Sequential and reinforcement learning for demand side management

ESSEC - WORKSHOP: « FORECASTING & OPTIMIZATION: STREAMLINING SUPPLY CHAINS »

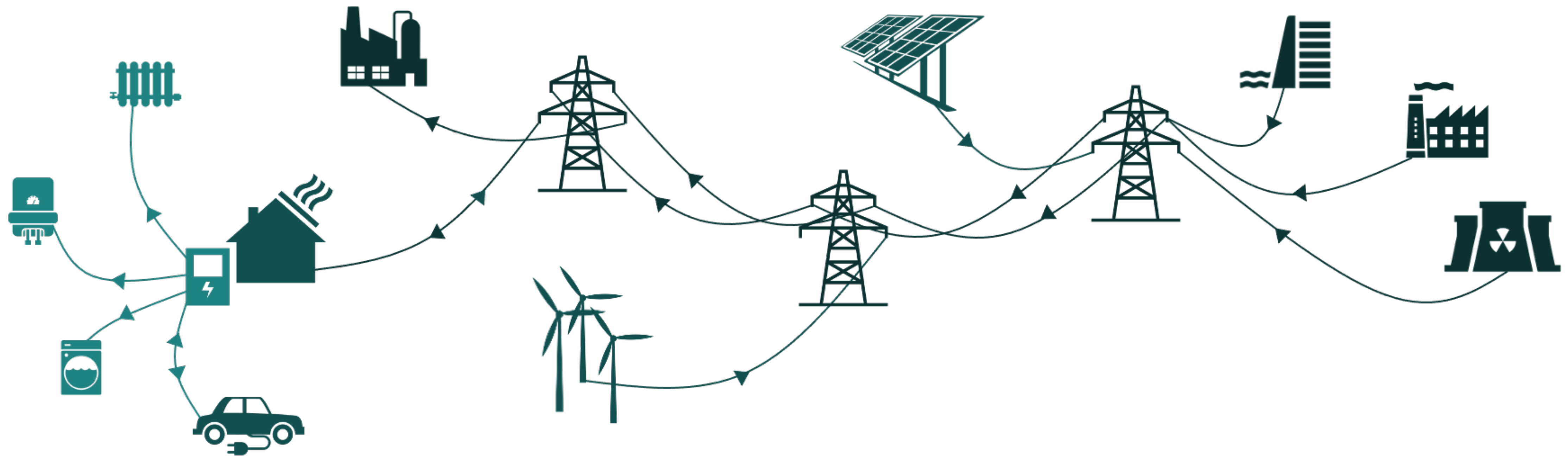
Margaux Brégère - December, 5th 2024

LPSM
LABORATOIRE DE PROBABILITÉS STATISTIQUE & MODÉLISATION

Introduction

Demand side management

Electricity is hard to store → production - demand balance must be strictly maintained



Current solution: forecast demand and adapt production accordingly

- Renewable energies development

→ production harder to adjust

- New (smart) meters → access to data and instantaneous communication

Prospective solutions: manage demand

→ Demand Response: Send incentive signals

→ Demand Despatch: Control flexible devices

Stochastic Bandit Algorithms for Demand Response

Gilles Stoltz

Yannig Goude

Pierre Gaillard

Demand side management with incentive signals

The environment (consumer behavior) is discovered through interactions (incentive signal choices) → Reinforcement learning

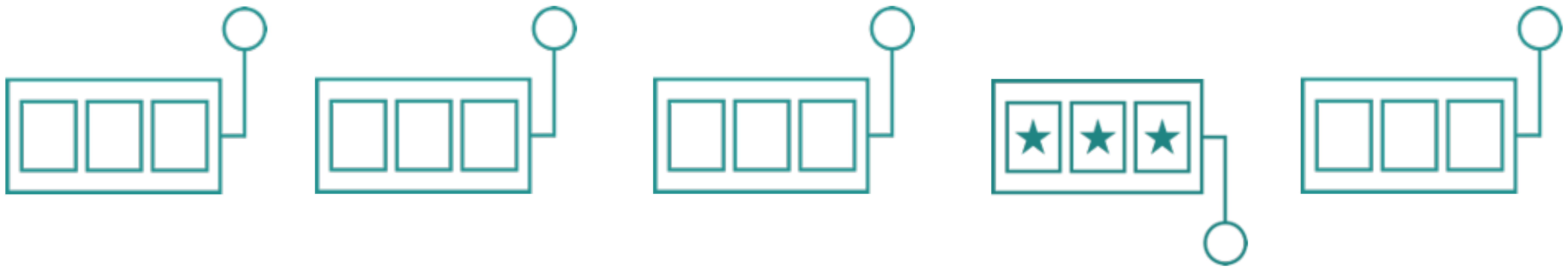
How to develop automatic solutions to chose incentive signals dynamically?

Exploration: Learn consumer behavior

Exploitation: Optimize signal sending

« Smart Meter Energy Consumption Data in London Households »

Stochastic multi-armed bandits



Stochastic multi-armed bandits

In a multi-armed bandit problem, a gambler facing a row of K slot machines (also called one-armed bandits) has to decide which machines to play to maximize her reward

Exploration:Test many arms

Exploitation: Play the « best » arm

Exploration - Exploitation trade-off

Stochastic multi-armed bandit

Each arm k is defined by an unknown probability distribution ν_k

For $t = 1, \ldots, T$

- Pick an arm $I_t \in \{1, \ldots, K\}$
- Receive a random reward Y_t with $Y_t \,|\, I_t = k \sim \nu_k$

Maximize the cumulative reward ⇔ Minimize the regret, i.e., the difference, in expectation, between the cumulative reward of the best strategy and that of ours:

$$R_T = T \max_{k=1,\ldots,K} \mu_k - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right], \text{ with } \mu_k = \mathbb{E}[\nu_k]$$

A good bandit algorithm has a sub-linear regret: $\frac{R_T}{T} \to 0$

Upper Confidence Bound algorithm[1]

