# Leader-follower MDP models with factored state space and many followers - followers abstraction, structured dynamics and state aggregation

Régis Sabbadin, Anne France Viet

*ECAI 2016*
*G.A. Kaminka et al. (Eds.)*

*doi:10.3233/978-1-61499-672-9-116*

# Leader-Follower MDP Models with Factored State Space and Many Followers – Followers Abstraction, Structured Dynamics and State Aggregation

**Régis Sabbadin**[1] and **Anne-France Viet**[2]

**Abstract.** The *Leader-Follower Markov Decision Processes* (LF-MDP) framework extends both Markov Decision Processes (MDP) and Stochastic Games. It provides a model where an agent (the *leader*) can influence a set of other agents (the *followers*) which are playing a stochastic game, by modifying their immediate reward functions, but not their dynamics. It is assumed that all agents act selfishly and try to optimize their own long-term expected reward. Finding equilibrium strategies in a LF-MDP is hard, especially when the joint state space of followers is factored. In this case, it takes exponential time in the number of followers. Our theoretical contribution is threefold. First, we analyze a natural assumption (*substitutability of followers*), which holds in many applications. Under this assumption, we show that a LF-MDP can be solved exactly in polynomial time, when deterministic equilibria exist for all games encountered in the LF-MDP. Second, we show that an additional assumption of *sparsity* of the problem dynamics allows us to decrease the exponent of the polynomial. Finally, we present a *state-aggregation approximation*, which decreases further the exponent and allows us to approximately solve large problems. We empirically validate the LF-MDP approach on a class of realistic animal disease control problems. For problems of this class, we find deterministic equilibria for all games. Using our first two results, we are able to solve the exact LF-MDP problem with 15 followers (compared to 6 or 7 in the original model). Using state-aggregation, problems with up to 50 followers can be solved approximately. The approximation quality is evaluated by comparison with the exact approach on problems with 12 and 15 followers.

## 1 Introduction

The *Leader-Follower Markov Decision Processes* (LF-MDP) framework [20] is a framework which has been recently proposed to extend both Markov Decision Processes (MDP) [14] and Stochastic Games (SG) [18, 4]. In a LF-MDP, an agent (the *leader*) partially controls the reward functions of several *followers* acting selfishly in a stochastic game, to optimize their long-term expected reward. However, the leader does not influence the dynamics of the stochastic game, which is only governed by the followers' actions. Some recent applications of the LF-MDP framework include management in organizations [21, 13].

Many real-life problems exist where a set of followers act selfishly on a dynamical system in order to maximize their own long-term profit (in a game-theoretic fashion) while a leader, not acting directly on the system, fixes the rules of the game so that game equilibria favor its own long term objective. The following are intuitive examples of such problems:

- **Carbon tax:** here, the leader fixes a carbon emission tax level (as a modifiable rate of the total carbon emissions), while followers (firms) can take costly measures to decrease their own carbon emission rate. The followers are the only ones to emit carbon, but their rewards/costs are functions of other followers actions (through global carbon emissions) and leader's actions (through taxes). The leader has its own profit function, which depends on total carbon emission as well as total taxes paid by followers.
- **Soccer league:** In a soccer league, clubs (followers) want to maximize both the number of points they score during a whole season (which determines their ranking) and their financial profit. The league (leader) does not own a club, but wishes to maximize its own profit (through taxes on clubs' benefits) and the interest of the championship (which helps generating profit). The league's actions consist in modifying game rules, introducing salary cap, changing tax level, etc. But these do not change directly the state of the system (ranking, points...).
- **Animal health management:** Individual farmers (followers) breed cattle in an area where some disease can spread. Their aim is to maximize their own profit. Control actions (depopulation, treatment) can be applied by followers, but at some cost. The leader can decide on financial incentives to control. These cost him money if followers apply control actions, but the reduction in the disease spread rewards the leader [16].

Even though solution algorithms have been proposed for LF-MDP, based on dynamic programming [16] or reinforcement learning [21], these do not scale to the case where the followers' joint state space is factored, except under very drastic assumptions (no more than two states for each follower in [16]). Even in the case where the state space is not factored, one has to solve multiple instances of $n$ players games, where $n$ is the number of followers, which can only be done in time exponential in $n$.

In this article, we consider LF-MDP in which $n$ is higher than in usual LF-MDP ($> 10$) and where the joint state space is a product of followers state spaces. After reviewing the LF-MDP model in Section 2, we show in Section 3 that under some natural assumptions about the followers (substitutability, structured dynamics), we can find equilibrium strategies for the leader and followers in time polynomial in $n$. Substitutability of followers, in particular, occurs when the transition and reward functions of each follower do not depend on

[1] MIAT, INRA,Toulouse, France, mail: Regis.Sabbadin@toulouse.inra.fr
[2] BIOEPAR, INRA, Oniris, Nantes, France, email: anne-france.viet@oniris-nantes.fr

the "labeling of other followers". When this holds, we show that the equilibrium policies of followers which are in the same state are also the same. This suggests that a LF-MDP with substitutable followers can be replaced with smaller a LF-MDP, where the number of followers is reduced to the size of the followers' state space. However, we show that, unfortunately, the solution of the reduced LF-MDP is different from the one of the original LF-MDP in general, except when solution polices are deterministic. Fortunately, our experiments show that this seems to occur very often in practice...

While polynomial in $n$, time complexity of reduced LF-MDP is still exponential in the size of the state space of each follower, which keeps them hard to solve, especially when $n$ is large (>15). We thus present an approximation of the solution through *state aggregation* which decreases the value of the exponent of $n$ in time complexity. Finally, in Section 5, we present an illustration of the approach based on a realistic problem of coordination of farmers to limit the spread of the Porcine Reproductive and Respiratory Syndrome within a group of farms [11]. It is used, in particular, to empirically validate the quality of approximate policies obtained through state aggregation.

## 2 The Leader-Follower MDP model

### 2.1 Definition of the LF-MDP model

#### 2.1.1 States, actions, transitions and rewards

A single leader/multiple followers finite-horizon MDP model [20] is a multiple time steps decision process involving one *leader* and $n$ *followers*. It is defined, in the finite horizon case, as[3]: $\mathcal{M} =< n, \Sigma, A^L, \{A_i^F\}_{i=1..n}, T, r^L, \{r_i^F\}_{i=1..n}, H >$, where:

- $\Sigma$ is the joint state space of the leader and the followers. It can have a very general form and can be factored, e.g. as $\Sigma = S^L \times S_1^F \times \ldots \times S_n^F$.
- $A^L = \{1, \ldots, m\}$ is the finite leader action space.
- $A_i^F = \{1, \ldots, p_i\}$ is the finite action space of follower $i$. For sake of notational simplicity, we will consider in this paper that all followers have the same action space $A^F = \{1, \ldots, p\}$.
- $T : \Sigma \times (A^F)^n \times \Sigma \to [0, 1]$ is the joint state transition function. $T\left(\sigma' | \sigma, \{a_i^F\}_{i=1..n}\right)$ is the probability to transition from state $\sigma$ to state $\sigma'$, when the actions of the followers are set to $a_F = \{a_i^F\}_{i=1..n}$. Note that the leader's actions do not influence transition probabilities.
- $r^L : \Sigma \times A^L \times (A^F)^n \to \Re$ is the leader instant reward function.
- $r_i^F : \Sigma \times A^L \times A^F \to \Re$ is the instant reward function of follower $i$.
- $H$ is the horizon of the problem.

