Session 2 : Approches mathématiques

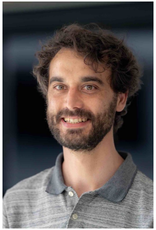
Online Convex Optimisation for Demand-Side Management

Pierre Gaillard (LJK)

Collaborators

This work was carried out as part of the Cifre PhD at EDF R&D of B. Marin Moreno.

Bianca M. Moreno



Pierre Gaillard
Inria, LJK

Margaux Brégère
EDF R&D

Nadia Oudjane
EDF R&D

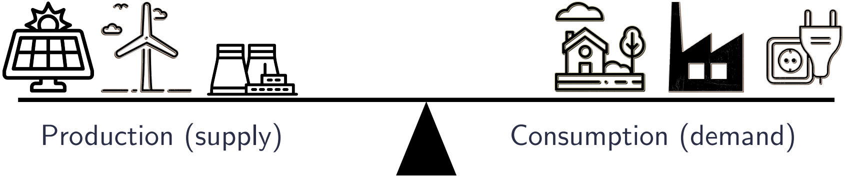
Thanks to Bianca for many of these slides.

EnergyAlps
Université Grenoble Alpes

Balancing the power grid

Electricity is hard to store

$\longrightarrow$ production - demand balance must be maintained

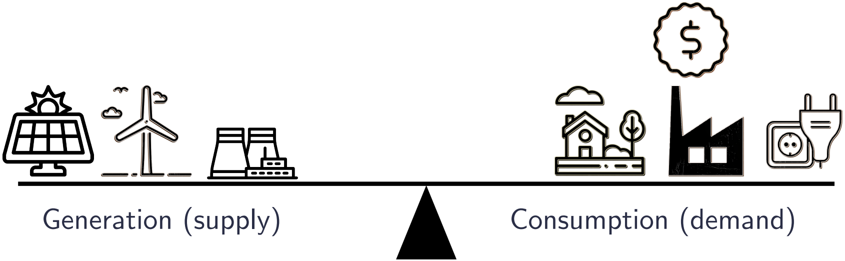


Current solution: forecast demand and adapt production accordingly

- Adapting production is hard:
  - Integration of renewable energy $\rightarrow$ intermittent nature
  - Energy importation $\rightarrow$ costly alternative

Demand-Side Management

- Prospective solutions: manage demand instead



- Send incentive signals (prices)
- Control flexible devices

Control of flexible devices

- TCLs: Thermostatically Controlled Loads
  - Electrical heating or cooling elements controlled by a thermostat: water-heaters, ar conditioners, refrigerators, etc
  - Flexible loads
- New Smart meters
  - Access to data and instantaneous communication

Control of flexible devices: example of water heaters

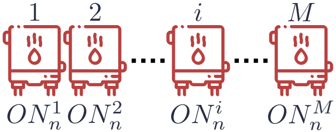
Goal: Control the average consumption of $M \gg 1$ water-heaters

Control of flexible devices: example of water heaters

Goal: Control the average consumption of $M \gg 1$ water-heaters

Individual consumption (time step n)

Average consumption (time step n)

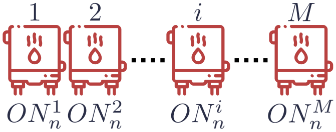


$$\Longrightarrow \frac{1}{M}\sum_{i=1}^{M} ON_n^i$$

Control of flexible devices: example of water heaters

Goal: Control the average consumption of $M \gg 1$ water-heaters

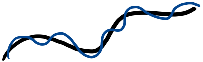
Individual consumption (time step n)



Average consumption (time step n)

$$\Longrightarrow \frac{1}{M}\sum_{i=1}^{M} ON_n^i$$

in order to track a reference consumption profile (γ_n) by sending a control signal (π_n)

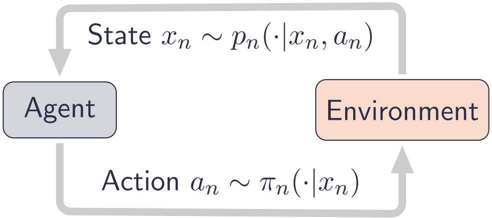


$$\pi_n \Longrightarrow \begin{cases} \text{device } 1 \to ON_n^1 \\ \vdots \\ \text{device } i \to ON_n^i \\ \vdots \\ \text{device } M \to ON_n^M \end{cases} \Longrightarrow \underbrace{\frac{1}{M}\sum_{i=1}^{M} ON_n^i}_{\text{average cons.}} \approx \underbrace{\gamma_n}_{\text{target}}$$

