Mean Field Optimization in Discrete Time

Nicolas Gast and Bruno Gaujal

Grenoble University **INRIA**

Warwick – May, 2012

Outline

Empirical Measure

Markovian model

- N objects $(O_1(t) \dots O_N(t))$.
- $(O_1(t), \dots O_N(t))$ is a finite homogeneous discrete time Markov chain over S^N.
- Dynamics is invariant under any permutation of the objects.

Empirical measure : $M^N(t) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \delta_{O_n(t)}$.

Under permutation invariance, $M^N(t)$ is a finite homogeneous discrete time Markov chain over $\mathscr{P}_N(\mathcal{S})$ the set of probability measures p on $\{1 \dots S\}$, such that $Np(i) \in \mathbb{N}$ for all $i \in \mathcal{S}$.
When N goes to infinity, it converges to $\mathscr{P}(\mathcal{S})$ the set of probability measures on $\mathcal{S}$.

Context

Context The system of objects evolves depending on their common environment (context).
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.

Actions and Policies

Action The action space $\mathcal{A}$ is assumed to be a compact subset of $\mathbb{R}^k$.

Kernel For an action $a \in \mathcal{A}$ and an environment $C \in \mathbb{R}^d$, each object evolves independently of the others, according to a Transition matrix

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

$K_{i,j}(a, C)$ is continuous in a and C.

Policy $\pi = (\pi_1 \pi_2 \dots)$ specifies the action taken at each time step. When the state space is finite, deterministic policies are dominant:
$\pi_t : \mathscr{P}(\mathcal{S}) \times \mathbb{R}^d \to \mathcal{A}$ is deterministic.
The variables $M_\pi^N(t), C_\pi^N(t)$ denote the state of the system at time t under policy π.
$(M_\pi^N(t), C_\pi^N(t))_{t \geq 0}$ is a sequence of random variables on $\mathscr{P}_N(\mathcal{S}) \times \mathbb{R}^d$.

Reward functions

To each state $(M(t), C(t))$, we associate a reward $r_t(M, C)$ (invariant by permutation of the objects).
In the finite-horizon case, the controller maximizes 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 π is:

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

In the infinite-horizon discounted case, let $0 \leq \delta < 1$, the δ-discounted reward associated to the policy π is the quantity:

$$V_{(\delta),\pi}^N(M_0^N, C_0^N) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{\infty} \delta^t r_t(M_\pi^N(t), C_\pi^N(t))\right].$$

Main Assumptions

(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 $\|C\| = \sup_i |C_i|$.

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:

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

Let π be a policy and consider a realization of the sequence $(M^N(t), C^N(t))$. At time t, a controller that uses policy π, 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 (1). The quantity $m_{A_\pi^N}(t)$ is a random variable depending on the sequence A_π^N (and is deterministic once A_π^N is fixed).

Convergence Theorem

Theorem (Controlled mean field)

Under (A1,A3) and π, $\exists\, \mathcal{E}_t(\epsilon, x), \lim_{\epsilon\to 0, x\to 0} \mathcal{E}_t(\epsilon, x) = 0$ *s.t.* $\forall t$*:*

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

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

$$\mathcal{E}_0(\epsilon, \ell) \stackrel{\text{def}}{=} \ell;$$

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

Proof.

By induction on t: at each step, the system stays close to the deterministic approximation with high probability. □

Convergence Theorem (II)

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

Under assumptions (A1,A3,A4),

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

Optimal Mean Field

The reward of the deterministic system starting at $m(0), c(0)$ under the sequence of action a is:

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

optimal cost: $v_*(m(0), c(0)) \stackrel{\text{def}}{=} \max_{a \in \mathcal{A}^T}\{v_a(m(0), c(0))\}$. An argmax sequence in this equation is not unique. In the following, a^* will be one of such sequence and will be called the sequence of *optimal limit actions*.

Theorem (Convergence of the optimal reward)

Under (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: A.s.,*

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

Proof

Let a^* be optimal for the deterministic limit:
$\lim_{N\to\infty} V_{a^*}^N\left(M^N(0), C^N(0)\right) = v_{a^*}(m(0), c(0)) = v_*(m(0), c(0))$.

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

Conversely, let π_*^N be optimal for the stochastic system and $A_{\pi_*^N}^N$ the corresponding actions. It is suboptimal for the deterministic limit: $v_*(m(0), c(0)) \geq v_{A_{\pi_*^N}^N}(m(0), c(0))$.

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

Discussion

- Determinitic optimal cost is the limit of the optimal cost.
- deterministic policy is asymptotically optimal.
- As N grows, the reward of the constant policy $a_0^*, \ldots, a_{t-1}^*$ converges to the optimal reward: the value of information vanishes.
- **Adaptive policy**: $\mu_t^*(m(t), c(t))$ 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 $M^N(0), C^N(0)$:

$$\lim_N V_{\mu^*}^N\left(M^N(0), C^N(0)\right) = \lim_N V_{a^*}^N\left(M^N(0), C^N(0)\right) = \lim_N v_*(m(0), c(0)).$$

However, the policy μ^* is not necessarily continuous and $M_{\mu^*}^N, C_{\mu^*}^N$ may not have limits with N.

