Reimagining Demand-Side Management with Mean Field Learning

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

Reimagining Demand-Side Management with Mean Field Learning

Bianca Marin Moreno
Inria[*], EDF R&D[†]

Margaux Brégère
Sorbonne Université[‡], EDF R&D[†]

Pierre Gaillard
Inria[*]

Nadia Oudjane
EDF R&D[†]

Abstract

Integrating renewable energy into the power grid while balancing supply and demand is a complex issue, given its intermittent nature. Demand side management (DSM) offers solutions to this challenge. We propose a new method for DSM, in particular the problem of controlling a large population of electrical devices to follow a desired consumption signal. We model it as a finite horizon Markovian mean field control problem. We develop a new algorithm, MD-MFC, which provides theoretical guarantees for convex and Lipschitz objective functions. What distinguishes MD-MFC from the existing load control literature is its effectiveness in directly solving the target tracking problem without resorting to regularization techniques on the main problem. A non-standard Bregman divergence on a mirror descent scheme allows dynamic programming to be used to obtain simple closed-form solutions. In addition, we show that general mean-field game algorithms can be applied to this problem, which expands the possibilities for addressing load control problems. We illustrate our claims with experiments on a realistic data set.

1 Introduction

Climate change is a complex problem, for which action takes many forms. Of the top 20 solutions identified by Foley et al. [2020] to reverse global warming, six are related to the energy sector, including integrating renewables into the electricity system and increasing the number of electric vehicles as a primary mode of transportation. In addition, the study managed by RTE [2022] showed that achieving carbon neutrality by 2050 in the French electricity scenario requires a decrease in final energy consumption and strong growth in renewable energy.

However, it is extremely difficult to make these solutions economically viable while also scaling them up. The intermittent nature of renewable energy sources can cause significant fluctuations in energy demand and supply, which may impact the balance of the power grid. Current solutions to keep the system in balance rely heavily on fossil fuel power plants, which have significant environmental costs, or on energy imports, which have capital and operating costs. Demand Side Management (DSM) are strategies to reduce energy acquisition costs and associated penalties by continuously monitoring energy consumption and managing devices [Bakare et al., 2023], which provides flexibility and improves the reliability of energy systems. Yet, implementing DSM solutions is challenging, as it involves large-scale data processing and near real-time scenarios. For this reason, machine learning solutions have recently emerged to solve DSM problems [Antonopoulos et al., 2020] with examples ranging from using multi-armed bandits to develop pricing solutions [Brégère et al., 2019], to deep learning models for smart charging of electric vehicles [López et al., 2019].

Preprint. Under review.

The goal of this paper is to make a new contribution to this field by proposing a new solution to a DSM problem concerning the control of thermostatically controled loads (TCLs: flexible appliances such as water heaters, air conditioners, refrigerators, etc). The aim is to control the aggregate power consumption of a large population of water heaters in order to follow a target consumption profile. To this end, we consider a finite time horizon Markovian mean field control (MFC) problem, and we propose a new algorithm based on mirror descent. We also adapt other mean field learning algorithms from the literature for this purpose.

Contributions In this paper, we propose and compare two new approaches to solve a DSM problem: a new algorithm, MD-MFC, for general Markovian MFC problems, and an adaptation of existing algorithms in the mean field learning literature for game problems. First we provide a modeling of the management system in question as a Markov decision process (MDP) in Section 2. The literature review of previous modeling and solutions to the load control problem, as well as a discussion of the main ingredient of our algorithms, mean field learning, are postponed to Subection 2.3. Our main results are stated in Section 3: we introduce the MD-MFC algorithm for a Markovian MFC problem, and we prove a convergence rate of order $1/\sqrt{K}$, where K is the number of iterations, by linking it to a mirror descent [Nemirovski and Yudin, 1983] scheme. This implies a non-trivial reformulation of a non-convex problem in a measure space into a convex problem. A good choice of non-standard regularization allows dynamic programming [Bertsekas, 2005]. This results in the first algorithm that efficiently and directly solves the target tracking problem without resorting to regularization techniques in the main problem. Section 4 illustrates the results with simulations based on a realistic data set [Albouys et al., 2019]. A series of future works concludes the paper.

2 Setting and model

Our framework consists in modeling the random dynamics of a population of water heaters in order to control their average consumption to follow a target signal. From now on, for a finite set S, we define Δ_S to be the simplex of dimension $|S|$, the cardinal of S.

2.1 Randomized controlled dynamics for one water heater

Let us consider a discretisation of the time for $n = 1, ..., N$. At each time step 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.

We call the uncontrolled dynamics the nominal dynamics. A water heater that follows the nominal dynamics [Bušić and Meyn, 2016] obeys 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.

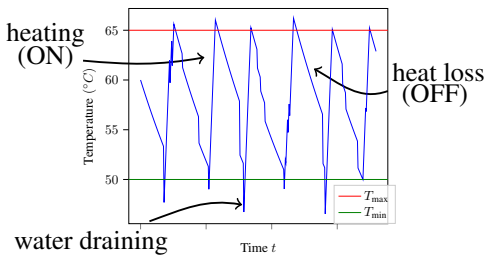


Figure 1: Temperature evolution of a water heater following the nominal dynamics.

