**Noname manuscript No.**
(will be inserted by the editor)

# A Mean Field Approach for Optimization in Discrete Time

**Nicolas Gast · Bruno Gaujal**

the date of receipt and acceptance should be inserted later

**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 discrete time 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 made 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.

**Keywords** Mean Field, Markov Decision Process, Brokering.

**CR Subject Classification** G.3 · C.4 · F.2.2

Nicolas Gast
Grenoble Université and LIG, 51 av Jean Kuntzman, 38330 Montbonnot
E-mail: nicolas.gast@imag.fr

Bruno Gaujal
INRIA and LIG, 51 av Jean Kuntzman, 38330 Montbonnot
E-mail: bruno.gaujal@imag.fr

## 1 Introduction

The general context of this paper is the optimization of the behavior of 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. At each step, the state of each object changes according to a probability kernel $K$, 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.

The seminal work of Kurtz (see for example [16]) initiated a long stream of work on the use of mean field techniques in performance evaluation. Several other papers [2, 7] 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 system in continuous time. The control and the optimization of systems with an intensity that goes to zero are investigated in [12]. In the present paper, the intensity is bounded away from zero so that time remains discrete at the limit. This requires a different approach to construct the limit.

In [9], 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 processes, the class of systems studied in [9] corresponds to the case where this family is reduced to a unique process and no decision can be made. Here, we show that under similar conditions as in [9], 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 [6]. 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.

Actually, the contributions of this paper concern three types of systems. The first one is a class of Markov decision processes where the sequence of actions is fixed. In some sense, their study could bowl down to considering classical Markovian systems. The second one is the class of MDPs where the policy is fixed. The last type of results concerns MDP under optimal control, where optimization issues come into the picture. The second type (controlled systems) plays a central role in this paper because the results on the other two types of systems can be seen as corollaries of the theorems established for them. Indeed, the results on systems with fixed sequences of actions are simple consequences of theorems on controlled systems by considering the special case of constant policies. As for the results on optimal control, they are obtained by taking the supremum over all policies in controlled system.

On a practical point of view, all this allows one to compute the optimal policy in a deterministic limit system which can often be done efficiently, and then to use this policy in the original random system as a good approximation of the optimal policy, which cannot be computed 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 [18]. 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 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 [11]. 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 $\mathcal{S} = \{1, \ldots, S\}$. Time is discrete and the state of the $n$th 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 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 $i$th component of $M^N(t)$ is

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

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 $\mathscr{P}_N(S)$. For each $N$, $\mathscr{P}_N(\mathcal{S})$ is a finite set. When $N$ goes to infinity, it converges to $\mathscr{P}(\mathcal{S})$ 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):

$$C^N(t+1) = g(C^N(t), M^N(t+1), a_t),$$

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

### 2.1 Actions and policies

At each time $t$, the system's state is $M^N \in \mathscr{P}_N(\mathcal{S})$. 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 $K(a, C)$ such that the probability that for any $n$, a object goes from state $i$ to state the $j$ is $K_{i,j}(a, C)$:

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

The evolutions of objects are supposed to be independent once $C$ is given. Moreover, we assume that $K_{i,j}(a, C)$ is continuous in $a$ and $C$. The assumption of independence of the users is a rather common assumption in mean field models [9]. However other papers [2,7] have shown that similar results can be obtained using asymptotic independence only (see [13] for results of this type). However, the kernel $K$ is not assumed to be irreducible. This allows for several classes of objects interacting under context $C$, as in [5,9].

Here, the focus is on Markov Decision Processes theory and on the computation of optimal policies. A policy $\pi = (\pi_1 \pi_2 \dots)$ specifies the decision rules to be used at each time slot. A decision rule $\pi_t$ is a procedure that provides an action at time $t$. In general, $\pi_t$ is a random measurable function that depends on the events $(M(0), C(0)) \dots (M(t), C(t))$ 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 [19]. In what follows, we will only focus on them and a policy $\pi$ will represent a sequence of functions $(\pi_t)_{t \geq 0}$ where each function $\pi_t : \mathscr{P}(\mathcal{S}) \times \mathbb{R}^d \to \mathcal{A}$ is deterministic.

For any policy $\pi$, 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 $\pi$. $(M_\pi^N(t), C_\pi^N(t))_{t \geq 0}$ is a sequence of random variables on $\mathscr{P}_N(\mathcal{S}) \times \mathbb{R}^d$.

### 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 infinite-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, $r_T(M^N(t), C^N(t))$. The expected reward of a policy $\pi$ 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.

Throughout the document, for all $(m, c) \in \mathscr{P}(\mathcal{S}) \times \mathbb{R}^d$, $\|(m, c)\|$ denotes the $L^\infty$ norm of the vector $(m, c) \in \mathbb{R}^{S+d}$: $\|(m, c)\| = \max(|m_1| \dots |m_s|, |c_1| \dots |c_d|)$.

(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 $K(a, C)$.

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

(A3) **Continuity of** $K, g, r$ – the mappings $(C, a) \mapsto K(a, C)$, $(C, M, a) \mapsto g(C, M, a)$ and $(M, C) \mapsto r_t(M, C)$ are continuous, Lipschitz continuous on all compact set.

(A4) **Almost sure initial state** – Almost surely, the initial measure $M^N(0), C^N(0)$ converges to a deterministic value $m(0), c(0)$. Moreover, there exists $B < \infty$ such that almost surely $\left\|C^N(0)\right\| \leq B$ where $\|C\| = \sup_i |C_i|$.

To simplify the notations, we choose the functions $K$ 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 $C^N(0)$ converges almost surely, $C^N(0)$ is almost surely bounded. Here we had a bound $B$ which is uniform on all events in order to be sure that the variable $C^N(0)$ is dominated by an integrable function. As $g$ is continuous and the sets $\mathcal{A}$ and $\mathscr{P}(\mathcal{S})$ are compact, this shows that for all $t$, there exists $B_t < \infty$ such that

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

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 grows (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 $N^{-1}\sqrt{\log(N)}$.

Because of Equation 1, $(M^N(t), C^N(t))$ always stays in a compact space when $t \in \{0 \dots T\}$. By assumption (A3), this implies that $K$, $g$ and $r$ are Lipschitz continuous and we denote by $L_K$, $L_g$ and $L_r$ their Lipschitz constants.

### 3.1 Controlled mean field

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

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

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 $i$ that moves to state $j$.

Let $\pi$ 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 $\pi$, will apply the action $A_\pi^N(t) \stackrel{\text{def}}{=} \pi_t(M_\pi^N(t), C_\pi^N(t))$. 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 $A_\pi^N$ (and is deterministic once $A_\pi^N$ is fixed).

The following theorem is the main result of convergence, showing that as $N$ grows, the gap between the stochastic system $M_\pi^N, C_\pi^N$ 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 $\pi$, 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 $t$:*

$$\mathbb{P}\left(\sup_{s\le t} \left\| (M_\pi^N(s), C_\pi^N(s)) - (m_{A_\pi^N}(s), c_{A_\pi^N}(s)) \right\| \ge \mathcal{E}_t(\epsilon, \epsilon_0^N)\right) \le 2tS^2 \exp(-2N\epsilon^2),$$

*where*

$$\begin{aligned} \epsilon_0^N &\stackrel{\text{def}}{=} \left\| (M^N(0), C^N(0)) - (m(0), c(0)) \right\|; \\ \mathcal{E}_0(\epsilon, \delta) &\stackrel{\text{def}}{=} \delta; \\ \mathcal{E}_{t+1}(\epsilon, \delta) &\stackrel{\text{def}}{=} \left(S\epsilon + (2 + L_K)\,\mathcal{E}_t(\epsilon, \delta) + L_K \mathcal{E}_t(\epsilon, \delta)^2\right) \max(1, L_g). \end{aligned}$$

*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.1.

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),*

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

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

### 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 $M_a^N(t), C^N(t)$ 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 $\mathbb{E}\left(r_t(M_a^N(t), C_a^N(t))\right)$ 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,A2, A3,A4), there exists a function $\mathcal{E}(N, \epsilon)$ such that:*

- $\lim_{N\to\infty,\epsilon\to 0} \mathcal{E}(N, \epsilon) = 0$,
- *for all policy $\pi$:*

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

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

*Proof*

$$\begin{aligned}
&\left|V_\pi^N\left(M^N(0),C^N(0)\right)-\mathbb{E}\left(v_{A_\pi^N}\left(m_{A_\pi^N}(0),c_{A_\pi^N}(0)\right)\right)\right| \qquad (3)\\
&\quad=\left|\mathbb{E}\left(\sum_{t=1}^{T} r_t(M_\pi^N(t),C_\pi^N(t))-r_t(m_{A_\pi^N}(0),c_{A_\pi^N}(0))\right)\right|\\
&\quad\le L_r\mathbb{E}\left(\max_{t\le T}\left\|(M_\pi^N(t),C_\pi^N(t))-(m_{A_\pi^N}(0),c_{A_\pi^N}(0))\right\|\right)
\end{aligned}$$

where $L_r$ is a Lipschitz constant of the function $r$.

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

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

where $D\stackrel{\text{def}}{=}\sup_{x\in B}\|x\|$ with $B$ a bounded set such that $\mathbb{P}\left((M_\pi^N(t),C_\pi^N(t))\in B\right)=1$ (see the remark above Equation (1)).

