# Target Tracking for Contextual Bandits:
## Application to Demand Side Management

Margaux Brégère

Joint work with Gilles Stoltz (Univ. Paris-Sud), Yannig Goude (EDF R&D) and Pierre Gaillard (Inria)

April, 5. 2019

8ème Rencontres Jeunes Statisticiens

# Introduction

Electricity is hard to store

- Maintain balance between production and demand at any time

**Current solution:** Forecast consumption and adapt production accordingly

- Renewable energies are subject to climate, making production hard to adjust
- New communication tools (smart meters) lead to data access and instantaneous communication

**Future solution:** Send incentive signals (electricity tariff variations) to manage demand response

How to optimize these signals learning from clients behaviors?

# Introduction

Learn from clients behaviors & Optimize tariffs sending
Exploration - Exploitation
trade-off

- Apply **contextual-bandit** theory to demand side management by offering price incentives

# Bandit Models


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.

# Bandit Models


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.

# Bandit Models

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.

# Stochastic Multi-Armed-Bandit Problem

Each arm (slot machine) $k$ has an unknown mean reward $\mu_k$
The mean reward of the best one is noted $\mu_{k^*}$

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

- **Picks** a machine $I_t \in \{1, \ldots, K\}$
- **Receives** a reward $g_{t,I_t}$, with $\mathbb{E}[g_{t,I_t} \mid I_t] = \mu_{I_t}$

Maximizing the expected cumulative reward = Minimizing pseudo-regret

Mean reward if the best machine is known

$$\mathbf{R_T = T\,\mu_{k^*} - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right]}$$

Mean reward of the strategy

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

# Bandit algorithm: Upper-Confidence-Bound (UCB) (Lai et al. 1985)

Upper-Confidence-Bound strategy: explore and exploit sequentially all along the experiment

- **Build** a confidence interval on the mean $\mu_k$ based on past observations

  Empirical reward: $\hat{\mu}_{k,t-1} = \frac{1}{N_{k,t-1}} \sum_{s=1}^{t-1} g_s \, 1_{\{I_s=k\}}$ with $N_{k,t-1} = \sum_{s=1}^{t-1} 1_{\{I_s=k\}}$

