A Mirror Descent Approach for Mean Field Control applied to Demande-Side Management

Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard, Nadia Oudjane

A MIRROR DESCENT APPROACH FOR MEAN FIELD CONTROL APPLIED TO DEMANDE-SIDE MANAGEMENT

Bianca Marin Moreno
Inria*, EDF R&D†

Margaux Brégère
Sorbonne Université‡, EDF R&D†

Pierre Gaillard
Inria*

Nadia Oudjane
EDF R&D†

ABSTRACT

We consider a finite-horizon Mean Field Control problem for Markovian models. The objective function is composed of a sum of convex and Lipschitz functions taking their values on a space of state-action distributions. We introduce an iterative algorithm which we prove to be a Mirror Descent associated with a non-standard Bregman divergence, having a convergence rate of order $1/\sqrt{K}$. It requires the solution of a simple dynamic programming problem at each iteration. We compare this algorithm with learning methods for Mean Field Games after providing a reformulation of our control problem as a game problem. These theoretical contributions are illustrated with numerical examples applied to a demand-side management problem for power systems aimed at controlling the average power consumption profile of a population of flexible devices contributing to the power system balance.

1 Introduction

This paper attempts to solve finite-horizon mean field control (MFC) problems for non-stationary Markovian models where an infinite number of agents cooperate to optimize a common goal. For a finite set S, we define Δ_S to be the simplex of dimension $|S|$, the cardinal of S. At each time step $n \in [1, ..., N]$, an agent is in a state $x_n \in \mathcal{X}$, where $\mathcal{X}$ is a finite state space, and chooses randomly an action a_n in a finite set $\mathcal{A}$ according to a policy $\pi_n(\cdot|x_n) \in \Delta_{\mathcal{A}}$, then moving to a next state x_{n+1} according to a transition kernel $p(\cdot|x_n, a_n) \in \Delta_{\mathcal{X}}$ (this probability function could depend on time, but for now we consider the homogeneous case as restoring the time dependence is a straightforward procedure). We suppose that all the agents are homogeneous and follow the same policy sequence $\pi := (\pi_n)_{1\leq n\leq N} \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$. We denote by $\mu_n^\pi \in \Delta_{\mathcal{X}\times\mathcal{A}}$ the state-action distribution at time step n common to all agents when they follow the sequence of policies π. We seek to find π that minimizes a cost of the form

$$F(\mu^\pi) := \sum_{n=1}^{N} f_n(\mu_n^\pi) \tag{1}$$

where the cost functions $f_n : \Delta_{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}$ are convex and Lipschitz with respect to some norm $\|\cdot\|$.

Decision-making problems in mean field models are a popular way to formulate stochastic optimization problems in many applications, ranging from robotics Shiri et al. [2019], Elamvazhuthi and Berman [2019] to finance Achdou et al. [2014], Casgrain and Jaimungal [2018], energy management De Paola et al. [2019], Bušić and Meyn [2019], epidemic modelling Lee et al. [2021] and, more recently, machine learning E et al. [2018], Ruthotto et al. [2020], Fouque and Zhang [2020], Lin et al. [2021]. In this paper we are especially motivated by demand-side management (DSM) applications in power systems. How we use energy has never been a more important topic than it is today due to the challenges of energy transition and geopolitical shifts in supply. These and other challenges cause significant fluctuations in energy supply, impacting the balance of the energy grid and potentially causing power outages. DSM is a strategy to manage the portion of demand that is flexible by controlling, for example, thermostatically controlled loads (TCLs: flexible devices such as water heaters, air conditioning, refrigerators, etc). Its goal is to optimize users' consumption, thus saving energy and reducing costs. However, efficient exploitation of the large number of TCLs

*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
†EDF Lab, 7 bd Gaspard Monge, 91120 Palaiseau, France
‡Sorbonne Université LPSM, Paris, France

requires the development of new algorithmic tools. In this paper, we develop a new iterative mean field algorithm to control the average consumption of a population of water heaters in order to follow a given target consumption profile. In this case, an agent is a water heater and its state space consists of its operating state (whether it is on or off) and its average temperature, with the possible actions being to maintain or change their operating state (turn them on or off).

Finding a minimizer π^* of (1) is not straightforward because the function $\pi \mapsto F(\mu^\pi)$ is generally non-convex in π. In this paper, we introduce an algorithm whose iterative scheme is similar to a proximal point algorithm over a convex subset of all the state-action distribution sequences $\mu := (\mu_n)_{1\leq n\leq N} \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$; these are induced by some policy sequence π which we penalize. We prove that our penalization term is in fact a non-standard Bregman divergence on the set of state-action distribution sequences μ. Therefore the algorithm is an instance of the mirror descent algorithm (implying a convergence rate of order $1/\sqrt{K}$, with K being the number of iterations).

Contributions and outline In Section 3, we provide a convex reformulation of Problem (1). This allows us to propose an original algorithm in Section 4, which we show to be a mirror descent with a non-standard Bregman divergence. Its iterations can be easily computed by dynamic programming (DP) Bertsekas [2005], which makes it, to our knowledge, the first low complexity algorithm with a proof of convergence in discrete iterations for the type of control problem considered. In addition, we explore connections between learning methods for mean field games and optimization approaches for mean field control problems such as Problem (1).

Finally, in Sections 5 and 6, we propose an application to the problem of controlling the average consumption of a population of water heaters modeled by a Markovian model. In Section 5 we give an original modelling for the DSM problem, and in Section 6 we illustrate and compare our new algorithm with the mean field learning methods discussed throughout the article.

2 Related Work

Demand-side management Controlling the sum of the consumption of a large number of TCLs started being investigated around 1980 by Ihara and Schweppe [1981], Malhame and Chong [1985], Mortensen and Haggerty [1988] establishing the first physically based modeling for a TCL population. In the works of Kizilkale and Malhame [2013, 2014], the difficulty due to the large number of devices is circumvented by a mean field approximation. In Le Floch et al. [2018], a mean field assumption is also considered to control the charging of a large fleet of electrical vehicles, leading to optimal control of partial differential equation problems. In the particular setting of stochastic control, Séguret et al. [2020] proposed an iterative stochastic algorithm providing a decentralized solution where each agent computes locally its own feedback control.

For water heater control, Cammardella et al. [2019] consider a quadratic objective and a Kullback-Leibler penalty allowing a Lagrangian approach that learns both the control and the probability transition kernel. However, their approach does not allow for situations where part of the state is uncontrolled, so that uncertainties induced by consumer behaviors (such as water withdrawals) must be neglected and modeled as deterministic. More recently, Cammardella et al. [2021] propose to take into account the uncontrolled stochastic environment in the Kullback-Leibler quadratic control framework by adding constraints on the probability transition kernel. However, to handle these new constraints, they need to add new dual variables, which leads to a high-dimensional dual problem, thus significantly increasing the complexity of their algorithm.

On the other hand, load control can also be done indirectly, with electricity consumers being encouraged to alter their consumption when necessary, typically by reducing it at peak hours and increasing it at off-peak hours. For example, Brégère et al. [2019] build dynamic pricing systems using multi-armed bandit methods.

Mean field learning Mean field games (MFG) have been introduced by Lasry and Lions [2007] and Huang et al. [2006] to tackle the issue of games with a large number of symmetric and anonymous players, by passing to the limit of an infinite number of players interacting through the population distribution. Although MFG focuses on finding Nash equilibria (NE), social optima on cooperative setting have also been studied under the term of mean field control (MFC) [Bensoussan et al., 2013]. Although MFG and MFC problems look similar, they in general have different solutions [Carmona et al., 2013]. Nevertheless, Lasry and Lions [2007] have pointed out that in some cases a mean field NE is also the solution to a different optimal control problem, see for example Alasseur et al. [2020] that applies this principle to energy production.

Lately, there has been interest in combining reinforcement learning [Sutton and Barto, 2018] with MFG and MFC. Some work has focused on studying stochastic methods based on neural network approximations Carmona and Laurière [2021a, 2022, 2021b]. These methods are based on the knowledge of the model. Carmona et al. [2019], Guo et al. [2019], Angiuli et al. [2021] focus on learning solutions without full knowledge of the model. Iterative learning methods

such as fictitious play and online mirror descent have been adapted to the MFG scenario in Perrin et al. [2020] and Pérolat et al. [2022]. Geist et al. [2022] show an equivalence between Frank Wolfe's classical optimization algorithm [Frank and Wolfe, 1956] and the fictitious play for potential structured games. We compare some of these algorithms with our new MFC algorithm on the DSM problem.

Learning on Markovian models The finite horizon Markovian model we consider can also be generalized as a loop-free stochastic shortest path (SSP) by considering a larger state space that includes time steps. There is significant work in the literature regarding algorithms for the online SSP problem, in particular [Even-Dar et al., 2009, Rosenberg and Mansour, 2019, Dick et al., 2014, Zimin and Neu, 2013]. Our work is also closely related to the work in [Neu et al., 2017, Geist et al., 2019] that propose general analysis on regularized Markov decision processes. In particular, Neu et al. [2017] is the first to explore the non-standard conditional entropy of the state-action distribution as a regularization, showing that we often arrive at approximate instances of mirror descent. In comparison, an important difference between these works and ours is that we additionally have to deal with the mean field framework.

3 Finite horizon mean field problems

We start by describing the MFG and MFC problems, specifying the difference between them. Then, we propose a reformulation of the MFC Problem (1) in a convex setting.

3.1 Mean field games and mean field control

MFGs were introduced to tackle decision-making problems involving an infinite number of homogeneous players interacting through the population distribution. It can be solved by focusing on the optimal policy of a representative player in response to the behavior of the population. In contrast, the search for a social optimum in a cooperative setting is known as the MFC problem.

In the MFG framework, at every time step $1 \leq n \leq N$, the player receives a reward given by a function $r_n : \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}$. The third argument corresponds to the current population's state-action distribution $\mu_n \in \Delta_{\mathcal{X}\times\mathcal{A}}$. The goal of a representative player is to find a sequence of policies π that maximises the expected sum of rewards when the population distributions sequence is given by $\mu := (\mu_n)_{1\leq n\leq N}$ and the initial state-action pair is sampled from a fixed distribution μ_0,

$$J_{\mu_0}(\pi, \mu) := \mathbb{E}_\pi \left[\sum_{n=1}^{N} r_n(x_n, a_n, \mu_n) \right], \tag{2}$$

where $\mathbb{E}_\pi$ is the expectation on the trajectories $(x_n, a_n)_{1\leq n\leq N}$ induced by the policy sequence π and the transition kernel p, and initialized by $(x_0, a_0) \sim \mu_0$.

Definition 3.1 (Distribution induced by a policy π)**.** Given an initial distribution μ_0 fixed, the state-action distributions sequence induced by the policy sequence $\pi = (\pi_n)_{1\leq n\leq N}$ is denoted $\mu^\pi = (\mu_n^\pi)_{1\leq n\leq N}$ and is defined recursively by

$$\begin{aligned}
\mu_0^\pi(x', a') &= \mu_0(x', a') \\
\mu_{n+1}^\pi(x', a') &= \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_n^\pi(x, a) p(x'|x, a) \pi_{n+1}(a'|x').
\end{aligned}$$

A MFG Nash equilibrium (NE) is defined as a pair $(\hat{\pi}, \hat{\mu})$ such that $\hat{\pi} := \arg\max_\pi J_{\mu_0}(\pi, \hat{\mu})$ and $\hat{\mu} := \mu^{\hat{\pi}}$ is the distribution induced by the policy $\hat{\pi}$. In contrast with the MFG problem, the MFC problem is an optimisation problem where there is no competition among players, but cooperation. The MFC problem seeks to find the optimal behavior of a population so as to maximize a reward averaged over the whole population. In mathematical terms, it seeks to solve $\pi^* := \arg\max_\pi J_{\mu_0}(\pi, \mu^\pi)$. In general, MFG and MFC problems have different solutions.

3.2 Reformulation into a convex framework

Consider the set of state-action distributions sequences initialized at $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$ and satisfying a specific constrained evolution given by

$$\mathcal{M}_{\mu_0} := \left\{ \mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N \Big| \sum_{a'\in\mathcal{A}} \mu_{n+1}(x', a') = \sum_{x\in\mathcal{X}, a\in\mathcal{A}} p(x'|x, a)\mu_n(x, a) \,, \forall x' \in \mathcal{X},\ \forall n \in [0, ..., N] \right\}. \tag{3}$$

The set $\mathcal{M}_{\mu_0}$ describes the sequences of state-action distribution respecting the dynamics of the Markov model. Furthermore, this set is convex [Cammardella et al., 2021].

