A Mean Field Approach for Optimization in Particles Systems and Applications

Nicolas Gast, Bruno Gaujal

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Mean Field Approach for Optimization in Particles Systems and Applications

Nicolas Gast — Bruno Gaujal

N° 6877 — version 2

initial version March 2009 — revised version June 2009

Thème NUM

ISSN 0249-6399 ISRN INRIA/RR--6877--FR+ENG

A Mean Field Approach for Optimization in Particles Systems and Applications

Nicolas Gast , Bruno Gaujal

Thème NUM — Systèmes numériques
Équipe-Projet MESCAL

Rapport de recherche n° 6877 — version 2 — initial version March 2009 — revised version June 2009 — 22 pages

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

In the finite horizon case or with a discounted cost and an infinite horizon, we show that when the number of particles becomes large, the optimal cost of the system converges almost surely to the optimal cost of a deterministic system (the "optimal mean field"). Convergence also holds for optimal policies.

We further provide insights on the speed of convergence by proving several central limits theorems for the cost and the state of the Markov decision process with explicit formulas for the variance of the limit Gaussian laws.

Then, our framework is applied to a brokering problem in grid computing. The optimal policy for the limit deterministic system is computed explicitly. 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 (such as Join the Shortest Queue) by measuring its asymptotic gain.

Key-words: Markov Decision Processes, Mean Field, Optimization, Particles System, Grid Broker

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Une approche champ moyen pour l'optimisation dans les systèmes de particules et ses applications

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

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

De plus, nous donnons un aperçu de la vitesse de convergence en prouvant plusieurs théorèmes de la limite centrale pour le coût ainsi que l'état moyen du processus en donnant des formules explicites pour la variance des lois gaussiennes limites.

Enfin, ce modèle est appliqué à un problème de gestionnaire de ressources dans des grilles de calcul. Nous donnons un algorithme explicite pour calculer la politique optimale de la limite puis plusieurs simulations avec un nombre variable de processeurs sont étudiées. Nous comparons les performances de la politique optimale de la limite appliquée au système initiale avec plusieurs politiques classiques, (telles que joindre la file la plus courte). Nous mesurons le gain asymptotique, ainsi que le seuil à partir duquel elle surpasse les politiques classiques.

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

1 Introduction

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

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

Several papers ([3], [6]) study the limit behavior of Markovian systems in the case of vanishing intensity (the expected number of transitions per time slot is $o(N)$). In these cases, the system converges to a differential system in continuous time. In the case considered here, time remains discrete at the limit. This requires a rather different approach to construct the limit.

In [8], discrete time systems are considered and the authors show that under certain conditions, as N grows large, a Markovian system made of N particles converges to a deterministic system. Since a Markov decision process can be seen as a family of Markovian kernels, the class of systems studied in [8] corresponds to the case where this family is reduced to a unique kernel and no decision can be made. Here, we show that under similar conditions as in [8], a Markov decision process also converges to a deterministic one. More precisely, we show that the optimal costs (as well as the corresponding states) converge almost surely to the optimal costs (resp. the corresponding states) of a deterministic system (the "optimal mean field").

On a practical point of view, this allows one to compute the optimal policy in a deterministic system which can often be done very efficiently, and then to use this policy in the original random system as a good approximation of the optimal policy, which cannot be computed efficiently because of the curse of dimensionality. This is illustrated by an application of our framework to optimal brokering in computational grids. We consider a set of multi-processor clusters (forming a computational grid, like EGEE [1]) and a set of users submitting tasks to be executed. A central broker assigns the tasks to the clusters (where tasks are buffered and served in a fifo order) and tries to minimize the average processing time of all tasks. Computing the optimal policy (solving the associated MDP) is known to be hard [13]. Numerical computations can only be carried up to a total of 10 processors and two users. However, our approach shows that when the number of processors per cluster and the number of users submitting tasks grow, the system converges to a mean field deterministic system. For this deterministic mean field system, the optimal brokering policy can be explicitly computed. Simulations reported in Section 4 show that, using this policy over a grid with a growing number of processors, makes performance converge to the optimal sojourn time in a deterministic system, as expected. Also, simulations show that this deterministic static policy outperforms classical dynamic policies such as Join the Shortest Queue, as soon as the total number of processors and users is over 50.

In general, how good the deterministic approximation is and how fast convergence takes place can also be estimated. For that, we provide bounds on the speed of convergence by proving of central limit theorem for the state of the system under the optimal policy as well as for the cost function.

2 Notations and definitions

The system is composed of N particles. There are S possible states for each particle, the state space is denoted by $\mathcal{S}=\{1,\dots,S\}$. The state of the nth particle at time t is denoted $X_n^N(t)$. We assume that the particles are distinguishable only through their state and that the dynamics of the system is homogeneous in N. In other words, this means that the behavior of the system only depends on in the proportion of particles in every state i. For all $i \in \mathcal{S}$, $\left(\mathrm{M}_t^N\right)_i \stackrel{\text{def}}{=} \sum_{n=1}^N \mathbf{1}_{X_n^N(t)=i}$ is the proportion of particles in state i and we denote by M_t^N the vector $(\left(\mathrm{M}_t^N\right)_1 \dots \left(\mathrm{M}_t^N\right)_S)$. The set of possible values for M^N is the set of probability measures p on $\{1\dots S\}$, such that $Np(i) \in \mathbb{N}$ for all $i \in \mathcal{S}$, denoted by $\mathcal{P}_N(S)$. For each N, $\mathcal{P}_N(S)$ is a finite set. When N goes to infinity, it converges to $\mathcal{P}(\mathcal{S})$ the set of probability measures on $\mathcal{S}$.

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

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

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

2.1 Actions and policies

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

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

The evolutions of particles are supposed to be independent once C is given. Moreover, we assume that $K_{i,j}(a, \mathrm{C})$ is continuous in a and C. The assumption of independence of the users is a rather common assumption in mean field models [8]. However other papers [3, 6] have shown that similar results can be obtained using asymptotic independence only (see [10] for results of this type).

Here, the focus is on Markov Decision Processes theory and on the computation of optimal policies. A policy $\Pi = (\Pi_1 \dots \Pi_t \dots)$ specifies the decision rules to be used at each time slot. A decision rule Π_t is a procedure that provides an action at time t. In general, Π_t is a random measurable function that depends on the events $((\mathrm{M}_1, \mathrm{C}_1) \dots (\mathrm{M}_t, \mathrm{C}_t))$ 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, therefore we will only focus on them [14].

2.2 Reward functions

To each possible state (M, C) of the system at time t, we associate a reward $\mathrm{r}_t(\mathrm{M}, \mathrm{C})$. 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 focus on two problems: finite-horizon reward and discounted reward.

In the finite-horizon case, we want to maximize the sum of the rewards over all time $t < T$ plus a final reward that depends on the final state, $\mathrm{r}_T(\mathrm{M}_T^N, \mathrm{C}_T^N)$. The expected reward of the policies $\Pi_0, \ldots, \Pi_{T-1}$ is:

$$V^N_{\Pi_0 \ldots \Pi_T}(\mathrm{M}_0^N, \mathrm{C}_0^N) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T-1} \mathrm{r}_t(\mathrm{M}_t^N, \mathrm{C}_t^N) + \mathrm{r}_T(\mathrm{M}_T^N, \mathrm{C}_T^N)\right],$$

where the expectation is taken over all possible $(\mathrm{M}_t^N, \mathrm{C}_t^N)$ when the actions are $\Pi_t(\mathrm{M}_t^N, \mathrm{C}_t^N)$, for all t.