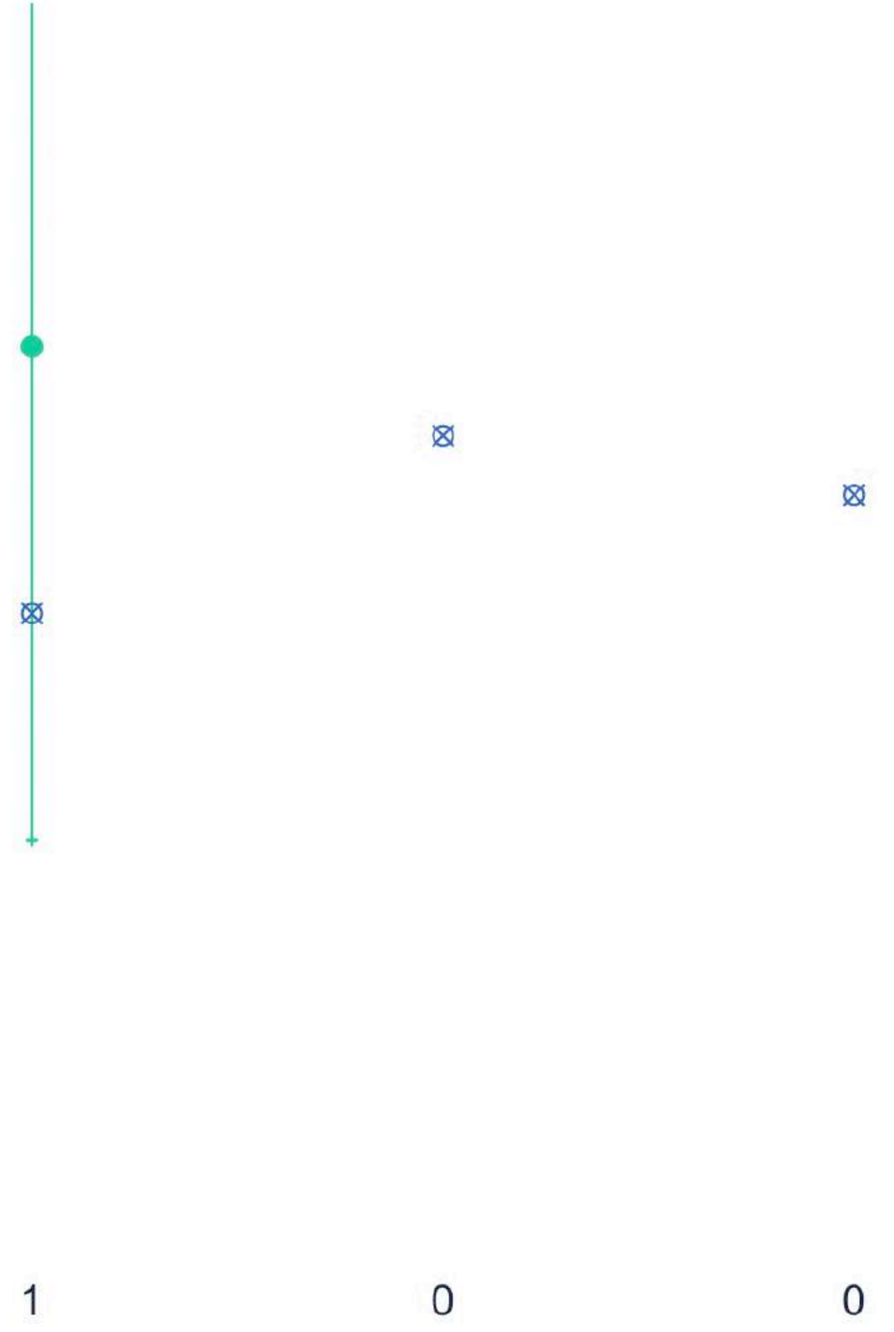
  With probability at least $1 - t^{-3}$ (Hoeffding-Azuma Inequality)

  $$\mu_k \in \left[ \hat{\mu}_k - \sqrt{\frac{2 \log t}{N_{k,t-1}}} , \hat{\mu}_k + \sqrt{\frac{2 \log t}{N_{k,t-1}}} \right]$$

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

$$I_t = \arg\max_{k \in \{1,\dots,K\}} \hat{\mu}_{k,t-1} + \sqrt{\frac{2 \log t}{N_{k,t-1}}}$$ which ensures $R_T \lesssim \sqrt{T K \log T}$

T = 1



# Stochastic Linear Bandits

There is a unknown parameter vector $\theta \in \mathbb{R}^K$
The reward is linear in the "arm vector"

At each round $t = 1, \dots, T$ the gambler

- **Picks** a vector $p_t \in \mathcal{P} \subset \Delta_K = \{(p_1, \dots, p_K) \in [0,1]^K \big| \sum_k p_k = 1\}$

- **Receives** a reward $g_{t,p_t}$, with $\mathbb{E}[g_{t,p_t} | p_t] = p_t^T \theta$

$$\mathbf{R_T} = \mathbf{T} \times \mathbf{p^{\star}}^{\mathbf{T}} \boldsymbol{\theta} - \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{p}_t^{\mathbf{T}} \boldsymbol{\theta}\right]$$

# Bandit algorithm: LinUCB

(Li et al., 2010)

