INRIA

A Mean Field Approach for Optimization in Discrete Time

Nicolas Gast — Bruno Gaujal

N° 6877 — version 4

Thème NUM

R apport de recherche

A Mean Field Approach for Optimization in Discrete Time

Nicolas Gast , Bruno Gaujal

Abstract: This paper investigates the limit behavior of Markov decision processes made of independent objects evolving in a common environment, when the number of objects (N) goes to infinity.

In the finite horizon case, we show that when the number of objects becomes large, the optimal cost of the system converges to the optimal cost of a mean field limit system that is deterministic. Convergence also holds for optimal policies.

We further provide bounds on the speed of convergence by proving second order results that resemble central limits theorems for the cost and the state of the Markov decision process, with explicit formulas for the limit. These bounds (of order $1/\sqrt(N)$) are proven to be tight in a numerical example.

One can even go further and get convergence of order $\sqrt{\log N}/N$ to a stochastic system composed of the mean field limit and a Gaussian term.

Our framework is applied to a brokering problem in grid computing. Several simulations with growing numbers of processors are reported. They compare the performance of the optimal policy of the limit system used in the finite case with classical policies by measuring its asymptotic gain.

Several extensions are also discussed. In particular, for infinite horizon cases with discounted costs, we show that first order limits hold and that second order results also hold as long as the discount factor is small enough. As for infinite horizon cases with non-discounted costs, examples show that even the first order limits may not hold.

Key-words: Markov Decision Processes, Mean Field, Optimization, Brokering

Une approche champ moyen pour l'optimisation en temps discret

Résumé : Cet article examine le comportement limite de processus de décision Markovien constitués de particules indépendantes évoluant dans un environnement commun, lorsque le nombre de particules tend vers l'infini.

Dans le cas où on s'intéresse à un coût à horizon fini ou dans le cas d'un coût à horizon infini avec décote, nous montrons que lorsque le nombre de particules devient grand, le coût optimal du système converge presque sûrement vers le coût optimal du système déterministe. La convergence vaut également pour les politiques optimales.

De plus, nous donnons un aperçu de la vitesse de convergence en prouvant plusieurs théorèmes de la limite centrale pour le coût ainsi que l'état moyen du processus en donnant des formules explicites pour la variance des lois gaussiennes limites. Un exemple numérique montre que ces bornes sont atteintes et que la vitesse de convergence est bien $1/\sqrt{N}$. Nous montrons aussi que l'on peut raffiner la convergence en s'intéressant au système déterministe plus un bruit gaussien. Dans ce cas on atteint une convergence en $\sqrt{\log(N)}/N$.

Le modèle est appliqué à un problème de gestionnaire de ressources dans des grilles de calcul. Nous comparons les performances de la politique optimale de la limite appliquée au système initiale avec plusieurs politiques classiques. Nous mesurons le gain asymptotique, ainsi que le seuil à partir duquel elle surpasse les politiques classiques.

Plusieurs extensions sont aussi présentées. Dans le cas d'un horizon infini et d'un coût actualisé, nous montrons un théorème de convergence au premier ordre ainsi qu'un théorème du second ordre si le facteur d'actualisation est suffisamment petit. Dans le cas d'un horizon infini et d'un coût non actualisé, nous présentons plusieurs exemples et contre-exemples qui montrent que la convergence n'a pas toujours lieu.

Mots-clés : Processus de décision Markovien, Champ moyen, Optimisation, Gestionnaire de ressource

1   Introduction

The general context of this paper is the optimization of the behavior of controlled Markovian systems, namely Markov Decision Processes composed by a large number of objects evolving in a common environment.

Consider a discrete time system made of N objects, N being large, that evolve randomly and independently (according to a transition probability kernel K). At each step, the state of each object changes according to a probability kernel, depending on the environment. The evolution of the environment only depends on the number of objects in each state. Furthermore, at each step, a central controller makes a decision that changes the transition probability kernel. The problem addressed in this paper is to study the limit behavior of such systems when N becomes large and the speed of convergence to the limit.

Several papers ([2], [6]) study the limit behavior of Markovian systems in the case of vanishing intensity (the expected number of transitions per time slot is $o(N)$). In these cases, the system converges to a differential system in continuous time. In our case, time remains discrete at the limit. This requires a different approach to construct the limit.

In [8], such discrete time systems are considered and the authors show that under certain conditions, as N grows large, a Markovian system made of N objects converges to a deterministic system. Since a Markov decision process can be seen as a family of Markovian kernels, the class of systems studied in [8] corresponds to the case where this family is reduced to a unique kernel and no decision can be made. Here, we show that under similar conditions as in [8], a Markov decision process also converges to a deterministic one. More precisely, we show that the optimal costs (as well as the corresponding states) converge almost surely to the optimal costs (resp. the corresponding states) of a deterministic system (the "optimal mean field"). These first order results are very similar to the results proved independently in [5]. Additionally, the quality of the deterministic approximation and the speed of convergence can also be estimated. For that, we provide second order results giving bounds on the speed of convergence under the form of central limit theorems for the state of the system as well as for the cost function.

On a practical point of view, this allows one to compute the optimal policy in a deterministic system which can often be done very efficiently, and then to use this policy in the original random system as a good approximation of the optimal policy, which cannot be computed efficiently because of the curse of dimensionality. This is illustrated by an application of our framework to optimal brokering in computational grids. We consider a set of multi-processor clusters (forming a computational grid, like EGEE [1]) and a set of users submitting tasks to be executed. A central broker assigns the tasks to the clusters (where tasks are buffered and served in a fifo order) and tries to minimize the average processing time of all tasks. Computing the optimal policy (solving the associated MDP) is known to be hard [15]. Numerical computations can only be carried up to a total of 10 processors and two users. However, our approach shows that when the number of processors per cluster and the number of users submitting tasks grow, the system converges to a mean field deterministic system. For this deterministic mean field system, the optimal brokering policy can be explicitly computed. Simulations reported in Section 4 show that, using this policy over a grid with a growing number of processors, makes performance converge to the optimal sojourn time in a deterministic system, as expected. Also, simulations show that this deterministic static policy outperforms classical dynamic policies such as Join the Shortest Queue, as soon as the total number of processors and users is over 50.

This paper is an extended version of [10]. Several Theorems (such as Theorem 1) are stronger than in the conference version, Others are new (such as Theorem 9). New extensions as well as new counter-examples (given in section 5) increase the set of systems that can be optimized using this approach on one hand and show the limitation of the validity of optimal mean field on the other.

2   Notations and definitions

The system is composed of N objects. Each object evolves in a finite state space $\mathcal{S} = \{1, \ldots, S\}$. The state of the nth object at time $t \in \mathbb{N}$ is denoted $X_n^N(t)$. We assume that the objects are distinguishable only through their state and that the dynamics of the system is homogeneous in N. This means that the behavior of the system only depends on in the proportion of objects in every state i. We call $M^N(t)$ the empirical measure of the collection of objects (X_n^N). $M^N(t)$ is a vector with S components and the ith component of $M^N(t)$ is

$$M_i^N(t) \overset{\text{def}}{=} N^{-1} \sum_{n=1}^{N} \mathbf{1}_{X_n^N(t)=i},$$

the proportion of objects in state i. The set of possible values for M^N is the set of probability measures p on $\{1 \ldots S\}$, such that $Np(i) \in \mathbb{N}$ for all $i \in \mathcal{S}$, denoted by $\mathcal{P}_N(S)$. For each N, $\mathcal{P}_N(S)$ is a finite set. When N goes to infinity, it converges to $\mathcal{P}(\mathcal{S})$ the set of probability measures on $\mathcal{S}$.

The system of objects evolves depending on their common environment. We call $C(t) \in \mathbb{R}^d$ the context of the environment. Its evolution depends on the empirical measure $M^N(t)$, itself at the previous time slot and the action a_t chosen by the controller (see below):

$$\mathrm{C}^N(t+1) = g(C^N(t), M^N(t+1), a_t),$$

where $g : \mathcal{P}_N(\mathcal{S}) \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ is a continuous function.

2.1   Actions and policies

At each time t, the system's state is $M^N \in \mathcal{P}_N(\mathcal{S})$. The decision maker may choose an action a from the set of possible actions $\mathcal{A}$. $\mathcal{A}$ is assumed to be a compact set (finite or infinite). In practical examples, $\mathcal{A}$ is often finite or a compact subset of $\mathbb{R}^k$. The action determines how the system will evolve. For an action $a \in \mathcal{A}$ and an environment $C \in \mathbb{R}^d$, we have a transition probability kernel $K(a, \mathrm{C})$ such that the probability that for any n, a object goes from state i to state the j is $K_{i,j}(a, \mathrm{C})$:

$$\mathbb{P}(X_n^N(t+1) = j | X_n^N(t) = i, a_t = a, \mathrm{C}^N(t) = \mathrm{C}) = K_{i,j}(a, \mathrm{C}).$$

The evolutions of objects are supposed to be independent once C is given. Moreover, we assume that $K_{i,j}(a, \mathrm{C})$ is continuous in a and C. The assumption of independence of the users is a rather common assumption in mean field models [8]. However other papers [2, 6] have shown that similar results can be obtained using asymptotic independence only (see [11] for results of this type).

Here, the focus is on Markov Decision Processes theory and on the computation of optimal policies. A policy $\pi = (\pi_1 \pi_2 \ldots)$ specifies the decision rules to be used at each time slot. A decision rule π_t is a procedure that provides an action at time t. In general, Π_t is a random measurable function that depends on the events $(M(0), C(0) \ldots M(t), C(t))$ but it can be shown that when the state space is finite and the action space is compact, then deterministic Markovian policies (*i.e.* that only depends deterministically on the current state) are dominant, *i.e.* are as good as general policies [16]. In what follows, we will only focus on them and a policy π will represent a sequence of functions $(\pi_t)_{t \geq 0}$ where each function $\pi_t : \mathcal{P}(\mathcal{S}) \times \mathbb{R}^d \to \mathcal{A}$ is deterministic.

For any policy π, we call the variables $M_\pi^N(t), C_\pi^N(t)$ will denote the state of the system at time t when the controller applies the policy π. $(M_\pi^N(t), C_\pi^N(t))_{t \geq 0}$ is a sequence of random variables on $\mathcal{P}_N(\mathcal{S}) \times \mathbb{R}^d$.

2.2   Reward functions

To each possible state $(M(t), C(t))$ of the system at time t, we associate a reward $r_t(M, C)$. The reward is assumed to be continuous in M and C. This function can be either seen as a reward – in that case the controller wants to maximize the reward –, or as a cost – in that case the goal of the

controller is to minimize this cost. In this paper, we will mainly focus on finite-horizon reward. Extensions to finite horizon reward (discounted and average reward) are discussed in Section 5.

In the finite-horizon case, the controller aims at maximizing the expectation of the sum of the rewards over all time $t < T$ plus a final reward that depends on the final state, $r_T(\mathrm{M}^N(t), \mathrm{C}^N(t))$. The expected reward of a policy π is:

$$V_\pi^N(M^N(0), C^N(0)) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T-1} r_t\left(M_\pi^N(t), C_\pi^N(t)\right) + r_T\left(M_\pi^N(T), C_\pi^N(T)\right)\right],$$

where the expectation is taken over all possible $(M_\pi^N(t), C_\pi^N(t))$ when the action chosen at time t is $\pi_t(M_\pi^N(t), C_\pi^N(t))$, for all t.

The goal of the controller is to find a policy that maximizes the expected reward and we call V_*^N the maximum expected reward.

$$V_*^N\left(M^N(0), C^N(0)\right) \stackrel{\text{def}}{=} \sup_\pi V_\pi^N\left(M^N(0), C^N(0)\right).$$

2.3 List of assumptions

Here is the list of the assumptions under which all our results will hold, together with some comments on their tightness and their degree of generality and applicability.

(A1) **Independence of the users, Markov system** – If at time t if the environment is C and the action is a, then the behavior of each object is independent of other objects and its evolution is Markovian with a kernel $K(a, \mathrm{C})$.

(A2) **Compact action set** – The set of action $\mathcal{A}$ is a compact metric space.

(A3) **Continuity of K, g, r** – the mappings $(\mathrm{C}, a) \mapsto K(a, \mathrm{C})$, $(\mathrm{C}, \mathrm{M}, a) \mapsto g(\mathrm{C}, \mathrm{M}, a)$ and $(\mathrm{M}, \mathrm{C}) \mapsto r_t(\mathrm{M}, \mathrm{C})$ are continuous, lipschitz continuous on all compact space.

(A4) **Almost sure initial state** – Almost surely, the initial measure $\mathrm{M}_0^N, \mathrm{C}_0^N$ converges to a deterministic value m_0, c_0. Moreover, there exists $B < \infty$ such that almost surely $\left\|\mathrm{C}_0^N\right\|_\infty \leq B$ where $\|C\|_\infty = \sup_i |C_i|$.

To simplify the notations, we choose the functions C and g not to depend on time. However as the proofs will be done for each time step, they also hold if the functions are time-dependent (in the finite horizon case).

Also, K, g and r do not to depend on N, while this is the case in most practical cases. Adding a uniform continuity assumption on these functions for all N will make all the proofs work the same.

Here are some comments on the uniform bound B on the initial condition (A4). In fact, as C_0^N converges almost surely, C_0^N is almost surely bounded. Here we had a bound B which is uniform on all events in order to be sure that the variable C_0^N is dominated by an integrable function. As g is continuous and the sets $\mathcal{A}$ and $\mathcal{P}(\mathcal{S})$ are compact, this shows that for all t, there exists $B_t < \infty$ such that

$$\left\|\mathrm{C}^N(t)\right\|_\infty \leq B_t. \tag{1}$$

Finally, in the Markov decision process literature, the reward function often depends on the action taken. To simplify the notations, we choose to take the reward independent of the action but again the proofs are the same in that case.

3 Finite time convergence and optimal policy

In this section, we focus on optimizing the finite horizon reward. T is fixed throughout all this section and the aim of the controller is to find a policy to maximize:

$$V_\pi^N(M^N, C^N) = \mathbb{E}\left[\sum_{t=1}^{T} r\left(M_\pi^N(t), C_\pi^N(t)\right)\right].$$

The infinite horizon case will be treated in Section 5.2.

This section contains the main results of this paper. There are four main results. Theorem 1 states the convergence of the controlled system to a deterministic limit. Next, we show that the optimal reward for the limit is asymptotically optimal as the size of the system grow (Theorem 5) and we characterize the speed of convergence towards this limit (Theorem 7) which is basically of order $1/\sqrt{N}$. Finally (Theorem 9) shows that a Gaussian approximation of the deterministic limit system leads to a better error of order $\frac{\sqrt{\log(N)}}{N}$.

3.1   Controlled mean field

Let $a = a_0, a_1 \ldots$ be a sequence of actions. We define the deterministic variables $m_a(t), c_a(t)$ starting in $m_a(0), c_a(0) \stackrel{\text{def}}{=} m(0), c(0) \in \mathcal{P}(\mathcal{S}) \times \mathbb{R}^d$ by induction on t:

$$\begin{array}{rcl} m_a(t+1) & = & m_a(t) K(a, c_a(t)) \\ c_a(t+1) & = & g\left(c_a(t), m_a(t+1), a_t\right). \end{array} \tag{2}$$

Here, $(m_a(t), c_a(t))$ corresponds to a deterministic approximation of the stochastic system (M^N, C^N) assuming that instead of having a probability K_{ij} for an object to go from state i to state j, there is a proportion K_{ij} of objects in state i moves to state j.

Let π be a policy and consider a realization of the sequence $(M^N(t), C^N(t))$. At time t, a controller that applies the policy π, will apply the action $A_\pi^N(t) \stackrel{\text{def}}{=} \pi(M_\pi^N(t), C_\pi^N(t))$. The actions $A_\pi^N(t)$ form a random sequence depending on the sequence $(M_\pi^N(t), C_\pi^N(t))$. To this random sequence, corresponds a deterministic approximation of M^N, C^N, namely $m_{A_\pi^N}(t)$ defined by Equation (2). The quantity $m_{A_\pi^N}(t)$ is a random variable depending on the sequence A_π^N (and is deterministic once A_π^N is fixed).