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 $\theta_n \geq T_{\text{max}}$. Conversely, choosing action 0 means turning the heater off except when $\theta_n \leq T_{\text{min}}$. The nominal dynamics deterministically chooses action 0 if the heater is off and 1 if it is on, independent of the heater's temperature. 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 is computed by $\theta_{n+1} := T(m_n, \theta_n, \epsilon_n)$ where T is a function defined as the solution of Equation (16) of Appendix D.1 and ϵ_n is the random variable corresponding to a water-withdraw event at time step n. The action a_n is sampled with probability $\pi_n(\cdot|x_n) \in \Delta_{\mathcal{A}}$, and the next operating state is given by $m_{n+1} := M(a_n, \theta_{n+1})$, where

$$M(a_n, \theta_{n+1}) := a_n \mathbb{1}_{\theta_{n+1} \in [T_{\text{min}}, T_{\text{max}}]} + \mathbb{1}_{\theta_{n+1} < T_{\text{min}}}.$$

Hence, the probability kernel for each time step n is given by

$$p_{n+1}(x_{n+1}|x_n, a_n) := \mathbb{P}\big(\theta_{n+1} = T(m_n, \theta_n, \epsilon_n)|\theta_n, m_n\big)\mathbb{P}\big(m_{n+1} = M(a_n, \theta_{n+1})|a_n, \theta_{n+1}\big).$$

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

2.2 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 2.1. 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 $p_{\text{max}} = 1$ 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 divided by the number of devices). Our goal is to solve the problem

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

where we have chosen to work with a quadratic loss.

Let $\mu := (\mu_n)_{n \in [0,\dots,N]}$ such that μ_n is the state-action distribution of the entire population of heaters at time n. We denote by μ^π a state-action distribution sequence induced by a policy sequence π such as in Definition 2.1.

Definition 2.1 (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

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

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

$$\begin{aligned}
\varphi : \quad \mathcal{X} &\to \mathbb{R} \\
(m, \theta) &\mapsto m.
\end{aligned} \tag{2}$$

For such a function φ, when $M \to \infty$, the mean field approximation [Jabin and Wang, 2017] of Problem (1) consists in the main mean field control problem considered in the paper, and is given by

$$\min_{\pi} F(\mu^\pi) := \sum_{n=1}^{N} f_n(\mu_n^\pi), \tag{3}$$

where $f_n(\mu_n^\pi) := (\mu_n^\pi(\varphi) - \gamma_n)^2$.

It is important to mention that the algorithms and results presented in Section 3 for solving Problem (3) remain valid for any general finite horizon Markovian MFC problem with finite state and action spaces, where the cost functions f_n are convex and Lipschitz with respect to the $\|\cdot\|_1$ norm.

2.3 Literature discussion

Decision-making problems formulated as such mean-field models are a popular framework for 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 modeling 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]. Thus, although this paper focuses on a demand management problem, our results also provide a new approach to solving problems in many other areas.

Load control 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.

For water heater control, Cammardella et al. [2019] use a quadratic objective and a Kullback-Leibler (KL) penalty allowing a Lagrangian approach that learns both the control and the probability transition kernel, but cannot handle uncontrolled state parts, so uncertainties like water withdrawals must be modeled as deterministic. More recently, Cammardella et al. [2021] takes into account the uncontrolled stochastic environment in the KL quadratic control framework by adding constraints on the probability transition kernel. It is the KL penalty on the main problem that allows them to obtain their main results. However, there is a trade-off between adding the KL and obtaining a good target tracking curve. We therefore propose to solve directly the same quadratic control framework but without the KL penalty. We have successfully provided the first algorithm for direclty solving the target tracking problem.

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].

Lately, 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. Similarly, we show an equivalence between our MFC problem and potential games, that open up a new range of solutions to the DSM problem considered using the above studies.

3 Main results: algorithmic approaches

3.1 Building a new algorithm

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

$$\mathcal{M}_{\mu_0} := \Big\{ \mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N \Big| \sum_{a'\in\mathcal{A}} \mu_{n+1}(x', a') = \sum_{x\in\mathcal{X}, a\in\mathcal{A}} p_{n+1}(x'|x, a)\mu_n(x, a)\ , \forall x' \in \mathcal{X}, \forall n \in [0, ..., N] \Big\}. \tag{4}$$

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].

Proposition 3.1. *Let $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$. The application $\pi \mapsto \mu^\pi$ is a surjection from $(\Delta_\mathcal{A})^{\mathcal{X}\times N}$ to $\mathcal{M}_{\mu_0}$.*

The idea of the proof of Proposition 3.1, reported to Appendix A, 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 $\rho_n(x) := \sum_{a\in\mathcal{A}} \mu_n(x,a)$. Let $\mathcal{M}^*_{\mu_0}$ denotes the subset of $\mathcal{M}_{\mu_0}$ where the corresponding policies π are such that $\pi_n(a|x) \neq 0$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$ and $1 \leq n \leq N$. We define the regularization function $\Gamma : \mathcal{M}_{\mu_0} \times \mathcal{M}^*_{\mu_0} \to \mathbb{R}$ as

$$\Gamma(\mu^\pi, \mu^{\pi'}) := \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu^\pi{}_n(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi'_n(a|x)}\right)\right]. \tag{5}$$

Before giving a solution to Problem (3), we consider the following auxiliary optimization problem, which will later help us build the new algorithm. This iterative scheme is possible thanks to Proposition 3.1 which guarantees the existence of a strategy π for any $\mu \in \mathcal{M}_{\mu_0}$. Here, k represents an iteration:

$$\mu^{k+1} \in \underset{\mu^\pi\in\mathcal{M}_{\mu_0}}{\arg\min}\left\{\langle \nabla F(\mu^k), \mu^\pi\rangle + \frac{1}{\tau_k}\Gamma(\mu^\pi, \mu^k)\right\}. \tag{6}$$

We consider $\tau_k > 0$ and $\langle \nabla F(\mu^k), \mu^\pi\rangle := \sum_{n=1}^{N}\langle \nabla f_n(\mu^k_n), \mu^\pi_n\rangle$. At iteration $k+1$, we want to find μ^π by minimizing a linearization of F around μ^k, the distribution sequence found at the previous iteration, and at the same time penalizing the distance between π generating μ^π and π^k generating μ^k. Choosing this non-standard regularization Γ in Equation (5) instead of the traditional KL divergence on marginal state-action distributions is what enables us to obtain a simple closed-form solution for the iterative scheme. Later we show that Γ is a Bregman divergence. Thus, the use of Γ brings a significant improvement to the solution of MFC problems because it allows to obtain low complexity solutions with theoretical bounds, as we will prove later.

