(Online) Convex Optimization for Demand-Side Management: Application to Thermostatically Controlled Loads

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

(Online) Convex Optimization for Demand-Side Management: Application to Thermostatically Controlled Loads

Bianca Marin Moreno[1,2,3*], Margaux Brégère[2,4], Pierre Gaillard[1], Nadia Oudjane[2,3]

[1]Univ. Grenoble Alpes, INRIA, CNRS, Grenoble INP, LJK, Grenoble, 38000, France.
[2]EDF R&D, Osiris, Palaiseau, 91200, France.
[3]FiME, Laboratoire de Finance des Marchés de l'Energie, Dauphine, CREST, EDF R&D, Paris, France.
[4] Université de Paris, Sorbonne Université, CNRS, LPSM, Paris, France.

*Corresponding author(s). E-mail(s): bianca.marin-moreno@inria.fr;
Contributing authors: margaux.bregere@edf.fr; pierre.gaillard@inria.fr; nadia.oudjane@edf.fr;

Abstract

To counter the challenge of integrating fluctuating renewables into the grid, devices like thermostatically controlled loads (water-heaters, air conditioners, etc) offer flexible demand. However, efficiently controlling a large population of these devices to track desired consumption signals remains a complex challenge. Existing methods lack convergence guarantees and computational efficiency, or resort to regularization techniques instead of tackling the target tracking problem directly. This work addresses these drawbacks. We propose to model the problem as a finite horizon episodic Markov decision process, enabling us to adapt convex optimization algorithms with convergence guarantees and computational efficiency. This framework also extends to online learning scenarios, where daily control decisions are made without prior knowledge of consumer behavior and with daily-changing target profiles due to fluctuations of energy production and inflexible consumption. We introduce a new algorithm, called Online Target Tracker (OTT), the first online learning load control method, for which we prove sub-linear regret. We demonstrate our claims with realistic experiments. This combination of optimization and learning lays the groundwork for more dynamic and efficient load control methods.

Keywords: Thermostatically controlled loads, Online learning, Convex optimization, Markov decision process

1 Introduction

The fight against climate change heavily relies on our capability to completely decarbonize the electricity supply, considering that the energy sector is the primary source of greenhouse gas emissions into the atmosphere [1]. However, it is extremely difficult to make these solutions economically viable while also scaling them up. As large-scale electricity storage is costly and relies on inefficient systems, it is crucial to maintain a strict balance between electricity supply and demand at all times. The intermittent nature of renewable energy sources can cause significant fluctuations in energy 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.

To address these issues, Demand-Side Management (DSM) [2] offers a set of strategies that utilities use to control electricity demand by continuously monitoring energy consumption and managing devices, reducing energy acquisition costs and improving 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 [3] with examples ranging from using multi-armed bandits to develop pricing solutions [4], to deep learning models for smart charging of electric vehicles [5].

This paper presents a new approach using convex optimization and learning to exploit the potential of thermostatically controlled loads (TCLs) as a flexible energy storage solution to mitigate the impact of intermittent renewable energy sources. Thermostatic loads are electrically powered devices that regulate temperature within a specified range, like air conditioners, heaters, refrigerators and freezers. Our main objective is to design signals to control the on-off states of a large population of TCLs, ensuring that their combined energy consumption follows a predetermined target consumption profile. We present three new contributions to the field of load control:

- Using a Markov decision process (MDP) framework, we effectively model the load control problem. This allows us to apply existing convex optimization algorithms like Mirror Descent with a specialized regularization function [6] and Fictitious play for mean field games [7] to solve the TCL control problem. These algorithms provide closed-form solutions and do not use regularization techniques, so their convergence results apply directly to the load problem. This is a novelty compared to existing approaches [8]. In addition, their convergence guarantees are $O(1/K)$, where K represents the number of iterations.
- We introduce an original algorithm, called Online Target Tracker (OTT), when the target consumption profile is allowed to change daily due to variations in energy supply and inflexible consumption, and when consumer behavior is not known in advance and must therefore be learned. OTT is the first online learning algorithm for the load control problem. The regret of OTT can be decomposed

into two components. The first term, reflecting the uncertainty in consumer behavior, scales as $\tilde{O}(\sqrt{T})$, where T is the number of days. The second term, accounting for daily variations in the target consumption profile, requires novel analysis. We demonstrate that OTT achieves a regret bound of $O(\log(T))$ for the dynamic target term, improving on the traditional online learning bounds of $\tilde{O}(\sqrt{T})$ [9] using optimism [10] and specific characteristics of the load problem.

- We empirically illustrate the performance of the proposed offline algorithms and the new online algorithm, OTT, on the problem of controlling the average consumption of a large population of water-heaters. To this end, we describe a controlled model of water-heaters and numerically simulate their consumption using a realistic dataset [11].[1]

1.1 Problem Context

For any $s \in \mathbb{N}$, we define $[s] := \{1, \dots, s\}$. We call $\Delta_{\mathcal{S}}$ the simplex on any finite set $\mathcal{S}$, and denote by $|\mathcal{S}|$ its size.

We model the On-Off switching dynamic of a TCL as a loop-free episodic Markov decision process (MDP) with a finite horizon [12]. We divide a day in discrete time steps $n \in [N]$. At time n, the device is at a state $x_n = (m_n, \theta_n) \in \mathcal{X} := \{0,1\} \times \Theta$ where m_n is the operating state (On if 1, Off if 0), and θ_n represents its temperature. It chooses an action $a_n \in \mathcal{A} := \{0,1\}$ (to turn or stay On if 1, or Off if 0) according to a probability vector $\pi_n(\cdot|x_n) \in \Delta_{\mathcal{A}}$, then moving to a next state x_{n+1} according to a transition kernel $p_{n+1}(\cdot|x_n, a_n) \in \Delta_{\mathcal{X}}$ encompassing the device dynamics and external uncertainties from human behavior (e.g., hot water withdraws for water-heaters). The control unit aims to find a policy $\pi := (\pi_n)_{n\in[N]}$ to be sent to each device inducing an average consumption as close as possible to a target consumption profile $\gamma := (\gamma_n)_{n\in[N]}$ over the course of the day, without interfering with consumer behavior.

Addressing decision-making problems involving a large number of agents is a complex task. To tackle this problem, we employ a mean field approximation [13, 14], which consists of considering a continuous population of devices. This simplification is justified by assuming all devices have the same transition kernel. This enables us to express the average consumption curve over a day using the state-action distribution sequence induced when the devices, with dynamics p, adhere to the policy π, i.e., $\mu^{\pi,p} := (\mu_n^{\pi,p})_{n\in[N]}$.

For a target consumption profile $\gamma := (\gamma_n)_{n\in[N]}$, we quantify the one-day loss incurred by the control unit, when all devices follow the sequence of policies π, by $F(\mu^{\pi,p}; \gamma)$, such that $F(\cdot;\gamma) : [0,1]^{N\times|\mathcal{X}|\times|\mathcal{A}|} \to \mathbb{R}$ is a convex function capturing the distance between the target curve γ and the average consumption computed with $\mu^{\pi,p}$. In this paper, we consider two variants of the target tracking problem:

Offline optimization problem

Thanks to the mean field limit, we can consider the problem of minimizing $F(\cdot;\gamma)$, for a fixed target consumption profile γ, over the space of policies inducing state-action distributions, assuming that the probability transition kernel p illustrating the

[1] All the code to reproduce the empirical results is available at: https://github.com/biancammoreno/tcl-online-control

dynamics of a water-heater is known in advance, as in Eq. (1),

$$\min_{\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X} \times N}} F(\mu^{\pi,p}; \gamma)\,. \tag{1}$$

Online learning problem

We consider an extension of Problem (1) to the online learning scenario, where the goal is to compute a sequence of policies $(\pi^t)_{t\in[T]}$ for T days in order to minimize the total loss:

$$L_T := \sum_{t=1}^{T} F(\mu^{\pi^t,p}; \gamma^t)\,, \tag{2}$$

where the target consumption profile γ^t can change daily due to variations in energy supply and inflexible consumption, and is revealed to the learner at the start of episode t. In order to encompass more realistic scenarios, we assume that consumer behavior is unknown, which means that the probability transition kernel p is also partially unknown and must then be learned. Therefore, the control unit's main objective is to minimize its total loss while learning consumer behavior.

1.2 Related Work

In the 1980s, pioneering models for thermostatically controlled load (TCL) switching dynamics were developed by [15–17], paving the way for load control in demand-side management. Today, different types of load control can be envisaged. Centralized load control, where the central directly commands the On-Off operational state of each device, is not scalable for large populations. Non-centralized approaches can be divided into two categories: *Distributed*: Decisions at a TCL are computed based on information exchanged with its neighbors, such as in [18–20]. This technique is not well-suited for discrete decisions, making it less practical. *Decentralized*: The central computes a control command and sends it to all TCLs, and each device behaves locally following this control [21]. For a comprehensive overview see [22].

Managing a large population of devices is challenging. The works of [23, 24] were the first to use a mean field approximation to circumvent this difficulty in load control. In this paper we consider a decentralized mean field approach with randomized policies, i.e., the policy determines the probability of switching each device On or Off. Initial work using randomized policies by [25] and [26] required solving non-convex optimization problems. More recently, [27] proposed a method using a quadratic objective and a Kullback-Leibler (KL) penalty, enabling a Lagrangian approach optimizing both the control policy and the probability transition kernel. However, this method cannot handle uncontrolled noises, so uncertainties like water drains must be modeled deterministically.

Closest to our work, [8] considers the uncontrolled stochastic environment in the KL quadratic control framework by adding constraints on the probability transition kernel. The KL penalty added to the objective function is essential to their main results. However, there is a trade-off between adding the KL penalty and obtaining a good target tracking curve. In contrast, our approach enables the solution of any convex problem without the need to modify the objective function through regularization

penalties. Besides, we also address the situation where the distribution of the stochastic environment is not known and has to be learned.

The way we formulate the load control Problem (1) is also known as the Concave Utility Reinforcement Learning (CURL) problem outside of the load control literature [28]. The CURL problem extends the Reinforcement Learning (RL) task [29], and recent approaches to tackling this more general framework include [6, 28, 30, 31]. In the mean field community, [32] has shown that all the algorithms for solving mean field games in model-based RL can be applied to solving CURL.

To the best of our knowledge, we are the first to consider the online learning version of the load control problem, where the probability transition dynamics are unknown and the target is allowed to change at each day. Modeling devices' dynamic as a Markov decision process (MDP) enables us to efficiently tackle the online learning problem. Online MDPs have mostly been studied in specific cases of CURL, e.g. [33, 34], rather than in its general form, and draw inspiration from online learning problems [9]. The firsts to propose a sub-linear regret algorithm to the online CURL framework are [6]. Building upon their approach, we develop an original algorithm to tackle the online load control task. Due to the specifics of the loading problem, we obtain a sub-linear regret of order $O(\log(T))$ for the regret term associated with dynamic target consumption profiles, where T represents the number of days, improving the bound of [6] of $\tilde{O}(\sqrt{T})$.