Consider the sequence of convex functions $(f_n)_{1\leq n\leq N}$, with $f_n : \Delta_{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}$, and recall that $F : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ is such that $F(\mu) := \sum_{n=1}^N f_n(\mu_n)$. In our settings we consider the finite-horizon MFC problem given by $\min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} F(\mu^\pi)$. Given the intrinsic dependence of the sequence of distributions μ^π and the sequence of policies π given by Definition 3.1, this minimisation problem is not necessarily convex on π, and in general the gradient $\nabla_\pi F(\mu^\pi)$ cannot be estimated - for more details on how one could still work with the gradient of $\nabla_\pi F(\mu^\pi)$, see Zhang et al. [2020]. We therefore seek an efficient reformulation of our optimization framework.

Proposition 3.2. *Let $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$. The application $\pi \mapsto \mu^\pi$ is a bijection from $(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ to $\mathcal{M}_{\mu_0}$.*

The proof of Proposition 3.2 is reported to Appendix A. The idea is that one can retrieve the policy sequence π inducing the state-action distribution sequence μ by taking $\pi_n(a|x) = \frac{\mu_n(x,a)}{\rho_n(x)}$, where $\rho_n(x) := \sum_{a\in\mathcal{A}} \mu_n(x,a)$. This proposition provides us with an efficient reformulation of the original control problem , i.e.

$$\min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} F(\mu^\pi) \equiv \min_{\mu\in\mathcal{M}_{\mu_0}} F(\mu). \tag{4}$$

4 Algorithmic approaches

We first introduce the new optimization algorithm for the MFC Problem (4) and provide a convergence result. Then, we formulate the MFC problem as a game and propose other algorithms from the MFG literature to solve it.

4.1 A mirror descent approach for mean field control

Here we propose a new iterative algorithm for MFC problems such as the one in Equation (4). It requires the solution of a simple dynamic programming problem at each iteration. The analysis shows a convergence rate of order $1/\sqrt{K}$ where K is the number of iterations, and is given by the mirror descent convergence [Beck and Teboulle, 2003].

We consider the following iterative scheme, where k represents an iteration. At iteration $k+1$, we want to find μ^π minimizing a linearization of F around μ^k, the distribution sequence found at the previous iteration, but penalizing the distance between π generating μ^π and π^k generating μ^k (recall that the uniqueness of a π generating μ^π is given by Proposition 3.2):

$$\mu^{k+1} \in \underset{\mu^\pi\in\mathcal{M}_{\mu_0}}{\arg\min} \left\{ \langle \nabla F(\mu^k), \mu^\pi\rangle + \frac{1}{\tau_k}\sum_{n=1}^N \mathbb{E}_{(x,a)\sim\mu_n^\pi(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi_n^k(a|x)}\right)\right] \right\} \tag{5}$$

where $\langle \nabla F(\mu^k), \mu^\pi\rangle := \sum_{n=1}^N \langle \nabla f_n(\mu_n^k), \mu_n^\pi\rangle$. The term $\langle \nabla F(\mu^k), \mu^\pi\rangle$ can be interpreted in a MFG setting as the expected sum of rewards of a representative agent when the population behaves like μ^k, the agent follows the strategy π, and the reward at each time step n derives from the potential f_n, i.e. for all $x_n \in \mathcal{X}$, $a_n \in \mathcal{A}$, $\mu_n \in \Delta_{\mathcal{X}\times\mathcal{A}}$,

$$r_n(x_n, a_n, \mu_n) := -\nabla f_n(\mu_n)(x_n, a_n). \tag{6}$$

Then, using the definition in Equation (2), we get

$$J_{\mu_0}(\pi, \mu^k) = \sum_{n=1}^N \left\langle r_n(\cdot,\cdot,\mu_n^k), \mu_n^\pi\right\rangle = -\left\langle \nabla F(\mu^k), \mu^\pi\right\rangle,$$

which allows us to re-write an iteration as

$$\mu^{k+1} \in \underset{\mu\in\mathcal{M}_{\mu_0}}{\arg\max} \left\{ J_{\mu_0}(\pi, \mu^k) - \frac{1}{\tau_k}\sum_{n=1}^N \mathbb{E}_{(x,a)\sim\mu_n(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi_n^k(a|x)}\right)\right] \right\}. \tag{7}$$

The problem in Equation (7) can be solved through dynamic programming Bertsekas [2005]. We build a Bellman recursion as enunciated by Theorem 4.1 below.

Theorem 4.1. *Let $k \geq 0$. The solution of Problem* (7) *is $\mu^{k+1} = \mu^{\pi^{k+1}}$ (as in Definition 3.1), where for all $1 \leq n \leq N$, and $(x, a) \in \mathcal{X} \times \mathcal{A}$,*

$$\pi_n^{k+1}(a|x) := \frac{\pi_n^k(a|x)\exp\left(\tau_k \tilde{Q}_n^k(x,a)\right)}{\sum_{a'\in\mathcal{A}} \pi_n^k(a'|x)\exp\left(\tau_k \tilde{Q}_n^k(x,a')\right)}$$

where $\tilde{Q}$ is a regularized Q-function satisfying the following recursion

$$\begin{cases} \tilde{Q}_N^k(x,a) = r_N(x,a,\mu_N^k) \\ \tilde{Q}_n^k(x,a) = \max\limits_{\pi_{n+1}\in(\Delta_{\mathcal{A}})^{\mathcal{X}}} \Bigg\{ r_n(x,a,\mu_n^k) + \sum\limits_{x'} p(x'|x,a) \\ \qquad\qquad \sum\limits_{a'} \pi_{n+1}(a'|x') \Big[-\frac{1}{\tau_k} \log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) + \tilde{Q}_{n+1}^k(x',a') \Big] \Bigg\}, \quad \forall 1 \le n \le N. \end{cases} \tag{8}$$

Proof. See Appendix B.1. □

Notice that the value $\pi_{n+1} \in (\Delta_{\mathcal{A}})^{\mathcal{X}}$ maximizing the equation to find $\tilde{Q}_n^k$ in the Recursion (8) is given by π_{n+1}^{k+1}. The algorithm may thus be summarized as follows.

Algorithm 1 Mirror descent approach to MFC

Input: number of iterations K, initial sequence of policies $\pi^0 \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ such that $\mu^0 := \mu^{\pi_0}$, initial state-action distribution μ_0 (always fixed), sequence of non-negative learning rates $(\tau_k)_{k\le K}$.
for $k = 0, ..., K-1$ **do**
  $\mu^k = \mu^{\pi^k}$ as in Definition 3.1.
  $\tilde{Q}_N^k(x,a) = r_N(x,a,\mu_N^k)$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$.
  for $n = N, ..., 1$ **do**
    $\forall (x,a) \in \mathcal{X}\times\mathcal{A}$:
    $\pi_n^{k+1}(a|x) = \frac{\pi_n^k(a|x)\exp\left(\tau_k \tilde{Q}_n^k(x,a)\right)}{\sum_{a'} \pi_n^k(a'|x)\exp\left(\tau_k \tilde{Q}_n^k(x,a')\right)}$.
    $\tilde{Q}_{n-1}^k(x,a)$ using the recursion in Equation (8).
  end for
end for
return π^K

4.2 Convergence properties of the algorithm

We now present a result on the convergence rate of Algorithm 1. For ease of notation, for any probability measure $\eta \in \Delta_E$, whatever the (finite) space E, we introduce the neg-entropy function

$$\phi(\eta) := \sum_{x\in E} \eta(x)\log\eta(x).$$

Proposition 4.2. *Let $\mu, \mu' \in \mathcal{M}_{\mu_0}$ with marginals given by $\rho, \rho' \in (\Delta_{\mathcal{X}})^N$, induced by the policy sequences π, π' respectively. The divergence $\Gamma : \mathcal{M}_{\mu_0}\times\mathcal{M}_{\mu_0} \to \mathbb{R}$ given by*

$$\Gamma(\mu,\mu') := \sum_{n=1}^N \mathbb{E}_{(x,a)\sim\mu_n^{\pi}(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi_n'(a|x)}\right)\right]$$

is a Bregman divergence induced by the function

$$\psi(\mu) := \sum_{n=1}^N \phi(\mu_n) - \sum_{n=1}^N \phi(\rho_n).$$

On the proof of Proposition 4.2 in Appendix B.2 we also see that Γ can be written as the difference between the Kullback-Leibler divergence on the state-action distributions and the Kullback-Leibler divergence on the marginal state distribution, and also as the Kullback-Leibler divergence on the joint distributions. Proposition 4.2 shows that the penalization term at each iteration scheme given by Equation (5) is a Bregman divergence. Thus, the iteration considered is in fact an iteration of the mirror descent algorithm (see Nemirovski and Yudin [1983]). This allows us to state the following result

Theorem 4.3. *Consider a sequence of functions $(f_n)_{1\leq n\leq N}$ with $f_n : \Delta_{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}$, convex and Lipschitz with respect to a norm $\|\cdot\|$, with l_n being the Lipschitz constant. Define $F : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ as $F(\mu) := \sum_{n=1}^N f_n(\mu_n)$. Hence, F is also convex and Lipschitz with constant $L := \left(\sum_{n=1}^N l_n^2\right)^{1/2}$. Consider $\mathcal{M}_{\mu_0}$ the convex set on Equation* (3)*. Applying K iterations of Algorithm 1, such that μ^* is a minimizer of F in $\mathcal{M}_{\mu_0}$, and with, for each $1 \leq k \leq K$,*

$$\tau_k := \frac{\sqrt{2\Gamma(\mu^*, \mu^0)}}{L}\frac{1}{\sqrt{k}},$$

gives the following convergence rate

$$\min_{0\leq s\leq K} F(\mu^s) - F(\mu^*) \leq L\frac{\sqrt{2\Gamma(\mu^*, \mu^0)}}{\sqrt{K}}.$$

Proof. See Appendix B.3. □

4.3 Potential games

We provide an equivalence between the control Problem (4) and a game problem in order to be able to compare Algorithm 1 with learning algorithms for MFG in the literature. For that, let us define a game with the same transition probability p, and with reward defined as in Equation (6). We call this type of game a potential game.

Proposition 4.4. *The strategy π^* is a minimizer of Problem* (4) *if and only if, (μ^{π^*}, π^*) is a NE of the MFG defined with reward as in Equation* (6)*. Furthermore, this game is monotone (and strictly monotone if f_n is strictly convex for all $1 \leq n \leq N$. See Definition A.1 in Appendix A).*

The proof of Proposition 4.4 is in Appendix A and is similar to the results introduced by Geist et al. [2022]. It connects the optimality conditions of Problem (4) and a NE, and shows that convexity and monotonicity are equivalent. If the optimization problem is (strictly) convex, the (unique) existence of an optimizer implies the (unique) existence of a NE. Thus, the notion of monotonicity when the reward depends on the state-action distribution provides the (unique) existence of a NE in the case of a potential game. These results are what allows us to use algorithms of learning on MFGs to solve MFC problems and to compare such game algorithms with optimisation approaches. Here we list some examples of learning in iterative MFG algorithms that are compared to the Algorithm 1 in Section 6.

Fictitious play vs. Frank-Wolfe The algorithm of fictitious play (FP) adapted to the MFG scenario (see Algorithm 2 in Appendix C) introduced by Perrin et al. [2020] is shown to be analogous to the classical optimisation algorithm of Frank-Wolfe [Frank and Wolfe, 1956], described by Algorithm 3 in Appendix C, in Geist et al. [2022] when applied to a potential game. With the additional hypothesis that f_n be two times differentiable with Lipschitz gradient for all $1 \leq n \leq N$, this provides the mean field adaption of fictitious play with a convergence rate of $1/K$, the same as for Frank-Wolfe [Bubeck et al., 2015], where K is the number of iterations.

Online mirror descent for MFG Online mirror descent for MFG (OMD) is an iterative algorithm for computing the NE of a game first introduced by Pérolat et al. [2022], inspired by the online mirror descent regret minimization algorithm [Shalev-Shwartz, 2012]. Algorithm 4 in Appendix C describes OMD for MFG. In Pérolat et al. [2022] the authors provide a proof of convergence for the case of continuous iterations. There is a similarity between OMD for MFG and Algorithm 1. It can be interpreted that to predict the $k+1$-th policy at time step n, OMD for MFG uses the state-action value function induced by the previous iteration policy π^k, whereas Algorithm 1 uses the regularized state-action value function induced by the current iteration policy π^{k+1} computed for times $n+1$ through N. A precise analysis of the difference between both approaches is left to further work.

5 Application: demand-side management

We corroborate our theoretical claims with a numerical simulation applied to the problem of controlling the average power consumption profile of a population of water heaters.

5.1 Randomized controlled dynamics for one water heater

Let us consider a time window $[t_0, t_0+T]$, and a discretisation of the time such that $t_n = t_0 + n\delta_t$ for $n = 1, ..., N$, and $\delta_t = T/N$ is the time step. At each time step t_n (which for convenience will be denoted n), the state of a water heater is described by a variable $x_n = (m_n, \theta_n) \in \mathcal{X} := \{0, 1\} \times \Theta$, where m_n indicates the operating state of the heater

(ON if 1, OFF if 0), and θ_n represents the average temperature of the water in the tank. For the sake of simplification we consider only temperatures inside a finite set Θ.