Let for all $(x_n, a_n, \mu_n) \in \mathcal{X}\times\mathcal{A}\times\Delta_{\mathcal{X}\times\mathcal{A}}$, $r_n(x_n, a_n, \mu_n) := -\nabla f_n(\mu_n)(x_n, a_n)$. We show in Theorem 3.2 that, due to the choice of penalizing strategies, the iterative scheme in Equation (6) can be solved through dynamic programming [Bertsekas, 2005] by building a Bellman recursion:

Theorem 3.2. *Let $k \geq 0$. The solution of Problem* (6) *is $\mu^{k+1} = \mu^{\pi^{k+1}}$ (as in Definition 2.1), where for all $1 \leq n \leq N$, and $(x,a) \in \mathcal{X}\times\mathcal{A}$,*

$$\pi^{k+1}_n(a|x) := \frac{\pi^k_n(a|x)\exp\left(\tau_k \tilde{Q}^k_n(x,a)\right)}{\sum_{a'\in\mathcal{A}}\pi^k_n(a'|x)\exp\left(\tau_k\tilde{Q}^k_n(x,a')\right)}, \tag{7}$$

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

$$\begin{cases} \tilde{Q}^k_N(x,a) = r_N(x,a,\mu^k_N) \\ \tilde{Q}^k_n(x,a) = \max\limits_{\pi_{n+1}\in(\Delta_\mathcal{A})^\mathcal{X}}\Bigg\{ r_n(x,a,\mu^k_n) + \sum\limits_{x'} p_{n+1}(x'|x,a) \\ \qquad\qquad \sum\limits_{a'}\pi_{n+1}(a'|x')\left[-\frac{1}{\tau_k}\log\left(\frac{\pi_{n+1}(a'|x')}{\pi^k_{n+1}(a'|x')}\right) + \tilde{Q}^k_{n+1}(x',a')\right]\Bigg\}, \quad \forall 1 \leq n \leq N. \end{cases} \tag{8}$$

Proof. See Appendix B.1. □

Notice that the value $\pi_{n+1} \in (\Delta_\mathcal{A})^\mathcal{X}$ maximizing the equation to find $\tilde{Q}^k_n$ in the Recursion (8) is given by π^{k+1}_{n+1}. We can then build the MD-MFC method in Algorithm 1. Note that Algorithm 1 is well defined because the policy update in Equation (7) ensures that each iteration remains in $\mathcal{M}^*_{\mu_0}$.

3.2 Convergence properties of the algorithm

We present a result on the convergence rate of Algorithm 1.

Algorithm 1 MD-MFC

Input: number of iterations K, initial sequence of policies $\pi^0 \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ such that $\mu^0 := \mu^{\pi_0} \in \mathcal{M}^*_{\mu_0}$, initial state-action distribution μ_0 (always fixed), sequence of non-negative learning rates $(\tau_k)_{k\leq K}$.
for $k = 0, ..., K-1$ **do**
  $\mu^k = \mu^{\pi^k}$ as in Definition 2.1.
  $\tilde{Q}^k_N(x,a) = r_N(x,a,\mu^k_N)$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$.
  for $n = N, ..., 1$ **do**
    $\forall (x,a) \in \mathcal{X}\times\mathcal{A}$:
    $\pi^{k+1}_n(a|x) = \frac{\pi^k_n(a|x)\exp\left(\tau_k \tilde{Q}^k_n(x,a)\right)}{\sum_{a'}\pi^k_n(a'|x)\exp\left(\tau_k\tilde{Q}^k_n(x,a')\right)}$.
    $\tilde{Q}^k_{n-1}(x,a)$ using the recursion in Equation (8).
  end for
end for
return π^K

Theorem 3.3. *Let π^* be a minimizer of Problem* (3)*. Applying K iterations of Algorithm 1 to this problem, with, for each $1 \leq k \leq K$,*

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

gives the following convergence rate

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

Proof. The proof consists in showing that Algorithm 1 is a mirror descent scheme applied to

$$\min_{\mu\in\mathcal{M}_{\mu_0}} F(\mu), \tag{9}$$

that is equivalent to Problem (3) as a direct consequence of Proposition (3.1), and that the new Problem (9) does satisfy the necessary hypothesis for mirror descent convergence [Beck and Teboulle, 2003] with a non-standard Bregman divergence. The strength of this result is showing that the complex non-convex Problem (3) can be solved using a classical optimization algorithm, which, with the right choice of regularizer, has an efficient solution thanks to dynamic programming.

Let us start by showing that Γ is indeed a Bregman divergence. For ease of notation, for any probability measure $\eta \in \Delta_E$, whatever the (finite) space E, we introduce the neg-entropy function, with the convention that $0\log(0) = 0$,

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

Proposition 3.4. *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 Γ is a Bregman divergence induced by the function*

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

Also, Γ is 1-strongly convex with respect to the $\sup_{1\leq n\leq N}\|\cdot\|_1$ norm.

The proof is in Appendix B.2 and consists in showing and exploring that the Γ divergence taking values on the marginal state-action distributions is in fact the KL divergence on the joint distribution.