2 General Problem Formulation

Let us consider a collection of $\overline{M}$ electrical devices indexed by i. Recall that we model the switching dynamics of a device as an episodic MDP, where each episode is a day divided into N time steps, and the evolution of each device is independent of the evolution of the other devices. At the start of a day, each initial state-action pair of a device is sampled from a fixed distribution $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$. At time step n, a device i in state $x_n^i = (m_n^i, \theta_n^i) \in \mathcal{X}$, where $m_n^i \in \{0,1\}$ is its operational state, and θ_n^i its internal temperature, chooses the action $a_n^i \sim \pi_n(\cdot|x_n^i)$ by means of a policy $\pi_n : \mathcal{X} \to \Delta_{\mathcal{A}}$, and moves to the next state x_{n+1}^i determined by the equation

$$x_{n+1}^i := g(x_n^i, a_n^i, \varepsilon_n^i), \tag{3}$$

where g comprises the physical dynamics of the device and can be computed by approximating an ordinary differential equation (see Section 5), and $(\varepsilon_n^i)_{n\in[N]}$ is an independent sequence of external noises, independent from one device to another, with $\varepsilon_n^i \sim h_n(\cdot)$ for h_n a distribution (e.g., hot water withdrawal for a water-heater). The MDP's probability transition kernel can then be expressed, for all $(x, a, x') \in \mathcal{X}\times\mathcal{A}\times\mathcal{X}$ and $n \in [N]$, as $p_n(x'|x,a) := \mathbb{P}\big(g(x,a,\varepsilon_n) = x'\big)$. All results hold true regardless of whether g is dependent on the time step n. We omit this term in our TCL example because the physical equation of a device remains constant throughout the day.

We assume that all devices are 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 electrical device is $p_{\max} = 1$ so that the average consumption is equal to the proportion of devices at state On. Note

that $\bar{m}_n$ depends on the policy π that the devices follow, thus we can denote it as $\bar{m}_n(\pi)$. Let $\gamma = (\gamma_n)_{n\in[N]} \in [0,1]^N$ be the target consumption profile (for example, the energy production at each time step divided by the number of devices). The control unit's goal is to solve

$$\min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} \mathbb{E}\left[\sum_{n=1}^{N} f_n(\bar{m}_n(\pi);\gamma_n)\right], \tag{4}$$

where f_n represents a general convex and Lipschitz loss, estimating the deviation of the average actual consumption from the target consumption profile at time step n.

We define $\mu^{\pi,p} := (\mu_n^{\pi,p})_{n\in[N]}$ the state-action distribution sequence induced when all devices follow a sequence of strategies $\pi := (\pi_n)_{n\in[N]}$ in the MDP with probability kernel $p := (p_n)_{n\in[N]}$, for all $(x',a') \in \mathcal{X}\times\mathcal{A}$ and all $n \in [N]$, recursively as below:

$$\begin{aligned}
\mu_0^{\pi,p}(x',a') &:= \mu_0(x',a') \\
\mu_n^{\pi,p}(x',a') &:= \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \mu_{n-1}^{\pi,p}(x,a)p_n(x'|x,a)\pi_n(a'|x').
\end{aligned} \tag{5}$$

We denote the empirical state-action distribution as $\hat{\mu}^{\pi,p} := (\hat{\mu}_n^{\pi,p})_{n\in[N]}$, where at time step $n \in [N]$, for all $(x,a) \in \mathcal{X}\times\mathcal{A}$,

$$\hat{\mu}_n^{\pi,p}(x,a) = \frac{1}{\overline{M}}\sum_{i=1}^{\overline{M}} \mathbb{1}_{\{(x_n^i,a_n^i)=(x,a)\}}.$$

For a function $\varphi : \mathcal{X} \to \mathbb{R}$, we define $\mu_n^{\pi,p}(\varphi) := \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \mu_n^{\pi,p}(x,a)\varphi(x)$. We are particularly interested in a function φ such that $\mu_n(\varphi)$ gives us the average consumption of our electrical devices' population. Thus, we consider for now $\varphi : (m,\theta) \mapsto m$.

The homogeneity assumption (same probability kernel p for all devices) implies that there is convergence when $\overline{M} \to \infty$, i.e. $\lim_{\overline{M}\to\infty} \hat{\mu}^{\pi,p} = \mu^{\pi,p}$. We can therefore consider the mean-field approximation of Eq. (4), which brings us to the main optimization problem in this paper already given by Eq. (1):

$$\min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} F(\mu^{\pi,p};\gamma) := \sum_{n=1}^{N} f_n(\mu_n^{\pi,p}(\varphi);\gamma_n).$$

Here are some examples of loss functions that the control unit can consider:

- Quadratic: $F(\mu^{\pi,p};\gamma) := \sum_{n=1}^{N}(\mu_n^{\pi,p}(\varphi) - \gamma_n^t)^2$.
- Kullback-Leibler divergence: $F(\mu^{\pi,p};\gamma) := \sum_{n=1}^{N} \mu_n^{\pi,p}(\varphi)\log\left(\frac{\mu_n^{\pi,p}(\varphi)}{\gamma_n}\right)$.

Offline optimization setting (Section 3):

To address the offline optimization Problem (1) where the dynamics p and the target are known, we reformulate the problem as a convex optimization problem on the

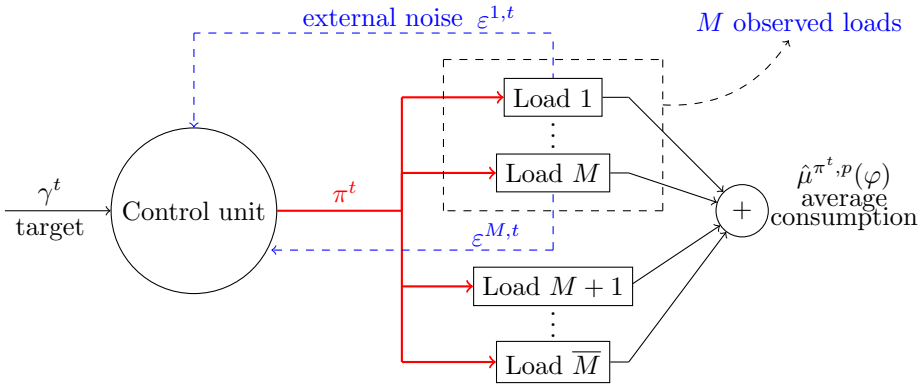


Fig. 1: Control unit's general framework.

space of probability measures. We then propose an iterative protocol, where at each iteration $k \in [K]$ the control unit updates the policy using one iteration of a convex optimization algorithm we denote by $\mathcal{F}$, depending on the previous policy π^{k-1}, the model dynamics p, and the objective function F, i.e. $\pi^k := \mathfrak{F}(\pi^{k-1}, p, F)$. Section 3 presents examples of such algorithms, including Fictitious Play for mean-field games [7] (equivalent to Frank-Wolfe for potential games [32]) and Mirror Descent with a non-standard Bregman divergence [6], which are specifically tailored to our model and provide quasi-explicit solutions at each iteration. Moreover, these algorithms guarantee that the gap between the optimal and the current policy, $\min_{k\in[K]} F(\mu^{\pi^k,p}) - F(\mu^{\pi^*,p})$, vanishes as the number of iterations K increases.

Online learning setting (Section 4):

On the online learning setting, the control unit's problem is to compute a sequence of policies $(\pi^t)_{t\in[T]}$ over T days, where the target reference profile γ^t may change from day to day but is known beforehand at the beginning of each day. The goal is to minimize the total loss as defined in Eq. (2). The performance is measured in comparison to the optimal policy for each target, using the following regret:

$$R_T := \sum_{t=1}^{T} F(\mu^{\pi^t,p}; \gamma^t) - \sum_{t=1}^{T} \min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} F(\mu^{\pi,p}; \gamma^t). \tag{6}$$

We consider the dynamics of Eq. (3). The physical dynamics of each device, represented by the function g, are assumed to be known. However, the consumer behavior, characterized by the external noise distribution h_n, is unknown. Therefore, the control unit must simultaneously optimize the objective function and learn the noise distribution through observations. The online protocol for the control unit is detailed in Algorithm 1, and illustrated in Fig. 1.

At each day t, the control unit chooses a policy π^t, sends it to all devices, observes the external noise of a sub-population of M devices, $(\varepsilon_1^{i,t}, \dots, \varepsilon_N^{i,t})$ for $i \in [M]$, computes an estimate $\hat{p}^{t+1}$ of the probability kernel using the observations, receives the

Algorithm 1 Control unit's online protocol

Input: initial state-action distribution μ_0, initial strategy sequence π^1.
for $t = 1, \dots, T$ **do**
  for $i = 1, \dots, M$ **do**
    the i-th device starts at $(x_0^{i,t}, a_0^{i,t}) \sim \mu_0(\cdot)$
    for $n = 1, \dots, N$ **do**
      environment draws new state $x_n^{i,t} \sim p_n(\cdot|x_{n-1}^{i,t}, a_{n-1}^{i,t})$
      control unit observes device's i external noise $\varepsilon_n^{i,t}$
      deivce i chooses an action $a_n^{i,t} \sim \pi_n^t(\cdot|x_n^{i,t})$
    end for
  end for
  control unit computes, for all $n \in [N]$, new estimate $\hat{p}_n^{t+1}$ from data $(\varepsilon_n^{i,s})_{s\in[t],j\in[M]}$

  next day's target γ^{t+1} is exposed
  control unit computes $\pi^{t+1} = \mathfrak{F}(\pi^t, \hat{p}^{t+1}, F(\cdot; \gamma^{t+1}))$ and send to all devices
end for
return $(\pi^t)_{t\in[T]}$

next day's target γ^{t+1}, and calculates the policy for the next day by applying the auxiliary problem $\mathfrak{F}$ on $\pi^t, F(\cdot; \gamma^{t+1})$, and $\hat{p}^{t+1}$. To compute a strategy sequence with sub-linear regret the control unit faces two challenges: how to estimate p with $\hat{p}^{t+1}$ from the data and how to define the auxiliary optimization problem $\mathfrak{F}$. In Section 4, we introduce the first algorithm for the online target tracking problem and we show it achieves sub-linear regret.

Remark 2.1. *In this paper, we consider the special case of a population of water-heaters when we describe the model in Section 5 and the experiments in Section 6. However, it should be noted that all results remain valid for other types of electrical devices whose dynamics are those of the Eq. (3), e.g. all TCLs, electric vehicle batteries, etc.*

3 Offline Optimization Setting

We now turn to algorithms for solving the main Problem (1), which are further simulated in Section 6 in the water-heater's case using the model in Section 5.

