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 Γ^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 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 [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 problem, 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

N. Gast is with EPFL IC-LCA2 – Lausanne, Switzerland. Email: nicolas.gast@epfl.ch, Tel: +41 21 69 31254. Corresponding author.

B. Gaujal is with INRIA Grenoble–Rhône-Alpes and LIG – 51 av. Jean-Kuntzmann, 38330 Montbonnot, France. Email: bruno.gaujal@imag.fr. Tel: +33 4 76 61 20 31. Fax: +33 4 76 61 20 99

J.-Y. Le Boudec is with EPFL IC-LCA2 – Lausanne, Switzerland. Email: jean-yves.leboudec@epfl.ch. Tel: +41 21 69 36631.

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. Malhamé 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 Tembiné 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 $\mathcal{S} = \{1 \dots S\}$. 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:

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

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

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. Let Γ^N be the transition kernel, namely Γ^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:

$$\begin{aligned}\mathcal{P}\{X^N(k{+}1){=}y_1 \dots y_N | X^N(k){=}x_1 \dots x_N, A^N(k){=}a\} \\ = \Gamma^N\left(x_1 \dots x_N, y_1 \dots y_N, a\right)\end{aligned} \tag{2}$$

We assume that

$(A0)$ Objects are observable only through their states

in particular, the controller can observe the collection of all states $X_1^N, X_2^N, \dots$, but not the identities $n = 1, 2, \dots$. 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 $\mathfrak{S}^N$ be the set of permutations of $\{1, 2, \dots, N\}$. By a slight abuse of notation, for $\sigma \in \mathfrak{S}^N$ and $x \in \mathcal{S}^N$ we also denote with $\sigma(x)$ the collection of object states after the permutation, i.e. $\sigma(x) \stackrel{\text{def}}{=} \left(x_{\sigma^{-1}(1)} \dots x_{\sigma^{-1}(N)}\right)$. The requirement is that

$$\Gamma^N(\sigma(x), \sigma(y), a) = \Gamma^N(x, y, a) \tag{3}$$

for all $x, y \in \mathcal{S}^N$, $\sigma \in \mathfrak{S}^N$ 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. $(\mathcal{A}, d)$ 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 π_k might depend on the sequence of past and present states of the process X^N, however, it it 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, π_k is a function $\mathcal{P}(\mathcal{S}) \to \mathcal{A}$. Let $M^N_\pi(k)$ denotes the occupancy measure of the system at time k when the controller applies policy π.

The controller focusses 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 π is the expected gain over the horizon $[0; H^N]$ starting from m_0 when applying the policy π. It is defined by

$$V^N_\pi(m) \stackrel{\text{def}}{=} \mathbb{E}\Big(\sum_{k=0}^{H^N-1} r^N(M^N_\pi(k), \pi(M^N_\pi(k))) + r_f(M^N_\pi(H^N)) \Big| M^N_\pi(0) = m \Big). \tag{4}$$

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

$$V^N_*(m) = \sup_\pi V^N_\pi(m) \tag{5}$$

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 $F^N(m, a)$:

$$F^N(m,a) \stackrel{\text{def}}{=} \mathbb{E}\big(M^N(k+1) - M^N(k) \,\big|\, M^N(k) = m, A^N(k) = a\big). \tag{6}$$

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 $a \in \mathcal{A}$: $|F^N(m,a)/I(N) - f(m,a)| \le I_0(N)$. 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 $(\hat{M}^N(t))_{t\in\mathbb{R}^+}$ 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 occupancy measure 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 $\le x$). 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 $\hat{M}^N_\pi$ 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 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 α and an initial condition m_0, we consider the following ordinary integral equation for $m(t)$, $t \in \mathbb{R}^+$:

$$m(t) - m(0) = \int_0^t f(m(s), \alpha(s)) ds. \tag{7}$$

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 α, this equation satisfies the Cauchy Lipschitz condition and 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: 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 α and starting from $m(0) = m_0$:

$$v_\alpha(m_0) \stackrel{\text{def}}{=} \int_0^T r\left(\phi_s(m_0, \alpha), \alpha(s)\right) ds + r_f(\phi_T(m_0, \alpha)). \tag{9}$$

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 $v_*(m_0)$:

$$v_*(m_0) = \sup_\alpha v_\alpha(m_0), \tag{10}$$

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.

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

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 $\langle m, m'\rangle = \sum_{i=1}^S m_i m'_i$ 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) Γ^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 π, the number of objects that perform a transition between time slot k and $k+1$ per time slot $\Delta_\pi^N(k)$ satisfies

$$\begin{aligned} \mathbb{E}\left(\Delta_\pi^N(k) \middle| M_\pi^N(k) = m\right) &\leq NI_1(N) \\ \mathbb{E}\left(\Delta_\pi^N(k)^2 \middle| M_\pi^N(k) = m\right) &\leq N^2 I(N) I_2(N) \end{aligned}$$

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

b) (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

$$\left\| \frac{1}{I(N)} F^N(m,a) - f(m,a) \right\| \leq I_0(N) \tag{11}$$

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

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

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

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

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 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:*

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

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 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 this 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^N_α, 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). The value for this system is defined by

$$V^N_\alpha(m_0) \stackrel{\text{def}}{=} \mathbb{E}\Big(\sum_{k=0}^{H^N} r\big(M^N_\alpha(k), \alpha(kI(N))\big) + r_f(M^N_\alpha(H^N))\Big| M^N_\alpha(0) = m_0\Big). \quad (12)$$

Our next result is the convergence of M^N_α and of the value:

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

$$\mathbb{P}\Big\{\sup_{0\le t\le T} \left\|\hat{M}^N_\alpha(t) - \phi_t(m_0,\alpha)\right\| > \big[\left\|M^N(0) - m_0\right\| + I'_0(N,\alpha)T + \epsilon\big]e^{L_1 T}\Big\} \le \frac{J(N,T)}{\epsilon^2} \quad (13)$$

and

$$\left|V^N_\alpha\left(M^N(0)\right) - v_\alpha(m_0)\right| \le B'\left(N, \left\|M^N(0) - m_0\right\|\right) \quad (14)$$

where J, I'_0 and B' 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 $\lim_{N\to\infty,\delta\to 0} B'(N,\delta) = 0$.