#### 2.1.2 Policies of the leader and the followers

As usual in finite horizon sequential decision problems, we assume that agents choose their actions at time step $t$ according to non-stationary *policies*, $\delta_t^L, \{\delta_{t,i}^F\}_{i=1\ldots n}$. We will focus on *Markovian*, *stochastic* policies. $\delta_t^L(a^L|\sigma)$ is the probability that $a^L \in A^L$ is chosen by the leader at time $t$, given current state $\sigma \in \Sigma$. $\delta_{t,i}^F(a_i^F|\sigma, a^L)$ is the probability that $a_i^F \in A^F$ is chosen by follower $i$ at time $t$, given current state $\sigma$ and after having observed the current action $a^L$ of the leader.

Policies are *deterministic*, when $\delta_t^L, \{\delta_{t,i}^F\}_{i=1\ldots n}$ take value in $\{0, 1\}$. In this case, we write $a^L = \delta_t^L(\sigma)$ or $a_i^F = \delta_{t,i}^F(\sigma, a^L)$.

[3] Transitions and rewards are considered stationary for the sake of notational simplicity, but the results can be easily extended to the non-stationary case.

#### 2.1.3 Values of policies, equilibrium policies

Let $\Delta = \left\{\delta_t^L, \{\delta_{t,i}^F\}_{i=1\ldots n}\right\}_{t=1\ldots H}$ be a given joint policy of the leader and the followers. The *values* $Q_\Delta^L$ and $Q_\Delta^{F,i}$ to the leader and the followers are defined as follows, in every joint state and time step:

$$Q_\Delta^L(\sigma, t) = E\Big[\sum_{t'=t}^{H} r_{t'}^L \mid \Delta, \sigma\Big], \quad (1)$$

$$Q_\Delta^{F,i}(\sigma, t) = E\Big[\sum_{t'=t}^{H} r_{t',i}^F \mid \Delta, \sigma\Big]. \quad (2)$$

Solving a LF-MDP consists in finding an *equilibrium joint policy*, $\Delta^* = \left\{\delta_t^{L*}, \{\delta_{t,i}^{F*}\}_{i=1\ldots n}\right\}_{t=1\ldots H}$, for the leader and the followers[4].

**Definition 1 (LF-MDP equilibrium joint policy)** *$\Delta^* = \left\{\delta_t^{L*}, \{\delta_{t,i}^{F*}\}_{i=1\ldots n}\right\}_{t=1\ldots H}$ is an equilibrium policy if and only if it verifies, $\forall t, \delta_t^L, \{\delta_{t,i}^F\}, \sigma$:*

$$Q_{\Delta^*}^L(\sigma, t) \geq Q_{\Delta^{*\downarrow\delta^L}}^L(\sigma, t), \forall \delta^L, \quad (3)$$

$$Q_{\Delta^*}^{F,i}(\sigma, t) \geq Q_{\Delta^{*\downarrow\delta_i^F}}^{F,i}(\sigma, t), \forall i, \delta_i^F. \quad (4)$$

*$\Delta^{*\downarrow\delta^L}$ (resp. $\Delta^{*\downarrow\delta_i^F}$) is the set of policies where the $\delta_t^{L*}$ (resp. $\delta_{t,i}^{F*}$) have been replaced with arbitrary policy $\delta_t^L$ (resp. $\delta_{t,i}^F$), $\forall t$ ($\forall i$).*

In [20], extending results of [4] from stochastic games to LF-MDP, it was shown that there exist at least one Markovian equilibrium joint policy in which the leader equilibrium policies are deterministic. Such an equilibrium policy can be computed by a *backward induction* type algorithm [14], interleaving Nash equilibria computation steps for the followers and backward induction steps for the leader, at each time step.

### 2.2 LF-MDP solution algorithm

Let $\mathcal{M}$ be a LF-MDP. An equilibrium joint policy, $\Delta^*$ can be computed backward by the following algorithm [16]:

#### 2.2.1 Final time step

At the final time step, $H$, any follower $i$ applying action $a_i^F \in A^F$ while the state is $\sigma$ and the leader action is $a^L$, receives an immediate reward $r_i^F(\sigma, a^L, a_i^F)$, regardless of the other followers' actions. Therefore, $\delta_{H,i}^{F*}$is deterministic and:

$$\delta_{H,i}^{F*}(\sigma, a^L) \in \arg\max_{a_i^F \in A^F} r_i^F(\sigma, a^L, a_i^F) \text{ and}$$

$$Q_{\Delta^*, a^L}^{F,i}(\sigma, H) = \max_{a_i^F \in A^F} r_i^F(\sigma, a^L, a_i^F), \forall (\sigma, a^L). \quad (5)$$

We define the expected immediate reward of the leader at time $t \in \{1, \ldots, H\}$, for a joint followers stochastic policy $\delta_t^F = \{\delta_{t,i}^F\}_{i=1..n}$:

$$r_{\delta_t^F}^L(\sigma, a^L) = \sum_{a^F} \Big(\prod_{i=1}^{n} \delta_{t,i}^F(a_i^F|\sigma, a^L)\Big) r^L(\sigma, a^L, a^F). \quad (6)$$

[4] In the following, we use the terms *equilibrium joint policy* (or *equilibrium policy*, for short) for a solution of a LF-MDP and *Nash equilibrium* for the solution of a normal form game, in order to avoid confusion between the two notions.

Then, for the leader at time step $H$:

$$\begin{aligned} \delta_H^{L*}(\sigma) &\in \arg\max_{a^L \in A^L} r^L_{\delta_H^{F*}}(\sigma, a^L), \\ Q^L_{\Delta^*}(\sigma, H) &= \max_{a^L \in \mathcal{A}^L} r^L_{\delta_H^{F*}}(\sigma, a^L), \forall \sigma. \end{aligned} \quad (7)$$

#### 2.2.2 Induction step

A followers' joint equilibrium policy at time step $t$, given subsequent time steps joint equilibrium policies, is defined inductively as stochastic Nash equilibria[5] of normal form $n$-players games $(\forall \sigma, a^L)$ [10], where each player's action belongs to $A^F$.

