Sequential and reinforcement learning for demand side management

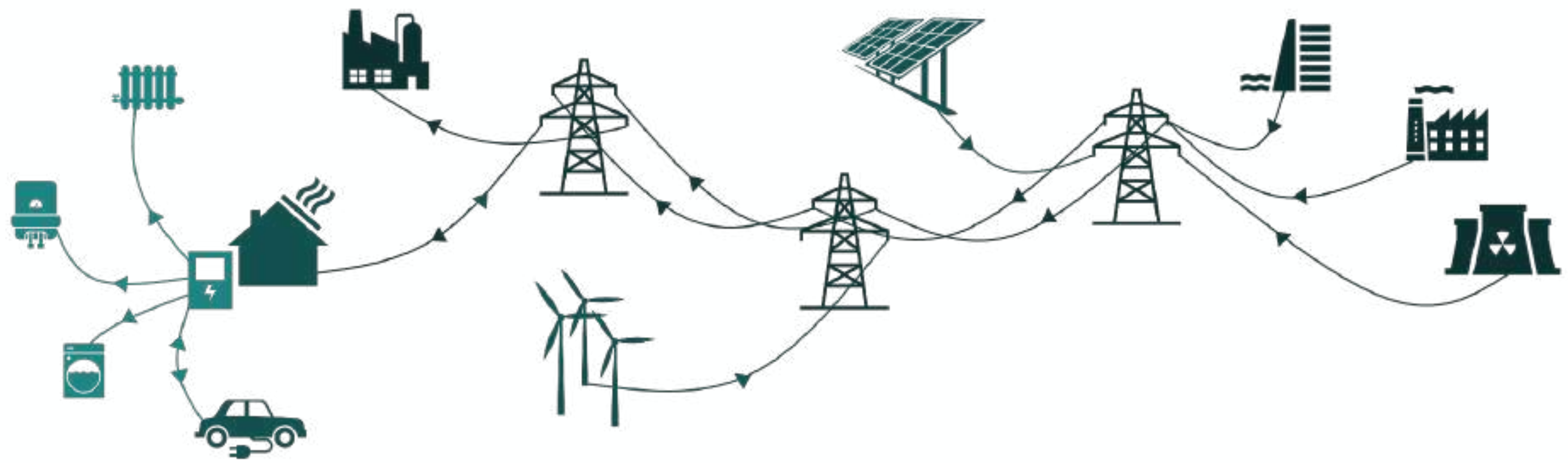
Paris Women in Machine Learning & Data Science @Criteo

Margaux Brégère - April, 24th 2024

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

→ Send incentive signals (prices)

→ Control flexible devices

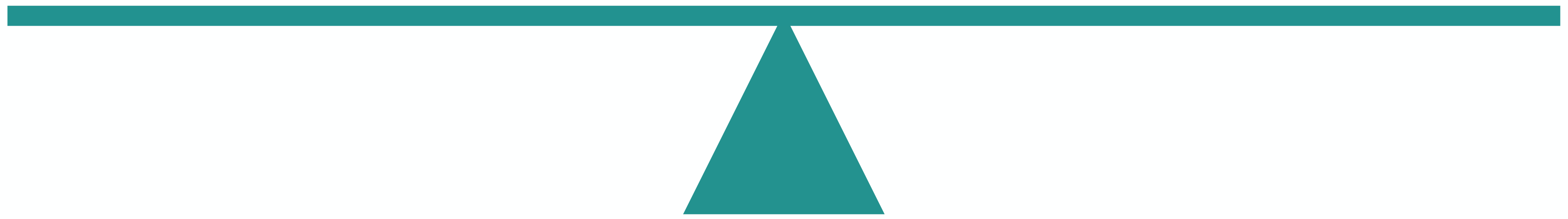
Demand side management with incentive signals

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

How to develop automatic solutions to chose incentive signals dynamically?