In particular, if $\lim_{N\to\infty} M^N_\pi(0) = m_0$ almost surely [resp. in probability] then $\lim_{N\to\infty} V^N_\alpha\left(M^N(0)\right) = v_\alpha(m_0)$ 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 when 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 α^ϵ such that $v_*(m) = \sup_\alpha v_\alpha(m) \le v_{\alpha^\epsilon} + \epsilon$. Combining Theorem 2 and Theorem 3, we have that $\lim_{N\to\infty} V^N_{\alpha^\epsilon} = v_{\alpha^\epsilon} \ge v_* - \epsilon = \lim_{N\to\infty} V^N_* - \epsilon$. This shows that α^ϵ is optimal up to 2ϵ for N large enough. This shows the following corollary:

Corollary 4 (Asymptotically Optimal Policy)**.** *If α_* is an optimal action function for the limiting system and if $\lim_{N\to\infty} M^N(0) = m_0$ almost surely [resp. in probability], then we have:*

$$\lim_{N\to\infty} \left|V^N_{\alpha_*} - V^N_*\right| = 0,$$

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 $\phi_t(m_0, A^N_\pi)$ 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 π. The process M^N_π is defined on some probability space Ω. To each $\omega \in \Omega$ corresponds a trajectory $M^N_\pi(\omega)$, and for each $\omega \in \Omega$, we define an action function $A^N_\pi(\omega)$. This random function is piecewise constant on each interval $[kI(N), (k+1)I(N))$ ($k \in \mathbb{N}$) and is such that $A^N_\pi(\omega)(kI(N)) \stackrel{\text{def}}{=} \pi_k(M^N(k))$ 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 any action function α, $\phi_t(m_0,\alpha)$ is the solution of the ODE (7). For every ω, $\phi_t(m_0, A^N_\pi(\omega))$ is the solution of the limiting system with action function $A^N_\pi(\omega)$, i.e.

$$\phi_t(m_0, A^N_\pi(\omega)) = m_0 + \int_0^t f(\phi_s(m_0, A^N_\pi(\omega)), A^N_\pi(\omega)(s))ds.$$

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

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 π.

Theorem 5. *Under Assumptions (A0) to (A3), for any $\epsilon > 0$, $N \ge 1$ and any policy π:*

$$\mathbb{P}\Big\{\sup_{0\le t\le T} \left\|\hat{M}^N_\pi(t) - \phi_t(m_0, A^N_\pi)\right\| > \big[\left\|M^N(0) - m_0\right\| + I_0(N)T + \epsilon\big]e^{L_1 T}\Big\} \le \frac{J(N,T)}{\epsilon^2} \quad (15)$$

where $\hat{M}_\pi^N$ 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 π be a policy and A_π^N the sequence of actions corresponding to a trajectory M_π^N as we just defined. Eq.(9) defines the value for the deterministic limit when applying a sequence of actions. This defines a random variable $v_{A_\pi^N}(m_0)$ that corresponds to the value over the limit system when using A_π^N as action function. The random part comes from A_π^N. $\mathbb{E}\left[v_{A_\pi^N}(m_0)\right]$ designates the expectation of this value over all possible A_π^N. A first consequence of Theorem 5 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 6 (Uniform convergence of the value)**.** *Let A_π^N be the random action function associated with M_π^N, as defined earlier. Under Assumptions (A0) to (A3),*

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

where B is defined in Section V-A.

Note that $\lim_{N\to\infty,\delta\to 0}B(N,\delta)=0$; in particular, if $\lim_{N\to\infty}M_\pi^N(0)=m_0$ 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 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 $U^N(m,t)=\sup_\pi\mathbb{E}\left[\sum_{k=t/I(N)}^{T/I(N)}r^N(M_\pi^N(k)):M^N(t)=m\right]$. The optimal value can be computed by a discrete dynamic programming algorithm [21] by setting $U^N(m,T)=r_f(m)$ and

$$U^N(m,t)=\sup_{a\in\mathcal{A}}\mathbb{E}\Big[r^N(m,a)+U^N(M^N(t+I(N)),t+I(N))\Big|\bar{M}^N(t)=m,A^N(t)=a\Big].$$

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:

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

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[1] 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_π^N of the policy π obtained by Algorithm 1 is asymptotically optimal:

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

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 Γ^N and let $M^N(0)$ be the initial occupancy measure;
- Compute the limit f of the drift of Γ^N;
- Solve the HJB equation (16) on $[0,HI(N)]$. This provides an optimal control function $\alpha_*(M_0^N,t)$;
- Construct a discrete control π for the discrete system: the action to be taken under state $M^N(k)$ at step k is

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

- **return** π;

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 value of the adaptive policy π' should be larger than the value of the static policy π 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

[1]if there is no optimal control of the HJB equation (16), one can replace the optimal control function α_* used in both algorithms by an action function that is $1/N$-optimal. This will still provide an asymptotically optimal policy.

Algorithm 2: Algorithm constructing an adaptive policy 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 (16) 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 $\Gamma_{\pi'}^N$;
k:= k+1;
until k=H;
return π';

limiting ODE is quite simple and can be optimized in closed analytical form.