The nominal dynamics [Bušić and Meyn, 2016] follow a cyclic ON/OFF decision rule with a deadband to ensure that the temperature is between a lower limit $T_{\min}$ and an upper limit $T_{\max}$. Thus, if the water heater is turned on, it heats water with the maximum power until its temperature exceeds $T_{\max}$. Then, the heater turns off and the water temperature decreases until it reaches $T_{\min}$, where the heater turns on again and a new cycle begins. The nominal dynamics at a discretized time is illustrated in Figure 1. The temperature at each time step is calculated by approximating an ordinary differential equation (ODE) depending on the current operating state of the heater and the hot water drawn at each time step (see Appendix D.1). We assume that the event of a water withdrawn is random and independent at each time step with a known probability distribution.

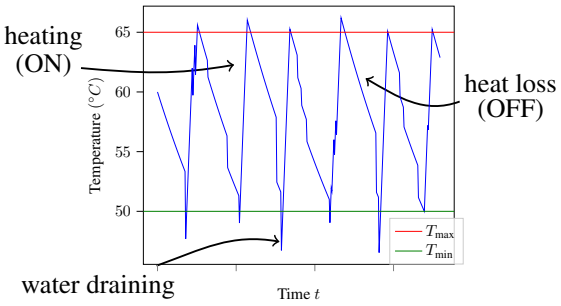


Figure 1: Temperature evolution of a water heater following the nominal dynamics.

In order to have a controllable model, we fit the nominal dynamics of a water heater to a Markov decision process. The finite state space is given by $\mathcal{X}$, and we consider an action space given by $\mathcal{A} := \{0, 1\}$. At time step n, choosing action 1 means turning the heater on except when $\theta_n \geq T_{\max}$. Conversely, choosing action 0 means turning the heater off except when $\theta_n \leq T_{\min}$. The nominal dynamics deterministically chooses action 0 if the heater is off and 1 if it is on. Unlike the nominal dynamics, we want to consider stochastic strategies for choosing actions. If the heater is in state $x_n = (m_n, \theta_n)$, the next temperature θ_{n+1} is computed by Equation (22), the action a_n is sampled with probability $\pi_n(\cdot|x_n)$, and the next operating state is given by

$$m_{n+1} = a_n \mathbb{1}_{\theta_{n+1} \in [T_{\min}, T_{\max}]} + \mathbb{1}_{\theta_{n+1} < T_{\min}}. \tag{9}$$

Thus if $\theta_{n+1} \in [T_{\min}, T_{\max}]$, the action $a_n \sim \pi_n(\cdot|x_n)$ defines the next operating state of the heater. For more details on the dynamics of a water heater see Appendix D.1.

5.2 Optimisation problem

Consider a population of M water heaters indexed by i and described at time step n by $X_n^i = (m_n^i, \theta_n^i)$ following the randomized dynamics described in Subsection 5.1. We suppose all water heaters to be homogeneous, i.e. they have the same dynamics, and follow the same policy π. Let $\bar{m}_n := \frac{1}{M} \sum_{i=1}^{M} m_n^i$ denote the average consumption. We assume for simplicity that the maximum power of each water heater is $p_{\max} = 1$ so that the average consumption is equal to the proportion of heaters at state ON. Note that $\bar{m}_n$ depends on the policy π that the water heaters follow, thus we can denote it as $\bar{m}_n(\pi)$. Let $\gamma = (\gamma_n)_{1 \leq n \leq N} \in [0, 1]^N$ be our target consumption profile (for example, the energy production at each time step). Our goal is to solve the problem

$$\min_{\pi \in (\Delta_{\mathcal{A}}^{\mathcal{X}})^N} \mathbb{E}\left[\sum_{n=1}^{N} (\bar{m}_n(\pi) - \gamma_n)^2\right], \tag{10}$$

where we have chosen to work with a quadratic loss.

Recall that $\mu := (\mu_n)_{n \in [0,\dots,N]}$ and μ_n is the state-action distribution of the entire population of heaters at time n. For a function $\varphi : \mathcal{X} \to \mathbb{R}$, we define $\mu_n(\varphi) := \sum_x \varphi(x) \mu_n(x, a)$ for all $1 \leq n \leq N$. We are particularly interested in a function φ such that $\mu_n(\varphi)$ gives us the average consumption of our water heater's population. Thus, we consider from now on

$$\begin{aligned} \varphi : \quad & \mathcal{X} \to \mathbb{R} \\ & (m, \theta) \mapsto m. \end{aligned} \tag{11}$$

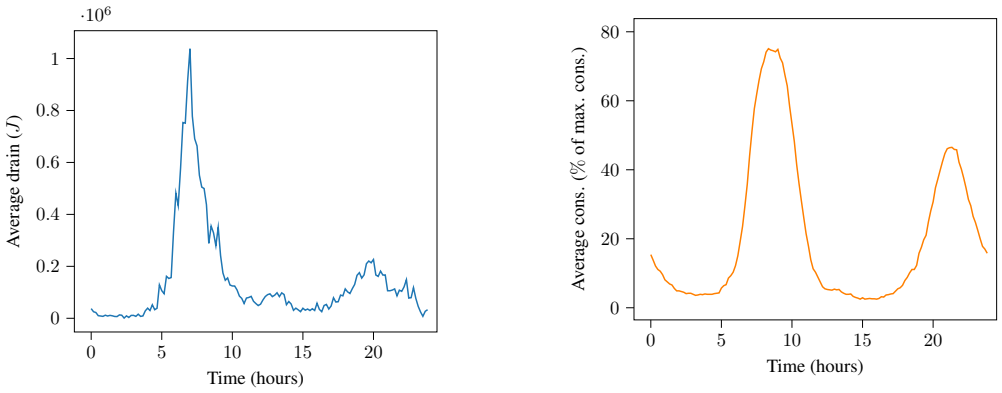
For such a function φ, when $M \to \infty$, the mean field approximation [Jabin and Wang, 2017] of Problem (10) is given by $\min_\pi \sum_{n=1}^N f_n(\mu_n^\pi)$, where $f_n(\mu_n^\pi) := (\mu_n^\pi(\varphi) - \gamma_n)^2$. In Proposition 3.2 we proved a bijection between the sets $\mathcal{M}_{\mu_0}$ and $(\Delta_\mathcal{A})^{\mathcal{X}\times N}$. Let $F : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ such that $\mu \mapsto F(\mu) := \sum_{n=1}^N f_n(\mu_n)$. Our main optimisation problem is then given by Equation (4).

In Appendix D.2 we show that we can apply Theorem 4.3 on the convergence of Algorithm 1 in this specific framework. In Subsection 4.3 we showed that for a control problem such as in Equation (4), we can define a potential game by taking a reward as in Equation (6), where here $\nabla f_n(\mu_n)(x,a) = 2(\mu_n(\varphi) - \gamma_n)\varphi(x)$. Hence, we can reframe the MFC Problem of controlling the average consumption of a population of water heaters as a MFG and apply the FP and OMD algorithms presented in Subsection 4.3.

6 Experiments

6.1 Simulating the nominal dynamics

To simulate the nominal dynamics, we use the nominal model presented in Appendix D.1 and data from the SMACH (*Simulation Multi-Agents des Comportements Humains*) platform [Albouys et al., 2019] to approximate the probability of having a water withdrawal for each time step. In addition, we take a time frequency $\delta_t = 10$ minutes, and a temperature deadband with $T_{\text{min}} = 50°C$ and $T_{\text{max}} = 65°C$. For more details on how the simulations are performed, see Appendix D.3. Figure 2 shows the simulation of the average drain and power consumption of 10^4 water heaters following the nominal dynamics over the period of one week day respectively. The states (operating state and temperature) are randomly initialized for each water heater. During the hours of the day with a peak of hot water withdrawal we also have a peak on the energy consumption, as all water heaters are turned on to compensate for the temperature loss.



(a) Average drain in Joules

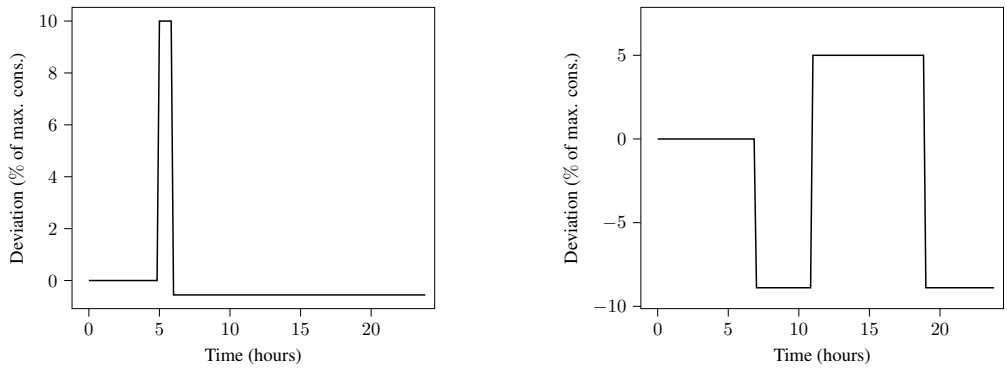
(b) Average consumption

Figure 2: Average drain and power consumption for a simulation of 10^4 water heaters over a period of one day.

The target signal $\gamma = (\gamma_n)_{1\leq n\leq N}$ is built as a sum of a baseline $b = (b_n)_{n\leq N}$ and a deviation signal $\lambda = (\lambda_n)_n$, $\gamma = \lambda + b(w)$, where $b(w)$ is the nominal dynamics obtained by simulating the water heaters (as in Figure 2), and w represents a random initialization of their states. If the deviation is zero, the average consumption is equal to the baseline. The deviation signal should have zero energy on the time considered for the simulations, i.e. $\sum_{n=0}^N \lambda_n = 0$, in order to ensure a stationary process. We consider the two deviation signals illustrated in Figure 3: an one-hour deviation between 5 and 6 in the morning, as well as an eight-hour gap where we increase consumption during off-peak hours and decrease it during peak hours.

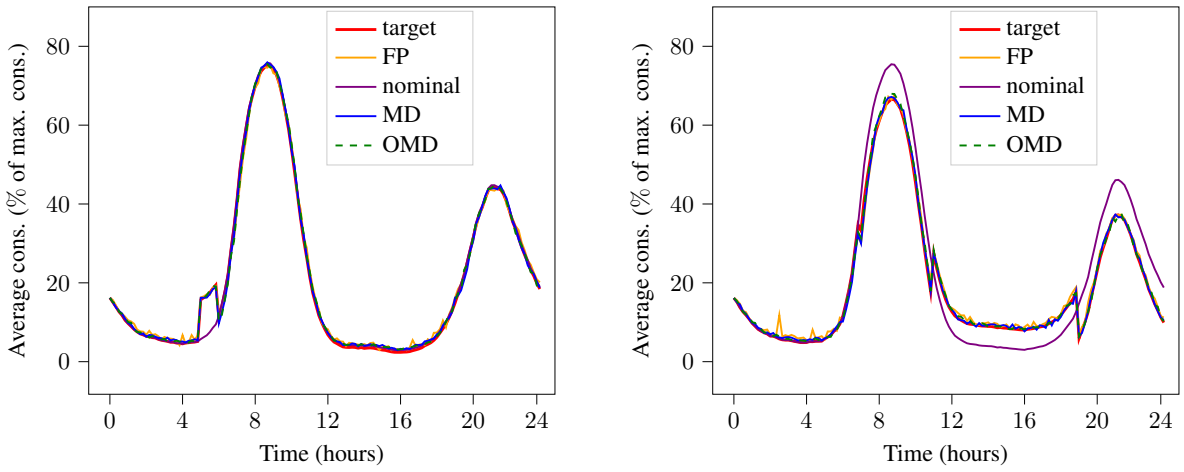
6.2 Results

For the sake of brevity, we refer to Algorithm 1 as MD (Mirror Descent). For a population of water heaters following the randomized dynamics we compare the optimal policy sequence obtained after 100 iterations of the three algorithms considered in the article: MD, OMD for MFG, and Fictitious Play for MFG (FP). At each iteration, we compute a policy sequence of size 144 (number of time steps equal to one day divided into 10 minute intervals). The heater's state space $\mathcal{X}$ is of size $2 * 41$ (two ON/OFF operating states times 41 possible temperatures - integers from the environment temperature $T_{\text{amb}} = 25$ to $T_{\text{max}} = 65$), and its action space $\mathcal{A}$ is of size 2. We simulate each policy on 10^4 water heaters



(a) One hour step. (b) Eight hours step.

Figure 3: Deviation signals $(\lambda_n)_{n \leq N}$



(a) Target with one hour step dev. (b) Target with eight hours step dev.