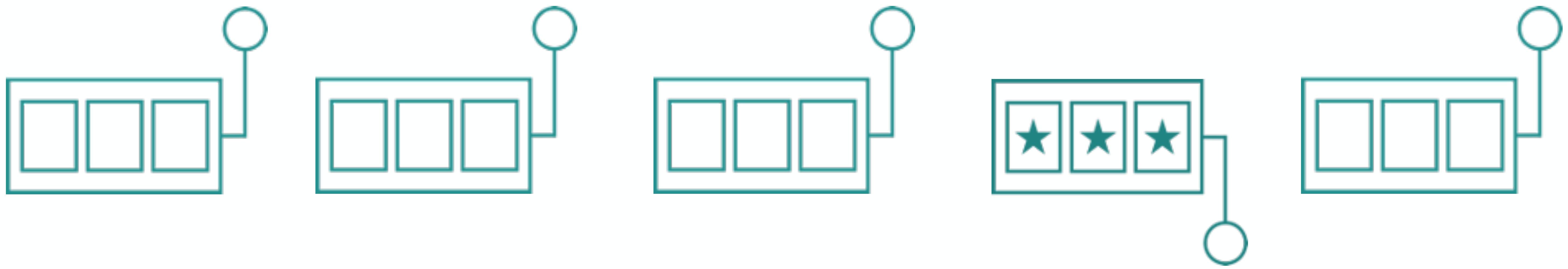
Exploration: Learn consumer behavior

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 $Y_t \mid I_t = k \sim \nu_k$

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: $\dfrac{R_T}{T} \to 0$

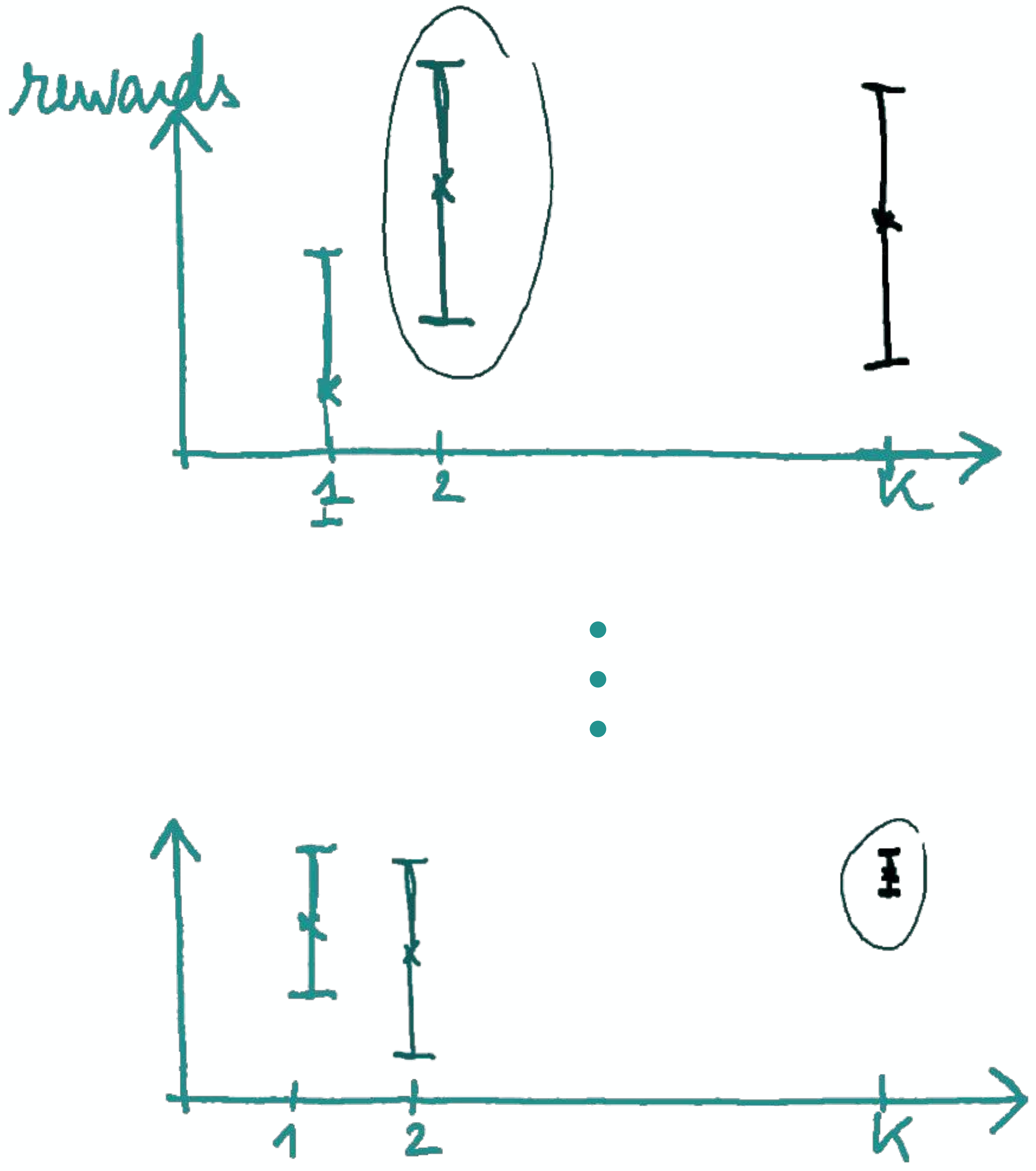
Upper Confidence Bound algorithm[1]