Let us define $\mathcal{E}(N,\epsilon_0^N)\stackrel{\text{def}}{=}\inf_{\epsilon>0} L_r\max_{t\le T}\mathcal{E}_t(\epsilon,\epsilon_0^N)\left(1-T^2S^2\exp(-2\epsilon N)\right)+DT^2S^2\exp(-2\epsilon N)$. 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:*

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

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

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

Conversely, let $\pi_*^N$ 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 $\epsilon_0^N \stackrel{\text{def}}{=} \left\|(M^N(0), C^N(0)) - (m(0), c(0))\right\|$. Since, $\lim_{N\to\infty} \mathcal{E}(N, \epsilon_0^N) = 0$, $\limsup_{N\to\infty} V_*^N \left(M^N(0), C^N(0)\right) \leq v_*(m(0), c(0))$.

This result has several practical consequences. Recall that the sequence of actions $a_0^* \dots a_{T-1}^*$ 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 as $N$ grows, the reward of the constant policy $a_0^*, \dots, a_{t-1}^*$ converges to the optimal reward. This implies that the difference between the reward of the best complete information policies and the best incomplete information policies vanishes. 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 $M^N(0), C^N(0)$:

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

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 $a^*$. Moreover, the system becomes very sensitive to random perturbations and therefore harder to analyze: the policy $\mu^*$ is not necessarily continuous and $M_\mu^N, C_\mu^N$ may not have a limit. In Section 4, a comparison between the performances of $a^*$ and $\mu^*$ is provided over an example and we see that the performance of $\mu^*$ is much better, especially for small 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 $1/\sqrt{N}$.

**Proposition 6** *Under assumptions (A1,A2,A3,A4), there exist constants $\beta_t, \beta'_t$ and a sequence of constants $e_t^N$ only dependent on the parameters of the system such that:*

$$\sqrt{N}\mathbb{E}\left(\left\|(M_\pi^N(t), C_\pi^N(t)) - (m_\pi(t), c_\pi(t))\right\| : \epsilon_0^N\right) \leq \beta_t + \beta'_t\sqrt{N}\epsilon_0^N + e_t^N \tag{5}$$

*where :*

- $\mathbb{E}\left(\cdot : \epsilon_0^N\right)$ *designates the expectation knowing* $\epsilon_0^N$.
- $\epsilon_0^N \overset{\text{def}}{=} \left\|\left(M^N(0), C^N(0)\right) - (m(0), c(0))\right\|$;
- $\beta_t, \beta'_t$ *are defined by* $\beta_0 = 0$, $\beta'_0 = 1$ *and for all* $t \geq 0$*:*

$$\begin{aligned}\beta_{t+1} &= \max\{1, L_g\}\left(\left(S + L_K + 1\right)\beta_t + \frac{S}{2}\right);\\ \beta'_{t+1} &= \max\{1, L_g\}\left(S + L_K + 1\right)\beta'_t;\end{aligned}$$

- *There exists a constant* $C > 0$ *such that* $e_0^N = 0$ *and*

$$e_{t+1}^N = \max\{1, L_g\}\left(\left(S + L_K + 1\right)e_t^N + C\sqrt{\frac{\log(N)}{N}}\right).$$

*In particular, for all t:* $\lim_{N\to\infty} e_t^N = 0$.

*Proof* See appendix A.2.

An almost direct consequence of the previous result is the next theorem.

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

- *For any policy $\pi$:*

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

-

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

This theorem is the main result of this section. The previous result (Theorem 5) says that $\limsup_{N\to\infty} V_T^{*N}(M^N(0), C^N(0)) = \limsup_{N\to\infty} V_{a_0^*\ldots a_{T-1}^*}^N(M^N(0), C^N(0)) = v_*(m(0), c(0))$. 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 $\sqrt{N}$. When the parameters of the system are not Lipschitz but differentiable, this results can be improved by showing that the term in $1/\sqrt{N}$ has a Gaussian law (see Theorem 8).

*Proof* For any policy $\pi$, the expected reward of the stochastic system and the expected reward of the deterministic limit under actions $A_\pi^N$ 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,and $f$ is a real function Lipschitz of constant $L$ then $\mathbb{E}\left(|f(X) - f(Y)| \leq L\,\|X - y\|\right)$.

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 $m(0), c(0)$).

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

where $\epsilon_0^N \stackrel{\text{def}}{=} \left\|M^N(0), C^N(0) - m(0), c(0)\right\|$ and $\pi_*^N$ 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:

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

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

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.

(A4-bis) **Initial Gaussian variable** – There exists a Gaussian vector $G_0$ of mean 0 with covariance $\Gamma_0$ such that the vector $\sqrt{N}((M^N(0), C^N(0)) - (m(0), c(0)))$ (with $S+d$ components) converges almost surely to $G_0$.

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

This differentiability condition is slightly stronger than the Lipschitz condition and is indeed false in many cases because of boundary conditions. The initial state condition is slightly stronger that (A4) but remains very natural. For example, if $C^N(0)$ if fixed to some $c(0)$ and if the initial states $X_1^N \ldots X_N^N$ of all objects are independent and identically distributed (*i.i.d.*), then $\sqrt{N}((M^N(0), C^N(0)) - (m(0), c(0)))$ converges in law to a Gaussian variable $G$ of the same covariance as $X_1^N$ – this is just the multidimensional central limit theorem, see for example Theorem 9.6 of Chapter 2 of [10]. 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 $t$:*

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

*Moreover if $\Gamma_t$ is the covariance matrix of $G_t$, then:*

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

*where for all $1 \leq i, j \leq S$ and $1 \leq k, \ell \leq d$: $(P_t)_{ij} = K_{ij}(a_t, c(t))$, $(Q_t)_{kj} = \sum_{i=1}^{S} m_i \frac{\partial K_{ij}}{\partial c_k}(a_t, c(t))$, $(F_t)_{ik} = \frac{\partial g_k}{\partial m_i}(m_{t+1}, c(t))$, $(H_t)_{k\ell} = \frac{\partial g_k}{\partial c_\ell}(m(t), c(t))$, $(D_t)_{jj} = \sum_{i=1}^{n} m_i (P_t)_{ij}(1 - (P_t)_{ij})$ 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 A.3.

### 3.4 Beyond square-root convergence

So far, we have 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 have shown 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 an even better approximation of the initial system. The purpose of this part is to show that this approximation is indeed better than mean field, in the sense that it leads to an error on the reward of order $\frac{\sqrt{\log N}}{N}$ instead of $\frac{1}{\sqrt{N}}$.

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

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

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

where $A^N(t) \stackrel{\text{def}}{=} \pi_t(\widetilde{M}_\pi^N(t), \widetilde{C}_\pi^N(t))$ and $G_t(a, \widetilde{C}_\pi^N(t))$ is a sequence of *i.i.d.* Gaussian random variables 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 $G_t(a,C)$ 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 $j \neq k$:

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

Notice that $\widetilde{M}^N$ is not a positive measure anymore, but a signed element of $\mathbb{R}^S$. The Lipschitz functions $r_t$ and $g$ are originally only defined for positive vectors but can be extended to $\mathbb{R}^S \times \mathbb{R}^d$ (theorem of Kirszbraun, [21]) with the same Lipschitz constants.

In the following, the process $(\widetilde{M}^N_\pi(t), \widetilde{C}^N_\pi(t))$ is called the *mean field Gaussian approximation* of $(M^N_\pi(t), C^N_\pi(t))$. As for the definition of $V^N_\pi\left(M^N(0), C^N(0)\right)$, we define the expected reward of the mean field Gaussian approximation by:

$$W^N_\pi\left(M^N(0), C^N(0)\right) = \mathbb{E}\left(\sum_{t=1}^{T} r_t\left(\widetilde{M}^N_\pi(t), \widetilde{C}^N_\pi(t)\right)\right).$$

The optimal cost of the mean field Gaussian approximation starting from the point $\left(M^N(0), C^N(0)\right)$ is $W^N_*\left(M^N(0), C^N(0)\right) = \sup_\pi W^N_\pi\left(M^N(0), C^N(0)\right)$. The following result shows that the Gaussian approximation is indeed a very accurate approximation of the original system.

**Theorem 9** *Under assumptions (A1,A2,A3,A4), there exists a constant $H$ independent of $M^N, C^N$ such that*

*(i) for all sequence of actions $a = a_1 \dots a_T$:*

$$\left|V^N_a(M^N, C^N) - W^N_a(M^N, C^N)\right| \leq H\frac{\sqrt{\log(N)}}{N}.$$

*(ii)*

$$\left|V^N_*(M^N, C^N) - W^N_*(M^N, C^N)\right| \leq H\frac{\sqrt{\log(N)}}{N}.$$

*Proof* The proof is detailed in Appendix A.4.