- **Estimate** parameters θ (Ridge regression) based on past observations

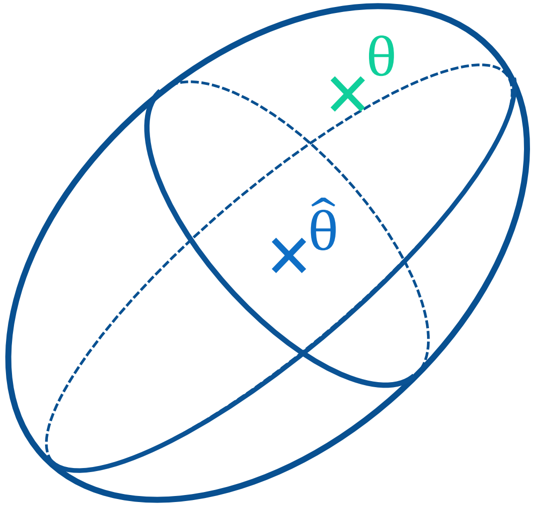
$$\hat{\theta}_{t-1} = \arg\min_{\hat{\theta}} \sum_{s=1}^{t-1} \left(g_{s,p_s} - p_s^T \hat{\theta}\right)^2 + \lambda \left\|\hat{\theta}\right\|^2$$

$$\hat{\theta}_{t-1} = V_{t-1}^{-1} \sum_{s=1}^{t-1} g_{s,p_s} p_s \quad \text{with } V_{t-1} = \lambda I_K + \sum_{s=1}^{t-1} p_s p_s^T$$

- **Build** confidence set for θ with high probability

$$\left\|\hat{\theta}_{t-1} - \theta\right\|_{V_{t-1}} \leq B_t \quad \text{with } B_t \propto \sqrt{\log t}$$

thus, $\left\|p^T\theta - p^T\hat{\theta}_{t-1}\right\| \leq B_t \|p\|_{V_{t-1}^{-1}}$




- **Be optimistic**

$$p_t = \arg\max_{p \in \mathcal{P}} p^T \hat{\theta}_{t-1} + B_t \|p\|_{V_{t-1}^{-1}}$$ which ensures $$R_T \lesssim \sqrt{TK \log^3 T}$$

# Stochastic Bandits **with context**

There is a set of contextual variables $\mathcal{X}$
Each arm (slot machine) $k$ has an unknown mean reward $\boldsymbol{\mu_k(x), x \in \mathcal{X}}$

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

- **Observes** a context $x_t$
- **Picks** a machine $I_t \in \{1, \ldots, K\}$
- **Receives** a reward $g_{t,I_t}$, with $\mathbb{E}[g_{t,I_t} \mid I_t] = \mu_{I_t}(x_t)$

$$R_T = \sum_{t=1}^{T} \mu_{k_t^\star}(x_t) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}(x_t)\right]$$

# Stochastic Linear Bandits **with context**

There is a unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{\mathbf{d}}$
The reward is linear in the feature vectors

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

- **Observes** a context $x_t$, a set $\mathcal{P} \subset \Delta^K$ of arms and feature vectors $\boldsymbol{\phi}(\mathbf{x_t}, \mathbf{p}) \in \mathbb{R}^{\mathbf{d}}$, $p \in \mathcal{P}$

The vector $\phi(x_t, p)$ summarizes information of both the context $x_t$ and arm $p$.

- **Picks** a vector $p_t \in \mathcal{P}$

- **Receives** a reward $g_t$, with $\mathbb{E}[g_t | p_t] = \phi(x_t, p_t)^T \theta$

$$R_T = \sum_{t=1}^{T} \phi(x_t, \mathbf{p_t^\star})^T \theta - \mathbb{E}\left[\sum_{t=1}^{T} \phi(x_t, p_t)^T \theta\right]$$