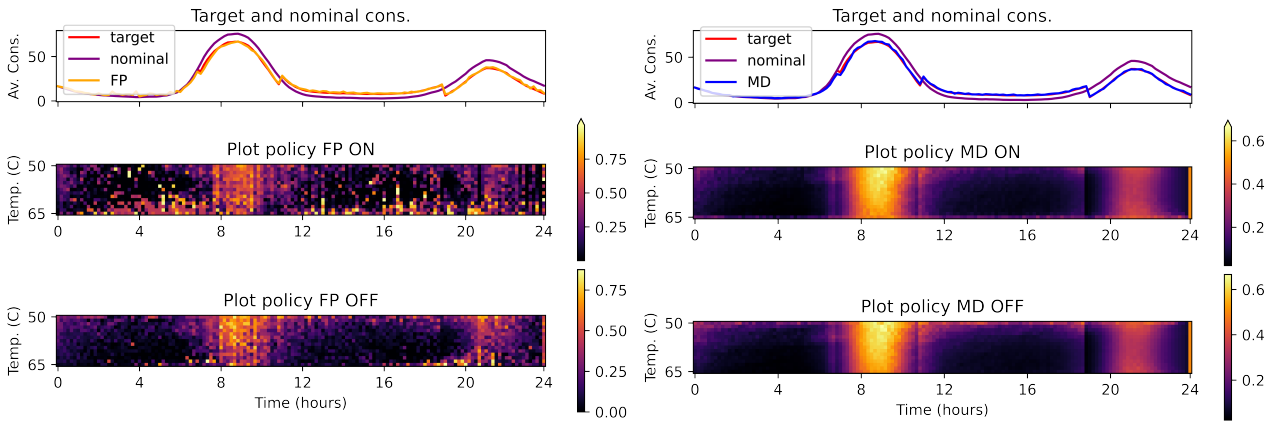
Figure 4: Simulation of the power consumption of 10^4 water heaters comparing the performance of policies obtained with Algorithm 1 (MD) and algorithms from the literature (FP and OMD), for targets constructed with the deviations of one hour [left] and eight hours [right]. We compare this with the nominal consumption (without deviation).

and analyze the average consumption curve. The water heater's initial state distribution is equal to the initial distribution of the nominal consumption. The distribution of actions is initialized uniformly.

In Figure 4, we compare the average consumption curves obtained. The consumption simulated by the best policies for all three algorithms appears to track the target better than the nominal consumption. As for the analysis in number of iterations, Figure 6 shows the logarithm of the objective function per iteration. The slope of the best approximation lines give us an empirical measure of convergence: we obtain that FP seems to converge empirically with a rate of $\sim 1/K^{5/2}$, while MD and OMD seem to converge empirically with a rate of $\sim 1/K^2$. The empirical convergence rates are better than the theoretical limits, $1/K$ for FP and $1/\sqrt{K}$ for MD.

To visualize the policies obtained with FP and MD we plot in Figure 5, at each time step [x axis], the probability of choosing the action 1 (ON) [colors] for all possible temperatures between $T_{\text{min}} = 50$ and $T_{\text{max}} = 65$ [y axis], when the current state is ON [up] or OFF [down]. The policies plots show how the probability of ignition is higher during times of the day when the target average consumption is higher. In addition, when the heater is in a low temperature state, the probability of ignition is also higher than when the heater is in a high temperature state. We can also see that MD returns a more regular policy than FP.

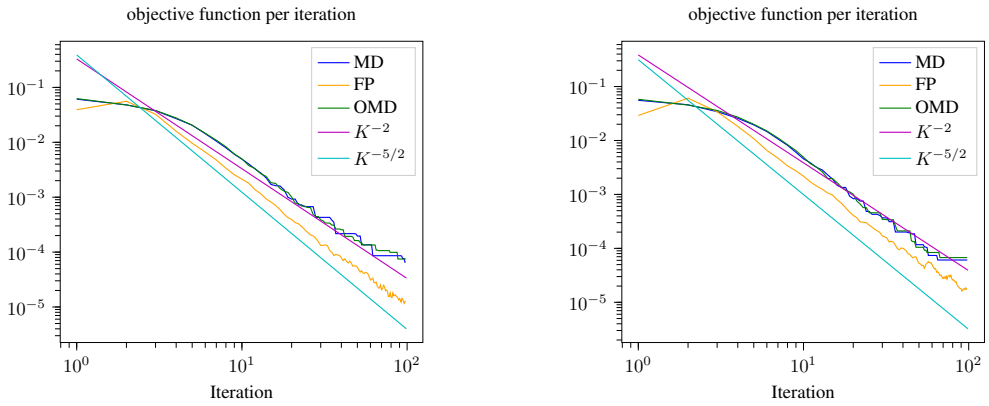
We also noticed that different initialization of MD lead to different policies: one can have several policy sequences inducing a fixed state distribution sequence ρ. This is further explored in Appendix E where we also focus on the number of ON/OFF switches that a device performs on average over the course of a day. In the case illustrated here, the average number of daily switches is 33, while the nominal dynamic (holding ON when ON and OFF when OFF within



(a) Policy Fictitious Play.

(b) Policy Mirror Descent.

Figure 5: [top] Target, average consumption obtained by the nominal policy and by the policy computed by FP (left) and MD (right). [middle] Probability of choosing the ON action when in the ON state. [bottom] Probability of choosing the ON action when in the OFF state. For all temperatures between $T_{\min} = 50$ and $T_{\max} = 65$ [y axis], over the course of a day with a time step of 10 minutes [x axis], for a target with a deviation step of eight hours.



(a) One hour step.

(b) Eight hours step.

Figure 6: Log-log plot of the objective function per iteration for each method when using a target with an one hour step [left] and eight hours step [right] deviations.

the temperature deadband $[T_{\min}, T_{\max}]$) averages only 3 switches per day. By initializing the algorithm with a policy that is a 0.1 deviation from the nominal policy, we find in Appendix E that the number of switches decreases to a daily average of 9.2 while still following the target curve. This is an important result, as a large number of switches can be detrimental to devices.

7 Further work:

The main objective for future work is to adapt the existing algorithms to real-time ones. So far, at each iteration, we optimize over all time steps up to the finite horizon. However, we would like to propose schemes where each iteration corresponds to a time step. In the context of power consumption control, this approach is interesting because the target curve we are trying to follow, predicted one day in advance, contains errors, and a real time control would allow a better system balance.

References

Hamid Shiri, Jihong Park, and Mehdi Bennis. Massive autonomous uav path planning: A neural network based mean-field game theoretic approach. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6,

2019. doi: 10.1109/GLOBECOM38437.2019.9013181.

Karthik Elamvazhuthi and Spring Berman. Mean-field models in swarm robotics: a survey. *Bioinspiration & Biomimetics*, 15(1):015001, November 2019. doi: 10.1088/1748-3190/ab49a4. URL https://dx.doi.org/10.1088/1748-3190/ab49a4. Publisher: IOP Publishing.

Y Achdou, FJ Buera, JM Lasry, PL Lions, and B Moll. Partial differential equation models in macroeconomics. *Philosophical transactions. Series A, Mathematical, physical, and engineering science*, 2014. doi: 10.1098/rsta.2013.0397.

Philippe Casgrain and Sebastian Jaimungal. Mean field games with partial information for algorithmic trading, 2018. URL https://arxiv.org/abs/1803.04094.

Antonio De Paola, Vincenzo Trovato, David Angeli, and Goran Strbac. A mean field game approach for distributed control of thermostatic loads acting in simultaneous energy-frequency response markets. *IEEE Transactions on Smart Grid*, 10(6):5987–5999, 2019. doi: 10.1109/TSG.2019.2895247.

Ana Bušić and Sean Meyn. Distributed control of thermostatically controlled loads: Kullback-leibler optimal control in continuous time. *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7258–7265, 2019. doi: 10.1109/CDC40024.2019.9029603.

Wonjun Lee, Siting Liu, Hamidou Tembine, Wuchen Li, and Stanley Osher. Controlling Propagation of Epidemics via Mean-Field Control. *SIAM Journal on Applied Mathematics*, 81(1):190–207, 2021. doi: 10.1137/20M1342690. URL https://doi.org/10.1137/20M1342690. _eprint: https://doi.org/10.1137/20M1342690.

Weinan E, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):10, December 2018. ISSN 2197-9847. doi: 10.1007/s40687-018-0172-y. URL https://doi.org/10.1007/s40687-018-0172-y.

Lars Ruthotto, Stanley J. Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020. doi: 10.1073/pnas.1922204117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1922204117. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1922204117.

Jean-Pierre Fouque and Zhaoyu Zhang. Deep Learning Methods for Mean Field Control Problems With Delay. *Frontiers in Applied Mathematics and Statistics*, 6, 2020. ISSN 2297-4687. doi: 10.3389/fams.2020.00011. URL https://www.frontiersin.org/articles/10.3389/fams.2020.00011.

Alex Tong Lin, Samy Wu Fung, Wuchen Li, Levon Nurbekyan, and Stanley J. Osher. Alternating the population and control neural networks to solve high-dimensional stochastic mean-field games. *Proceedings of the National Academy of Sciences*, 118(31):e2024713118, 2021. doi: 10.1073/pnas.2024713118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2024713118.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, Belmont, MA, USA, 3rd edition, 2005.

Satoru Ihara and Fred C. Schweppe. Physically based modeling of cold load pickup. *IEEE Transactions on Power Apparatus and Systems*, PAS-100(9):4142–4150, 1981. doi: 10.1109/TPAS.1981.316965.

R. Malhame and Chee-Yee Chong. Electric load model synthesis by diffusion approximation of a high-order hybrid-state stochastic system. *IEEE Transactions on Automatic Control*, 30(9):854–860, 1985. doi: 10.1109/TAC.1985.1104071.

R.E. Mortensen and K.P. Haggerty. A stochastic computer model for heating and cooling loads. *IEEE Transactions on Power Systems*, 3(3):1213–1219, 1988. doi: 10.1109/59.14584.

Arman Kizilkale and Roland Malhame. Mean field based control of power system dispersed energy storage devices for peak load relief. In *Proceedings of the IEEE Conference on Decision and Control*, pages 4971–4976, 12 2013. ISBN 978-1-4673-5717-3. doi: 10.1109/CDC.2013.6760669.

Arman C. Kizilkale and Roland P. Malhame. Collective target tracking mean field control for markovian jump-driven models of electric water heating loads. *IFAC Proceedings Volumes*, 47(3):1867–1872, 2014. ISSN 1474-6670. doi: https://doi.org/10.3182/20140824-6-ZA-1003.00630. URL https://www.sciencedirect.com/science/article/pii/S1474667016418859. 19th IFAC World Congress.

Caroline Le Floch, Emre Can Kara, and Scott Moura. Pde modeling and control of electric vehicle fleets for ancillary services: A discrete charging case. *IEEE Transactions on Smart Grid*, 9(2):573–581, 2018. doi: 10.1109/TSG.2016.2556643.

Adrien Séguret, Clémence Alasseur, J. Frédéric Bonnans, Antonio De Paola, Nadia Oudjane, and Vincenzo Trovato. Decomposition of high dimensional aggregative stochastic control problems, 2020. URL https://arxiv.org/abs/2008.09827.

Neil Cammardella, Ana Bušić, Yuting Ji, and Sean Meyn. Kullback-leibler-quadratic optimal control of flexible power demand. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 4195–4201, 2019. doi: 10.1109/CDC40024.2019.9029512.

Neil Cammardella, Ana Bušić, and Sean Meyn. Kullback-leibler-quadratic optimal control in a stochastic environment. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 158–165, 2021. doi: 10.1109/CDC45484.2021.9682943.

Margaux Brégère, Pierre Gaillard, Yannig Goude, and Gilles Stoltz. Target tracking for contextual bandits: Application to demand side management. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 754–763. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/bregere19a.html.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2:229–260, 2007.

Minyi Huang, Roland Malhame, and Peter Caines. Large population stochastic dynamic games: Closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Commun. Inf. Syst.*, 6, 01 2006. doi: 10.4310/CIS.2006.v6.n3.a5.

Alain Bensoussan, Jens Frehse, and Phillip Yam. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Rene Carmona, R. Delarue, and A. Lachapelle. Control of McKean–Vlasov dynamics versus mean field games. *Math Finan Econ*, 7:131–166, 2013. doi: https://doi.org/10.1007/s11579-012-0089-y.

Clemence Alasseur, Imen Taher, and Anis Matoussi. An extended mean field game for storage in smart grids. *Journal of Optimization Theory and Applications*, 184:1–27, 02 2020. doi: 10.1007/s10957-019-01619-3.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

René Carmona and Mathieu Laurière. Convergence Analysis of Machine Learning Algorithms for the Numerical Solution of Mean Field Control and Games I: The Ergodic Case. *SIAM Journal on Numerical Analysis*, 59 (3):1455–1485, 2021a. doi: 10.1137/19M1274377. URL https://doi.org/10.1137/19M1274377. _eprint: https://doi.org/10.1137/19M1274377.