Second order results

Theorem

Under (A1,A2,A3,A4), there exist constants γ *and* γ' *such that if* $\epsilon_0^N \stackrel{\text{def}}{=} \left\| M^N(0), C^N(0) - m(0), c(0) \right\|$ *For any policy* π*:*

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

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

Infinite horizon

(A5) 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 $K_t(a, C) = K(a, C)$.

(A6) Bounded reward $\sup_{M,C} r(M, C) \leq K < \infty$.

The rewards are discounted according to a discount factor $0 \leq \delta < 1$:

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

Infinite horizon (II)

Theorem ((Optimal discounted case))

Under (A1,A2,A3,A4,A5,A6), as N grows, 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:*
$v_*(m,c) = r(m,c) + \delta \sup_{a\in\mathcal{A}} \left\{ v_*(\Phi_a(m,c)) \right\}$.

Proof.

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

As $r < K$, the gap $\left|V^N_{T*} - V^N_*\right|$ is bounded independently of N, M, C:

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

Second order result

Proposition

Under (A1,A2,A3,A4,A5,A6) 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* $\max(1, L_g)(S + L_K + 1)\delta < 1$, *there exists constants* H *and* H' *s.t.*

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

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

Average Reward

The optimal average reward is

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

This raises the problem of the exchange of the limits $N \to \infty$ and $T \to \infty$.

Convergence to a global attractor

Let $f_a : B \to B$ denote the deterministic function corresponding to one step of the evolution of the deterministic limit under action a:

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

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.

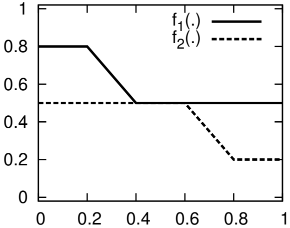
Proposition

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

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

Non-convergence in the controlled case

Consider a system with 2 states $\{0, 1\}$, where $C^N = M_0^N$ is the proportion of objects in state 0. Two actions (1 and 2) are possible, corresponding to a probability of transition from any state to 0 of $f_1(C)$ and $f_2(C)$ resp. Both f_1 and f_2 are piecewise linear functions:



Non-convergence in the controlled case(II)

The reward is equal to $|C^N - 1/2|$.

Both f_1 and f_2 have the same unique attractor, equal to $\{1/2\}$.

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.

An example: brokering in parallel queues

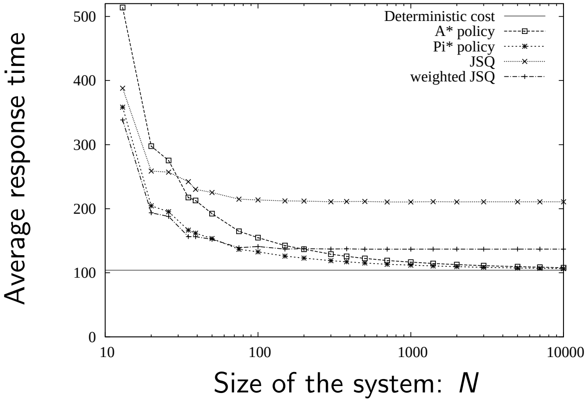
- P applications sending tasks to the broker (ON/OFF).
- C clusters (one buffer per cluster).
- K processors per clusters, using the push/pull model (ON/OFF).

Goal of the broker: allocate tasks to clusters to minimise the average response time.

Number of objects: $N = P + CK$, with a reducible transition matrix. Intensity is $O(1)$ so that the limit system lives in discrete time.

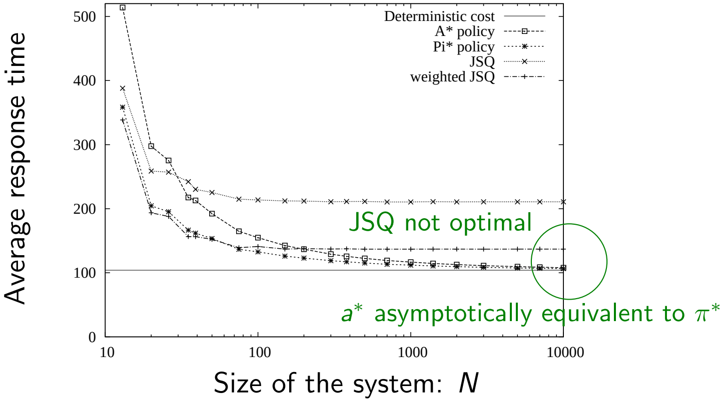
Simulations

- $V_{a^*}^N$ average response time of the optimal open loop policy: action at time t is $a^*(t)$.
- $V_{\pi^*}^N$ average response time of the optimal closed loop policy: action at time t is $\pi^*(t, M(t))$.