Let $0 \leq \delta < 1$, the discounted reward associated to δ and the policy $\Pi_0 \ldots \Pi_t \ldots$ is the quantity:

$$V^N_{(\delta),\Pi_0 \ldots}(\mathrm{M}_0^N, \mathrm{C}_0^N) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{\infty} \delta^t \mathrm{r}_t(\mathrm{M}_t^N, \mathrm{C}_t^N)\right].$$

Again, the expectation is taken over all possible $(\mathrm{M}_t^N, \mathrm{C}_t^N)$ when the actions at time t is $\Pi_t(\mathrm{M}_t^N, \mathrm{C}_t^N)$, for all t.

In both cases, the goal of the controller is to find a policy that maximizes the expected reward:

$$V^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) \stackrel{\text{def}}{=} \sup_{\Pi_1 \ldots \Pi_T} V^N_{\Pi_1 \ldots \Pi_T}(\mathrm{M}_0^N, \mathrm{C}_0^N),$$

$$V^{*N}_{(\delta)}(\mathrm{M}_0^N, \mathrm{C}_0^N) \stackrel{\text{def}}{=} \sup_{\Pi_1 \ldots} V^N_{(\delta),\Pi_1 \ldots}(\mathrm{M}_0^N, \mathrm{C}_0^N).$$

2.3 Summary of the assumptions

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

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

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

(A3) **Continuity of** K, g, r – the mappings $(\mathrm{C}, a) \mapsto K(a, \mathrm{C})$, $(\mathrm{C}, \mathrm{M}, a) \mapsto g(\mathrm{C}, \mathrm{M}, a)$ and $(\mathrm{M}, \mathrm{C}) \mapsto \mathrm{r}_t(\mathrm{M}, \mathrm{C})$ are continuous deterministic functions, uniformly continuous in a.

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

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

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

Here are some comments on the uniform bound B on the initial condition (A4). In fact, as C_0^N converges almost surely, C_0^N is almost surely bounded. Here we had a bound B which is uniform on all events in order to be sure that the variable C_0^N is dominated by an integrable

function. As g is continuous and the sets $\mathcal{A}$ and $\mathcal{P}(\mathcal{S})$ are compact, this shows that for all t, there exists $B_t < \infty$ such that

$$\|\mathrm{C}_t^N\|_\infty \leq B_t. \tag{1}$$

Finally, in many cases the rewards also depend on the action. This is not the case here, at a small loss of generality.

3 Convergence results and optimal policy

In the case where there is no control, one can adapt the results proved in [8] to show that when N goes to infinity, the system converges almost surely to a deterministic one. In our case, this means that if the actions are fixed, the system converges.

For any fixed action a and any value $\mathrm{M} \in \mathcal{P}_N(\mathcal{S})$, we define the random variable $\Phi_a^N(\mathrm{M}, \mathrm{C})$ that corresponds to the state of the system M', C' after one iteration started from M, C. For $m \in \mathbb{P}(\mathcal{S})$, we define $\Phi_a(m, c)$ the (deterministic) value corresponding to one iteration of the mean field system: $\Phi_a(m_t, c_t) = (m_{t+1}, c_{t+1})$ where

$$\begin{aligned} m_{t+1} &= m_t.K(a, c_t) \\ c_{t+1} &= g(m_{t+1}, c_t). \end{aligned}$$

We call $\Phi^N_{a_0 \dots a_{T-1}}$ (resp. $\Phi_{a_0 \dots a_{T-1}}$) the compositions of $\Phi^N_{a_0}, \dots, \Phi^N_{a_{T-1}}$ (resp. of $\Phi_{a_0} \dots \Phi_{a_{T-1}}$).

In [8], the system is homogeneous in time. However, the proofs are done for each step time and the results still hold without time homogeneity. With our notations, theorem 4.1 of [8] says that if the actions are $a_0 \dots a_{T-1}$, and if the initial state converges almost surely, then the system of size N converges almost surely.

Theorem 1 (Mean Field Limit, th. 4.1 of [8])**.** *Under assumptions (A1,A3,A4), if the controller takes the actions a_t at time t, then for any fixed T:*

$$(\mathrm{M}_t^N, \mathrm{C}_t^N) \xrightarrow{a.s} \Phi_{a_0 \dots a_{T-1}}(m_0, c_0).$$

In the following, we will first show that if we fix the actions, the total reward of the system converges when N grows, then we will show that the optimal reward also converges.

3.1 Finite horizon model

In this section, the horizon T is fixed, the infinite horizon case will be treated in Section 3.3. Using the same notation and hypothesis as in Theorem 1, we define the reward of the deterministic system starting at m_0, c_0 under the actions $a_0, \dots, a_{t-1}$:

$$v_{a_0 \dots a_{t-1}}(m_0, c_0) = \sum_{t=1}^{T} \mathrm{r}_t(\Phi_{a_0 \dots a_{t-1}}(m_0, c_0)).$$

For any t, if the action taken at instant t is fixed equal to a_t, then $(\mathrm{M}_t^N, \mathrm{C}_t^N)$ converges almost surely to (m_t, c_t). Since the reward at time t is continuous, this means that the finite-horizon expected reward converges as N grows large:

Lemma 2 (Convergence of the reward)**.** *Under assumptions (A1,A3,A4), if the controller takes actions $a_0 \dots a_{T-1}$, the finite-horizon expected reward of the stochastic system converges to the finite-horizon reward of the deterministic system:*

$$\lim_{N \to \infty} V^N_{a_0 \dots a_{t-1}}(\mathrm{M}_0^N, \mathrm{C}_0^N) = v_{a_0, \dots, a_{t-1}}(m_0, c_0) \quad \text{a.s.}$$

Proof. For all t, $(\mathrm{M}_t^N, \mathrm{C}_t^N)$ converges almost surely to (m_t, c_t). Since the reward at time t is continuous in (M, C), then $\mathrm{r}_t(\mathrm{M}_t^N, \mathrm{C}_t^N) \xrightarrow{a.s} \mathrm{r}_t(m_t, c_t)$. Moreover, as (M, C) are bounded (see Equation (1)), the dominated convergence theorem shows that $\mathbb{E}[\mathrm{r}_t(\mathrm{M}_t^N, \mathrm{C}_t^N)]$ goes to $\mathrm{r}_t(m_t, c_t)$ which concludes the demonstration. □

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 $v_T^*(m_0, c_0) \stackrel{\text{def}}{=} \sup_{a_0 \dots a_{T-1}} \{v_{a_0 \dots a_{T-1}}(m_0, c_0)\}$. Since the action set is compact, this supremum is a maximum: there exist $a_0^* \dots a_{T-1}^*$ such that $v_T^*(m_0, c_0) = v_{a_0^* \dots a_{T-1}^*}(m_0, c_0)$. In fact, in many cases there are more than one optimal action sequence. In the following, $a_0^* \dots a_{T-1}^*$ is one of them, and will be called the sequence of *optimal limit actions*.

Theorem 3 (Convergence of the optimal reward)**.** *Under assumptions (A1,A2,A3,A4), as N goes to infinity, the optimal reward of the stochastic system converges to the optimal reward of the deterministic limit system: almost surely,*

$$\lim_{N \to \infty} V_T^{*N}(M_0^N, C_0^N) = \lim_{N \to \infty} V_{a_0^* \dots a_{T-1}^*}^N(M_0^N, C_0^N) = v_T^*(m_0, c_0)$$

In words, this theorem says that, at the limit, the reward of the optimal policy under full information $V^*{}_T^N(M_0^N, C_0^N)$ is the same as the reward obtained when the optimal limit actions $(a_0^* \dots a_{T-1}^*)$ are used in the original system, both being equal to the optimal reward of the limit deterministic system, $v_T^*(m_0, c_0)$.