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

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



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

UCB regret bound

The empirical means based on past rewards are:

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

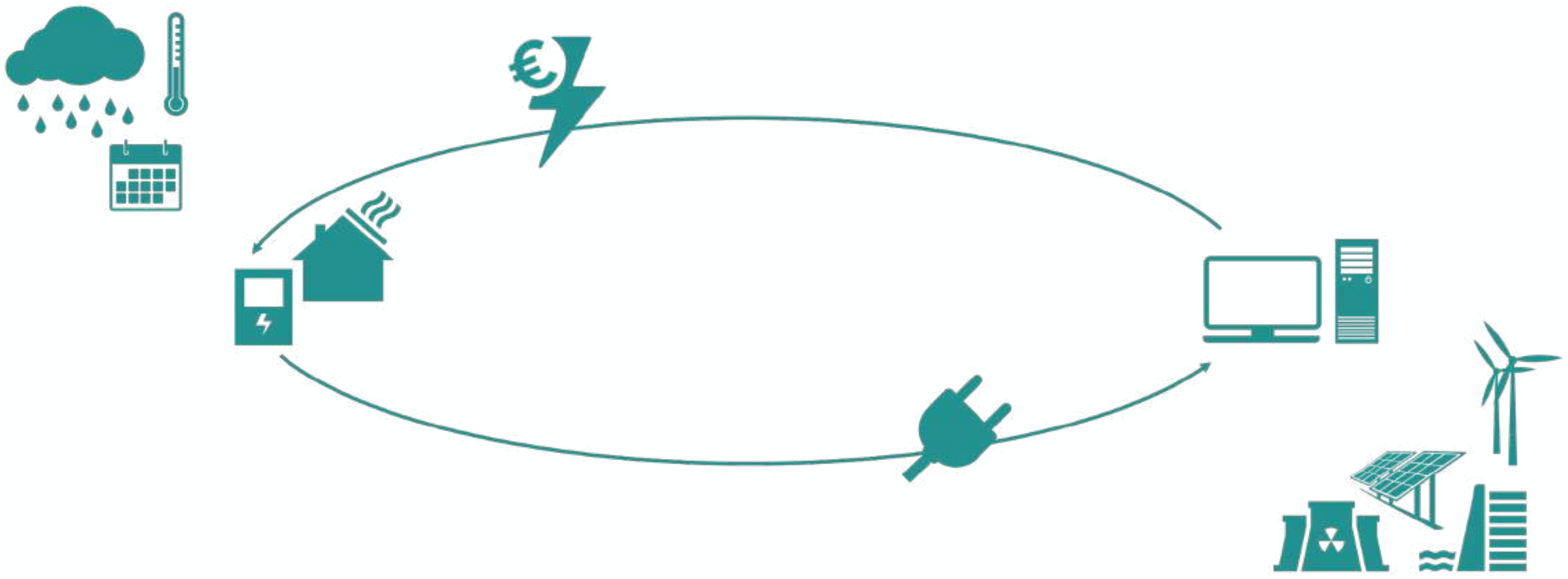