The game value to player $i$ of joint action $a^F$ in state $\sigma$ at time $t$ under leader action $a^L$ and assuming that a joint equilibrium policy is applied at subsequent time steps is defined as:

$$\begin{aligned} G^t_{\sigma,a^L,\Delta^*}(i, a^F) &= r^F_i(\sigma, a^L, a^F_i) \\ &+ \sum_{\sigma'} T\left(\sigma' | \sigma, a^F\right) Q^{F,i}_{\Delta^*}(\sigma', t+1). \end{aligned} \quad (8)$$

Let $\{\alpha_1^*, \ldots, \alpha_n^*\}$ be a solution of the game $G^t_{\sigma,a^L,\Delta^*}$ ($\alpha_i^*$ is a probability distribution over $A^F$). A followers joint equilibrium policy is given by: $\delta^{F*}_{t,i}(a^F_{t,i} | \sigma, a^L) = \alpha_i^*(a^F_i)$ and

$$Q^{F,i}_{\Delta^*,a^L}(\sigma, t) = \sum_{a^F} \Big( \prod_{j=1}^{n} \alpha_j^*(a^F_j) \Big) G^t_{\sigma,a^L,\Delta^*}(i, a^F). \quad (9)$$

Since a followers' Nash equilibrium is determined from action $a^L$ through Equation (9), the leader optimal policies can be computed as the solutions of a non-stationary Markov Decision Process $< \Sigma, A^L, \{T_{\delta_t^{F*}}\}, \{r^L_{\delta_t^{F*}}\}_{t=1..H}, H >$, where the $\{r^L_{\delta_t^{F*}}\}_{t=1..H}$ have been defined previously and

$$T_{\delta_t^{F*}}(\sigma' | \sigma, a^L) = \sum_{a^F} \prod_{j=1}^{n} \delta^{F*}_{t,j}(a^F_j | \sigma, a^L) T(\sigma' | \sigma, a^F). \quad (10)$$

Functions $T_{\delta_t^{F*}}$ and $r^L_{\delta_t^{F*}}$ are determined before being required at each step of the backward induction algorithm.

Optimal policies $\delta_t^{L*}(\sigma)$ and value functions $Q^{L*}_{\Delta^*}(\sigma, t)$ are also computed backward:

$$\begin{aligned} \delta_t^{L*}(\sigma) &\in \arg\max_{a^L \in \mathcal{A}^L} \Big\{ r^L_{\delta_t^{F*}}(\sigma, a^L) \\ &+ \sum_{\sigma' \in \Sigma} T_{\delta_t^{F*}}(\sigma' | \sigma, a^L) Q^{L*}_{\Delta^*}(\sigma', t+1) \Big\}, \\ Q^L_{\Delta^*}(\sigma, t) &= \max_{a^L \in \mathcal{A}^L} \Big\{ r^L_{\delta_t^{F*}}(\sigma, a^L) \\ &+ \sum_{\sigma' \in \Sigma} T_{\delta_t^{F*}}(\sigma' | \sigma, a^L) Q^{L*}_{\Delta^*}(\sigma', t+1) \Big\}. \end{aligned} \quad (11)$$

### 2.3 Computational complexity considerations

The various steps of the generic LF-MDP solution algorithm described above have different time and space complexities.

[5] Note that since game solutions are involved, there may be more than one Nash equilibrium, leading to different equilibrium values. When solving a LF-MDP, one is usually interested into finding a single such Nash equilibrium.

#### 2.3.1 Step 1: Normal form games generation

To compute a followers Nash equilibrium, we need to build normal form games $G^t_{\sigma,a^L,\Delta^*}$ (Equation 8). Each game has $O\left(n \times |A^F|^n\right)$ elements and there are $|\Sigma| \times |A^L|$ games. The time complexity to generate them all is thus $O\left(n \times |A^F|^n \times |\Sigma| \times |A^L|\right)$, but we require to store only one such game at a time.

#### 2.3.2 Step 2: Followers policies computation and storage

Followers Nash equilibria are solutions of the games computed above. Storing followers' optimal policies requires space in $O(n \times |\Sigma| \times |A^F| \times |A^L|)$. Finding an (approximate) Nash equilibrium in a game is a hard task[6] by itself [2].

#### 2.3.3 Step 3: Leader transition and reward computation

Transition tables $T_{\delta_t^{F*}}$ are computed through Equation 10. They require $O\left(|\Sigma|^2 \times |A^L|\right)$ space to store and time $O\left(n \times |A^F|^n \times |\Sigma|^2 \times |A^L|\right)$ to compute. The reward functions $r_{\delta_t^{F*}}$ require $O\left(|\Sigma| \times |A^L|\right)$ space to store and time $O\left(n \times |A^F|^n \times |\Sigma| \times |A^L|\right)$ to compute, using Equation 6.

#### 2.3.4 Step 4: Leader dynamic programming step

$\delta_t^{L*}$ requires $O(|\Sigma|)$ space to store and $O\left(|\Sigma|^2 \times |A^L|\right)$ to compute (Equation 11).

#### 2.3.5 What can we do to decrease space and time complexity?

Given these complexity considerations, one can notice that the time and space complexities of all steps of solving a LF-MDP are at least either exponential in $n$ or at least linear in $|\Sigma|$ (or both). In the case where the joint state of the problem $\sigma \in \Sigma$ is factored[7], for example when $\Sigma = \left(S^F\right)^n$, $|\Sigma|$ is itself exponential in $n$.

Next, we explore the property of *substitutability of followers* in LF-MDP problems. Under this property, the above steps can be performed, exactly or approximately, at a lower complexity cost. It allows *state abstraction*, a classical property of factored MDP [6, 3, 9]. It also allows *followers abstraction*, i.e. a potential reduction of the *number of players* of all considered games. Since the complexity of solving games (and LF-MDP) is exponential in the number of followers, this may induce an important reduction in time (and space) complexity. Furthermore, in some cases, followers abstraction leads to an exact solution of the LF-MDP.

We will consider two other complexity reduction approaches.
(i) Structured followers dynamics: We will exploit the *sparsity* of the transition matrix of each follower, to reduce further LF-MDP solution complexity, without adding new approximations.
(ii) Joint state aggregation: An additional state abstraction approach will allow us to design an approximate LF-MDP solution method that scales to problem with 50-100 followers.

Substitutability and aggregation are particularly legible properties when the leader policy must be expressed in a simple and intelligible way and when followers' states are imperfectly observed. It is

[6] This problem is PPAD-complete, where PPAD is a specific complexity class, "believed" to strictly include P.

[7] In the most general case, $\Sigma = S^L \times S^F_1 \times \ldots \times S^F_n$, but we will give up the dependency on $S^L$ to simplify notations.

common in human-based systems (economical or social), where the leader often has to consider aggregate states of "anonymous" followers.

# 3 Exploiting problem structure to decrease the complexity of solving LF-MDP

## 3.1 Followers substitutability

### 3.1.1 State space reduction through substitutability

In Section 2, we considered *global state* $\sigma = (s_1, \ldots, s_n)$. but, in many applications, followers are *substitutable*: in the view of the leader, all followers in the same state behave identically.