Proof. For all N and $0 \leq t \leq T$ and $(\mathrm{M}, \mathrm{C}) \in \mathbb{P}_N(\mathcal{S}) \times \mathbb{R}^d$, let us define by induction on t the function $V_{t \dots T}^{*N}$:

$$\begin{aligned} V_{T \dots T}^{*N}(\mathrm{M}, \mathrm{C}) &= \mathrm{r}_T(\mathrm{M}, \mathrm{C}) \\ V_{t \dots T}^{*N}(\mathrm{M}, \mathrm{C}) &= \mathrm{r}_t(\mathrm{M}, \mathrm{C}) + \sup_{a \in \mathcal{A}} \mathbb{E}_{\mathrm{M},\mathrm{C}}[V_{t+1 \dots T}^{*N}(\Phi_a^N(\mathrm{M}, \mathrm{C}))]. \end{aligned} \tag{2}$$

where the expectation $\mathbb{E}_{\mathrm{M},\mathrm{C}}[\cdot]$ is taken over all possible values of $\Phi_a^N(\mathrm{M}, \mathrm{C})$ given (M, C). Also notice that $V_{t \dots T}^{*N}(\mathrm{M}, \mathrm{C})$ is the maximal expected reward between time t and time T starting in (M, C) and therefore $V_{0 \dots T}^{*N} = V_T^{*N}$.

Let us also define for the limit system, $v_{t \dots T}^*$ similarly (by removing the expectation):

$$\begin{aligned} v_{T \dots T}^*(m, c) &= \mathrm{r}_T(m, c) \\ v_{t \dots T}^*(m, c) &= \mathrm{r}_t(m, c) + \sup_{a \in \mathcal{A}} \Big[v_{t+1 \dots T}^*\big(\Phi_a(m, c)\big) \Big], \end{aligned} \tag{3}$$

and let $\Pi_t^*(m, c)$ be an action that maximize the sup in the previous equation (it exists because of (A2): $\mathcal{A}$ is compact).

We will show by induction on $t < T$ that $V_{t \dots T}^{*N}(\cdot, \cdot)$ is continuous (note that since $\mathrm{M} \in \mathcal{P}^N(\mathcal{S})$ is discrete the continuity in M is trivial) and that we can define an optimal policy $\Pi_t^{*N}(\mathrm{M}, \mathrm{C})$, such that:

$$V_{t \dots T}^{*N}(\mathrm{M}, \mathrm{C}) = \mathrm{r}_t(\mathrm{M}, \mathrm{C}) + \mathbb{E}\big[V_{t+1 \dots T}^{*N}(\Phi^N_{\Pi_t^{*N}(\mathrm{M},\mathrm{C})}(\mathrm{M}, \mathrm{C}))\big]. \tag{4}$$

For $t = T$, the assumption holds by the continuity of r (A3).

Let us assume that it holds for $t + 1 \leq T$. By assumption (A3), the mapping g and the kernel K are continuous in a thus if $\{a(k)\}_{k \in \mathbb{N}}$ is a sequence of action converging to a, $\Phi_{a(k)}^N$ converges (in law) to Φ_a^N. As $V_{t+1 \dots T}^{*N}$ is continuous, $a \mapsto \mathbb{E}[V_{t+1 \dots T}^{*N}(\Phi_a^N(\mathrm{M}, \mathrm{C}))]$ is continuous. Using this continuity and the compacity of $\mathcal{A}$, the optimal action $\Pi_t^{*N}(\mathrm{M}, \mathrm{C}) \in \mathcal{A}$ exists. The

functions r, g, K are uniformly continuous in a, therefore the convergence of the continuity of the function $a \mapsto \sup_a \mathbb{E}[V^{*N}_{t+1\dots T}(\Phi^N_a(\mathrm{M},\mathrm{C}))]$ is uniform in M, R. This shows that $(M,R) \mapsto \sup_a \mathbb{E}[V^{*N}_{t+1\dots T}(\Phi^N_a(\mathrm{M},\mathrm{C}))]$ is continuous and the property for all t is proved.

Let us now prove by induction on t that for all sequences $(\mathrm{M}^N,\mathrm{C}^N)$ converging almost surely to (m,c), $v^{*N}_{t\dots T}(\mathrm{M}^N,\mathrm{C}^N) \xrightarrow{a.s} v^*_{t\dots T}(m,c)$. This is clearly true for $t{=}T$. Assume that it holds for some $t{+}1{\leq}T$ and let us call $a^*_t \dots a^*_{T-1}$ a sequence of optimal actions for the deterministic limit. Lemma 2 shows that $V^N_{a^*_t\dots a^*_{T-1}}(\mathrm{M}^N,\mathrm{C}^N) \xrightarrow{a.s} v_{a^*_t\dots a^*_{T-1}}(m,c) = v^*_{t\dots T}(m,c)$. In particular, this shows the second inequality (which holds a.s.) of the following equation:

$$\begin{aligned}\liminf V^{*N}_{t\dots T}(\mathrm{M}^N,\mathrm{C}^N) &\geq \liminf V^N_{a^*_t\dots a^*_{T-1}}(\mathrm{M}^N,\mathrm{C}^N)\\ &= v^*_{t\dots T}(m,c).\end{aligned} \tag{5}$$

Let a^{*N} be a sequence of actions maximizing the expectation in (2). As $\mathcal{A}$ is compact, there exists a subsequence $a^{*\psi(N)}$ converging to a value a. Again by lemma 2, the lim sup of $\mathrm{r}(\mathrm{M}^{\psi(N)},\mathrm{C}^{\psi(N)}){+}\mathbb{E}[V^{*\psi(N)}_{t+1\dots T}(\Phi^{\psi(N)}_a(\mathrm{M}^{\psi(N)},\mathrm{C}^{\psi(N)}))]$ converges a.s. to $\mathrm{r}(m,c){+}v^*_{t+1}(\Phi_a(m,c)) \leq v^*_{t\dots T}(m,c)$. Using both inequalities, this shows that $V^{*\psi(N)}_{t\dots T}(\mathrm{M}^{\psi(N)},\mathrm{C}^{\psi(N)}) \xrightarrow{a.s} v^*(m,c)$.

To conclude the proof, remark that since the limit system is deterministic and takes the values $(m_0,c_0),\dots,(m_t,c_t)$, fixing the policy at time t to the action $a^*_t \stackrel{\text{def}}{=} \Pi^*(m_t,c_t)$ achieves the optimal reward. □

This result has several practical consequences. Recall that the limit actions $a^*_0 \dots a^*_{T-1}$ is a sequence of optimal actions in the limit case, *i.e.* such that $v_{a^*_0\dots a^*_{t-1}}(m,c) = v^*_T(m,c)$. This result proves that in the limit case, the optimal policy does not depend on the state of the system. This also shows that incomplete information policies are as good as complete information policies. However, the state $(\mathrm{M}^N_t,\mathrm{C}^N_t)$ is not deterministic and on one trajectory of the system, it could be quite far from its deterministic limit (m_t,c_t). In the proof of proposition 2, we also defined the policy $\Pi^*_t(\mathrm{M}^N_t,\mathrm{C}^N_t)$ which is optimal for the deterministic system starting at time t in state m_t, r_t. The least we can say is that this strategy is also asymptotically optimal, that is:

$$\lim_{N\to\infty} V^N_{\Pi^*_0\dots\Pi^*_T}(\mathrm{M},\mathrm{C}) = \lim_{N\to\infty} V^N_{a^*_0\dots a^*_T}(\mathrm{M},\mathrm{C}).$$

In practical situations, using this policy 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 $a^*_0 \dots a^*_{T-1}$. An other one is that the system becomes very sensitive to random perturbations: the policy Π^* is not necessarily continuous and may not have a limit. In Section 4, a comparison between the performances of $a^*_0 \dots a^*_{T-1}$ and $\Pi^*_0 \dots \Pi^*_{T-1}$ is provided over an example.

3.2 Central Limit Theorems

In this part we prove central limit theorems for interacting particles. This result provides estimates on the speed of convergence to the mean field limit. 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 particles. 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 $1/\sqrt{N}$. These properties are formalized in theorems 5 and 4 respectively.