3.1 Reformulation of the Control Unit's Objective

Problem (1) is not convex in π. We therefore reformulate the control unit's objective to obtain a convex problem. For that we define

$$\mathcal{M}_{\mu_0}^p := \Big\{ \mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N \Big| \sum_{a'\in\mathcal{A}} \mu_n(x', a') = \sum_{x\in\mathcal{X}, a\in\mathcal{A}} p_n(x'|x, a)\mu_{n-1}(x, a) \ , \forall x' \in \mathcal{X}, \forall n \in [N] \Big\}, \tag{7}$$

as the set of state-action distribution sequences satisfying the Bellman-flow in the MDP with transition kernel p and initial state-action distribution μ_0. For any $\mu \in \mathcal{M}_{\mu_0}^p$, there exists a strategy π such that $\mu^{\pi,p} = \mu$. It suffices to take $\pi_n(a|x) = \mu_n(x,a)/\sum_{a\in\mathcal{A}} \mu_n(x,a)$ when the normalization factor is non-zero, and arbitrarily defined otherwise [12]. To ensure the uniqueness of π given μ, we make the convention that $\pi_n(a|x) = \frac{1}{|\mathcal{A}|}$ whenever the normalization factor is zero. We therefore have the equivalence

$$\min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} F(\mu^\pi;\gamma) \equiv \min_{\mu\in\mathcal{M}_{\mu_0}^p} F(\mu;\gamma), \tag{8}$$

for any target curve γ. Note that the optimization problem over μ is convex.

3.2 Algorithmic approaches

We now consider different auxiliary optimization problems $\mathfrak{F}$, as discussed in Section 2, to solve Problem (8). We assume that the probability kernel p is known (i.e. g and $(h_n)_{n\in[N]}$ are known in the dynamics of Eq. (3)) and, to minimize notations, we let $\mu^\pi := \mu^{\pi,p}$ and $\mathcal{M}_{\mu_0} := \mathcal{M}_{\mu_0}^p$.

MD-CURL [6]:

MD-CURL is an algorithm based on Mirror Descent to solve the general CURL problem by using a non standard Bregman divergence. For $\tau_k > 0$ a positive learning rate, MD-CURL considers as an auxiliary problem computing at iteration $k+1$,

$$\mu^{k+1} \in \operatorname*{arg\,min}_{\mu^\pi\in\mathcal{M}_{\mu_0}} \Big\{ \tau_k \langle \nabla F(\mu^k;\gamma), \mu^\pi\rangle + \Gamma(\mu^\pi,\mu^k) \Big\}, \tag{9}$$

where $\langle \nabla F(\mu^k;\gamma), \mu^\pi\rangle := \sum_{n=1}^N \langle \nabla f_n(\mu_n^k;\gamma_n), \mu_n^\pi\rangle$, and the regularization function Γ is defined as

$$\Gamma(\mu^\pi,\mu^{\pi'}) := \sum_{n=1}^N \mathbb{E}_{(x,a)\sim\mu_n^\pi(\cdot)} \left[\log\left(\frac{\pi_n(a|x)}{\pi'_n(a|x)}\right)\right]. \tag{10}$$

Thanks to the choice of Γ as the regularization function, [6] demonstrated that Problem (9) can be efficiently solved using dynamic programming. The solution is given by $\mu^{k+1} = \mu^{\pi^{k+1}}$, as shown in Eq. (5), with $\pi_n^{k+1}(a|x) = ExpTwist\big(\pi^k, Q^k, \tau_k\big)_n(x,a)$, where

$$ExpTwist\big(\pi^k, Q^k, \tau_k\big)_n(x,a) := \frac{\pi_n^k(a|x)\exp\big(\tau_k Q_n^k(x,a)\big)}{\sum_{a'}\pi_n^k(a'|x)\exp\big(\tau_k Q_n^k(x,a')\big)}. \tag{11}$$

Here $Q^k := (Q_n^k)_{n\in[N]}$ is a regularized Q-function [29] computed backwardly in time alternating with policy updates, using the gradient of the objective function evaluated at the previous policy $\nabla F(\mu^{\pi^k};\gamma)$, the previous policy π^k, and the probability transition kernel p, such that

$$Q_N^k(x,a) = -\nabla f_N(\mu_N^k)(x,a)$$

$$Q_n^k(x,a) = \max_{\pi_{n+1}\in(\Delta_{\mathcal{A}})^{\mathcal{X}}} \Bigg\{ -\nabla f_n(\mu_n^k)(x,a) + \sum_{x'} p_{n+1}(x'|x,a) \sum_{a'} \pi_{n+1}(a'|x') \Bigg[-\frac{1}{\tau_k} \log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) + Q_{n+1}^k(x',a') \Bigg] \Bigg\}. \tag{12}$$

In short, we denote $Q^k := \mathcal{Q}\big(\nabla F(\mu^{\pi^k};\gamma), \pi^k, p\big)$.

In addition, [6] proves that Γ is a Bregman divergence induced by

$$\psi(\mu) := \sum_{n=1}^{N} \phi(\mu_n) - \sum_{n=1}^{N} \phi(\rho_n), \tag{13}$$

where $\rho_n(x) := \sum_{a\in\mathcal{A}} \mu_n(x,a)$, and ϕ is the neg-entropy function defined for any probability measure $\eta \in \Delta_E$, whatever the (finite) space E, with the convention that $0\log(0) = 0$, by $\phi(\eta) := \sum_{x\in E} \eta(x)\log\eta(x)$. Therefore, if F is convex and Lipschitz, MD-CURL has a convergence rate of $O\big(1/\sqrt{K}\big)$.

Optimistic MD-CURL:

Just as [10] develops the Optimistic Mirror Descent algorithm, we can also define the Optimistic MD-CURL approach, a faster version of MD-CURL, by solving at iteration $k+1$

$$\begin{aligned} \mu^{k+1} &\in \arg\min_{\mu\in\mathcal{M}_{\mu_0}} \big\{\tau\langle\nabla F(\nu^k;\gamma),\mu\rangle + \Gamma(\mu,\nu^k)\big\} \\ \nu^{k+1} &\in \arg\min_{\nu\in\mathcal{M}_{\mu_0}} \big\{\tau\langle\nabla F(\mu^{k+1};\gamma),\nu\rangle + \Gamma(\nu,\nu^k)\big\}, \end{aligned} \tag{14}$$

where Γ is defined as in Eq. (10). With the additional assumption that F is smooth (Lipschitz gradient) for all γ targets, the convergence rate of Optimistic MD-CURL is of the order of $O(1/K)$.

Potential games:

The works of [32, 35] relate the optimality conditions of optimization problems to the concept of Nash equilibrium in games. It is then possible to provide an equivalence between Problem (1) and a mean field game (MFG) problem by considering a game whose reward is given at each time step n by $-\nabla f_n(\mu_n;\gamma)(x_n,a_n)$ for all $(x_n,a_n,\mu_n) \in \mathcal{X}\times\mathcal{A}\times\Delta_{\mathcal{X}\times\mathcal{A}}$. Thus, we can also consider algorithms for MFGs in episodic MDPs as auxiliary problems for solving the target tracking task. In Section 6 we simulate the Fictitious Play algorithm for MFG (FP-MFG) of [7] (equivalent of Frank-Wolf for potential games), defined in Algorithm 3 at Appendix A. For F convex and smooth, FP-MFG has a convergence rate of order $O(1/K)$.

4 Online Learning Setting: a Novel Approach

We consider the online variant of Problem (1) where the control unit must compute a sequence of policies every day through T days while facing unknown external noises and changing targets. We first propose two baseline algorithms with sub-linear regret inspired by [6], but that do not fully exploit the structure of the target tracking problem. We then propose a new algorithm, called Online Target Tracker (OTT), specifically designed for the online load control scenario. Unlike the methods proposed in [6], OTT uses the concept of predictable sequences introduced in [10]. This enables OTT to exploit the predictability of future stochastic targets, specific to the load problem, resulting in better convergence rates. We achieve this by adapting the optimistic MD-CURL algorithm, from Section 3, to handle online scenarios with dynamically changing targets and an unknown transition kernel.

4.1 Learning the External Noises

Since the control unit does not know the noise dynamics due to consumers' behavior, it has to estimate it from observation. Recall that we denote by M the number of devices observed at each day and that the dynamics follow Eq. (3). Let δ_x be the Dirac distribution centered in x. For all $n \in [N]$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$, we define $\hat{p}_n^1(\cdot|x, a) = \frac{1}{|\mathcal{X}|}$, and

$$\hat{p}_n^{t+1}(\cdot|x, a) := \frac{1}{Mt} \sum_{j=1}^{M} \delta_{g(x,a,\epsilon_n^{j,t})}(\cdot) + \frac{t-1}{t} \hat{p}_n^t(\cdot|x, a), \tag{15}$$

as the empirical probability transition kernel estimated on day $t + 1$. When g is invertible the external noises can be determined by observing the subpopulation's state-action trajectory and inverting the function g.

4.2 Algorithmic Approaches

Operating under the online protocol outlined in Algorithm 1, the control unit updates its estimate $\hat{p}^{t+1}$ based on the observed noises at the end of each day t using Eq. (15). Upon receiving the target consumption profile signal γ^{t+1} for the subsequent day, the control unit calculates the corresponding policy π^{t+1}. This section presents novel approaches for calculating the policy under different conditions imposed on the possible values of the target consumption profile. To ease notations, let $\mathcal{M}_{\mu_0}^t := \mathcal{M}_{\mu_0}^{\hat{p}^t}$.

Intuitive approach:

Although the target consumption profiles change daily due to variations in energy production and flexible consumption, we assume that these profiles are known to the control unit before the policy is calculated. Our first approach is to build an online algorithm that, at day $t \in [T]$, runs K iterations of one of the offline methods in Section 3 on the set $\mathcal{M}_{\mu_0}^t$, for the corresponding target γ^t. This approach can deal with targets encompassing any value in $\mathbb{R}^N$. However, it is computationally costly. For

example, MD-CURL with a convergence rate of $O(1/\sqrt{K})$, requires $K = T$ iterations to achieve a regret of $O(\sqrt{T})$, resulting in a computational complexity of order $O(T^2)$. Faster methods like FP-MFG or optimistic MD-CURL still require $K = \sqrt{T}$ iterations, leading to a computational complexity of order $O(T^{3/2})$.

To improve computational complexity we consider a new framework. Assume that the target consumption profile γ is restricted to a finite set $\Upsilon := \{\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_J\}$, where J denotes the total number of permissible targets. At each day t, the target can take any value from Υ, known to the control unit before the day begins.

Classic online CURL:

A naive algorithm for this framework is to apply a classic online learning algorithm for CURL problems, like Greedy MD-CURL from [6], for each target $\boldsymbol{\gamma}_j$, for $j \in [J]$. At each day t, we perform one iteration of Greedy MD-CURL over $\mathcal{M}^t_{\mu_0}$, initialized with the best policy computed the last time target j appeared. This approach achieves a regret of order $O(\sqrt{JT}\log(T))$, and computational complexity of order $O(T)$.

Classic online learning algorithms like Greedy MD-CURL aim for sub-linear regret when facing adversarial objective functions revealed only at each episode's end [9]. Unlike this scenario, our framework assumes targets are known beforehand. This allows us to achieve better results than classical online learning lower bounds [36]. We present a novel algorithm that leverages this advantage while upholding computational efficiency.