Initialization: pick each arm once

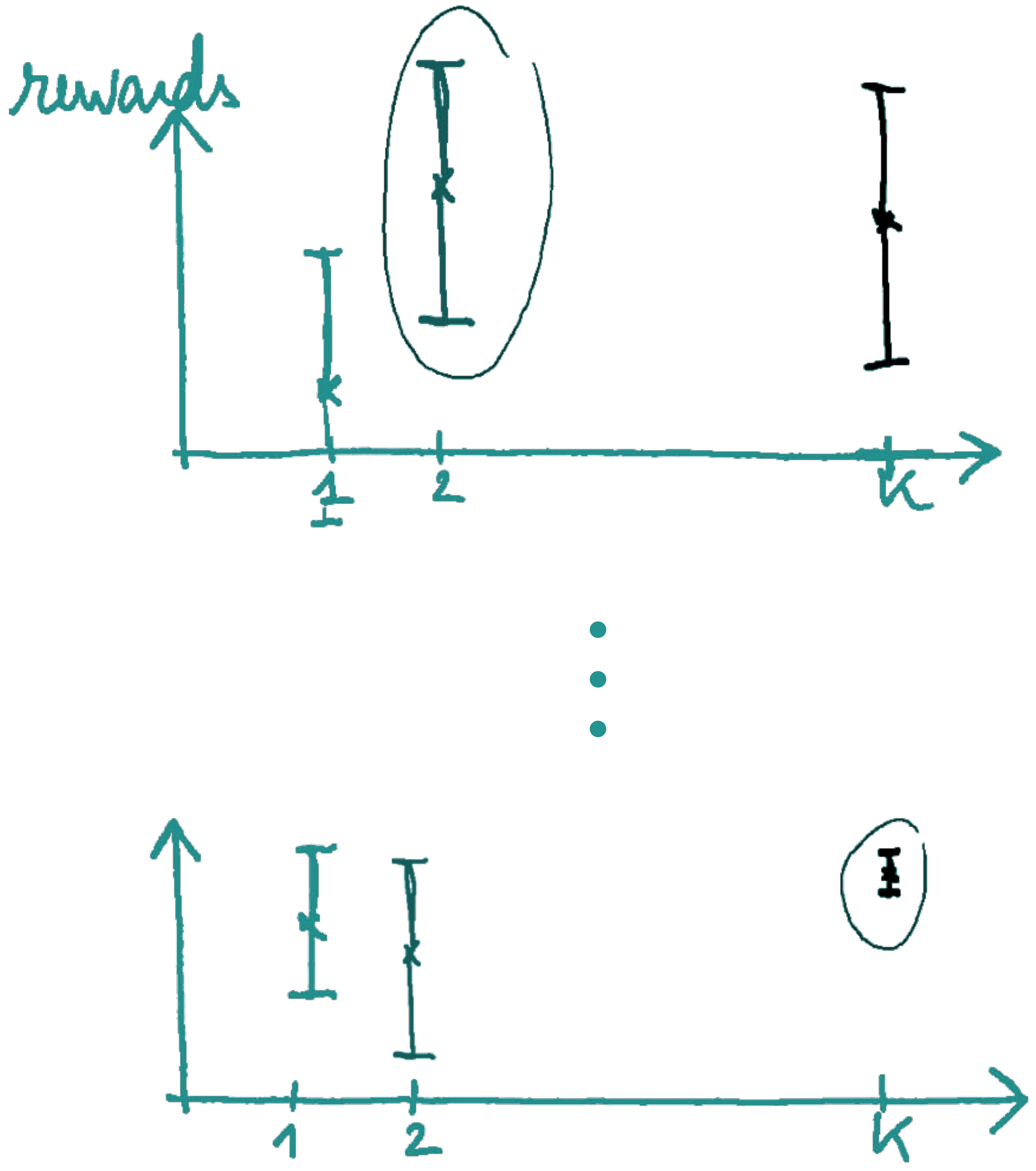
For $t = K + 1, \ldots, T$:

- Estimate the expected reward of each arm k with $\hat{\mu}_{k,t-1}$ (empirical mean of past rewards)

- Build some confidence intervals around these estimations: $\mu_k \in \left[\hat{\mu}_{k,t-1} - \alpha_{k,t}\, , \hat{\mu}_{k,t-1} + \alpha_{k,t}\right]$ with high probability

- Be optimistic and act as if the best possible probable reward was the true reward and choose the next arm accordingly

$$I_t \in \arg\max_k \left\{ \hat{\mu}_{k,t-1} + \alpha_{k,t} \right\}$$



[1] Finite-time analysis of the multi-armed bandit problem, Peter Auer, Nicolo Cesa-Bianchi, Paul Fischer, Machine learning, 2002

UCB regret bound

The empirical means based on past rewards are:

$$\hat{\mu}_{k,t-1} = \frac{1}{N_{k,t-1}} \sum_{s=1}^{t-1} Y_s \mathbf{1}_{\{I_s = k\}} \quad \text{with} \quad N_{k,t-1} = \sum_{s=1}^{t-1} \mathbf{1}_{\{I_s = k\}}$$