To prove these results, we will need additional assumptions (A4-bis) and (A5) or (A5-bis).

(A4-bis) **Initial Gaussian variable** – There exists a Gaussian vector G_0 of mean 0 with covariance Γ_0 such that the vector $\sqrt{N}((\mathrm{M}_0^N, \mathrm{C}_0^N)-(m_0, c_0))$ (with $S+d$ components) converges in law to G_0. (This is denoted as $\sqrt{N}((\mathrm{M}_0^N, \mathrm{C}_0^N) - (m_0, c_0)) \xrightarrow{\mathcal{L}} G_0$). This assumption also includes (A4), *i.e.* almost sure convergence of the initial state.

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

(A5-bis) **Differentiability in** $a_0 \dots a_{T-1}$ – Let (m_t, c_t) be the deterministic limit of the system if the controller takes the actions $a_0 \dots a_{T-1}$ then for all $i, j \in \mathcal{S}$, the functions g, K_{ij} and r_t are differentiable in the points (m_t, c_t).

These assumptions are slightly stronger than (A3) and (A4) but remain very natural. (A4-bis) is clearly necessary for Theorems 5 and 4 to hold. The differentiability condition implies that if the gap between M_t and m_t is of order $1/\sqrt{N}$, it remains of the same order at time $t+1$. For Theorem 5, (A5-bis) is necessary but can be replaced by a Lipschitz continuity condition for Theorem 4. This will be further discussed in Section 4.2.

Theorem 4 (Central limit theorem for costs)**.** *Under assumptions (A1,A2,A3,A4bis,A5), (i)- there exists constants $\beta$and γ such that for all x:*

$$\begin{aligned}\limsup_{N\to\infty} \mathbb{P}(\sqrt{N}\Big|V_T^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) - v_T^*(m_0, c_0)\Big| \geq x) \\ \leq \mathbb{P}(\beta\|G_0\|_\infty + \gamma \geq x);\end{aligned} \tag{6}$$

(ii)- there exist constants $\beta', \gamma' > 0$ such that for all x:

$$\begin{aligned}\limsup_{N\to\infty} \mathbb{P}(\sqrt{N}\Big|V_T^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) - V^N_{a_0^*\dots a_{T-1}^*}(\mathrm{M}_0^N, \mathrm{C}_0^N)\Big| \\ \geq x) \leq \mathbb{P}(\beta'\|G_0\|_\infty + \gamma' \geq x);\end{aligned} \tag{7}$$

where $\|G'\|_\infty = \sup_i |G'_i|$.

This theorem is the main result of this section. The previous result (Theorem 3) says that $\limsup_{N\to\infty} V_T^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) = \limsup_{N\to\infty} V^N_{a_0^*\dots a_{T-1}^*}(\mathrm{M}_0^N, \mathrm{C}_0^N) = v^*_{t\dots T}(m_0, c_0)$. This new theorem says that both the gap between the cost under the optimal policy and of the cost when using the limit actions *(i)* or the gap between the latter cost and the optimal cost of the limit system *(ii)* are random variables that decrease to 0 with speed $\sqrt{N}$ and have Gaussian laws. Actually, a stronger result (using almost sure convergence instead of convergence in law) will be shown in Corollary 8. A direct consequence of this result is that there exists a constant γ'' such that:

$$\mathbb{E}\Big[\sqrt{N}|V_T^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) - v_T^*(m_0, c_0)|\Big] \to \gamma'' \tag{8}$$

The rest of this section is devoted to the proof of this theorem. A first step in the proof of Theorem 4 is a central limit theorem for the states, which has an interest by its own.

Theorem 5 (Mean field central limit theorem)**.** *Under assumption (A1,A2,A3,A4bis,A5-bis), 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:*

$$\begin{aligned}\sqrt{N}((\mathrm{M}_0^N, \mathrm{C}_0^N) - (m_0, c_0), \dots, (\mathrm{M}_t^N, \mathrm{C}_t^N) - (m_t, c_t)) \\ \xrightarrow{\mathcal{L}} G_0, \dots, G_t.\end{aligned} \tag{9}$$

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

$$\Gamma_{t+1} = \left[\begin{array}{c|c} P_t & F_t \\ \hline Q_t & H_t \end{array}\right]^{tr} \Gamma_t \left[\begin{array}{c|c} P_t & F_t \\ \hline Q_t & H_t \end{array}\right] + \left[\begin{array}{c|c} D_t & 0 \\ \hline 0 & 0 \end{array}\right] \tag{10}$$

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

Proof. Let us assume that the Equation (9) holds for some $t \geq 0$.

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

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

Let us call $P_t^N \stackrel{\text{def}}{=} K(a_t,\mathrm{C}_t^N)$. According to lemma 6 there exists a Gaussian variable H_t independent of G_t and of covariance D such that we can replace M_{t+1}^N (without changing M_t and C_t) by a random variables $\widetilde{\mathrm{M}}_{t+1}^N$ with the same laws such that:

$$\sqrt{N}(\widetilde{\mathrm{M}}_{t+1}^N - \mathrm{M}_t^N P_t^N) \xrightarrow{a.s} H_t. \tag{11}$$

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

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

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

$$\begin{aligned}\lim_{N\to\infty} m_t\sqrt{N}(P_t^N - P_t)_j &= \sum_{i=1}^{S} m_{t_i} \sum_{k=1}^{d} \frac{\partial K_{ij}}{\partial c_{t_k}}(a_t,c_t)(XU)_{S+k} \\ &= \sum_{k=1}^{d} Q_{kj}(XU)_{S+k}.\end{aligned}$$

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

$$H_t + \sum_{k=1}^{d} Q_{kj}(XU)_{S+k} + \sum_{i=1}^{S} (XU)_i P_{ij} \tag{12}$$

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

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

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

For $1 \leq j, j' \leq S$, $\mathrm{M}'_{j,j'}$ is the expectation of (12) taken in j times (12) taken in j'. Using the facts that $\mathbb{E}[(XU)_{S+k}(XU)_{S+k'}] = \mathrm{C}_{kk'}$, $\mathbb{E}[(XU)_{S+k}(XU)_i] = O_{ik}$ and $\mathbb{E}[(XU)_i(XU)_{i'}] = \mathrm{M}_{ii'}$, this leads to:

$$\begin{aligned}\mathrm{M}'_{j,j'} &= \mathbb{E}[H_j H'_j] + \sum_{k,k'} Q_{kj}Q_{k'j'}\mathrm{C}_{kk'} + \sum_{k,i'} Q_{kj}O_{i'k}P_{i'j'} \\ &\quad + \sum_{i,k'} Q_{k'j'}O_{ik'}P_{ij} + \sum_{i,i'} P_{ij}\mathrm{M}_{ii'}P_{i'j} \\ &= D_{jj'} + (Q^T\mathrm{C}Q)_{jj'} + (Q^TO^TP)_{jj'} + (P^TOQ)_{jj'} + (P^T\mathrm{M}P)_{jj'}.\end{aligned}$$

By similar computation, we can write similar equations for O' and C' that lead to Equation (10). □

Lemma 6. *Let M^N be a sequence of random measure on $\{1,\dots,S\}$ and P^N a sequence of random stochastic matrices on $\{1,\dots,S\}$ such that $(\mathrm{M}^N, P^N) \xrightarrow{a.s} (m,p)$. Let $(U_{ik})_{1\leq i\leq S, k\geq 1}$ be a collection of* iid *random variables following the uniform distribution on $[0;1]$ and independent of P^N and M^N and let us define Y^N: for all $1 \leq j \leq S$:*

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

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

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

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

$$\left\{\begin{array}{lcll} D_{jj} & = & \sum_i m_i p_{ij}(1-p_{ij}) & \\ D_{jk} & = & -\sum_i m_i p_{ij}p_{ik} & (j \neq k). \end{array}\right. \tag{13}$$