René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: II – the finite horizon case. *The Annals of Applied Probability*, 32:4065–4105, 12 2022.

René Carmona and Mathieu Laurière. Deep learning for mean field games and mean field control with applications to finance, 2021b. URL https://arxiv.org/abs/2107.04568.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning, 2019. URL https://arxiv.org/abs/1910.12802.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/030e65da2b1c944090548d36b244b28d-Paper.pdf.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Lauriere. Reinforcement learning for mean field games, with applications to economics. Preprint, 2021. URL https://arxiv.org/abs/2106.13755.

Sarah Perrin, Julien Perolat, Mathieu Lauriere, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13199–13213. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/995ca733e3657ff9f5f3c823d73371e1-Paper.pdf.

Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up mean field games with online mirror descent. In *Proceedings of the 39th International Conference on Machine Learning*, ICML'22, 03 2022.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 489–497, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3 (1-2):95–110, 1956. doi: https://doi.org/10.1002/nav.3800030109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800030109.

Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/40538442.

Aviv Rosenberg and Yishay Mansour. Online Stochastic Shortest Path with Bandit Feedback and Unknown Transition Function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/a0872cc5b5ca4cc25076f3d868e1bdf8-Paper.pdf.

Travis Dick, András György, and Csaba Szepesvári. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page I–512–I–520. JMLR.org, 2014.

Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. In *Deep Reinforcement Learning Symposium*, NIPS'17, 05 2017.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2160–2169. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, may 2003. ISSN 0167-6377. doi: 10.1016/S0167-6377(02)00231-6. URL https://doi.org/10.1016/S0167-6377(02)00231-6.

A Nemirovski and D Yudin. *Problem complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, feb 2012. ISSN 1935-8237. doi: 10.1561/2200000018. URL https://doi.org/10.1561/2200000018.

Ana Bušić and Sean Meyn. Distributed Randomized Control for Demand Dispatch. In *55th IEEE Conference on Decision and Control (CDC)*, Proceedings of 55th IEEE Conference on Decision and Control, December 2016.

Pierre-Emmanuel Jabin and Zhenfu Wang. Mean Field Limit for Stochastic Particle Systems. In Nicola Bellomo, Pierre Degond, and Eitan Tadmor, editors, *Active Particles, Volume 1 : Advances in Theory, Models, and Applications*, pages 379–402. Springer International Publishing, Cham, 2017. ISBN 978-3-319-49996-3. doi: 10.1007/978-3-319-49996-3_10. URL https://doi.org/10.1007/978-3-319-49996-3_10.

Jérémy Albouys, Nicolas Sabouret, Yvon Haradji, Mathieu Schumann, and Christian Inard. SMACH: Multi-agent Simulation of Human Activity in the Household. In Yves Demazeau, Eric Matson, Juan Manuel Corchado, and Fernando De la Prieta, editors, *Advances in Practical Applications of Survivable Agents and Multi-Agent Systems: The PAAMS Collection*, pages 227–231, Cham, 2019. Springer International Publishing. ISBN 978-3-030-24209-1.

Diogo A. Gomes and Vardan K. Voskanyan. Extended deterministic mean-field games. *SIAM Journal on Control and Optimization*, 54(2):1030–1055, 2016. doi: 10.1137/130944503. URL https://doi.org/10.1137/130944503.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787.

A Missing proofs

A.1 Proof of Proposition 3.2

Proof. Consider a fixed initial state-action distribution $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$. Let $\mu \in \mathcal{M}_{\mu_0}$ and define $\rho = (\rho_n)_{1\leq n\leq N}$ such that for all $x \in \mathcal{X}$, $\rho_n(x) = \sum_a \mu_n(x,a)$ (the associated state distribution). Define a policy sequence $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ such that $\pi_n(a|x) = \frac{\mu_n(x,a)}{\rho_n(x)}$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$. We want to show that $\mu^\pi = \mu$ for this policy π. We reason by induction. For $n = 0$, $\mu_0^\pi = \mu_0$ by definition. Suppose $\mu_n^\pi = \mu_n$, thus for $n+1$ and for all $(x',a') \in \mathcal{X}\times\mathcal{A}$

$$\begin{aligned}
\mu_{n+1}^\pi(x',a') &= \sum_{x,a} p(x'|x,a)\mu_n^\pi(x,a)\pi_{n+1}(a'|x') \\
&= \sum_{x,a} p(x'|x,a)\mu_n(x,a)\frac{\mu_{n+1}(x',a')}{\rho_{n+1}(x')} \\
&= \sum_{a} \mu_{n+1}(x',a)\frac{\mu_{n+1}(x',a')}{\rho_{n+1}(x')} \\
&= \rho_{n+1}(x')\frac{\mu_{n+1}(x',a')}{\rho_{n+1}(x')} \\
&= \mu_{n+1}(x',a'),
\end{aligned}$$

where the first equality comes from Definition 3.1, the second equality comes from the induction assumption and the way we defined the strategy π, and the third comes from the assumption that $\mu \in \mathcal{M}_{\mu_0}$.

□

A.2 Proof of Proposition 4.4

Proof. The convexity of each f_n for $1 \leq n \leq N$, and the convexity of the set $\mathcal{M}_{\mu_0}$ ensure the existence of a minimizer of Problem (4) satisfying the optimality conditions. Also, Proposition 3.2 shows, for a fixed initial state-action distribution μ_0, a bijection between the sets $(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ and $\mathcal{M}_{\mu_0}$.

Let (μ^*, π^*), where $\mu^* = \mu^{\pi^*}$, be a Nash equilibrium.

As introduced in Section 3, a Nash equilibrium (μ^*, π^*) satisfies $\pi^* = \arg\max_\pi J(\pi, \mu^*)$ by definition. In other words,

$$J(\pi^*, \mu^{\pi^*}) \geq J(\pi, \mu^{\pi^*}) \quad \forall \pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}. \tag{12}$$

Expanding the terms of the sum of expected rewards and using the definition of reward in a potential game, we obtain that

$$\begin{aligned}
J(\pi, \mu^{\pi^*}) &= \mathbb{E}_\pi\left[\sum_{n=1}^N r_n(x_n, a_n, \mu_n^{\pi^*})\right] \\
&= \sum_{n=1}^N \sum_{x\in\mathcal{X}, a\in A} r_n(x, a, \mu_n^{\pi^*})\mu_n^\pi(x,a) \\
&= \sum_{n=1}^N -\left\langle \nabla f_n(\mu_n^{\pi^*}), \mu_n^\pi \right\rangle.
\end{aligned}$$

Similarly,

$$J(\pi^*, \mu^{\pi^*}) = \sum_{n=1}^N -\left\langle \nabla f_n(\mu_n^{\pi^*}), \mu_n^{\pi^*} \right\rangle.$$

Thus, the Nash equilibrium condition in Inequality (12) entails

$$\sum_{n=1}^{N} \left\langle \nabla f_n(\mu_n^{\pi^*}), \mu_n^{\pi^*} - \mu_n^{\pi} \right\rangle \leq 0. \tag{13}$$

As f_n is convex for all $n \in \{1, ..., N\}$, this yields

$$\sum_{n=1}^{N} f_n(\mu_n^{\pi^*}) - f_n(\mu_n^{\pi}) \leq 0. \tag{14}$$

Thus, π^* satisfies the optimality conditions of Problem (4). We then proved that if (π^*, μ^*) is a NE with $\mu^* = \mu^{\pi^*}$, then π^* is an optimum of Problem (4).

On the other way around, if π^* is a minimizer of Problem (4) then it satisfies Inequality (14) for all $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X} \times N}$. Again, by convexity of $(f_n)_{1 \leq n \leq N}$, π^* also satisfies Inequality (13). Following the same calculations backwards, we obtain that π^* then satisfies Inequality (12), and by definition is then a NE. This concludes the first part of the proof.

The second part concerns the monotonicity of the game, defined below for the mean field game framework.

Definition A.1 (Monotonicity)**.** According to Lasry and Lions [2007], a game where the reward depends on the population's state-action distribution (sometimes called "extended MFG" in the literature, see Gomes and Voskanyan [2016]) is (strictly) monotone if for any state-action distributions $\nu, \nu' \in \Delta_{\mathcal{X} \times \mathcal{A}}$ with $\nu \neq \nu'$,

$$\int_{\mathcal{X}, \mathcal{A}} [r(x, a, \nu) - r(x, a, \nu')] d(\nu - \nu')(x, a) \leq 0, \ (< 0).$$

Back to the proof, consider μ, μ' two distributions over $\mathcal{X} \times \mathcal{A}$. As the result should be true to all n, we omit the time step index for the computations. Recall that the reward is of the form $r(x, a, \mu) = -\nabla f(\mu)(x, a)$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, with f a convex function. Then,

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{A}} [r(x, a, \mu) - r(x, a, \mu')] d(\mu - \mu')(x, a) &= \int_{\mathcal{X} \times \mathcal{A}} [\nabla f(\mu')(x, a) - \nabla f(\mu)(x, a)] d(\mu - \mu')(x, a) \\ &= \langle \nabla f(\mu') - \nabla f(\mu), \mu - \mu' \rangle \leq 0 \end{aligned}$$

where the last inequality comes from the convexity of f. □

B Missing proofs: algorithm 1 scheme and convergence rate

By abuse of notations, for any probability measure $\eta \in \Delta_E$ whatever the finite space E on which it is defined we introduce the neg-entropy function

$$\phi(\eta) := \sum_{x \in E} \eta(x) \log \eta(x),$$

to which we associate the Bregman divergence D, also known as the Kullback-Leibler divergence, such that for any pair $(\eta, \nu) \in \Delta_E \times \Delta_E$,

$$D(\eta, \nu) := \phi(\eta) - \phi(\nu) - \langle \phi'(\nu), \eta - \nu \rangle.$$

Observe that to any $\mu = (\mu_n)_{1 \leq n \leq N} \in \mathcal{M}_{\mu_0}$ one can associate a unique probability mass function on $(\mathcal{X} \times \mathcal{A})^N$ denoted by $\mu_{1:N}$ such that $\mu_{1:N}$ is *generated* by the strategy $\pi = (\pi_n)_{1 \leq n \leq N}$ associated with μ which is determined by

$$\pi_n(a|x) = \frac{\mu_n(x, a)}{\rho_n(x)},$$

where ρ_n denotes the marginal probability distribution on $\mathcal{X}$ associated with μ_n i.e., for all $x \in \mathcal{X}$

$$\rho_n(x) := \sum_{a \in \mathcal{A}} \mu_n(x, a).$$

Before proving Theorems (4.1) and (4.3) we state and prove a Lemma which is key to proving both theorems.

Lemma B.1. *For any probability mass functions $\mu_{1:N}, \mu'_{1:N} \in \mathcal{P}\big((\mathcal{X}\times\mathcal{A})^N\big)$ generated by π, π' respectively with the same initial state-action distribution, i.e. $\mu_0 = \mu'_0$, we have*

$$\begin{aligned} D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu_n(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi'_n(a|x)}\right)\right] \\ &= \sum_{n=1}^{N} D(\mu_n, \mu'_n) - \sum_{n=0}^{N} D(\rho_n, \rho'_n) \end{aligned} \tag{15}$$

Proof. For each $1 \le n \le N$, let us define a transition matrix P^{π_n} for all $x, x' \in \mathcal{X}$ and $a, a' \in \mathcal{A}$,

$$P^{\pi_n}(x', a'|x, a) := p(x'|x, a)\pi_n(a'|x').$$

Given Definition 3.1, for any randomized policy the state-action distributions evolve according to linear dynamics

$$\mu_n(x', a') = \langle \mu_{n-1}(\cdot), P^{\pi_n}(x', a'|\cdot)\rangle.$$

Any randomized policy π gives a probability mass function $\mu_{1:N}$ that is Markovian:

$$\mu_{1:N}(\vec{y}) = \mu_0(y_0)P^{\pi_1}(y_1|y_0)...P^{\pi_N}(y_N|y_{N-1}), \tag{16}$$

where $\vec{y}$ represents the elements of $(\mathcal{X}\times\mathcal{A})^{N+1}$ such that $y_i = (x_i, a_i)$ for all $0 \le i \le N$. Note that $\mu_n(y_n)$ is the marginal probability mass function.