With Hoeffding-Azuma Inequality, we get

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

And be optimistic ensures that

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

Modeling demand side management



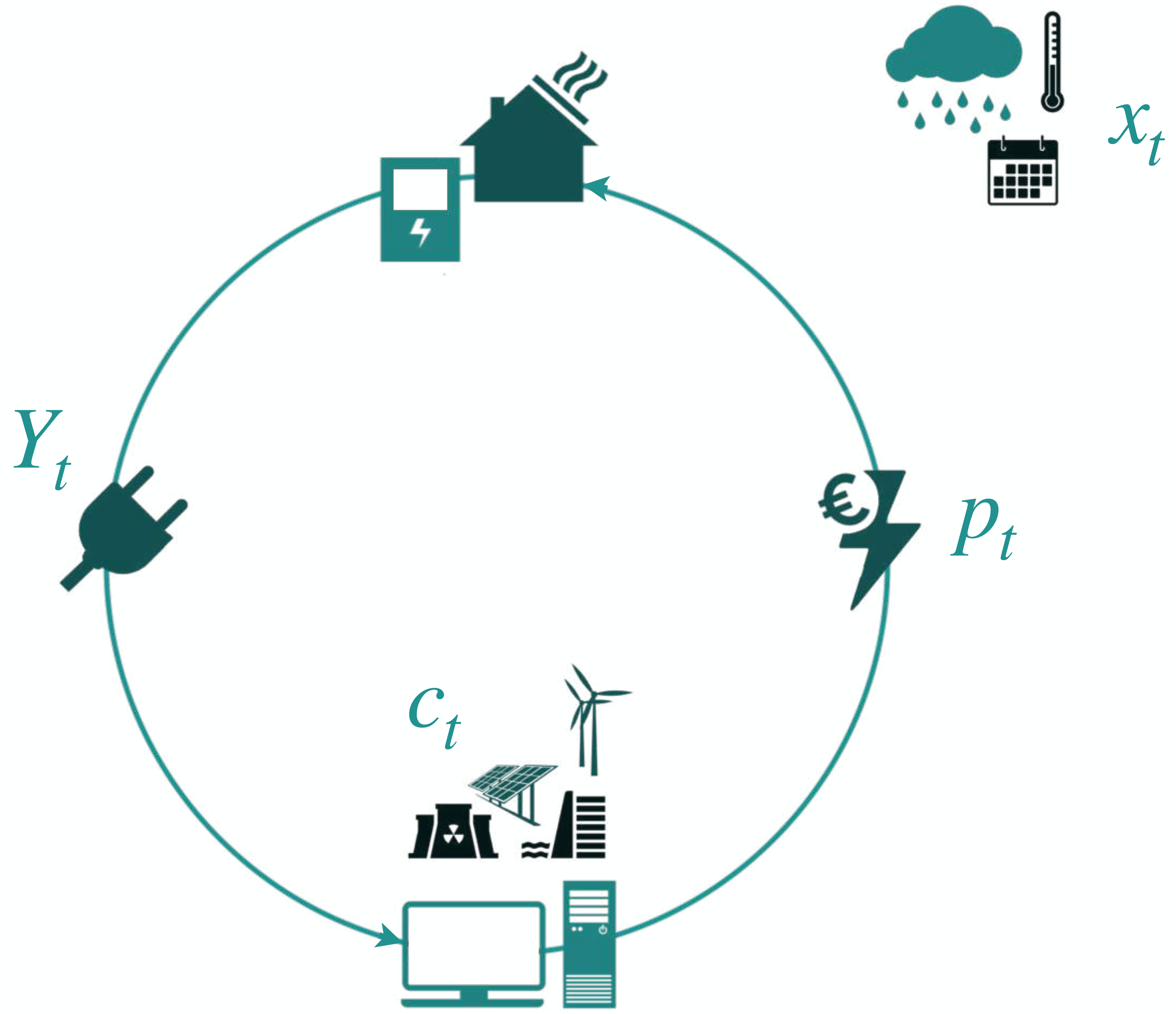
Demand side management with incentive signals