## 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 [17,3,4].

Here, time is discrete and the $A$ sources follow a discrete on/off model: for each source $j \in \{1 \dots A\}$, let $(Y_j(t)) \stackrel{\text{def}}{=} 1$ if the source is on (*i.e.* it sends a task between $t$ and $t+1$) and 0 if it is off. The total number of tasks sent between $t$ and $t+1$ is $Y(t) \stackrel{\text{def}}{=} \sum_j Y_j(t)$. Each queue $i \in \{1 \dots d\}$ is composed of $P_i$ processors, and all of them work at speed $\mu_i$ when available. Each processor $j \in \{1 \dots P_i\}$ of the queue $i$ can be either *available* (in that case we set $X_{ij}(t) \stackrel{\text{def}}{=} 1$) or *broken* (in that case

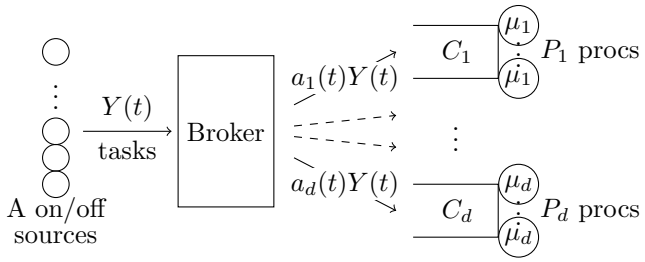



**Fig. 1** The routing system

$X_{ij}(t) \stackrel{\text{def}}{=} 0$). The total number of processors available in the queue $i$ between $t$ and $t+1$ is $X_i(t) \stackrel{\text{def}}{=} \sum_j X_{ij}(t)$ and we define $B_i(t)$ 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 $a(t) \in \mathscr{P}\left(\{1 \dots Y_t\}^d\right)$ and routes each of the $Y(t)$ tasks to queue $i$ with probability $a_i(t)$. The system is represented figure 1. The number of tasks in the queue $i$ (buffer size) evolves according to the following relation:

$$B_i(t+1) = \Big(B_i(t) - \mu_i X_i(t) + a_i(t)Y(t)\Big)^+. \tag{8}$$

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 $r_t(B) \stackrel{\text{def}}{=} \sum_i B_i(t)$. As we consider a finite horizon, we should decide a cost for the remaining tasks in the queue. In our simulations, we choose $r_T(B) \stackrel{\text{def}}{=} \sum_i B_i(T)$.

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 [23]. Here, indexability may help to compute near optimal policies by solving one MDP for each queue [23,22]. 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: $a_d(t)$ is the probability that a task is sent to queue $d$ at time $t$.
- The environment of the system depends on the vector $B(t) = (B_1(t) \dots B_d(t))$, giving the number of tasks in queues $q_1, \dots q_d$ at time $t$. The time evolution of the i-th component is

$$B_i(t+1) = g_i(B(t), M^N(t+1), a(t)) \stackrel{\text{def}}{=} \Big(B_i(t) - \mu_i X_i(t) + a_i(t)Y(t)\Big)^+.$$

The shared environment is represented by the context $C^N(t) \stackrel{\text{def}}{=} (\frac{B_1(t)}{N} \dots \frac{B_d(t)}{N})$.

- 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 $\mu^*$, when $N$ goes to infinity the system converges to a deterministic system with optimal cost. An explicit computation of the policies $a^*$ and $\mu^*$ 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 $\mu_i X_i(t)$ and $Y(t)$ converge almost surely to deterministic values that we call $x_i(t)$ and $y(t)$. If $y_i(t)$ is the number of tasks distributed to the $i$th queue at time $t$, $c_i(t+1) = (c_i(t) + y_i(t) - x_i(t))^+$. The deterministic optimization problem is to compute

$$\min_{y^1(1)\dots y^d(T)} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{d} c_i(t) \text{ with } \left\{ \begin{array}{l} c_i(t+1) = (c_i(t) + y_i(t) - x_i(t))^+ \\ \sum_i y_i(t) = y(t) \end{array} \right\} \right. . \tag{9}$$

Let us call $w_i(t)$ the work done by the queue $i$ at time $t$: $w_i(t) = c_i(t) - c_i(t-1) + y_i(t-1)$. 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:

$$\sum_{i=1}^{d} c_i(t) = \sum_{i=1}^{d} c_i(0) - \sum_{u \leq t, i} w_i(t).$$

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. See [11] for more details.

The principle of the algorithm is the following.

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

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 examples

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 and the time horizon $T$ equals 40. 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 $a_0^* \dots a_{T-1}^*$ 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 $\mu^*$ – here, the optimal policy $\mu^*$ for the deterministic case was first computed. When the stochastic system is in state $(M^N(t), C^N(t))$ at time $t$, we apply the action $\mu_t^*(M^N(t), C^N(t))$ 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 task is routed (deterministically) in the shortest queue. In W-JSQ, a task 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 $V_*^N(M(0), C(0))$. This figure illustrates Theorem 5: if we apply $a^*$ or $\mu^*$, 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 $\mathbb{E}\left(A^N\right) = \frac{N}{N_0} A$, 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

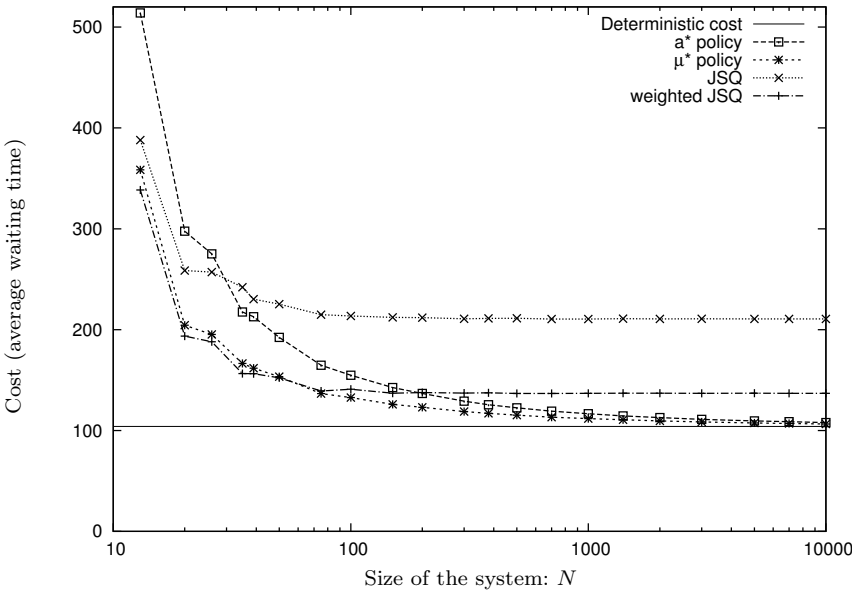



**Fig. 2** Expected cost of the policies $a^*$, $\mu^*$, JSQ and W-JSQ for different values of $N$.

conjecture [20] that says that for a given load, the average queue length decreases when the arrival and service processes are more deterministic.

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 $\mu^*$ 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 $\mu^*$ (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 $\delta$ defined in Equation (7).

Besides comparing $a^*$ and $\mu^*$ 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.

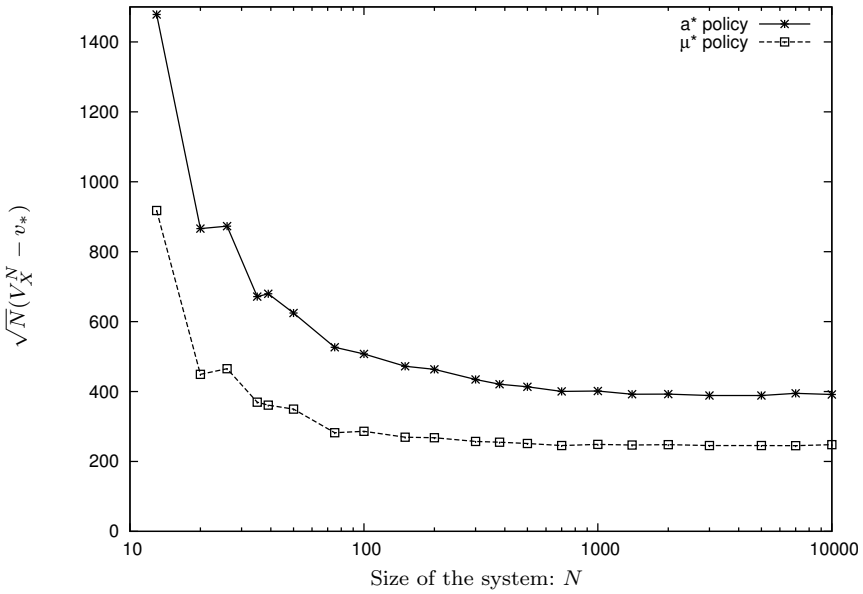



**Fig. 3** Speed of convergence of the policies $X = a^*$ or $\mu^*$ for different values of $N$.

## 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.

### 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 framework 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 $\mathcal{X}^N(t) = (X_1^N(t) \dots X_N^N(t))$ 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 $i$th object. We assume that $\mathcal{A}$ is finite. The reward and the evolution of the context is defined as before and we call $V_{od*}^N(\mathcal{X}^N(0), C^N(0))$ the optimal reward of this system over a finite horizon $[0; T]$ where *od* stands for "Object-Dependent"-actions.

As before, $M^N(0) \stackrel{\text{def}}{=} N^{-1} \sum_{n=1}^N \delta_{X_n^N(0)}$ is the empirical measure of $\mathcal{X}^N(0)$. We consider our original problem in which we replace the action set $\mathcal{A}$ by $\mathscr{P}(\mathcal{A})^{\mathcal{S}}$. 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 $K(a, C)$.