Consider $\mu, \mu' \in \mathcal{M}_{\mu_0}$ the state-action distribution sequences induced by π, π' respectively (i.e, $\mu = \mu^\pi$ and $\mu' = \mu^{\pi'}$). Thus, computing the relative entropy between the probability mass functions $\mu_{1:N}, \mu'_{1:N}$ gives

$$\begin{aligned} D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{\vec{y}} \mu_{1:N}(\vec{y}) \log\left(\frac{\mu_{1:N}(\vec{y})}{\mu'_{1:N}(\vec{y})}\right) \\ &= \sum_{y_0,\dots,y_N} \mu_{1:N}(\vec{y}) \log\left(\frac{\mu_0(y_0)P^{\pi_1}(y_1|y_0)...P^{\pi_N}(y_N|y_{N-1})}{\mu'_0(y_0)P^{\pi'_1}(y_1|y_0)...P^{\pi'_N}(y_N|y_{N-1})}\right) \\ &= \sum_{y_0,\dots,y_N} \mu_{1:N}(\vec{y}) \sum_{i=1}^{N} \log\left(\frac{P^{\pi_i}(y_i|y_{i-1})}{P^{\pi'_i}(y_i|y_{i-1})}\right). \end{aligned}$$

Where

$$\begin{aligned} \sum_{i=1}^{N} \log\left(\frac{P^{\pi_i}(y_i|y_{i-1})}{P^{\pi'_i}(y_i|y_{i-1})}\right) &= \sum_{i=1}^{N} \log\left(\frac{p(x_i|x_{i-1}, a_{i-1})\pi_i(a_i|x_i)}{p(x_i|x_{i-1}, a_{i-1})\pi'_i(a_i|x_i)}\right) \\ &= \sum_{i=1}^{N} \log\left(\frac{\pi_i(a_i|x_i)}{\pi'_i(a_i|x_i)}\right). \end{aligned}$$

Thus,

$$\begin{aligned} D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{\vec{y}} \mu_{1:N}(\vec{y}) \sum_{i=1}^{N} \log\left(\frac{\pi_i(a_i|x_i)}{\pi'_i(a_i|x_i)}\right) \\ &= \sum_{\vec{y}} \mu_0(y_0)P^{\pi_1}(y_1|y_0)...P^{\pi_N}(y_N|y_{N-1}) \sum_{i=1}^{N} \log\left(\frac{\pi_i(a_i|x_i)}{\pi'_i(a_i|x_i)}\right) \\ &= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_i(x, a) \log\left(\frac{\pi_i(a|x)}{\pi'_i(a|x)}\right). \end{aligned}$$

Where for the last equality we used that

$$\sum_{y_0,\dots,y_{i-1}} \mu_0(y_0)P^{\pi_1}(y_1|y_0)...P^{\pi_i}(y_i|y_{i-1}) = \sum_{y_i} \mu_i(y_i)$$

and for a fixed y_i,

$$\sum_{y_{i+1},\dots,y_N} P^{\pi_{i+1}}(y_{i+1}|y_i)...P^{\pi_N}(y_N|y_{N-1}) = 1.$$

This proves the first equality of the Lemma. We now prove the second. For this, we recall that Proposition 3.2 gives a unique relation between a state-action distribution sequence $\mu \in \mathcal{M}_{\mu_0}$ and the policy sequence $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ inducing it by taking for all $1 \leq i \leq N$, $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$\pi_i(a|x) = \frac{\mu_i(x,a)}{\rho_i(x)},$$

where ρ is the marginal on the states of μ. Using this relation, we have then that

$$\begin{aligned}
D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_i(x,a) \log\left(\frac{\pi_i(a|x)}{\pi'_i(a|x)}\right) \\
&= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_i(x,a) \log\left(\frac{\mu_i(a|x)}{\rho_i(x)} \frac{\rho'_i(x)}{\mu'_i(a|x)}\right) \\
&= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_i(x,a) \log\left(\frac{\mu_i(a|x)}{\mu'_i(a|x)}\right) - \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_i(x,a) \log\left(\frac{\rho_i(x)}{\rho'_i(x)}\right) \\
&= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_i(x,a) \log\left(\frac{\mu_i(a|x)}{\mu'_i(a|x)}\right) - \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \rho_i(x) \log\left(\frac{\rho_i(x)}{\rho'_i(x)}\right) \\
&= \sum_{i=1}^{N} D(\mu_i, \mu'_i) - \sum_{i=1}^{i} D(\rho_i, \rho'_i)
\end{aligned}$$

which concludes the proof. □

B.1 Proof of Theorem 4.1: formulation of Algorithm 1

Proof. At each iteration we seek to solve

$$\mu^{k+1} \in \underset{\mu^\pi \in \mathcal{M}_{\mu_0}}{\arg\min} \left\{ \langle \nabla F(\mu^k), \mu^\pi \rangle + \frac{1}{\tau_k} \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu_n(\cdot)} \left[\log\left(\frac{\pi_n(a|x)}{\pi_n^k(a|x)}\right) \right] \right\} \tag{17}$$

where recall that $\langle \nabla F(\mu^k), \mu^\pi \rangle := \sum_{n=1}^{N} \langle \nabla f_n(\mu_n^k), \mu_n^\pi \rangle$.

Using the definition of the reward in a potential game and the associated expected sum of rewards J_{μ_0}, we reformulated this problem in Section 4 of the main paper as follows

$$\mu^{k+1} \in \underset{\mu\in\mathcal{M}_{\mu_0}}{\arg\max} \left\{ J_{\mu_0}(\pi, \mu^k) - \frac{1}{\tau_k} \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu_n(\cdot)} \left[\log\left(\frac{\pi_n(a|x)}{\pi_n^k(a|x)}\right) \right] \right\}. \tag{18}$$

Now, we use the optimality principle to solve this optimization problem with an algorithm backward in time. Remember that the initial distribution μ_0 is always fixed. The equivalence between solving a minimization problem on sequences of state-action distributions in $\mathcal{M}_{\mu_0}$ and on sequences of policies in $(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ (see Proposition 3.2), allows us to reformulate Problem (18) on $\mathcal{M}_{\mu_0}$ into a problem on $(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$, thus

$$
\begin{aligned}
(18) &= \max_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} \sum_{n=0}^{N}\sum_{x,a}\mu_n^{\pi}(x,a)r_n(x,a,\mu_n^k) - \frac{1}{\tau_k}\sum_{n=1}^{N}\sum_{x,a}\mu_{n-1}^{\pi}(x,a)\sum_{x',a'}p(x'|x,a)\pi_n(a'|x')\log\left(\frac{\pi_n(a'|x')}{\pi_n^k(a'|x')}\right)\\
&= \max_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} \sum_{n=0}^{N}\sum_{x,a}\mu_n^{\pi}(x,a)\left[r_n(x,a,\mu_n^k) - \frac{1}{\tau_k}\sum_{x',a'}p(x'|x,a)\pi_{n+1}(a'|x')\log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right)\right]\\
&= \max_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} \mathbb{E}_{\pi}\left[r_N(x_N,a_N,\mu_N^k) + \sum_{n=0}^{N-1} r_n(x_n,a_n,\mu_n^k) - \frac{1}{\tau_k}\sum_{x',a'}p(x'|x_n,a_n)\pi_{n+1}(a'|x')\log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right)\right].
\end{aligned}
$$

Let us define a regularized version of the state-action value function that we denote by $\tilde{Q}^k$, such that for all $1\leq i\leq N$, $(x,a)\in\mathcal{X}\times\mathcal{A}$,

$$
\begin{aligned}
\tilde{Q}_i^k(x,a) = \max_{\pi_{i+1:N}\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N-i}} \mathbb{E}_{\pi}\Bigg[r_N(x_N,a_N,\mu_N^k) + \sum_{n=i}^{N-1}\Bigg\{r_n(x_n,a_n,\mu_n^k)\\
- \frac{1}{\tau_k}\sum_{x',a'}p(x'|x_n,a_n)\pi_{n+1}(a'|x')\log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right)\Bigg\}\Bigg|(x_i,a_i)=(x,a)\Bigg],
\end{aligned}
\tag{19}
$$

where $\pi_{i+1:N}=\{\pi_{i+1},...,\pi_N\}$.

First, note that $\mathbb{E}_{(x,a)\sim\mu_0(\cdot)}[\tilde{Q}_0^k(x,a)] = (18)$. Moreover, the optimality principle states that this regularized state-action value function satisfies the following recursion

$$
\begin{cases}
\tilde{Q}_N(x,a) = r_N(x,a,\mu_N^k)\\
\tilde{Q}_i(x,a) = \max_{\pi_{i+1}\in(\Delta_{\mathcal{A}})^{\mathcal{X}}}\left\{r_i(x,a,\mu_i^k) + \sum_{x'}p(x'|x,a)\sum_{a'}\pi_{i+1}(a'|x')\left[-\frac{1}{\tau_k}\log\left(\frac{\pi_{i+1}(a'|x')}{\pi_{i+1}^k(a'|x')}\right) + \tilde{Q}_{i+1}(x',a')\right]\right\}.
\end{cases}
$$

Thus, to solve (18) we compute backwards in time, i.e. for $i=N-1,...,0$, for all $x\in\mathcal{X}$,

$$
\pi_{i+1}^{k+1}(\cdot|x)\in\operatorname*{arg\,max}_{\pi(\cdot|x)\in\Delta_{\mathcal{A}}}\left\{\left\langle\pi(\cdot|x),\tilde{Q}_{i+1}^k(x,\cdot)\right\rangle - \frac{1}{\tau_k}D\big(\pi(\cdot|x),\pi_{i+1}^k(\cdot|x)\big)\right\},
$$

where D is the Kullback-Leibler divergence.

The solution of this optimisation problem for each time step i can be found by writing the Lagrangian function $\mathcal{L}$ associated. Let λ be the Lagrangian multiplier associated to the simplex constraint. For simplicity, let $\pi_x := \pi(\cdot|x)$, $\pi_x^k := \pi_{i+1}^k(\cdot|x)$ and $\tilde{Q}_x^k := \tilde{Q}_{i+1}^k(x,\cdot)$. Thus,

$$
\mathcal{L}(\pi_x,\lambda) = \langle\pi_x,\tilde{Q}_x^k\rangle - \frac{1}{\tau_k}D(\pi_x,\pi_x^k) - \lambda\left(\sum_{a\in\mathcal{A}}\pi_x(a)-1\right).
$$

Taking the gradient of the Lagrangian with respect to $\pi_x(a)$ for each $a\in\mathcal{A}$ gives

$$
\frac{\partial\mathcal{L}}{\partial\pi_x(a)} = \tilde{Q}_x^k(a) - \frac{1}{\tau_k}\log\left(\frac{\pi_x(a)}{\pi_x^k(a)}\right) - \frac{1}{\tau_k} - \lambda,
$$

and thus

$$
\frac{\partial\mathcal{L}}{\partial\pi_x(a)} = 0 \Longrightarrow \quad \pi_x(a) = \pi_x^k(a)\exp\Big(\tau_k\tilde{Q}_x^k(a) - 1 - \tau_k\lambda\Big).
$$

Applying the simplex constraint, $\sum_{a\in\mathcal{A}}\pi_x(a)=1$, we find the value of the Lagrangian multipler λ, and we get for all $a\in\mathcal{A}$

$$
\pi_x(a) = \frac{\pi_x^k(a)\exp\Big(\tau_k\tilde{Q}_x^k(a)\Big)}{\sum_{a'\in\mathcal{A}}\pi_x^k(a')\exp\Big(\tau_k\tilde{Q}_x^k(a')\Big),}
$$

which proves the theorem.