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

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

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
- $\text{noise}(p_t) = p_t^{\mathrm{T}}\varepsilon_t$ 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: $\mathbb{E}\left[\left(Y_t - c_t\right)^2 \,\middle|\, \text{past}, x_t, p_t\right] = \left(\phi(x_t, p_t)^\mathrm{T}\theta - c_t\right)^2 + p_t^\mathrm{T}\Sigma p_t$

☞ 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: $\hat{\ell}_{p,t} = \left(\phi(x_t, p)^\mathrm{T}\hat{\theta}_{t-1} - c_t\right)^2 + p^\mathrm{T}\hat{\Sigma}_\tau p$
- Get confidence bound on these estimations: $|\hat{\ell}_{p,t} - \ell_p| \leq \alpha_{p,t}$
- Select price levels optimistically:

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

Regret bound[2]

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

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

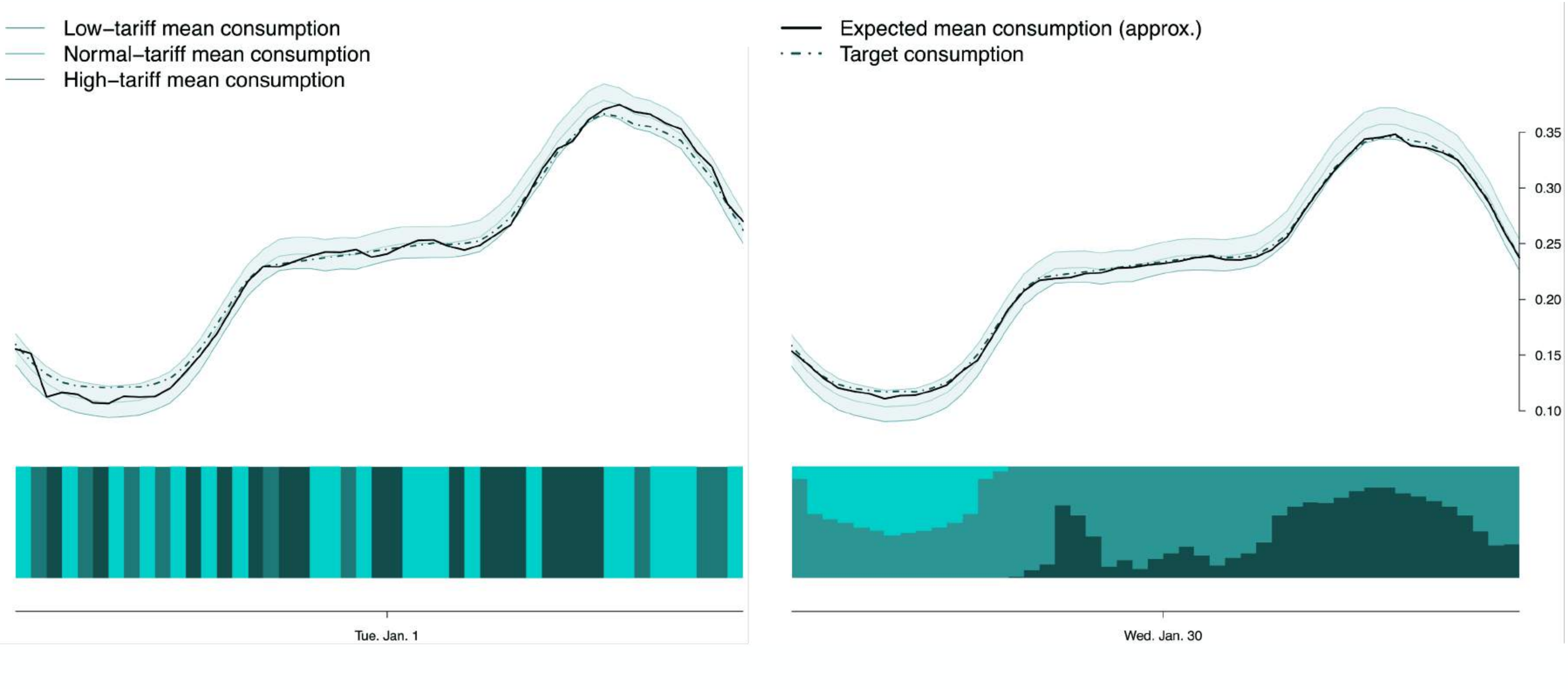
- Deviation inequalities on $\hat{\theta}_t$[3] and on $\hat{\Sigma}_\tau$
- Inspired from LinUCB regret bound analysis[4]