Compared to the problem in which the action sent is object-dependent, the action of the controller in the latter case is 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 $g, K, \mathcal{A}, M^N(0), C^N(0)$ satisfy assumptions (A1,A2,A3,A4), then the object-dependent reward $V^N_{od*}$ converges to the deterministic limit:*

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

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

*Proof (Sketch of proof)* To each $N$-uple $a = a_1 \dots a_N \in \mathcal{A}^N$ 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

$$(p_{a,\mathcal{X}^N})_i \stackrel{\text{def}}{=} \frac{1}{\sum_{n=1}^N \mathbf{1}_{X_n^N=i}} \sum_{n=1}^N \mathbf{1}_{X_n^N=i}\delta_{a_n}.$$

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 $\mathcal{X}^N(t), C^N(t)$ and applying action $a$ or $p_{a,\mathcal{X}^N(t)}$ both lead to the same almost sure limit for $(M^N(t+1), C^N(t+1))$. Then one can show by induction on the time-horizon $T$ that the reward under a fixed sequence of action $V^N_{od*}(\mathcal{X}^N(0), C^N(0))$ is asymptotically equal to $V^N_*(M^N(0), C^N(0))$.

Then remarking that $(\mathscr{P}(\mathcal{A}))^{\mathcal{S}}$ also satisfies hypothesis (A2) and (A3) – it is compact and the mappings $g$, $K$ and $r$ are uniformly continuous if $p_a \in \mathscr{P}(\mathcal{A})$ – 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 $M, C, a$ $r_t(M, C) = r(M, C)$ and $K_t(a, C) = K(a, C)$.

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

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 $\pi$, the expected total discounted reward of $\pi$ is ($\delta$ is omitted in the notation):

$$V^N_\pi(M^N(0), C^N(0)) \stackrel{\text{def}}{=} \mathbb{E}^\pi\Big(\sum_{t=1}^\infty \delta^{t-1} r(M^N(t), C^N(t))\Big).$$

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

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

As $r$ is bounded by $K < \infty$, the gap between $V_{T*}^N$ and $V_*^N$ can be bounded independently of $N, M, C$:

$$\left| V_{T*}^N(M, C) - V_*^N(M, C) \right| \leq K \sum_{t=T+1}^{\infty} \delta^t = K \frac{\delta^{T+1}}{1-\delta}. \tag{10}$$

In particular, this shows that $V_{T*}^N$ converges uniformly in $(M, C)$ and $N$ to $V_*^N$ as $T$ goes to infinity:

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

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 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:*

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

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

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

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 $T$ goes to infinity.

**Proposition 12** *Under assumptions (A1,A2,A3,A4,A6,A7) and if the functions $c \mapsto K(a, c)$, $(m, c) \mapsto g(c, m, a)$ and $(m, c) \mapsto r(m, c)$ are Lipschitz with constants $L_K$,$L_g$ and $L_r$ satisfying $\max(1, L_g)(S + L_K + 1)\delta < 1$, the constants $H \stackrel{\text{def}}{=} L_r \sum_t \delta^t \beta_t$ and $H' \stackrel{\text{def}}{=} L_r \sum_t \delta^t \beta'_t$ yield*

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

*where $\epsilon_0^N \stackrel{\text{def}}{=} \left\| \left( M^N(0), C^N(0) \right) - (m(0), c(0)) \right\|$ and $\beta_t$ and $\beta'_t$ are defined in Proposition 6.*

*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,

$$\sqrt{N}\mathbb{E}\left(\left\|\left(M_\pi^N(t), C_\pi^N(t)\right) - \left(m_{A_\pi^N}(t), c_{A_\pi^N}(t)\right)\right\|\right) \leq \beta_t + \beta'_t\sqrt{N}\epsilon_0^N + e_t^N, \qquad (11)$$

the growth of the constants $\beta_t$ and $\beta'_t$ and $e_t^N$ is bounded by a factor $\max(1, L_g)(S + L_K + 1)$.

*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 converge if $\delta$ does not satisfy the assumptions of Proposition 12. The system is defined as follows:

- The state space is $\mathcal{S} = \{0, 1\}$ and the context is $C(t) \stackrel{\text{def}}{=} M_0^N(t)$ ($C$ is the mean number of particles in state 0). Therefore the interesting process is $C(t)$.
- For any object, the probability of going to state 0 is $f(C)$ (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(a).
- The starting point is $M_0^N(0) = C(0) = \frac{1}{2}$.
- The reward function is $r(M, C) = \left|C - \frac{1}{2}\right|$.

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 $\sum_{t=0}^{\infty}\delta^t\sqrt{N}\left|C^N(t) - \frac{1}{2}\right|$.

The process $C^N(t) - \frac{1}{2}$ is quite complicated. However for any finite horizon $[0; T]$, if $N$ is large enough, $C^N(t) - \frac{1}{2}$ is close 0 with high probability. During this time, $\sqrt{N}(C^N(t) - \frac{1}{2})$ is 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 $\delta \geq 1/L$, $\mathbb{E}\left(\sum_{i=0}^{t-1}\delta^t\left|X_t^N\right|\right)$ goes to $+\infty$ as $t$ goes to infinity. By choosing $T$ large enough, this shows that $\sum_{t=0}^{\infty}\delta^t\sqrt{N}\left|C^N(t) - \frac{1}{2}\right|$ is not bounded as $N$ grows.

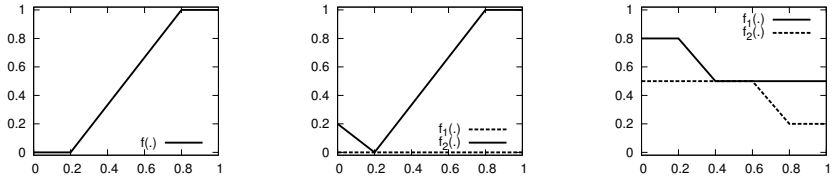
### 5.3 Average Reward

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])

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

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 $C(t)$ is the mean number of objects in state 1 ($C(t) = (M(t))_1$) and with a function $f:[0; 1]\to[0; 1]$ such that the

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




(a) Example 1 (b) Example 2 (c) Example 3

**Fig. 4** The transitions functions of Examples 1, 2 and 3 (from left to right). On each figure is draw the probability for an object to go in state 1 as a function $M_0^N$ for the different actions.

transition kernel $K$ is $K_{i1}(C) = f(C)$ for $i \in \mathcal{S}$. If $M_0^N(0) \xrightarrow{\text{a.s.}} m_0$ then for any fixed $t$, $M^N(t)$ converges to $f(f(\dots f(m(0))\dots))$. However, in general we may have $\lim_{t\to\infty} \lim_{N\to\infty} M^N(t) \neq \lim_{N\to\infty} \lim_{t\to\infty} M^N(t)$. 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 only takes values in $\{0; 1\}$. In some situations, such as Example 2, the optimal policy of the deterministic limit differs from the optimal policy for the stochastic system.

*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 equal to $f_1(C)$ and $f_2(C)$ respectively, defined by (see Figure 4(b)).

- $f_1(C) = 0$.
- $f_2$ is piecewise linear with $f_2(0) = 0.2, f_2(0.1) = 0, f_2(0.8) = 1, f_2(1) = 1$.

The reward function is $r(C) = |C - 0.1|$.

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 $[0; 0.5)$ 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 0.1).

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 something greater than 0.8 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 [8] 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 $C$ is bounded (from above and from below). This implies that the couple $(M, C)$ lives in a compact set $B \subset \mathbb{R}^{S+d}$. Let $f_a : B \to B$ denote the deterministic function corresponding to one step of the evolution of the

deterministic limit under action $a$. The definition of $f_a$ is given by Equation (2):

$$f_a(m,c) = (m',c') \text{ with } \begin{cases} m' &= m \cdot K(a,c) \\ c' &= g(c,m',a). \end{cases}$$

We say that a set $H$ is an attractor of the function $f_a$ if

$$\lim_{t\to\infty} \sup_{x\in B} d(f_a^t(x), H) = 0,$$

where $d(x,H)$ denotes the distance between a point $x$ and a set $H$ and $f_a^t(x)$ denotes $t$ iterations of $f_a$ applied to $x$: $f_a^t(x) = f_a(f_a(\dots f_a(x)))$.

The following proposition shows that as $t$ goes to infinity and $N$ goes to infinity, $(M^N(t), C^N(t))$ concentrates around the attractors of $f_a$.

**Proposition 13** *Under assumptions (A1,A2,A3), if the controller always chooses action $a$ then for any attractor $H$ of $f_a$ and for all $\epsilon > 0$:*

$$\lim_{N\to\infty} \limsup_{t\to\infty} \mathbb{P}\left(d\left(\left(M_a^N(t), C^N(t)\right), H\right) \geq \epsilon\right) = 0$$

*Proof* Let $\epsilon > 0$. Since $H$ is an attractor, there exists $T$ such that

$$\sup_{x\in H} d(f_a^T(x), H) \leq \epsilon/2.$$

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

$$d(X^N(t+T), H) \leq \left\| X^N(t+T) - f_a^T(X^N(t)) \right\| + d(f_a^T(X^N(t)), H).$$

By Theorem 1, the first part of this inequality is less than $\mathcal{E}_T(\delta, 0)$ with probability greater than $\exp(-2N\delta^2)$. Moreover, $\mathcal{E}_T(\delta,0)$ converges to 0 as $\delta$ goes to 0. Therefore, there exists $\delta$ such that $\mathcal{E}_T(\delta,0) < \epsilon/2$. This implies that for such $\delta$ and all $t \geq 0$,

$$\mathbb{P}\left(d(X^N(t+T), H) \geq \epsilon\right) \leq 2TS^2 \exp(-2N\delta^2),$$

which goes to 0 as $N$ goes to infinity.