Next, if f_n is convex and l_n Lipschitz with respect to the norm $\|\cdot\|_1$ for any $1 \leq n \leq N$, then F is also convex and Lipschitz with constant $L := (\sum_{n=1}^N l_n^2)^{1/2}$ (see Appendix B.3). The proof that our cost functions for the DSM model satisfy these assumptions is given in Appendix D.2. Since the set $\mathcal{M}_{\mu_0}$ is convex, we also satisfy the convexity assumptions for the convergence of the mirror descent. The rate of convergence is thus a direct consequence of the application of the proof of convergence of mirror descent for Problem (9). □

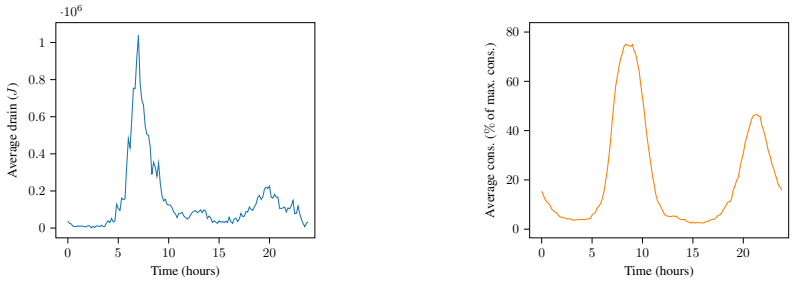
3.3 Potential games

In Appendix E we provide an equivalence between the MFC problem considered and a MFG by considering a game whose reward is given by $r_n(x_n, a_n, \mu_n) := -\nabla f_n(\mu_n)(x_n, a_n)$ for all $(x_n, a_n, \mu_n) \in \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}\times\mathcal{A}}$. We call this type of game a potential game. The work done in Geist et al. [2022] and Bonnans et al. [2021] relates the optimality conditions of optimization problems to the concept of Nash equilibrium in game problems. We do not go into details here, and leave more in-depth discussions to the Appendix section. The main purpose of this section is the realization that we can apply MFG algorithms to the DSM problem, which is a major breakthrough in this area because it opens up a new range of algorithms to this type of management system problems.

4 Experiments

4.1 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.

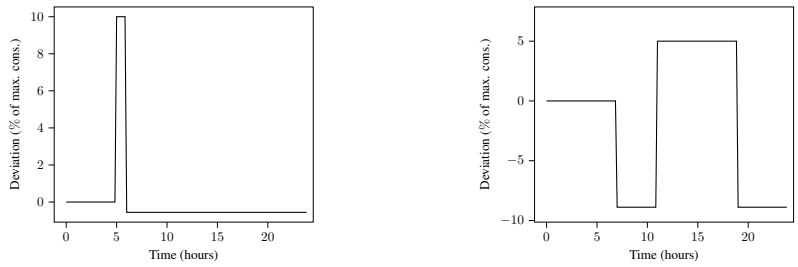


(a) Average drain in Joules

(b) Average consumption

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

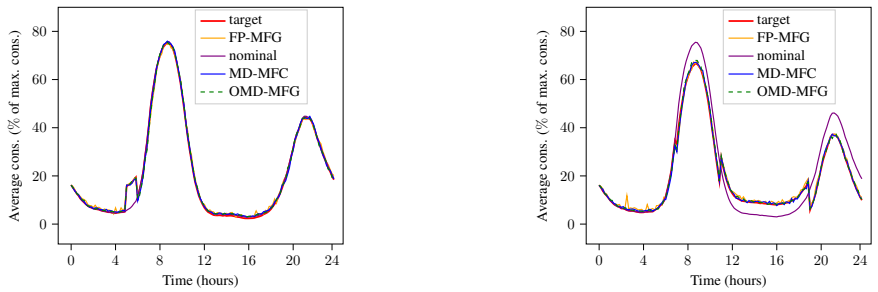
The target signal $\gamma = (\gamma_n)_{1\leq n\leq N}$ is built as a sum of a baseline $b = (b_n)_{n\leq N}$ and a deviation signal $\lambda = (\lambda_n)_n$, $\gamma = \lambda + b(w)$, where $b(w)$ is the nominal dynamics obtained by simulating the water heaters (as in Figure 2), and w represents a 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.



(a) One hour step.

(b) Eight hours step.

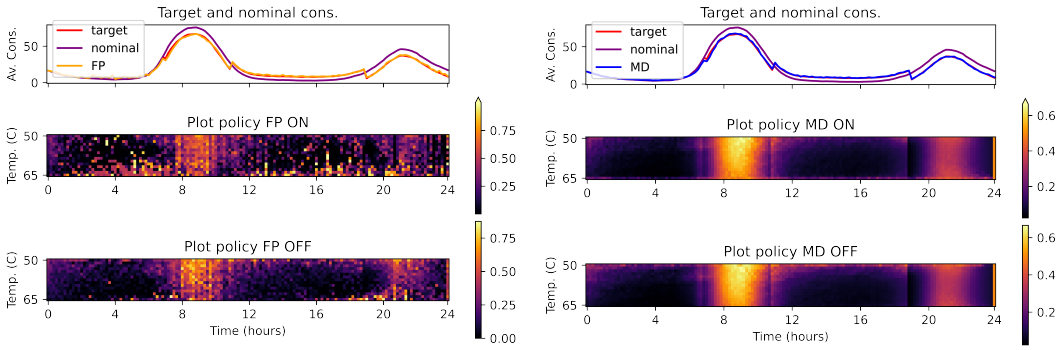
Figure 3: Deviation signals $(\lambda_n)_{n\leq N}$



(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 for the optimal policy computed through different algorithms, for targets constructed with the deviations of one hour [left] and eight hours [right]. We compare with the nominal consumption (without deviation).