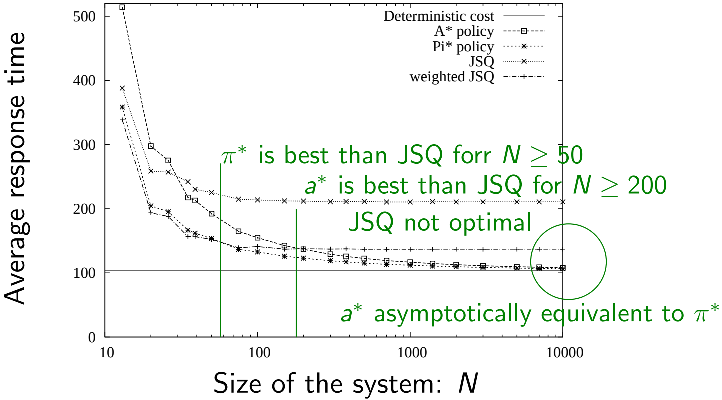
Simulations

- $V_{a^*}^N$ average response time of the optimal open loop policy: action at time t is $a^*(t)$.
- $V_{\pi^*}^N$ average response time of the optimal closed loop policy: action at time t is $\pi^*(t, M(t))$.



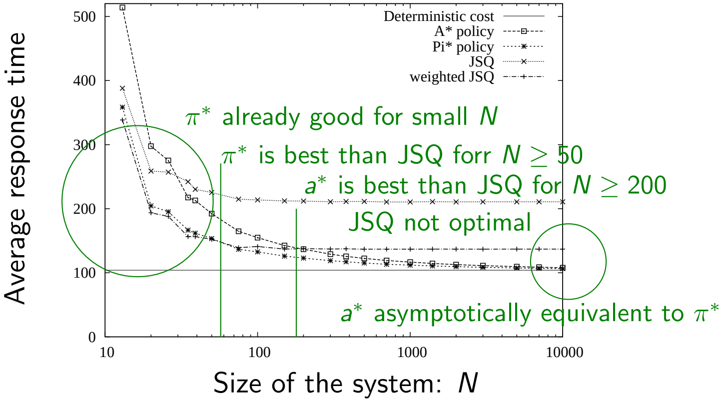
Simulations

- $V_{a^*}^N$ average response time of the optimal open loop policy: action at time t is $a^*(t)$.
- $V_{\pi^*}^N$ average response time of the optimal closed loop policy: action at time t is $\pi^*(t, M(t))$.



Simulations

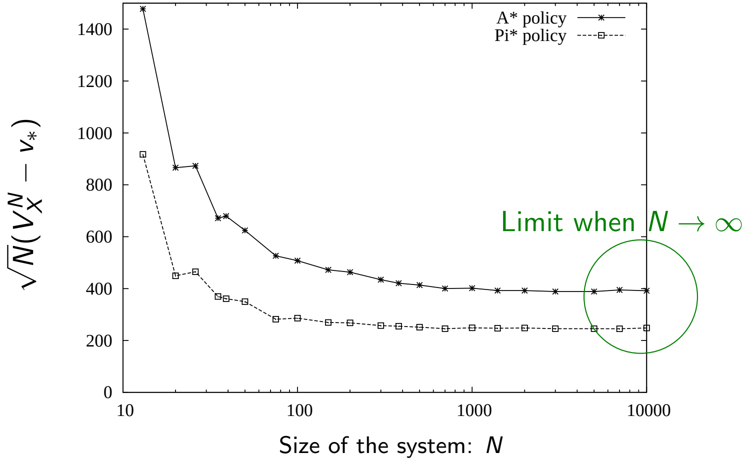
- $V^N_{a^*}$ average response time of the optimal open loop policy: action at time t is $a^*(t)$.
- $V^N_{\pi^*}$ average response time of the optimal closed loop policy: action at time t is $\pi^*(t, M(t))$.



Central limit theorem

We compute

$$\sqrt{N}(V^N_{\pi^*} - v_*) \quad \text{et} \quad \sqrt{N}(V^N_{a^*} - v_*)$$



Beyond deterministic limits

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

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

where $A^N(t) \stackrel{\text{def}}{=} \pi_t(\widetilde{M}^N_\pi(t), \widetilde{C}^N_\pi(t))$ and $G_t(a, \widetilde{C}^N_\pi(t))$ is a sequence of *i.i.d.* Gaussian random variables independent of all $\widetilde{M}^N_\pi(t'), \widetilde{C}^N_\pi(t')$ for $t' < t$. 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 $j \neq k$:

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

Beyond deterministic limits (II)

Theorem

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 $a = a_1 \dots a_T$:*

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

(ii)

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

Object-Dependent Actions

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 ith object.

We also construct a second system by replacing 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 endure action a according to the distribution p and evolves independently according to $K(a, C)$.

Object-Dependent Actions (II)

The difference between the 2 systems collapses as N grows. Other results, such as second order results, also hold.

Proposition

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 $\mathscr{P}(\mathcal{A})$.