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

**Corollary 14** *If the function $f_a$ has a unique attractor $(m_\infty, c_\infty)$, then*

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

In the controlled case, there is no simple positive result under assumptions that are easy to check in practice. In particular, assuming that all $f_a$ have the same attraction point, 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 $C^N = M_0^N$ is the proportion of objects in state 0. The only difference here is that there are two possible actions 1 and 2, corresponding to a probability of transition from any state to 0 of $f_1(C)$ and $f_2(C)$. Both $f_1$ and $f_2$ are piecewise linear functions taking the values:

– $f_1(x) = 0.8$ for $x \leq 0.2$, $0.5$ for $x > 0.4$;
– $f_2(x) = 0.5$ for $x \leq 0.6$, $0.2$ for $x > 0.6$.

Figure 4(c) shows the transition functions. The reward is set to $|C^N - 1/2|$.

Both $f_1$ and $f_2$ have the same attractor, equal to $\{1/2\}$. Moreover, one can prove that under any policy, $\lim_{N\to\infty} \lim_{t\to\infty} M_\pi^N(t)$ will converge to 0.5, leading to an average reward of 0 regardless of the initial condition. However, if the deterministic limit starts from the point $C^N(0) = .2$, then by choosing the sequence of actions $1, 2, 1, 2 \ldots$ the system will oscillate between 0.2 and 0.8, leading to an average reward of 0.3.

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 points: $0.2, 0.5$ and $0.8$.

## 6 Computational issues

Throughout this paper, we have shown that if the controller uses the optimal policy $\mu^*$ 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 [14] for a review on the subject). Imagine that a 2-approximation algorithm exists for the deterministic system, then, Theorem 5 proves that for all $\varepsilon$, 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

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

where $v_{t\ldots T*}$ denotes the optimal reward of the deterministic limit over finite horizon $\{t \ldots T\}$. 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:

$$V^N_{t\ldots T*}(M, C) = r_t(M, C) + \sup_{a\in\mathcal{A}} \mathbb{E}\left(V^N_{t+1\ldots T*}\left(\Phi^N_a(M, C)\right)\right).$$

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 $\phi$ 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 $|\mathscr{P}_N(\mathcal{S})|$ possible values of $M^N_{t+1}$ 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.

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

## 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. Andrea Bobbio, Marco Gribaudo, and Miklos Telek. Analysis of large scale interacting systems by mean field method. In *5th International Conference on Quantitative Evaluation of Systems(QEST)*, pages 215–224, St Malo, 2008.
6. C. Bordenave and V. Anantharam. Optimal control of interacting particle systems. Technical Report 00397327, CNRS Open-Archive HAL, 2007.
7. 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.
8. V. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.
9. 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.
10. R. Durrett. *Probability: theory and examples*. Wadsworth & Brooks/Cole, 1991.
11. 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.
12. N. Gast, B. Gaujal, and J.Y. Le Boudec. Mean field for Markov Decision Processes: from Discrete to Continuous Optimization. Technical report, INRIA, 2010.
13. 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.
14. D.S. Hochbaum. *Approximation algorithms for NP-hard problems*. PWS Publishing Co. Boston, MA, USA, 1996.
15. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
16. T. Kurtz. *Strong approximation theorems for density dependent Markov chains*. Stochastic Processes and their Applications. Elsevier, 1978.
17. 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).
18. Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queueing network control. *Math. Oper. Res.*, 24:293–305, 1999.
19. M.L. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.
20. 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.

21. J. T. Schwartz. *Nonlinear functional analysis*. Gordon and Breach Science Publishers, New York, 1969.
22. R. R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27:637–648, 1990.
23. 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.

## A Appendix: proofs

### A.1 Proof of Theorem 1 (controlled mean field)

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

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

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

Let $B_{inj}^N \stackrel{\text{def}}{=} \mathbf{1}_{H_{ij}(A_\pi^N(t),C^N(t))\le U_{i,n}^N(t)\le H_{ij+1}(A_\pi^N(t),C^N(t))}$. $(B_{inj}^N)_{i,n,N}$ are *i.i.d.* Bernoulli random variable with mean $\mathbb{E}\left(B_{inj}^N | A_\pi^N(t)=a, C^N(t)=c\right) = K_{ij}(a,c)$. Therefore, by Hoeffding's inequality (Inequality 2.3 of [15]), we have:

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

Therefore, the quantity $\left|\sum_{n=1}^{NM_i^N(t)} B_{inj}^N(t) - NM_i^N(t)K_{ij}(A_\pi^N(t),C^N(t))\right|$ is less than $N\epsilon$ for all $i,j$ with probability greater than $1-2S^2\exp(-2N\epsilon^2)$. If this holds for all $i,j$ and if $\left\|(M^N(t),C^N(t)) - (m(t),c(t))\right\| \le \epsilon_t$, then for all $1\le j\le S$, the gap at time $t+1$ between the $j$th component of $M^N$ $(M_j^N(t+1))$ and $m$ $(m_j(t+1))$ is:

$$\begin{aligned}
\left|M_j^N(t+1) - m_j(t+1)\right| &= \left|\sum_{i=1}^{S}\frac{1}{N}\sum_{n=1}^{NM_i^N(t)} B_{inj}^N(t) - m_i(t)K_{ij}(A_\pi^N(t),c(t))\right| \\
&\le \sum_{i=1}^{S}\frac{1}{N}\left|\sum_{n=1}^{NM_i^N(t)} B_{inj}^N(t) - NM_i^N(t)K_{ij}(A_\pi^N(t),C^N(t))\right| \\
&\quad + \sum_{i=1}^{S}\left|\left(M_i^N(t) - m_i(t)\right)\right| K_{ij}(A_\pi^N(t),C^N(t)) \\
&\quad + \sum_{i=1}^{S} m_i(t)\left|K_{ij}(A_\pi^N(t),C^N(t)) - K_{ij}(A_\pi^N(t),c(t))\right| \\
&\quad + \left|\sum_{i=1}^{S}\left(M_i^N(t) - m_i(t)\right)\left(K_{ij}(A_\pi^N(t),C^N(t)) - K_{ij}(A_\pi^N(t),c(t))\right)\right| \\
&\le S\epsilon + S\epsilon_t + L_K\epsilon_t + SL_K\epsilon_t^2. \qquad (13)
\end{aligned}$$

where we use (12) for the first part of the inequality, the fact that $K_{ij} \leq 1$ for the second and the fact that $\sum_{i=1}^{S} m_i(t) = 1$ and that $K$ is Lipschitz with constant $L_K$ for the third one.

Moreover, using the fact that $g$ is Lipschitz with constant $L_g$ ($\|g(c,m,a) - g(c',m',a)\| \leq L_g \|(c,m) - (c',m')\|$), we have

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

This implies that

$$\begin{aligned}\left\|(M^N(t+1), C^N(t+1)) - (c(t+1), m(t+1))\right\| &\leq (S\epsilon + S\epsilon_t + L_K\epsilon_t + SL_K\epsilon_t^2)\max(L_g, 1) \\ &\leq \mathcal{E}_{t+1}\left(\epsilon, \left\|M^N(0), C^N(0) - m(0), c(0)\right\|\right),\end{aligned}$$

where

$$\mathcal{E}_{t+1}(\epsilon, \delta) = \left(S\epsilon + (2 + L_K)\mathcal{E}_t(\epsilon, \delta) + L_K\mathcal{E}_t(\epsilon, \delta)^2\right)\max(1, L_g).$$

By a direct induction on $t$, this holds with probability greater than $2(t+1)S^2\exp(-2N\epsilon^2)$.

## A.2 Proof of Theorem 6 (second order result)

We are interested in comparing the behavior of $(M_\pi^N(t), C_\pi^N(t))$ and $(m_{A_\pi^N}(t), c_{A_\pi^N}(t))$ where $A_\pi^N$ is the sequence of actions taken by the controller under policy $\pi$ ($A_\pi^N$ 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 $A_\pi^N$. In order to improve the readability, we suppress the indexes $\pi$ and $A_\pi^N$. Variables $M_\pi^N(t), C_\pi^N(t), m_{A_\pi^N}(t), c_{A_\pi^N}(t)$ will be denoted $M^N(t), C^N(t), m(t), c(t)$, respectively.

We will show Equation (5) by induction on $t$. Let us first recall that Equation (5) is:

$$\sqrt{N}\mathbb{E}\left(\left\|\left(M^N(t), C^N(t)\right) - (m(t), c(t))\right\|\right) \leq \beta_t + \beta_t'\sqrt{N}\epsilon_0^N + 1 + e_t^N,$$

where $\epsilon_0^N \stackrel{\text{def}}{=} \mathbb{E}\left(\left\|\left(M^N(0), C^N(0)\right) - (m(0), c(0))\right\|\right)$.