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

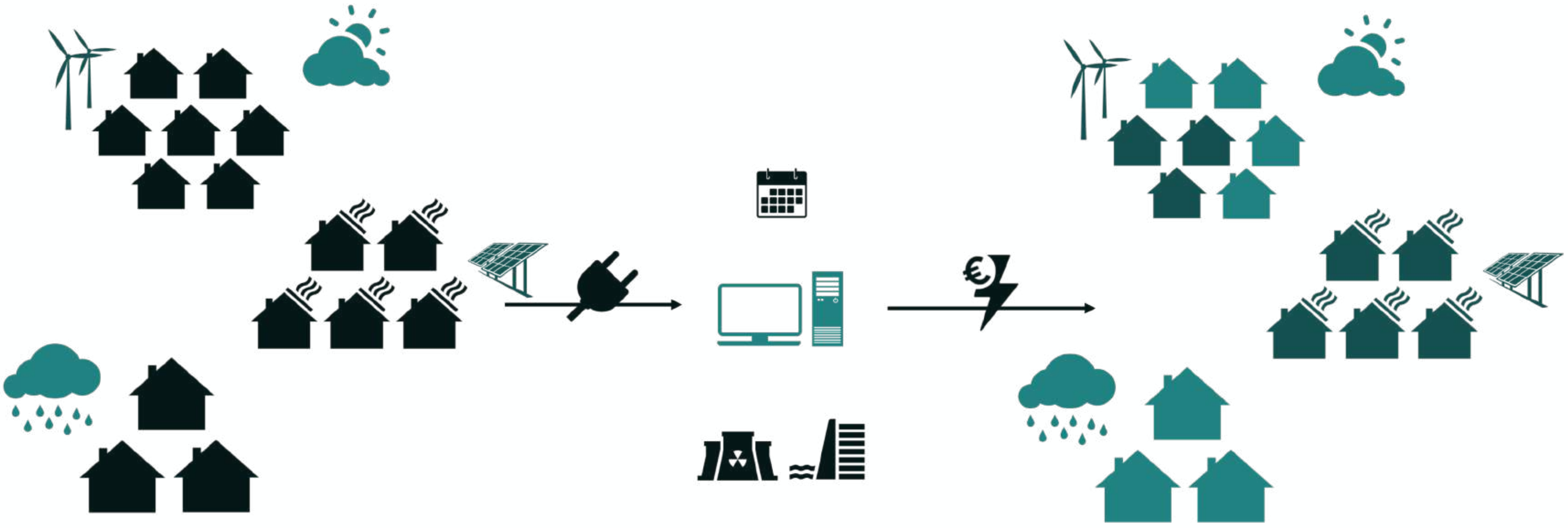
[3] Laplace's method on supermartingales: Improved algorithms for linear stochastic bandits, Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári, NeuRIPS, 2011