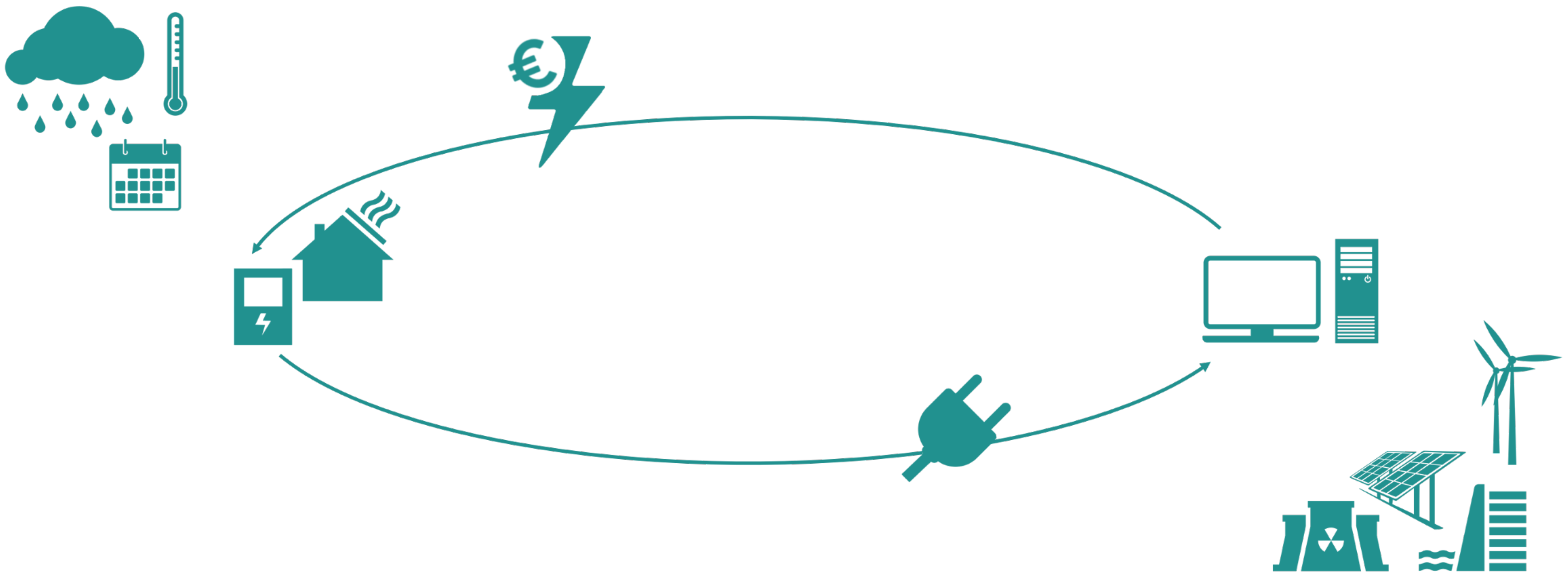
With Hoeffding-Azuma Inequality, we get

$$\mathbb{P}\left(\mu_k \in \left[\hat{\mu}_{k,t-1} - \alpha_{k,t}\,, \hat{\mu}_{k,t-1} + \alpha_{k,t}\right] \right) \geq 1 - t^{-3} \quad \text{with} \quad \alpha_{k,t} = \sqrt{\frac{2 \log t}{N_{k,t-1}}}$$

And be optimistic ensures that

$$R_T \lesssim \sqrt{TK \log T}$$

A Bandit Approach for Demand Response



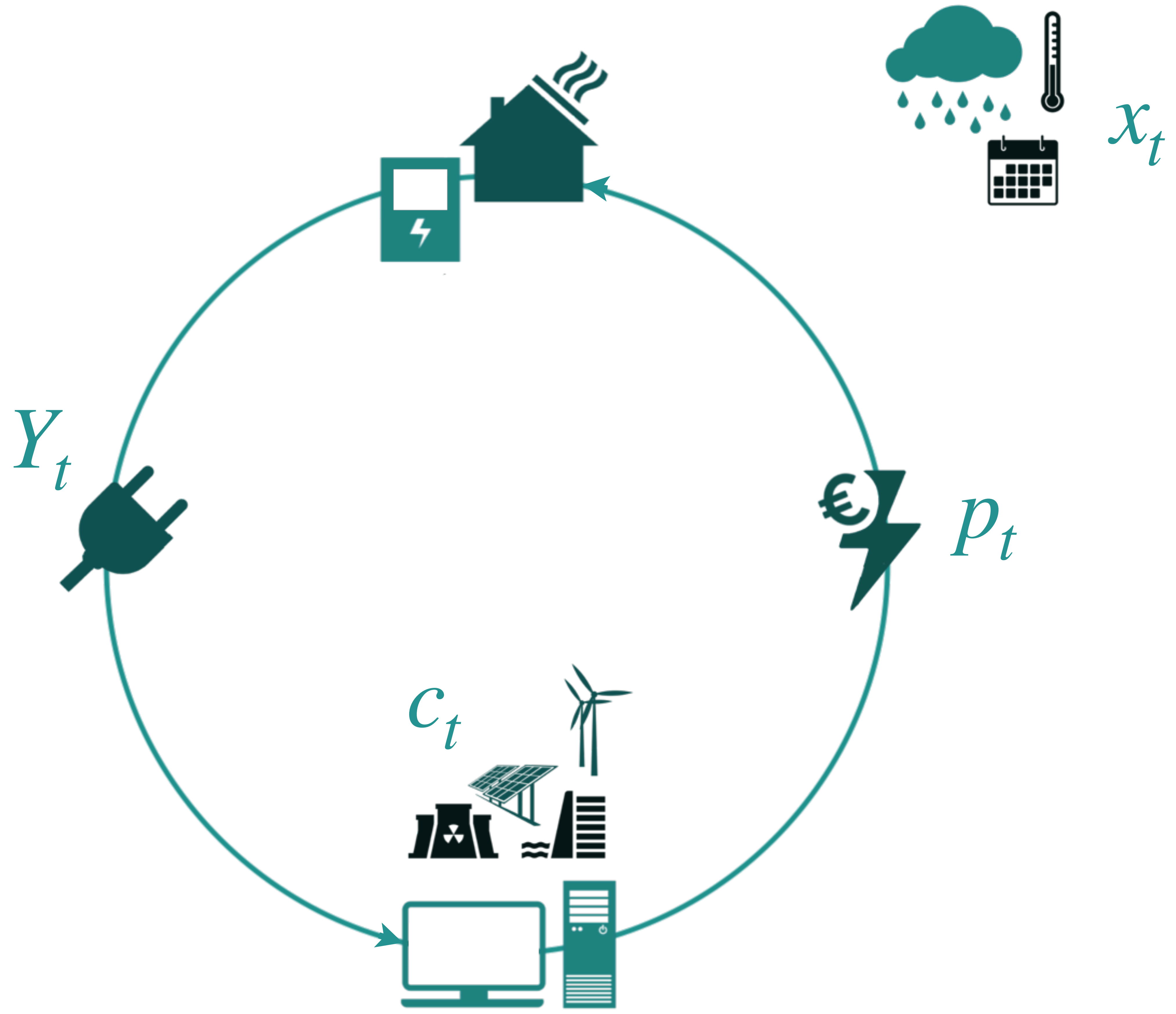
Demand side management with incentive signals