New algorithm: Online Target Tracker (OTT):

To achieve improved regret bounds while maintaining lower complexity, we introduce OTT. It considers the framework with only a finite number of targets and employs an online learning version of optimistic MD-CURL tailored to the online target tracking problem. Optimistic methods are uncommon in traditional online learning literature as they do not provide improved regret bounds for adversarial objective functions. However, in the context of online target tracking, as targets are known in advance, we can demonstrate that implementing optimistic methods enables us to improve the bound of the regret term responsible for handling varying targets from $\tilde{O}\big(\sqrt{T}\big)$ to $O\big(\log(T)\big)$.

Let t_j mark the index of the latest day, up to day t, when target $\boldsymbol{\gamma}_j$ was observed. If before the start of day $t+1$ the control unit observes target $\boldsymbol{\gamma}_j$, OTT solves

$$\begin{aligned} \mu^{t+1} &\in \operatorname*{arg\,min}_{\mu \in \mathcal{M}^{t+1}_{\mu_0}} \{\tau \langle \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j), \mu \rangle + \Gamma(\mu, \tilde{\nu}^{t_j})\} \\ \nu^{t+1} &\in \operatorname*{arg\,min}_{\nu \in \mathcal{M}^{t+1}_{\mu_0}} \{\tau \langle \nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j), \nu \rangle + \Gamma(\nu, \tilde{\nu}^{t_j})\}, \end{aligned} \tag{16}$$

where, $\mu^{t+1} := \mu^{\pi^{t+1}, \hat{p}^{t+1}}$, $\tilde{\nu}^{t_j} := \nu^{\tilde{\eta}^{t_j}, \hat{p}^t}$ with

$$\tilde{\eta}^s := (1 - \alpha_s)\eta^s + \tfrac{\alpha_s}{|\mathcal{A}|}, \tag{17}$$

for all $s \in [T]$, where $\alpha_s \in (0, 1/2)$ is an exploration parameter, and π^s, η^s are the policies inducing μ^s, ν^s respectively in the MDP $\mathcal{M}^s_{\mu_0}$ for all $s \in [T]$. A practical implementation of OTT is given in Algorithm 2.

Algorithm 2 OTT: Online Target Tracker

Input: number of days T, initial auxiliary policy $\eta^0 \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$, number of devices observed per day M, initial state-action distribution μ_0, learning rate τ, exploration parameters $(\alpha_t)_{t\in[T]}$.
Initialization: $\forall (x,a)$, $\hat{p}^1_n(\cdot|x,a) = g(x,a,0)$; $\forall j \in [J], t_j = 0$
for $t = 0, \dots, T-1$ **do**
  for $i = 1, \dots, M$ **do**
    i-th device starts at $(x_0^{i,t}, a_0^{i,t}) \sim \mu_0(\cdot)$
    for $n = 1, \dots, N$ **do**
      environment draws new state $x_n^{i,t} \sim p_n(\cdot|x_{n-1}^{i,t}, a_{n-1}^{i,t})$
      control unit observes device i's external noise $\varepsilon_n^{i,t}$
      device i takes an action $a_n^{i,t} \sim \pi_n^t(\cdot|x_n^{i,t})$
    end for
  end for
  update probability kernel estimate for all (x,a):
  $$\hat{p}_n^{t+1}(\cdot|x,a) := \frac{1}{Mt}\sum_{i=1}^{M} \delta_{g(x,a,\varepsilon_n^{i,t})} + \frac{t-1}{t}\hat{p}_n^t(\cdot|x,a)$$
  control unit receives next day target $\gamma^{t+1} \in \Upsilon$
  if $\gamma^{t+1} = \boldsymbol{\gamma}_j$ for $j \in [J]$ **then**
    compute next policy:
    $Q^{t+1} := \mathcal{Q}\big(\nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j), \tilde{\eta}^{t_j}, \hat{p}^{t+1}\big)$ as in Eq. (12)
    $\pi^{t+1} := ExpTwist\big(\tilde{\eta}^{t_j}, Q^{t+1}, \tau\big)$ as in Eq. (11)
    compute $\mu^{t+1} = \mu^{\pi^{t+1}, \hat{p}^{t+1}}$ as in Eq. (5)
    compute next auxiliary policy:
    $\tilde{Q}^{t+1} := \mathcal{Q}\big(\nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j), \tilde{\eta}^{t_j}, \hat{p}^{t+1}\big)$ as in Eq. (12)
    $\eta^{t+1} := ExpTwist\big(\tilde{\eta}^{t_j}, \tilde{Q}^{t+1}, \tau\big)$ as in Eq. (11)
    update count $t_j \leftarrow t+1$
  end if
  compute $\tilde{\eta}^{t+1}$ and $\tilde{\nu}^{t+1}$ as in Eq. (17)
end for
return $(\pi^t)_{t\in[T]}$

4.3 Online Target Tracker Regret Analysis

In this section, we prove the regret bound of OTT. For that, we use results from [6, 10] while also having to handle an online optimization problem with varying constraint sets. Let $\pi^{*,j} := \min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} F(\mu^{\pi,p}; \boldsymbol{\gamma}_j)$, for all $j \in [J]$. Define the map $v : [T] \to [J]$ such that $v(t) = j$ if $\gamma^t = \boldsymbol{\gamma}_j$. We start by decomposing the regret in Eq. (6) into three

terms,

$$
\begin{aligned}
R_T = &\sum_{t=1}^{T} F(\mu^{\pi^t,p};\boldsymbol{\gamma}_{v(t)}) - F(\mu^{\pi^t,\hat{p}^t};\boldsymbol{\gamma}_{v(t)}) \\
&+ \sum_{t=1}^{T} F(\mu^{\pi^t,\hat{p}^t};\boldsymbol{\gamma}_{v(t)}) - F(\mu^{\pi^{*,v(t)},\hat{p}^t};\boldsymbol{\gamma}_{v(t)}) \\
&+ \sum_{t=1}^{T} F(\mu^{\pi^{*,v(t)},\hat{p}^t};\boldsymbol{\gamma}_{v(t)}) - F(\mu^{\pi^{*,v(t)},p};\boldsymbol{\gamma}_{v(t)}) \\
&:= R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) + R_T^{policy} + R_T^{MDP}\big((\pi^{*,v(t)})_{t\in[T]}\big).
\end{aligned}
$$

The terms $R_T^{MDP}\big((\pi^t)_{t\in[T]}\big)$ and $R_T^{MDP}\big((\pi^{*,v(t)})_{t\in[T]}\big)$ account for the error incurred by the control unit due to its lack of knowledge of the true probability kernel. The term R_T^{policy} represents the cost of calculating sub-optimal policies with OTT. Subsections 4.4 and 4.5 bound each of these terms, yielding our main result:

Theorem 4.1 (Main result)**.** *Consider the target tracking problem in an episodic MDP with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, episodes of length N, probability kernel $p := (p_n)_{n\in[N]}$, T days, and targets γ^t arbitrarily chosen each day t from a finite set $\Upsilon := \{\boldsymbol{\gamma}_1,\dots,\boldsymbol{\gamma}_J\}$. For all $t \in [T]$, let $F^t(\cdot;\gamma^t) : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ be convex, L-Lipschitz and β-smooth with respect to the norm $\|\cdot\|_{\infty,1} := \sup_{n\in[N]} \|\cdot\|_1$. Then, with a probability of $1-\delta$, for any $\delta \in (0,1)$, OTT obtains, for $\tau = 1/\beta$ and $\alpha_t = 1/T$ for all $t \in [T]$,*

$$
\begin{aligned}
R_T \leq\ & 2NL\sqrt{\frac{2T}{M}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)} \\
&+ \beta\left(2N^2 J\log(|\mathcal{A}|T)\log\left(T\right) + 2N + JN\log(|\mathcal{A}|)\right).
\end{aligned}
$$

4.4 Bounding R_T^{MDP}

The same analysis holds for $R_T^{MDP}\big((\pi^{*,v(t)})_{t\in[T]}\big)$ and $R_T^{MDP}\big((\pi^t)_{t\in[T]}\big)$. It uses a concentration result from Proposition 5.5 of [6], assuring that, with probability $1-\delta$ for any $\delta \in (0,1)$, OTT obtains

$$
R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) \leq LN\sqrt{\frac{2T}{M}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}.
$$

4.5 Bounding R_T^{policy}

The R_T^{policy} bound is a novel result. Despite changing costs, prior target knowledge allows for novel improved results. The challenge lies in handling online optimization with dynamic constraints. The result is stated in Proposition 4.2 and its proof is in Appendix B.

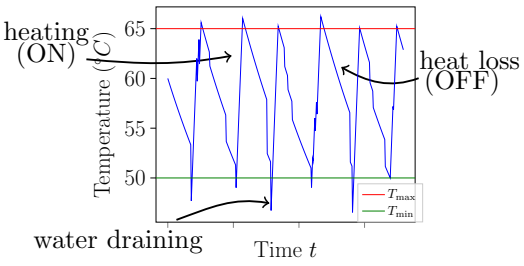


Fig. 2: Temperature evolution of a water-heater following the nominal dynamics.

Proposition 4.2. *Under the same hypothesis as in Theorem 4.1, OTT obtains,*

$$R_T^{policy} \leq \beta\Big(2N^2 J \log(|\mathcal{A}|T)\log\big(T\big) + 2N + JN\log(|\mathcal{A}|)\Big).$$

The main challenge of the proof is to deal with variable constraint sets $\mathcal{M}_{\mu_0}^t$. The policy played for target γ^j is updated only when this target appears. However, the probability kernel estimate is updated every day. Consequently, we need to bound the difference between the state-action distributions induced by the probability kernel estimate between two appearances of the same target. We must also guarantee a limit on $\|\nabla\psi(\mu^{\pi,p})\|_{\infty,1}$, the function inducing the Bregman divergence, justifying our construction in Eq. (17).

5 Water-Heater Model

5.1 Nominal Dynamics of a Water-Heater

We consider that a water-heater follows the dynamics described in Section 2. For simplicity's sake, we only consider temperatures within a finite set Θ. We call its uncontrolled dynamics the nominal dynamics. A water-heater that follows the nominal dynamics 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 $T_{\max}$ (also known as quality of service constraint). 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 are illustrated in Figure 2.

For a heater in state $x_n = (m_n, \theta_n)$ at time step n, the next temperature is computed by $\theta_{n+1} := \hat{T}(m_n, \theta_n, \varepsilon_n)$ where $\hat{T}$ is an approximation of the solution of the ordinary differential Eq. (18)

$$\begin{cases} \frac{dT(t)}{dt} = -\underbrace{\rho(T(t) - T_{\text{amb}})}_{\text{heat loss}} + \underbrace{\sigma m_n p_{\max}}_{\text{Joule effect}} - \underbrace{\tau(T(t) - T_{\text{in}})f(t)}_{\text{water drain}} \\ T(t_n) = \theta_n, \end{cases} \tag{18}$$

modelling the impact of the heat loss to the environment temperature T_{amb}, the Joule effect (heating), and hot water drainings (showers, dishwashing, etc). 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.