**Definition 2 (Substitutability of followers)**
*Followers are substitutable*[8] *in a LF-MDP $\mathcal{M}$ if and only if:*

- $S_i^F = S_j^F, A_i^F = A_j^F$ *and* $r_i^F = r_j^F, \forall i, j \in \{1..n\}^2$.
- *For any $\tau$ and $\tau_{-i}$, permutations of $\{1, \ldots, n\}$ where $\tau_{-i}$ leaves $i$ at its place ($\tau_{-i}(i) = i$), we have $T(\sigma'_\tau|\sigma_\tau, a^F_\tau) = T(\sigma'|\sigma, a^F)$, $r^L(\sigma_\tau, a^L, a^F_\tau) = r^L(\sigma, a^L, a^F)$ and $r_i^F(\sigma_{\tau_{-i}}, a^L, a_i^F) = r_i^F(\sigma, a^L, a_i^F)$.*

**Example 1** *Followers are substitutable when:*

- $T(\sigma'|\sigma, a^F) = \prod_{i=1}^n p(s'_i|s_i, f(\sigma), a_i^F)$,
- $r^L(\sigma, a^L, a^F) = \sum_{i=1}^n r_L(s_i, a^L, a_i^F)$ *and*
- $r^F(\sigma, a^L, a_i^F) = r_F(s_i, a^L, a_i^F)$,

*with $r_L$ and $r_F$ rewards functions and provided that function $f$ verifies $f(\sigma) = f(\sigma_\tau), \forall \tau$.*

**Proposition 1 (Substitutability of optimal policies)** *If followers are substitutable in a LF-MDP $\mathcal{M}$, then: $\delta_t^{L*}(\sigma) = \delta_t^{L*}(\sigma_\tau)$, $\delta_{t,i}^{F*}(\cdot|\sigma, a^L) = \delta_{t,i}^{F*}(\cdot|\sigma_{\tau_{-i}}, a^L)$, $\forall, \sigma, t, \tau, \tau_{-i}$. The same property holds for $Q^L_{\Delta^*}$ and $Q^{F,i}_{\Delta^*}$ functions.*

**Sketch of proof:** The proof follows the induction. Step $H$ is easy, using the invariance of $r^L$ and $r^F$ from which easily follows the substitutability of $r_{\delta_H^F}$, $\delta_{H,i}^{F*}$, $Q^{F,i}_{\Delta^*}(\cdot, H)$, $\delta_H^{L*}$ and $Q^L_{\Delta^*}(\cdot, H)$. The induction step goes as easily, once we have noticed that games $G^t_{\sigma, a^L, \Delta^*}$ are also substitutable. □

An important consequence of this proposition is that if a LF-MDP is substitutable, then it can be replaced with an equivalent LF-MDP which (reduced) state space, denoted $\Sigma^L$ is composed of the *equivalence classes* of $\Sigma$ for the "permutation" relation $\equiv$: $\sigma \equiv \sigma'$ iff $\exists \tau, \sigma' = \sigma_\tau$.

It is easy to show that any equivalence class can be represented by a *reduced global state*, modeling the number of followers in each of the $k = |S^F|$ states. This reduced global state can thus be represented by the tuple of integers:

$$c = (c_1, \ldots, c_k) \in \{0, \ldots, n\}^{|S^F|}, \text{ where } \sum_{h=1}^k c_h = n. \quad (12)$$

$\Sigma^L$ is thus the set of tuples satisfying Equation 12. The size of the reduced state space of the leader is $|\Sigma^L| = \binom{n+k-1}{k-1} = O(n^k)$, instead of $k^n$ for $|\Sigma|$.

The time and space complexities of the steps of the algorithm (Table 1) are now reduced, by replacing every occurrence of $|\Sigma|$, with $|\Sigma^L|$. Step 4 becomes polynomial in $n$.

[8] Note that the current definition has links with *stochastic bisimilarity* [6]. However, stochastic bisimilarity only handles factored state space, not factored action space.

### 3.1.2 Action space reduction through substitutability

Followers' substitutability implies the substitutability of followers policies $\delta_{t,i}^{F*}$. In the case where followers' policies are deterministic, then any two followers' actions are identical when the followers are in the same state. So, the action profiles of the $n$ followers, $(a_1, \ldots, a_n)$, are limited to profiles in which all followers in the same state perform the same action. In this way, a followers' action profile is of the form $d^F = (d_1^F, \ldots, d_k^F) \in (A^F)^k$, leading to decreased joint action space size ($|A^F|^k$ instead of $|A^F|^n$).

However, optimal policies of followers can be stochastic, meaning that two identical policies may lead to different actions choices. Thus, it may be that even under the substitutability assumption, a Nash equilibrium for the followers may give non-zero probabilities to all the $|A^F|^n$ potential profiles. For equilibrium computation, in order to limit the size of the joint action space, we make the approximation that even when followers policies are stochastic, followers in the same state actually implement the same action.

This suggests the following modifications to Equations 8 and 9, which are now used to compute games and followers policies of dimension $k$: $\forall h = 1, \ldots, k$,

$$\begin{aligned} G^t_{c,a^L,\Delta^*}(h, d^F) &= r_h^F(c, a^L, d_h^F) \\ &+ \sum_{c'} \bar{T}\left(c'|c, d^F\right) Q^{F,h}_{\Delta^*}(c', t+1). \end{aligned} \quad (13)$$

And, if $\{\alpha_1^*, \ldots, \alpha_k^*\}$ is a stochastic Nash equilibrium of the above game, $\delta_{t,h}^{F*}(d_h^F|c, a^L) = \alpha_h^*(d_h^F)$ and

$$Q^{F,h}_{\Delta^*,a^L}(c, t) = \sum_{d^F} \prod_{j=1}^k \alpha_j^*(d_j^F) G^t_{c,a^L,\Delta^*}(h, d^F). \quad (14)$$

We will see in the next subsection how $\bar{T}$ is defined, but first remark that with the above approximation it is assumed that actions for all followers are chosen in the following way : An action $d_h^F \in A^F$ is chosen at random for each $h \in 1, \ldots, k$, following distribution $\delta_{t,h}^{F*}(\cdot|c, a^L)$, then all followers $j$ which are in state $h$ apply the same action $d_h^F$.

When all Nash equilibria are deterministic, this assumption holds, even in the original LF-MDP. In all other cases, the computed equilibria are only approximate. Notice also that when replacing Equations 8 and 9 with Equations 13 and 14, the time complexities of the steps are reduced. For Step 1, for example, it becomes $O(k|A^F|^k|\Sigma^L||A^L|) = O(n^k)$. The space complexity of Step 2 becomes $O(n^k)$ and the games to solve are $k$-players games so their time complexity is independent on $n$. In Step 3, Equation 10 can be replaced with: $\forall c, c' \in \Sigma^L, a^L \in A^L$,

$$T_{\delta_t^{F*}}(c'|c, a^L) = \sum_{d^F} \prod_{h=1}^k \delta_{t,h}^{F*}(d_h^F|c, a^L) \bar{T}(c'|c, d^F). \quad (15)$$

The complexity of Step 3 also depends on the computation of $\bar{T}$, which we now describe.

### 3.1.3 Transitions in the reduced model

Making use of the reduction of the state and action spaces, we can compute an aggregate transition function $\bar{T} : \Sigma^L \times (A^F)^k \times \Sigma^L \to [0, 1]$. $\bar{T}(c'|c, d^F)$ is the probability to transition from any state $\sigma^c \in \Sigma$ "compatible" with $c$ to any state $\sigma^{c'} \in \Sigma$ "compatible"

with $c'$, when applying an action $a^F = (a_1^F, \ldots, a_n^F)$ "compatible" with $d^F = (d_1^F, \ldots, d_k^F)$.

To be more precise, $\bar{T}$ is defined as: $\forall c, c', d^F$,

$$\bar{T}(c'|c, d^F) = \sum_{\sigma' \models c'} T(\sigma'|\sigma^c, \ldots, \underbrace{d_h^F, \ldots, d_h^F}_{c_h}, \ldots), \qquad (16)$$

where $\sigma^c = (\underbrace{1, \ldots, 1}_{c_1}, \ldots, \underbrace{k, \ldots, k}_{c_k})$ is a full state compatible with $c$. $\sigma' \models c'$ means that if $\sigma' = (s'_1, \ldots, s'_n)$ then, $\forall h = 1, \ldots, k$, there are exactly $c'_h$ indices $i$ such that $s'_i = h$.