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

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

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

The first application of the mean field CLT is to show that it also works for the cost. Let us assume that the controller takes actions $a_0 \dots a_{T-1}$ and let us introduce the definition of $R^N_{a_0\dots a_{T-1}}(\mathrm{M}_0^N, \mathrm{C}_0^N) = \sum_{t=1}^T(\mathrm{r}_t(\mathrm{M}_t^N, \mathrm{C}_t^N))$ and $r_{a_0\dots a_{T-1}}(m_0,c_0) = \sum_{t=1}^T \mathrm{r}_t(m_t,c_t)$. Lemma 2, says that $R^N_{a_0\dots a_{T-1}}(\mathrm{M}_0^N, \mathrm{C}_0^N) \xrightarrow{a.s} r_{a_0\dots a_{T-1}}(m_0,c_0)$, the following results is more accurate:

Corollary 7 (Application of the CLT to reward)**.** *Under assumption (A1,A2,A3,A4-bis,A5-bis), if the controller takes the actions $a_0 \dots a_{T-1}$ and if we call $\mathbf{D}\mathrm{r}_t(m_t, c_t)$ the differential of $\mathrm{r}_t(\mathrm{M}, \mathrm{C})$ at the point (m_t, c_t), we have:*

$$\begin{aligned}\sqrt{N}(R^N_{a_0\dots a_{T-1}}(\mathrm{M}^N_0, \mathrm{C}^N_0) - r_{a_0\dots a_{T-1}}(m_0, c_0)) \\ \xrightarrow{\mathcal{L}} \textstyle\sum_{t=1}^T \mathbf{D}\mathrm{r}_t(m_t, c_t)G_t.\end{aligned} \tag{14}$$

Proof. Let $G_0 \dots G_T$ be the Gaussian variables defined in the central limit theorem. The proof of Theorem 5 says that one can replace $(\mathrm{M}^N_t, \mathrm{C}^N_t)$ by variables with the same law such that the convergence is almost sure. Let ω be an event such that $\lim_N \sqrt{N}((\mathrm{M}^N_t(\omega), \mathrm{C}^N_t(\omega)) - (m, c)_t)) = G_t(\omega)$. For this event, we have $\lim_{N\to\infty} \sqrt{N}(c_t(\mathrm{M}^N_t, \mathrm{C}^N_t) - \mathrm{r}_t(m_t, c_t)) = \mathbf{D}\mathrm{r}_t(m_t, c_t)G_t$ which leads to Equation (14) by using a Taylor expansion at order one. □

As the means of the Gaussian variables are 0, we have directly:

Corollary 8. *Under the same assumptions and if the convergence of the initial condition is almost sure ($(\mathrm{M}^N_0, \mathrm{C}^N_0) \xrightarrow{a.s} (m_0, c_0)$), one has:*

$$\begin{aligned}\sqrt{N}\Big|V^N_{a_0\dots a_{T-1}}(\mathrm{M}^N_0, \mathrm{C}^N_0) - v_{a_0\dots a_{T-1}}(m_0, c_0)\Big| \\ \leq_{N\to\infty} |\mathbf{D}\mathrm{r}_0(m_0, c_t)G_0| \quad \text{a.s.}\end{aligned} \tag{15}$$

Proof. $v^N_{a_0\dots a_{T-1}}(\mathrm{M}^N_0, \mathrm{C}^N_0) - v_{a_0\dots a_{T-1}}(m_0, c_0) = \mathrm{r}(\mathrm{M}^N_0, \mathrm{C}^N_0) - \mathrm{r}(m_0, c_0) + \mathbb{E}_{\mathrm{M}^N_0, \mathrm{C}^N_0}[r^N_{1\dots T}(\mathrm{M}^N_1, \mathrm{C}^N_1) - \mathrm{r}_{1\dots T}(m_1, c_1)]$. As $\sqrt{N}((\mathrm{M}^N_0, \mathrm{C}^N_0) - (m_0, c_0))$ converges almost surely, the first part of the sum can be upper bounded by $|\mathbf{D}\mathrm{r}_0(m_0, c_0)G_0|$. As for the second part of the sum, using the Berry-Esseen Theorem (Durrett 2.4.d [9]), one can refine Lemma 6 and show that the convergence is uniform. Therefore one can switch the expectation and the limit, the second part of the sum becomes $\mathbb{E}_{\mathrm{M}^N_0, \mathrm{C}^N_0}[\lim_{N\to\infty} \sqrt{N}(r^N_{1\dots T}(\mathrm{M}^N_1, \mathrm{C}^N_1) - \mathrm{r}_{1\dots T}(m_1, c_1))] =_{a.s} 0$ which proves Equation (15). □

We are now ready for the proof of Theorem 4.

of theorem 4. For a vector G, let us write $\|G\|_1 = \sum_i |G_i|$. Because of assumption (A4), there exists a compact set $\mathcal{B}$ such that for all t from 0 to T, $\mathrm{M}^N_t, \mathrm{C}^N_t$ will remain in $\mathcal{B}$.

Let us prove by induction on t from T to 0 that there exist $\beta_t, \gamma_t \in \mathbb{R}^+$ such that if there exists a Gaussian variable G_t satisfying $\sqrt{N}\big((\mathrm{M}^N_t, \mathrm{C}^N_t) - (m_t, c_t)\big) \xrightarrow{a.s} G_t$, then

$$\begin{aligned}\limsup_{N\to\infty} \sqrt{N}\Big|V^{*N}_{t\dots T}(\mathrm{M}^N_t, \mathrm{C}^N_t) - v^*_{t\dots T}(m_t, c_t)\Big| \\ \leq \beta_t\|G_t\|_\infty + \gamma_t.\end{aligned} \tag{16}$$

For $t = T$, Corollary 8 can be used to transform Equation (16) into $\sqrt{N}|\mathbf{D}\mathrm{r}_T(m_T, c_T)G_T| \leq \|\mathbf{D}\mathrm{r}_t(m_T, c_T)\|_1\|G_T\|_\infty$. Therefore, Inequality (16) is true if $\beta_T = \|\mathbf{D}\mathrm{r}_t(m_T, c_T)\|_1$ and $\gamma_T = 0$.

Let us assume that (16) holds for some $t+1 \leq T$ and that $\sqrt{N}\big((\mathrm{M}^N_t, \mathrm{C}^N_t) - (m_t, c_t)\big) \xrightarrow{a.s} G_t$. At time t, (16) can be upper bounded by:

$$\begin{aligned}&\sqrt{N}|\mathrm{r}_t(\mathrm{M}^N_t, \mathrm{C}^N_t) - \mathrm{r}_t(m_t, c_t)| \\ &+\sqrt{N}\Big|\sup_a \mathbb{E}_{\mathrm{M}^N_t, \mathrm{C}^N_t}[V^{*N}_{t\dots T}(\Phi^N_a(\mathrm{M}^N_t, \mathrm{C}^N_t))] \\ &\qquad - \sup_a v^*_{t\dots T}(\Phi_a(m_t, c_t))\Big|.\end{aligned}$$

The first part can be bounded by $\|\mathbf{D}\mathrm{r}_t(m_t, c_t)\|_1\|G_t\|_\infty$. The rest of the proof focuses in the second part of the sum. In the proof of Theorem 5, we showed that for all a (up to the replacement

of $\Phi_a^N(\mathrm{M}_t^N, \mathrm{C}_t^N)$ by a random variable with the same law), there exists a matrix P_a and a Gaussian variable G_a independent of G_t such that $\sqrt{N}(((\mathrm{M}_t^N, \mathrm{C}_t^N), (\mathrm{M}_{t+1}^N, \mathrm{C}_{t+1}^N)) - ((m_t, c_t), (m_{t+1}, c_{t+1})))$ converges almost surely to $(G_t, P_a G_t + G_a)$. Using the fact that $\sup_a f(a) - \sup_a g(a) \leq \sup_a (f(a) - g(a))$, the expectation can be upper bounded by:

$$\sup_a \sqrt{N} \mathbb{E}_{\mathrm{M}_t^N, \mathrm{C}_t^N} \left| V_{t+1\ldots T}^{*N}(\Phi_a^N(\mathrm{M}_t^N, \mathrm{C}_t^N)) - v_{t+1\ldots T}^*(\Phi_a(m_t, c_t)) \right|.$$