For $t = 1, \ldots, T$

- Observe a context x_t and a target c_t
- Choose price levels p_t
- Observe the resulting electricity demand

$$Y_t = f(x_t, p_t) + \mathrm{noise}(p_t)$$

and suffer the loss $\ell\left(Y_t, c_t\right)$



Assumptions:

- Homogenous population, K tariffs, $p_t \in \Delta_K$
- $f\left(x_t, p_t\right) = \phi\left(x_t, p_t\right)^{\mathrm{T}}\theta$ with ϕ a known mapping function and θ an unknown vector to estimate
- $\mathrm{noise}(p_t) = p_t^{\mathrm{T}}\varepsilon_t$ with $\mathbb{V}\left[\varepsilon_t\right] = \Sigma$
- $\ell\left(Y_t, c_t\right) = \left(Y_t - c_t\right)^2$

Bandit algorithm for target tracking

Under these assumptions: $\mathbb{E}\Big[\,(Y_t - c_t)^2 \,\Big|\, \text{past}, x_t, p_t\Big] = \big(\phi(x_t, p_t)^\mathrm{T}\theta - c_t\big)^2 + p_t^\mathrm{T}\Sigma p_t$

☞ Estimate parameters θ and Σ to estimate losses and reach a bias-variance trade-off

Optimistic algorithm:

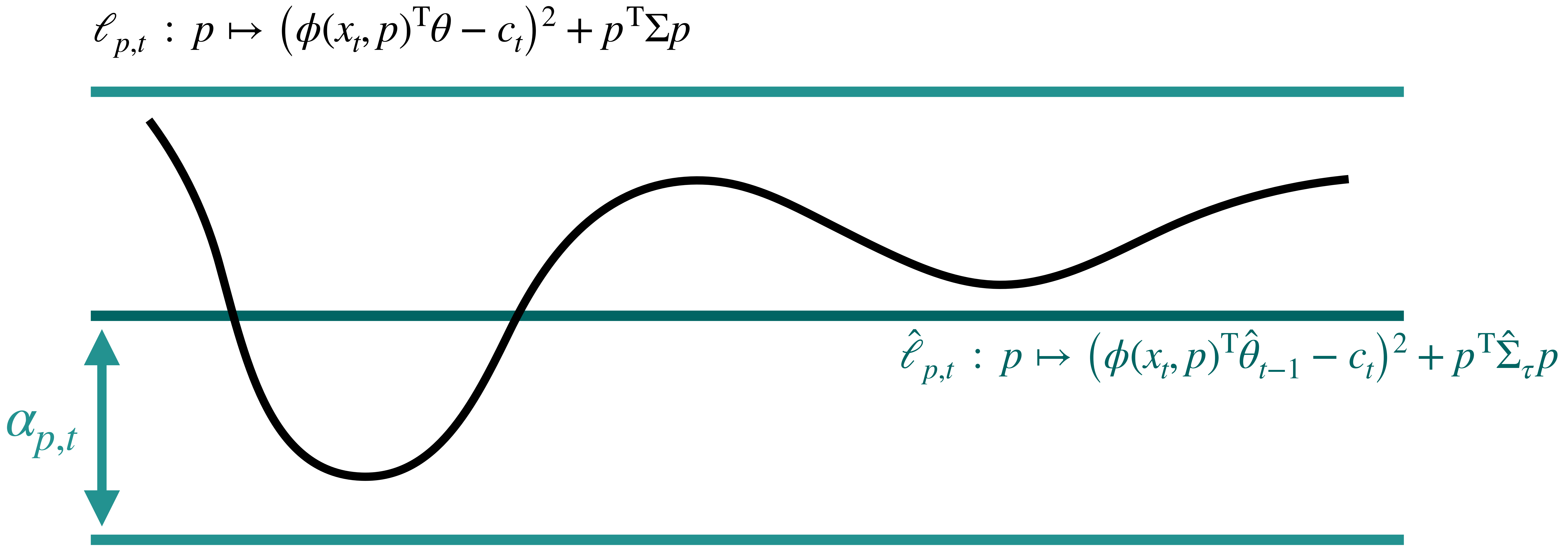
For $t = 1,\ldots,\tau$

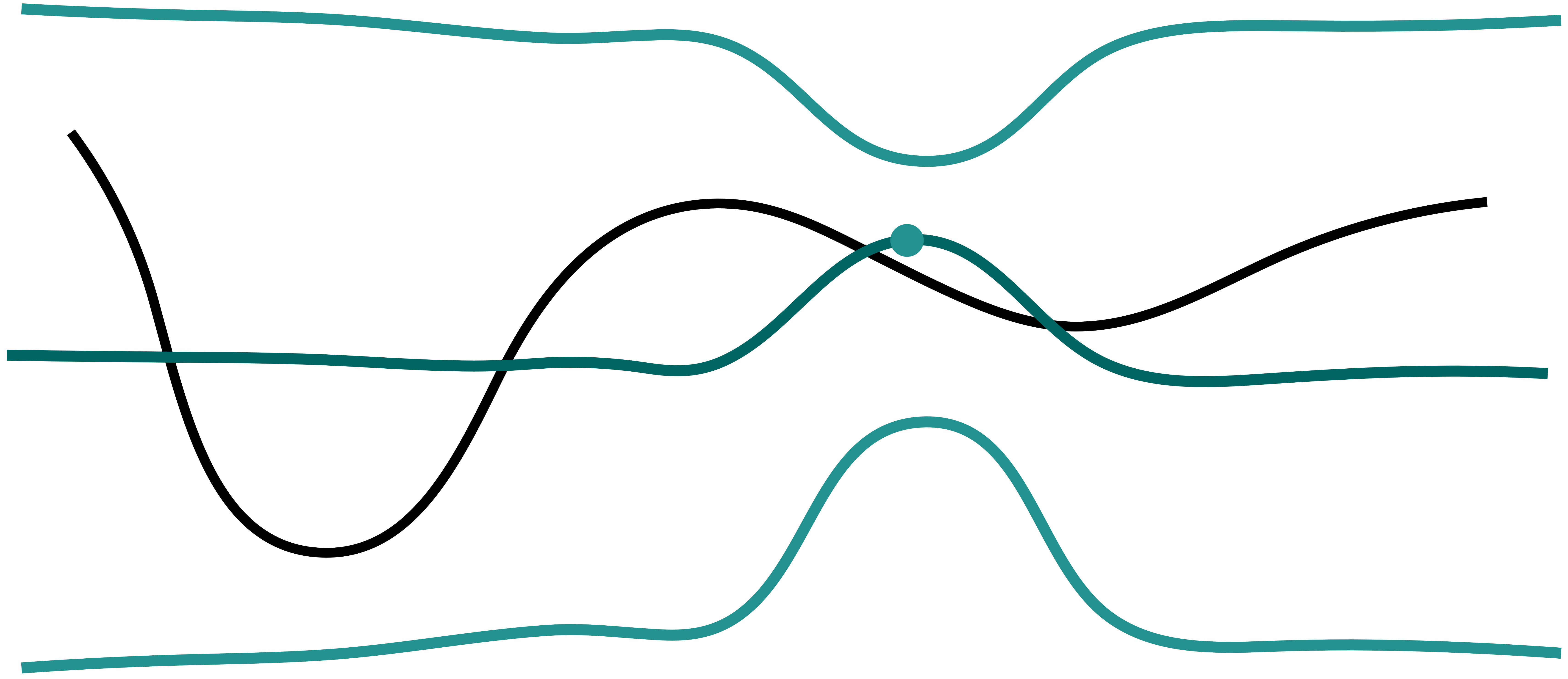
- Select price levels deterministically to estimate Σ offline with $\hat{\Sigma}_\tau$

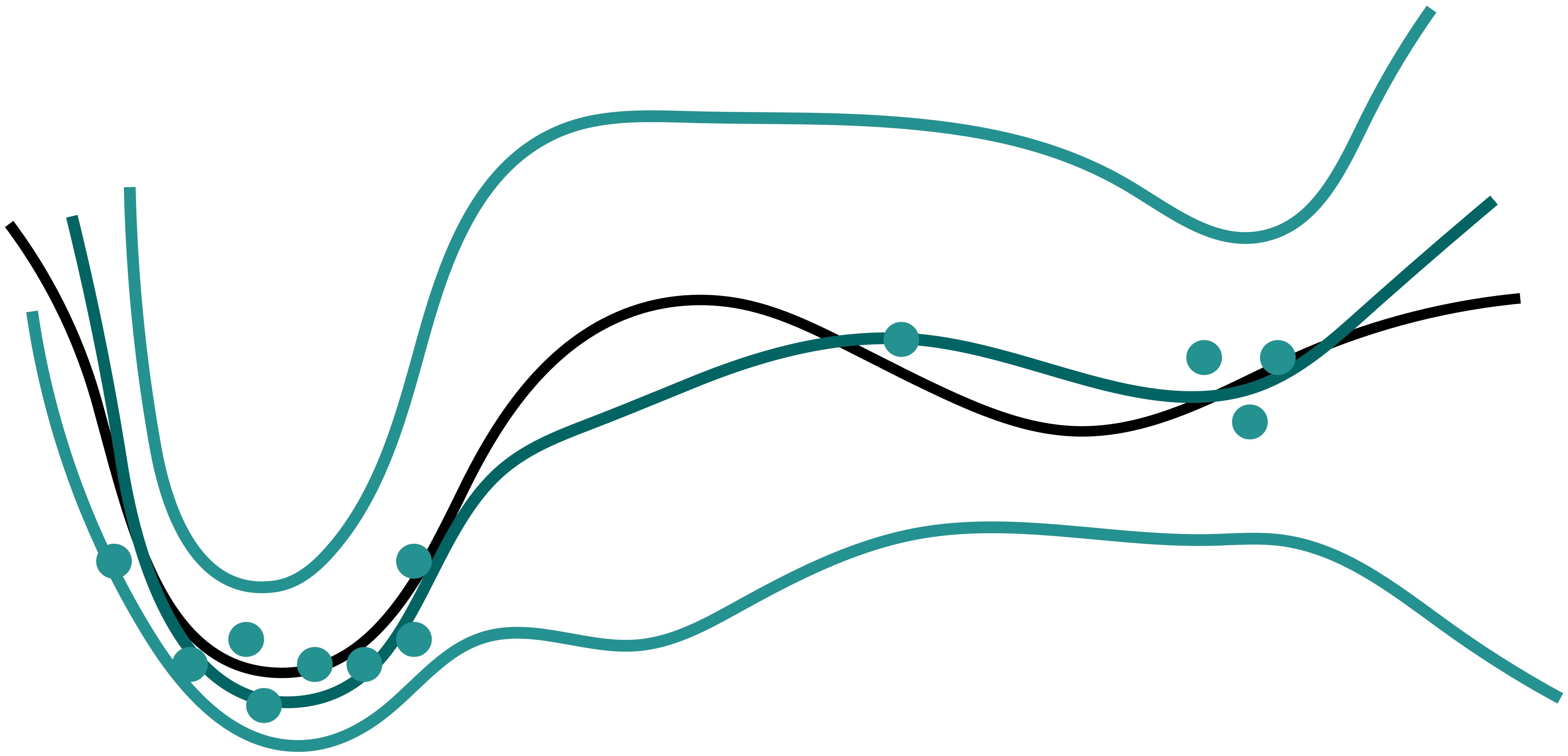
For $t = \tau + 1,\ldots,T$

- Estimate θ based on past observation with $\hat{\theta}_{t-1}$ thanks to a Ridge regression
- Estimate future expected loss for each price level p: $\hat{\ell}_{p,t} = \big(\phi(x_t, p)^\mathrm{T}\hat{\theta}_{t-1} - c_t\big)^2 + p^\mathrm{T}\hat{\Sigma}_\tau p$
- Get confidence bound on these estimations: $|\hat{\ell}_{p,t} - \ell_p| \le \alpha_{p,t}$
- Select price levels optimistically:

$$p_t \in \arg\min_p \big\{\hat{\ell}_{p,t} - \alpha_{p,t}\big\}$$


$\ell_{p,t} : p \mapsto \left(\phi(x_t, p)^{\mathrm{T}}\theta - c_t\right)^2 + p^{\mathrm{T}}\Sigma p$
$\hat{\ell}_{p,t} : p \mapsto \left(\phi(x_t, p)^{\mathrm{T}}\hat{\theta}_{t-1} - c_t\right)^2 + p^{\mathrm{T}}\hat{\Sigma}_\tau p$
$\alpha_{p,t}$