# Smart Meter Energy Consumption Data in London Households

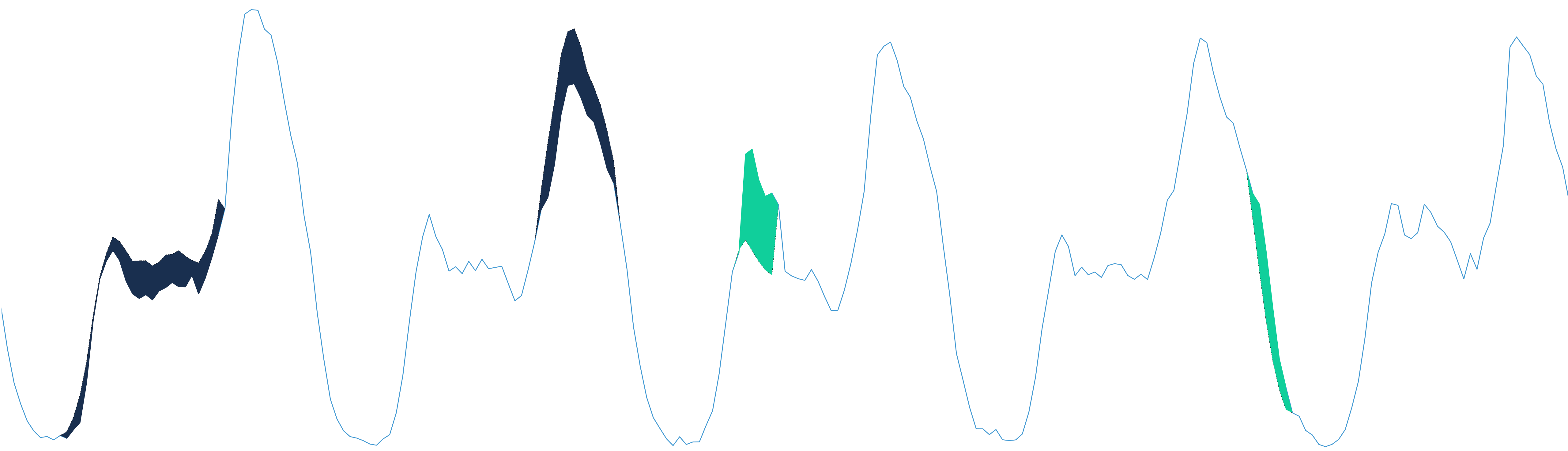
"Smart Meter Energy Consumption Data in London Households"
Public dataset - UK Power Networks

Individual consumption at half-an-hour intervals throughout 2013 of

~1 000 clients subjected to Dynamic Time of Use energy prices

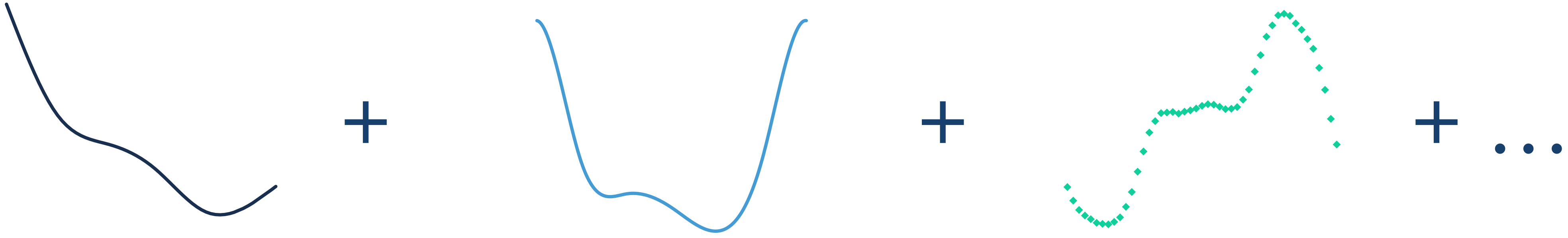
Three tariffs: **Low (L)**, **Normal (N)**, **High (H)**

# Tariff impact



# General Additive Model for power consumption

$$Y_t = f_1(\text{temperature}) + f_2(\text{position in the year}) + f_3(\text{hour}) + f_4(\text{tariff}) + \ldots + \text{noise}$$