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

Application



Extension: personalized demand side management



Others approaches and prospects

Control of flexible devices[4]

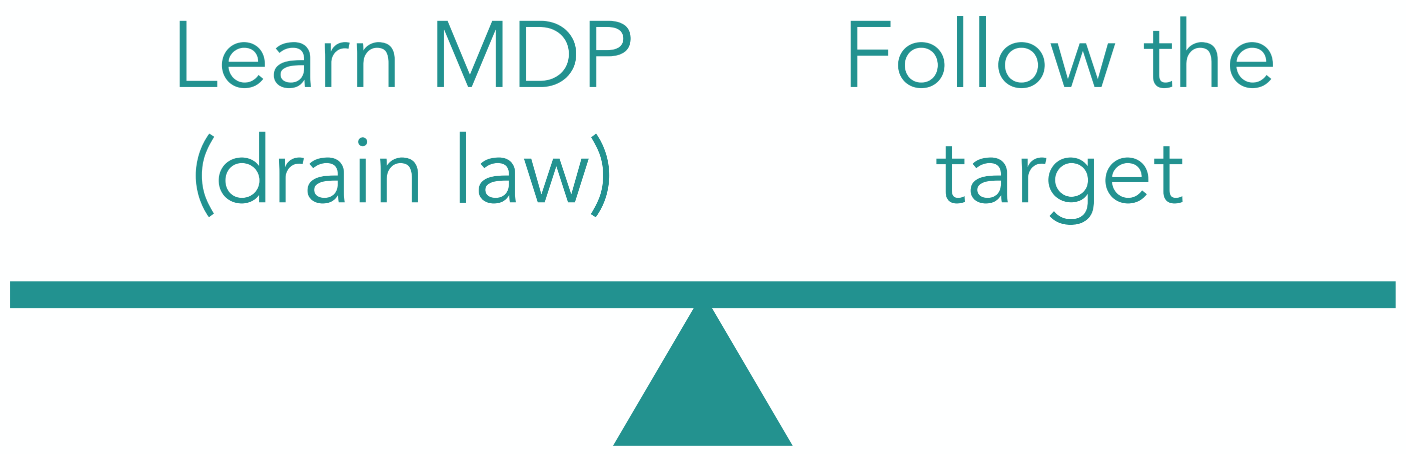
At each round $t = 1,\ldots,T$

- Observe a target c_t
- Send to all thermostatically controlled loads a probability of switching on $p_t \in [0,1]$
- Observe the demand

N water-heaters to be controlled without compromising service quality

Assumptions:

- N water-heaters with same characteristics
- Demand of water-heater i is zero if OFF and constant if ON
- State $x_{i,t} = \left(\text{Temperature}_t, \text{ON/OFF}_t\right)$ of water-heater i follows an unknown Markov Decision Process (MDP)
- It is possible to control demand if the MDP is known



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

Hyper-parameter optimization[5]