The following theorem is the main result of convergence, showing that as N grows, the gap between the stochastic system M_π^N, C_π^N and its deterministic limit $m_{A_\pi^N}, c_{A_\pi^N}$ vanishes (in probability) with a bound only depending on the initial condition.

Theorem 1 (Controlled mean field)**.** *Under assumptions (A1,A3), and if the controller applies the policy π, then there exists a sequence of functions $\mathcal{E}_t(\epsilon, x)$ such that $\lim_{\epsilon \to 0, x \to 0} \mathcal{E}_t(\epsilon, x) = 0$ and for all t:*

$$\mathcal{P}\left(\left\|(M_\pi^N(t), C_\pi^N(t)) - (m_{A_\pi^N}(t), c_{A_\pi^N}(t))\right\| \geq \mathcal{E}_t(\epsilon, \left\|M^N(0), C^N(0) - m(0), c(0)\right\|)\right)$$
$$\leq 2tS^2 \exp(-2N\epsilon).$$

Proof. The proof is done by induction on t. We show that at each time step, we stay close to the deterministic approximation with high probability. A detailed proof is given in Appendix A. $\square$

Assuming that the initial condition converges almost surely to $m(0), c(0)$, we can refined the convergence in law into an almost sure convergence:

Corollary 2. *Under assumptions (A1,A3,A4),*

$$\left\|(M_\pi^N(t), C_\pi^N(t)) - (m_\pi(t), c_\pi(t))\right\| \xrightarrow{\text{a.s.}} 0.$$

Proof. This proof is a direct corollary of Theorem 1 and the Borel-Cantelli Lemma. $\square$

3.2   Optimal mean field

Using the same notation and hypothesis as in the previous section, we define the reward of the deterministic system starting at $m(0), c(0)$ under the sequence of action a:

$$v_a(m(0), c(0)) \stackrel{\text{def}}{=} \sum_{t=1}^{T} r_t(m_a(t), c_a(t)).$$

If for any t, the action taken at instant t is fixed equal to a_t, we say that the controller applies the policy a. a can be viewed as a policy independent of the state M^N, C^N and $M_a^N(t), C_a^N(t)$ denotes the state of the system when applying the policy a. According to Corollary 2, the stochastic system $M_a^N(t), C_a^N(t)$ converges almost surely to $m_a(t), c_a(t)$. Since the reward at time t is continuous, this means that the finite-horizon expected reward converges as N grows large:

Lemma 3. (CONVERGENCE OF THE REWARD) *Under assumptions (A1,A3,A4), if the controller takes actions $a = (a_0, a_1 \dots)$, the finite-horizon expected reward of the stochastic system converges to the finite-horizon reward of the deterministic system when initial conditions converge. If $(M^N(0), C^N(0)) \to (m(0), c(0))$a.s. then*

$$\lim_{N \to \infty} V_a^N \left(M^N(0), C^N(0) \right) = v_a(m(0), c(0)) \quad \text{a.s..}$$

Proof. For all t, $(M_a^N(t), C_a^N(t))$ converges in probability to $(m_a(t), c_a(t))$. Since the reward at time t is continuous in (M, C), then $r_t(M_a^N(t), C_a^N(t))$ converges in probability to $r_t(m_a(t), c_a(t))$. Moreover, as (M, C) are bounded (see Equation (1)), the $\mathbb{E}[r_t(M_a^N(t), C_a^N(t))]$ goes to $r_t(m_a(t), c_a(t))$ which concludes the proof. □ □

The previous lemma can also be deduced from the following proposition.

Proposition 4 (Uniform convergence of reward). *Under assumptions (A1)-(A4), there exists a function $A(N, \epsilon)$ such that:*

- $\lim_{N \to \infty, \epsilon \to 0} \mathcal{E}(N, \epsilon) = 0$,

- *for all policy π:*

$$\left| V_\pi^N \left(M^N(0), C^N(0) \right) - \mathbb{E} \left[v_{A_\pi^N} \left(m_{A_\pi^N}(0), c_{A_\pi^N}(0) \right) \right] \right| \leq \mathcal{E}(N, \epsilon_0^N),$$

where $\epsilon_0^N \overset{\text{def}}{=} \left\| (M^N(0), C^N(0)) - (m(0), c(0)) \right\|$ and the expectation is taken on all possible values of A_π^N.

Proof.

$$\left| V_\pi^N \left(M^N(0), C^N(0) \right) - \mathbb{E} \left[v_{A_\pi^N} \left(m_{A_\pi^N}(0), c_{A_\pi^N}(0) \right) \right] \right| \tag{3}$$

$$= \left| \mathbb{E} \left[\sum_{t=1}^{T} r_t(M_\pi^N(t), C_\pi^N(t)) - r_t(m_{A_\pi^N}(0), c_{A_\pi^N}(0)) \right] \right|$$

$$\leq L_r \mathbb{E} \max_{t \leq T} \left\| (M_\pi^N(t), C_\pi^N(t)) - (m_{A_\pi^N}(0), c_{A_\pi^N}(0)) \right\|$$

where L_r is a Lipschitz constant of the function r.

According to Theorem 1, $\max_{t \leq T} \left\| (M_\pi^N(t), C_\pi^N(t)) - (m_{A_\pi^N}(0), c_{A_\pi^N}(0)) \right\| \geq \max_{t \leq T} \mathcal{E}_t(\epsilon, \epsilon_0^N)$ with probability $T^2 S^2 \exp(-2\epsilon N)$, where $\epsilon_0^N \overset{\text{def}}{=} \left\| (M^N(0), C^N(0)) - (m(0), c(0)) \right\|$. Computing the expectation on the events such that this is verified and the others, we get:

$$(3) \leq L_r \max_{t \leq T} \mathcal{E}_t(\epsilon, \epsilon_0^N) \left(1 - T^2 S^2 \exp(-2\epsilon N) \right) + DT^2 S^2 \exp(-2\epsilon N),$$

where $D \overset{\text{def}}{=} \sup_{x \in B} \|x\|$ where B is a bounded set such that $\mathcal{P} \left((M_\pi^N(t), C_\pi^N(t)) \in B \right) = 1$ (see (1)).

Let us define $\mathcal{E}(N, \epsilon_0^N) \overset{\text{def}}{=} \inf_{\epsilon > 0} L_r \max_{t \leq T} \mathcal{E}_t(\epsilon, \epsilon_0^N) \left(1 - T^2 S^2 \exp(-2\epsilon N) \right) + DT^2 S^2 \exp(-2\epsilon N)$. The function $\mathcal{E}(.,.)$ satisfies the two requirements of the theorem. □ □

Now, let us consider the problem of convergence of the reward under the optimal strategy of the controller. First, it should be clear that the optimal strategy exists for the limit system. Indeed, the limit system being deterministic, starting at state $(m(0), c(0))$, one only needs to know

the actions to take for all $(m(t), c(t))$ to compute the reward. The optimal policy is deterministic and $v_*(m(0), c(0)) \stackrel{\text{def}}{=} \sup_{a \in \mathcal{A}^T} \{v_a(m(0), c(0))\}$. Since the action set is compact, this supremum is a maximum: there exists a sequence of actions $a^* = a_0^* a_1^* \ldots$ – depending on $m(0), c(0)$ – such that $v_*(m(0), c(0)) = v_{a^*}(m(0), c(0))$. Such a sequence is not unique and in many cases there are multiple optimal action sequences. In the following, a^* design one of them and will be called the sequence of *optimal limit actions*.

Theorem 5. (CONVERGENCE OF THE OPTIMAL REWARD) *Under assumptions (A1, A2, A3, A4), if $\left\| (M^N(0), C^N(0)) - (m(0), c(0)) \right\|$ goes to 0 when N goes to infinity, the optimal reward of the stochastic system converges to the optimal reward of the deterministic limit system:*

$$\lim_{N \to \infty} V_*^N \left(M^N(0), C^N(0) \right) = \lim_{N \to \infty} V_{a^*}^N \left(M_0^N, C_0^N \right) = v_*(m(0), c(0)), \quad a.s.$$

In words, this theorem states two important results. Firstly, as N goes to infinity, the reward of the stochastic system goes to the reward of its deterministic limit. Secondly, the reward of the optimal policy under full information $V_*^N \left(M^N(0), C^N(0) \right)$ is asymptotically the same as the reward obtained when applying to the stochastic system a sequence of optimal actions of the deterministic limit, both being equal to the optimal reward of the limit deterministic system, $v_T^*(m_0, c_0)$.

Proof. Let a^* be a sequence of optimal actions for the deterministic limit starting at $m(0), c(0)$. Lemma 3 shows that $\lim_{N \to \infty} V_{a^*}^N \left(M^N(0), C^N(0) \right) = v_{a^*}(m(0), c(0)) = v_*(m(0), c(0))$. This shows that

$$\limsup_{N \to \infty} V_*^N \left(M^N(0), C^N(0) \right) \geq \limsup_{N \to \infty} V_{a^*}^N \left(M^N(0), C^N(0) \right) = v_*(m(0), c(0))$$

Reciprocally, let π_*^N be an optimal policy for the stochastic system and $A_{\pi_*^N}^N$ the (random) sequence of action $A_{\pi_*^N}^N(t) \stackrel{\text{def}}{=} \pi_*^N(M^N(t), C^N(t))$. This policy is suboptimal for the deterministic limit: $v_*(m(0), c(0)) \geq v_{A_{\pi_*^N}^N}(m(0), c(0))$. Using Proposition 4,

$$
\begin{aligned}
V_*^N \left(M^N(0), C^N(0) \right) = V_{\pi_*^N}^N \left(M^N(0), C^N(0) \right) & \leq & v_{A_{\pi_*^N}^N}(m(0), c(0)) + \mathcal{E}(N, \epsilon_0^N) \\
& \leq & v_*(m(0), c(0)) + \mathcal{E}(N, \epsilon_0^N)
\end{aligned}
$$

where $\mathcal{E}(.,.)$ is defined as in Proposition 4 and $\epsilon_0^N \stackrel{\text{def}}{=} \left\| (M^N(0), C^N(0)) - (m(0), c(0)) \right\|$.

As N goes to infinity, $\mathcal{E}(N, \epsilon_0^N)$ goes to 0.

This shows that $\liminf_{N \to \infty} V_*^N \left(M^N(0), C^N(0) \right) \geq v_*(m(0), c(0))$ and concludes the proof of the theorem. $\square$ $\square$

This result has several practical consequences. Recall that the sequence of actions $a_0^* \ldots a_{T-1}^*$ is a sequence of optimal actions in the limit case, *i.e.* such that $v_{a^*}(m, c) = v_*(m, c)$. This result proves that in the limit case, the optimal policy does not depend on the state of the system. This also shows that incomplete information policies are as good as complete information policies in the limit. However, the state $(M^N(t), C^N(t))$ is not deterministic and on one trajectory of the system, it could be quite far from its deterministic limit $(m(t), c(t))$. Let us also define the policy $\mu_t^*(m(t), c(t))$ which is optimal for the deterministic system starting at time t in state $m(t), c(t)$. The least we can say is that this strategy is also asymptotically optimal, that is for any initial state $M^N(0), C^N(0)$:

$$\lim_{N \to \infty} V_{\mu^*}^N \left(M^N(0), C^N(0) \right) = \lim_{N \to \infty} V_{a^*}^N \left(M^N(0), C^N(0) \right) = \lim_{N \to \infty} V_*^N(m(0), c(0)). \tag{4}$$

In practical situations, using this policy in the original system will decrease the risk of being far from the optimal state. On the other hand, using this policy has some drawbacks. The first one is that the complexity of computing the optimal policy for all states can be much larger than the

complexity of computing a^*. Another one is that the system becomes very sensitive to random perturbations and therefore harder to analyze: the policy μ^* is not necessarily continuous and M_μ^N, C_μ^N may not have a limit. In Section 4, a comparison between the performances of a^* and μ^* is provided over an example and we see that the performance of μ^* is much better, especially for low values of N.

3.3 Second order results

In this part we give bounds on the gap between the stochastic system and its deterministic limit. This result provides estimates on the speed of convergence to the mean field limit. These theorems have a flavor of central limit theorems in the sense that the convergence speed towards the limit is of order $1/\sqrt{N}$. This section contains two main results:

The first one is that when the control action sequence is fixed, the gap to the mean field limit decreases as the inverse square root of the number of objects. The second result states that the gap between the optimal reward for the finite system and the optimal reward for the limit system also decreases as fast as $1/\sqrt{N}$.

In all this section, $\leq_{\mathrm{st}}$ denotes the stochastic order: if X, Y are two real random variables, we say that $X \leq_{\mathrm{st}} Y$ if for all $x \in \mathbb{R}$, $P(X > x) \leq P(Y > x)$. For a vector $X \in \mathbb{R}^d$, we denote $\|X\|$ the L^1 norm of X: $\|X\| = \sum_{i=1}^d |X_i|$ where X_i is the ith coordinate of X.

Proposition 6. *Under assumptions (A1,A2,A3,A4) and if there exists a sequence δ_0^N such that $\lim_{N \to \infty} \mathbb{E} \left|\delta_0^N\right| = 0$ and $\left\|\left(M^N(0), C^N(0)\right) - (m(0), c(0))\right\| \leq_{\mathrm{st}} \delta_0^N$, then there exists a sequence $e(N) \to 0$ only dependent on the parameters of the system such that:*

$$\sqrt{N} \left\|\left(M_\pi^N(t), C_\pi^N(t)\right) - (m_\pi(t), c_\pi(t))\right\| \leq_{\mathrm{st}} \left(\beta_t |G| + \beta_t' \sqrt{N} \delta_0^N\right) \left(1 + \epsilon^N(t)\right) \tag{5}$$

where :

- *G is a Gaussian random variable of mean 0 and variance 1.*

- *$\beta_t, \beta_t' > 0$ is a sequence of constant with $\beta_t = 0$, $\beta_t' = 1$ and*

$$\begin{aligned} \beta_{t+1} &= \beta_t L_g (1 + \max(L_g, L_K)) + (1 + L_g) \\ \beta_{t+1}' &= \beta_t' L_g (1 + \max(L_g, L_K)) \end{aligned}$$

- *$\epsilon^N(t)$ is a sequence of random variables such that*

$$\mathbb{E} \left|\epsilon^N(t)\right| \leq e(N)$$

These functions only depend on the parameters of the system and not on the values of M^N, C^N or π^N.

Proof. See appendix B. □

Theorem 7. *Under assumptions (A1,A2,A3,A4), there exist constants d and e such that if $\epsilon_0^N \overset{\text{def}}{=} \left\|M^N(0), C^N(0) - m(0), c(0)\right\|$*

- *For any policy π:*

$$\sqrt{N} \left(V_\pi^N \left(M^N(0), C^N(0)\right) - \mathbb{E}\left[v_{A_\pi^n}^N (m(0), c(0))\right]\right) \leq d + e\epsilon_0^N.$$

-

$$\sqrt{N} \left(V_*^N \left(M^N(0), C^N(0)\right) - v_*^N (m(0), c(0))\right) \leq d + e\epsilon_0^N.$$

This theorem is the main result of this section. The previous result (Theorem 5) says that $\limsup_{N\to\infty} V_T^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) = \limsup_{N\to\infty} V_{a_0^*\dots a_{T-1}^*}^N(\mathrm{M}_0^N, \mathrm{C}_0^N) = v_{t\dots T}^*(m_0, c_0)$. This new theorem says that both the gap between the cost under the any policy for the original and the limit system and the gap between the optimal costs for both systems are two random variables that decrease to 0 with speed $\sqrt{N}$ and have Gaussian laws.