(a) Policy FP-MFG.

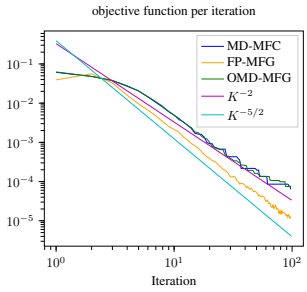
(b) Policy MD-MFC.

Figure 5: [top] Target, average consumption obtained by the nominal policy and by the policy computed by FP-MFG (left) and MD-MFC (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 $T_{\text{min}} = 50$ and $T_{\text{max}} = 65$ [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.

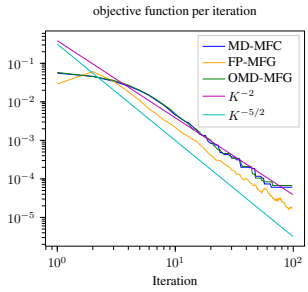
4.2 Results

For a population of water heaters following the randomized dynamics we compare the optimal policy sequence obtained after 100 iterations of MD-MFC, and two mean field game algorithms: Fictitious Play for MFG (FP-MFG) from Perrin et al. [2020] and Online Mirror Descent for MFG (OMD-MFG) from Pérolat et al. [2022] (see Algorithms 2 and 4 respectively in Appendix C). At each iteration, we compute a policy sequence of size 144 (number of time steps). The heater's state space $\mathcal{X}$ is of size $2 * 41$ (two ON/OFF operating states times 41 possible temperatures - integers from the ambient 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 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. The three algorithms have a memory complexity of order $N \times |\mathcal{X}| \times |\mathcal{A}|$, and a computational complexity of order $K \times N \times (|\mathcal{X}| \times |\mathcal{A}|)^2$.

In Figure 4, the consumption simulated by the best policies for all three algorithms appears to track the target better than the nominal consumption. This is not a surprise because all algorithms are supposed to converge to the same minima. However, they do so by finding different strategies and with different convergence rates. Figure 6 shows the logarithm of the objective function per iteration, and to visualize the policies obtained 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 that MD-MFC returns a more regular policy than FP-MFG.

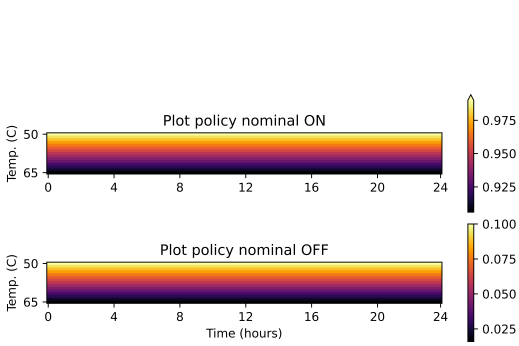


(a) One hour step.

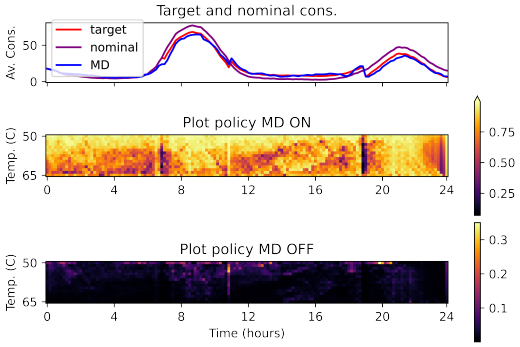


(b) Eight hours step.

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.



(a) Nominal policy deviation.



(b) Policy MD-MFC with nominal policy deviation.

Figure 7: [left] Initial policy sequence π^0 with a deviation of 0.1 from the nominal policy. [right] Output policy sequence of Algorithm 1 initialized with the policy at left.

Different initialisations impact the number of switches We noticed that different initialisations of MD-MFC lead to different policies. Given a state distribution sequence ρ, the policy generating this distribution is not necessarily unique. In particular, these policies, while providing the same ρ, may differ in terms of the average number of ON/OFF switches induced over the time horizon considered. In our model, no switching limit is assumed, but a large number of switches can be detrimental to the device. This non-uniqueness helps us reduce switch count without adding new constraints by finding multiple policies that achieve the right consumption and selecting the one with the fewest switches. This can also be useful for MFC problems in other areas, e.g. transaction costs in finance.

In the case illustrated here the average number of daily switches is 33, while the nominal dynamic averages only 3 switches per day. By initializing the MD-MFC algorithm with a policy that is a 0.1 deviation from the nominal policy as in Figure 7a, we find that the number of switches decreases to a daily average of 9.2 while still following the target curve, see Figure 7b. The same does not happen with FP-MFG, which makes it less interesting for the real-world scenarios we consider here.

Finally, Table 1 gives a global comparison between the three algorithms. FP-MFG converges faster but is not suitable for controlling switch count, being less interesting for the considered DSM problem, and needs a smooth objective function assumption. OMD-MFG is empirically as good as MD-MFC but lacks convergence proof for discrete cases.

5 Future work

Future work involves adapting existing algorithms to real-time algorithms, proposing schemes where each iteration corresponds to a time step. We further aim to generalize to a model-free scenario, learning user behavior on the fly while preserving privacy with partially observable states. Moreover, we believe we can extend our approach to the accelerated version of mirror descent [Krichene et al., 2015] providing a better theoretical convergence rate of order $1/K^2$.