Let us consider an arbitrary action a. The Berry-Esseen Theorem shows that $\sqrt{N}((\mathrm{M}_{t+1}^N, \mathrm{C}_{t+1}^N) - (m_{t+1}, c_{t+1})) - P_a G_t$ converges uniformly to G_a, therefore we can switch the limit in N and the expectation and by induction, it can be upper bounded by $\mathbb{E}_G[\gamma_t \|P_a G_t + G_a\|_\infty + \beta_{t+1}] \leq \beta_{t+1} \|P_a G_t\|_\infty + \gamma_t + \beta_t \mathbb{E}[\|G_a\|_\infty]$. As $\mathcal{A}$ is compact and $(\mathrm{M}_{t+1}^N, \mathrm{C}_{t+1}^N)$ remains in a compact set $\mathcal{B}$ (Equation (1)), $\sup_{a \in \mathcal{A}, (\mathrm{M},\mathrm{C}) \in \mathcal{B}} \|P_a\|_1 < \infty$ and $\sup_{a \in \mathcal{A}, (\mathrm{M},\mathrm{C}) \in \mathcal{B}} \mathbb{E}[\|G_a\|_\infty] < \infty$. Thus to obtain an uniform bound on all (M, C), taking $\beta_t \stackrel{\text{def}}{=} \beta_{t+1} \sup_{\mathcal{A},\mathcal{B}} \|P_a\|_1$ and $\gamma_t \stackrel{\text{def}}{=} \gamma_{t+1} + \beta_{t+1} \sup_{\mathcal{A},\mathcal{B}} \mathbb{E}[\|G_a\|_\infty]$ satisfy (16).

Assumption (A4bis) says that at time $t = 0$, $\sqrt{N}\big((\mathrm{M}_t^N, \mathrm{C}_t^N) - (m_t, c_t)\big) \to G_t$ holds in distribution. Using appropriate random variables $(\tilde{\mathrm{M}}_t^N, \tilde{\mathrm{C}}_t^N)$ with the same laws as $(\mathrm{M}_t^N, \mathrm{C}_t^N)$ makes this convergence almost sure so that the induction above holds from $t = 0$. This ends the proof for assertion i of the theorem.

As for assertion *ii*, it comes from the triangular inequality

$$\begin{aligned}
&\left| V_T^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) - V_{a_0^* \cdots a_T^*}^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) \right| \\
&\quad \leq \left| V_T^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) - v_T^*(m_0, c_0) \right| \\
&\qquad + \left| v_T^*(m_0, c_0) - V_{a_0^* \cdots a_T^*}^{*N}(\mathrm{M}_0^N, \mathrm{C}_0^N) \right|.
\end{aligned}$$

An upper bound on the first term of the right side comes from assertion i and the second term can be bounded using Corollary 8. This ends the proof. □

3.3 Infinite horizon discounted reward

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

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

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

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

$$V_\Pi^N(\mathrm{M}_0^N, \mathrm{C}_0^N) \stackrel{\text{def}}{=} \mathbb{E}_\Pi \Big[\sum_{t=1}^{\infty} \delta^{t-1} \mathrm{r}(\mathrm{M}_t^N, \mathrm{C}_t^N) \Big].$$

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

$$V_T^{*N}(\mathrm{M}_0, \mathrm{C}_0) \stackrel{\text{def}}{=} \sup_{\Pi} \mathbb{E}_\Pi\big[\sum_{t=1}^{T} \delta^{t-1} \mathrm{r}(\mathrm{M}_t, \mathrm{C}_t)\big].$$

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

$$\lim_{T\to\infty} \sup_{\mathrm{M},\mathrm{C}} \Big| V_T^{*N}(\mathrm{M}, \mathrm{C}) - V^{*N}(\mathrm{M}, \mathrm{C}) \Big| = 0. \tag{17}$$

Equation (17) is the key of the following analysis. Using this fact, we can prove the convergence when N grows large for fixed T and then let T go to infinity. Therefore with a very few changes in the proofs of Section 3.1, we have the following result:

Theorem 9 (Optimal discounted case)**.** *Under assumptions (A1,A2,A3,A4,A6,A7), as N grows large, the optimal discounted reward of the stochastic system converges to the optimal discounted reward of the deterministic system:*

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

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

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

3.3.1 Problems for other infinite horizon criteria

Again, 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. As for other other infinite-horizon criteria such as average reward or its variants, the average reward is (if it exists) $\lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\Pi \sum_{t=1}^{T} c(\mathrm{M}_t, \mathrm{C}_t)$.

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 particles in state 1 ($\mathrm{C}_t = (\mathrm{M}_t)_1$) and with a function $f{:}[0; 1]{\to}[0; 1]$ such that the transition kernel K is $K_{i1}(\mathrm{C}) = f(\mathrm{C})$ for $i \in \mathcal{S}$. If $\mathrm{M}_0^N(0) \xrightarrow{a.s} m_0$ then for any fixed t, M_t^N converges to $f(f(\dots f(m_0)\dots))$. Using techniques that can be found in [7], one can prove that as N grows large, $\lim_{t\to\infty} \mathrm{M}_t^N$ might converges to almost any subset of $L{\subset}[0; 1]$ such that $L = f(L)$. However, in general $\lim_{t\to\infty} \lim_{N\to\infty} \mathrm{M}_t^N \neq \lim_{N\to\infty} \lim_{t\to\infty} \mathrm{M}_t^N$. For example if $f(x) = x$, the deterministic system is constant while the stochastic system converges almost surely to a random variable (as a bounded Martingale) that takes values in $\{0; 1\}$.

Similar difficulties arise for the central limit theorem in the discounted case: the convergence depends on the behavior of the system when T tends to infinity.

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 [12, 4, 5].

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

$$B_{t+1}^i = \left(B_t^i - \mu_i X_t^i + a_t^i Y_t \right)^+. \tag{18}$$

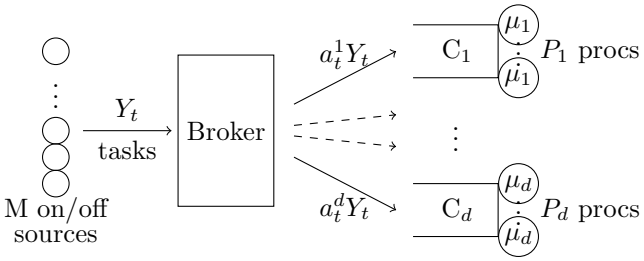

Figure 1: The routing system

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

This problem can be viewed as a multidimensional restless bandit problem where computing the optimal policy for the broker is known to be a hard problem [17]. Here, indexability may help to compute near optimal policies by solving one MDP for each queue [17, 16]. 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 particles evolving in a common environment.

- There are $N \stackrel{\text{def}}{=} A + \sum_{i=1}^d P_i$ "particles". Each particle 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 particle is an element of $\mathcal{S} = \left\{(x,e) | x \in \{s, q_1, \cdots, q_d\}, e \in \{\text{on}, \text{off}\}\right\}$. the population mix M is the proportion of sources in state on and the proportion of servers in state on, for each queue.
- The action of the controller are the routing choices of the broker: a_t^d is the probability that a task is sent to queue d at time t.