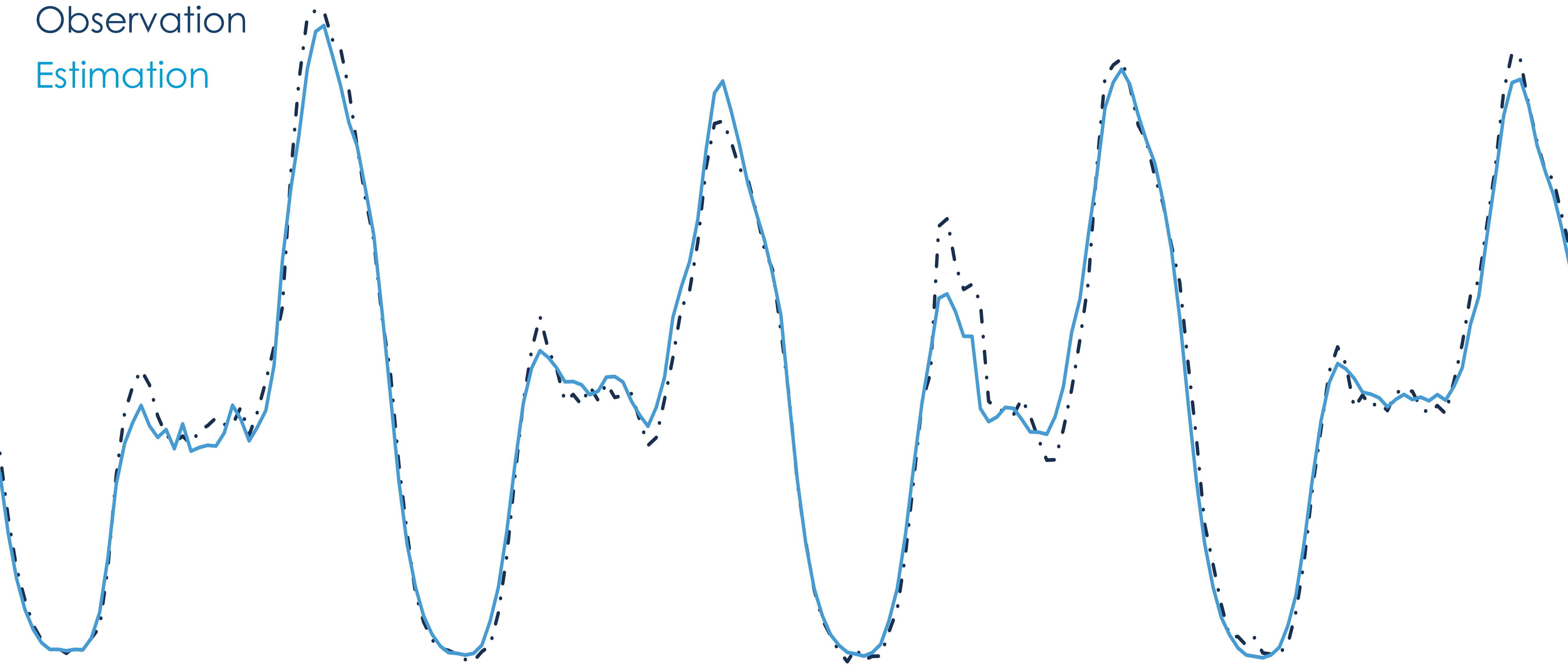


Temperature Position in the year Hour

→ There is a known transfer function $\phi$ and an unknown parameter $\theta$ such that

$$\mathbb{E}[\mathbf{Y}] = \boldsymbol{\phi}(\mathbf{X})^{\mathbf{T}}\boldsymbol{\theta}$$

# General Additive Model for power consumption



# Consumption modelling

**Assumption:**

- $K$ tariffs
- Homogenous population

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

- Observe a context $x_t \in \mathcal{X}$
- Choose proportions $\mathrm{p_t} \in \mathcal{P} \subset \Delta_K = \{(\mathrm{p_1}, \ldots, \mathrm{p_K}) \in [0,1]^K \,|\, \sum_k \mathrm{p_k} = 1\}$
- Observe the consumption $\mathbf{Y}_{\mathbf{t},\mathbf{p_t}} = \boldsymbol{\phi}(\mathbf{x_t}, \mathbf{p_t})^{\mathbf{T}} \boldsymbol{\theta} + \mathbf{p}_{\mathbf{t}}^{T} \boldsymbol{\varepsilon}_{\mathbf{t}}$

with $\mathbb{E}[\varepsilon_t] = (0, \ldots 0)^T$ and $\mathbb{V}[\varepsilon_t] = \Gamma \in \mathcal{M}_K(\mathbb{R})$