**Proposition 2** *$\bar{T}$ defined in Equation 16 is well-defined.*

**Proof:** By well-defined, we mean that $\bar{T}(c'|c, d^F)$ does not depend on the actual choice of $\sigma^c$ compatible with $c$. Indeed, this holds thanks to the substitutability of $T$. □

The space and time complexities of computing $\bar{T}$ are $O\left(|\Sigma^L|^2|A^F|^k\right)$ and $O\left(n|\Sigma||\Sigma^L||A^F|^k\right)$. Indeed, since we have to sum over all $\sigma'$ compatible with $c'$ and this for all $c'$ to compute $\bar{T}$, we still have to explore $\Sigma$ entirely once. Hence, computing $\bar{T}$ is still exponential in $n$, even though the exponent has been reduced from $2n$ to $n$.

However, for a fixed $c$ and given follower state $h$, the $c_h$ followers in state $h$ change their state to a new repartition $(c_h^1, \ldots, c_h^k)$ (with $c_h^1 + \ldots + c_h^k = c_h$) according to a multinomial distribution $\Gamma^h_{c_h, d_h^F}(c_h^1, \ldots, c_h^k)$ of parameters $\{p(h'|h, c, d_h^F)\}_{h'=1..k}$. Furthermore, $\forall c, c', d^F$,

$$\bar{T}(c'|c, d^F) = \sum_{\substack{(c_h^1, \ldots, c_h^k),\\ \sum_{h'=1}^k c_h^{h'} = c_h,\\ \sum_{h=1}^k c_h^{h'} = c'_{h'}}} \prod_{h=1}^{k} \Gamma^h_{c_h, d_h^F}(c_h^1, \ldots, c_h^k).$$

**Proposition 3** *The time complexity*[9] *of computing $\bar{T}$ is $O(n^k|\Sigma^L||A^F|^k) = O(n^{2k}|A^F|^k)$.*

**Sketch of proof:** For any $c = (c_1, \ldots, c_k)$, each $c_h$ can transition to $\binom{c_h+k-1}{k-1}$ $k$-tuples $(c_h^1, \ldots, c_h^k)$. Therefore, for fixed $c$ and $d^F$, we need to evaluate only $\sum_h \binom{c_h+k-1}{k-1}$ terms to compute the values $\bar{T}(c'|c, d^F)$, for all feasible $c'$. The result comes from the fact that $\sum_h \binom{c_h+k-1}{k-1} = O(n^k)$. □

### 3.1.4 Approximate reduced LF-MDP: Followers abstraction

The construction of the reduced versions $\bar{r}^L$ and $\bar{r}^F$ is immediate in a LF-MDP with substitutable followers. In the end, we get an approximate LF-MDP $\bar{\mathcal{M}} =< k, \Sigma^L, A^L, \prod_{h=1}^k A_h^F, \bar{T}, \bar{r}^L, \{\bar{r}_h^F\}, H >$, which is only exponential in $k$ (and no more in $n$) to solve. Note the following important proposition:

**Proposition 4** *In the case where the joint followers equilibrium policies of approximate LF-MDP $\bar{\mathcal{M}}$ are deterministic, they can be used to build deterministic joint equilibrium policies for the original LF-MDP $\mathcal{M}$. This also holds for the leader equilibrium policies which are always deterministic.*

**Sketch of proof:** Just note that the only approximation in the process we have described concerns the interpretation of stochastic Nash equilibria. When Nash equilibria are pure, there is a complete equivalence between $\mathcal{M}$ and $\bar{\mathcal{M}}$. □

In the view of Proposition 4, two facts should be highlighted: (i) Since games may have more than one equilibrium, we should return in priority deterministic equilibria of the games and (ii) Even though the above approach may not be exact in all cases, it is possible to check, after equilibrium policies have been computed in the reduced LF-MDP, whether these provide equilibrium policies in the original LF-MDP. It is enough to check that the games which have been solved for all $(t, c, a^L)$ have a deterministic equilibrium.

### 3.1.5 Connection with state / action abstraction in MDP and Stochastic Games

The state-space reduction approach that we have described in Section 3.1.1 is very close to state aggregation approaches in MDPs [9, 6, 3, 19] or in stochastic games [17]. We basically identify the case where state aggregation leads to an optimal solution in a LF-MDP (extending both MDP and SG cases). What is especially useful here, is that state aggregation leads to an exponential reduction of the state space.

More original is the form of *followers abstraction* in stochastic games that we propose in Section 3.1.2. State and action spaces abstraction have already been proposed in the field of game theory [5, 8] or in stochastic games [17]. However when action spaces are abstracted, as in [17] (the closest work to ours we have found), only followers action spaces $A^F$ are abstracted[10]. In the case we consider, action spaces $A^F$ are already small and need not be abstracted. Instead, we propose to "lump" players together, which is a way to abstract a game where the joint action space is large, due to the number of players and not to the size of individual action spaces. To our knowledge, this is the first proposition in that direction in (stochastic) game theory, or LF-MDP.

## 3.2 Structured dynamics of followers

Recall that a transition $c \to c'$ is obtained through the aggregation of $n$ individual changes of states $h \to h'$. Each transition has probability $p(h'|h, c, d_h^F)$ and is independent of the other transitions, $c$ and $d_h^F$ being given.

We can further decrease complexity when the dynamics is structured, i.e. when a follower in state $h$ can only transition to a few possible states, say at most $N_{Succ} < k$ states.

**Proposition 5** *When, in LF-MDP $\bar{\mathcal{M}}$, followers in any state $h$ transition to at most $N_{Succ}$ possible states, Computing $\bar{T}$ requires $O(n^{k+N_{Succ}}|A^F|^k)$ time.*

**Sketch of proof:** The number of terms that should be computed for a given pair $(c, d^F)$ is reduced to $\sum_h \binom{c_h+N_{Succ}}{N_{Succ}-1} = O(n^{N_{Succ}})$. □

# 4 State aggregation

State aggregation allows us to further decrease the exponent of $n$ in time complexities and suppress it from space complexities, but there is no performance guarantee anymore on the returned policies, and we must resort to experiments to empirically evaluate the merits of this approach (which we will do in the case study section). It consists

[9] The complexity results given in the paper are not the tightest possible, for ease of exposition. Here, for example, $\sum_h \binom{c_h+k-1}{k-1} = O(kn^{k-1}) = o(n^k)$.

[10] The authors considering stochastic games, there is no leader.