Setting and Model

Episodic Markov Decision Processes

- $(\mathcal{X}, \mathcal{A})$ finite state and action spaces; episodes of length N
- Agent starts at $(x_0, a_0) \sim \mu_0(\cdot)$
- At step $n \in \{1, \ldots, N\}$:
  - The agent: observes state x_n and takes action $a_n \sim \pi_n(\cdot|x_n)$
  - The environment: generates next state $x_{n+1} \sim p_n(\cdot|x_n, a_n)$

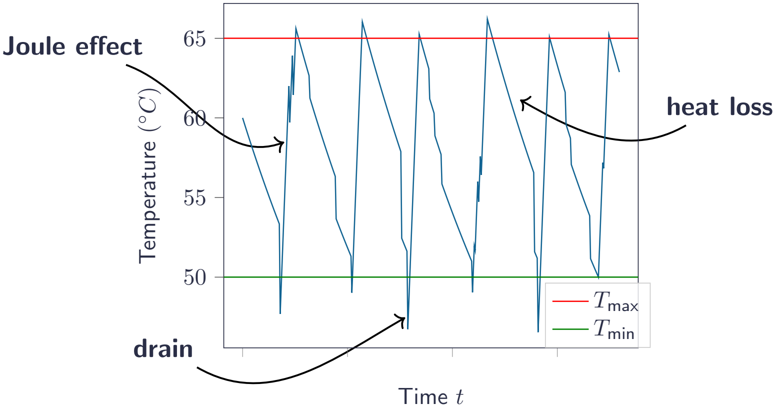


State-action distribution sequence:

$$\mu_n^{\pi,p}(x, a) = \mathbb{P}(x_n = x, a_n = a|\pi, p)$$

Water-heater uncontrolled dynamics

- $[T_{\min}, T_{\max}]$ = temperature deadband

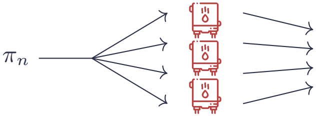


Example: water-heater modeled as an episodic MDP

- Agent $=$ water-heater; $x = (ON$, temperature); $n =$ an hour; $a =$ to turn/keep on/off;
- Dynamics: $x_{n+1} \sim p_n(\cdot | x_n, a_n)$ probability of water withdraws (showers, etc), and quality of service

Example: water-heater modeled as an episodic MDP

- Agent $=$ water-heater; $x = (ON,$ temperature); $n =$ an hour; $a =$ to turn/keep on/off;
- Dynamics: $x_{n+1} \sim p_n(\cdot | x_n, a_n)$ probability of water withdraws (showers, etc), and quality of service
- M water-heaters



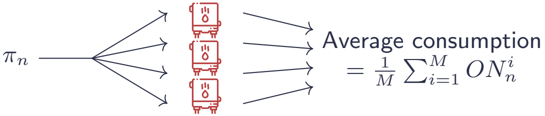
Average consumption $= \frac{1}{M} \sum_{i=1}^{M} ON_n^i$

- Goal: Given a consumption profile target γ, compute π such that

$$\frac{1}{M} \sum_{i=1}^{M} ON_n^i(\pi) \approx \gamma_n$$

Example: water-heater modeled as an episodic MDP

- Agent $=$ water-heater; $x = (ON$, temperature); $n =$ an hour; $a =$ to turn/keep on/off;
- Dynamics: $x_{n+1} \sim p_n(\cdot|x_n, a_n)$ probability of water withdraws (showers, etc), and quality of service
- M water-heaters



- Goal: Given a consumption profile target γ, compute π such that

$$\frac{1}{M}\sum_{i=1}^{M} ON_n^i(\pi) \approx \gamma_n$$

- Optimization problem:

$$\min_{\pi \in \Pi} \mathbb{E}\left[\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M} ON_n^i(\pi) - \gamma_n\right)^2\right]$$

Mean-field approximation

- Optimization problem:

$$\min_{\pi \in \Pi} \mathbb{E}\left[\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M} ON_n^i(\pi) - \gamma_n\right)^2\right]$$

Mean-field approximation

- Optimization problem:

$$\min_{\pi \in \Pi} \mathbb{E}\left[\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M} ON_n^i(\pi) - \gamma_n\right)^2\right]$$