□

B.2 Proof of Proposition 4.2: Γ is a Bregman divergence

Proof. Lemma B.1 states that

$$\Gamma(\mu, \mu') := \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu_n(\cdot)} \left[\log \left(\frac{\pi_n(a'|x')}{\pi_n^k(a'|x')} \right) \right] = \sum_{t=0}^{n} D(\mu'_t, \mu_t) - \sum_{t=0}^{n} D(\rho'_t, \rho_t).$$

Recall that ϕ is the negentropy and that D is the Bregman divergence induced by the negentropy. Define the function $\psi : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ such that

$$\psi(\mu) := \sum_{n=0}^{N} \phi(\mu_n) - \sum_{n=0}^{N} \phi(\rho_n).$$

Note that for $\mu, \mu' \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$ with marginals given by $\rho, \rho' \in (\Delta_{\mathcal{X}})^N$, using the second equality of Lemma B.1,

$$\psi(\mu) - \psi(\mu') - \langle \nabla\psi(\mu'), \mu - \mu' \rangle = \Gamma(\mu, \mu').$$

Thus, for Γ to be a Bregman divergence it is sufficient to show that ψ is a convex function. Recall that the marginal ρ is such that for each $1 \leq n \leq N$, and for all $x \in \mathcal{X}$, $\rho_n(x) = \sum_{a\in\mathcal{A}} \mu_n(x, a)$. Thus,

$$\begin{aligned}
\psi(\mu) &= \sum_n \left[\sum_{x,a} \mu_n(x,a) \log(\mu_n(x,a)) - \sum_x \rho_n(x) \log(\rho_n(x)) \right] \\
&= \sum_n \sum_{x,a} \mu_n(x,a) \log \left(\frac{\mu_n(x,a)}{\sum_{a'} \mu_n(x,a')} \right).
\end{aligned}$$

Computing the first order partial derivative of ψ with respect to $\mu_n(x, a)$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $1 \leq n \leq N$, we get

$$\begin{aligned}
\frac{\partial\psi}{\partial\mu_n(x,a)}(\mu) &= \log \left(\frac{\mu_n(x,a)}{\sum_{a'} \mu_n(x,a')} \right) + \mu_n(x,a) \frac{1}{\mu_n(x,a)} - \sum_{a'} \mu_n(x,a') \frac{1}{\sum_{a'} \mu_n(x,a')} \\
&= \log \left(\frac{\mu_n(x,a)}{\sum_{a'} \mu_n(x,a')} \right) \\
&= \log \left(\frac{\mu_n(x,a)}{\rho_n(x)} \right).
\end{aligned}$$

Now we apply the following convexity property [Boyd and Vandenberghe, 2004]: ψ is convex if and only if for all $\mu, \mu' \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$, $\langle \psi'(\mu) - \psi'(\mu'), \mu - \mu' \rangle \geq 0$. Indeed,

$$\begin{aligned}
\langle \psi'(\mu) - \psi'(\mu'), \mu - \mu' \rangle &= \sum_n \sum_{x,a} \left[\frac{\partial\psi}{\partial\mu_n(x,a)}(\mu) - \frac{\partial\psi}{\partial\mu_n(x,a)}(\mu') \right] \big(\mu_n(x,a) - \mu'_n(x,a)\big) \\
&= \sum_n \sum_{x,a} \left[\log \left(\frac{\mu_n(x,a)}{\rho_n(x)} \right) - \log \left(\frac{\mu'_n(x,a)}{\rho'_n(x)} \right) \right] \big(\mu_n(x,a) - \mu'_n(x,a)\big) \\
&= \sum_n \sum_{x,a} \log \left(\frac{\mu_n(x,a)}{\mu'_n(x,a)} \right) \big(\mu_n(x,a) - \mu'_n(x,a)\big) - \sum_n \sum_x \log \left(\frac{\rho_n(x)}{\rho'_n(x)} \right) \big(\rho_n(x) - \rho'_n(x)\big) \\
&\overset{(a)}{=} \sum_n D(\mu_n, \mu'_n) + D(\mu_n, \mu'_n) - D(\rho_n, \rho'_n) - D(\rho'_n, \rho_n) \\
&\overset{(b)}{=} \Gamma(\mu, \mu') + \Gamma(\mu', \mu) \\
&\overset{(c)}{=} D(\mu_{1:N}, \mu'_{1:N}) + D(\mu'_{1:N}, \mu_{1:N}) \overset{(d)}{\geq} 0,
\end{aligned}$$

where (a) comes from the definition of the Kullback-Leibler divergence D, (b) comes from the definition of Γ, (c) comes from Lemma B.1 and (d) comes from a property of Bregman divergences that they are always positive. As ψ is convex and induces the divergence Γ then Γ is a Bregman divergence. After writing this proof, we came across a different strategy to prove that Γ is a Bregman divergence that is presented in Appendix A of Neu et al. [2017]. □

B.3 Proof of Theorem 4.3

We just need to prove that if $(f_n)_{1\leq n\leq N}$ are convex and Lipschitz then so is F, which places our optimization problem under the convergence hypothesis of mirror descent. The rest of the proof follows from the fact that each iteration of Algorithm 1 is an iteration of mirror descent, so it benefits from the same convergence result whose proof can be verified in Beck and Teboulle [2003].

Convexity: F is convex as the sum of convex functions.

Lipschitz: First, with some abuse of notation, we consider that for all $\mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$ such that $\mu = (\mu_n)_{1\leq n\leq N}$ and that for any norm $\|\cdot\|$,

$$\|\nabla F(\mu)\|^2 = \langle \nabla F(\mu), \nabla F(\mu)\rangle := \sum_{n=1}^{N} \langle \nabla f_n(\mu_n), \nabla f_n(\mu_n)\rangle = \sum_{n=1}^{N} \|\nabla f_n(\mu_n)\|^2.$$

As f_n is convex and Lipschitz with respect to $\|\cdot\|$ with constant l_n, then from Lemma 2.6 of [Shalev-Shwartz, 2012], for all $\mu_n \in \Delta_{\mathcal{X}\times\mathcal{A}}$, $\|\nabla f_n(\mu_n)\|_* \leq l_n$, where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

Thus,

$$\|\nabla F(\mu)\|_*^2 = \sum_{n=1}^{N} \|\nabla f_n(\mu_n)\|_*^2 \leq \left(\sum_{n=1}^{N} l_n^2\right).$$

Therefore, F is Lipschitz with constant $L := \left(\sum_{n=1}^{N} l_n^2\right)^{1/2}$.

C Algorithms

Algorithm 2 Fictitious play for MFG (FP)

Input: number of iterations K, initial policy π^0.
Initialization: $\bar{\mu}^0 = \mu^{\pi^0}$ as in Definition 3.1.
for $k = 0, ..., K$ **do**
  $\pi^{k+1} \in \arg\max_\pi J(\pi, \bar{\mu}^k)$, best response against $\bar{\mu}^k$.
  $\bar{\mu}^{k+1} = \frac{1}{k+1}\mu^{\pi^{k+1}} + \frac{k}{k+1}\bar{\mu}^k$.
end for
Return: $\bar{\mu}^K$ and $\bar{\pi}^K$ s.t. $\bar{\pi}_n^K(a|x) := \sum_{k=0}^{K} \frac{\rho_n^{\pi^k}(x)\pi_n^k(a|x)}{\sum_{k=0}^{K}\rho_n^{\pi^k}(x)}$, $\left(\rho_n^{\pi^k}(x) := \sum_{a\in\mathcal{A}} \mu_n^{\pi^k}(x,a) \text{ for all } k \leq K\right)$.

Algorithm 3 Frank Wolfe

Input: number of iterations K, initial distribution μ^0, sequence $(\eta_k)_k$.
for $k = 0, ..., K$ **do**
  $\mu^k \in \arg\min_{\mu\in\mathcal{M}} \left\langle \mu, \nabla F(\bar{\mu}^k)\right\rangle_{|\mathcal{X}\times\mathcal{A}|}$.
  $\bar{\mu}^{k+1} = (1-\eta_{k+1})\bar{\mu}^k + \eta_{k+1}\mu^k$.
end for
Return: $\bar{\mu}^K$

The Online Mirror Descent for MFG algorithm uses the regular state-value function (or Q-function) at each iteration. It's definition is given by

$$Q_n^{\pi,\mu}(x,a) := \mathbb{E}_\pi\left[\sum_{i=n}^{N} r_i(x_i, a_i, \mu_i)\middle| x_n = x, a_n = a\right]. \tag{20}$$

Note that, considering an initial state-action distribution μ_0, $J_{\mu_0}(\pi,\mu) = \mathbb{E}_{(x,a)\sim\mu_0}[Q_0^{\pi,\mu}(x,a)]$. Furthermore, $Q^{\pi,\mu}$ is the solution of the backward equation, for all $n < N$, $(x,a) \in \mathcal{X}\times\mathcal{A}$:

$$\begin{cases} Q_N^{\pi,\mu}(x,a) = r_N(x,a,\mu_N) \\ Q_n^{\pi,\mu}(x,a) = r_n(x,a,\mu_n) + \sum_{x'} p(x'|x,a)\sum_{a'}\pi_{n+1}(a'|x')Q_{n+1}^{\pi,\mu}(x',a'). \end{cases} \tag{21}$$

Algorithm 4 OMD for MFG

Input: number of iterations K, $\pi^0 \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$.
for $k = 0, ..., K$ **do**
  $\mu^k := \mu^{\pi^k}$, as in Definition 3.1.
  $Q^k := Q^{\pi^k,\mu^k}$ as in Equation (21).
  $\pi_n^{k+1}(\cdot|x) := \arg\max_{\pi(\cdot|x)\in\Delta_{\mathcal{A}}} \langle Q_n^k(x,\cdot), \pi(\cdot|x)\rangle + \tau D\left(\pi(\cdot|x), \pi_n^k(\cdot|x)\right), \forall x \in \mathcal{X}, \forall n \leq N.$
end for
Return: μ^K, π^K

D Water heater application

D.1 Standard cycling behavior of one water heater

Let us consider a time window $[t_0, t_0 + T]$, and consider a discretisation of the time such that $t_n = t_0 + n\delta_t$ for $n = 0, ..., N$, and $\delta_t = T/N$ the time step. At each time step t_n (that for short we call n), the state of a water heater is described by a variable $X_n = (m_n, \theta_n) \in \{0, 1\} \times \mathbb{R}^+$, where m_n indicates the operating state of the heater (ON if 1, OFF if 0), and θ_n represents the average temperature of the water in the tank.

The evolution of the temperature in the next time step t_{n+1} is given by $\theta_{n+1} = \bar{T}_{t_{n+1}}^{t_n,m_n,\theta_n}$, where $t \mapsto \bar{T}_t^{t_n,m_n,\theta_n}$ is the solution of the ordinary differential equation (ODE) in Equation (22) on the interval $[t_n, t_{n+1}]$. This ODE models the impact of the heat loss to the environment temperature (T_{amb}), the Joule effect (heating) and water drains (hot water being withdrawn from the tanks for showers, taps, etc),

$$\begin{cases} \frac{dT(t)}{dt} = -\underbrace{\rho(T(t) - T_{\text{amb}})}_{\text{heat loss}} + \underbrace{\sigma m_n p_{\max}}_{\text{Joule effect}} - \underbrace{\tau(T(t) - T_{\text{in}})f(t)}_{\text{water drain}} \\ T(t_n) = \theta_n. \end{cases} \tag{22}$$

The parameters ρ, σ, τ are technical parameters of the water heater, $p_{\max}$ is the maximum power, T_{in} denotes the temperature of the cold water entering the tank, and $f(t)$ denotes the drain function.

The dynamics follow a cyclic ON/OFF decision rule with a deadband to ensure that the temperature is between a lower limit $T_{\min}$ and an upper limit $T_{\max}$. Thus, if the water heater is turned on, it heats water with the maximum capacity until its temperature exceeds $T_{\max}$. Then, the heater turns off. The water temperature then decreases until it reaches $T_{\min}$, then the heater turns on again and a new cycle begins. Therefore, the nominal dynamics at a discretized time is given by Equation (23) and is illustrated at Figure 1.

$$\begin{cases} \theta_{n+1} = \bar{T}_{t_{n+1}}^{t_n,m_n,\theta_n} \\ m_{n+1} = \begin{cases} m_n, & \text{if } \theta_{n+1} \in [T_{\min}, T_{\max}] \\ 0, & \text{if } \theta_{n+1} \geq T_{\max} \\ 1, & \text{if } \theta_{n+1} \leq T_{\min}. \end{cases} \end{cases} \tag{23}$$

Note that assuming the temperature set is finite prevents us from using the ODE on Equation (22) to compute the evolution of the mean temperature. In addition, we also have trouble computing the drain function $f(t)$, which in practice is not deterministic. Instead, we adapt this ODE to simplify our system. We start by making an Euler discretization of the ODE. We define a sequence $(d_n)_n$ denoting the amount of draining in liters at each time step. To decide whether hot water is drawn at each time step, we also consider a sequence $(\epsilon_n)_n$ of independent random variables following Bernoulli's laws of parameters $(q_n)_n$ respectively. The interest of having different parameters for each time step is to take into account the moments of the day when people are more inclined to use hot water (for taking a shower, doing the dishes, etc.). Assuming the existence of an independent water discharge at each time step is justified by assuming that the time frequency δ_t is large enough to contain all the time when hot water will be drawn from the water heater tank for a single use. In the interest of more realistic dynamics, we intend to weaken this assumption in future work. Therefore, we define

$$\theta'_{n+1} = \theta_n + \delta_t\big(- \rho\left(\theta_n - T_{\text{amb}}\right) + \sigma m_n p_{\max} - \epsilon_n \tau \left(\theta_n - T_{\text{in}}\right) d_n\big). \tag{24}$$

To tackle the finite-temperature state space problem, we assume that the space of possible temperatures Θ contains only integers from T_{amb} (the room temperature) up to $T_{\max}$, assuming that $T_{\text{amb}} < T_{\min}$ (it is reasonable to assume that the

ambient temperature is below the minimum temperature accepted for the heater). Given the dynamics of the operating state, θ_{n+1} never exceeds T_{max} (the heater turns off when it reaches T_{max} and when it is turned off, its temperature only decreases). On the other hand, drain may allow a temperature to be lower than T_{min}, but we assume that T_{amb} is small enough that the mean temperature is never lower than it. Therefore, we can take $\theta_{n+1} = \text{Round}(\theta'_{n+1})$, where

$$\text{Round}(\theta) = \begin{cases} \lfloor \theta \rfloor, & \text{if } B(\theta) = 0 \\ \lceil \theta \rceil, & \text{if } B(\theta) = 1, \end{cases}$$

and $B(\theta)$ is a random variable following a Bernoulli of parameter $\theta - \lfloor \theta \rfloor$. Thus, the closer θ is to its smallest nearest integer, the greater the probability that we approximate θ by it, and vice-versa. We perform stochastic rounding instead of deterministic to have an unbiased temperature estimator, i.e. $\mathbb{E}[\theta_{n+1}] = \theta'_{n+1}$.