The second example considers a classic virus problem. Although virus propagations concern discrete objects (individuals or devices), most work in the literature study a continuous approximation of the problem under the form of an ODE. The justification of passing from a discrete to a continuous model is barely mentioned in most papers (they mainly focus on the study of the ODE). Here we present a discrete dynamical system based on a simple stochastic mobility model for the individuals whose behavior converges to a classic continuous model. We show on a numerical example that the limiting problem provides a policy that is close to optimal, even for a system with a relatively small numbers of nodes.

Finally, the last example comes from routing optimization in a queuing network model of volunteer computing platforms. The purpose of this last example is to show that a discrete optimal control problem suffering from the curse of dimensionality can be replaced by a continuous optimization problem where an HJB equation must be solved over a much smaller state space.

1) Utility Provider Pricing: This is a simplified discrete Merton's problem. This 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: although the original system is difficult to solve even numerically when N is large, taking the limit when N goes to infinity makes it simple to solve, in an analytical form.

We consider a system made of a utility and N users; users can be either in state S (subscribed) or U (unsubscribed). The utility fixes their price $\alpha \in [0,1]$. Customers are asynchronous and revise their status according to independent time processes. An equivalent model is that at every time step (which size decreases as $1/N$), one customer is chosen randomly and revises her status: if she is in state U [resp. S], with probability $s(\alpha)$ [resp. $a(\alpha)$] she moves to the other state; $s(\alpha)$ is the probability of a new subscription, and $a(\alpha)$ is the probability of attrition. We assume $s(\cdot)$ decreases with α and $a(\cdot)$ increases. The instant reward at time t is $\alpha(t)x(t)$ and there is no final reward. If the price is large, the instant gain is large, but the utility loses customers, which eventually reduces the gain.

Within our framework, this problem can be seen as a Markovian system made of N objects (users) and one controller (the provider). The intensity of the model is $I(N)$=$1/N$. If the immediate profit is divided by N (this does not alter the optimal pricing policy) and if $x(t)$ is the fraction of objects in state S at time t and $\alpha(t) \in [0;1]$ is the action taken by the provider at time t, the mean field limit of the system is:

$$\begin{aligned} \frac{dx}{dt} &= -x(t)a(\alpha(t)) + (1-x(t))s(\alpha(t)) \\ &= s(\alpha(t)) - x(s(\alpha(t)) + a(\alpha(t)) \end{aligned} \tag{17}$$

and the rescaled profit over a time horizon T is $\int_0^T x(t)\alpha(t)dt$. Call $u_*(t,x)$ the optimal benefit over the interval $[t,T]$ if there is a proportion x of subscribers at time t. The Hamilton-Jacobi-Bellman equation is

$$\frac{\partial}{\partial t}u_*(t,x) + H\left(x, \frac{\partial}{\partial x}u_*(t,x)\right) = 0 \tag{18}$$

$$\text{with} \quad H(x,p) = \max_{\alpha\in[0,1]} \left[p(s(\alpha) - x(s(\alpha)+a(\alpha)) + \alpha x\right].$$

H can be computed under reasonable assumptions on the rates of subscription and attrition $s()$ and $a()$, which can then be used to show that there exists an optimal policy that is threshold based. To continue the rest of this illustration, we consider the radically simplified case where α can take only the values 0 and 1 and under the conditions $s(0) = a(1) = 1$ and $s(1) = a(0) = 0$, in which case the ODE becomes

$$\frac{dx}{dt} = -x(t)\alpha(t) + (1-x(t))(1-\alpha(t)) = 1 - x(t) - \alpha(t), \tag{19}$$

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 chose 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.

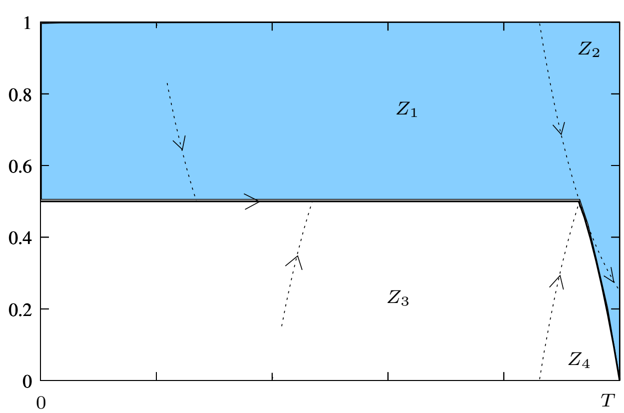


Figure 1. Evolution of the proportion of subscribers (y-axis) under the optimal pricing policy.

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 α_* and the performance of other policies for the stochastic system for small values of N. We show that α_* 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 β. 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 α/N where $\alpha \in [0; \alpha_{\max}]$ 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:

$$\mathbb{E}\left(D_\pi(T) + \frac{1}{NT}\sum_{k=1}^{NT} f(I_\pi(k))\right).$$

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.

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS - qS \\ \frac{dI}{dt} &= \beta IS - bI - \alpha(t)I \\ \frac{dD}{dt} &= \alpha(t)I \\ \frac{dR}{dt} &= bI + qS, \end{aligned} \tag{20}$$

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 α such that the damage function $D(T) + \frac{1}{T}\int_0^T f(I(t))dt$ is maximized under the constraint $0 \le \alpha(t) \le \alpha_{\max}$ (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 α_* is of bang-bang type: there exists $t_1 \in [0 \dots T)$ s.t.

$$\alpha_*(t) = \begin{cases} 0 & \text{for } 0 < t < t_1 \\ \alpha_{\max} & \text{for } t_1 < t < T \end{cases} \tag{21}$$

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 α_* of the worm is indeed asymptotically optimal when the number of objects goes to infinity.