Table 1: Comparing MD-MFC, OMD-MFG and FP-MFG

Algorithm	Convergence rate	Flexibility on applications	Convergence hypothesis
MD-MFC	$K^{-1/2}$	Yes (switches)	convex + Lispschitz
OMD-MFG	no proof	Yes (switches)	convex + Lispschitz
FP-MFG	K^{-1}	No	convex + Lipschitz + smooth

References

Jonathan Foley, Katherine Wilkinson, Chad Frischmann, Ryan Allard, João Gouveia, Kevin Bayuk, Mamta Mehra, Eric Toensmeier, Chris Forest, Tala Daya, Denton Gentry, Sarah Myhre, s. Karthik Mukkavilli, Abdulmutalib Yussuff, Ashok Mangotra, Phil Metz, Ariani Wartenberg, Chirjiv Anand, Marzieh Jafary, and Barbara Rodriguez. *The Drawdown Review (2020) - Climate Solutions for a New Decade*. 03 2020. doi: 10.13140/RG.2.2.31794.76487.

RTE. Futurs énergétiques 2050. Technical report, Le réseaux de transport d'électricité, 2022.

Mutiu Shola Bakare, Abubakar Abdulkarim, Mohammad Zeeshan, and Aliyu Nuhu Shuaibu. A comprehensive overview on demand side energy management towards smart grids: challenges, solutions, and future direction. *Energy Informatics*, 6(1):4, March 2023. ISSN 2520-8942. doi: 10.1186/s42162-023-00262-7. URL `https://doi.org/10.1186/s42162-023-00262-7`.

Ioannis Antonopoulos, Valentin Robu, Benoit Couraud, Desen Kirli, Sonam Norbu, Aristides Kiprakis, David Flynn, Sergio Elizondo-Gonzalez, and Steve Wattam. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, 130(C), 2020. URL `https://ideas.repec.org/a/eee/rensus/v130y2020ics136403212030191x.html`.

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`.

Karol Lina López, Christian Gagné, and Marc-André Gardner. Demand-side management using deep learning for smart charging of electric vehicles. *IEEE Transactions on Smart Grid*, 10(3): 2683–2691, 2019. doi: 10.1109/TSG.2018.2808247.

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

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

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.

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`.

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.

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.

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.

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.

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.

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.

J. Frédéric Bonnans, Pierre Lavigne, and Laurent Pfeiffer. Discrete potential mean field games: duality and numerical resolution, 2021. URL https://arxiv.org/abs/2106.07463.

Walid Krichene, Alexandre M. Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2845–2853, Cambridge, MA, USA, 2015. MIT Press.

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.

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.

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.

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.

A Missing proofs

A.1 Proof of Proposition 3.1

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). First, let us deal with the case where $\rho_n(x) \neq 0$. 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 $(x',a') \in \mathcal{X}\times\mathcal{A}$

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

where the first equality comes from Definition 2.1, the second equality comes from the induction assumption and the way we defined the strategy π, and the third comes from the assumption that $\mu \in \mathcal{M}_{\mu_0}$.

In the case $\rho_n(x) = 0$, we therefore have $\mu_n(x,a) = 0$ for all $a \in \mathcal{A}$, so any choice of $\pi_n(a|x)$ would work.

□

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, with the convention $0\log(0) = 0$,

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

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

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

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

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

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 $\mathcal{P}(\mathcal{X}\times\mathcal{A})^N$ 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

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

when $\rho_n(x) \neq 0$, otherwise we fix an arbitrary strategy $\pi_n(a|x) = \frac{1}{|\mathcal{A}|}$.

Before proving Theorems (3.2) and (3.3) we state and prove a Lemma which is key to proving both theorems.

Lemma B.1. *For any $\mu \in \mathcal{M}_{\mu_0}$ and $\mu' \in \mathcal{M}^*_{\mu_0}$, with associated 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*

$$\begin{aligned} D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu_n(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi'_n(a|x)}\right)\right] \\ &= \sum_{n=1}^{N} D(\mu_n, \mu'_n) - \sum_{n=0}^{N} D(\rho_n, \rho'_n) \end{aligned} \tag{10}$$

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 $a, a' \in \mathcal{A}$,

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

Given Definition 2.1, 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:

$$\mu_{1:N}(\vec{y}) = \mu_0(y_0)P^{\pi_1}(y_1|y_0)...P^{\pi_N}(y_N|y_{N-1}), \tag{11}$$

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

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

Where

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

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

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

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 3.1 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 3.2: formulation of Algorithm 1

Proof. At each iteration we seek to solve

$$\mu^{k+1} \in \operatorname*{arg\,min}_{\mu^\pi \in \mathcal{M}_{\mu_0}} \left\{ \langle \nabla F(\mu^k), \mu^\pi \rangle + \frac{1}{\tau_k} \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu_n(\cdot)} \left[\log\left(\frac{\pi_n(a|x)}{\pi_n^k(a|x)}\right) \right] \right\} \tag{12}$$

where recall that $\langle \nabla F(\mu^k), \mu^\pi \rangle := \sum_{n=1}^{N} \langle \nabla f_n(\mu_n^k), \mu_n^\pi \rangle$. We further use that $r_n(x_n, a_n, \mu_n) := -\nabla f_n(\mu_n)(x_n, a_n)$.

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 3.1), allows us to reformulate Problem (12) on $\mathcal{M}_{\mu_0}$ into a problem on $(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$, thus

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

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

$$
\begin{aligned}
\tilde{Q}_i^k(x,a) = \max_{\pi_{i+1:N} \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N-i}} \mathbb{E}_{\pi} \Bigg[& r_N(x_N,a_N,\mu_N^k) + \sum_{n=i}^{N-1} \Bigg\{ r_n(x_n,a_n,\mu_n^k) \\
& - \frac{1}{\tau_k} \sum_{x',a'} p_{n+1}(x'|x_n,a_n)\pi_{n+1}(a'|x') \log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) \Bigg\} \Bigg| (x_i,a_i) = (x,a) \Bigg],
\end{aligned} \tag{13}
$$

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

First, note that $\mathbb{E}_{(x,a)\sim\mu_0(\cdot)}[\tilde{Q}_0^k(x,a)] = (12)$. Moreover, the optimality principle states that this regularized state-action value function satisfies the following recursion

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

Thus, to solve (12) 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 KL 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

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

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 3.4

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 $1 \le n \le N$, 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 $1 \le n \le N$, 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,

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

where (a) comes from the definition of the KL 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].