**Table 1.** LF-MDP objects complexity, with and without suggested simplifications/approximations.

| | Naive | Substitutability | Aggregation |
|---|---|---|---|
| State space size | $\|\Sigma\| = O\left(k^n\right)$ | $\|\Sigma^L\| = O\left(n^k\right)$ | $\|\overline{\Sigma^L}\| = O\left(K^k\right)$ |
| Action space size | $\|A^F\|^n$ | $\|A^F\|^k$ | $\|A^F\|^k$ |
| Joint state transition | $\|T\| = O\left(k^{2n}\|A^F\|^n\right)$ | $\|\bar{T}\| = O\left(n^{2k}\|A^F\|^k\right)$ \| $\|\bar{T}\| = O\left(n^{k+N_{Succ}}\|A^F\|^k\right)$ | $\|\hat{T}\| = O\left(K^{2k}\|A^F\|^k\right)$ |
| Single game size | $\|G^t_{\sigma,a^L,\Delta^*}\| = O\left(n\|A^F\|^n\right)$ | $\|G^t_{c,a^L,\Delta^*}\| = O\left(k\|A^F\|^k\right)$ | $\|G^t_{\kappa,a^L,\Delta^*}\| = O\left(k\|A^F\|^k\right)$ |
| Nb games / time step | $nb = \|A^F\|^n\|A^L\|$ | $nb = \|A^F\|^k\|A^L\|$ | $nb = K^k\|A^L\|$ |

in considering a partition of the set $\{0,\ldots,n\}$ into $K+2$ intervals, where $K$ is an integer dividing $N$: $I_0 = \{0\}$, $I_{K+1} = \{n\}$ and $I_i = \{\frac{(i-1)n}{K}+1,\ldots,\frac{i.n}{K}\}, \forall i = 1,\ldots,K$. Then, we define *aggregate states* as tuples of integers $(\kappa_1,\ldots,\kappa_k) \in \{0,\ldots,K\}^k$. These correspond to every combinations of sets $I_{\kappa_1} \times \ldots \times I_{\kappa_k}$, such that there exists a reduced state $c \in \Sigma^L, c_h \in I_{\kappa_h}, \forall h = 1,\ldots,k$. Let $\overline{\Sigma^L}$ denote the set of aggregate states. Obviously, $|\overline{\Sigma^L}| = O(K^k)$, which is now independent of $n$.

We define two functions, relating reduced and aggregate states:

- $\kappa(c) \in \overline{\Sigma^L}$ is the unique aggregate state which is compatible with the reduced state $c \in \Sigma^L$,
- $\hat{c}(\kappa) \in \Sigma^L$ is a *representative* of the aggregate state $\kappa \in \overline{\Sigma^L}$, defined as the compatible reduced state which components $\hat{c}_h$ are the "closest possible" to the centers of intervals $I_{\kappa_h}$. These can be computed, e.g. by solving small size integer linear programs (one for each $\kappa$).

Finally, it is possible to generate a LF-MDP on the aggregate state space. The transition function $\widehat{T} : \overline{\Sigma^L} \times \left(A^F\right)^k \times \overline{\Sigma^L} \to [0,1]$ is defined as

$$\widehat{T}(\kappa'|\kappa, d^F) = \sum_{c',\kappa(c')=\kappa'} \bar{T}(c'|\hat{c}(\kappa), d^F). \tag{17}$$

Reward functions $\hat{r}^L$ and $\hat{r}^F_h$ are defined accordingly: $\hat{r}^L(\kappa, a^L, d^F) = \bar{r}^L(\hat{c}(\kappa), a^L, d^F)$ and $\hat{r}^F_h(\kappa, a^L, d^F_h) = \bar{r}^L(\hat{c}(\kappa), a^L, d^F_h)$.

**Proposition 6** *The time complexity of computing $\widehat{T}$ is in $O(|A^F|^k K^k n^k)$.*

The proof is obvious, using the form of Equation 17. □

Complexity still depends on $n$, due to the sum over all $c'$, but all other steps are independent on $n$. Note that since we are interested in probabilities to transition to aggregate states $\kappa'$ and not reduced states $c'$, we could estimate these by simulating the transitions. So doing, the sample size would be a function of $|\overline{\Sigma^L}|$ and not $|\Sigma^L|$, and thus independent on $n$.

Solution policies of this LF-MDP will assign the same policies to all reduced states compatible with the same aggregate state, and of course to the full states compatible with these reduced states. This may result in a loss of quality of the returned strategies. In Section 5, we illustrate the impact of state aggregation on a disease control problem.

The contributions of this paper to LF-MDP complexity reduction are summarized in Table 1.

## 5 Case study

In order to demonstrate the practical interest of exploiting substitutability, structured dynamics and state aggregation in LF-MDP, we will focus on a case study concerning a problem of coordination of farmers to limit the spread of the Porcine Reproductive and Respiratory Syndrome (PRRS) within a group of farms [11].

This problem, as many others in animal health management (or other applications, listed in the introduction) involves some followers and one leader that has the ability to indirectly control their dynamics. The followers each own one herd that may be infected by an endemic disease (PRRS). PRRS is an endemic, non-regulated disease, meaning that treating infected herds is non-compulsory. However, in order to limit the spread of the disease, which impacts pig growth and meat production, farmers associations may propose financial incentives, to limit the cost of treatments to farmers.

This is typically a case where the followers directly influence the dynamics of the system (disease spread) through management, while the leader (the Association) influences their reward functions (through incentives). The farmers are assumed to maximize their own long-term profit, while the association maximizes its own, including both an infection-spread related reward and incentive costs.

### 5.1 The LF-MDP model of PRRS spread control

#### *5.1.1 The model*

We consider a group of $n$ farmers taking decisions for their own herd (followers). PRRS is modeled with a compartmental approach, with five compartments ($k = |S^F| = 5$):

- $S$: Susceptible (=non-infected).
- $S_b$: Susceptible with management (biosecurity measure).
- $I$: Infected by PRRS.
- $I_0$: Infected, with control measure starting. Not fully efficient yet
- $I_C$ Infected, controlled (efficient).

Actions are different in each herd state, but at most two actions ('do nothing', 'manage') are available in each state.

These govern the transition probabilities from each state[11] (see Figure 1):

- $S$: A $S$ herd becomes $I$ (infected) with probability $\beta_S$. Else, either stays $S$ (if action=do nothing) or becomes $S_b$ (action=manage).
- $S_b$: For a $S_b$ herd, only 'manage' action is available. Transition probability to $I$ is reduced to $\nu\beta_S$.
- $I$: Transitions are deterministic: to $I$ (do nothing) or to $I_0$ (manage).
- $I_0$: Only 'manage' action is available: transition to $I_C$ with probability $\psi$, modeling a stochastic sojourn time in state $I_0$.
- $I_C$: Transitions are deterministic: To $I_C$ (do nothing) or to $S$ (manage=depopulation).

[11] Self-transition probabilities are omitted, for sake of readability.

**Table 2.** Values of parameters in the three sets used for $n = 15$. In addition, $\psi = 0.5$, $\beta = (0, 0, 0.08, 0.06, 0.01)$, $\beta_{out} = 0.005$ and $L^L = red \times L^F$.

| Set | $\nu$ | $L^F$ | $c^F$ | $c^L$ | $perc$ | $red$ |
|---|---|---|---|---|---|---|
| 2001 | 0.5 | (0,0,6,5,4) | (4,1,4,2,101) | (0,3) | 0.5 | 0.75 |
| 824 | 0.73 | (0,0,8.76,5.84,2.92) | (8.53,1.46,12.79,2.92,147.46) | (0,4.38) | 0.26 | 0.73 |
| 131 | 0.7 | (0,0,4.8,5.6,2.8) | (7.84,1.4,11.76,2.8,101.4) | (0,4.2) | 0.7 | 0.7 |

Note that while $\nu$ and $\psi$ are constants, $\beta_S$ is a function of the state of all herds. We assume that, the group of herds being tightly linked (geographically, but also by sales and purchases), $\beta_S$ only depends on the total numbers of herds in each of the 5 states (equation 18). This implies the substitutability of the transition model.

$$\beta_S(c) = \frac{1}{n}\sum_{h=1}^{k} \beta(h)c(h) + \beta_{out} \tag{18}$$

where the $\beta(h)$ and $\beta_{out}$ are real-valued parameters.