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:

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

The central limit theorem shows that $\sqrt{N}(M^N(t+1) - M^N(t)P^N(t))$ converges in law to a Gaussian vector of mean 0 and covariance $D$ where $D$ is defined in Equation (15). Moreover,

by Lemma 18, there exists a constant $\alpha_2 > 0$ such that

$$\begin{aligned}\sqrt{N}\mathbb{E}\left(\left\|M^N(t+1) - M^N(t)P^N(t)\right\|\right) &\leq \sum_i^S \sqrt{N}\mathbb{E}\left(\left|M^N(t+1)_i - (M^N(t)P^N(t))_i\right|\right) \\ &\quad + \alpha_2\sqrt{\frac{\log(N)}{N}} \\ &= \sum_{i=1}^S \sqrt{m_i p_{ij}(1-p_{ij})} + \alpha_2\sqrt{\frac{\log(N)}{N}} \\ &\leq \frac{S}{2} + \alpha_2\sqrt{\frac{\log(N)}{N}}. \end{aligned} \tag{14}$$

The other terms can be bounded using similar ideas as for proving Equation (13) in the proof of Theorem 1. Since $p_{ij} \leq 1$ and $\sum_{i=1}^S m_i(t) = 1$, we have for all $1 \leq j \leq S$:

$$\left|\left(M^N(t) - m(t)\right)P^N(t)\right|_j \leq \sum_{i=1}^S \left|M^N(t) - m(t)\right|_i P^N_{ij}(t) \leq S\left\|M^N(t) - m(t)\right\|; \tag{15}$$

$$\left|m(t)\left(P^N(t) - p(t)\right)\right|_j \leq \sum_{j=1}^S m_j(t)\left|P^N_{ij}(t) - p_{ij}(t)\right| \leq L_K\left\|C^N(t) - c(t)\right\|. \tag{16}$$

Combining Equations (14), (15) and (16), we get

$$\begin{aligned}\sqrt{N}\mathbb{E}\left(\left\|M^N(t+1) - m(t+1)\right\|\right) &\leq \frac{S}{2} + (S + L_K)\left\|(M^N(t), C^N(t)) - (m(t), c(t))\right\| \\ &\quad + \alpha_2\sqrt{\frac{\log(N)}{N}}.\end{aligned}$$

Since $C^N(t+1) = g(C^N(t), M^N(t), a)$ where $(c, m) \mapsto g(c, m, a)$ is a deterministic Lipschitz function with constant $L_g$, we have:

$$\begin{aligned}\left\|C^N(t+1) - c(t+1)\right\| &\leq L_g \max\left(\left\|M^N(t+1) - m(t+1)\right\|, \left\|C^N(t) - c(t)\right\|\right) \\ &\leq L_g\left(\left\|M^N(t+1) - m(t+1)\right\| + \left\|C^N(t) - c(t)\right\|\right) \\ &\leq L_g\Bigg(\frac{S}{2} + (S + L_K + 1)\left\|\left(M^N(t), C^N(t)\right) - (m(t), c(t))\right\| \\ &\quad + \alpha_2\sqrt{\frac{\log(N)}{N}}\Bigg).\end{aligned}$$

Using the induction hypothesis, this implies that

$$\begin{aligned}&\sqrt{N}\mathbb{E}\left(\left\|\left(M^N(t+1), C^N(t+1)\right) - (m(t+1), c(t+1))\right\|\right) \\ &\qquad \leq \max\{1, L_g\}\left(\frac{S}{2} + (S + L_K + 1)\left(\beta_t + \beta'_t\epsilon^N_0 + e^N_t\right) + \alpha_2\sqrt{\frac{\log(N)}{N}}\right).\end{aligned}$$

## A.3 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 measures on $\{1, \ldots, S\}$ and $P^N$ a sequence of random stochastic matrices on $\{1, \ldots, S\}$ such that $(M^N, P^N) \xrightarrow{\text{a.s.}} (m, p)$.*

*Let* $(U_{ik})_{1\leq i\leq S, k\geq 1}$ *be a collection of* i.i.d. *random variables following the uniform distribution on* $[0;1]$ *and independent of* $P^N$ *and* $M^N$ *and let us define* $Y^N$ *such that for all* $1\leq j\leq S$,

$$Y_j^N \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{S}\sum_{k=1}^{NM_i^N} \mathbf{1}_{\sum_{l<j} P_{il}^N < U_{ik} \leq \sum_{l\leq j} P_{il}^N}.$$

*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*

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

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

$$\begin{cases} D_{jj} = \sum_i m_i p_{ij}(1-p_{ij}) \\ D_{jk} = -\sum_i m_i p_{ij} p_{ik} \qquad (j\neq k). \end{cases} \tag{17}$$

*Proof* 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 $\Omega$ and $\Omega'$. For all $(\omega,\omega')\in\Omega\times\Omega'$, we define $X_\omega^N(\omega') \stackrel{\text{def}}{=} \sqrt{N}(Y^N(\omega,\omega') - M^N(\omega)P^N(\omega))$.

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

Therefore for almost all $\omega$, there exists a random variable $\overline{X}_\omega^N$ with the same law as $X_\omega^N$ that converges $\omega'$-almost surely to $G(\omega')$ (see [10] for details on that representation). Let $Z^N(\omega,\omega') \stackrel{\text{def}}{=} M^N(\omega)P^N(\omega) + \frac{1}{N}\overline{X}_\omega^N(\omega')$. By construction of $\overline{X}_\omega^N$, for almost all $\omega$, $Z^N(\omega,.)$ 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 $\overline{Z}^N(\omega,.)$ that has the same distribution as $Y^N(\omega)$ for all $\omega$ and that converges $(\omega,\omega')$-almost surely to $G$.

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 $\sqrt{N}((M^N, C^N)(t) - (m,c)(t))$ converges in law to $G_t$, there exists another probability space and random variables $\overline{M}^N$ and $\overline{C}^N$ with the same distribution as $M^N$ and $C^N$ such that $\sqrt{N}((\overline{M}^N, \overline{C}^N)(t) - (m,c)(t))$ converges almost surely to $G_t$ [10]. In the rest of the proof, by abuse of notation, we will write $M$ and $C$ instead of $\overline{M}$ and $\overline{C}$ and then we assume that $\sqrt{N}((M^N(t), C^N(t)) - (m,c)(t)) \xrightarrow{\text{a.s.}} G_t$.

$G_t$ 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 $\Delta$ of size $(S+d)\times(S+d)$ such that $G_t = \Delta U$.

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

$$\sqrt{N}(\overline{M}(t+1)^N - M(t)^N P_t^N) \xrightarrow{\text{a.s.}} H_t. \tag{18}$$

In the following, by abuse of notation we will also write $M$ instead of $\overline{M}$. Therefore we have

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

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

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

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

$$H_t + \sum_{k=1}^{d} Q_{kj}(\Delta U)_{S+k} + \sum_{i=1}^{S} (\Delta U)_i P_{ij} \tag{19}$$

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

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

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

For $1 \le j, j' \le S$, $\Delta'_{j,j'}$ is the expectation of (19) taken in $j$ times (19) taken in $j'$. Using the facts that
$\mathbb{E}\left((\Delta U)_{S+k}(\Delta U)_{S+k'}\right) = C_{kk'}$, $\mathbb{E}\left((\Delta U)_{S+k}(\Delta U)_i\right) = O_{ik}$ and $\mathbb{E}\left((\Delta U)_i(\Delta U)_{i'}\right) = \Delta_{ii'}$, this leads to:

$$\Delta'_{j,j'} = \mathbb{E}\left(H_j H'_j\right) + \sum_{k,k'} Q_{kj}Q_{k'j'}C_{kk'} + \sum_{k,i'} Q_{kj}O_{i'k}P_{i'j'}$$
$$+ \sum_{i,k'} Q_{k'j'}O_{ik'}P_{ij} + \sum_{i,i'} P_{ij}\Delta_{ii'}P_{i'j}$$
$$= D_{jj'} + (Q^T C Q)_{jj'} + (Q^T O^T P)_{jj'} + (P^T O Q)_{jj'} + (P^T \Delta P)_{jj'}.$$

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

## A.4 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 $B_i$ be independent Bernoulli random variables ($i.e$ $P(B_i = 1) = 1 - P(B_i = 0) = p$) and let $Y^N = \frac{1}{N}\sum_{i=1}^{N} B_i$. 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 $\sigma^2 = \mathbb{E}\left((X_0 - p)^2\right)$. 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 $L$. The quantity $d_N$ is called the Wasserstein distance between $Y^N$ and $Z^N$.

Let $F_{N,p} : \mathbb{R} \to [0:1]$ and $F_p$ be respectively the CDF (Cumulative Distribution Function) of $\sqrt{N}(Y^N - p)$ and of the standard normal distribution: $F_{N,p}(x) = P(\sqrt{N}(Y^N - p) \le x)$ 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:

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

where $F^{-1}_{N,p}(U) \stackrel{\text{def}}{=} \min\{y : F_{N,p}(y) \geq U\}$.