In order to compute $\hat{T}$, we start by making an Euler discretization as in Eq. (19). We define a sequence of independent random variables $(\varepsilon_n)_{n\in[N]}$ denoting the amount of draining in liters at each time step. The independence of drains at each time step is justified by assuming that the time δ between two steps is large enough to contain all the time when hot water will be drawn from the water-heater tank for a single use.

$$\overline{T}(m_n, \theta_n, \varepsilon_n) := \theta_n + \delta\big(- \rho\,(\theta_n - T_{\text{amb}}) + \sigma m_n p_{\max} - \varepsilon_n \tau\,(\theta_n - T_{\text{in}})\big) \tag{19}$$

Let the finite set of possible temperatures, Θ, consists of integers ranging from T_{amb} to $T_{\max}$, where $T_{\text{amb}} < T_{\min}$. This assumption is reasonable since the ambient temperature is typically lower than the minimum acceptable temperature for the heater. Due to the dynamics of the operating state, θ_{n+1} never exceeds $T_{\max}$ (the heater turns Off when it reaches $T_{\max}$ and its temperature only decreases when it is turned Off). Hence, we take $\theta_{n+1} = \hat{T}(m_n, \theta_n, \varepsilon_n) := \text{Round}(\overline{T}(m_n, \theta_n, \varepsilon_n))$, 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.

The full nominal dynamics at a discretized time is then given by Eq. (20),

$$\begin{cases} \theta_{n+1} = \hat{T}(m_n, \theta_n, \varepsilon_n) \\ m_{n+1} = \begin{cases} m_n, & \text{if } \theta_{n+1} \in [T_{\min}, T_{\max}] \\ 0, & \text{if } \theta_{n+1} \geq T_{\max} \\ 1, & \text{if } \theta_{n+1} \leq T_{\min}. \end{cases} \end{cases} \tag{20}$$

5.2 Randomized Controlled Dynamics of a Water-Heater

The action space of the MDP is 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 deterministicaly 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. Hence, we define a randomized policy $\pi := (\pi_n)_{n\in[N]}$ such that action a_n is sampled with probability $\pi_n(\cdot|x_n) \in \Delta_{\mathcal{A}}$, conditioned in the current state x_n. The next operating state is now given by $m_{n+1} := M(a_n, \theta_{n+1})$,

where

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

Moreover 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.

Casting into the general framework for TCL dynamics as in Eq. (3),

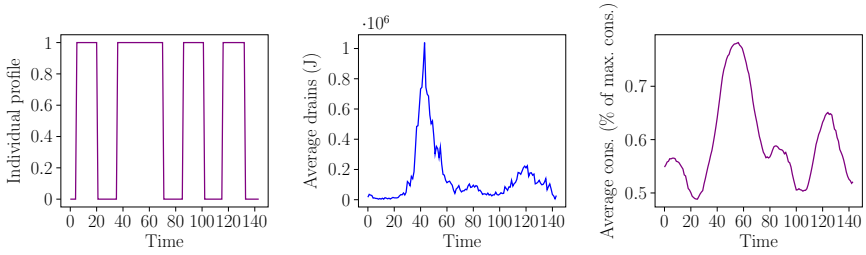
$$x_{n+1} = g(x_n, a_n, \varepsilon_n) := \big(M_n(a_n, \hat{T}(x_n, \varepsilon_n)), \hat{T}(x_n, \varepsilon_n)\big). \tag{21}$$

6 Experiments with water-heaters

6.1 Simulating the Nominal Dynamics

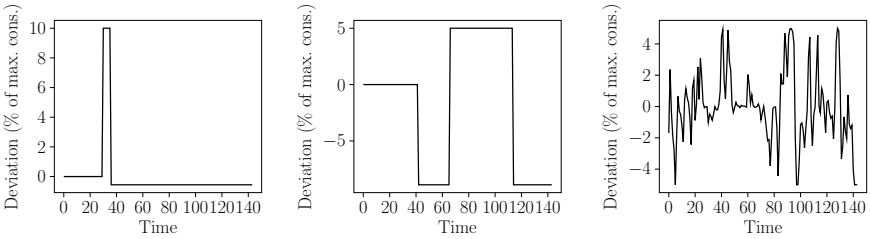
To simulate the nominal dynamics, we employ the nominal model presented in Section 5 in conjunction with data obtained from the SMACH (*Simulation Multi-Agents des Comportements Humains*) platform [11]. This data encompasses the simulation of water withdrawals made by over 5,132 water-heaters over a single day during the summer of 2018. We set a time frequency of $\delta_t = 10$ minutes to capture the entire duration of single hot water withdrawal events, thereby satisfying the independence of external noise assumption. The temperature deadband is defined as $T_{\min} = 50°C$ and $T_{\max} = 65°C$. The values of the parameters ρ, σ, τ and $p_{\max}$ used in the temperature Eq. (19) are computed in Eq. (22) using the variables introduced in Tables C1 and C2 in Appendix C. We take $T_{\text{amb}} = 25°C$ and $T_{\text{in}} = 18°C$. Figure 3 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.

$$\begin{aligned} \rho &= \frac{\text{coefLoss} * 3600}{\text{capWater} * \text{denWater} * \text{vol/height}} \quad \text{(fraction of heat loss by hour)} \\ \sigma &= (\text{vol} * \text{denWater} * \text{capWater})^{-1} \\ \tau &= (\text{vol} * \text{denWater})^{-1} \end{aligned} \tag{22}$$



(a) Individual profile. (b) Average drain in Joules. (c) Average consumption.

Fig. 3: Individual profile, average drain and power consumption for a simulation of 10^4 water-heaters over a period of one day.



(a) One hour step. (b) Eight hours step. (c) TSO balancing signal.

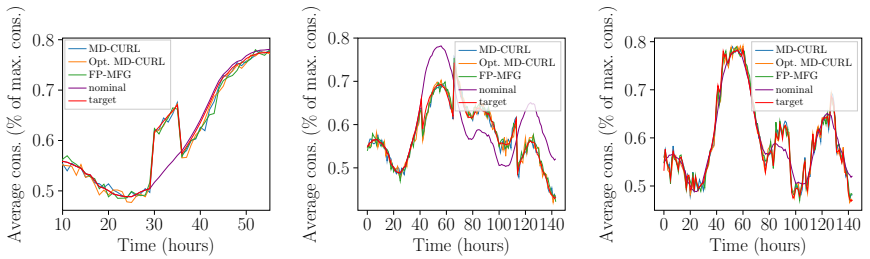
Fig. 4: Deviation signals $(\lambda_n)_{n\leq N}$.

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, 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 control process that can be repeated over many episodes. Electricity consumption is then shifted in time (from one moment of the day to another), but daily consumption remains the same. We consider the three deviation signals illustrated in Figure 4: a one hour step deviation, an eight hours step deviation (during peak hours), and a transmission system operator (TSO) balancing signal.

For all experiments we have $N = 144$, $|\mathcal{X}| = 82$ (two operational states On/Off times 41 possible temperatures - integers from the ambient temperature $T_{\text{amb}} = 25$ to $T_{\max} = 65$), $|\mathcal{A}| = 2$. On every plot the policy is simulated over 5000 heaters.

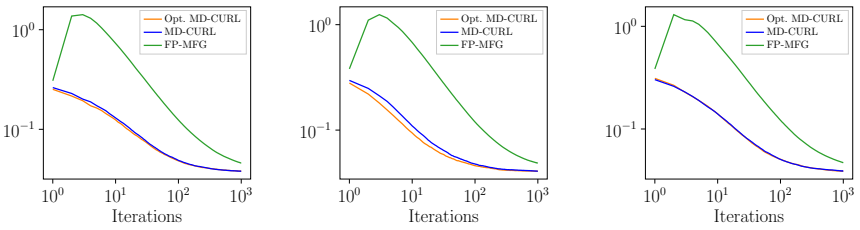
6.2 Offline Setting

We start by simulating the algorithms introduced for the offline setting in Section 3.



(a) One hour step. (b) Eight hours step. (c) TSO balancing signal.

Fig. 5: Average consumption profile of 10^4 heaters following the policy output by each offline algorithm from Section 3 over 10^3 iterations.



(a) One hour step. (b) Eight hours step. (c) TSO balancing signal.

Fig. 6: Log plot of the objective function per iteration for each offline method in Section 3 for each deviation.

In Figure 5, the consumption predicted by the optimal policies for all algorithms in Section 3 closely tracks the actual target consumption, as expected. This convergence to the same optimal solution is achieved by each algorithm through different strategies and at varying rates. Figure 6 shows the log objective function per iteration, revealing that MD-CURL and optimistic MD-CURL converge faster than FP-MFG.

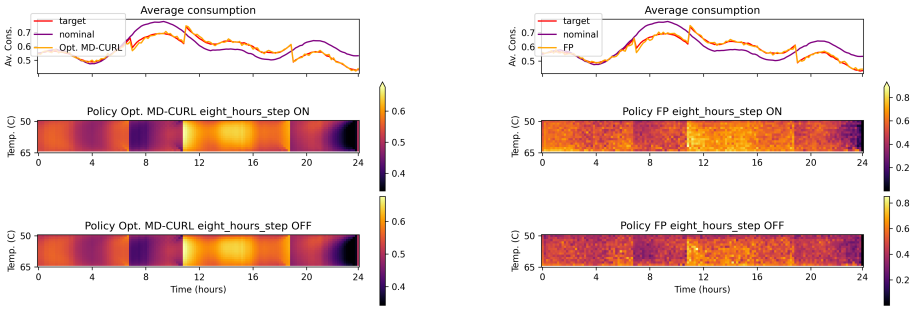
To visualize the learned policies, Figure 7 plots the probability of choosing the On action (represented by color intensity) at each time step (x-axis) for all possible temperatures (y-axis), with $T_{\min}$ at the top, when the current state is either On (top) or Off (bottom). This plot focuses on optimistic MD-CURL and FP-MFG in the eight-hour deviation scenario. Notably, optimistic MD-CURL exhibits a more smooth policy compared to FP. Note also that the probability of switching On or Off is independent of the current temperature.

Different initializations of the algorithm impact switch counts.

Our model does not assume a On/Off switching limit for each water-heater, but excessive switches can harm the device. An advantage of optimistic MD-CURL is that different initializations lead to distinct policies, generating the same average consumption profile but differing in the average number of On/Off switches over the time horizon considered. This allows us to reduce the number of switches without new constraints by discovering several policies that achieve the desired consumption profile and selecting the one that generates the fewest switches. This is only possible thanks to MD-CURL's regularization term. For example, in Figure 7a, the algorithm is initialized with the uniform policy, and has an average of 33 switches per day. The nominal policy has in average only 3 switches per day. Initializing optimistic MD-CURL with a policy deviating by 0.1 from the nominal policy significantly reduces daily switches to an average of 9.2 while still following the target curve, see Figure 8. Note also that for this initialization, the probabilities vary with temperature, unlike in the uniform initialization.