- Mean Field Limit ($M \gg 1$):

$$\frac{1}{M}\sum_{i=1}^{M} ON_n^i(\pi) \xrightarrow[M\to\infty]{} \mathbb{E}_{\mu_n^{\pi,p}}[ON_n]$$

Mean-field approximation

- Optimization problem:

$$\min_{\pi\in\Pi} \mathbb{E}\left[\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M} ON_n^i(\pi) - \gamma_n\right)^2\right]$$

- Mean Field Limit ($M \gg 1$):

$$\frac{1}{M}\sum_{i=1}^{M} ON_n^i(\pi) \xrightarrow[M\to\infty]{} \mathbb{E}_{\mu_n^{\pi,p}}[ON_n]$$

- Goal after mean-field limit:

$$\min_{\pi\in\Pi}\left\{F(\mu^{\pi,p}) := \sum_{n=1}^{N}\left(\mathbb{E}_{\mu_n^{\pi,p}}[ON_n] - \gamma_n\right)^2\right\}$$

EnergyAlps
Université Grenoble Alpes

Reinforcement Learning (RL)

- Environment generates a loss $\ell = (\ell_n)_{n\in[N]}$, $\ell_n : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

RL $\min_{\pi\in\Pi} \mathbb{E}\left[\sum_{n=1}^{N} \ell_n(x_n, a_n) \middle| \pi, p \right]$

Convex RL (CURL)

- For any convex loss F over the state-action distribution

CURL $\min_{\pi\in\Pi} F(\mu^{\pi,p})$

Reinforcement Learning (RL)

- Environment generates a loss $\ell = (\ell_n)_{n\in[N]}$, $\ell_n : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

$$\text{RL} \min_{\pi\in\Pi} \langle \ell, \mu^{\pi,p} \rangle$$

Convex RL (CURL)

- For any convex loss F over the state-action distribution

$$\text{CURL} \min_{\pi\in\Pi} F(\mu^{\pi,p})$$

Examples:

- Pure RL exploration: $F(\mu^{\pi,p}) := \langle \mu^{\pi,p}, \log(\mu^{\pi,p}) \rangle$
- Imitation learning: $F(\mu^{\pi,p}) := D_f(\mu^{\pi,p}, \mu^*)$, where D_f is a Bregman divergence induced by a function f

Reinforcement Learning (RL)

- Environment generates a loss $\ell = (\ell_n)_{n \in [N]}$, $\ell_n : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

RL $\min_{\pi \in \Pi} \langle \ell, \mu^{\pi,p} \rangle$

Convex RL (CURL)

- For any convex loss F over the state-action distribution

CURL $\min_{\pi \in \Pi} F(\mu^{\pi,p})$

Examples:

- Pure RL exploration: $F(\mu^{\pi,p}) := \langle \mu^{\pi,p}, \log(\mu^{\pi,p}) \rangle$
- Imitation learning: $F(\mu^{\pi,p}) := D_f(\mu^{\pi,p}, \mu^*)$, where D_f is a Bregman divergence induced by a function f

CURL's non-linearity invalidates classical Bellman equations requiring new algorithms

Today:

- Optimization task: how to solve

$$\min_{\pi \in \Pi} F(\mu^{\pi,p})$$

when $p = (p_n)_n$ and F are known and fixed?

Today:

- Optimization task: how to solve

$$\min_{\pi \in \Pi} F(\mu^{\pi,p})$$

when $p = (p_n)_n$ and F are known and fixed?

- Online Learning task: Real-world tasks are non-stationary!

Today:

- Optimization task: how to solve

$$\min_{\pi\in\Pi} F(\mu^{\pi,p})$$

  when $p = (p_n)_n$ and F are known and fixed?
- Online Learning task: Real-world tasks are non-stationary!
  - Fluctuations of energy production: $\Rightarrow$ changing F^t

Today:

- Optimization task: how to solve

$$\min_{\pi\in\Pi} F(\mu^{\pi,p})$$

  when $p = (p_n)_n$ and F are known and fixed?
- Online Learning task: Real-world tasks are non-stationary!
  - Fluctuations of energy production: $\Rightarrow$ changing F^t
  - Unknown and non-stationary consumer behavior: $\Rightarrow$ changing p^t

Today:

- Optimization task: how to solve

$$\boxed{\min_{\pi \in \Pi} F(\mu^{\pi, p})}$$

when $p = (p_n)_n$ and F are known and fixed?