Proof. For any policy π, the expected reward of the stochastic system and the expected reward of the deterministic limit under actions A_π^N are:

$$V_\pi^N\left(M^N(0), C^N(0)\right) = \sum_{t=1}^T \mathbb{E}\left[r\left(M_\pi^N(t), C_\pi^N(t)\right)\right]$$

$$\mathbb{E}\left[v_{A_\pi^n}^N\left(m(0), c(0)\right)\right] = \sum_{t=1}^T \mathbb{E}\left[r\left(m_{A_\pi^N}(t), c_{A_\pi^N}(t)\right)\right].$$

The first part of the theorem comes directly corollary of Proposition 6 and the fact that if X and Y are two stochastic variables with $\|X - Y\| \leq_{\mathrm{st}} Z$ and f a real Lipschitz function of constant L, then

$$\mathbb{E}\left|f(X) - f(Y)\right| \leq L\mathbb{E}\|X - Y\| \leq \mathbb{E}|Z|.$$

The second part is proved as Theorem 5 by bounding both part of the inequality (for readability, the following equation is written suppressing the dependence in $M^N(0), C^N(0)$ and $m(0), c(0)$).

$$V_*^N = \sup V_\pi^N = V_{\pi_*^N}^N \leq \mathbb{E}v_{A_{\pi_*^N}^N} + (d + e\epsilon_0)/\sqrt{N} \leq v_* + (d + e\epsilon_0)/\sqrt{N},$$

where $\epsilon_0 \stackrel{\mathrm{def}}{=} \left\|M^N(0), C^N(0) - m(0), c(0)\right\|$ and π_*^N is the optimal policy of the stochastic system of size N. The first inequality comes from the first part of this theorem, the second from the fact that v_* is the optimal reward of the deterministic system.

The other inequality is similar:

$$v_* = v_{a^*} \leq V_{a^*}^N + (d + e\epsilon_0)/\sqrt{N} \leq V_*^N + (d + e\epsilon_0)/\sqrt{N},$$

where a^* is the sequence of optimal actions of the deterministic system. $\square$

In the case where the parameter are differentiable and not just only Lipschitz, the Proposition 6 can be refined into Theorem 8 which is a central limit theorem for the states.

(A5) **Continuous differentiability** – For all t and all $i, j \in \mathcal{S}$, all functions g, K_{ij} and r_t are continuously differentiable.

(A4-bis) **Initial Gaussian variable** – There exists a Gaussian vector G_0 of mean 0 with covariance Γ_0 such that the vector $\sqrt{N}(M^N(0), C^N(0) - m(0), c(0))$ (with $S+d$ components) converges almost surely to G_0.

This differentiability condition is slightly stronger than the Lipschitz condition and is indeed false in many cases because of boundary condition. The initial state condition is slightly stronger that (A4) but remains very natural. For example, if the initial states of all objects are chosen randomly and independently of the others, then $\sqrt{N}((M^N(0), C^N(0)) - (m_0, c_0))$ converges in law to a Gaussian variable G. The fact that we assumed that the convergence holds almost surely rather that in law is just a technical matter: we can replace the variables $M^N(0), C^N(0)$ by random variables with the same law that converge almost surely.

Theorem 8. (Mean field central limit theorem) *Under assumption (A1,A2, A3,A4bis,A5), if the actions taken by the controller are $a_0 \dots a_{T-1}$, there exist Gaussian vectors of mean 0, $G_1 \dots G_{T-1}$ such that for every t:*

$$\sqrt{N}((\mathrm{M}^N(0), \mathrm{C}^N(0)) - (m(0), c(0)), \dots, (\mathrm{M}^N(t), \mathrm{C}^N(t)) - (m(t), c(t)))$$
$$\xrightarrow{\mathcal{L}} G_0, \dots, G_t. \tag{6}$$

Moreover if Γ_t is the covariance matrix of G_t, then:

$$\Gamma_{t+1} = \left[\begin{array}{c|c} P_t & F_t \\ \hline Q_t & H_t \end{array}\right]^{tr} \Gamma_t \left[\begin{array}{c|c} P_t & F_t \\ \hline Q_t & H_t \end{array}\right] + \left[\begin{array}{c|c} D_t & 0 \\ \hline 0 & 0 \end{array}\right], \tag{7}$$

where for all $1 \le i, j \le S$ and $1 \le k, \ell \le d$: $(P_t)_{ij} = K_{ij}(a_t, c(t))$,
$(Q_t)_{kj} = \sum_{i=1}^S m_i \frac{\partial K_{ij}}{\partial c_k}(a_t, c(t))$, $(F_t)_{ik} = \frac{\partial g_k}{\partial m_i}(m_{t+1}, c(t))$, $(H_t)_{k\ell} = \frac{\partial g_k}{\partial c_\ell}(m(t), c(t))$, $(D_t)_{jj} = \sum_{i=1}^n m_i(P_t)_{ij}(1 - (P_t)_{ij})$ and $(D_t)_{jk} = -\sum_{i=1}^n m_i(P_t)_{ij}(P_t)_{ik}$ $(j \neq k)$.

Proof. The proof is done by induction on t. We show that each time step:

- a new Gaussian error independent of the past is created by the Markovian evolution of the objects.

- Since all of the evolution parameters are differentiable, the Gaussian error of time t is scaled by a linear transformation.

The proof is detailed in Appendix C. $\qquad\square$

3.4 Beyond square-root convergence

We proved that as N grows, the system gets closer to a deterministic one: if at time t the system is in state $M^N(t)$, then at time $t+1$, the state of the system is close to $M^N(t)K(t)$. Moreover, we showed that the optimal policy for the deterministic limit is asymptotically optimal for the stochastic system as well and we give bounds for the speed of convergence. The mean field central limit theorem (Theorem 8) shows that $M^N(t+1) \approx M^N(t)K(t) + \frac{1}{\sqrt{N}}G(t)$. This should be a better approximation of the initial system. The purpose of this part is to show that this approximation is indeed better than the mean field, in the sense that it leads to an error for the cost of order $\frac{\sqrt{\log N}}{N}$ instead of $\frac{1}{\sqrt{N}}$.

For any policy π and any initial condition $M^N(0), C^N(0)$ of the original process, let us define a coupled process $\widetilde{M}_\pi^N(t), \widetilde{C}_\pi^N(t)$ as follows:

- $\widetilde{M}_\pi^N(0), \widetilde{C}_\pi^N(0) \stackrel{\text{def}}{=} M^N(0), C^N(0)$

- for $t \ge 0$:

$$\begin{aligned} \widetilde{M}_\pi^N(t+1) &\stackrel{\text{def}}{=} \widetilde{M}_\pi^N(t)K(A^N(t), \widetilde{C}_\pi^N(t)) + G(A^N(t), \widetilde{C}_\pi^N(t)) \\ \widetilde{C}_\pi^N(t+1) &\stackrel{\text{def}}{=} g(\widetilde{C}_\pi^N(t), \widetilde{M}_\pi^N(t+1), A^N(t)) \end{aligned}$$

where $A^N(t) \stackrel{\text{def}}{=} \pi(\widetilde{M}_\pi^N(t), \widetilde{C}_\pi^N(t))$ and $G(a, \widetilde{C}_s^N(t))$ is a Gaussian random variable independent of all $\widetilde{M}_\pi^N(t'), \widetilde{C}_\pi^N(t')$ for $t' < t$ corresponding to the error added by the random evolution of the objects between time t and time $t+1$. The covariance of $G(a, C)$ is a $S \times S$ matrix $D(a, C)$ where if we denote $P_{ij} \stackrel{\text{def}}{=} K_{ij}(a, C)$, then for all $j \neq k$:

$$D_{jj}(a, C) = \sum_{i=1}^n m_i P_{ij}(1 - P_{ij}) \quad \text{and} \quad D_{jk}(a, C) = -\sum_{i=1}^n m_i P_{ij} P_{ik}$$

In the following, the process $\widetilde{M}_{pi}^N(t), \widetilde{C}_{pi}^N(t)$ is called the *mean field Gaussian approximation* of $M_\pi^N(t), C_\pi^N(t)$. Similarly to the definition of $V_\pi^N(M^N(0), C^N(0))$, we define the expected reward of the mean field Gaussian approximation by:

$$W_\pi^N(M^N(0), C^N(0)) = \mathbb{E}\left[\sum_{t=1}^T r\left(\widetilde{M}_\pi^N(t), \widetilde{C}_\pi^N(t)\right)\right].$$

The optimal cost of the mean field Gaussian approximation is $W_*^N(M^N(0), C^N(0)) = \sup_\pi W_\pi^N(M^N(0), C^N(0))$. The following result shows that $\widetilde{M}_\pi^N(t), \widetilde{C}_\pi^N(t)$ is indeed a very accurate approximation of $M_\pi^N(t), C_\pi^N(t)$.
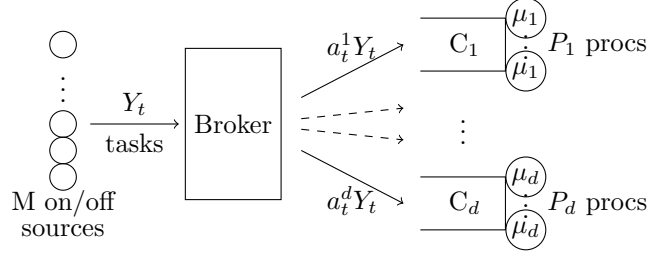
Figure 1: The routing system

Theorem 9. *Under assumptions (A1,A2,A3,A4), there exists a constant D independent of M^N, C^N such that for all sequence of action $a = a_1 \ldots a_T$:*

$$\left| V_a^N(M^N, C^N) - W_a^N(M^N, C^N) \right| \leq D \frac{\sqrt{\log(N)}}{N}.$$

$$\left| V_*^N(M^N, C^N) - W_*^N(M^N, C^N) \right| \leq D \frac{\sqrt{\log(N)}}{N}.$$

Proof. The proof is detailed in Appendix D.   □   □

4   Application to a brokering problem

To illustrate the usefulness of our framework, let us consider the following model of a brokering problem in computational grids. There are A application sources that send tasks into a grid system and a central broker routes all theses tasks into d clusters (seen as multi-queues) and tries to minimize the total waiting time of the tasks. A similar queuing model of a grid broker was used in [14, 3, 4].

Here, time is discrete and the A sources follow a discrete on/off model: for each source $j \in \{1 \ldots A\}$, let $(Y_t^j) \stackrel{\text{def}}{=} 1$ if the source is on (*i.e.* it sends a tasks between t and $t + 1$) and 0 if it is off. The total number of packets sent between t and $t + 1$ is $Y_t \stackrel{\text{def}}{=} \sum_j Y_t^j$. Each queue $i \in \{1 \ldots d\}$ is composed of P_i processors, and all of them work at speed μ_i when available. Each processor $j \in \{1 \ldots P_i\}$ of the queue i can be either *available* (in that case we set $X_t^{ij} \stackrel{\text{def}}{=} 1$) or *broken* (in that case $X_t^{ij} \stackrel{\text{def}}{=} 0$). The total number of processors available in the queue i between t and $t + 1$ is $X_t^i \stackrel{\text{def}}{=} \sum_j X_t^{ij}$ and we define B_t^i to be the total number of tasks waiting in the queue i at time t. At each time slot t, the broker (or controller) allocates the Y_t tasks to the d queues: it chooses an action $a_t \in \mathcal{P}(\{1 \ldots Y_t\}^d)$ and routes each Y_t packets in queue i with probability a_t^i. The system is represented figure 1. The number of tasks in the queue i (buffer size) evolves according to the following relation:

$$B_{t+1}^i = \left(B_t^i - \mu_i X_t^i + a_t^i Y_t \right)^+. \tag{8}$$

The cost that we want to minimize is the sum of the waiting times of the tasks. Between t and $t + 1$, there are $\sum_i B_i^t$ tasks waiting in the queue, therefore the cost at time t is $r_t(B) \stackrel{\text{def}}{=} \sum_i B_t^i$. As we consider a finite horizon, we should decide a cost for the remaining tasks in the queue. In our simulations, we choose $r_T(B) \stackrel{\text{def}}{=} \sum_i B_T^i$.

This problem can be viewed as a multidimensional restless bandit problem where computing the optimal policy for the broker is known to be a hard problem [19]. Here, indexability may help to compute near optimal policies by solving one MDP for each queue [19, 18]. However the complexity remains high when the number of processors in all the queues and the number of sources are large.

4.1 Mean field limit

This system can be modeled using the framework of objects evolving in a common environment.

- There are $N \stackrel{\text{def}}{=} A + \sum_{i=1}^{d} P_i$ "objects". Each object can either be a source (of type s) or a server (belonging to one of the queues, $q_1 \cdots q_d$), and can either be "on" or "off". Therefore, the possible states of one object is an element of $\mathcal{S} = \big\{ (x, e) | x \in \{s, q_1, \cdots, q_d\}, e \in \{\text{on}, \text{off}\} \big\}$. the population mix M is the proportion of sources in state on and the proportion of servers in state on, for each queue.

- The action of the controller are the routing choices of the broker: a_t^d is the probability that a task is sent to queue d at time t.

- The environment of the system depends on the vector $B_t = (B_{t_1} \ldots Bt_d)$, giving the number of tasks in queues $q_1, \ldots q_d$ at time t. The time evolution of the i-th component is

$$B_{t+1_i} = g_i(B_t, \mathrm{M}_{t+1}^N, a_t) \stackrel{\text{def}}{=} \left(B_{t_i} - \mu_i X_t^i + a_t^i Y_t \right)^+.$$

  The shared environment is represented by the context $C_t^N \stackrel{\text{def}}{=} (\frac{B_{t_1}}{N} \ldots \frac{B_{t_d}}{N})$.

- Here, the transition kernel can be time dependent but is independent of a and C. The probability of a object to go from a state $(x, e) \in \mathcal{S}$ to $(y, f) \in \mathcal{S}$ is 0 if $x \neq y$ (a source cannot become a server and vice-versa). If $x = y$ then $K_{(x,\text{on}),(x,\text{off})}(a, C)(t)$ as well as $K_{(x,\text{off}),(x,\text{on})}(a, C)(t)$ are arbitrary probabilities.

Here is how a system of size N is defined. A preliminary number of sources A_0 as well as a preliminary number P_i of servers per queue is given, totaling in N_0 objects. For any N, a system with N objects is composed of $\lfloor A_0 N / N_0 \rfloor$ (resp. $\lfloor P_i N / N_0 \rfloor$) objects that are sources (resp. servers in queue i). The remaining objects (to reach a total of N) are allocated randomly with a probability proportional to the fractional part of A/N_0 and $P_i N / N_0$ so that the mean number of objects that are sources is A/N_0 and the mean number of objects that are servers in queue i is $P_i N / N_0$. Then, each of these objects changes state over time according to the probabilities $K_{u,v}(a, C)(t)$. At time $t = 0$, a object is in state "on" with probability one half.

One can easily check that this system satisfies Assumptions (A1) to (A4) and therefore one can apply the convergence Theorem 5 that shows that if using the policies a^* or μ^*, when N goes to infinity the system converges to a deterministic system with optimal cost. An explicit computation of the policies a^* and μ^* is possible here and is postponed to Section 4.2. Also note that Assumption (A4-bis) on the convergence of the initial condition to a Gaussian variable is true since the random part of the initial state is bounded by $\frac{N_0}{N}$ and $\sqrt{N} \frac{N_0}{N}$ goes to 0 as N grows.