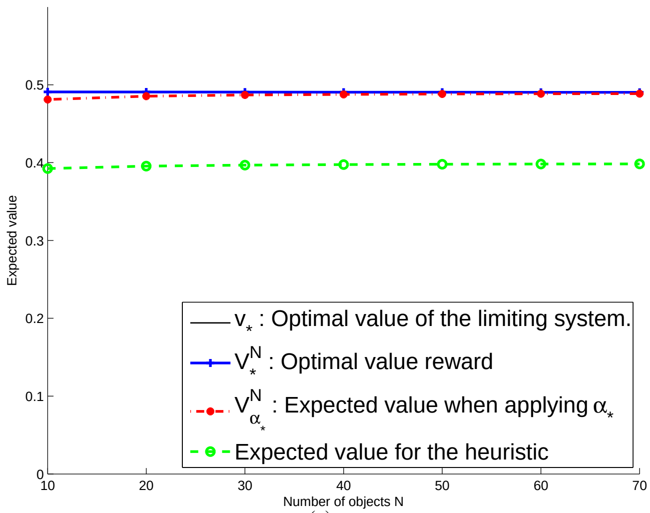
We investigated numerically the performance of α_* against various infection policies for small values of the number of nodes in the system N. These results are reported on Figure 2, where we compare four values:

- v_* – 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 α_* 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 ν for the whole time. The curve on the figure is drawn for the optimal ν (recomputed for each parameter N).

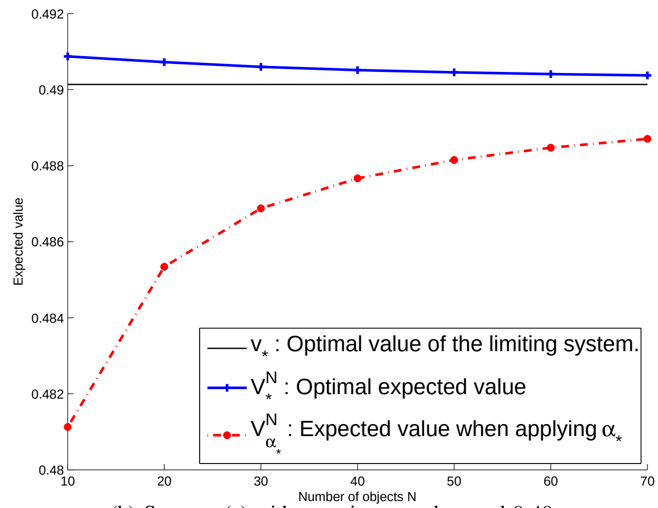
We implemented a simulator that follows strictly the model of infection described earlier in this part. We chose parameters similar to those used in [15]: the parameter for the evolution of the system are $\beta = .6$, $q = .1$, $b = .1$, $\alpha_{\max} = 1$ and the damage function to be optimized is $D(T) + \frac{1}{T}\int_0^T I^2(t)dt$ 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: $\alpha_*(t) = \mathbf{1}_{\{t>4.9\}}$.

Theorem 2 shows that α_* 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 α_* 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



(a)



(b) Same as (a) with $y-$axis zoomed around 0.49

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 $v_{\alpha_*}^N$ 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.

we show the two quantities V_*^N, $V_{\alpha_*}^N$ and their common limit v_*. This figure shows that the convergence is indeed very fast. Other numerical experiments indicate that this is true for a large panel of parameters. Although these figures seem to indicate that $V_{\alpha_*}^N \leq v_* \leq V_*^N$, this is not true in general, for example adding $5D(t)$ to the damage function leads to $V_{\alpha_*}^N \leq V_*^N \leq v_*$ ($V_{\alpha_*}^N$ is always less than V_*^N by definition of V_*^N).

3) Brokering Problem: Finally, let us consider a model of a 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 queuing 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 $x \in \{\text{on}, \text{off}\}$. At each time step, an active user **s**ends one job with probability p_s^N and becomes **i**nactive with probability p_i/N. An inactive user sends no jobs to the system and becomes **o**n with probability p_o/N.

There are C clusters in the system. Each cluster c contains Q_c^N 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 μ_c/N at this time slot. A resource gets **b**roken with probability p_b/N. In that case, it discards all the packets of its buffer. A broken resource becomes **v**alid with probability p_v/N.

At each time step, the broker takes an action $a \in \mathcal{P}(\{1 \dots C\})$ and sends the packets it received to the clusters according to the distribution a. A packet sent to cluster c joins the queue of one of the resources, k; according to a local rule (for example chosen uniformly among the Q_c^N resources composing the cluster). If the queue of resource k 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.

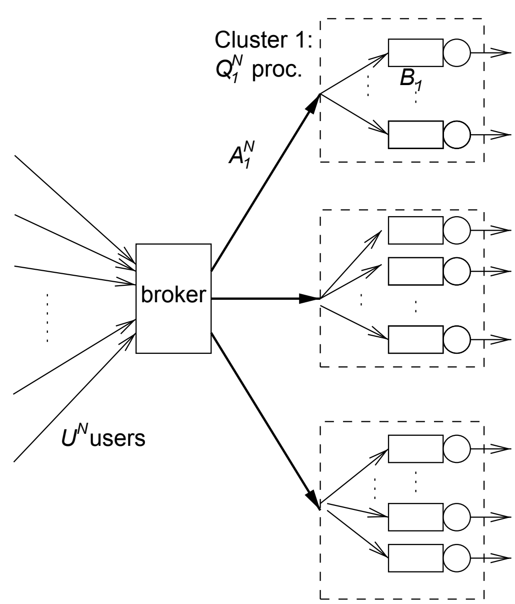


Figure 3. The brokering problem in a desktop grid system, such as Boinc

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_c^N, 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 $m_o(t)$, that 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 denotes the fraction of users (both active or inactive) and q_c the fraction of processors in cluster c. These fractions are constant (independent of time) and satisfy $m + q_1 + \cdots + q_C = 1$. We get the following equations:

$$\frac{dm_o}{dt} = -p_i m_o(t) + p_o(m - m_o) \tag{22}$$

$$\frac{dq_{c,0}}{dt} = p_a b_c - \frac{\alpha_c(t) p_s m_o}{q_c} q_{c,0} + \mu_c q_{c,1} - p_b q_{c,0} \tag{23}$$

$$\begin{aligned}\frac{dq_{c,j}}{dt} &= \frac{\alpha_c(t) p_s m_o}{q_c} (q_{c,j-1}(t) - q_{c,i}(t)) \\ &\quad + \mu_c(q_{c,j+1} - q_{c,j}) - p_b q_{c,j}\end{aligned} \tag{24}$$

$$\frac{dq_{c,J_c}}{dt} = \frac{\alpha_c(t) p_s m_a}{q_c} q_{c,J_c-1} - \mu_c q_{c,J_c} - p_b q_{c,J_c} \tag{25}$$

$$\frac{db_c}{dt} = -p_v b_c(t) + p_b \sum_{j=0}^{J_c} q_{c,j}. \tag{26}$$

where (23) and (25) hold for each cluster c and (24) holds for each cluster c and for all $j \le J_c$. The cost associated to the action function α is:

$$\begin{aligned}\int_0^T \Big[&\sum_{c=1}^{C} \sum_{j=1}^{J_c} j q_{c,j}(t) + \gamma\Big(\sum_{c=1}^{C} p_b \sum_{j=1}^{J_c} j q_{c,j}(t) \\ &+ \sum_{c=1}^{C} \frac{\alpha_c(t) p_s m_o(t)}{q_c} (q_{c,J_c}(t) + b_c(t))\Big)\Big] dt\end{aligned} \tag{27}$$

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 γ 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 $(J+2)C$ 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

$$\begin{aligned}I_0'(N,\alpha) \overset{\text{def}}{=} &I_0(N) + I(N) K e^{(K-L_1)T} \\ &\cdot \left(\frac{K_\alpha}{2} + 2\left(1 + \min(1/I(N), p)\right) \|\alpha\|_\infty \right)\end{aligned}$$

$$\begin{aligned}J(N,T) \overset{\text{def}}{=} 8T \Big\{ &L_1^2 \left[I_2(N) I(N)^2 + I_1(N)^2 \left(T + I(N)\right) \right] \\ &+ S^2 \left[2I_2(N) + I(N) \left(I_0(N) + L_2\right)^2 \right] \Big\}\end{aligned}$$

$$\begin{aligned}B(N,\delta) \overset{\text{def}}{=} &I(N) \|r\|_\infty + K_r \sqrt{2} I_1(N) \\ &+ K_r \left(\delta + I_0(N) T\right) \left(e^{L_1 T} + \frac{e^{L_1 T} - 1}{L_1} \right) \\ &+ \frac{3}{2^{\frac{1}{3}}} \left[e^{L_1 T} + \frac{e^{L_1 T} - 1 + I(N)/2}{L_1} \right]^{\frac{2}{3}} \\ &\cdot K_r^{\frac{2}{3}} \|r\|_\infty^{\frac{1}{3}} J(N,T)^{\frac{1}{3}} (T+1)^{\frac{2}{3}}\end{aligned}$$

$B'(N,\delta)$ 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 $\mathcal{P}$ be the set of probabilities on $\mathcal{S}$ and $\mu^N : \mathcal{S}^N \to \mathcal{P}$ defined by $\mu^N(x)_i = \frac{1}{N} \sum_{n=1}^N \mathbf{1}_{x_n = i}$ for all $i \in \mathcal{S}$. Also let $\mathcal{P}^N$ be the image set of μ^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 $x, x' \in \mathcal{S}^N$, if $\mu^N(x) = \mu^N(x')$ there exists some $\sigma \in \mathfrak{S}^N$ 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 $x, x' \in \mathcal{S}^N$, with $\mu^N(x) = \mu^N(x')$. There is at least one coordinate, say i, such that $x_i' = x_1$, because there is the same number of occurrences of $s = x_1$ in both x and x'. Let $y = x_2 ... x_N$ and $y' = x_1' ... x_{i-1}' x_{i+1}' ... x_N'$. Then $\mu^{N-1}(y) = \mu^{N-1}(y')$, therefore there exists some $\tau \in \mathfrak{S}^{N-1}$ such that $y' = \tau(y)$. Define σ by $\sigma(1) = i$, $\sigma(j) = \tau(j) + \mathbf{1}_{\tau(j) > i}$, for $j \ge 2$, so that $x' = \sigma(x)$. Clearly σ is a permutation of $\{1, ..., N\}$. ∎

Let $f : \mathcal{S}^N \to E$ where E is some arbitrary set. We say that f is invariant under $\mathfrak{S}^N$ if $f \circ \sigma = f$ for all $\sigma \in \mathfrak{S}^N$. 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 $f : \mathcal{S}^N \to E$ is invariant under $\mathfrak{S}$ then there exists $\bar{f} : \mathcal{P}^N \to E$ such that $\bar{f} \circ \mu^N = f$.*