D.2 Proof of convergence of Algorithm 1 applied to the water heater optimisation problem

Corollary D.1. *Consider the mean field problem of controlling the average power consumption profile of a population of water heaters in order to track a reference signal defined in the main paper as*

$$\min_{\mu \in \mathcal{M}_{\mu_0}} F(\mu)$$

where $F(\mu) := \sum_{n=1}^{N} f_n(\mu_n)$ with $f_n(\mu_n) := (\mu_n(\varphi) - \gamma_n)^2$, φ is the function defined in Equation (11)*, $\mathcal{M}_{\mu_0}$ is the set defined in Equation* (3)*, and $(\gamma_n)_{1 \leq n \leq N}$ is the target (recall that $\gamma_n \in [0, 1]$ for all $1 \leq n \leq N$).*

Algorithm 1 applied for K iterations to this minimisation problem with

$$\tau_k = \frac{\sqrt{2\Gamma(\mu^*, \mu^0)}}{2} \frac{1}{\sqrt{k}},$$

converges with rate

$$\min_{0 \leq s \leq K} F(\mu^s) - F(\mu^*) \leq 2 \frac{\sqrt{2\Gamma(\mu^*, \mu^0)}}{\sqrt{K}}.$$

Proof. The proof is a consequence of Theorem 4.3 applied to this problem. For that, we need to show that F is convex and Lipschitz with respect to the L_1 norm $\| \cdot \|_1$.

Convexity for all $n \leq N$, each f_n is given by

$$f_n(\mu_n) = \left(\sum_{x,a} \mu_n(x, a) \varphi(x) - \gamma_n \right)^2.$$

Let g be a real function such that $g(x) = (x - \gamma_n)^2$. The function g is convex and non-decreasing on $\mathbb{R}_+$.

Let $h : \mathbb{R}^{|\mathcal{X} \times \mathcal{A}|} \to \mathbb{R}$, such that $h(\mu_n) = \sum_{x,a} \mu_n(x, a) \varphi(x)$. Note that $\frac{\partial h}{\partial \mu_n(x,a)}(\mu_n) = \varphi(x)$. Thus, for any $\mu_n, \mu'_n \in \Delta_{\mathcal{X} \times \mathcal{A}}$,

$$h(\mu_n) - h(\mu'_n) = \left(\sum_{x,a} \big(\mu_n(x, a) - \mu'_n(x, a) \big) \varphi(x) \right) = \langle \nabla h(\mu'_n), \mu_n - \mu'_n \rangle,$$

therefore, the function h is also convex. As $f_n(\mu_n) = g(h(\mu_n))$, then f_n is convex as g and h are convex, and g is non decreasing in a univariate domain [Boyd and Vandenberghe, 2004]. Finally, as F is the sum of convex functions, then F is also convex.

Lipschitz As we already showed that F is convex, to show that F is Lipschitz with respect to the $\| \cdot \|_1$ norm, it suffices to show that the sup-norm $\| \cdot \|_\infty$ of F' is bounded (the sup-norm is the dual norm of the L_1 norm). This result can be found in Lemma 2.6 of Shalev-Shwartz [2012].

For any $\mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$,

$$\begin{aligned}
\|F'(\mu)\| &= \sup_{n\leq N,(x,a)\in\mathcal{X}\times\mathcal{A}} |f_n'(\mu_n)(x,a)| \\
&= 2 \sup_{n\leq N,(x,a)\in\mathcal{X}\times\mathcal{A}} |\mu_n(\varphi) - \gamma_n||\varphi(x)| \\
&= 2 \sup_{n\leq N,(x,a)\in\mathcal{X}\times\mathcal{A}} \left| \sum_{x',a'} \mu_n(x',a')\varphi(x') - \gamma_n \right| |\varphi(x)| \\
&= 2 \sup_{n\leq N,x\in\mathcal{X}} |\langle \rho_n, \varphi\rangle||\varphi(x)| \\
&\leq 2\|\varphi\|_\infty^2.
\end{aligned}$$

Thus, F is Lipschitz with respect to the L_1 norm with Lipschitz constant $L = 2\|\varphi\|_\infty^2$. In our particular case φ is bounded by 1 (see its definition in Equation (11)), hence $L = 2$.

We apply Theorem 4.3 with $L = 2$ to conclude.

□

D.3 Simulation of the nominal behavior of a water heater

Here we explain in details how the nominal dynamics are simulated in order to obtain the results in Section 6.

To simulate the nominal dynamics we use the nominal model presented in Equation (23) with the average temperature evolution introduced in Equation (24). To compute the sequences $(d_n)_n$ and $(q_n)_n$ regarding the amount of draining in liters and the probability of having a water withdrawal for each time step, respectively, we use data from the SMACH (*Simulation Multi-Agents des Comportements Humains*) platform [Albouys et al., 2019], which simulates power consumption of people in their homes separated by appliance. The data we use simulates the consumption of 5132 water heaters at a time step of one minute over a week in the summer of 2018.

Since we want a time step large enough to contain all the time that hot water will be drawn from the water heater tank for a single use, we take $\delta_t = 10$ minutes instead of one minute (as initially provided by the data). Therefore we transform the data to contain for each water heater the average discharge over each 10 minute interval. To compute d_n, we take the average discharge in liters over all water heaters with a water withdrawal during this time step. To calculate $(q_n)_n$, we calculate the percentage of water heaters with a water withdrawal over the entire population for each time step. The values of the parameters ρ, σ, τ and p_{max} are computed in Equation (25) using the variables introduced in Tables 1 and 2. We take $T_{\text{min}} = 50°C$, $T_{\text{max}} = 65°C$, $T_{\text{amb}} = 25°C$ and $T_{\text{in}} = 18°C$.

Table 1: Water heater intrinsic parameters.

Volume	0.2m^3
Height	1.37m
EI (thickness of isolation)	$\frac{0.035}{4}$m
p_{max}	$3600 * 2200$W (in one hour)

Table 2: Other parameters specifications to compute Equation 25.

denWater (water density)	1000 kg m^{-3}
capWater (water capacity)	4185 J kg^{-1} K^{-1}
CI (heat conductivity)	0.033 W/(m K)
coefLoss (loss coeff.)	$\frac{\text{CI}}{\text{EI}} * 2 * 3.14\sqrt{\frac{\text{vol}*3.14}{\text{height}}}$

$$\begin{aligned}
\rho &= \frac{\text{coefLoss} * 3600}{\text{capWater} * \text{denWater} * \text{vol/height}} \quad \text{(fraction of heat loss by hour)} \\
\sigma &= (\text{vol} * \text{denWater} * \text{capWater})^{-1} \\
\tau &= (\text{vol} * \text{denWater})^{-1}.
\end{aligned} \tag{25}$$

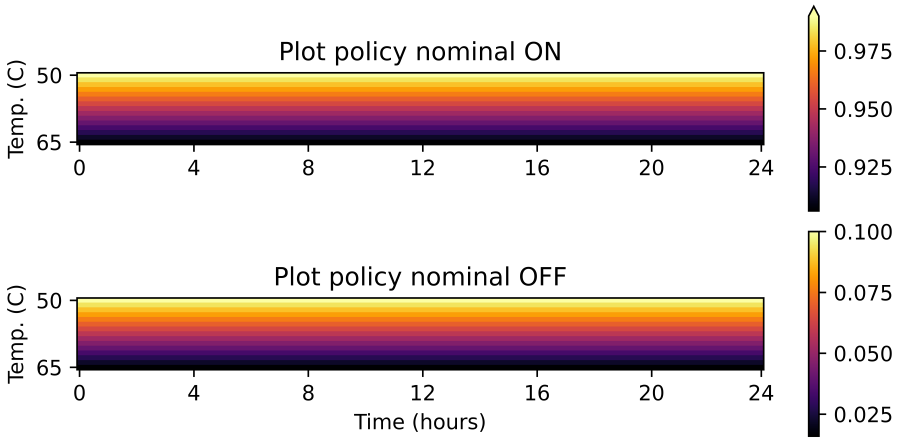


Figure 7: Initial policy sequence π^0 with a difference of 0.1 from the nominal policy.

E Discussion on different initialisation

We argue that given a state distribution sequence ρ, the policy generating this distribution is not necessarily unique. Moreover, as $\rho_n(x) = \sum_{a \in \mathcal{A}} \mu_n(x, a)$ for all $1 \leq n \leq N$ and for all $x \in \mathcal{X}$, the state-action distribution sequence μ with marginal ρ is not unique.

Recall that in the water heater application the iterative algorithms explored seek to find a policy sequence π inducing a population flow μ^π such that the distribution of heaters at state ON would be as close as possible to a reference signal $(\gamma_n)_{1 \leq n \leq N}$ at each time step:

$$f_n(\mu_n) := \left(\sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \mu_n(x, a) \varphi(x) - \gamma_n \right)^2$$

with φ defined in Equation (11). The non-uniqueness of the state distribution ρ given a policy π and a state-action distribution μ^π is particularly interesting when controlling the average consumption of a population of water heaters. In our model, we do not make any assumption on the number of ON/OFF switches that a water heater should have. However, this is an important constraint because a large number of switches can be detrimental to the device. The non-uniqueness of a policy allows us to formulate the problem of decreasing the number of switches as that of finding a policy that induces the right consumption while performing the smallest number of switches.

For example, suppose that at some time step n, half of the water heaters are ON and the other half are OFF. Suppose that at the next time step $n + 1$, the target indicates that we would like to still have half of the heaters to be ON and the other half to be OFF. A policy inducing this behavior could be constructed in several ways: we could have a probability 1 of keeping the heaters in the same state, or a probability 1 of turning ON the heaters that were OFF and turning OFF the heaters that were ON. Both policies result in the same proportion of heaters ON/OFF at the end. However, in the first case, no switching is done, while in the second case, all heaters must be switched.

In particular, different policy initialization on Algorithm 1 lead to different best policies. In the main body, results are computed by initializing each algorithm with the uniform policy over the action space for each state and time step. The average daily switch count is 33 in this case, while that the nominal dynamics (to keep ON when ON and to keep OFF when OFF within the temperature deadband $[T_{\min}, T_{\max}]$) only do in average 3 switches in a day. Figure 8 plots the probability of turning ON while in state ON/OFF for the policy computed by Algorithm 1 when the initial policy, illustrated at Figure 7, is a deviation of 0.1 of the nominal policy. In this case the number of switches decrease to a daily average of 9.2 and still seems to track the target curve. Further work involves continuing to explore how to find a policy that leads to the smallest number of switches while following the reference target.

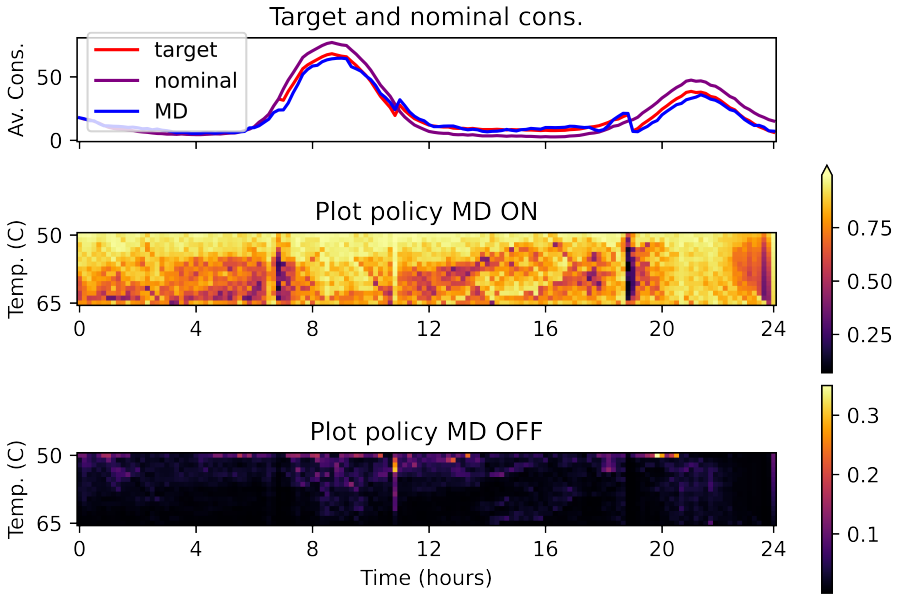


Figure 8: Output policy sequence of Algorithm 1 initialized with the policy at Figure 7