- Online Learning task: Real-world tasks are non-stationary!
  - Fluctuations of energy production: $\Rightarrow$ changing F^t
  - Unknown and non-stationary consumer behavior: $\Rightarrow$ changing p^t
- How to compute $(\pi^t)_{t \in [T]}$ minimizing

$$\boxed{\sum_{t=1}^{T} F^t(\mu^{\pi^t, p^t})}$$

when $p = (p_n)_n$ and F are unknown and may change?

EnergyAlps
Université Grenoble Alpes

Optimization task: offline CURL

Problem reformulation

$$\min_{\pi \in \Pi} F(\mu^{\pi,p})$$

⚠ gradient on π? convexity?

Problem reformulation

$$\min_{\pi \in \Pi} F(\mu^{\pi,p})$$

⚠ gradient on π? convexity?

$$\implies \min_{\mu \in ?} F(\mu)$$

☺ gradient on μ! convexity!

Problem reformulation

$$\min_{\pi \in \Pi} F(\mu^{\pi,p})$$

gradient on π? convexity?

$$\implies \min_{\mu \in ?} F(\mu)$$

gradient on μ! convexity!

$$\mathcal{M}^p_{\mu_0} := \left\{ (\mu_n)_n \middle| \sum_{a'} \mu_n(x', a') = \sum_{x,a} p_n(x'|x,a)\mu_{n-1}(x,a) \right\}$$

$\mu \in \mathcal{M}^p_{\mu_0} \longrightarrow \pi \in \Pi$ such that $\mu^{\pi,p} = \mu$

Iterative scheme

- Consider the following iterative scheme at iteration k

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

- where Γ is a non-standard regularization

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

Iterative scheme

- Consider the following iterative scheme at iteration k

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

- where Γ is a non-standard regularization

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

- Dynamic Programming yielding in a simple closed-form solution for (1): $\mu^{k+1} := \mu^{\pi^{k+1}}$ such that

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

Convergence analysis

Theorem

Let π^* be a minimizer of Offline CURL and K be the number of iterations, thus

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

Convergence analysis

Theorem

Let π^* be a minimizer of Offline CURL and K be the number of iterations, thus

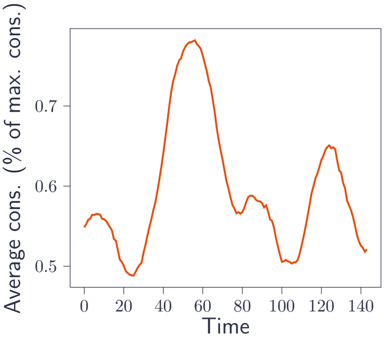
$$\min_{0 \le s \le K} F(\mu^{\pi^s}) - F(\mu^{\pi^*}) \le O\Big(\frac{\sqrt{\Gamma(\mu^{\pi^*}, \mu^0)}}{\sqrt{K}}\Big)$$

Proof idea:

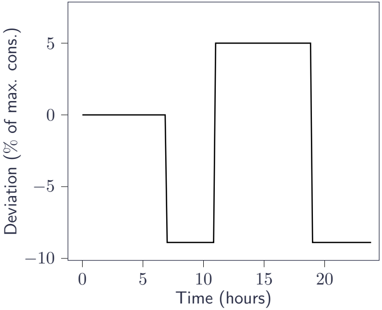
- Show Γ is a Bregman divergence and is 1-strongly convex with respect to the $\sup_{1 \le n \le N} \|\cdot\|_1$ norm
- The classic convergence proof of mirror descent applies Beck and Teboulle 2003

EnergyAlps Université Grenoble Alpes

Target = uncontrolled dynamics + deviation



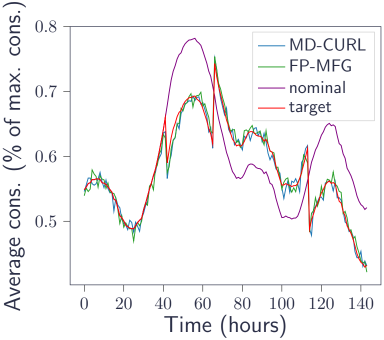
(a) Average consumption.



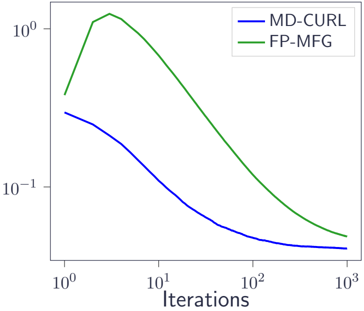
(b) Eight hours step deviation signal.