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

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

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

- Here, the transition kernel can be time dependent but is independent of a and C. The probability of a particle 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 particles. For any N, a system with N particles is composed of $\lfloor A_0 N/N_0 \rfloor$ (resp. $\lfloor P_i N/N_0 \rfloor$) particles that are sources (resp. servers in queue i). The remaining particles (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 particles that are sources is A/N_0 and the mean number of particles that are servers in queue i is $P_i N/N_0$. Then, each of these particles changes state over time according to the probabilities $K_{u,v}(a, C)(t)$. At time $t = 0$, a particle is in state "on" with probability one half.

It should be clear that this system satisfies Assumptions (A1) to (A4) and therefore one can apply the convergence theorem 3 to this system that shows that if using the policies a^* or Π^*, when N goes to infinity the system converges to a deterministic system with optimal cost. An explicit computation of the policies a^* and Π^* is possible here and is postponed to Section 4.3.

4.2 CLT applicability

As for the central limit theorem, 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. Unfortunately Assumption (A5) does not hold since the function g is not differentiable when $\mathrm{C}_t^i - \mu_i X_t^i + a_t^i Y_t = 0$. However, as mentioned in the beginning of section 3.2 the differentiability condition in Assumption (A5) can be replaced by a Lipschitz continuity condition. Let us consider Assumption (A5-ter):

(A5-ter) **Continuous Lipschitz** – For all t and all $i, j \in \mathcal{S}$, all functions g, K_{ij} and r_t are Lipschitz continuous on all compact sets of their domain.

This assumption is weaker than (A5) since, if a function is C^1, it is Lipschitz on every compact set (with Lipschitz constant $\sup ||f'||$). In the example, function g has a right-derivative and a left-derivative at all points and therefore satisfies (A5-ter). The central limit theorem 4 should apply here as well:

Theorem 10. *Theorem 4 still holds when replacing (A5) by (A5-ter).*

(Sketch of the proof). The proof is very similar to the one of 4 and we just sketch the main differences.

As seen at the end of section 2.3, all variables are almost surely bounded. By assumption (A5-ter), all functions are Lipschitz, thus let $L_g, L_K, L_{\mathrm{r}_t}$ be the Lipschitz constants on the compact space $\mathcal{B}$ (see Equation (1)) for g, K and r_t respectively and $L = \max\{L_g, L_K, L_{\mathrm{r}_t}\}$. The main

idea is to replace all equalities in the proof of all CLT theorems by inequalities. For instance, in Theorem 5, Equation (9) is replaced by the following statement: for all $x_1 \dots x_t \in \mathbb{R}^t$,

$$\begin{aligned} \limsup_N \mathbb{P}\big(\sqrt{N}(\|(\mathrm{M}_0^N, \mathrm{C}_0^N) - (m_0, c_0)\|_\infty, \dots, \\ \|(\mathrm{M}_t^N, \mathrm{C}_t^N) - (m_t, c_t)\|_\infty\big) \geq (x_1 \dots x_t)\big) \\ \leq \mathbb{P}((\|G_0\|_\infty, \dots, \|G_t\|_\infty) \leq (x_1 \dots x_t)) \end{aligned} \tag{19}$$

where the variables G_t have covariance $\Gamma_t = L^2\Gamma_{t-1} + D_{t-1}$. The other steps in the proof can be changed in almost the same way. Formula (14) in Corollary 7 is replaced by

$$\begin{aligned} \sqrt{N}|R^N_{a_0 \dots a_{T-1}}(\mathrm{M}_0^N, \mathrm{C}_0^N) - r_{a_0 \dots a_{T-1}}(m_0, c_0)| \\ \leq_{st} \sum_{t=0}^T L\|G_t\|_\infty \end{aligned} \tag{20}$$

and Formula (15) of Corollary 8 by

$$\begin{aligned} \sqrt{N}\Big|V^N_{a_0 \dots a_{T-1}}(\mathrm{M}_0^N, \mathrm{C}_0^N) - v_{a_0 \dots a_{T-1}}(m_0, c_0)\Big| \\ \leq \alpha\|G_0\|_\infty + \delta, \quad \text{a.s.} \end{aligned} \tag{21}$$

where α and δ are constants depending on L. □

4.3 Optimal policy for the deterministic limit

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

$$\min_{y_1^1 \dots y_T^d} \{ \sum_{t=1}^T \sum_{i=1}^d c_t^i \text{ with } \begin{array}{l} c_{t+1}^i = (c_t^i + y_t^i - x_t^i)^+ \\ \sum_i y_t^i = y_t \end{array} \}. \tag{22}$$

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

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

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

The principle of the algorithm is the following.

1. The processors in all queues, which are "on" at time t with a speed μ are seen as slots of size μ.
2. At each time t, y_t units of tasks have to be allocated. This is done in a greedy fashion by filling up the empty slots starting from time t. Once all slots at time t are full, slots at time $t+1$ are considered and are filled up with the remaining volume of tasks, and so forth up to time T.
3. The remaining tasks that do not fit in the slots before T are allocated in an arbitrary fashion.

See figure 2 for an illustration of the execution of the algorithm on an example. It should be clear that the algorithm is linear in the number of slots nk and that this algorithm computes an optimal allocation.

Time t	0	1	2	3	4	5	6
y_t (tasks)	8	1	0	1	7	6	6
Queue 1	X	τ_0	τ_0	τ_3	τ_4	τ_4	τ_6
	X	τ_0	τ_0		τ_4	τ_5	
	τ_0				τ_4		
					τ_4		
Queue 2			X	X		τ_5	τ_6
						τ_5	
Queue 3	X	τ_0	τ_1		τ_4	τ_5	τ_6
	X	τ_0			τ_4	τ_5	τ_6
		τ_0				τ_5	
Optimal allocation	5	.	.	1	5	1	1+2
	.	.	.	.	.	2	1
	3	1	.	.	2	3	2

Figure 2: This figure presents an example of an execution of the algorithm. We consider a case with 3 queues. At $t = 0$ (resp. $1, ..., 6$) there are 8 (resp. $1, 0, 1, 7, 6, 6$) packets arriving in the system. Each processor has speed 1 and the processors in state "off" are represented by grey cells (for example, at time 0, there are respectively $3, 0$ and 2 processors available in queue $1, 2$ and 3). All queues start at time 0 with 2 packets. The top part of the table shows at which time a packet will be processed while the bottom part shows the corresponding optimal allocation (`X` represent tasks present in the queues before $t = 0$; A label τ_i in a slot of queue j at time t represents one task arriving at time i allocated to queue j that will be processed at time t. The number of slots with label τ_i should be equal to y_i; At the end, 2 packets cannot be allocated in empty slots. They are routed arbitrarily (in queue 1)).

4.4 Numerical example

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

1. Deterministic policy a^* – to obtain this curve, the optimal actions $a_0^* \dots a_{T-1}^*$ 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 Π^* – here, the optimal policy Π^* for the deterministic case was first computed. When the stochastic system is in state $(\mathrm{M}_t^N, \mathrm{C}_t^N)$ at time t, we apply the action $\Pi_t^*(\mathrm{M}_t^N, \mathrm{C}_t^N)$ and the corresponding cost up to time T is reported.

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

The results are reported in Figures 3 and 4.