4.2 Optimal policy for the deterministic limit

As the evolution of the sources and of the processors does not depend on the environment, for all i, t, the quantities $\mu_i X_t^i$ and Y_t converge almost surely to deterministic values that we call x_t^i and y_t. If y_t^i is the number of packets distributed to the ith queue at time t, $c_{t+1}^i = (c_t^i + y_t^i - x_t^i)^+$. The deterministic optimization problem is to compute

$$\min_{y_1^1 \cdots y_T^d} \Big\{ \sum_{t=1}^{T} \sum_{i=1}^{d} c_t^i \text{ with } \begin{array}{l} c_{t+1}^i = (c_t^i + y_t^i - x_t^i)^+ \\ \sum_i y_t^i = y_t \end{array} \Big\}. \tag{9}$$

Let us call w_t^i the work done by the queue i at time t: $w_t^i = c_t^i - c_{t-1}^i + y_{t-1}^i$. The sum of the size of the queues at time t does not depend on with queue did the job but only on the quantity of work done:

$$\sum_{i=1}^{d} c_t^i = \sum_{i=1}^{d} c_0^i - \sum_{u \leq t, i} w_t^i$$

Therefore to minimize the total cost, we have to maximize the total work done by the queues. Using this fact, the optimal strategy can be computed by iteration of a greedy algorithm.

The principle of the algorithm is the following.

1. The processors in all queues, which are "on" at time t with a speed μ are seen as slots of size μ.

2. At each time t, y_t units of tasks have to be allocated. This is done in a greedy fashion by filling up the empty slots starting from time t. Once all slots at time t are full, slots at time $t + 1$ are considered and are filled up with the remaining volume of tasks, and so forth up to time T.

3. The remaining tasks that do not fit in the slots before T are allocated in an arbitrary fashion.

4.3 Numerical example

We consider a simple instance of the resource allocation problem with 5 queues. Initially, they have respectively $1, 2, 2, 3$ and 3 processors running at speed $.5, .1, .2, .3$ and $.4$ respectively. There are 3 initial sources. The transition matrices are time dependent and are chosen randomly before the execution of the algorithm – that is they are known for the computation of the optimal policy and are the same for all experiments. We ran some simulations to compute the expected cost of different policies for various sizes of the system. We compare different policies:

1. Deterministic policy a^* – to obtain this curve, the optimal actions $a_0^* \ldots a_{T-1}^*$ that the controller must take for the deterministic system have been computed. At time t, action a_t^* is used regardless of the currently state, and the cost up to time T is displayed.

2. Limit policy μ^* – here, the optimal policy μ^* for the deterministic case was first computed. When the stochastic system is in state $(\mathrm{M}^N(t), \mathrm{C}^N(t))$ at time t, we apply the action $\mu_t^*(\mathrm{M}^N(t), \mathrm{C}^N(t))$ and the corresponding cost up to time T is reported.

3. Join the Shortest Queue (JSQ) and Weighted Join the Shortest Queue (W-JSQ) – for JSQ, each packet is routed (deterministically) in the shortest queue. In W-JSQ, a packet is routed in the queue whose weighted queue size $B_i/(\mu_i X_i)$ is the smallest.

The results are reported in Figures 2 and 3.

A series of several simulations for with different values of N was run. The reported values in the figures are the mean values of the waiting time over 10000 simulations for small values of N and around 200 simulations for big values of N. Over the whole range for N, the 95% confidence interval is less than 0.1% for the expected cost – figure 2 – and less than 5% for the central limit theorem – figure 3.

Figure 2 shows the average waiting time of the stochastic system when we apply the different policies. The horizontal line represents the optimal cost of the deterministic system $v^*(m_0, c_0)$ which is probably less than $V_*^N(\mathrm{M}(0), \mathrm{C}(0))$. This figure illustrates Theorem 5: if we apply a^* or μ^*, the cost converges to $v_*(m(0), c(0))$.

In Figure 2, one can see that for low values of N, all the curves are not smooth. This behavior comes from the fact that when N is not very large with respect to N_0, there are at least $\lfloor \frac{N}{N_0} A \rfloor$ (resp. $\lfloor \frac{N}{N_0} P_i \rfloor$) objects that are sources (resp. processors in queue i) and the remaining objects are distributed randomly. The random choice of the remaining states are chosen so that $\mathbb{E}[A^N] = \frac{N}{N_0} A$, but the difference $A^N - N N_0 A$ may be large. Therefore, for some N the load of the system is much higher than the average load, leading to larger costs. As N grows, the proportion of remaining objects decreases and the phenomena becomes negligible.

A second feature that shows in Figure 2, is the fact that on all curves, the expected waiting times are decreasing when N grows. This behavior is certainly related to Ross conjecture [17] that says that for a given load, the average queue length decreases when the arrival and service processes are more deterministic.
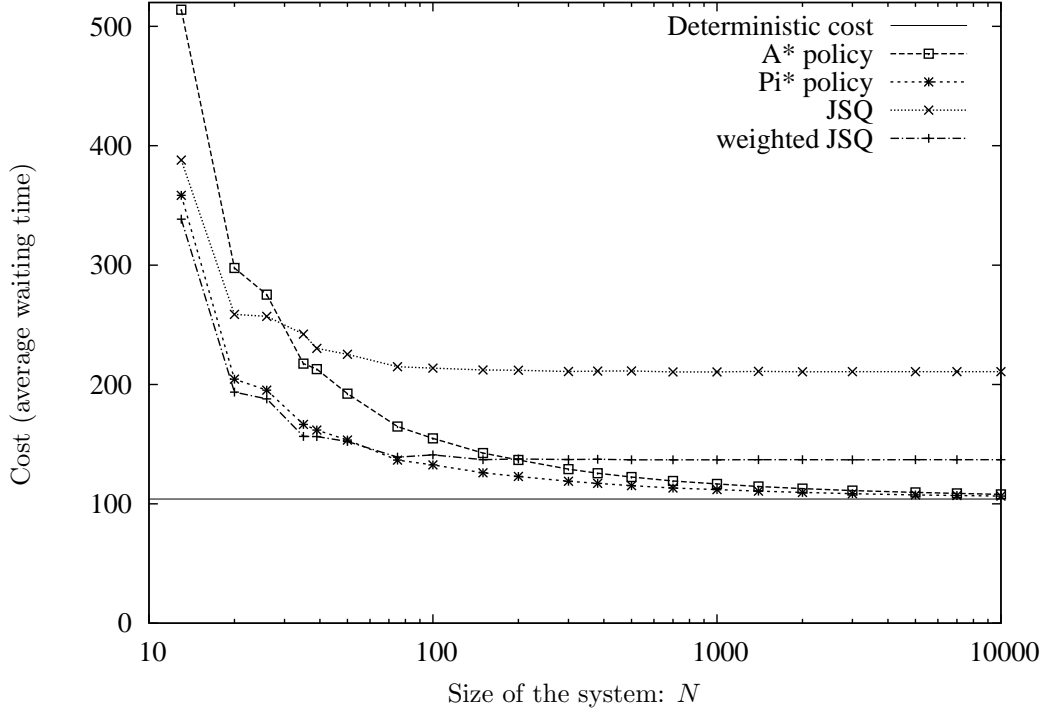
Figure 2: Expected cost of the policies a^*, μ^*, JSQ and W-JSQ for different values of N.

Finally, the most important information on this figure is the fact that the optimal deterministic policy and the optimal deterministic actions perform better than JSQ and weighted JSQ as soon as the total number of elements in the system is over 200 and 50 respectively. The performance of the deterministic policy a^* is quite far from W-JSQ and JSQ for small values of N, and it rapidly becomes better than JSQ ($N \geq 30$) and W-JSQ ($N \geq 200$). Meanwhile the behavior of μ^* is uniformly good even for small values of N.

Figure 3 illustrates Theorem 7 which says that the speed of convergence towards the limit is of order $\sqrt{N}$. On the y-axis, $\sqrt{N}$ times the average cost of the system minus the optimal deterministic cost is plotted. One can see that the gap between the expected cost of the policy μ^* (resp. a^*) and the deterministic cost $v_*(m(0), c(0))$ is about $250/\sqrt{N}$ (resp. $400/\sqrt{N}$) when N is large. This shows that the speed of convergence of $1/\sqrt{N}$ is a tight bound. This should be an upper bound on the constant δ defined in Equation (7).

Besides comparing a^* and μ^* to other heuristics, it would be interesting to compare it to the optimal policy of the stochastic system, whose cost is $V_*^N(M(0), C(0))$. One way to compute this optimum would be by using Bellman fixed point equation. However to do so, one needs to solve it for all possible values of M and C. In this example, C can be as large as the length of the five queues and each object's state can vary in {on,off}. Therefore even with $N = 10$ and if we only compute the cost for queues of size less than 10, this leads to $2^N 10^5 \approx 10^8$ states which is hard to handle even with powerful computers.

5   Extensions and Counter-Examples

This part is devoted to several extensions of the previous results as well as to some counter-examples showing the limitations of the mean field model.
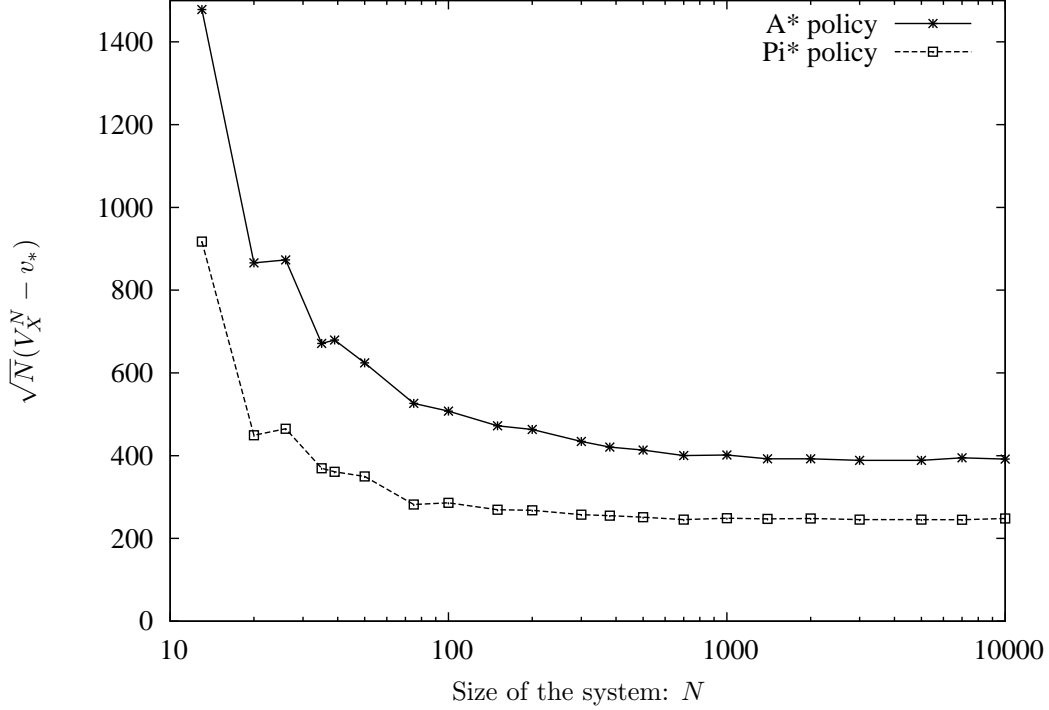
Figure 3: Speed of convergence of the policies $X = a^*$ or μ^* for different values of N.

5.1 Object-Dependent Actions

Up to now, we have assumed that the controller takes the same action for all objects. Here, we show that our frame work can also be used in the case where the controller can take a different action for each object, depending on its state but also on the object itself.

For that, we consider the following new system. The state of the system is the states of the N objects $\mathcal{X}^N = (X_1^N \dots X_N^N)$ and the state of the context. At each time step, the controller chooses an N-uple of actions $a_1 \dots a_N \in \mathcal{A}$ and uses the action a_i for the ith object. We assume that $\mathcal{A}$ is finite. The reward and the evolution of the context is defined as before and we call $V_{od*}^N(\mathcal{X}^N, \mathrm{C}^N)$ the optimal reward of this system over a finite horizon $[0; T]$ where od stands for "Object-Dependent"-actions.

As before, $\mathrm{M}^N(0) \stackrel{\text{def}}{=} N^{-1} \sum_{n=1}^N \delta_{X_n^N(0)}$ is the empirical measure of $\mathcal{X}^N$. We consider our original problem in which we replace the action set $\mathcal{A}$ by $\mathcal{P}(A)^{\mathcal{S}}$. An action is a $\mathcal{S}$-uple $(p_1 \dots p_S)$. If the controller takes the action p, then an object in state i will choose an action a according to the distribution p and evolves independently according to $K(a, \mathrm{C})$.

Compared to the problem in which the action sent is object-dependent, the action of the controller in the latter case are more constrained since it can not choose which object or even the exact number of objects receiving a particular action. However, as we see in Proposition 10, this difference collapses as N grows. Other results, such as second order results, also hold.

Proposition 10. *If $g, K, \mathcal{A}, \mathrm{M}_0^N, \mathrm{C}_0^N$ satisfy assumptions (A1,A2,A3,A4), the object-dependent reward V_{od*}^N converges to the deterministic limit:*

$$\lim_{N \to \infty} V_{od*}^N(\mathcal{X}^N, \mathrm{C}^N) = \lim_{N \to \infty} V_*^N(\mathrm{M}^N, \mathrm{C}^N) = v_*(m(0), c(0))$$

where the deterministic limit has an action set $\mathcal{P}(\mathcal{A})$.

Sketch of proof. To each N-uple $a = a_1 \ldots a_N \in \mathcal{A}^N$ and each vector $\mathcal{X}^N \in \mathcal{S}^N$, we associate a $\mathcal{S}$-uple of probability measures on the set $\mathcal{A}$ defined by $(p_{a,\mathcal{X}^N})_i \overset{\text{def}}{=} \frac{1}{\sum_{n=1}^N \mathbf{1}_{X_n^N = i}} \sum_{n=1}^N \mathbf{1}_{X_n^N = i} \delta_{a_n}$. For $b \in \mathcal{A}$, $(p_{a,\mathcal{X}^N})_i(b)$ represents the average number of objects in state i that received the action b.

One can show that starting from $\mathcal{X}^N(t), \mathrm{C}^N(t)$ and applying action a or $p_{a,\mathcal{X}^N(t)}$ both lead to the same almost sure limit for $\mathrm{M}^N(t+1), \mathrm{C}^N(t+1)$. Then one can show by induction on the time-horizon T that the reward under a fixed sequence of action $V_{od*}^N(\mathcal{X}^N, \mathrm{C}^N)$ is asymptotically equal to $V_*^N(\mathrm{M}^N, \mathrm{C}^N)$.

Then remarking that $(\mathcal{P}(\mathcal{A}))^{\mathcal{S}}$ also satisfies hypothesis (A2) and (A3) – it is compact and the mappings g, K and r are uniformly continuous if $p_a \in \mathcal{P}(\mathcal{A})$ – we can apply the rest of the results and show that the reward converges to the deterministic counterpart. □ □

5.2 Infinite horizon discounted reward

In this section, we prove first and second order results for infinite-horizon discounted Markov decision processes. As in the finite case, we will show that when N grows large, the maximal expected discounted reward converges to the one of the deterministic system and the optimal policy is also asymptotically optimal. To do this , we need the following new assumptions:

(A6) **Homogeneity in time** – The reward r_t and the probability kernel K_t do not depend on time: there exists r, K such that, for all $\mathrm{M}, \mathrm{C}, a \; r_t(\mathrm{M}, \mathrm{C}) = r(\mathrm{M}, \mathrm{C})$ and $K_t(a, \mathrm{C}) = K(a, \mathrm{C})$.

(A7) **Bounded reward** – $\sup_{\mathrm{M},\mathrm{C}} r(\mathrm{M}, \mathrm{C}) \leq K < \infty$.