Now we prove that Γ is 1-strongly convex with respect to the $\sup_{1\leq n\leq N} \|\cdot\|_1$ norm. By Lemma B.1,

$$\begin{aligned}
\Gamma(\mu, \mu') &= \sum_{n=1}^N D(\mu_n, \mu'_n) - \sum_{n=1}^N D(\rho_n, \rho'_n) \\
&= D(\mu_{1:N}, \mu'_{1:N}) \\
&\geq 2\|\mu_{1:N} - \mu_{1:N}\|_{\text{TV}}^2 \\
&= \frac{1}{2}\|\mu_{1:N} - \mu'_{1:N}\|_1^2,
\end{aligned}$$

the last inequality being a consequence of Pinsker's inequality. The norm $\|\cdot\|_{\text{TV}}$ stands for the total variation norm. Let y represent an element of $(\mathcal{X}\times\mathcal{A})^{N+1}$ such that $y_i \in \mathcal{X}\times\mathcal{A}$ for all $1 \leq i \leq N$. Observe that

$$\begin{aligned}
\|\mu_{1:N} - \mu'_{1:N}\|_1 &= \sum_{y \in (\mathcal{X}\times\mathcal{A})^{N+1}} |\mu_{1:N}(y) - \mu'_{1:N}(y)| \\
&\geq \sum_{y_n \in \mathcal{X}\times\mathcal{A}} \left| \sum_{y_s \in \mathcal{X}\times\mathcal{A},\, s\neq n} \big(\mu_{1:N}(y) - \mu'_{1:N}(y)\big) \right| \\
&= \sum_{y_n \in \mathcal{X}\times\mathcal{A}} |\mu_n(y_n) - \mu'_n(y_n)| \quad \text{for all } n \in \{1, \cdots, N\}.
\end{aligned}$$

In particular,

$$\|\mu_{1:N} - \mu'_{1:N}\|_1 \geq \sup_{1\leq n\leq N} \|\mu_n - \mu'_n\|_1 .$$

This implies that

$$\Gamma(\mu, \mu') \geq \frac{1}{2} \sup_{1\leq n\leq N} \|\mu_n - \mu'_n\|_1^2 ,$$

proving that Γ is 1-strongly convex with respect to the $\sup_{1\leq n\leq N} \|\cdot\|_1$ norm.

□

B.3 Complements of the proof of Theorem 3.3

Proof. Here we prove that if $(f_n)_{1\leq n\leq N}$ are convex and Lipschitz with respect to the L1-norm, then so is F. **Convexity:** F is convex as the sum of convex functions.

Lipschitz: Let $\mu, \mu' \in (\mathcal{X} \times \mathcal{A})^N$. As f_n is Lipschitz with respect to $\|\cdot\|_1$ with constant l_n, then $|f_n(\mu_n) - f_n(\mu'_n)| \leq l_n \|\mu_n - \mu_n\|_1$ for all $1 \leq n \leq N$. Therefore,

$$\begin{aligned}
|F(\mu) - F(\mu')| &= \left| \sum_{n=1}^{N} f_n(\mu_n) - f_n(\mu'_n) \right| \\
&\leq \sum_{n=1}^{N} |f_n(\mu_n) - f_n(\mu'_n)| \\
&\leq \sum_{n=1}^{N} l_n \|\mu_n - \mu'_n\|_1 \\
&\leq \left(\sum_{n=1}^{N} l_n^2 \right)^{1/2} \left(\sum_{n=1}^{N} \|\mu_n - \mu'_n\|_1^2 \right)^{1/2} \\
&\leq L \|\mu - \mu'\|_1,
\end{aligned}$$

where we use Cauchy-Schwarz in the second to last inequality. Therefore, F is Lipschitz with respect to the L1-norm 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 2.1.
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

$$Q_n^{\pi,\mu}(x, a) := \mathbb{E}_\pi \left[\sum_{i=n}^{N} r_i(x_i, a_i, \mu_i) \middle| x_n = x, a_n = a \right]. \tag{14}$$

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 $n < N$, $(x, a) \in \mathcal{X} \times \mathcal{A}$:

$$\begin{cases} Q_N^{\pi,\mu}(x, a) = r_N(x, a, \mu_N) \\ Q_n^{\pi,\mu}(x, a) = r_n(x, a, \mu_n) + \sum_{x'} p(x'|x, a) \sum_{a'} \pi_{n+1}(a'|x') Q_{n+1}^{\pi,\mu}(x', a'). \end{cases} \tag{15}$$

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 2.1.
  $Q^k := Q^{\pi^k, \mu^k}$ as in Equation (15).
  $\pi_n^{k+1}(\cdot|x) := \arg\max_{\pi(\cdot|x)\in\Delta_{\mathcal{A}}} \langle Q_n^k(x,\cdot), \pi(\cdot|x)\rangle + \tau D\left(\pi(\cdot|x), \pi_n^k(\cdot|x)\right), \forall x \in \mathcal{X}, \forall n \leq N.$
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 frequency. 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 (16) 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),

$$\begin{cases} \frac{dT(t)}{dt} = -\underbrace{\rho(T(t) - T_{\text{amb}})}_{\text{heat loss}} + \underbrace{\sigma m_n p_{\max}}_{\text{Joule effect}} - \underbrace{\tau(T(t) - T_{\text{in}}) f(t)}_{\text{water drain}} \\ T(t_n) = \theta_n. \end{cases} \tag{16}$$