# Protocol: Target tracking for contextual bandits

**Input:**

- Transfer function $\phi: \mathcal{X} \times \mathcal{P} \rightarrow \mathbb{R}^d$

**Unknown parameters:**

- Transfer parameter $\theta \in \mathbb{R}^d$ and covariance matrix $\Gamma \in \mathcal{M}_K(\mathbb{R})$

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

- Observe a context $x_t \in \mathcal{X}$ and a **target** $c_t$
- Choose a vector $p_t \in \mathcal{P} \subset \Delta_K = \{(p_1, \ldots, p_K) \in [0,1]^K \mid \sum_k p_k = 1\}$
- Observe a resulting consumption $Y_{t,p_t} = \phi(x_t, p_t)^T \theta + p_t^T \varepsilon_t$ with $\mathbb{V}(\varepsilon_t) = \Gamma$
- Suffer a **loss** $\ell_t = \left(Y_{t,p_t} - c_t\right)^2$

# Minimize pseudo-regret – Estimate losses

**Aim:** Minimize the pseudo-regret (compare to the best strategy)

$$\mathbf{R_T} = \sum_{t=1}^{T} \ell_{t,p_t} - \sum_{t=1}^{T} \min_{p \in \mathcal{P}} \ell_{t,p}$$

with $\ell_{t,p} = \mathbb{E}\left[\left(Y_{t,p} - c_t\right)^2\right] = \left(\phi(x_t, p)^T \theta - c_t\right)^2 + p^T \Gamma p$

- Reach a bias-variance trade-off
- Estimate parameters $\theta$ and $\Gamma$ to estimate losses !

# Optimistic algorithm for tracking target with context

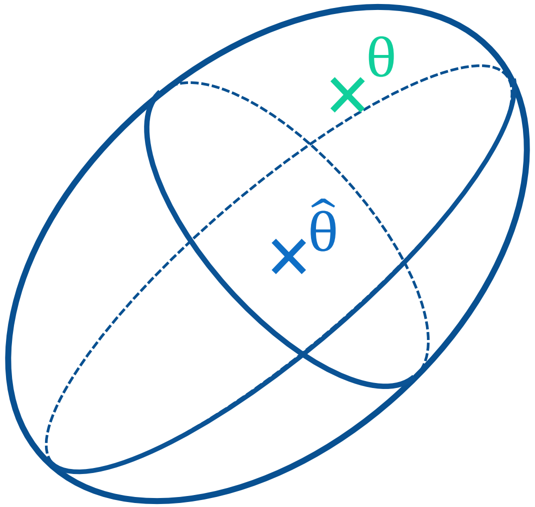
- **Estimate parameters** $\theta$ (Ridge regression) and $\Gamma$ ($\hat{\Gamma}_{t-1}$ provided in the article)

$$\hat{\theta}_{t-1} = \arg\min_{\hat{\theta}} \sum_{s=1}^{t-1} \left(Y_{s,p_s} - \phi(x_s, p_s)^T \hat{\theta}\right)^2 + \lambda \|\hat{\theta}\|^2$$

$$\hat{\theta}_{t-1} = V_{t-1}^{-1} \sum_{s=1}^{t-1} Y_{s,p_s} \phi(x_s, p_s) \quad \text{with } V_{t-1} = \lambda I_d + \sum_{s=1}^{t-1} \phi(x_s, p_s) \phi(x_s, p_s)^T$$

- **Build confidence sets** for $\theta$ and $\Gamma$

$$\|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}} \leq B_t \text{ and } \|\hat{\Gamma}_{t-1} - \Gamma\|_\infty \leq \gamma_t$$



# Optimistic algorithm for tracking target with context

- **Estimate the future loss** $\ell_{t,p}$ for each price level

As $\ell_{t,p} = \mathbb{E}\left[\left(Y_{t,p} - c_t\right)^2\right] = \left(\phi(x_t, p)^T\theta - c_t\right)^2 + p^T\Gamma p$

$$\hat{\ell}_{t,p} = \left(\phi(x_t, p)^T\hat{\theta}_{t-1} - c_t\right)^2 + p^T\hat{\Gamma}_{t-1}p$$