The homogeneity in time is clearly necessary as we are interested in infinite-time behavior. Assuming that the cost is bounded might seem strong but it is in fact very classical and holds in many situation, for example when C is bounded. The rewards are discounted according to a discount factor $0 \leq \delta < 1$: if the policy is π, the expected total discounted reward of π is (δ is omitted in the notation):

$$V_\pi^N(\mathrm{M}^N(0), \mathrm{C}^N(0)) \overset{\text{def}}{=} \mathbb{E}_\pi \Big[\sum_{t=1}^\infty \delta^{t-1} r(\mathrm{M}^N(t), \mathrm{C}^N(t)) \Big].$$

Notice that Assumption (A7) implies that this sum remains finite. The optimal total discounted reward V_*^N is the supremum on all policies. For $T \in \mathbb{N}$, the optimal discounted finite-time reward until T is

$$V_{T*}^N(\mathrm{M}(0), \mathrm{C}(0)) \overset{\text{def}}{=} \sup_\pi \mathbb{E}_\pi \Big[\sum_{t=1}^T \delta^{t-1} r(\mathrm{M}(t), \mathrm{C}(t)) \Big].$$

As r is bounded, one can show that it converges uniformly in (M, C) to V_*^N:

$$\lim_{T \to \infty} \sup_{\mathrm{M},\mathrm{C}} \left| V_{T*}^N(\mathrm{M}, \mathrm{C}) - V_*^N(\mathrm{M}, \mathrm{C}) \right| = 0. \tag{10}$$

Equation (10) is the key of the following analysis. Using this fact, we can prove the convergence when N grows large for fixed T and then let T go to infinity. Therefore with a very few changes in the proofs of Section 3.2, we have the following result:

Theorem 11. (Optimal discounted case) *Under assumptions (A1,A2,A3,A4,A6, A7), as N grows large, the optimal discounted reward of the stochastic system converges to the optimal discounted reward of the deterministic system:*

$$\lim_{N \to \infty} V_*^N(\mathrm{M}^N, \mathrm{C}^N) =_{a.s} v_*(m, c),$$

where $v_*(m, c)$ satisfies the Bellman equation for the deterministic system:

$$v_*(m, c) = r(m, c) + \delta \sup_{a \in \mathcal{A}} \Big\{ v_*(\Phi_a(m, c)) \Big\}.$$

The first order of convergence for the discounted cost is a direct consequence of the finite time behavior convergence. However, when considering second order results, similar difficulties as in the infinite horizon case arise and the convergence rate depends on the behavior of the system when T tends to infinity.

Let assume that at time 0, $\sqrt{N}\left(M^N(0), C^N(0) - m(0), c(0)\right)$ converges to a Gaussian variable G_0 of covariance Γ_0. We have seen that if we fix the sequence of actions, this implies that at time t, $\sqrt{N}\left(M^N(t), C^N(t) - m(t), c(t)\right)$ converges to a random variable G_t of covariance Γ_t with

$$\Gamma_{t+1} = H_t^{tr}\Gamma_t H_t + B_t$$

where H_t, B_t are square matrices depending on C, see Theorem 8 for a precise definition of H_t, B_t.

The gap between the discounted reward at time t and the mean field limit is, $\sqrt{N}\left|\mathbb{E}[M^N(0), C^N(0)] - m(0), c(0)\right| \leq \sqrt{\|\Gamma_t\|}$. This shows that

$$\sqrt{N}\left\|V_*^N(\mathrm{M}^N(0), \mathrm{C}^N(0)) - v_*(m(0), c(0))\right\|_\infty \leq \sum_{t=0}^\infty \delta^t \sqrt{\|\Gamma_t\|}.$$

We will see that the growth of Γ_t can be bounded by $\|\Gamma_{t+1}\| \leq L^2 \|\Gamma_t\| + 1$, showing that the previous sum converges as soon as $\delta < \frac{1}{L}$. On the other hand, Example 1 displays an example of divergence in the case $\delta \geq \frac{1}{L}$.

Proposition 12. *under assumptions (A1,A2,A3,A4bis, A6,A7) and if the functions $m \mapsto K(a, m)$ and $m \mapsto g(c, m)$ are Lipschitz of constant L_K, L_g such that $L_g(1 + \max(L_K, L_g))\delta \leq 1$, there exist constants $K, K' > 0$ such that:*

$$\lim_{N\to\infty} \sqrt{N}\left\|V_*^N(\mathrm{M}^N(0), \mathrm{C}^N(0)) - v_*(m(0), c(0))\right\|_\infty \leq K + K'\sqrt{N}\epsilon_0^N$$

where $\epsilon_0^N \overset{\text{def}}{=} \left\|\left(\mathrm{M}^N(0), \mathrm{C}^N(0)\right) - (m(0), c(0)\|\right)$

Proof. (sketch) This result is a direct consequence of Proposition 6 and can be proved similarly to Theorem 7. In particular, it uses the fact that in Equation (5) of Proposition 6,

$$\sqrt{N}\left\|\left(M_\pi^N(t), C_\pi^N(t)\right) - (m_\pi(t), c_\pi(t))\right\| \leq_{\text{st}} \left(\beta_t |G| + \beta_t'\sqrt{N}\delta_0^N\right)(1 + \epsilon^N(t)), \tag{11}$$

the growth of the constants β_t and β_t' is bounded by a factor $L_g(1 + \max(L_g, L_K))$. □   □

Example 1. *This example is a system without control. We show that even in this simple case, $\sum_{t=0} \delta^t \sqrt{N}\left(r(M^N(t)) - r(m(t))\right)$ does not converges if δ does not satisfies the assumptions of Proposition 12. The system is defined as follows:*

- *The state space is $\mathcal{S} = \{0, 1\}$ and the context is $\mathrm{C}(t) \overset{\text{def}}{=} \mathrm{M}_0^N(t)$ (C is the mean number of particles in state 0). Therefore the interesting process is $\mathrm{C}(t)$.*

- *For any particle, the probability of going to state 0 is $f(\mathrm{C})$ (independent of the state) where f is a piecewise linear function with $f(0) = f(\alpha) = 0$ and $f(1 - \alpha) = f(1) = 1$ for a number $\alpha < \frac{1}{2}$. The transition function is depicted on the left of Figure 4*

- *The starting point is $\mathrm{M}_0^N(0) = \mathrm{C}(0) = \frac{1}{2}$.*

- *The reward function is $r(\mathrm{M}, \mathrm{C}) = \left|\mathrm{C} - \frac{1}{2}\right|$.*

The deterministic limit of M^N starting in $\frac{1}{2}$ is constant equal to $\frac{1}{2}$. Therefore, the gap between the discounted reward of the stochastic system and the discounted reward of the limit normalized by $\sqrt{N}$ is $\sum_{t=0}^\infty \delta^t \sqrt{N}\left|\mathrm{C}^N(t) - \frac{1}{2}\right|$.

The process $\mathrm{C}^N(t) - \frac{1}{2}$ is quite complicated. However for any finite horizon $[0; T]$, if N is large enough, $\mathrm{C}^N(t) - \frac{1}{2}$ is close $\frac{1}{2}$ with high probability. During this time, $\sqrt{N}(\mathrm{C}^N(t) - \frac{1}{2})$ is
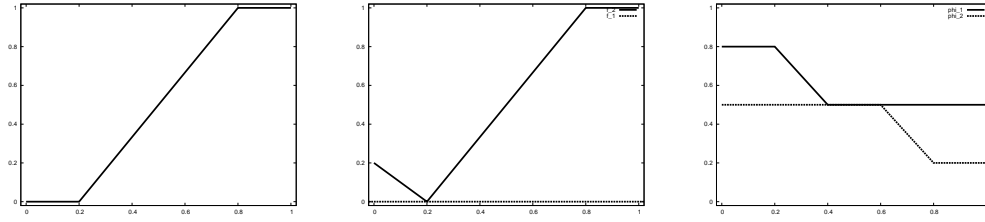
Figure 4: The transitions functions of examples 1, 2 and 3 (from left to right): on the left figure is draw the probability for each objects to go in state 1 as a function under the only action as a function of M_0^N. On the center and right figures are drawn the two possible probabilities as a function of the actions.

close to the process $(Y^N(t))_{t\in\mathbb{N}}$ where $Y^N(t)$ is defined by $Y^N(0) = 0$ and $X_{t+1}^N = LX_t^N + G_t$ where L is the Lipschitz constant of f, $L \stackrel{\text{def}}{=} \frac{1}{1-2\alpha}$, and G_t are i.i.d Gaussian variables of mean 0 and variance $1/4$. As sum of i.i.d Gaussian variables, X_t^N is a Gaussian variable of variance $\sum_{i=0}^{t-1} L^{2i} = L^{2t}\frac{1-L^{-2t}}{L^2-1}$. Therefore, if $\delta \geq 1/L$, $\mathbb{E}\sum_{i=0}^{t-1} \delta^t |X_t^N|$ goes to $+\infty$ as t goes to infinity. By choosing T large enough, this shows that $\sum_{t=0}^{\infty} \delta^t \sqrt{N} |C^N(t) - \frac{1}{2}|$ is not bounded as N grows.

5.3   Average Reward

The discounted problem is very similar to the finite case because the total reward mostly depends on the rewards during a finite amount of time. Here, we consider another infinite-horizon criterion, namely the average reward. The optimal average reward is (if it exists[1])

$$V_{av*}^N = \lim_{T\to\infty} \frac{1}{T} V_{T*}(\mathrm{M}(0), \mathrm{C}(0)).$$

This raises the problem of the exchange of the limits $N \to \infty$ and $T \to \infty$. Consider a case without control with two states $\mathcal{S}=\{0; 1\}$ and $\mathrm{C}(t)$ is the mean number of objects in state 1 ($\mathrm{C}(t) = (\mathrm{M}(t))_1$) and with a function $f:[0;1]\to[0;1]$ such that the transition kernel K is $K_{i1}(\mathrm{C}) = f(\mathrm{C})$ for $i \in \mathcal{S}$. If $\mathrm{M}_0^N(0) \xrightarrow{\text{a.s.}} m_0$ then for any fixed t, $\mathrm{M}^N(t)$ converges to $f(f(\ldots f(m_0)\ldots))$. However, in general we may have $\lim_{t\to\infty} \lim_{N\to\infty} \mathrm{M}^N(t) \neq \lim_{N\to\infty} \lim_{t\to\infty} \mathrm{M}^N(t)$. For example if $f(x) = x$, the deterministic system is constant while the stochastic system converges almost surely to a random variable (as a bounded Martingale) that takes values in $\{0; 1\}$. Also, in some examples the optimal stochastic policy is not the same as the optimal deterministic policy.

Example 2. *We consider a similar example as Example 1: there are two states and $C^N = M_0^N$ (the proportion of objects in state 0). We consider two possible actions, say 1 and 2, corresponding to a probability to go from any state to 0 is equal to $f_1(C)$ and $f_2(C)$ respectively, defined by (see Figure 4, center).*

- $f_1(C) = 0$.

- f_2 *is piecewise linear with $f_2(0) = .2$, $f_2(.1) = 0$, $f_2(.8) = 1$, $f_2(1) = 1$.*

The reward function is $r(C) = |C - .1|$.

For the deterministic system, applying action 1 always makes the system converge to 0 while applying action 2 makes the system converge to .2 if we start in $[0; .5)$ and 1 in $(0.5; 1]$. Therefore, if we start in $[0; .5)$, the average reward of the deterministic system is maximized under action 1 (it gives .1).

For the stochastic system, applying action 1 makes the system converge to 0. However, applying action 2 makes the system converge to 1: there is a small but positive probability that M^N goes to

---

[1]If it does not exist, one may replace this limit by $\lim\sup$ or $\lim\inf$

something greater than .8 at each step which makes the system go to 1. *Therefore, if we start in* $[0; .5)$, *it is better to apply the action* 2, *which is different from the optimal policy of the limit.*

In the case without control, Proposition 13 gives the condition under which the ergodic behavior of the system with N finite converges to the asymptotic behavior of its deterministic limit. This result is similar to the classical results of stochastic approximation theory concerning differential equations limit (see [7] for example). However, no general results for the controlled problem is presented here since the condition to apply theses results are too restrictive to be applied in practical situations (see Example 3) for an example where many assumptions are verified but where we cannot exchange the limits.

Let us assume that the context variable C remains bounded. In particular the possible values for M, C remains in a compact space $B \subset \mathbb{R}^{d+S}$. Let $f : B \to B$ be a deterministic function. Let us define the sequence x_s^x by

$$
\begin{aligned}
x_s^x(s) &= x \\
x_s^x(t+1) &= f(x_s^x(t)) \text{ for } t \geq s.
\end{aligned}
$$

We say that a set H is an attractor of the function f if:

- For all $x \in B$:
$$
\lim_{t \to \infty} d(f(x_0^x(t)), H) = 0
$$

  where $d(x, H)$ denotes the distance between a point x and a set H.

- There exists a neighborhood H^ϵ of H such that the convergence is uniform in $x \in H^\epsilon$:
$$
\lim_{t \to \infty} \sup_{x \in H^\epsilon} d(f(x_0^x(t)), H) = 0
$$

For an action $a \in \mathcal{A}$, we denote by Φ_a the deterministic function corresponding to one iteration of the deterministic limit. Using this notation, we have the following theorem:

Proposition 13. *For any attractor H of Φ_a, and if the controller always choose the action a, for all $\epsilon > 0$:*
$$
\lim_{N \to \infty} \limsup_{t \to \infty} \mathcal{P}\left(d\left(\left(M_a^N(t), C^N(t)\right), H\right) \geq \epsilon\right) = 0
$$

Proof. Let $\epsilon > 0$ and let T be such that

$$
\sup_{x \in H^\epsilon} d(f(x_0^x(T)), H) \leq \epsilon.
$$

For all $t \in \mathbb{N}$, using the triangular inequality, we have

$$
d(X^N(t+T), H) \leq \left\| X^N(t+T) - x_t^{X^N(t)}(t+T) \right\| + d(x(t+T), H).
$$

By Theorem 1,

$$
\mathcal{P}\left(d(X^N(t+T), x_t^{X^N(t)}(t+T)) \geq \mathcal{E}_T(\epsilon, 0)\right) \leq 2 \exp(-2N\epsilon^2)
$$

which goes to 0 as N goes to infinity.

Therefore,
$$
\mathcal{P}\left(d\left(\left(M_a^N(t), C^N(t)\right), H\right) \geq \mathcal{E}_T(\epsilon, 0) + \epsilon\right) \leq 2 \exp(-2N\epsilon^2)
$$

Let ϵ' be such that $\mathcal{E}_T(\epsilon', 0) + \epsilon' < \epsilon$. This concludes the proof.   □   □

We say that a point x is an attractor of a mapping Φ is $\{x\}$ is an attractor of Φ. As a direct corollary of Proposition 13, we have:

Corollary 14. *If the function Φ_a has a unique attractor m_∞, c_∞, then*

$$\lim_{N\to\infty} \limsup_{t\to\infty} \left\| \left(\mathrm{M}^N(t), \mathrm{C}^N(t)\right) - (m_\infty, c_\infty) \right\| = 0 \quad \text{in probability.}$$

In the controlled case, there is no positive results with assumptions that are easy to verified in practice. In particular, assuming that for all action a, there exists a unique attractor point of Φ_a does not ensure that the average reward converges to its deterministic counterpart as Example 3 shows.

Example 3. *As in the example 1, we consider a system with 2 states where $\mathrm{C}^N = \mathrm{M}_0^N$ is the proportion of objects in state 0. The only difference here is that there is two possible actions 1 and 2, corresponding to a probability of transition from any state to 0 of $f_1(\mathrm{C})$ and $f_2(\mathrm{C})$. Both f_1 and f_2 are piecewise linear functions taking the values:*

- *$f_1(x) = .8$ for $x \le .2$, $.5$ for $x > .4$;*

- *$f_2(x) = .5$ for $x \le .6$, $.2$ for $x > .6$.*

Figure 4,right shows the transition functions. The reward is set to $|\mathrm{C}^N - 1/2|$.

