epsilon$-optimal). Theorem 3 shows that $\lim_{N \to \infty} V_\pi^N(M^N(0)) = v_\pi(m_0) - \epsilon$ a.s. This shows that $\liminf_{N \to \infty} V_\pi^N(M^N(0)) \geq v_\pi(m_0) - \epsilon$; this holds for every $\epsilon > 0$ thus $\liminf_{N \to \infty} V_\pi^N(M^N(0)) \geq v_\pi(m_0)$ a.s., which establishes the second inequality in Eq. (37), on a set of probability 1.

Now, let $B(N, \delta)$ be as in Th. 6, $\epsilon > 0$ and $\pi^N$ such that $V_\pi^N(M^N(0)) \leq V_\pi^N(M^N(0)) + \epsilon$. By Th. 6, $V_\pi^N(M^N(0)) \leq E\left(v_A^N(m_0) + B(N, \delta)^N \leq v_\pi(m_0) + B(N, \delta)^N \right)$ where

$$
\delta^N := \|M^N(0) - m_0\|. Thus V_\pi^N(M^N(0)) \leq v_\pi(m_0) + B(N, \delta)^N + \epsilon. If further $\delta^N \to 0$ a.s. it follows that $\limsup_{N \to \infty} V_\pi^N(M^N(0)) \leq v_\pi(m_0) + \epsilon$ a.s. for every $\epsilon > 0$, thus $\limsup_{N \to \infty} V_\pi^N(M^N(0)) \leq v_\pi(m_0)$ a.s.

### VI. Conclusion and Perspectives

There are several natural questions arising from this work. One concerns the convergence of optimal policies. Optimal policies $\pi^N_\pi$ of a stochastic systems with $N$ objects may not be unique, they may also exhibit thresholds and therefore be discontinuous. This implies that $M^N$ will not converge to an ODE but to a differential inclusion in general [9]. In some particular cases, such as the best response dynamics studied in [11], limit theorems can nevertheless be obtained, at the cost of a much greater complexity. In full generality
however, this problem is still open and definitely deserves further investigations.

The second question concerns the time horizon. In this paper we have focused on the finite horizon case. Actually, most results and in particular theorems 2 and 3, remain valid with an infinite horizon with discount. Here is the key point that makes everything work. When the rewards \( r(s, a) \) are bounded, for a given discount \( \beta < 1 \) and any \( \varepsilon > 0 \), there exists a finite \( T \) such that the discounted value of a policy \( \pi \) can be decomposed into the value over horizon \( T \) plus a term less than \( \varepsilon \):

\[
\sum_{t=0}^{\infty} \beta^t r(M^N(t), \pi(M^N(t))) \leq \sum_{t=0}^{T} \beta^t r(M^N(t), \pi(M^N(t))) + \varepsilon.
\]

Therefore, the main result of this paper, which states that a policy that is optimal in the mean field limit is near-optimal for the system with \( N \) objects, also holds in the infinite horizon discounted case.

As for the infinite horizon without discount or average reward cases, convergence of the value when \( N \) goes to infinity is not guaranteed in general. Finding natural assumptions under which convergence holds is also one of our goals for the future.

REFERENCES


Nicolas Gast is currently a post-doctoral fellow at EPFL. He graduated from Ecole Normale Supérieure, Paris, in 2007 where he obtained the Agregation in Mathematics in 2007. He received a Ph.D. degree in Computer science from the University of Grenoble and INRIA (in Grenoble, France) in 2010. His main interests are in stochastic modeling and control of large systems.

Bruno Gaujal is a Research Director in INRIA Rhones-Alpes, head of the large-scale computing group, MESCAL. He has held several positions in AT&T Bell Labs, INRIA Sophia-Antipolis, Loria and Ecole Normale Supérieure in Lyon. His main interests are in performance evaluation, optimization and control of large discrete event dynamic systems with applications to telecommunications networks and large computing infrastructures.

Jean-Yves Le Boudec is full professor at EPFL and fellow of the IEEE. He graduated from Ecole Normale Superieure de Saint-Cloud, Paris, where he obtained the Agregation in Mathematics in 1980 and received his doctorate in 1984 from the University of Rennes, France. From 1984 to 1987 he was with INSA/IRISA, Rennes. In 1987 he joined Bell Northern Research, Ottawa, Canada, as a member of scientific staff in the Network and Product Traffic Design Department. In 1988, he joined the IBM Zurich Research Laboratory where he was manager of the Customer Premises Network Department. In 1994 he joined EPFL as associate professor.

His interests are in the performance and architecture of communication systems. In 1984, he developed analytical models of multiprocessor, multiple bus computers. In 1990 he invented the concept called “MAC emulation” which later became the ATM forum LAN emulation project, and developed the first ATM control point based on OSPF. He also launched public domain software for the interworking of ATM and TCP/IP under Linux. He proposed in 1998 the first solution to the failure propagation that arises from common infrastructures in the Internet. He contributed to network calculus, a recent set of developments that forms a foundation to many traffic control concepts in the internet, and co-authored a book on this topic. He is also the author of the book "Performance Evaluation" (2010). He received the IEEE millennium medal, the Infocom 2005 Best Paper award, the CommSoc 2008 William R. Bennett Prize and the 2009 ACM Sigmetrics Best Paper award.

He is or has been on the program committee or editorial board of many conferences and journals, including Sigcom, Sigmetrics, Infocom, Performance Evaluation and ACM/IEEE Transactions on Networking.