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

The sequence of actions a_k is given and N is fixed. We are thus given a time-inhomogeneous Markov chain X^N on $\mathcal{S}^N$, 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 $\sigma \in \mathfrak{S}^N$ and any states x, y we have

$$G_k(\sigma(x), \sigma(y)) = G_k(x,y) \tag{28}$$

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

Pick some arbitrary test function $\varphi : \mathcal{S}^N \to \mathbb{R}$ and fix some time $k \geq 1$; we will now compute $\mathbb{E}\left(\varphi(M^N(k)) \middle| \mathcal{F}(k-1)\right)$. Because M^N is a function of X^N and X^N is a Markov chain, $\mathbb{E}\left(\varphi(M^N(k)) \middle| \mathcal{F}(k-1)\right)$ is a function, say ψ, of $X^N(k-1)$. We have, for any fixed $\sigma \in \mathfrak{S}^N$:

$$
\begin{aligned}
\psi(x) &\stackrel{\text{def}}{=} \sum_{y \in \mathcal{S}^N} G_k(x,y)\varphi(\mu^N(y)) \\
&= \sum_{y \in \mathcal{S}^N} G_k(x,\sigma(y))\varphi(\mu^N(\sigma(y))) \\
&= \sum_{y \in \mathcal{S}^N} G_k(x,\sigma(y))\varphi(\mu^N(y)) \\
\psi(\sigma(x)) &= \sum_{y \in \mathcal{S}^N} G_k(\sigma(x),\sigma(y))\varphi(\mu^N(y)) \\
&= \sum_{y \in \mathcal{S}^N} G_k(x,y))\varphi(\mu^N(y))
\end{aligned}
$$

where the last equality is by Eq.(28). Thus $\psi(\sigma(x)) = \psi(x)$ and by Lemma 2 there exists some function $\bar{\psi}$ such that $\psi(x) = \bar{\psi}\left(\mu^N(x)\right)$, i.e.

$$\mathbb{E}\left(\varphi(M^N(k)) \middle| \mathcal{F}(k-1)\right) = \bar{\psi}\left(M^N(k-1)\right)$$

In particular, the variable $\mathbb{E}\left(\varphi(M^N(k)) \middle| \mathcal{F}(k-1)\right)$ is $\mathcal{G}(k-1)-$measurable. Now $\mathbb{E}\left(\varphi(M^N(k)) \middle| \mathcal{G}(k-1)\right)$ is equal to

$$
\begin{aligned}
&\mathbb{E}\left(\mathbb{E}\left(\varphi(M^N(k)) \middle| \mathcal{F}(k-1)\right) \middle| \mathcal{G}(k-1)\right) \\
&= \mathbb{E}\left(\bar{\psi}\left(M^N(k-1)\right) \middle| \mathcal{G}(k-1)\right) = \bar{\psi}\left(M^N(k-1)\right)
\end{aligned}
$$

which expresses that M^N is a Markov chain.

C. Proof of Theorem 5

The proof is inspired by the method in [3]. The main idea of the proof is to bound $\left\|M_\pi^N(k) - \phi_{kI(N)}(m_0, A_\pi^N)\right\|$ by

$$\left\|M_\pi^N(k) - M^N(0) - \sum_{j=0}^{k-1} f^N(j)\right\| \tag{29}$$

$$+ \left\|M^N(0) + \sum_{j=0}^{k-1} f^N(j) - \phi_{kI(N)}(m_0, A_\pi^N)\right\| \tag{30}$$

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 line (29) is bounded with high probability using a Martingale argument (Lemma 4) and the second part (30) 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 $W_{k+1} \stackrel{\text{def}}{=} \left\|M_\pi^N(k+1) - M_\pi^N(k)\right\|$. Thus,

$$\left\|\hat{M}_\pi^N(t) - \bar{M}_\pi^N(t)\right\| \leq W_{k+1} \tag{31}$$

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

$$Z_\pi^N(k) \stackrel{\text{def}}{=} M_\pi^N(k) - M^N(0) - \sum_{j=0}^{k-1} F^N(M_\pi^N(j), \pi_j(M_\pi^N(j))) \tag{32}$$

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 $\hat{M}_\pi^N(t)$ is equal to

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

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:

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

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

$$
\begin{aligned}
\left\|\hat{M}_\pi^N(t) - \phi_t(m_0, A_\pi^N)\right\| \leq \left\|M_\pi^N(0) - m_0\right\| + \left\|\hat{Z}_\pi^N(t)\right\| \\
+L_1 \int_0^t \left\|\hat{M}_\pi^N(s) - \phi_s(m_0, A_\pi^N)\right\| ds \\
+I_0(N)t + L_1 I(N) \sum_{k=0}^{\lfloor t/I(N) \rfloor} W_{k+1}
\end{aligned}
$$

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

$$
\begin{aligned}
\Omega_1 &= \left\{\omega \in \Omega : \sup_{0 \leq k \leq \frac{T}{I(N)}} \sum_{j=0}^{k} W_{k+1} > b_1\right\}, \\
\Omega_2 &= \left\{\omega \in \Omega : \sup_{0 \leq k \leq \frac{T}{I(N)}} \left\|Z_\pi^N(k)\right\| > b_2\right\}.
\end{aligned}
$$

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 $M^N(k)$ and $M^N(k+1)$, we have $\mathbb{E}\left(W(k+1) \middle| \mathcal{F}_k\right) = \mathbb{E}\left(W(k+1) \middle| M_\pi^N(k)\right)$. Moreover, since $\Delta_\pi^N(k)$ is the number of objects that change state during one time step, we have $W(k+1) \leq \sqrt{2}N^{-1}\Delta_\pi^N(k)$. Using Assumption (A1), this shows that $\mathbb{E}\left(W(k+1) \middle| \mathcal{F}_k\right) \leq \sqrt{2}I_1(N)$ and $\mathbb{E}\left(W(k+1)^2 \middle| \mathcal{F}_k\right) \leq 2I(N)I_2(N)$. Therefore applying Lemma 3 with $n = T/I(N)$ shows that for any $b_1 > 0$:

$$\mathbb{P}(\Omega_1) \quad \leq \quad \frac{2T}{b_1^2}\left[I_2(N) + \frac{I_1(N)^2}{I(N)^2}(T + I(N))\right] \tag{33}$$

Moreover, we show in Lemma 4 that:

$$\mathbb{P}(\Omega_2) \leq 2S^2 \frac{T}{b_2^2}\left[2I_2(N) + I(N)\left[(I_0(N) + L_2)\right]^2\right] \tag{34}$$