Both f_1 and f_2 have only one accumulation point $\{1/2\}$. Moreover, one can prove that under any policy, $\lim_{N\to\infty} \lim_{t\to\infty} \mathrm{M}_\pi^N(t)$ will converges to $.5$, leading to an average reward of 0 regardless of the initial condition. However, if the deterministic limit starts from the point $\mathrm{C}^N(0) = .2$, then by choosing the sequence of action $1, 2, 1, 2 \ldots$ the system will oscillate between $.2$ and $.8$, leading to an average reward of $.3$.

This is caused by the fact that even if f_1 and f_2 have the same unique attractor, $f_1 \circ f_2$ has 3 accumulation point: $.2, .5$ and $.8$.

6 Computational issues

Throughout the paper, we have shown that if the controller uses the optimal policy μ^* of the deterministic limit of the finite real system, the expected cost will be close to the optimal one (Theorem 5). Moreover, Theorem 7 gives a bound on the error that we make. However to apply these results in practice, a question remains: how difficult is it to compute the optimal limit policy?

The first answer comes straight from the example. In many cases, even if the stochastic system is extremely hard to solve, the deterministic limit is often much simpler. The best case of course is, as in the example of section 4, when one can compute the optimal policy. If one can not compute it, there might also exist approximation policies with bounded error (see [12] for a review on the subject). Imagine that a 2-approximation algorithm exists for the deterministic system, then, Theorem 5 proves that for all ε, this algorithm will be a $(2+\varepsilon)$-approximation for the stochastic system if N is large enough. Finally, heuristics for the deterministic system can also be applied to the stochastic version of the system.

If none of this works properly, one can also compute the optimal deterministic policy by "brute-force" computations using the equation

$$v_{t\ldots T*}(m, c) = r_t(m, c) + \sup_a v_{t+1\ldots T*}(\Phi_a(m, c)).$$

In that case, an approximation of the optimal policy is obtained by discretizing the state space and by solving the equation backward (from $t = T$ to $t = 0$), to obtain the optimal policy for all states. The brute force approach can also be applied directly on the stochastic equation using:

$$V_{t\ldots T*}^N(\mathrm{M}, \mathrm{C}) = r_t(\mathrm{M}, \mathrm{C}) + \sup_{a\in\mathcal{A}} \mathbb{E}_{\mathrm{M},\mathrm{C}} \left[V_{t+1\ldots T*}^N \left(\Phi_a^N(\mathrm{M}, \mathrm{C}) \right) \right].$$

However, solving the deterministic system has three key advantages. The first one is that the size of the discretized deterministic system may have nothing to do with the size of the original state

space for N objects: it depends mostly on the smoothness of functions g and ϕ rather than on N. The second one is the suppression of the expectation which might reduce the computational time by a polynomial factor[1] by replacing the $|\mathbb{P}_N(\mathcal{S})|$ possible values of M_{t+1}^N by 1. The last one is that the suppression of this expectation allows one to carry the computation going forward rather than backward. This latter point is particularly useful when the action set and the time horizon are small.

7   Conclusion and future work

In this paper, we have shown how the mean field framework can be used in an optimization context: the results known for Markov chains can be transposed almost unchanged to Markov decision processes. We further show that the convergence to the mean field limit in both cases (Markovian and Markovian with controlled variables) satisfies a central limit theorem, providing insight on the speed of convergence.

We are currently investigating several extensions of these results. A possible direction is to consider stochastic systems where the event rate depends on N. In such cases the deterministic limits are given by differential equations and the speed of convergence can also be studied.

References

[1] EGEE: Enabling Grids for E-sciencE.

[2] M. Benaïm 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] V. Berten and B. Gaujal. Brokering strategies in computational grids using stochastic prediction models. *Parallel Computing*, 2007. Special Issue on Large Scale Grids.

[4] V. Berten and B. Gaujal. Grid brokering for batch allocation using indexes. In *Euro-FGI NET-COOP*, Avignon, France, june 2007. LNCS.

[5] C. Bordenave and V. Anantharam. Optimal control of interacting particle systems. Technical Report 00397327, CNRS Open-Archive HAL, 2007.

[6] C. Bordenave, D. McDonald, and A. Proutiere. A particle system in interaction with a rapidly varying environment: Mean field limits and applications. *Arxiv preprint math.PR/0701363*, 2007.

[7] V. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[8] J.Y.L. Boudec, D. McDonald, and J. Mundinger. A Generic Mean Field Convergence Result for Systems of Interacting Objects. *QEST 2007.*, pages 3–18, 2007.

[9] R. Durrett. *Probability: theory and examples*. Wadsworth & Brooks/Cole, 1991.

[10] N. Gast and B. Gaujal. A mean field approach for optimization in particle systems and applications. *Fourth International Conference on Performance Evaluation Methodologies and Tools, ValueTools*, 2009.

[11] C. Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability*, 37:198–211, 2000.

[12] D.S. Hochbaum. *Approximation algorithms for NP-hard problems*. PWS Publishing Co. Boston, MA, USA, 1996.

---

[1] The size of $\mathbb{P}_N(\mathcal{S})$ is the binomial coefficient $(N+1+S, S) \sim_{N \to \infty} \frac{N^S}{S!}$

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[14] J. Palmer and I. Mitrani. Optimal and heuristic policies for dynamic server allocation. *Journal of Parallel and Distributed Computing*, 65(10):1204–1211, 2005. Special issue: Design and Performance of Networks for Super-, Cluster-, and Grid-Computing (Part I).

[15] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queueing network control. *Math. Oper. Res.*, 24:293–305, 1999.

[16] M.L. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

[17] T. Rolski. Comparison theorems for queues with dependent interarrival times. In *Lecture Notes in Control and Information Sciences*, volume 60, pages 42–71. Springer-Verlag, 1983.

[18] R. R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27:637–648, 1990.

[19] P. Whittle. *A celebration of applied probability*, volume 25A, chapter Restless Bandits: activity allocation in a changing world, pages 287–298. J. Appl. Probab. Spec., 1988.

Appendix

A   Proof of Theorem 1 (controlled mean field)

Let $U_{i,n}^N(t)$ be a collection of *iid* random variables uniformly distributed on $[0;1]$. Let $\Pi : \mathcal{P}(\mathcal{S}) \times \mathbb{R}^d \mapsto a$ be a policy. The evolution of the process M^N, C^N can be defined as follows:

$$
\begin{aligned}
M_j^N(t+1) &= \frac{1}{N} \sum_{i=1}^{S} \sum_{n=1}^{NM_i^N(t)} \mathbf{1}_{H_{ij}(A_\pi^N(t),C^N(t)) \leq U_{i,n}^N(t) \leq H_{ij+1}(A_\pi^N(t),C^N(t))} \\
C^N(t+1) &= g(A_\pi^N(t), M^N(t+1), C^N(t))
\end{aligned}
$$

where $H_{ij}(a,C) \overset{\text{def}}{=} \sum_{\ell=1}^{j-1} K_{ij}(a,C)$ and $A_\pi^N(t) \overset{\text{def}}{=} \Pi_t(M^N(t), C^N(t))$.

Let $B_{inj}^N \overset{\text{def}}{=} \mathbf{1}_{H_{ij}(A_\pi^N(t),C^N(t)) \leq U_{i,n}^N(t) \leq H_{ij+1}(A_\pi^N(t),C^N(t))}$. $(B_{inj}^N)_{i,n,N}$ are *iid* bernouilli random variable with mean $\mathbb{E}[B_{inj}^N | A_\pi^N(t) = a, C^N(t) = c] = K_{ij}(a,c)$. Therefore, by Hoeffding's inequality (Inequality 2.3 of [13]), we have:

$$
\mathcal{P}\left(\left\| \sum_{n=1}^{NM_i^N(t)} B_{inj}^N(t) - NM_i^N(t) K_{ij}(A_\pi^N(t), C^N(t)) \right\| \geq NM_i^N(t) \frac{\epsilon}{M_i^N(t)} \right) \leq 2\exp(-2N\frac{\epsilon}{M_i^N(t)})
$$
$$
\leq 2\exp(-2N\epsilon) \quad (12)
$$

Moreover, if $m \in \mathcal{P}(\mathcal{S}), c \in \mathbb{R}^d$, we have:

$$
\begin{aligned}
\left\| \frac{1}{N} \sum_{i=1}^{NM_i^N(t)} B_{inj}^N(t) - m_i K_{ij}(A_\pi^N(t), c) \right\| &\leq \left\| \frac{1}{N} \sum_{i=1}^{NM_i^N(t)} B_{inj}^N(t) - M_i^N(t) K_{ij}(A_\pi^N(t), C^N(t)) \right\| \\
&+ \left\| (M_i^N(t) - m_i) K(A_\pi^N(t), C_t^N) \right\| \\
&+ \left\| m_i (K(A_\pi^N(t), C_t^N) - K(A_\pi^N(t), c)) \right\| \\
&+ \left\| (M_i^N(t) - m_i)(K(A_\pi^N(t), C_t^N) - K(A_\pi^N(t), c)) \right\|
\end{aligned}
$$

If $\left\|M^N(t), C^N(t) - m, c\right\| \leq \epsilon_t$ and (12) is satisfied, using that K is lipschitz of constant L_K, we have (with probability greater than $1 - 2\exp(-2N\epsilon)$):

$$\left\|\frac{1}{N}\sum_{i=1}^{NM_i^N(t)} B_{inj}^N(t) - m_i K_{ij}(a,c)\right\| \leq \epsilon + \epsilon_t + L_K\epsilon_t + L_K\epsilon_t^2. \tag{13}$$

This implies that with probability greater than $(1 - 2\exp(-2N\epsilon))^{S^2} \geq 1 - 2S^2\exp(-2N\epsilon)$, for all i:

$$\left\|M_i^N(t+1) - m_i(t+1)\right\| \leq \epsilon + \epsilon_t + L_K\epsilon_t + L_K\epsilon_t^2.$$

Assuming that (13) holds, then

$$\begin{aligned}
\left\|C^N(t+1) - c(t+1)\right\| &= \left\|g(a, C^N(t), M^N(t)) - g(a, c(t), m(t))\right\| \\
&\leq (\epsilon_t + \epsilon + \epsilon_t + L_K\epsilon_t + L_K\epsilon_t^2)L_g.
\end{aligned}$$

Let us define now $\epsilon_{t+1} = (\epsilon_t + \epsilon + \epsilon_t + L_K\epsilon_t + L_K\epsilon_t^2)\max(1, L_g)$ with $\epsilon_0 = \left\|M^N(0) - m(0)\right\|$. Then for all $t \geq 0$:

$$\begin{aligned}
\mathcal{P}\left(\left\|M^N(t+1), C^N(t+1) - m(t+1), c(t+1)\right\| \geq \epsilon_{t+1} : \left\|M^N(t), C^N(t) - m(t), c(t)\right\| \leq \epsilon_t\right) \\
\leq 2S^2\exp(-2N\epsilon).
\end{aligned}$$

This proves that

$$\mathcal{P}\left(\left\|M^N(t), C^N(t) - m(t), c(t)\right\| \geq \mathcal{E}_t(\epsilon, \left\|M^N(0), C^N(0) - m(0), c(0)\right\|)\right) \leq 2tS^2\exp(-2N\epsilon),$$

where $\mathcal{E}_t(\epsilon, \left\|M^N(0), C^N(0) - m(0), c(0)\right\|)$ is equal to ϵ_t defined previously. $\qquad\square$

B Proof of Theorem 6 (second order result)

We are interesting in comparing the behavior of $M_\pi^N(t), C_\pi^N(t)$ and $m_{A^N}(t), c_{A_\pi^N}(t)$ where A_π^N is the sequence of actions taken by the controller under policy π (A_π^N is a random variable depending on the values of $M_\pi^N(t), C_\pi^N(t)$) and $m_{A^N}(t), c_{A_\pi^N}(t)$ corresponds to the value of the deterministic limit when the controller apply the sequence of action A_π^N. In order to improve the readability, we suppress the indexes π and A_π^N: the variables $M_\pi^N(t), C_\pi^N(t), m_{A_\pi^N}(t), c_{A_\pi^N}(t)$ will be denoted $M^N(t), C^N(t), m(t), c(t)$.

We show Equation (5) by induction on t. Recall that Equation (5) is:

$$\sqrt{N}\left\|\left(M^N(t), C^N(t)\right) - (m(t), c(t))\right\| \leq_{\text{st}} \left(\beta_t|G| + \beta_t'\sqrt{N}\delta_0^N\right)(1 + \epsilon^N(t))$$

For $t = 0$, it is satisfied by taking $\beta_t = 0$, $\beta_t' = 1$ and $\epsilon^N(0) = 0$.

Assume now that Equation (5) is true for some $t \geq 0$. Let $P^N(t) \stackrel{\text{def}}{=} K(A^N(t), C^N(t))$ and $p t) \stackrel{\text{def}}{=} K(A^N(t), c(t))$. $P^N(t)$ corresponds to the transition matrix at time t of the objects in the system of size N, $p(t)$ is its deterministic counterpart. $\left\|M^N(t+1) - m(t+1)\right\|$ can be decomposed in:

$$\begin{aligned}
\left\|M^N(t+1) - m(t+1)\right\| &\leq \left\|M^N(t+1) - M^N(t)P^N(t)\right\| + \left\|\left(M^N(t) - m(t)\right)p(t)\right\| \\
&+ \left\|m^N(t)\left(P^N(t) - p(t)\right)\right\| + \left\|\left(M^N(t) - m(t)\right)\left(P^N(t) - p(t)\right)\right\|
\end{aligned}$$

The proof of Lemma 18 can be adapted to show that $\left\|M^N(t+1) - M^N(t)P^N(t)\right\| \leq_{\text{st}} |G| + e^N$ where G is a Gaussian variable and e^N is such that $\mathbb{E}[e^N] \leq C\sqrt{\log(N)/N}$.

For the second term, and using the definition of the L_1 norm, we have:

$$\left\|\left(M^N(t) - m(t)\right)p(t)\right\| = \sum_{i,j}\left(M^N(t) - m(t)\right)_i p_{ij} = \left\|M^N(t) - m(t)\right\|.$$

The third term is:

$$\left\| m^N(t) \left(P^N(t) - p(t) \right) \right\| = \left\| P^N(t) - p(t) \right\| \le L_K \left\| C^N(t) - c(t) \right\|$$

Recall that the sequence δ_0^N is such that $\left\| \left(M^N(0), C^N(0) \right) - (m(0), c(0)) \right\| \le_{\text{st}} \delta_0^N$ and $\lim \left| \delta_0^N \right| = 0$. Denoting $d^N(t) \stackrel{\text{def}}{=} \left\| \left(M^N(t) - m(t) \right) \left(P^N(t) - p(t) \right) \right\|$, we have by the induction assumption:

$$\sqrt{N} d^N(t) \quad \le_{\text{st}} \quad \frac{1}{\sqrt{N}} \left(\beta_t + \beta_t' \sqrt{N} \delta_0^N \right)^2 \left(1 + \epsilon^N(t) \right)^2$$

$$\le_{\text{st}} \quad \frac{1}{\sqrt{N}} \left(\beta_t^2 (1 + \epsilon^N(t)) \right)^2 + \delta_0^N \sqrt{N} \beta_t' \left(\frac{\beta_t}{\sqrt{N}} + \beta_t' \delta_0^N \right) \left(1 + \epsilon^N(t) \right)^2$$

By the induction assumption and the fact that $\mathbb{E} \left| \delta_0^N \right| \to 0$, the first term of the sum as well as the factor after δ_0^N are uniformly[1] bounded by random variables converging to 0 in the L_1 norm. Both of these terms are included in $\epsilon^N(t+1)$.