- Nb of water-heaters = 10^4
- Time horizon = one day
- Time step = 10 minutes

Results



(a) Consumption simulation



(b) Objective function

- FP-MFG = Fictitious Play for mean field games Perrin et al. 2020

Online learning task: unknown but fixed dynamics

Online setting: unknown but fixed dynamics

Assumptions:

- T episodes
- Unknown, but fixed, dynamics $(p_n)_{n\in[N]}$
- Adversarial objective functions F^t

Online Protocol

- π^1 *initial policy,* μ_0 *initial state-action distribution:*
- *for each episode* $t \in \{1, \ldots, T\}$*:*
  - $(x_0^t, a_0^t) \sim \mu_0(\cdot)$
  - *for each time step* $n \in \{1, \ldots, N\}$*:*
    - *agent moves to* $x_n^t \sim p_n(\cdot|x_{n-1}^t, a_{n-1}^t)$
    - *choose* $a_n^t \sim \pi_n^t(\cdot|x_n^t)$
  - *observe* F^t *(full-information)*
  - *update probability transition estimate* $\hat{p}^{t+1}$
  - *compute next policy* π^{t+1}

Questions:

- How to compute the probability transition estimate $\hat{p}^t$?
- How to compute the next policy π^{t+1}?

Exploration: play a policy that explores

Exploitation: play the current optimal policy

Questions:

- How to compute the probability transition estimate $\hat{p}^t$?
- How to compute the next policy π^{t+1}?

Exploration: play a policy that explores

Exploitation: play the current optimal policy

Performance measure: Minimize the regret

$$R_T(\pi) := \sum_{t=1}^{T} F^t(\mu^{\pi^t,p}) - \min_{\pi\in\Pi} \sum_{t=1}^{T} F^t(\mu^{\pi,p}).$$

Computing $\hat{p}^t$

- $N_n^t(x,a) = \#(x,a)$ is visited at time step n up to episode t
- $M_n^t(x'|x,a) = \#$ event above is followed by a transition to x'

$$\hat{p}_n^t(x'|x,a) = \frac{M_n^t(x'|x,a)}{\max\left\{1, N_n^t(x,a)\right\}}$$

Proposition (Neu, Gyorgy, and Szepesvari 2012)

For any $\delta \in (0,1)$

$$\|p_n(\cdot|x,a) - \hat{p}_n^t(\cdot|x,a)\|_1 \leq \sqrt{\frac{4|\mathcal{X}|\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)}{\max\left\{1, N_n^t(x,a)\right\}}}$$

holds, with probability at least $1-\delta$ *simultaneously for all* (n,x,a,t).

Computing π^{t+1}

Mirror descent with $\hat{p}^{t+1}$ (as in Offline CURL)

$$\mu^{t+1} := \arg\min_{\mu \in \mathcal{M}_{\mu_0}^{\hat{p}^{t+1}}} \left\{ \tau \langle \nabla F^t(\mu^{\pi^t, \hat{p}^t}), \mu \rangle + \Gamma(\mu, \mu^t) \right\}$$

Computing π^{t+1}

Define a bonus vector:

$$b_n^t(x,a) \propto \frac{1}{\sqrt{\max\{1, N_n^{t+1}(x,a)\}}}.$$

Computing π^{t+1}

Define a bonus vector:

$$b_n^t(x,a) \propto \frac{1}{\sqrt{\max\{1, N_n^{t+1}(x,a)\}}}.$$

Solve at each episode

$$\mu^{t+1} := \underset{\mu \in \mathcal{M}_{\mu_0}^{\hat{p}^{t+1}}}{\arg\min} \left\{ \tau \langle \nabla F^t(\mu^{\pi^t,\hat{p}^t}) - b^t, \mu \rangle + \Gamma(\mu, \mu^t) \right\}$$

Regret analysis

Theorem (Online CURL with exploration)

With probability at least $1-\delta$, Mirror Descent with the exploration bonus achieves

$$R_T(\pi) = \tilde{O}\big(LN^3|\mathcal{X}|\sqrt{|\mathcal{A}|T}\big)$$

Main challenges:

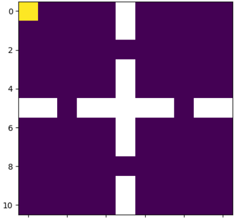
- Mirror descent with changing constraint sets
- Building the exploration bonus