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

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

By Grönwall's lemma: $\left\|\hat{M}_\pi^N(t) - \phi_t(m_0, A_\pi^N)\right\|$ is less than

$$\left[\left\|M_\pi^N(0) - m_0\right\| + \epsilon + I_0(N)T\right] e^{L_1 t} \tag{35}$$

and this is true for all $\omega \in \Omega \setminus (\Omega_1 \cup \Omega_2)$. We apply the union bound $\mathbb{P}(\Omega_1 \cup \Omega_2) \le \mathbb{P}(\Omega_1) + \mathbb{P}(\Omega_2)$ 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 $k \in \mathbb{N}$: $\mathbb{E}\left(W_{k+1} \middle| \mathcal{F}_k\right) \le \alpha$ and $\mathbb{E}\left(W_{k+1}^2 \middle| \mathcal{F}_k\right) \le \beta$. Then for all $n \in \mathbb{N}$ and $b > 0$:*

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

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

$$\mathbb{E}\left(Y_n^2\right) \le n\beta + n(n+1)\alpha^2 \tag{36}$$

Now, because $W_{n+1} \ge 0$:

$$\begin{aligned}\mathbb{E}\left(Y_{n+1}^2 \middle| \mathcal{F}_n\right) &\ge \left(\mathbb{E}\left(Y_{n+1} \middle| \mathcal{F}_n\right)\right)^2 \\ &= \left(Y_n + \mathbb{E}\left(W_{n+1} \middle| \mathcal{F}_n\right)\right)^2 \ge Y_n^2\end{aligned}$$

thus Y_n^2 is a non-negative sub-martingale and by Kolmogorov's inequality:

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

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

Lemma 4. *Define Z_π^N as in Eq.(32). For all $N \ge 2$, $b > 0$, $T > 0$ and all policy π:*

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

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

$$L_k \stackrel{\text{def}}{=} \langle h, Z_\pi^N(k)\rangle.$$

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

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

By Assumption (A2):

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

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

$$\begin{aligned}&\|h\|^2\left[N^{-2}2\mathbb{E}\left(\Delta_\pi^N(k)^2 \middle| M_\pi^N(k)\right) + \left[(I_0(N) + L_2)\, I(N)\right]^2\right] \\ &\le \|h\|^2\left[2I(N)I_2(N) + \left[(I_0(N) + L_2)\, I(N)\right]^2\right]\end{aligned}$$

Applying Kolmogorov's inequality for martingales, we obtain

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

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

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

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

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

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

D. Proof of Theorem 6

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

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

Since $H^N = \lfloor T/N \rfloor$, the latter term is bounded by $I(N)\|r\|_\infty$. Moreover, by assumption (A1) on $\Delta_\pi^N(H^N)$, we have $\left|r_f(m_{A_\pi^N}(H^N I(N))) - r_f(m_{A_\pi^N}(T))\right| \le K_r \left\|m_{A_\pi^N}(H^N I(N)) - m_{A_\pi^N}(T)\right\| \le K_r\sqrt{2}I_1(N)$.

Let $\epsilon > 0$ and $\Omega_0 = \Omega_1 \cup \Omega_2$ where Ω_1, Ω_2 are as in the proof of Theorem 5. 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):

$$\begin{aligned}
&\left|V_\pi^N(M^N(0)) - \mathbb{E}\left[v_{A_\pi^N}(m_0)\right]\right| \\
&\leq I(N)\left\|r\right\|_\infty + K_r\sqrt{2}I_1(N) + \frac{2(T+1)\left\|r\right\|_\infty J(N,T)}{\epsilon^2} \\
&+ K_r\mathbb{E}\Big(1_{\omega\notin\Omega_0}\Big(\int_0^T \left\|\bar{M}_\pi^N(s) - m_{A_\pi^N}(s)\right\| ds \\
&+ r_f(M_\pi^N(H^N)) - r_f(m(H^N I(N)))\Big)\Big).
\end{aligned}$$

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

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

where

$$\begin{aligned}
B_\epsilon(N,\delta) &\stackrel{\text{def}}{=} I(N)\left\|r\right\|_\infty + K_r\sqrt{2}I_1(N) + \frac{K_rI(N)}{2L_1}\epsilon \\
&+ \frac{2(T+1)\left\|r\right\|_\infty J(N,T)}{\epsilon^2} \\
&+ K_r\left(\delta + I_0(N)T + \epsilon\right)\left(\frac{e^{L_1T}-1}{L_1} + e^{L_1T}\right)
\end{aligned}$$

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

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

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 $B(N,\delta)$.

E. 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 5, on the set $\Omega \setminus (\Omega_1 \cup \Omega_2)$, for every $t \in [0;T]$:

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

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

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

where $d_1 \stackrel{\text{def}}{=} \int_0^T \left\|\alpha(t) - \bar{\alpha}^N(t)\right\| dt$. Therefore, using Grönwall's inequality, we have $u(t) \leq Kd_1e^{KT}$. By Lemma 5, this shows Eq.(13). The rest of the proof is as for Theorem 6.