The problem is a bit more complex: curves vary with time t

Regret bound[3]

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(Y_t - c_t\right)^2 \; - \min_p \left(Y(p) - c_t\right)^2\right] = \sum_{t=1}^{T} \left(\phi(x_t, p_t)^{\mathrm{T}}\theta - c_t\right)^2 + p_t^{\mathrm{T}}\Sigma p_t - \sum_{t=1}^{T} \min_p \left(\phi(x_t, p)^{\mathrm{T}}\theta - c_t\right)^2 + p^{\mathrm{T}}\Sigma p$$

Theorem

For proper choices of confidence levels $\alpha_{p,t}$ and number of exploration rounds τ, with high probability $R_T \leq \mathcal{O}(T^{2/3})$

If Σ is known, $R_T \leq \mathcal{O}\left(\sqrt{T} \ln T\right)$

Elements of proof

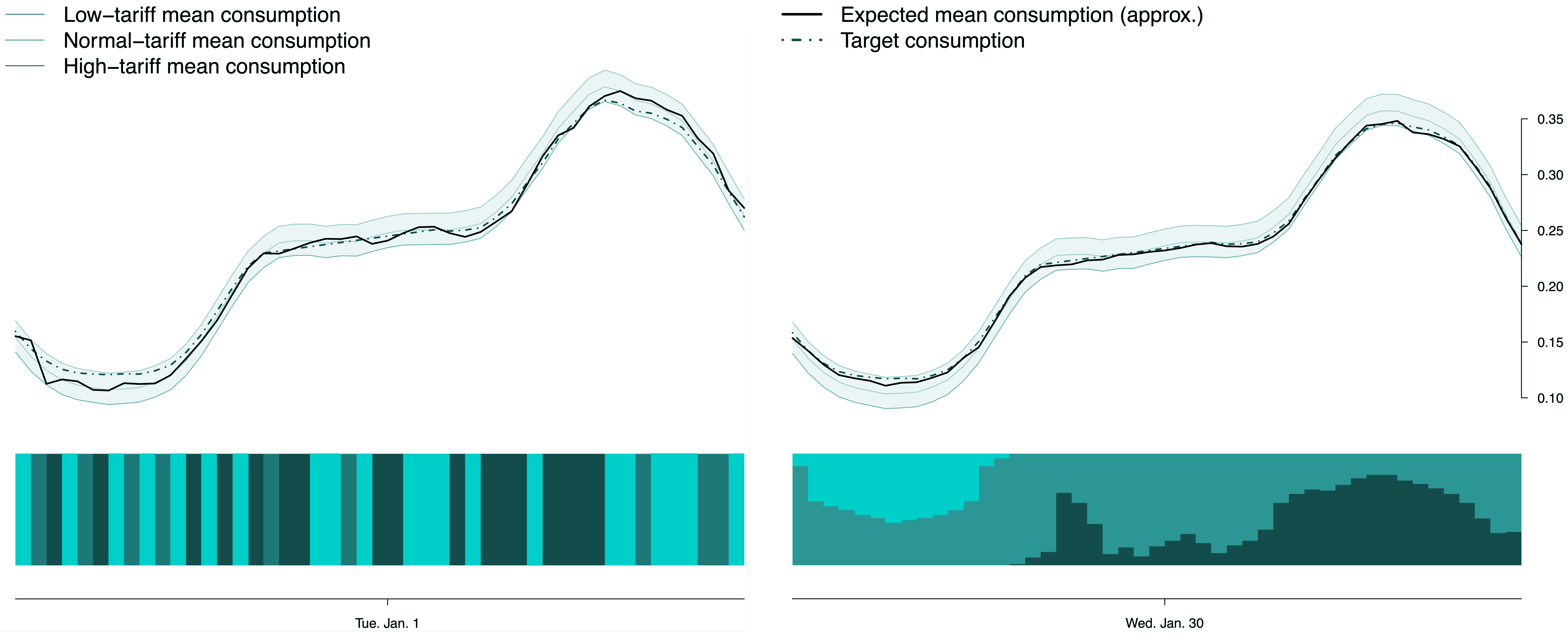
- Deviation inequalities on $\hat{\theta}_t$[4] and on $\hat{\Sigma}_\tau$
- Inspired from LinUCB regret bound analysis[5]

[3] Target Tracking for Contextual Bandits : Application to Demand Side Management, Margaux Brégère, Pierre Gaillard, Yannig Goude and Gilles Stoltz, ICML, 2019