The reward functions are the following:

$$r^L(\sigma, a^L, a^F) = -c^L(a^L) - \sum_{i=1}^{n} c^F(s_i)q^L(a^L, a_i^F) - L^L(s_i),$$

$$r^F(\sigma, a^L, a_i^F) = -E_{s'_i}\left[L^F(s'_i)\right] - \sum_{i=1}^{n} c^F(s_i)q^F(a^L, a_i^F),$$

where

- $c^L$ is the cost of the leader action,
- $L^X(s_i)$ are the losses of the herd in state $s_i$ for either the farmer if $X = F$ or the leader if $X = L$,
- $c^F(s_i)q^L(a^L, a_i^F)$ is the amount of the action cost of a herd in state $s_i$ payed by the leader,
- $c^F(s_i)q^F(a^L, a_i^F)$ is the amount of the action cost of a $s_i$ herd payed by the farmer.

Where $q^L(a^L, a_i^F)$ and $q^F(a^L, a_i^F)$ are defined by:

- $q^L(a^L, a_i^F) = perc$ if $a^L = 1$ and $a_i^F = 1$,
- $q^L(a^L, a_i^F) = 0$ else;
- $q^F(a^L, a_i^F) = 1$ if $a^L = 0$ and $a_i^F = 1$,
- $q^F(a^L, a_i^F) = 1 - perc$ if $a^L = 1$ and $a_i^F = 1$,
- $q^F(a^L, a_i^F) = 0$ else.

It can be easily checked that they are also substitutable (note that expectation $E_{s'_i}[\cdot]$ is taken with respect to a substitutable transition function).

### 5.1.2 The experiments

The comparison with the exact model was run in Matlab for $H = 10$ with $n = 12$, $K \in \{3, 4, 6\}$ and $n = 15$, $K \in \{3, 5\}$. For $n = 12$ and $K = 3$, we generated 1000 sets of parameters values with varying values of costs and losses. For most sets, the leader policy consisted in always doing nothing. For the comparison with the exact model (using the substitutability assumption), we selected 17 scenarios. For comparison with the case $n = 15$, due to the computing time to solve the exact model, we explored only the three sets of parameter values described in Table 2.

We evaluated the impact of the number of classes, $K$, on various model results at the leader level. As far as the followers policies were concerned, we simply checked whether or not they were deterministic.

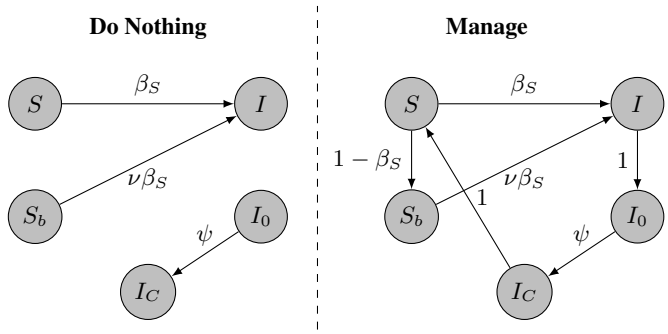


**Figure 1.** Transitions between follower's states.

To compare the leader optimal policies computed with the different $K$ (noted $\delta_K$) with the global optimal policy $\delta^*$, we computed 3 indicators :

- $Last\_Diff$: The absolute difference between times of the last leader management, between the aggregate and global optimal policies.
- $\#Diff$: Number of time steps for which both policies were not equal.
- $Max\_Gap$: Maximum (over all time steps) proportion of states for which both policies differ.

Different policies may lead to similar distributions over states at each time step. So, in order to further evaluate our approximation, we compared the distributions over states at time step $H$ obtained when applying $\delta^*$ or $\delta_K$. To compute these distributions, we have to choose an initial distribution (time step 1). We considered two initial distributions : (i) $\Gamma_U$ uniform over all states, and (ii) $\Gamma_E$ uniform only on "highly infected" states (states where around $40\%$ of followers are in state $S$ or $S_b$ and around $40\%$ are in state $I_C$). To compare the distributions resulting from $\delta_K$ and $\delta^*$, we computed the Battacharyya distances [1] between them, denoted $DB\_U$ and $DB\_E$ for $\Gamma_U$ and $\Gamma_E$ respectively.

As different policies may be equivalent in terms of values, we evaluated the impact of the approximation on the objective function. We computed a distance derived from the root mean square error (RMSE):

$$RMSE\_x = \sqrt{\sum_{c \in \Sigma^L} ((V^K(c) - V^*(c))^2 \times \Gamma_x)}.$$

with $x = U$ or $x = E$ and $V^K(.)$ is the value function of policy $\delta^K$.

## 5.2 Experimental results

### 5.2.1 Empirical complexity reduction

The decrease in state space size using aggregation of course depends on $K$ and $N$ (Table 3). One should note that state aggregation does not partition the state space into uniform clusters (Fig. 2). We will see in the following section that this does not have a dramatic impact on the performance of the computed policies.

**Table 3.** Size of the leader states space according to the value of K ($|\Sigma^L|$ if $K = n$ or $|\overline{\Sigma^L}|$ else; $-$ if not applicable)

| $n$ | $K = n$ | $K = 3$ | 4 | 5 | 6 | 10 | 25 |
|---|---|---|---|---|---|---|---|
| 12 | 1,820 | 236 | 361 | - | 745 | - | - |
| 15 | 3,876 | 251 | - | 631 | - | - | - |
| 20 | 10,626 | - | 496 | 806 | - | 10 | - |
| 50 | 316,251 | - | - | 876 | - | 6,376 | 79,101 |
| 100 | 4,598,251 | - | - | 876 | - | 6,376 | 114,526 |

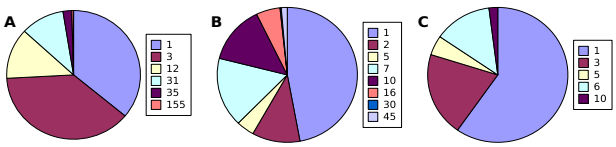


**Figure 2.** For $n = 12$, repartition of the number of global states per aggregate state for (A) K=3, (B) K=4 and (C) K=6.

### 5.2.2 Comparison results

Followers' optimal policies were found deterministic in all parameters sets tested. The leader policies varied when $K$ varied. When considering the time steps where the action 'manage' was retained by the leader in at least one state, we had two possible profiles for the leader policy: management only at one time step (for example for parameters set 824) or management in several time steps (for example the sets 2001 and 131). In the tested sets, these profiles were kept when changing $K$ and $n$. In the three tested configurations, the last management time step was the same for all $K$, both for $n = 12$ and $n = 15$ ($Last\_Diff = 0$). Still, the policies differed in some states when $K$ varied ($\#Diff > 0$). The proportion of states where the policies differed also varied ($Max\_Gap$ between $1.5\%$ and $30\%$ for the tested sets).