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

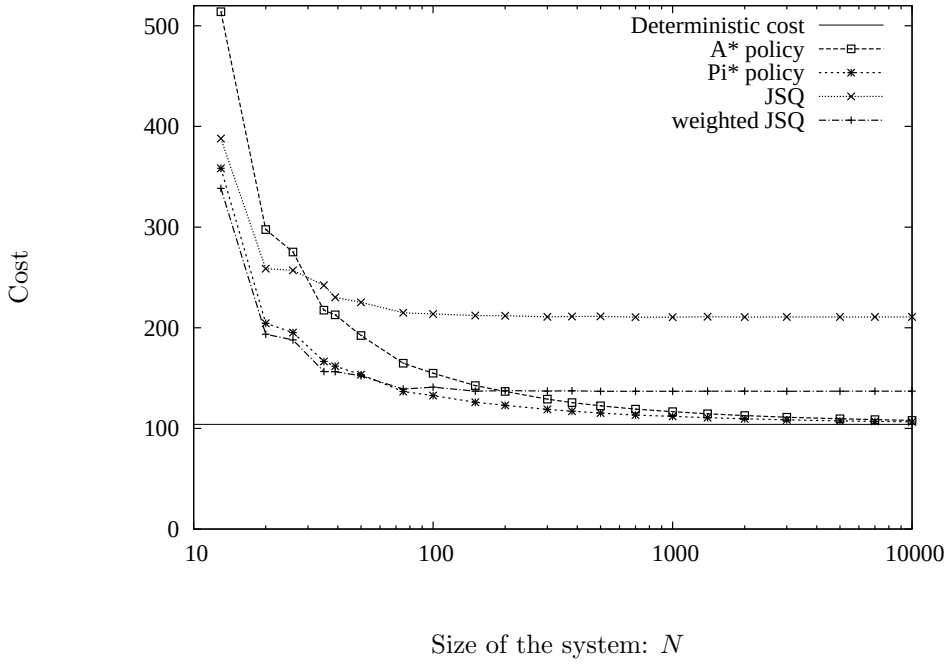

Figure 3: Expected cost of the policies a^*, Π^*, JSQ and W-JSQ for different values of N.

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 3 – and less than 5% for the central limit theorem – figure 4.

Figure 3 shows the average waiting time of the stochastic system when we apply the different policies. The horizontal line represents the optimal cost of the deterministic system $v^*(m_0, c_0)$ which is probably less than $V^{*N}(\mathrm{M}_0, \mathrm{C}_0)$. This figure illustrates Theorem 3: if we apply a^* or Π^*, the cost converges to $v^*(m_0, c_0)$.

In Figure 3, 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. $\lfloor \frac{N}{N_0} P_i \rfloor$) particles that are sources (resp. processors in queue i) and the remaining particles are distributed randomly. The random choice of the remaining states are chosen so that $\mathbb{E}[A^N] = \frac{N}{N_0} A$, but the difference $A^N - N N_0 A$ may be large. Therefore, for some N the load of the system is much higher than the average load, leading to larger costs. As N grows, the proportion of remaining particles decreases and the phenomena becomes negligible.

A second feature that shows in Figure 3, is the fact that on all curves, the expected waiting times are decreasing when N grows. This behavior is certainly related to Ross conjecture [15] 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 Π^* is uniformly good even for small values of N.

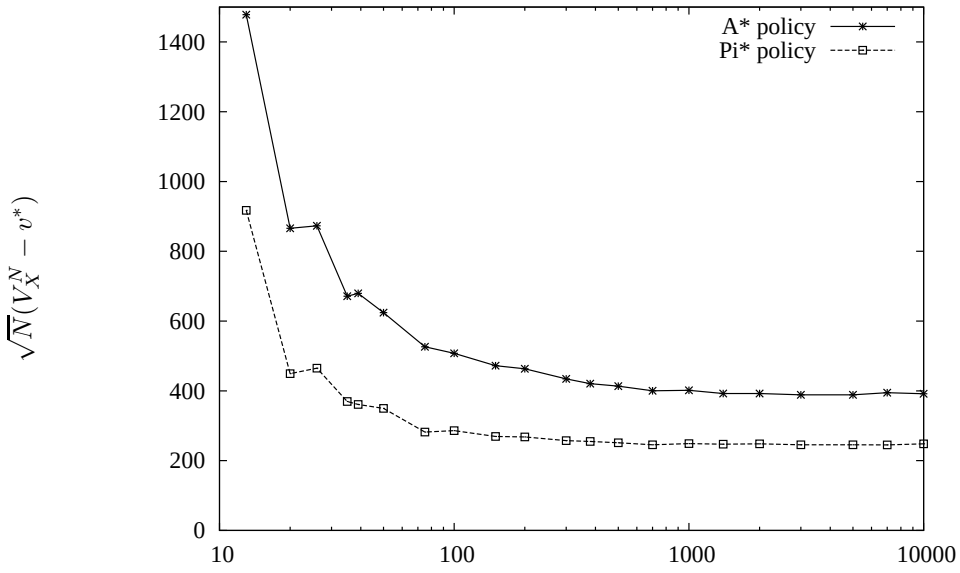

Figure 4: Speed of convergence of the policies $X = a^*$ or Π^* for different values of N.

The figure 4 illustrates Theorem 4 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 Π^* (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 should be an upper bound on the constant δ defined in Equation (21).

Besides comparing a^* and Π^* to other heuristics, it would be interesting to compare it to the optimal policy of the stochastic system, whose cost is $V^{*N}(\mathrm{M}, \mathrm{C})$. One way to compute this optimum would be by using Equation (3). 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 particle's state can vary in {on,off}. Therefore even with $N = 10$ and if we only compute the cost for queues of size less than 10, this leads to $2^N 10^5 \approx 10^8$ states which is hard to handle even with powerful computers.

5 Computational issues

Throughout the paper, we have shown that if the controller uses the optimal policy Π^* of the deterministic limit of the finite real system, the expected cost will be close to the optimal one (Theorem 3). Moreover, Theorem 4 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 [11] for a review on the subject). Imagine that a 2-approximation algorithm exists for the deterministic system, then, Theorem 3 proves that for all ε, this algorithm will be a $(2+\varepsilon)$-approximation for the stochastic system if N is large enough. Finally, heuristics for the deterministic system can also be applied to the stochastic version of the system.

If none of this works properly, one can also compute the optimal deterministic policy by "brute-force" computations using Equation (3): $v^*_{t\dots T}(m,c) = \mathrm{r}_t(m,c) + \sup_a v^*_{t+1\dots T}(\Phi_a(m,c))$. In that case, an approximation of the optimal policy is obtained by discretizing the state space and by solving the equation backward (from $t = T$ to $t = 0$), to obtain the optimal policy for all states. The brute force approach can also be applied directly on the stochastic equation using (2): $V^{*N}_{t\dots T}(\mathrm{M},\mathrm{C}) = \mathrm{r}_t(\mathrm{M},\mathrm{C}) + \sup_{a\in\mathcal{A}} \mathbb{E}_{\mathrm{M},\mathrm{C}}\left[V^{*N}_{t+1\dots T}\big(\Phi^N_a(\mathrm{M},\mathrm{C})\big)\right]$. 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 particles: it depends mostly on the smoothness of functions g and ϕ rather than on N. The second one is the suppression of the expectation which might reduce the computational time by a polynomial factor[1] by replacing the $|\mathbb{P}_N(\mathcal{S})|$ possible values of M^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.

6 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. First, if one allows the actions to depend on the particles, it seems natural that the limit behavior of such systems is the same as the limit behavior of systems where the actions are random variables and that they both converge to mean field system whose cost is averaged. Another 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] V. Anantharam and C. Bordenave. Optimal control of interacting particle systems. *Private Communication*, 2008.

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

[3] M. Benaım and J.Y. Le Boudec. A Class Of Mean Field Interaction Models for Computer and Communication Systems. *To appear in Performance Evaluation*.

[4] Vandy Berten and Bruno Gaujal. Brokering strategies in computational grids using stochastic prediction models. *Parallel Computing*, 2007. Special Issue on Large Scale Grids.

[5] Vandy Berten and Bruno Gaujal. Grid brokering for batch allocation using indexes. In *Euro-FGI NET-COOP*, Avignon, France, 2007. LNCS.

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

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

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

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

[10] Carl 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.

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

[12] Jennie Palmer and Isi 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).

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

[14] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. New York, NY, USA, 1994.

[15] 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.

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

[17] 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., j. gani edition, 1988.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399