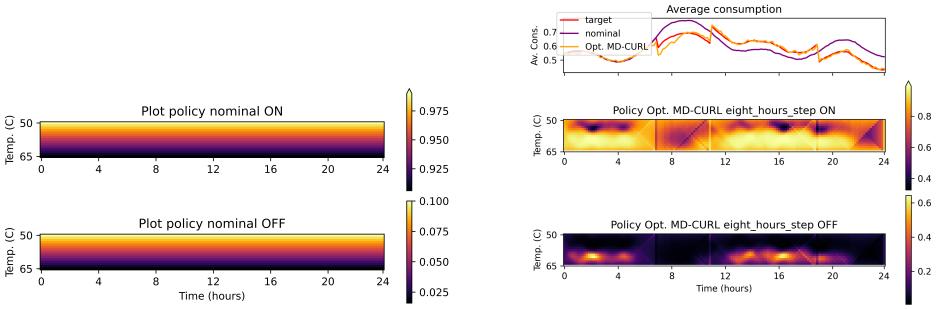
6.3 Online Setting

We investigate the performance of OTT in the online setting where water withdrawal noise is unknown and needs to be learned. Given a finite set of three possible target



(a) Opt. MD-CURL (b) FP

Fig. 7: Optimal policies for Optimisitic MD-CURL and FP over 10^3 iterations for the target with eight hours step deviation.



(a) Initial policy (b) Opt. MD-CURL

Fig. 8: [left] Initial policies π^0 as a deviation of 0.1 from the nominal policy. [right] Optimal policies for Optimistic MD-CURL initialized with π^0 at left.

deviations (as shown in Figure 4), we pre-compute the target sequence for all days uniformly at random. We use $T = 10^3$ days and a sub-population of $M = 100$ water-heaters out of 5132 is observed per day to learn the external noise.

Figure 9 compares the log regret per iteration for OTT (blue), OTT with known noise dynamics (green), and OTT with unknown noise dynamics, where the learner never learns the noise distribution (orange), i.e. $\hat{p}_n^t(\cdot|x,a) = \delta_{g_n(x,a,0)}$ for all (x,a). Figure 10 shows the average consumption profile induced by the best policies for each target. We see that OTT learning the dynamics quickly matches OTT with known noise dynamics, and that never learning the noise is sub-optimal. While our model simplifies external noise by assuming homogeneity for all water-heaters, the training data we use does not. This data simulates realistic water consumption patterns in diverse households. Interestingly, our experiments reveal that despite this simplification in the model, we can still generate effective policies that meet the target even under more realistic conditions.

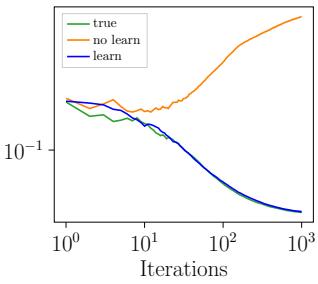
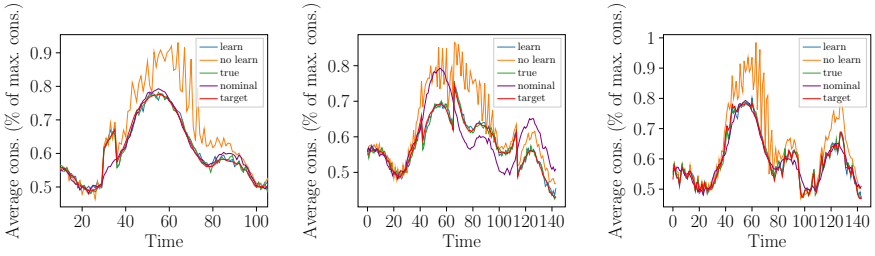


Fig. 9: Log plot of the objective function per iteration for OTT [blue], OTT with known [green] and unknown [red] noise dynamics.



(a) One hour step. (b) Eight hours step. (c) TSO balancing signal.

Fig. 10: Average consumption profile over 10^4 heaters following the policy output by OTT [blue], OTT with known noise distribution [green], and OTT with $\hat{p}_n^t(\cdot|x,a) = \delta_{g(x,a,0)}$ [orange].

7 Conclusion and Future Works

In this paper, we propose to model the problem of controlling a large population of TCLs as a loop-free episodic Markov decision process. We show that this formulation makes it possible to adapt classical optimization algorithms to the load problem achieving closed-form solutions with convergence guarantees. The proposed solution is improved over previous approaches because it requires no additional regularization in the main problem and can be applied to any convex and Lipschitz objective functions. This Markovian formulation also gives rise to more realistic settings, such as the online learning scenario, for which we develop a new algorithm, OTT, dealing with variable target consumption profiles and unknown consumer behavior. We prove OTT achieves $O(\log(T))$ regret, where T is the number of episodes. We also validate our claims using realistic simulations of a population of water-heaters.

A future direction is to adapt our algorithm to the case where probability transitions are not stationary, taking into account sudden changes in human behavior due to abrupt temperature changes or episodes such as the Covid 19 pandemic. Furthermore, one of the limitations of our method is that it only works in finite state spaces, forcing

us to discretize the temperature. Future work also involves generalizing our work to continuous state spaces, using function approximations and/or model-free algorithms.

Appendix A Fictitious Play for MFG Algorithm

Algorithm 3 FP-MFG [7]

Input: number of iterations K, initial policy π^0.
Initialization: $\bar{\mu}^0 = \mu^{\pi^0}$ as in Eq. 5.
for $k = 0, ..., K$ **do**
$\pi^{k+1} \in \arg\max_{\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} \sum_{n=1}^{N} \sum_{x,a} \mu_n^{\pi}(x,a)\big(-\nabla f_n(\bar{\mu}_n^k)(x,a)\big)$, best response against $\bar{\mu}^k$ computed backwards in time.
$\bar{\mu}^{k+1} = \frac{1}{k+1}\mu^{\pi^{k+1}} + \frac{k}{k+1}\bar{\mu}^k$.
end for
Return: $\bar{\mu}^K$ and $\bar{\pi}^K$ s.t. $\bar{\pi}_n^K(a|x) := \sum_{k=0}^{K} \frac{\rho_n^{\pi^k}(x)\pi_n^k(a|x)}{\sum_{k=0}^{K}\rho_n^{\pi^k}(x)}$, $\big(\rho_n^{\pi^k}(x) := \sum_{a\in\mathcal{A}} \mu_n^{\pi^k}(x,a)$ for all $k \leq K\big)$.

Appendix B Proof of Proposition 4.2: Upper Bound of R_T^{policy}

Proof. Recall that an iteration of OTT solves Problem (16) whenever $\gamma^{t+1} = \boldsymbol{\gamma}_j$, i.e.,

$$\begin{aligned} \mu^{t+1} &\in \arg\min_{\mu \in \mathcal{M}_{\mu_0}^{t+1}} \{\tau \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j), \mu\rangle + \Gamma(\mu, \tilde{\nu}^{t_j})\} \\ \nu^{t+1} &\in \arg\min_{\nu \in \mathcal{M}_{\mu_0}^{t+1}} \{\tau \nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j), \nu\rangle + \Gamma(\nu, \tilde{\nu}^{t_j})\}, \end{aligned} \tag{$*$}$$

where we let π^t be the policy inducing μ^t and η^t be an auxiliary policy inducing ν^t, both in $\mathcal{M}_{\mu_0}^t$, i.e., $\mu^t := \mu^{\pi^t, \hat{p}^t}$ and $\nu^t := \nu^{\eta^t, \hat{p}^t}$, in the sense of Eq. (5). Also, recall that $\tilde{\eta}^t := (1-\alpha_t)\eta^t + \frac{\alpha_t}{|\mathcal{A}|}$ for all $t \in [T]$, and that we assume $\alpha_t \in (0, 1/2)$. We also let $\tilde{\nu}^t := \nu^{\tilde{\eta}^t, \hat{p}^t}$.

We let $\|\cdot\|_1$ be the L_1 norm, and for all $v := (v_n)_{n\in[N]}$, such that $v_n \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ we define $\|v\|_{\infty,1} := \sup_{n\in[N]} \|v_n\|_1$. Also, for all $\zeta := (\zeta_n)_{n\in[N]}$ with $\zeta \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, we define by $\|\zeta\|_{1,\infty} := \sup_v\{|\langle\zeta, v\rangle|, \|v\|_{\infty,1} \leq 1\} = \sum_{n=1}^{N} \|\zeta_n\|_\infty$ the respective dual norm.

Preliminaries

: We begin by examining Problem (16) restated in Eq. ($*$). Since F is convex and the set $\mathcal{M}_{\mu_0}^{t+1}$ is also convex, the optimality conditions imply that for all $u \in \mathcal{M}_{\mu_0}^{t+1}$,

$$\begin{aligned} \langle \tau \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j) + \nabla\psi(\mu^{t+1}) - \nabla\psi(\tilde{\nu}^{t_j}), u - \mu^{t+1}\rangle \geq 0 \\ \langle \tau \nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j) + \nabla\psi(\nu^{t+1}) - \nabla\psi(\tilde{\nu}^{t_j}), u - \nu^{t+1}\rangle \geq 0, \end{aligned}$$

where ψ is the function inducing the Bregman divergence Γ defined in Eq. (13).

Rearranging the terms and using the three points identity of Bregman divergences (see Chapter 11.2 of [9]) we get that

$$\begin{aligned}\tau\langle\nabla F(\tilde{\nu}^{t_j};\boldsymbol{\gamma}_j),\mu^{t+1}-u\rangle &\le \langle\nabla\psi(\mu^{t+1})-\nabla\psi(\tilde{\nu}^{t_j}),u-\mu^{t+1}\rangle \\ &= \Gamma(u,\tilde{\nu}^{t_j})-\Gamma(u,\mu^{t+1})-\Gamma(\mu^{t+1},\tilde{\nu}^{t_j}),\end{aligned} \tag{B1}$$

$$\begin{aligned}\tau\langle\nabla F(\mu^{t+1};\boldsymbol{\gamma}_j),\nu^{t+1}-u\rangle &\le \langle\nabla\psi(\nu^{t+1})-\nabla\psi(\tilde{\nu}^{t_j}),u-\nu^{t+1}\rangle \\ &= \Gamma(u,\tilde{\nu}^{t_j})-\Gamma(u,\nu^{t+1})-\Gamma(\nu^{t+1},\tilde{\nu}^{t_j}).\end{aligned} \tag{B2}$$

R_T^{policy} ***analysis:***

We now analyse R_T^{policy}. The regret term is given by

$$\begin{aligned}R_T^{policy} &:= \sum_{t=1}^{T} F(\mu^{\pi^t,\hat{p}^t};\boldsymbol{\gamma}_{v(t)})-F(\mu^{\pi^{*,v(t)},\hat{p}^t};\boldsymbol{\gamma}_{v(t)}) \\ &= \sum_{j=1}^{J}\sum_{t\in[T]:\gamma^t=\boldsymbol{\gamma}_j} F(\mu^{\pi^t,\hat{p}^t};\boldsymbol{\gamma}_j)-F(\mu^{\pi^{*,j},\hat{p}^t};\boldsymbol{\gamma}_j).\end{aligned}$$