$C^N(t+1)$ is a deterministic function depending on $M^N(t+1)$ and $C^N(t)$. Thus:

$$\left\| C^N(t+1) - c(t+1) \right\| = \left\| g \left(M^N(t+1), C^N(t) \right) - g \left(m(t+1), c(t) \right) \right\|$$

$$\le L_g \left(\left\| M^N(t+1) - m(t+1) \right\| + \left\| C^N(t) - c(t) \right\| \right)$$

Using that, the quantity $\left\| \left(M^N(t+1), C^N(t+1) \right) - (m(t+1), c(t+1)) \right\|$ is bounded by:

$$\left\| M^N(t+1) - m(t+1) \right\| (1 + L_g) + L_g \left\| C^N(t) - c(t) \right\|$$

$$\le \left(|G| + \epsilon^N + \left\| M^N(t) - m(t) \right\| + L_K \left\| C^N(t) - c(t) \right\| \right) (1 + L_g) + L_g \left\| C^N(t) - c(t) \right\| + d^N(t)$$

$$\le \left\| \left(M^N(t), C^N(t) \right) - (m(t), c(t)) \right\| L_g \left(1 + \max(L_g L_K) \right) + (|G| + \epsilon^N)(1 + L_g) + d^N(t).$$

This concludes the proof of the theorem. $\qquad\square$

C   Proof of Theorem 8 (mean field central limit theorem)

We first start by a technical lemma.

Lemma 15. *Let* M^N *be a sequence of random measure on* $\{1, \ldots, S\}$ *and* P^N *a sequence of random stochastic matrices on* $\{1, \ldots, S\}$ *such that* $(\mathrm{M}^N, P^N) \xrightarrow{\text{a.s.}} (m, p)$.
Let $(U_{ik})_{1 \le i \le S, k \ge 1}$ *be a collection of* iid *random variables following the uniform distribution on* $[0; 1]$ *and independent of* P^N *and* M^N *and let us define* Y^N: *for all* $1 \le j \le S$:

$$Y_j^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{S} \sum_{k=1}^{N \mathrm{M}_i^N} \mathbf{1}_{\sum_{l < k} P_{il}^N < U_{ik} \le \sum_{l \le k} P_{il}^N}$$

then there exists a Gaussian vector G *independent of* M^N *and* P^N *and a random variable* Z^N *with the same law as* Y^N *such that*

$$\sqrt{N}(Z^N - \mathrm{M}^N P^N) \xrightarrow{\text{a.s.}} G.$$

Moreover the covariance of the vector G *is the matrix* D:

$$\begin{cases} D_{jj} & = & \sum_i m_i p_{ij}(1 - p_{ij}) \\ D_{jk} & = & -\sum_i m_i p_{ij} p_{ik} \qquad (j \ne k). \end{cases} \tag{14}$$

Proof. The fact that $\sqrt{N}(Z^N - \mathrm{M}^N P^N)$ converges to a Gaussian variable is a direct consequence of the central limit theorem. Here we only prove the independence of G and M^N, P^N.

---

[1]*i.e.* not depending on the values of the process but only on the parameters of the system

As (M^N, P^N) and $(U_{ik})_{1 \leq i \leq S, k \geq 1}$ are independent, they can be viewed as functions on independent probability space Ω and Ω'. For all $(\omega, \omega') \in \Omega \times \Omega'$, we define $X_\omega^N(\omega') \overset{\text{def}}{=} \sqrt{N}(Y^N(\omega, \omega') - \mathrm{M}^N(\omega)P^N(\omega))$.

By assumption, for almost all $\omega \in \Omega$, $(\mathrm{M}^N(\omega), P^N(\omega))$ converges to (m, p). A direct computation shows that, when N grows, the characteristic function of X_ω^N converges to $\exp(-\frac{1}{2}\xi^T \sum_{i=1}^S m_i C_i \xi)$. Therefore for almost all ω, X_ω^N converges in law to G, a Gaussian random variable on Ω'.

Therefore for almost all ω, there exists a random variable $\widetilde{X}_\omega^N$ with the same law as X_ω^N that converges ω'-almost surely to $G(\omega')$. Let $Z^N(\omega, \omega') \overset{\text{def}}{=} \mathrm{M}^N(\omega)P^N(\omega) + \frac{1}{N}\widetilde{X}_\omega^N(\omega')$. By construction of $\widetilde{X}_\omega^N$, for almost all ω, $Z^N(\omega, .)$ has the same distribution as $Y^N(\omega)$ and $\sqrt{N}(Z^N - Y^N P^N) \xrightarrow{\omega, \omega'-a.s} G$. Thus there exists a function $\widetilde{Z}^N(\omega, .)$ that has the same distribution as $Y^N(\omega)$ for all ω and that converges (ω, ω')-almost surely to G. $\square$ $\square$

We are now ready for the proof of Theorem 8.

Proof. Let us assume that the Equation (6) holds for some $t \geq 0$.

As $\sqrt{N}((\mathrm{M}^N, \mathrm{C}^N)(t) - (m, c)(t))$ converges in law to G_t, there exists another probability space and random variables $\widetilde{\mathrm{M}}^N$ and $\widetilde{\mathrm{C}}^N$ with the same distribution as M^N and C^N such that $\sqrt{N}((\widetilde{\mathrm{M}}^N, \widetilde{\mathrm{C}}^N)(t) - (m, c)(t))$ converges almost surely to G_t [9]. In the rest of the proof, by abuse of notation, we will write M and C instead of $\widetilde{\mathrm{M}}$ and $\widetilde{\mathrm{C}}$ and then we assume that $\sqrt{N}((\mathrm{M}^N, \mathrm{C}^N)(t) - (m, c)(t)) \xrightarrow{\text{a.s.}} G_t$.

G_t being a Gaussian vector, there exists a vector of $S+d$ independent Gaussian variables $U = (u_1, \ldots, u_{S+d})^T$ and a matrix X of size $(S+d) \times (S+d)$ such that $G_t = XU$.

Let us call $P_t^N \overset{\text{def}}{=} K(a_t, \mathrm{C}^N(t))$. According to lemma 15 there exists a Gaussian variable H_t independent of G_t and of covariance D such that we can replace $\mathrm{M}^N(t+1)$ (without changing $\mathrm{M}(t)$ and $\mathrm{C}(t)$) by a random variables $\widetilde{\mathrm{M}}^N(t+1)$ with the same laws such that:

$$\sqrt{N}(\widetilde{\mathrm{M}}(t+1)^N - \mathrm{M}(t)^N P_t^N) \xrightarrow{\text{a.s.}} H_t. \tag{15}$$

In the following, by abuse of notation we write M instead of $\widetilde{\mathrm{M}}$. Therefore we have

$$\sqrt{N}(\mathrm{M}^N(t+1) - m(t)P_t) = \sqrt{N}\Big(\mathrm{M}(t+1) - \mathrm{M}^N(t)P_t^N + m(t)(P_t^N - P_t) +$$
$$(\mathrm{M}^N(t) - m(t))P_t + (\mathrm{M}^N(t) - m(t))(P_t^N - P_t)\Big)$$
$$\xrightarrow{\text{a.s.}} H_t + m(t) \lim_{N \to \infty} \sqrt{N}(P_t^N - P_t) + \lim_{N \to \infty} \sqrt{N}(\mathrm{M}^N(t) - m(t))P_t.$$

By assumption, $\lim \sqrt{N}(\mathrm{M}^N(t) - m(t))_i = (XU)_i$. Moreover, the first order Taylor expansion with respect to all component of C gives a.s.

$$\lim_{N \to \infty} m(t)\sqrt{N}(P_t^N - P_t)_j = \sum_{i=1}^S m_i(t) \sum_{k=1}^d \frac{\partial K_{ij}}{\partial c_{t_k}}(a_t, c(t))(XU)_{S+k}$$
$$= \sum_{k=1}^d Q_{kj}(XU)_{S+k}.$$

Thus, the jth component of $\sqrt{N}(\mathrm{M}^N(t+1) - m(t)P_t)$ goes to

$$H_t + \sum_{k=1}^d Q_{kj}(XU)_{S+k} + \sum_{i=1}^S (XU)_i P_{ij} \tag{16}$$

Using similar ideas, we can prove that $\sqrt{N}(\mathrm{C}_k^N(t+1) - c_k(t+1))$ converges almost surely to $\sum_{i=0}^S \frac{\partial g_k}{\partial m_i}(XU)_i + \sum_{\ell=0}^d \frac{\partial g_k}{\partial c_{t_\ell}}(XU)_{S+\ell}$. Thus $\sqrt{N}((\mathrm{M}^N(t+1), \mathrm{C}^N(t+1)) - (m(t+1), c(t+1)))$ converges almost surely to a Gaussian vector.

Let us write the covariance matrix at time t and time $t+1$ as two bloc matrices:

$$\Gamma_t = \left[\begin{array}{c|c} X & O \\ \hline O^T & \text{C} \end{array}\right] \text{ and } \Gamma_{t+1} = \left[\begin{array}{c|c} X' & O' \\ \hline O'^T & \text{C}' \end{array}\right].$$

For $1 \le j, j' \le S$, $X'_{j,j'}$ is the expectation of (16) taken in j times (16) taken in j'. Using the facts that
$\mathbb{E}[(XU)_{S+k}(XU)_{S+k'}] = \text{C}_{kk'}$, $\mathbb{E}[(XU)_{S+k}(XU)_i] = O_{ik}$ and $\mathbb{E}[(XU)_i(XU)_{i'}] = X_{ii'}$, this leads to:

$$\begin{aligned} X'_{j,j'} &= \mathbb{E}[H_j H'_j] + \sum_{k,k'} Q_{kj} Q_{k'j'} \text{C}_{kk'} + \sum_{k,i'} Q_{kj} O_{i'k} P_{i'j'} \\ &\quad + \sum_{i,k'} Q_{k'j'} O_{ik'} P_{ij} + \sum_{i,i'} P_{ij} X_{ii'} P_{i'j} \\ &= D_{jj'} + (Q^T \text{C}Q)_{jj'} + (Q^T O^T P)_{jj'} + (P^T OQ)_{jj'} + (P^T XP)_{jj'}. \end{aligned}$$

By similar computation, we can write similar equations for O' and C' that lead to Equation (7). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

D Proof of Theorem 9 (third order results)

In order to prove Theorem 9, we start with a result on the sharpness of the approximation of the sum of Bernoulli random variable by a Gaussian distribution (Lemma 16).

Let X_i be independent Bernoulli random variables (*i.e* $P(X_i = 1) = 1 - P(X_i = 0) = p$) and let $Y^N = \frac{1}{N} \sum_{i=1}^N X_i$. We know that in a sense, Y^N is close to $Z^N = p + \frac{1}{\sqrt{N}} G$ with G a normal random variable of variance $\sigma^2 = \mathbb{E}[(X_0 - p)^2]$. We want to compute an asymptotic development of the quantity:

$$d_N = \left| \mathbb{E}\left[f\left(Y^N\right) \right] - \mathbb{E}\left[f\left(Z^N\right) \right] \right|$$

where f is a Lipschitz function of Lipschitz constant L.

Let $F_{N,p} : \mathbb{R} \to [0:1]$ and F_p be respectively the CDF (Cumulative Distribution Function) of $\sqrt{N}(Y^N - x)$ and of the standard normal distribution: $F_{N,p}(x) = P(\sqrt{N}(Y^N - x) \le x)$ and $F_p(x) = P(G \le x)$ where G is a normal variable of mean 0 and variance $\sigma^2 = p(1-p) = \mathbb{E}\left[(X_0 - p)^2\right]$. Let U be a random variable uniformly distributed in $[0;1]$. d_N can be rewritten as:

$$d_N = \left| \mathbb{E}\left[f(x + \frac{\sigma}{\sqrt{N}} F_{N,p}^{-1}(U)) - f(x + \frac{\sigma}{\sqrt{N}} F_p^{-1}(U)) \right] \right| \le \frac{L\sigma}{\sqrt{N}} \mathbb{E}\left[|F_{N,p}^{-1}(U) - F_p^{-1}(U)| \right]$$

Therefore, the problem becomes to get an estimation of $\mathbb{E}\left[|F_{N,p}^{-1}(U) - F_p^{-1}(U)| \right]$.

Lemma 16. *There exists a constant C_1 independent of N, L, p such that if U is a random variable uniformly distributed on $[0;1]$ and $F_{N,p}$ and F be the cumulative distribution functions defined previously, then for N big enough,*

$$\mathbb{E}\left[|F_{N,p}^{-1}(U) - F_p^{-1}(U)| \right] \le C_1 \sqrt{\frac{\log(N)}{N}} \tag{17}$$

Proof. For more simplicity, in this proof, we omit the index p when writing $F_{N,p}$ and F_p.

$$\mathbb{E}\left[|F_N^{-1}(U) - F^{-1}(U)| \right] = \int_0^1 |F_N^{-1}(u) - F^{-1}(u)| du \tag{18}$$

The Berry-Essen theorem (see for example part 2.4.d of [9]) shows that $\sup_{y \in \mathbb{R}} |F_N(\sigma y) - F(\sigma y)| \le \frac{3\rho}{\sigma^3 \sqrt{N}}$ where $\rho = \mathbb{E}\left[|X_0 - p|^3 \right]$. As F and F_N are increasing, for all $u \in (\frac{3\rho}{\sigma^2 \sqrt{N}}; 1 - \frac{3\rho}{\sigma^2 \sqrt{N}})$, we have:

$$F^{-1}(u - \frac{3\rho}{\sigma^2 \sqrt{N}}) \le F_N^{-1}(u) \le F^{-1}(u + \frac{3\rho}{\sigma^2 \sqrt{N}}).$$

Using these remarks, the symmetry of the function F and the fact that $F^{-1}(u+\epsilon) - F^{-1}(u) \geq F^{-1}(u) - F^{-1}(u-\epsilon)$ for $u > 1/2$, (18) is less than

$$2 \left(\int_{\frac{1}{2}}^{K_N} \left(F^{-1}(u + \frac{3\rho}{\sigma^2\sqrt{N}}) - F^{-1}(u) \right) du + \int_{K_N}^1 |F_N^{-1}(U) - F^{-1}(U)| du \right) \tag{19}$$

for all constant $K_N \in (\frac{1}{2}; 1 - \frac{3\rho}{\sigma^2\sqrt{N}})$.

The function F^1 is continuously differentiable, therefore the mean value theorem says that there exists $v(u) \in \left(u; u + \frac{3\rho}{\sigma^2\sqrt{N}} \right)$ such that

$$F^{-1}(u + \frac{3\rho}{\sigma^2\sqrt{N}}) - F^{-1}(u) = \frac{3\rho}{\sigma^2\sqrt{N}}(F^{-1})'(v(u)) \leq \frac{3\rho}{\sigma^2\sqrt{N}}(F^{-1})'(u + \frac{3\rho}{\sigma^2\sqrt{N}}).$$

Thus, the first part of inequality (19) is bounded by

$$
\begin{aligned}
\int_{\frac{1}{2}}^{K_N} \left(F^{-1}(u + \frac{3\rho}{\sigma^2\sqrt{N}}) - F^{-1}(u) \right) du &\leq \frac{3\rho}{\sigma^2\sqrt{N}} \int_{\frac{1}{2}}^{K_N} (F^{-1})'(u + \frac{3\rho}{\sigma^2\sqrt{N}}) du \\
&= \frac{3\rho}{\sigma^2\sqrt{N}} \left(F^{-1}(K_N + \frac{3\rho}{\sigma^2\sqrt{N}}) - F^{-1}(\frac{1}{2}) \right) \\
&\leq \frac{3\rho}{\sigma^2\sqrt{N}} F^{-1}(K_N + \frac{3\rho}{\sigma^2\sqrt{N}}).
\end{aligned}
\tag{20}
$$

Using an integration by substitution with $x = F^{-1}(u)$ (and $F'(x)dx = du$) and an integration by part, we get:

$$
\begin{aligned}
\int_{K_N}^1 F^{-1}(u) du &= \int_{F^{-1}(K_N)}^\infty x F'(x) dx \\
&= \left[x \left(F^{-1}(x) - 1 \right) \right]_{F^{-1}(K_N)}^\infty - \int_{F^{-1}(K_N)}^\infty (F(x) - 1) \, dx \\
&= (1 - K_N) F^{-1}(K_N) + \int_{F^{-1}(K_N)}^\infty (1 - F(x)) \, dx
\end{aligned}
\tag{21}
$$

For $x \geq 1$, the tail of the distribution of a Gaussian variable satisfies:

$$\frac{1}{2x\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \leq \frac{x}{(1+x^2)\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \leq 1 - F(x) \leq \frac{1}{x\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \leq \exp(-\frac{x^2}{2}).$$

Moreover, for all $u \geq x \geq 1$ we have $u^2/2 - u \geq x^2/2 - x$ and $\int_x^\infty \exp(-u^2/2) du \leq \int_x^\infty \exp(-x^2/2 + x - u) du = \exp(-x^2/2)$. This leads to

$$\int_{F^{-1}(K_N)}^\infty 1 - F(x) dx \leq \exp(-\frac{F^{-1}(K_N)}{2}) \leq 2\sqrt{2\pi} F^{-1}(K_N)(1 - K_N)$$

for K_N such that $F_N^{-1}(K_N) \geq 1$.