The impact on the final distributions and expected values varied. For parameters set 824, the impact was null for the distributions and very low for the values. It can be explained by the fact that the policies differed in only one time step. For other parameters sets, the final distributions varied with $K$, but the variations were low. The impact of the initial distribution ($\Gamma_U$ or $\Gamma_E$) on the variation of expected value with different $K$ was not consistent between different parameters' sets and number of followers (Fig. 3). Overall, even though this conclusion should be taken with caution, given the small number of configurations tested, it seems that the approximation becomes better when $K$ increases, which seems logical.

## 6 Concluding remarks and future work

In this article, we have proposed approximation methods to compute solutions to LF-MDP problems. Even though these experiments

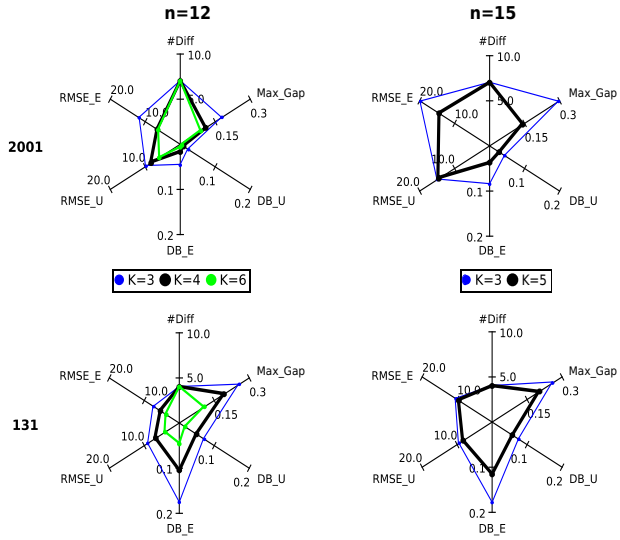


**Figure 3.** For parameter sets 2001 and 131 (table 2), comparison of the approximated results to the exact ones for various $K$ and $n$.

have not been reported in the paper, we observed that the suggested approximations permitted to solve approximately problems with up to 100 followers (and $K = 5$). The number of classes for state aggregation were observed in the three parameter sets to be correlated to the approximation quality on the case study whatever $n$, but this obviously has to be confirmed by further experiments.

Note that, in the proposed approach, the computed equilibria are equilibria in the reduced LF-MDP, not in the original one. However, we have a way to check *a posteriori* (apart in the case of state aggregation), whether all the returned equilibria of games are deterministic. If so, the solution is exact. If not, it is approximate. We leave for further research the question of approximability of LF-MDP equilibrium strategies, which is certainly worth considering and extends that of stochastic Nash equilibrium approximation.

One perspective of this work is to model the heterogeneity of followers, which can be done rather straightforwardly by "duplicating" followers state spaces and modeling followers types by different reward functions. Transitions are allowed between the duplicated state spaces if and only if followers can change their type over time.The new problem is still a LF-MDP, with more states, but potentially more "structure" as well. If followers keep the same "type" all along the problem, no transition is allowed between the duplicated state spaces.

As mentioned in the end of Section 4, the dependency on $n$ of time complexity could be suppressed if *simulation-based* approaches were used to approximate $\hat{T}$ (e.g. Bayesian RL [7]). Bayesian approaches have also been used in the framework of *Partially Observed MDP (POMDP)* [15]. Partial observability of followers states could be considered as well in LF-MDPs, leading to LF-POMDP models, mixing dec-POMDP with Bayesian games, in the line of [12], for example.

## Acknowledgments

This work was supported by the French Research Agency, program Investments for the future, project ANR-10-BINF-07 (MIHMES), by the European fund for the regional development FEDER and by the INRA Flagship program SMaCH. We also thank IJCAI and ECAI reviewers.

## REFERENCES


[1] A. Bhattacharyya, 'On a measure of divergence between two statistical populations defined by their probability distributions', *Bulletin of the Calcutta Mathematical Society*, **35**, 99–109, (1943).

[2] X. Cheng and X. Deng, 'Settling the complexity of 2-player nash equilibrium', in *Proceedings of FOCS*, (2006).

[3] N. Ferns, P. Panangaden, and D. Precup, 'Metrics for finite Markov decision processes', in *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, pp. 162–169. AUAI Press, (2004).

[4] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer, 1996.

[5] A. Gilpin and T. Sandholm, 'Lossless abstraction of imperfect information games', *Journal of the ACM*, **54**(5), (2007).

[6] R. Givan, T. Dean, and M. Greig, 'Equivalence notions and model minimization in Markov decision processes', *Artificial Intelligence*, **147**(1), 163–223, (2003).

[7] A. Guez, D. Silver, and P. Dayan, 'Efficient bayes-adaptive reinforcement learning using sample-based search', in *Proceedings of NIPS 2012*, (2012).

[8] J. Hawkin, R. Holte, and D. Szafron, 'Automated action abstraction of imperfect information extensive-form games', in *National Conference on Artificial Intelligence (AAAI)*, (2011).

[9] L. Li, T. J. Walsh, and M. L. Littman, 'Towards a unified theory of state abstraction for MDPs.', in *ISAIM*, (2006).

[10] R.B. Myerson, *Game Theory: Analysis of Conflict*, Harvard University Press, 1997.

[11] G. Nodelijk, 'Porcine reproductive and respiratory syndrome (prrs) with special reference to clinical aspects and diagnosis: A review', *Veterinary Quarterly*, **24**, 95–100, (2002).

[12] F. A. Oliehoek, M.T.J. Spaan, J.S. Dibangoye, and C. Amato, 'Heuristic search for identical payoff bayesian games', in *Proc. of AAMAS 2010*, pp. 1115–1122, (2010).

[13] E. L. Plambeck and S. A. Zenios, 'Performance-based incentives in a dynamic principal-agent model.', *Manufacturing & service operations management*, **2**, 240–263, (2000).

[14] M. Puterman, *Markov Decision Processes*, John Whiley and Son, 1994.

[15] S. Ross, J. Pineau, B. Chaib-Draa, and P. Kreitmann, 'A bayesian approach for learning and planning in partially observable Markov decision processes', *Journal of Machine Learning Research*, **12**, 1729–1770, (2011).

[16] R. Sabbadin and A.F. Viet, 'A tractable leader-follower MDP model for animal disease management', in *27th AAAI Conference on Artificial Intelligence*, (2013).

[17] T. Sandholm and S. Singh, 'Lossy stochastic game abstraction with bounds', in *EC '12 Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 880–897, (2012).

[18] L. S. Shapley, 'Stochastic games', *Proc. Nat. Academy of Science*, **39**, 1095–1100, (1953).

[19] J. Sorg and S. Singh, 'Transfer via soft homomorphisms', in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS*, (2009).

[20] K. Tharakunnel and S. Bhattacharyya, 'Leader-follower semi-Markov decision problems: theoretical framework and approximate solution', in *IEEE international conference on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 111–118, (2007).

[21] K. Tharakunnel and S. Bhattacharyya, 'Single-leader-multiple-follower games with boundedly rational agents', *Journal of Economic Dynamics and Control*, **33**, 1593–1603, (2009).