Assume $\gamma^{t+1}=\boldsymbol{\gamma}_j$ and define $r_{t+1} := F(\mu^{\pi^{t+1},\hat{p}^{t+1}};\boldsymbol{\gamma}_j)-F(\mu^{\pi^{*,t+1},\hat{p}^{t+1}};\boldsymbol{\gamma}_j)$ the instantaneous regret incurred in R_T^{policy} at day $t+1$. Using the convexity of F, and further decomposing the instantaneous regret we get

$$\begin{aligned}r_{t+1} \le \langle\nabla F(\mu^{t+1};\boldsymbol{\gamma}_j),\mu^{t+1}-\mu^{\pi^{*,j},\hat{p}^{t+1}}\rangle &= \underbrace{\langle\nabla F(\tilde{\nu}^{t_j};\boldsymbol{\gamma}_j),\mu^{t+1}-\nu^{t+1}\rangle}_{(i)} \\ &+\underbrace{\langle\nabla F(\mu^{t+1};\boldsymbol{\gamma}_j),\nu^{t+1}-\mu^{\pi^{*,j},\hat{p}^{t+1}}\rangle}_{(ii)}+\underbrace{\langle\nabla F(\mu^{t+1};\boldsymbol{\gamma}_j)-\nabla F(\tilde{\nu}^{t_j}),\mu^{t+1}-\nu^{t+1}\rangle}_{(iii)}.\end{aligned}$$

We apply the bounds of Eq. (B1) in (i) with $u=\nu^{t+1}$ and the bounds of Eq. (B2) in (ii) with $u=\mu^{\pi^{*,j},\hat{p}^{t+1}}$, therefore,

$$\begin{aligned}r_{t+1} &\le \frac{1}{\tau}\big(\Gamma(\nu^{t+1},\tilde{\nu}^{t_j})-\Gamma(\nu^{t+1},\mu^{t+1})-\Gamma(\mu^{t+1},\tilde{\nu}^{t_j})\big) \\ &\quad+\frac{1}{\tau}\big(\Gamma(\mu^{\pi^{*,j},\hat{p}^{t+1}},\tilde{\nu}^{t_j})-\Gamma(\mu^{\pi^{*,j},\hat{p}^{t+1}},\nu^{t+1})-\Gamma(\nu^{t+1},\tilde{\nu}^{t_j})\big) \\ &\quad+\underbrace{\langle\nabla F(\mu^{t+1};\boldsymbol{\gamma}_j)-\nabla F(\tilde{\nu}^{t_j};\boldsymbol{\gamma}_j),\mu^{t+1}-\nu^{t+1}\rangle}_{(iii)} \\ &= \frac{1}{\tau}\big(\Gamma(\mu^{\pi^{*,j},\hat{p}^{t+1}},\tilde{\nu}^{t_j})-\Gamma(\mu^{\pi^{*,j},\hat{p}^{t+1}},\nu^{t+1})-\Gamma(\nu^{t+1},\mu^{t+1})-\Gamma(\mu^{t+1},\tilde{\nu}^{t_j})\big) \\ &\quad+\underbrace{\langle\nabla F(\mu^{t+1};\boldsymbol{\gamma}_j)-\nabla F(\tilde{\nu}^{t_j};\boldsymbol{\gamma}_j),\mu^{t+1}-\nu^{t+1}\rangle}_{(iii)}.\end{aligned}$$

We now analyse the term (iii). For $\sigma > 0$ to be optimized later, Young's inequality yields

$$\begin{aligned}(iii) &= \langle \nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j) - \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j), \mu^{t+1} - \nu^{t+1} \rangle \\ &\leq \frac{\sigma}{2} \|\nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j) - \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j)\|_{1,\infty}^2 + \frac{1}{2\sigma} \|\mu^{t+1} - \nu^{t+1}\|_{\infty,1}^2.\end{aligned}$$

In [6], it is shown that the function ψ inducing the Bregman divergence Γ defined in Eq. (10) is 1-strongly convex with respect to the norm $\|\cdot\|_{\infty,1}$, therefore,

$$-\Gamma(\nu^{t+1}, \mu^{t+1}) \leq -\frac{1}{2} \|\mu^{t+1} - \nu^{t+1}\|_{\infty,1}^2.$$

Thus, by taking $\sigma = \tau$ in the Young's inequality we get that

$$(iii) - \frac{1}{\tau} \Gamma(\nu^{t+1}, \mu^{t+1}) \leq \frac{\tau}{2} \|\nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j) - \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j)\|_{1,\infty}^2.$$

Substituting this inequality in the instantaneous regret bound we get

$$\begin{aligned}r_{t+1} \leq \frac{1}{\tau} \big(&\Gamma(\mu^{\pi^*, j, \hat{p}^{t+1}}, \tilde{\nu}^{t_j}) - \Gamma(\mu^{\pi^*, j, \hat{p}^{t+1}}, \nu^{t+1}) - \Gamma(\mu^{t+1}, \tilde{\nu}^{t_j}) \big) \\ &+ \frac{\tau}{2} \|\nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j) - \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j)\|_{1,\infty}^2.\end{aligned}$$

Using that F is β-smooth with respect to $\|\cdot\|_{\infty,1}$, we get that

$$\|\nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j) - \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j)\|_{1,\infty}^2 \leq \beta^2 \|\mu^{t+1} - \tilde{\nu}^{t_j}\|_{1,\infty}^2.$$

Using again the strong convexity of the function ψ inducing the Bregamn divergence Γ,

$$-\Gamma(\mu^{t+1}, \tilde{\nu}^{t_j}) \leq -\frac{1}{2} \|\mu^{t+1} - \tilde{\nu}^{t_j}\|_{\infty,1}^2.$$

Taking $\tau = \frac{1}{\beta}$,

$$\begin{aligned}&-\frac{1}{\tau} \Gamma(\mu^{t+1}, \tilde{\nu}^{t_j}) + \frac{\tau}{2} \|\nabla F(\mu^{t+1}; \boldsymbol{\gamma}_j) - \nabla F(\tilde{\nu}^{t_j}; \boldsymbol{\gamma}_j)\|_{1,\infty}^2 \\ &\quad - \leq \frac{\beta}{2} \|\mu^{t+1} - \tilde{\nu}^{t_j}\|_{\infty,1}^2 + \frac{\beta}{2} \|\mu^{t+1} - \tilde{\nu}^{t_j}\|_{1,\infty}^2 = 0.\end{aligned}$$

Replacing this inequality in the instantaneous regret bound we obtain that

$$r_{t+1} \leq \beta \big(\Gamma(\mu^{\pi^*, j, \hat{p}^{t+1}}, \tilde{\nu}^{t_j}) - \Gamma(\mu^{\pi^*, j, \hat{p}^{t+1}}, \nu^{t+1}) \big).$$

Summing over all days,

$$R_T^{policy} \leq \beta \sum_{j=1}^{J} \sum_{t\in[T]:\gamma^t=\boldsymbol{\gamma}_j} \left(\Gamma(\mu^{\pi^{*,j},\hat{p}^t}, \tilde{\nu}^{(t-1)_j}) - \Gamma(\mu^{\pi^{*,j},\hat{p}^t}, \nu^t)\right)$$
$$= \beta \sum_{j=1}^{J} \sum_{s=1}^{T(j)} \left(\Gamma(\mu^{\pi^{*,j},\hat{p}^{j_s}}, \tilde{\nu}^{j_{s-1}}) - \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_s}}, \nu^{j_s})\right),$$

where $T(j)$ is the number of days that target j appeared until the end of day T, and j_s represents the day of the s-th occurrence of target j. In order to make this sum telescope we add and subtract $\Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \tilde{\nu}^{j_{s-1}})$,

$$R_T^{policy} \leq \beta \sum_{j=1}^{J} \underbrace{\sum_{s=1}^{T(j)} \left(\Gamma(\mu^{\pi^{*,j},\hat{p}^{j_s}}, \tilde{\nu}^{j_{s-1}}) - \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \tilde{\nu}^{j_{s-1}})\right)}_{A^j}$$
$$+ \beta \sum_{j=1}^{J} \underbrace{\sum_{s=1}^{T(j)} \left(\Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \tilde{\nu}^{j_{s-1}}) - \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_s}}, \nu^{j_s})\right)}_{B^j}.$$

We analyse the sum for each $j \in [J]$.

Step* 1*: Upper bound on A^j

Using the definition of a Bregman divergence induced by ψ we get that

$$A^j = \sum_{s=1}^{T(j)} \psi(\mu^{\pi^{*,j},\hat{p}^{j_s}}) - \psi(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}) - \langle \nabla\psi(\tilde{\nu}^{j_{s-1}}), \mu^{\pi^{*,j},\hat{p}^{j_s}} - \mu^{\pi^{*,j},\hat{p}^{j_{s-1}}} \rangle$$
$$\leq -\psi(\mu^{\pi^{*,j},\hat{p}^0}) + \sum_{s=1}^{T(j)} \|\nabla\psi(\tilde{\nu}^{j_{s-1}})\|_{1,\infty} \|\mu^{\pi^{*,j},\hat{p}^{j_s}} - \mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}\|_{\infty,1},$$

where we use that the first term telescopes, and we apply Holder's inequality to the second term. With the definition of Γ in Eq. (10), and given the definition of $\tilde{\eta}$ in Eq. (17), for each $n \in [N], (x,a) \in \mathcal{X} \times \mathcal{A}$, $|\nabla\psi(\tilde{\nu}^s)(n,x,a)| = |\log(\tilde{\eta}_n^s(a|x))| \leq \log(|\mathcal{A}|/\alpha_s)$ for all $s \in [T]$. From Lemma 5.6 of [6], we get that $\|\mu^{\pi^{*,j},\hat{p}^{j_s}} - \mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}\|_{\infty,1} \leq \frac{2N(j_s - j_{s-1})}{j_s}$. Plugging both results in the bound of A^j we get

$$\begin{aligned} A^j &\leq -\psi(\mu^{\pi^{*,j},\hat{p}^0}) + \sum_{s=1}^{T(j)} \log\left(\frac{|\mathcal{A}|}{\alpha_{j_s}}\right) \|\mu^{\pi^{*,j},\hat{p}^{j_s}} - \mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}\|_{\infty,1} \\ &\leq -\psi(\mu^{\pi^{*,j},\hat{p}^0}) + 2N^2 \sum_{s=1}^{T(j)} \log\left(\frac{|\mathcal{A}|}{\alpha_{j_s}}\right) \frac{(j_s - j_{s-1})}{j_s}. \end{aligned} \tag{B3}$$

Step 2: Upper bound on B^j

We start by adding and subtracting $\Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \nu^{j_{s-1}})$,

$$B^j = \underbrace{\sum_{s=1}^{T(j)} \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \tilde{\nu}^{j_{s-1}}) - \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \nu^{j_{s-1}})}_{B^j_{(i)}}$$
$$+ \underbrace{\sum_{s=1}^{T(j)} \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \nu^{j_{s-1}}) - \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_s}}, \nu^{j_s})}_{B^j_{(ii)}}.$$

We analyse each term from B's decomposition:
Using the definition of Γ in Eq. (10), we obtain that

$$\begin{aligned}
B^j_{(i)} &= \sum_{s=1}^{T(j)} \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \tilde{\nu}^{j_{s-1}}) - \Gamma(\mu^{\pi^{*,j},\hat{p}^{j_{s-1}}}, \nu^{j_{s-1}}) \\
&= \sum_{s=1}^{T(j)} \sum_{n,x,a} \mu_n^{\pi^{*,j},\hat{p}^{j_{s-1}}}(x,a) \left[\log\left(\frac{\pi_n^{*,j}(a|x)}{\tilde{\eta}_n^{j_{s-1}}(a|x)} \right) - \log\left(\frac{\pi_n^{*,j}(a|x)}{\eta_n^{j_{s-1}}(a|x)} \right) \right] \\
&= \sum_{s=1}^{T(j)} \sum_{n,x,a} \mu_n^{\pi^{*,j},\hat{p}^{j_{s-1}}}(x,a) \log\left(\frac{\eta_n^{j_{s-1}}(a|x)}{\tilde{\eta}_n^{j_{s-1}}(a|x)} \right) \\
&= \sum_{s=1}^{T(j)} \sum_{n,x,a} \mu_n^{\pi^{*,j},\hat{p}^{j_{s-1}}}(x,a) \log\left(\frac{\eta_n^{j_{s-1}}(a|x)}{(1-\alpha_{j_s})\eta_n^{j_{s-1}}(a|x) + \alpha_{j_s}/|\mathcal{A}|} \right) \\
&\leq N \sum_{s=1}^{T(j)} (-\log(1-\alpha_{j_s})) \leq 2N \sum_{s=1}^{T(j)} \alpha_{j_s},
\end{aligned}$$

where the last inequality is valid if $0 \leq \alpha_{j_s} \leq 1/2$.