Train a neural network is expensive and time-consuming

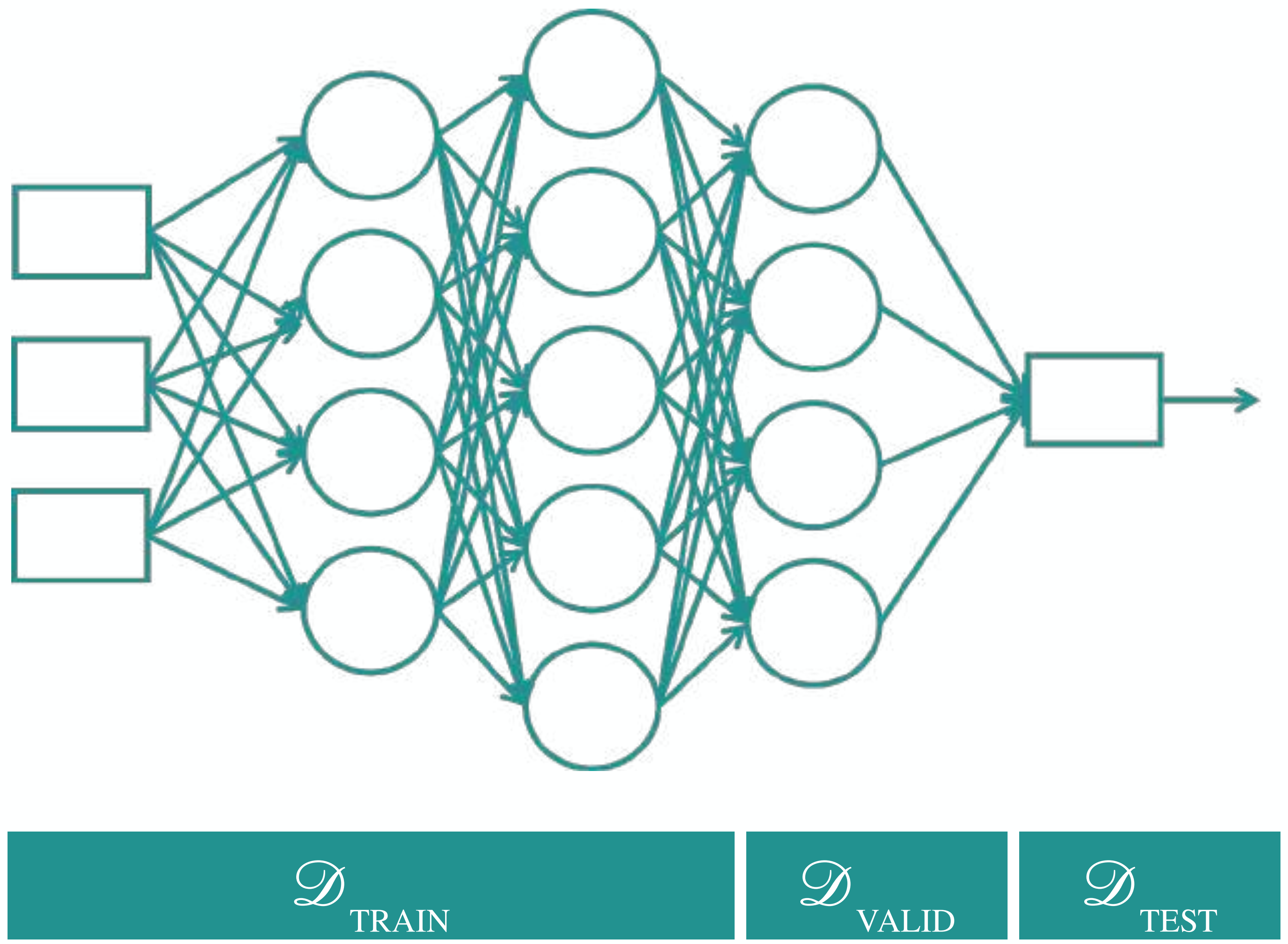
Aim: for a set of hyper-parameters Λ (number of neurons, activation functions etc.) and a budget T, find the best neural network:

$$\arg\min_{\lambda \in \Lambda} \ell\left(f_\lambda(\mathscr{D}_{\text{TEST}})\right)$$

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

- Choose hyper-parameters $\lambda_t \in \Lambda$
- Train network f_{λ_t} on $\mathscr{D}_{\text{TRAIN}}$
- Observe the forecast error $\ell_t = \ell\left(f_{\lambda_t}(\mathscr{D}_{\text{VALID}})\right)$

Output (best arm identification): $\arg\min_{f_{\lambda_t}} \ell\left(f_{\lambda_t}(\mathscr{D}_{\text{VALID}})\right)$



[5] A bandit approach with evolutionary operators for model selection : Application to neural architecture optimization for image classification, Margaux Brégère and Julie Keisler, 2024

That's all folks!

Questions