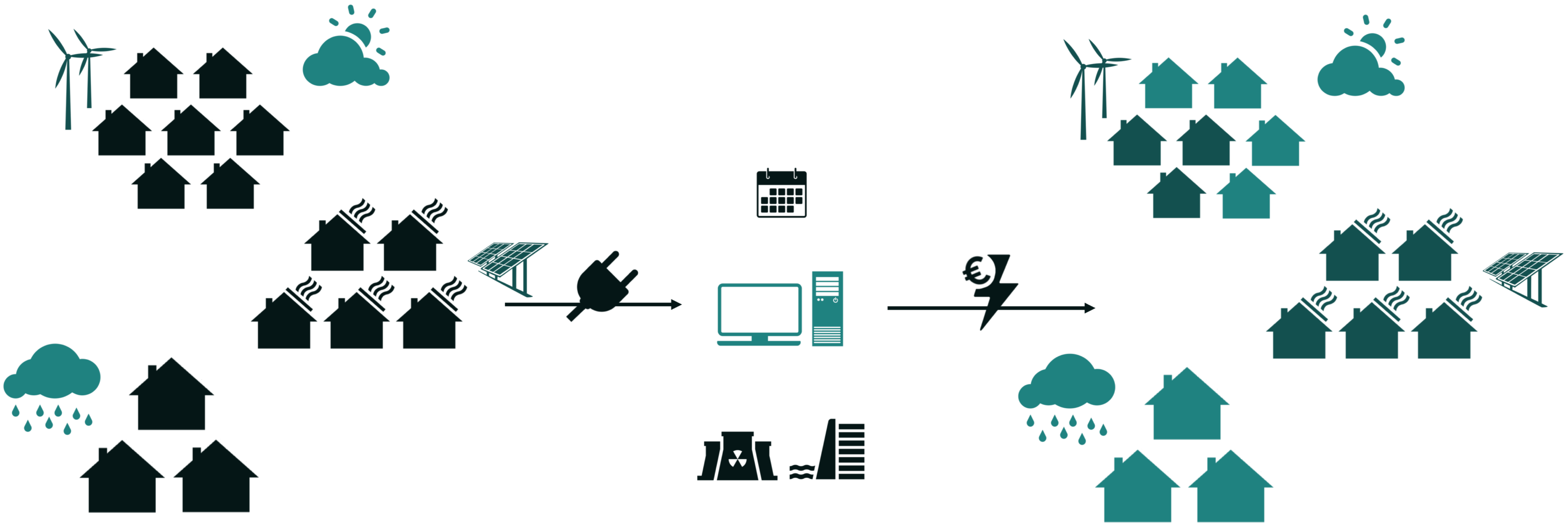
[4] Laplace's method on supermartingales: Improved algorithms for linear stochastic bandits, Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári, NeuRIPS, 2011

[5] Contextual bandits with linear payoff functions , Wei Chu, Li Lihong, Lev Reyzin, and Robert Schapire., JMLR 2011

Application



Extension: personalized demand side management



Online Optimization for Flexible Thermostatically Devices Control

Bianca Marin Moreno PhD

Mean Field Approach

Control of M water heaters with same characteristics without compromising service quality

For each water heater j, day t, time of the day n:

State: $x^n_{j,t} = \left(\mathrm{Temperature}^n_{j,t}, \mathrm{ON/OFF}^n_{j,t}\right)$

Action: $a^n_{j,t} = \left(\mathrm{Turn/Keep}^n\ \mathrm{ON/OFF}^n_{j,t}\right)$

New state depends on:

Temperature evolution (deterministic PDE)

+ Eventuel water drains (probabilistic law)

+ Action to turn/keep ON/OFF (service quality)

Markov Decision Process (MDP) p

Mean Field assumption ($M \to \infty$): Control the state-action distribution $\mu^{\pi,p}$ induces by a policy π in p

[5] (Online) Convex Optimization for Demand-Side Management: Application to Thermostatically Controlled Loads, Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard and Nadia Oudjane, 2024

Control with Mean Field Approach[5]

At each day $t = 1,\ldots,T$

For each water-heater $j = 1,\ldots,M$

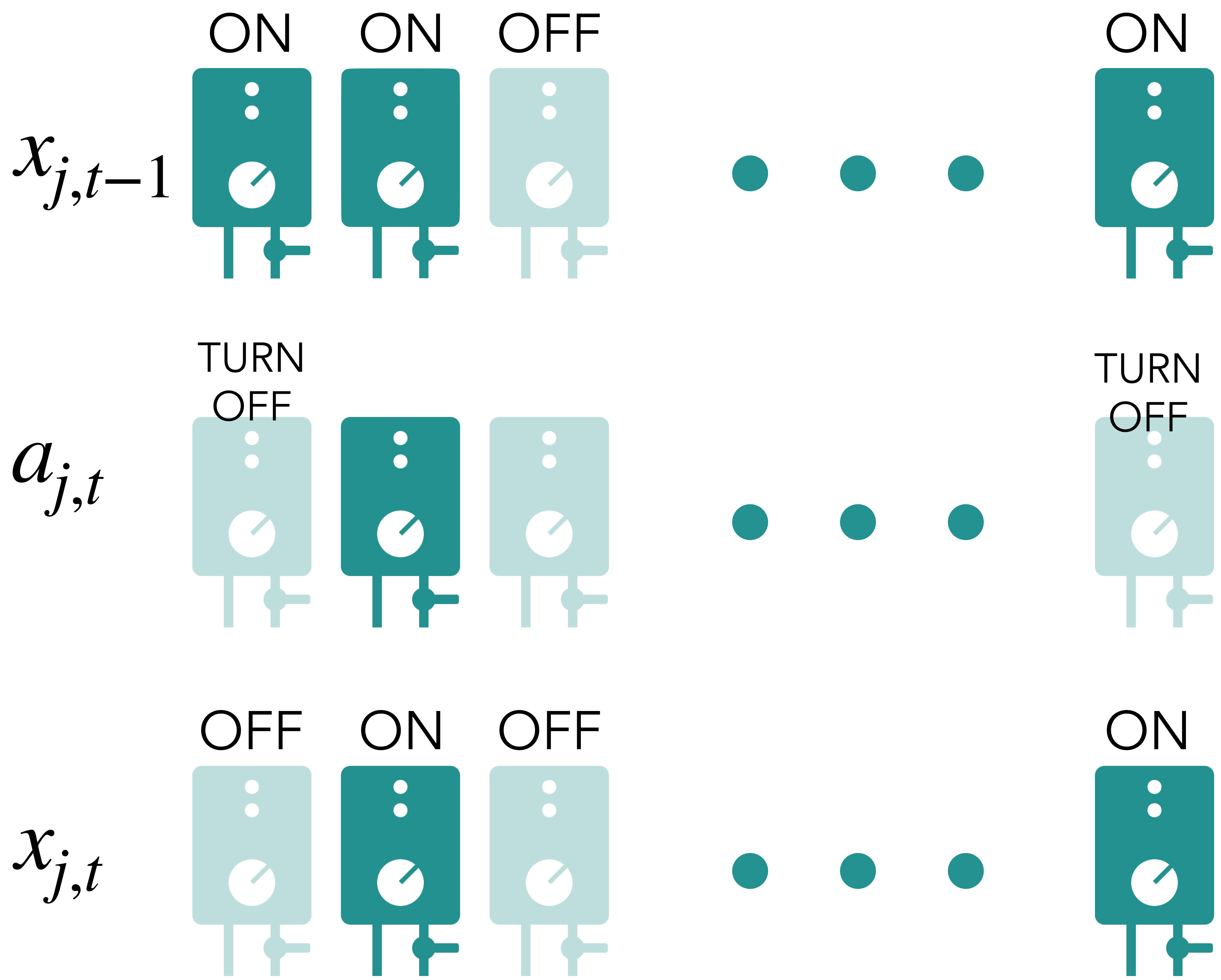
Initialization: $(x^0_{j,t}, a^0_{j,t}) \sim \mu_0$