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

**Lemma 16** *There exists a constant $\alpha_1$ 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,*

$$\mathbb{E}\left(|F^{-1}_{N,p}(U) - F_p^{-1}(U)|\right) \leq \alpha_1 \sqrt{\frac{\log(N)}{N}}, \tag{20}$$

*where $\alpha_1 < 356$.*

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

$$\mathbb{E}\left(|F_N^{-1}(U) - F^{-1}(U)|\right) = \int_0^1 |F_N^{-1}(u) - F^{-1}(u)|du \tag{21}$$

The Berry-Essen theorem (see for example part 2.4.d of [10]) shows that $\sup_{y\in\mathbb{R}} |F_N(\sigma y) - F(\sigma y)| \leq \frac{3\rho}{\sigma^3\sqrt{N}}$ 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:

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

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 $u > 1/2$, (21) is less than

$$2\left(\int_{\frac{1}{2}}^{k_N} \left(F^{-1}(u + \frac{3\rho}{\sigma^2\sqrt{N}}) - F^{-1}(u)\right) du + \int_{k_N}^1 |F_N^{-1}(u) - F^{-1}(u)|du\right) \tag{22}$$

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 (22) is bounded by

$$\begin{aligned}
\int_{\frac{1}{2}}^{k_N} \left(F^{-1}(u + \frac{3\rho}{\sigma^2\sqrt{N}}) - F^{-1}(u)\right) du &\leq \frac{3\rho}{\sigma^2\sqrt{N}} \int_{\frac{1}{2}}^{k_N} (F^{-1})'(u + \frac{3\rho}{\sigma^2\sqrt{N}})du \\
&= \frac{3\rho}{\sigma^2\sqrt{N}} \left(F^{-1}(k_N + \frac{3\rho}{\sigma^2\sqrt{N}}) - F^{-1}(\frac{1}{2})\right) \\
&\leq \frac{3\rho}{\sigma^2\sqrt{N}} F^{-1}(k_N + \frac{3\rho}{\sigma^2\sqrt{N}}).
\end{aligned} \tag{23}$$

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

$$\begin{aligned}
\int_{k_N}^1 F^{-1}(u)du &= \int_{F^{-1}(k_N)}^{\infty} xF'(x)dx \\
&= \left[x\left(F^{-1}(x) - 1\right)\right]_{F^{-1}(k_N)}^{\infty} - \int_{F^{-1}(k_N)}^{\infty} (F(x) - 1)\, dx \\
&= (1 - k_N)\, F^{-1}(k_N) + \int_{F^{-1}(k_N)}^{\infty} (1 - F(x))\, dx
\end{aligned} \tag{24}$$

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

$$\int_{F^{-1}(k_N)}^\infty 1-F(x)dx \leq \exp(-\frac{F^{-1}(k_N)}{2}) \leq 2\sqrt{2\pi}F^{-1}(k_N)(1-k_N) \tag{25}$$

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:

$$\int_{k_N}^1 F_N^{-1}(u)du \leq (1-k_N)F_N^{-1}(k_N) + \exp(-2F_N^{-1}(k_N)) \tag{26}$$

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 (22), (23), (24), (25) and (26), (21) is less than:

$$\begin{aligned}(21) &\leq (22)\\ &\leq 2\left[(23)+(24)+(26)\right]\\ &\leq 2\frac{3\rho}{\sigma^2\sqrt{N}}F^{-1}\left(k_N+\frac{3\rho}{\sigma^2\sqrt{N}}\right) + 2(1-k_N)(F_N^{-1}(k_N) + (1+2\sqrt{2\pi})F^{-1}(k_N))\\ &\quad +4\sqrt{2\pi}F^{-1}(k_N - \frac{3\rho}{\sigma^2\sqrt{N}})(1-k_N+\frac{3\rho}{\sigma^2\sqrt{N}})\end{aligned} \tag{27}$$

Let $k_N \stackrel{\text{def}}{=} 1 - 2\frac{3\rho}{\sigma^2\sqrt{N}}$. Since $\rho$ is the third moment of a Bernoulli variable of mean $p$ and $\sigma^2$ its variance, we have $\rho/\sigma^2 = p^2 + (1-p)^2 \in [.5;1]$. Moreover, the functions $F^{-1}$ and $F_N^{-1}$ are increasing. Using these facts, Equation (27) becomes

$$\begin{aligned}(21) &\leq F^{-1}\left(1-\frac{3\rho}{\sigma^2\sqrt{N}}\right)\left(2\frac{3\rho}{\sigma^2\sqrt{N}}\right)\left(2+4+8\sqrt{2\pi}+12\sqrt{2\pi}\right)\\ &\quad +4\frac{3\rho}{\sigma^2\sqrt{N}}F_N^{-1}(1-2\frac{3\rho}{\sigma^2\sqrt{N}}),\\ &\leq \frac{350}{\sqrt{N}}F^{-1}(1-\frac{3}{2\sqrt{N}}) + \frac{12}{\sqrt{N}}F_N^{-1}(1-\frac{3}{\sqrt{N}}),\end{aligned} \tag{28}$$

where we used the fact that $2\sqrt{2\pi} < 6$ and $6\sqrt{2\pi} < 16$.

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 $1 - 3/\sqrt{N}$ leads to:

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

Similar inequality for the tail of the normal distribution leads to $F^{-1}(y) \leq \sqrt{-2\log(1-y)}$ and $F^{-1}(1-3/(2\sqrt{N})) \leq \sqrt{\log(N)}$.

This shows that:

$$(21) \leq 356\frac{\sqrt{\log(N)}}{\sqrt{N}}.$$

This bound could be improved, in particular by a more precise analysis of (28), but fulfills for our needs.

In the case where we sum only a fraction $\delta_N$ of the $N$ Bernoulli variables, the result still holds.

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

$$\mathbb{E}\left(\left|\sqrt{\delta_N}F^{-1}_{N\delta_N,p}(U) - \sqrt{b}F_p^{-1}(U)\right|\right) \leq \alpha_2\sqrt{\frac{\log(N)}{N}}$$

*where $\alpha_2 = \alpha_1 + \max(\alpha', \sqrt{\alpha'}+2)$*

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

$$\begin{aligned}\mathbb{E}\left(\left|\sqrt{\delta_N}F^{-1}_{N\delta_N}(U) - \sqrt{b}F^{-1}(U)\right|\right) &\leq \mathbb{E}\left(\left|\sqrt{\delta_N}F^{-1}_{N\delta_N}(U) - \sqrt{\delta_N}F^{-1}(U)\right|\right)\\ &\quad + \mathbb{E}\left(\left|\left(\sqrt{\delta_N}-\sqrt{b}\right)F^{-1}(U)\right|\right)\\ &\leq \alpha_1\sqrt{\delta_N}\sqrt{\frac{\log(N\delta_N)}{N\delta_N}} + \mathbb{E}\left(\left|\sqrt{\delta_N}-\sqrt{b}\right|\right)\mathbb{E}(G)\end{aligned}$$

The first part of the inequality comes from the Lemma 16 and is less than $\alpha_1\frac{\log(N)}{N}$ since $\delta_N \leq 1$. The last part comes from the fact that $U$ and $\delta_N$ are independent. Moreover, the variance of $G$ is the variance of a Bernoulli variable, so $\mathbb{E}(|G|) \leq 1/4$.

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

$$\begin{aligned}\mathbb{E}\left(\left|\sqrt{\delta_N}-\sqrt{b}\right|\right) &\leq \mathbb{E}\left(\frac{\left|\delta_N - \sqrt{b}\right|}{\sqrt{\delta_N}+\sqrt{b}}\right) \leq \mathbb{E}\left(\frac{\left|\delta_N-\sqrt{b}\right|}{\sqrt{b}}\right)\\ &\leq \sqrt{N}\mathbb{E}\left(\left|\delta_N - \sqrt{b}\right|\right) \leq C'\sqrt{\frac{\log(N)}{N}}.\end{aligned}$$

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

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

This shows the inequality.

The following lemma uses the previous results in the case of multidimensional Bernoulli variables. A multidimensional Bernoulli variable $B$ of parameter $(p_1 \dots p_S)$ is a unit vector $\mathbf{e}_j$ on $\mathbb{R}^S$ (its $j$-th component equals 1 and all others equal 0). The probability that $B = \mathbf{e}_j$ is $p_j$.