Similarly, $\int_{F^{-1}(K_N)}^\infty 1 - F_N(x) dx$ can be bounded by the same method using Hoeffding's inequality $F_N(x) \leq \exp(-2x^2)$ and we get:

$$\int_{K_N}^1 F_N^{-1}(u) du \leq (1 - K_N) F_N^{-1}(K_N) + \exp(-2F_N^{-1}(K_N)) \tag{22}$$

with

$$
\begin{aligned}
\exp\left(-2F_N^{-1}(K_N) \right) &\leq \exp\left(-2F^{-1}(K_N - \frac{3\rho}{\sigma^2\sqrt{N}}) \right) \\
&\leq 2\sqrt{2\pi} F^{-1}(K_N - \frac{3\rho}{\sigma^2\sqrt{N}})(1 - K_N + \frac{3\rho}{\sigma^2\sqrt{N}})
\end{aligned}
$$

Combining (20), (21) and (22), (18) is less than:

$$
\begin{aligned}
(18) \quad \leq \quad & \frac{3\rho}{\sigma^2\sqrt{N}}F^{-1}\left(K_N + \frac{3\rho}{\sigma^2\sqrt{N}}\right) + (1 - K_N)(F_N^{-1}(K_N) + (1 + 2\sqrt{2\pi})F^{-1}(K_N)) \\
& + 2\sqrt{2\pi}F^{-1}(K_N - \frac{3\rho}{\sigma^2\sqrt{N}})(1 - K_N + \frac{3\rho}{\sigma^2\sqrt{N}})
\end{aligned}
\tag{23}
$$

Let $K_N \stackrel{\text{def}}{=} 1 - 2\frac{3\rho}{\sigma^2\sqrt{N}}$ Hoeffding's inequality $1 - F_N(x) \leq \exp(-2x^2)$ shows that $F_N^{-1}(y) \leq \sqrt{-\log(1 - y)/2}$. Applying this formula to $K_N + \frac{3\rho}{\sigma^2\sqrt{N}}$ leads to:

$$
F_N^{-1}(K_N + \frac{3\rho}{\sigma^2\sqrt{N}}) \leq \sqrt{-\log(\frac{3\rho}{\sigma^2\sqrt{N}})/2} = \sqrt{\frac{\log(N)}{4} + \frac{\log(\sigma^2/3\rho)}{2}}
$$

Similar inequality for the tail of the normal distribution leads to $F^{-1}(y) \leq \sqrt{-2\log(1 - y)}$.

This proves the existence of the constant C of the Equation (17). To conclude, we have to show that the constant is independent of N, L, p. The Equation (23) leads to an inequality of the following form.

$$
(18) \leq \sqrt{\frac{\log(N)}{N}}\frac{\rho}{\sigma^2}(\ldots)
$$

where $\ldots$ represents other terms depending on some constant and the quantity $3\rho/\sigma^2$. The constants are independent of N, L by construction. The only remaining term is ρ/σ^2. As the X_i are Bernoulli random variable of parameter p, we have $\rho/\sigma^2 = p^2 + (1 - p)^2 \in [0; 1]$. This shows that this constant is bounded independently of the value of p. $\qquad\square \qquad\qquad\square$

In the case where we sum only a fraction B_N of the N Bernouilli variables, the result still holds.

Lemma 17. *Let $0 \leq B_N \leq 1$ be a random variable and $b \in [0; 1]$ such that $E|B_N - b| \leq C'\sqrt{\log(N)}/N$ and U a random variable uniformly distributed on $[0; 1]$ independent of B_N. Then:*

$$
\mathbb{E}\left|\sqrt{B_N}F_{NB_N,p}^{-1}(U) - \sqrt{b}F_p^{-1}(U)\right| \leq C_2\sqrt{\frac{\log(N)}{N}}
$$

where $C_2 = C_1 + \max(C', \sqrt{C'} + 2)$

Proof. Again, to ease the notations, we omit to write p in $F_{N,p}$ and F_p.

$$
\begin{aligned}
\mathbb{E}\left|\sqrt{B_N}F_{NB_N}^{-1}(U) - \sqrt{b}F^{-1}(U)\right| \quad \leq \quad & \mathbb{E}\left|\sqrt{B_N}F_{NB_N}^{-1}(U) - \sqrt{B_N}F^{-1}(U)\right| \\
& + \mathbb{E}\left|\left(\sqrt{B_N} - \sqrt{b}\right)F^{-1}(U)\right| \\
\leq \quad & C_1\sqrt{B_N}\sqrt{\frac{\log(NB_N)}{NB_N}} + \mathbb{E}\left|\sqrt{B_N} - \sqrt{b}\right|\mathbb{E}G
\end{aligned}
$$

The first part of the inequality comes from the Lemma 16 and is less than $C_1\frac{\log(N)}{N}$ since $B_N \leq 1$. The last part comes from the fact that U and B_N are independent. Moreover, the variance of G is the variance of a Bernoulli variable, so $\mathbb{E}|G| \leq 1/4$.

To bound $\mathbb{E}\left|\sqrt{B_N} - \sqrt{b}\right|$, we distinguish two cases. If $\sqrt{b} \geq 1/\sqrt{N}$, we have:

$$
\mathbb{E}\left|\sqrt{B_N} - \sqrt{b}\right| \leq \mathbb{E}\frac{\left|B_N - \sqrt{b}\right|}{\sqrt{B_N} + \sqrt{b}} \leq \mathbb{E}\frac{\left|B_N - \sqrt{b}\right|}{\sqrt{b}} \leq \sqrt{N}\mathbb{E}\left|B_N - \sqrt{b}\right| \leq C'\sqrt{\frac{\log(N)}{N}}.
$$

If $\sqrt{b} \leq 1/\sqrt{N}$, we have:

$$
\mathbb{E}\left|\sqrt{B_N} - \sqrt{b}\right| \leq \mathbb{E}\sqrt{B_N} + \sqrt{b} \leq \mathbb{E}\sqrt{|B_N - b|} + 2\sqrt{b} \leq \sqrt{C'\frac{\sqrt{\log(N)}}{N}} + \frac{2}{\sqrt{N}}.
$$

This shows the inequality. $\qquad\square \qquad\qquad\square$

The following lemma generalizes in a sense the previous results in the case of multidimensional bernoulli variables. A multidimensional bernoulli variable X_i of parameter $(p_1 \ldots p_d)$ is a vector on $\mathbb{R}^d$ that has exactly one component equal to 1 and all others equal 0. The probability that its jth component is equal to 1 is p_j.

Lemma 18. *Let X_i be* i.i.d *multidimensional Bernoulli random vectors on $\mathbb{R}^S$. Assume that X_i are almost surely bounded and $\rho = \mathbb{E}[\|X_i - x\|^3] < \infty$. Let f be a lipschitz function on $\mathbb{R}^S$ with constant Lipschitz L. Let us define $S_N \stackrel{\text{def}}{=} \sum_{i=1}^N X_i$. Then there exists a constant C_3 such that:*

$$\left| \mathbb{E}\left[f(S_N)\right] - \mathbb{E}\left[f(x + \frac{1}{\sqrt{N}}G)\right] \right| \leq C_3 L \frac{\sqrt{\log(N)}}{N},$$

where G is a random Gaussian vector of the same covariance as X_1 and mean 0.

Proof. For $p \in [0;1]$, $A \in \mathbb{N}$, we define the function $F_{A,p}(x) = \mathcal{P}\left(\frac{1}{\sqrt{A}} \left(\sum_{i=1}^A B_i - p \right) \leq x \right)$ where B_i are *iid* Bernoulli variable – *i.e.* $\mathcal{P}(B_i = 1) = 1 - \mathcal{P}(B_i = 0) = p$. Similarly, we define $F_p(x) = \mathcal{P}(G \geq x)$ where G is a normal variable of mean 0 and variance $p(1 - p)$.

Let $U_1 \ldots$ be *iid* uniform variables on $[0;1]$ and let us define the variables T^N and H^N:

$$T_1^N = p_1 + \frac{1}{\sqrt{N}}(F_{N,p_1})^{-1}(U_1)$$

$$T_k^N = \frac{p_k}{1 - \sum_{j<k} p_j}(1 - \sum_{j<k} T_j^N) + \sqrt{\frac{\sum_{j<k} T_j^N}{N}} F^{-1}_{N(1-\sum_{j<k} T_j^N), \frac{p_k}{1-\sum_{j<k} p_j}}(U_k)$$

$$H_1 = F_{p_1}^{-1}(U)$$

$$H_k = -\frac{p_k}{1 - \sum_{j<k} p_j} \sum_{j<k} H_j + \sqrt{\sum_{j<k} p_j} F^{-1}_{\frac{p_k}{1-\sum_{j<k} p_j}}(U_k)$$

S^N is constructed by choosing for each bernoulli vector X_i which of this component is equal to 1. T^N has the same law as S^N, the difference is that we first choose the vectors X_i with the first component equal to 1, then the ones with the second component,...:

- For each of the N vectors X_i, their first component is equal to 1 with probability p_1.

- For the remaining $N - T_1$ vectors, their second component is equal to 1 with probability $p_2/(p_2 + \cdots + p_S) = p_2/(1 - p_1)$.

- The process is iterated until all vectors have been chosen.

The variables H are constructed by the same construction. It is straightforward to prove that H as the same law as G by:

1. Showing that H is a Gaussian vector.

2. Computing the covariance matrix of H.

Therefore, T^N and H are respectively distributed with the same law as the variables S^N and G and we have:

$$
\begin{aligned}
\left| \mathbb{E}\left[f(S^N)\right] - \mathbb{E}\left[x + \frac{1}{\sqrt{N}}G\right] \right| &\leq \mathbb{E}\left[\|f(T^N) - f(H)\|\right] \\
&\leq L\mathbb{E}\left[\|T^N - H\|\right] \\
&\leq L \sum_{d=1}^k \mathbb{E}\left|T_d^N - H_d\right|
\end{aligned}
$$

The next step of the proof is to bound

$$\mathbb{E}\left|\sqrt{\frac{\sum_{j<k} T_j^N}{N}} F^{-1}_{N(1-\sum_{j<k} T_j^N), \frac{p_k}{1-\sum_{j<k} p_j}}(U_k) - \sqrt{\sum_{j<k} p_j} F^{-1}_{\frac{p_k}{1-\sum_{j<k} p_j}}(U_k)\right|$$

By a simple induction on k and using Lemma 17, it can be shown that this quantity is less than $C^{(k)}\frac{\sqrt{N}}{N}$ with $C^{(k)} = C_1 + \max(C^{(k-1)}, \sqrt{C^{(k-1)} + 2})$. C_3 is equal to $C^{(S)}$ where S is the dimension of M^N. $\qquad\square$ $\qquad\qquad\square$

We are now ready for the proof of Theorem 9.

Proof of Theorem 9. Let a be a sequence of actions. We define by a backward induction on t the function $W^N_{t...T}(.)$ which is the expected reward of the mean field Gaussian approximation between t and T:

$$\begin{array}{rcl} W^N_{T...T,a}\left(M^N(t), C^N(t)\right) & = & 0 \\ W^N_{t...T,a}\left(M^N(t), C^N(t)\right) & = & r\left(M^N(t), C^N(t)\right) + \mathbb{E}\left[W^N_{t...T,a}\left(\widetilde{M}^N_a(t+1), \widetilde{C}^N_a(t+1)\right)\right] \end{array} \quad (24)$$

where $\widetilde{M}^N_a(t+1), \widetilde{C}^N_a(t+1)$ is the mean field Gaussian approximation starting at time t in from $M^N(t), C^N(t)$ and after one time step during which the controller took action a_t. Similarly, we define $V^N_{t...T,a}\left(M^N(t), C^N(t)\right)$ the expected reward between t and T for the original system. We want to prove by induction that there exist constants C_t such that for any t:

$$\left|V^N_{t...T}(M^N, C^N) - W^N_{t...T}(M^N, C^N)\right| \le C_t \frac{\sqrt{\log(N)}}{N}. \quad (25)$$

The constant C_t may depend on the parameters of the system (such as the Lipschitz constants of the function g, K, r) but not on the values of M^N, C^N.

Equation (25) is clearly true for $t = T$ by taking $C_T = 0$. Let us now assume that (25) holds for some $t + 1 \le T$. By a backward induction on t, one can show that $W^N_{t...T,a}(.,.)$ is Lipschitz for some constant L_{W_t}. $\left|V^N_{t...T}(M^N, C^N) - W^N_{t...T}(M^N, C^N)\right|$ can be written:

$$\begin{array}{l} \left|\mathbb{E}\left[V^N_{t+1...T,a}\left(M^N_a(t+1), C^N_a(t+1)\right)\right] - \mathbb{E}\left[W^N_{t+1...T,a}\left(\widetilde{M}^N_a(t+1), \widetilde{C}^N_a(t+1)\right)\right]\right| \\ \le \left|\mathbb{E}\left[W^N_{t+1...T,a}\left(M^N_a(t+1), C^N_a(t+1)\right) - W^N_{t+1...T,a}\left(\widetilde{M}^N_a(t+1), \widetilde{C}^N_a(t+1)\right)\right]\right| \\ \quad + \left|\mathbb{E}\left[W^N_{t+1...T,a}\left(M^N_a(t+1), C^N_a(t+1)\right) - V^N_{t+1...T,a}\left(M^N_a(t+1), C^N_a(t+1)\right)\right]\right| \\ \le C_3 L_{W_t} L_g \frac{\sqrt{\log(N)}}{N} + C_t \frac{\sqrt{\log(N)}}{N}. \end{array} \quad (26)$$

The fist part of the second inequality comes from Lemma 18 applied to the function $m \mapsto W^N_{t+1...T,a}\left(m, g(a, C^N(t), m)\right)$ that is lipschitz of constant $L_{W_t} L_g$. The second part comes from the hypothesis of induction. This concludes the proof in the case of a fixed sequence of actions.

The proof for $V^N_* - W^N_*$ is very similar. The first step of the proof is to write a similar equation as (24) for W^N_* and V^N_* which can be computed by a backward induction:

$$\begin{array}{rcl} W^N_{T...T,*}\left(M^N(t), C^N(t)\right) & = & 0 \\ W^N_{t...T,*}\left(M^N(t), C^N(t)\right) & = & r\left(M^N(t), C^N(t)\right) + \sup_a \mathbb{E}\left[W^N_{t...T,a}\left(\widetilde{M}^N_a(t+1), \widetilde{C}^N_a(t+1)\right)\right]. \end{array}$$

We then get equations similar to (26) but with a $\sup_a$ before both expectation. The supremum can be removed using the fact that for any functions f and g: $|\sup_a f(a) - \sup_a g(a)| \le \sup_a |f(a) - g(a)|$. $\qquad\square$ $\qquad\qquad\square$