It is easy to see that the term $B^j_{(ii)}$ telescopes, therefore

$$B^j_{(ii)} \leq \Gamma(\mu^{\pi^{*,j},\hat{p}^0}, \nu^0).$$

Plugging all into B^j's upper bound, we obtain that

$$B^j \leq 2N \sum_{s=1}^{T(j)} \alpha_{j_s} + \Gamma(\mu^{\pi^{*,j},\hat{p}^0}, \nu^0). \tag{B4}$$

Final step

By replacing the bounds on Eq.s (B3) and (B4), and using the result from Lemma $D.2$ of [6] that $\Gamma(\mu^{\pi^{*,j},\hat{p}^0}, \nu^0) - \psi(\mu^{\pi^{*,j},\hat{p}^0}) \leq N \log(|\mathcal{A}|)$, we obtain that

$$\begin{aligned} R_T^{policy} &\leq \beta \sum_{j=1}^{J} (A^j + B^j) \\ &\leq \beta \sum_{j=1}^{J} \left(2N^2 \sum_{s=1}^{T(j)} \log\left(\frac{|\mathcal{A}|}{\alpha_{j_s}}\right) \frac{(j_s - j_{s-1})}{j_s} + 2N \sum_{s=1}^{T(j)} \alpha_{j_s} + N \log(|\mathcal{A}|) \right). \end{aligned}$$

Note that

$$\begin{aligned} \sum_{s=1}^{T(j)} \frac{j_s - j_{s-1}}{j_s} &= \sum_{s=1}^{T(j)} \int_{j_{s-1}}^{j_s} \frac{1}{j_s} du \\ &\leq \sum_{s=2}^{T(j)} \int_{j_{s-1}}^{j_s} \frac{1}{u} du + 1 \\ &\leq \log\left(\frac{j_{T(j)}}{j_1}\right) + 1 \\ &\leq 1 + \log(T). \end{aligned}$$

Therefore, if we further take $\alpha_t = 1/T$ for all $t \in [T]$, we obtain that

$$\begin{aligned} R_T^{policy} &\leq \beta \sum_{j=1}^{J} \left(2N^2 \log(|\mathcal{A}|T) \log(T) + 2N \frac{T(j)}{T} + N \log(|\mathcal{A}|) \right) \\ &\leq \beta \left(2N^2 \log(|\mathcal{A}|T) J \log\left(T\right) + 2N + JN \log(|\mathcal{A}|) \right), \end{aligned}$$

where we use that $\sum_{j=1}^{J} T(j) = T$. □ □

Appendix C Numerical parameters for simulations

Table C1: 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 C2: Other parameters specifications to compute Eq. 22.

denWater (water density)	10^3 kg m^{-3}
capWater (water capacity)	4185 J kg^{-1} K^{-1}
CI (heat conductivity)	0.033 W/(m K)
coefLoss (loss coeff.)	$\frac{\text{CI}}{\text{EI}} * 2 * 3.14\sqrt{\frac{\text{vol}*3.14}{\text{height}}}$

References

[1] Ritchie, H., Roser, M., Rosado, P.: Co2 and greenhouse gas emissions. Our World in Data (2020). https://ourworldindata.org/co2-and-greenhouse-gas-emissions

[2] Bakare, M.S., Abdulkarim, A., Zeeshan, M., Shuaibu, A.N.: A comprehensive overview on demand side energy management towards smart grids: challenges, solutions, and future direction. Energy Informatics **6**(1), 4 (2023)

[3] Antonopoulos, I., Robu, V., Couraud, B., Kirli, D., Norbu, S., Kiprakis, A., Flynn, D., Elizondo-Gonzalez, S., Wattam, S.: Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. Renewable and Sustainable Energy Reviews **130**(C) (2020)

[4] Brégère, M., Gaillard, P., Goude, Y., Stoltz, G.: Target tracking for contextual bandits: Application to demand side management. In: Proceedings of the 36th International Conference on Machine Learning (ICML), vol. 97, pp. 754–763 (2019)

[5] López, K.L., Gagné, C., Gardner, M.-A.: Demand-side management using deep learning for smart charging of electric vehicles. IEEE Transactions on Smart Grid **10**(3), 2683–2691 (2019)

[6] Marin Moreno, B., Gaillard, P., Oudjane, N., Brégère, M.: Efficient model-based concave utility reinforcement learning through greedy mirror descent. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2024)

[7] Perrin, S., Perolat, J., Lauriere, M., Geist, M., Elie, R., Pietquin, O.: Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications. In: Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), vol. 33, pp. 13199–13213 (2020)

[8] Cammardella, N., Bušić, A., Meyn, S.: Kullback-leibler-quadratic optimal control in a stochastic environment. In: Proceedings of the IEEE Conference on Decision and Control (CDC), pp. 158–165 (2021)

[9] Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, UK (2006)

[10] Rakhlin, A., Sridharan, K.: Optimization, learning, and games with predictable sequences. In: Proceedings of the 26th International Conference on Neural Information Processing Systems (NeurIPS), pp. 3066–3074 (2013)

[11] Albouys, J., Sabouret, N., Haradji, Y., Schumann, M., Inard, C.: SMACH: Multi-agent Simulation of Human Activity in the Household. In: Advances in Practical Applications of Survivable Agents and Multi-Agent Systems: The PAAMS Collection, pp. 227–231. Springer, Cham (2019)

[12] Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming, 1st edn. John Wiley & Sons, Inc., USA (1994)

[13] Lasry, J.-M., Lions, P.-L.: Mean field games. Japanese Journal of Mathematics **2**, 229–260 (2007)

[14] Huang, M., Malhame, R., Caines, P.: Large population stochastic dynamic games: Closed-loop mckean-vlasov systems and the nash certainty equivalence principle. Commun. Inf. Syst. **6** (2006)

[15] Ihara, S., Schweppe, F.C.: Physically based modeling of cold load pickup. IEEE Transactions on Power Apparatus and Systems **PAS-100**, 4142–4150 (1981)

[16] Malhame, R., Chong, C.-Y.: 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)

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

[18] Burger, E.M., Moura, S.J.: Generation following with thermostatically controlled

loads via alternating direction method of multipliers sharing algorithm. Electric Power Systems Research **146**, 141–160 (2017)

[19] Kim, S.-J., Giannakis, G.B.: Scalable and robust demand response with mixed-integer constraints. IEEE Transactions on Smart Grid **4**(4), 2089–2099 (2013)

[20] Franceschelli, M., Pilloni, A., Gasparri, A.: Multi-agent coordination of thermostatically controlled loads by smart power sockets for electric demand side management. IEEE Transactions on Control Systems Technology **29**(2), 731–743 (2021)

[21] Seguret, A., Oudjane, N., Le Corre, T.: A decentralized algorithm for a mean field control problem of piecewise deterministic markov processes. ESAIM: Probability and Statistics **28** (2023)

[22] Coffman, A., Bušić, A., Barooah, P.: A unified framework for coordination of thermostatically controlled loads. Automatica **152**, 111002 (2023)

[23] Kizilkale, A., Malhame, R.: 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 (CDC), pp. 4971–4976 (2013)

[24] Kizilkale, A.C., Malhame, R.P.: Collective target tracking mean field control for markovian jump-driven models of electric water heating loads. IFAC Proceedings Volumes **47**(3), 1867–1872 (2014)

[25] Bušić, A., Meyn, S.: Distributed randomized control for demand dispatch. In: IEEE 55th Conference on Decision and Control (CDC), pp. 6964–6971 (2016)

[26] Chen, Y., Hashmi, M.U., Mathias, J., Bušić, A., Meyn, S.: In: Meyn, S., Samad, T., Hiskens, I., Stoustrup, J. (eds.) Distributed Control Design for Balancing the Grid Using Flexible Loads, pp. 383–411. Springer, New York, NY (2018)

[27] Cammardella, N., Bušić, A., Ji, Y., Meyn, S.: Kullback-leibler-quadratic optimal control of flexible power demand. In: IEEE 58th Conference on Decision and Control (CDC), pp. 4195–4201 (2019)

[28] Hazan, E., Kakade, S., Singh, K., Van Soest, A.: Provably efficient maximum entropy exploration. In: International Conference on Machine Learning (ICML), vol. 97, pp. 2681–2691 (2019)

[29] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA (2018)

[30] Zhang, J., Koppel, A., Bedi, A.S., Szepesvari, C., Wang, M.: Variational policy gradient method for reinforcement learning with general utilities. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural

Information Processing Systems (NeurIPS), vol. 33, pp. 4572–4583 (2020)

[31] Laurière, M., Perrin, S., Geist, M., Pietquin, O.: Learning Mean Field Games: A Survey (2022)

[32] Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., Pietquin, O.: Concave utility reinforcement learning: The mean-field game viewpoint. In: International Conference on Autonomous Agents and Multiagent Systems, pp. 489–497 (2022)

[33] Rosenberg, A., Mansour, Y.: Online convex optimization in adversarial Markov decision processes. In: Proceedings of the 36th International Conference on Machine Learning (ICML), vol. 97, pp. 5478–5486 (2019)

[34] Pasztor, B., Bogunovic, I., Krause, A.: Efficient model-based multi-agent mean-field reinforcement learning. Trans. Mach. Learn. Res. **2023** (2021)

[35] Bonnans, J.F., Lavigne, P., Pfeiffer, L.: Discrete potential mean field games: duality and numerical resolution. Mathematical Programming **202**, 241–278 (2023)

[36] Hazan, E.: Introduction to Online Convex Optimization, 2nd edn. Adaptive Computation and Machine Learning series. MIT Press, USA (2022)