**Lemma 18** *Let $m \in \mathscr{P}_N(\mathcal{S})$ and let $(B_k^i)_{1\leq i\leq S, k\geq 1}$ be independent multidimensional Bernoulli random vectors on $\{0;1\}^S$ such that $\mathbb{P}\left(B_k^i = \mathbf{e}_j\right) = p_{ij}$. Let $f$ be a Lipschitz function on $\mathbb{R}^S$ with constant $L$. Let us define $Y^N \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^S\sum_{k=1}^{m_iN} B_k^i$. Then there exists a constant $\alpha_3$ such that:*

$$\left|\mathbb{E}\left(f(Y^N)\right) - \mathbb{E}\left(f(mP + \frac{1}{\sqrt{N}}G)\right)\right| \leq \alpha_3 L\frac{\sqrt{\log(N)}}{N},$$

*where $mP$ is the vector $(mP)_j = \sum_{i=1}^S m_i p_{ij}$ and $G$ is a random Gaussian vector of mean 0 and covariance matrix $D$ defined by Equation (15) ($D_{jj} = \sum_i m_i p_{ij}(1-p_{ij})$ and $D_{jk} = -\sum_i m_i p_{ij} p_{ik}$ for $j \neq k$).*

*Proof* Let $(b_k^{ij})_{i,j\leq S, k\geq 1}$ be a collection of independent Bernoulli random variables of parameters $p_{ij}/(1-\sum_{\ell<j}p_{i\ell})$ (the parameter of $b_k^{i1}$ is $p_{i1}$). The Bernoulli vector $B_k^i$ is equal in law to $\bar{B}_k^i$ (denoted $B_k^i \stackrel{\mathcal{L}}{=} \bar{B}_k^i$) where

$$\bar{B}_k^i = b_k^{i1}\mathbf{e}_1 + (1-b_k^{i1})\left(b_k^{i2}\mathbf{e}_2 + (1-b_k^{i2})\left(b_k^{i3}\mathbf{e}_3 + (1-b_k^{i3})(\dots)\right)\right).$$

Indeed, $\bar{B}_k^i$ is a vector with only one component equal to 1 and

$$\mathbb{P}\left(\bar{B}_k^i = \mathbf{e}_j\right) = (1-p_{i1})(1-\frac{p_{i2}}{1-p_{i1}})(1-\frac{p_{i3}}{1-p_{i1}-p_{i2}})\dots\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}} = p_{ij}.$$

Let $T_{ij}^N \stackrel{\text{def}}{=} N^{-1}\sum_{k=1}^{m_iN}\mathbf{1}_{\{B_k^i=\mathbf{e}_j\}}$ be the proportion of objects going from state $i$ to state $j$. By definition of $b_k^{ij}$, $T_{ij}^N$ can be written

$$\begin{aligned}
T_{ij}^N &\stackrel{\mathcal{L}}{=} \frac{1}{N}\sum_{k=1}^{m_iN}(1-b_k^{i1})(1-b_k^{i2})\dots(1-b_k^{i,j-1})b_k^{ij}\\
&\stackrel{\mathcal{L}}{=} \frac{1}{N}\sum_{1\le k\le m_iN \text{ s.t. } (b_k^{i1}=0)\wedge\dots\wedge(b_k^{i,j-1}=0)} b_k^{ij}\\
&\stackrel{\mathcal{L}}{=} \frac{1}{N}\sum_{k=1}^{m_iN-\sum_{\ell<j}T_{i\ell}} b_k^{i\ell}\\
&\stackrel{\mathcal{L}}{=} \frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}\left(m_i-\sum_{\ell<j}T_{i\ell}^N\right) + \sqrt{\frac{m_i-\sum_{\ell<j}T_{i\ell}^N}{N}}F^{-1}_{m_iN-N\sum_{\ell<j}T_{i\ell}^N,\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}}(U_{ij}).
\end{aligned}$$

where the function $F$ is defined by $F_{A,q}(x) \stackrel{\text{def}}{=} \mathbb{P}\left(\frac{1}{\sqrt{A}}\left(\sum_{i=1}^A c_i - Aq\right)\le x\right)$ where $q\in[0;1]$, $A\in\mathbb{N}$ and $c_i$ are scalar *i.i.d.* Bernoulli variables – *i.e.* $\mathbb{P}(c_i=1) = 1-\mathbb{P}(c_i=0) = p$. By construction, $(\sum_{i=1}^S T_{ij}^N)_{1\le j\le S}$ has the same law as $Y^N$.

The variable $H$ is constructed similarly. Denoting $F_p(x) \stackrel{\text{def}}{=} \mathbb{P}(G\ge x)$ where $G$ is a normal variable of mean 0 and variance $p(1-p)$, we define the variables $H_{ij}^N$ by:

$$\begin{aligned}
H_{i1} &\stackrel{\text{def}}{=} F^{-1}_{p_{i1}}(U_{i1})\\
H_{ij} &\stackrel{\text{def}}{=} -\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}\sum_{\ell<j}H_{i\ell} + \sqrt{\sum_{\ell<j}p_{i\ell}}F^{-1}_{\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}}(U_{ij})
\end{aligned}$$

It is direct to prove that $H$ has the same law as $G$ by showing that $H$ is a Gaussian vector and by computing the covariance matrix of $H$.

Using this representation of $Y^N$ and $G$ by $T^N$ and $H$, we have:

$$\begin{aligned}
\left|\mathbb{E}\left(f(Y^N)\right) - \mathbb{E}\left(mP+\frac{1}{\sqrt{N}}G\right)\right| &\le \mathbb{E}\left(\left\|f(T^N) - f(mP+\frac{1}{\sqrt{N}}H)\right\|\right)\\
&\le L\mathbb{E}\left(\left\|T^N - mP - \frac{1}{\sqrt{N}}H\right\|\right)\\
&\le L\sum_{i=1}^S\sum_{j=1}^S\mathbb{E}\left(\left|T_{ij}^N - m_ip_{ij} - H_{ij}\right|\right)
\end{aligned}$$

The next step of the proof is to bound $\mathbb{E}\left|T_{ij}^N - m_ip_{ij} - H_{ij}\right|$ which is less than:

$$\begin{aligned}
\mathbb{E}\left|\sqrt{\frac{m_i-\sum_{\ell<j}T_{i\ell}^N}{N}}F^{-1}_{m_iN-N\sum_{\ell<j}T_{i\ell}^N,\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}}(U_{ij}) - \sqrt{\sum_{\ell<j}p_{i\ell}}F^{-1}_{\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}}(U_{ij})\right|&\\
+\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}\mathbb{E}\left|\sum_{\ell<j}(m_ip_{i\ell} - T_{i\ell}^N + H_{i\ell})\right|&
\end{aligned}$$

where we used the fact that $\frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}m_i - m_ip_{ij} = \frac{p_{ij}}{1-\sum_{\ell<j}p_{i\ell}}\sum_{\ell<j}m_ip_{i\ell}$.

By induction on $j$ and using Lemma 17, this quantity is less than $\alpha^{(j)}\frac{\sqrt{N}}{N}$ where the constant $\alpha^{(j)} \stackrel{\text{def}}{=} \alpha^{(j-1)} + \alpha_1 + \max(\alpha^{(j-1)}, \sqrt{\alpha^{(j-1)}}+2)$. The constant $\alpha_3$ is equal to $S^2\alpha^{(S)}$.

We are now ready for the proof of Theorem 9.

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

$$\begin{aligned} W^N_{T\dots T,a}\left(M^N(t),C^N(t)\right) &= 0 \\ W^N_{t\dots T,a}\left(M^N(t),C^N(t)\right) &= r\left(M^N(t),C^N(t)\right) + \mathbb{E}\left(W^N_{t\dots T,a}\left(\widetilde{M}^N_a(t+1),\widetilde{C}^N_a(t+1)\right)\right) \end{aligned} \tag{29}$$

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 $V^N_{t\dots T,a}\left(M^N(t),C^N(t)\right)$ the expected reward between $t$ and $T$ for the original system. We want to prove by induction that there exist constants $\gamma_t$ such that for any $t$:

$$\left|V^N_{t\dots T}(M^N,C^N) - W^N_{t\dots T}(M^N,C^N)\right| \leq \gamma_t \frac{\sqrt{\log(N)}}{N}. \tag{30}$$

The constant $\gamma_t$ may depend on the parameters of the system (such as the Lipschitz constants of the functions $g,K,r$) but not on the value of $(M^N,C^N)$.

Equation (30) is clearly true for $t=T$ by taking $\gamma_T=0$. Let us now assume that (30) holds for some $t+1\leq T$. By a backward induction on $t$, one can show that $W^N_{t\dots T,a}(.,.)$ is Lipschitz for some constant $L_{W_t}$. $\left|V^N_{t\dots T}(M^N,C^N) - W^N_{t\dots T}(M^N,C^N)\right|$ can be written:

$$\begin{aligned} &\left|\mathbb{E}\left(V^N_{t+1\dots T,a}\left(M^N_a(t+1),C^N_a(t+1)\right)\right) - \mathbb{E}\left(W^N_{t+1\dots T,a}\left(\widetilde{M}^N_a(t+1),\widetilde{C}^N_a(t+1)\right)\right)\right| \\ &\quad\leq \left|\mathbb{E}\left(W^N_{t+1\dots T,a}\left(M^N_a(t+1),C^N_a(t+1)\right) - W^N_{t+1\dots T,a}\left(\widetilde{M}^N_a(t+1),\widetilde{C}^N_a(t+1)\right)\right)\right| \\ &\qquad+\left|\mathbb{E}\left(W^N_{t+1\dots T,a}\left(M^N_a(t+1),C^N_a(t+1)\right) - V^N_{t+1\dots T,a}\left(M^N_a(t+1),C^N_a(t+1)\right)\right)\right| \\ &\quad\leq \alpha_3 L_{W_t} L_g \frac{\sqrt{\log(N)}}{N} + \gamma_t \frac{\sqrt{\log(N)}}{N}. \end{aligned} \tag{31}$$

The fist part of the second inequality comes from Lemma 18 applied to the function $m\mapsto W^N_{t+1\dots T,a}\left(m,g(a,C^N(t),m)\right)$ 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 (29) for $W^N_*$ and $V^N_*$ which can be computed by a backward induction:

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

We then get equations similar to (31) but with a $\sup_a$ before both expectation. The sup can be removed using the fact that for any functions $f$ and $g$: $|\sup_a f(a) - \sup_a g(a)| \leq \sup_a |f(a)-g(a)|$.