Lemma 5. *If α is a piecewise Lipschitz continuous action function on $[0;T]$, of constant K_α, and with at most p discontinuity points, then $\int_0^T d(\alpha(t),\bar{\alpha}^N(t))dt$ is bounded by*

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

Proof of lemma 5: 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 $[iI(N),(i+1)I(N))$:

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

If α has no discontinuity point on $[iI(N),(i+1)I(N))$, then $\int_{iI(N)}^{(i+1)I(N)} d(\alpha(s),\alpha(iI(N)))ds \leq \int_0^{I(N)} K_\alpha s ds \leq \frac{K_\alpha}{2}I(N)^2$. If α has one or more discontinuity points on $[iI(N),(i+1)I(N))$, then $\int_{iI(N)}^{(i+1)I(N)} d(\alpha(s)\alpha(iI(N)))ds \leq 2\left\|\alpha\right\|_\infty I(N)$. There are at most $\min(1/I(N),p)$ intervals $[iI(N),(i+1)I(N)]$ that have discontinuity points which shows that

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

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

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

$$\limsup_{N\to\infty} V_*^N(M^N(0)) \leq v_*(m_0) \leq \liminf_{N\to\infty} V_*^N(M^N(0)) \quad (37)$$

Let $\epsilon > 0$ and α be an action function such that $v_\alpha(m_0) \geq v_*(m_0) - \epsilon$ (such an action is called $\epsilon-$optimal). Theorem 3 shows that $\lim_{N\to\infty} V_\alpha^N(M^N(0)) = v_\alpha(m_0) \geq v_*(m_0) - \epsilon$ a.s. This shows that $\liminf_{N\to\infty} V_*^N(M^N(0)) \geq \lim_{N\to\infty} V_\alpha^N(M^N(0)) \geq v_*(m_0) - \epsilon$; this holds for every $\epsilon > 0$ thus $\liminf_{N\to\infty} V_*^N(M^N(0)) \geq v_*(m_0)$ 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 π^N such that $V_*^N(M^N(0)) \leq V_{\pi^N}^N(M^N(0)) + \epsilon$. By Th. 6, $V_{\pi^N}^N(M^N(0)) \leq \mathbb{E}\left(v_{A_{\pi^N}^N}(m_0)\right) + B(N,\delta^N) \leq v_*(m_0) + B(N,\delta^N)$ where $\delta^N \stackrel{\text{def}}{=} \left\|M^N(0) - m_0\right\|$. Thus $V_*^N(M^N(0)) \leq v_*(m_0) + B(N,\delta^N) + \epsilon$. If further $\delta^N \to 0$ a.s. it follows that $\limsup_{N\to\infty} V_*^N(M^N(0)) \leq v_*(m_0) + \epsilon$ a.s. for every $\epsilon > 0$, thus $\limsup_{N\to\infty} V_*^N(M^N(0)) \leq v_*(m_0)$ a.s.

VI. Conclusion and Perspectives

There are several natural questions arising from this work. One concerns the convergence of optimal policies. Optimal policies π_*^N of a stochastic systems with N objects may not be unique, they may also exhibit thresholds and therefore be discontinuous. This implies that $M_{\pi_*^N}^N$ 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 π can be decomposed into the value over horizon T plus a term less than ε:

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

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

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

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

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

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

[5] V. Berten and B. Gaujal. Grid brokering for batch allocation using indexes. In Springer, editor, *Network Control and Optimization*, volume 4465 of *LNCS*, 2007.

[6] W. Chen, D. Huang, A.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 *Proceedings of the 48th IEEE Conference on Decision and Control. CDC/CCC 2009.*, pages 3575–3580. IEEE, 2009.

[7] R. Cole. Initial studies on worm propagation in manets for future army combat systems. Technical report, Pentagon Reports, 2004.

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

[9] N. Gast and B. Gaujal. Mean field limit of non-smooth systems and differential inclusions. *ACM SIGMETRICS Performance Evaluation Review*, 38(2):30–32, 2010.

[10] N. Gast and B. Gaujal. A mean field approach for optimization in discrete time. *Discrete Event Dynamic Systems*, 21:63–101, 2011.

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

[12] M. Huang, P. E. Caines, and R. P. Malhame. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Com. Inform. Systems*, 6:221–252, 2006.

[13] M. Huang, P. E. Caines, and R. P. Malhame. Nash certainty equivalence in large population stochastic dynamic games: connections with the physics of interacting particle systems. In *45th IEEE Conference on Decision and Control*, pages 4921–4926, San Diego, 2006.

[14] M. Huang, P. E. Caines, and R. P. Malhame. Large-population cost-coupled lqg problems with nonuniform agents: individual-mass behavior and decentralized ε-nash equilibria. *IEEE Transactions Automatic Control*, 52, 2007.

[15] M.H.R. Khouzani, S. Sarkar, and E. Altman. Maximum damage malware attack in mobile wireless networks. In *IEEE Infocom*, San Diego, 2010.

[16] D. Kondo, B. Javadi, P. Malecot, F. Cappello, and D. Anderson. Cost-benefit analysis of cloud computing versus desktop grids. In *18th International Heterogeneity in Computing Workshop*, Rome, 2009.

[17] T. Kurtz. Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of Applied Probability*, 7:49–58, 1970.

[18] J.-M. Lasry and P.-L. Lions. Mean field games. *Japan. Journal Math.*, 2007.

[19] J.Y. Le Boudec, D. McDonald, and J. Mundinger. A generic mean field convergence result for systems of interacting objects. In *Proceedings of QEST 2007*, pages 3–18, 2007.

[20] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res.*, 24(2):292–305, 1999.

[21] M.L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

[22] S. Tanachaiwiwat and A. Helmy. Vaccine: War of the worms in wired and wireless networks. In *IEEE INFOCOM*, 2006.

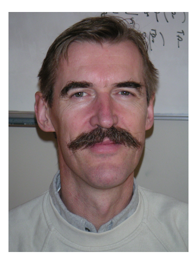
[23] H. Tembine, J.-Y. Le Boudec, R. El-Azouzi, and E. Altman. Mean field asymptotic of markov decision evolutionary games and teams. In *Gamenets*, 2009.

[24] H. Tembine, P. Vilanova, and M. Debbah. Noisy mean field stochastic games with network applications. Technical Report 342-P-11867-83, Supelec, 2010.

[25] 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.

Nicolas Gast is currently a post-doctoral fellow at EPFL. He graduated from École 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 École 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 millenium 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 Sigcomm, Sigmetrics, Infocom, Performance Evaluation and ACM/IEEE Transactions on Networking.