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 (17) and is illustrated at Figure 1.

$$\begin{cases} \theta_{n+1} = \bar{T}_{t_{n+1}}^{t_n, m_n, \theta_n} \\ m_{n+1} = \begin{cases} m_n, & \text{if } \theta_{n+1} \in [T_{\min}, T_{\max}] \\ 0, & \text{if } \theta_{n+1} \geq T_{\max} \\ 1, & \text{if } \theta_{n+1} \leq T_{\min}. \end{cases} \end{cases} \tag{17}$$

Note that assuming the temperature set is finite prevents us from using the ODE on Equation (16) 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

$$\theta'_{n+1} = \theta_n + \delta_t\big(- \rho\left(\theta_n - T_{\text{amb}}\right) + \sigma m_n p_{\max} - \epsilon_n \tau \left(\theta_n - T_{\text{in}}\right) d_n\big). \tag{18}$$

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 $T_{\text{amb}} < T_{\text{min}}$ (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 Complements of the proof of Theorem 3.3 for the DSM problem

We show that the cost function considered in Problem (3) concerning the water heater optimisation problem 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].

Lipschitz As f_n is convex for all $1 \leq n \leq N$, to show that it is Lipschitz with respect to the $\|\cdot\|_1$ norm, it suffices to show that the sup-norm $\|\cdot\|_\infty$ of ∇f_n 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 $\mu_n \in \Delta_{\mathcal{X}\times\mathcal{A}}$,

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

Thus, f_n is Lipschitz with respect to the L_1 norm with Lipschitz constant $l_n = 2\|\varphi\|_\infty^2$. In our particular case φ is bounded by 1 (see its definition in Equation (2)), hence $l_n = 2$ for all $1 \leq n \leq N$.

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 4.2.

To simulate the nominal dynamics we use the nominal model presented in Equation (17) with the average temperature evolution introduced in Equation (18). 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 (19) using the variables introduced in Tables 2 and 3. We take $T_{\text{min}} = 50°C$, $T_{\text{max}} = 65°C$, $T_{\text{amb}} = 25°C$ and $T_{\text{in}} = 18°C$.

Table 2: Water heater intrinsic parameters.

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

Table 3: Other parameters specifications to compute Equation 19.

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.)	$\frac{\text{CI}}{\text{EI}} * 2 * 3.14\sqrt{\frac{\text{vol}*3.14}{\text{height}}}$

$$
\begin{aligned}
\rho &= \frac{\text{coefLoss} * 3600}{\text{capWater} * \text{denWater} * \text{vol/height}} \quad \text{(fraction of heat loss by hour)} \\
\sigma &= (\text{vol} * \text{denWater} * \text{capWater})^{-1} \\
\tau &= (\text{vol} * \text{denWater})^{-1}.
\end{aligned} \tag{19}
$$

E Potential games discussion

In Subsection 3.3 we mention an equivalence between the control Problem (3) and a game problem in order to be able to compare Algorithm 1 with learning algorithms for MFG in the literature. For this, we use results similar to those of Geist et al. [2022] and we refer to it for further definitions on a MFG structure and the notion of Nash equilibrium (NE).

In a mean field game problem, 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,

$$
J_{\mu_0}(\pi, \mu) := \mathbb{E}_\pi \left[\sum_{n=1}^{N} r_n(x_n, a_n, \mu_n) \right]. \tag{20}
$$

Let us define a game with the same transition probability p, and with reward defined as

$$r_n(x_n, a_n, \mu_n) := -\nabla f_n(\mu_n)(x_n, a_n) \tag{21}$$

for all $(x_n, a_n, \mu_n) \in \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}\times\mathcal{A}}$.

Proposition E.1. *The strategy π^* is a minimizer of Problem* (3) *if and only if, (μ^{π^*}, π^*) is a NE of the MFG defined with reward as in Equation* (21)*. Furthermore, this game is monotone (and strictly monotone if f_n is strictly convex for all $1 \leq n \leq N$. See Definition E.2).*

This Proposition connects the optimality conditions of Problem (3) 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.

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 (3) satisfying the optimality conditions. Also, Proposition 3.1 shows, for a fixed initial state-action distribution μ_0, a surjection 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.

By definition, a Nash equilibrium (μ^*, π^*) satisfies $\pi^* = \arg\max_\pi J(\pi, \mu^*)$. In other words,

$$J(\pi^*, \mu^{\pi^*}) \geq J(\pi, \mu^{\pi^*}) \quad \forall \pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}. \tag{22}$$

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 (22) entails

$$\sum_{n=1}^{N} \left\langle \nabla f_n(\mu_n^{\pi^*}), \mu_n^{\pi^*} - \mu_n^{\pi} \right\rangle \leq 0. \tag{23}$$

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

$$\sum_{n=1}^{N} f_n(\mu_n^{\pi^*}) - f_n(\mu_n^{\pi}) \leq 0. \tag{24}$$

Thus, π^* satisfies the optimality conditions of Problem (3). We then proved that if (π^*, μ^*) is a NE with $\mu^* = \mu^{\pi^*}$, then π^* is an optimum of Problem (3).

On the other way around, if π^* is a minimizer of Problem (3) then it satisfies Inequality (24) 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 (23). Following the same calculations backwards, we obtain that π^* then satisfies Inequality (22), 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 E.2 (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. □