For each instant of the day $n = 1,\ldots,N$

Send to all water heaters action $a^n_{j,t} \sim \pi^n_t(\cdot \mid x^n_{j,t})$

Loss function $F_t(\mu^{\pi_t,p})$ is exposed

Compute $\pi_{t+1} = (\pi^1_{t+1}, \ldots \pi^N_{t+1})$



Aim: Find $\pi^\star \in \operatorname{argmin}_\pi \sum_{t=1}^{T} F_t\left(\mu^{\pi,p}\right)$

with F_t the quadratic difference between the consumption for all water-heaters and the target at t

[5] (Online) Convex Optimization for Demand-Side Management: Application to Thermostatically Controlled Loads, Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard and Nadia Oudjane, 2024

CURL in online learning scenario[6]

Mirror-Descent approach for CURL (convex reinforcement learning) when p and $F_t = F$ are known: $\pi^{\text{MD}}(F, p)$

At each day $t = 1, \dots, T$

For each water-heater $j = 1, \dots, M$

Initialization: $(x^0_{j,t}, a^0_{j,t},) \sim \mu_0$

For each instant of the day $n = 1, \dots, N$

Send to all water heaters action $a^n_{j,t} \sim \pi^n_t(\cdot \,|\, x^n_{j,t})$

Loss function $F_t(\mu^{\pi_t, p})$ is exposed

Update the estimation of the MDP using the new observations: $\hat{p}_{t+1} = \dfrac{1}{M(t+1)} \sum_{j=1}^{M} \delta_j + \dfrac{t}{t+1} \hat{p}_t$

Act if $F_{t+1} = F_t$ and compute $\pi_{t+1} = \pi^{\text{MD}}(F_t, \hat{p}_{t+1})$

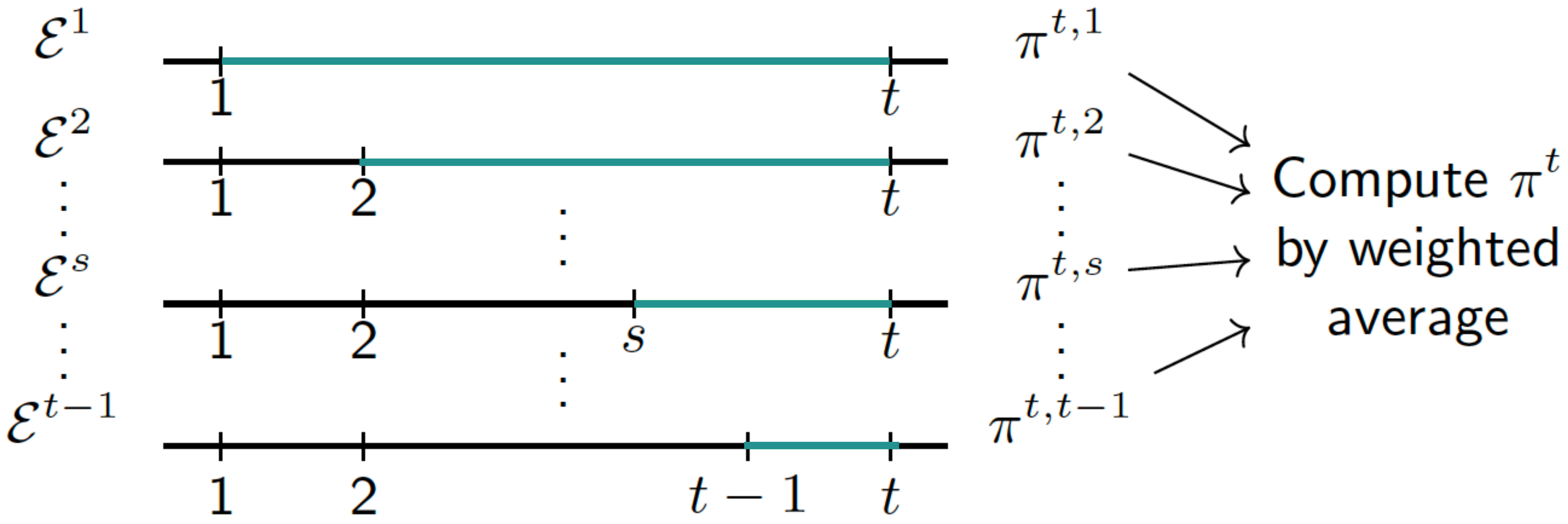
[6] Efficient Model-Based Concave Utility Reinforcement Learning through Greedy Mirror Descent, Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard and Nadia Oudjane, AISTAT 2024

Extension: non-stationary MDP[7]

At each day $t = 1,\ldots,T$

Restart previous algorithm $\mathscr{E}$ from the beginning: $\mathscr{E}^t$

Define a new policy by averaging the t policies $\pi_t = \sum_{s=1}^{t} \omega_{s,t}\pi_t^s$



[7] MetaCURL: Non-stationary Concave Utility Reinforcement Learning, Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard and Nadia Oudjane, NeurIPS, 2024

That's all folks!