Environment

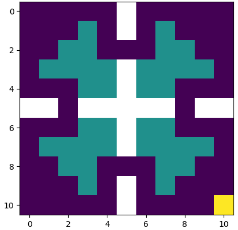
- 11×11 four-room grid world
- Actions $=$ up, down, left, right, still
- ε_n $=$ external noise

$$x_{n+1} = x_n + a_n + \varepsilon_n$$

Objective:



(a) Initial distribution



(b) Objective (reward in yellow, constraints in blue)

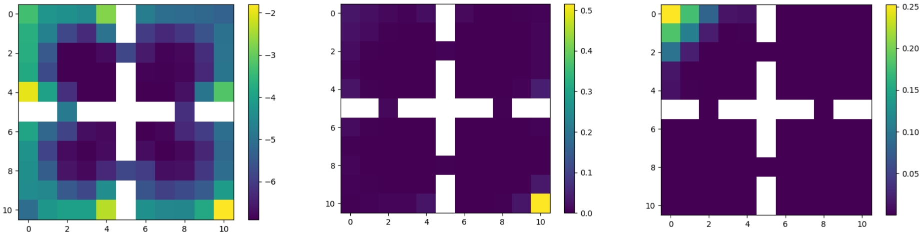
Results

Constrained MDP task after 1000 iterations.

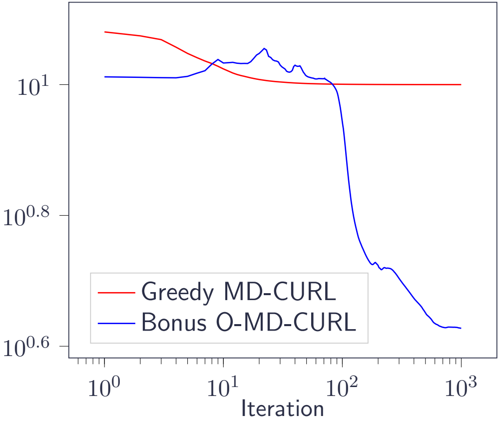
Greedy MD-CURL mean distributions over all steps $n \in [40]$

onus O-MD-CURL distribution at last step $n = 40$

Greedy MD-CURL distribution at last step $n = 40$



Average Regret



EnergyAlps Université Grenoble Alpes

Works of Bianca on the subject

Application to demand side management:

Bianca Marin Moreno et al. (2023). "(Online) Convex Optimization for Demand-Side Management: Application to Thermostatically Controlled Loads".

Learn fixed policy π with unknown fixed dynamics p, evolving adversarial losses F^t:

Bianca Marin Moreno, Khaled Eldowa, et al. (2025). "Online Episodic Convex Reinforcement Learning".

Learn time-varying policies π^t with unknown non-stationary dynamics p_t and evolving adversarial losses F^t

Bianca Marin Moreno, Margaux Brégère, et al. (2024). "MetaCURL: Non-stationary Concave Utility Reinforcement Learning".

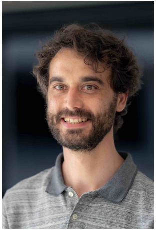
How to avoid episodic restarts?

Bianca Marin Moreno, Pierre Gaillard, et al. (2025). "Online Markov Decision Processes with Terminal Law Constraints".

Thank you for your attention!

Questions?

Bianca M. Moreno



Pierre Gaillard
Inria, LJK

Margaux Brégère
EDF R&D

Nadia Oudjane
EDF R&D

EnergyAlps
Université Grenoble Alpes

References I

- Beck, Amir and Marc Teboulle (2003). “Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization”.
- Marin Moreno, Bianca et al. (2023). “(Online) Convex Optimization for Demand-Side Management: Application to Thermostatically Controlled Loads”.
- Moreno, Bianca Marin, Margaux Brégère, et al. (2024). “MetaCURL: Non-stationary Concave Utility Reinforcement Learning”.
- Moreno, Bianca Marin, Khaled Eldowa, et al. (2025). “Online Episodic Convex Reinforcement Learning”.
- Moreno, Bianca Marin, Pierre Gaillard, et al. (2025). “Online Markov Decision Processes with Terminal Law Constraints”.
- Neu, Gergely, Andras Gyorgy, and Csaba Szepesvari (2012). “The adversarial stochastic shortest path problem with unknown transition probabilities”.
- Perrin, Sarah et al. (2020). *Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications*.