- **Get a confidence bound for losses** for each $p$ thanks to $B_t$ and $\gamma_t$

$$\left\|\hat{\ell}_{t,p} - \ell_{t,p}\right\| \leq \alpha_{t,p}$$

# Optimistic algorithm for tracking target with context

Inspired from Lin-UCB (Li et al. 2010)

- **Estimate parameters** $\theta$ and $\Gamma$ from observations ($\hat{\Gamma}_{t-1}$ provided in the article)
- **Estimate the future loss** $\ell_{t,p}$ for each price level

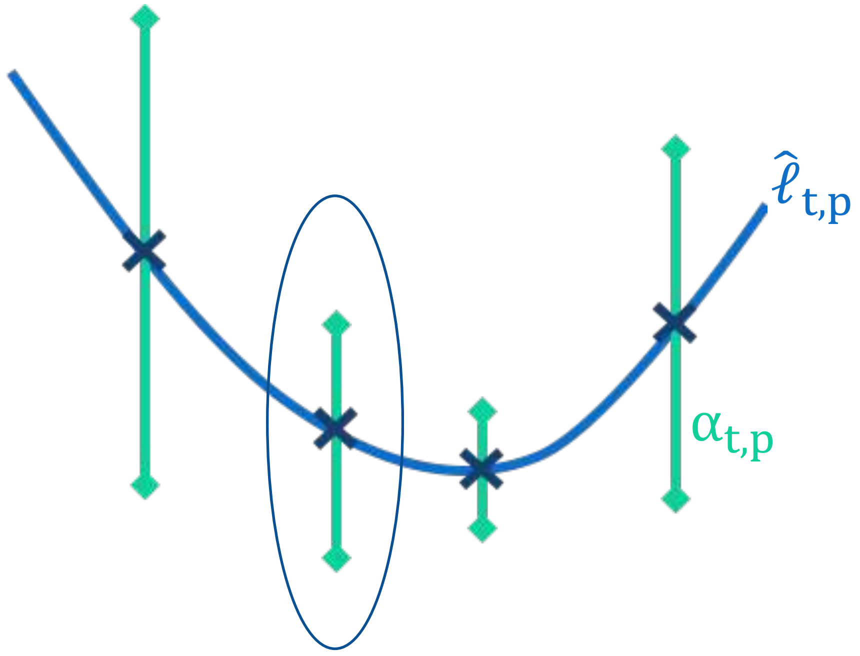
$$\hat{\ell}_{t,p} = \left(\phi(x_t, p)^T \hat{\theta}_{t-1} - c_t\right)^2 + p^T \hat{\Gamma}_{t-1} p$$

- **Build confidence sets** for $\theta$ and $\Gamma$
- **Get a confidence bound for losses** for each $p$

$$\left\|\hat{\ell}_{t,p} - \ell_{t,p}\right\| \leq \alpha_{t,p}$$

- Select price level **optimistically**

$$\mathbf{p_t} \in \operatorname*{\mathbf{arg\,min}}_{\mathbf{p} \in \mathcal{P}} \left\{\hat{\ell}_{\mathbf{t,p}} - \alpha_{\mathbf{t,p}}\right\}$$



# Theoretical guarantee

**Theorem**

For proper choices of confidence levels $\alpha_{t,p}$, $B_t, \gamma_t$ and regularisation $\lambda$, with probability at least $\mathbf{1-\delta}$ the regret is upper bounded as

$$R_T = \sum_{t=1}^{T} \ell_{t,p_t} - \sum_{t=1}^{T} \min_{p \in \mathcal{P}} \ell_{t,p_t} \lesssim \mathbf{T^{2/3}} \ln^2 (T/\delta) \sqrt{\ln(1/\delta)}$$

**Limitation**

The optimization problem $p_t \in \arg\min_{p \in \mathcal{P}} \{\hat{\ell}_{t,p} - \alpha_{t,p}\}$ is nonconvex and hard to solve.

- Restrict $\mathcal{P}$

# Back to data !

- "Smart-Meter Energy Consumption Data in London Households"

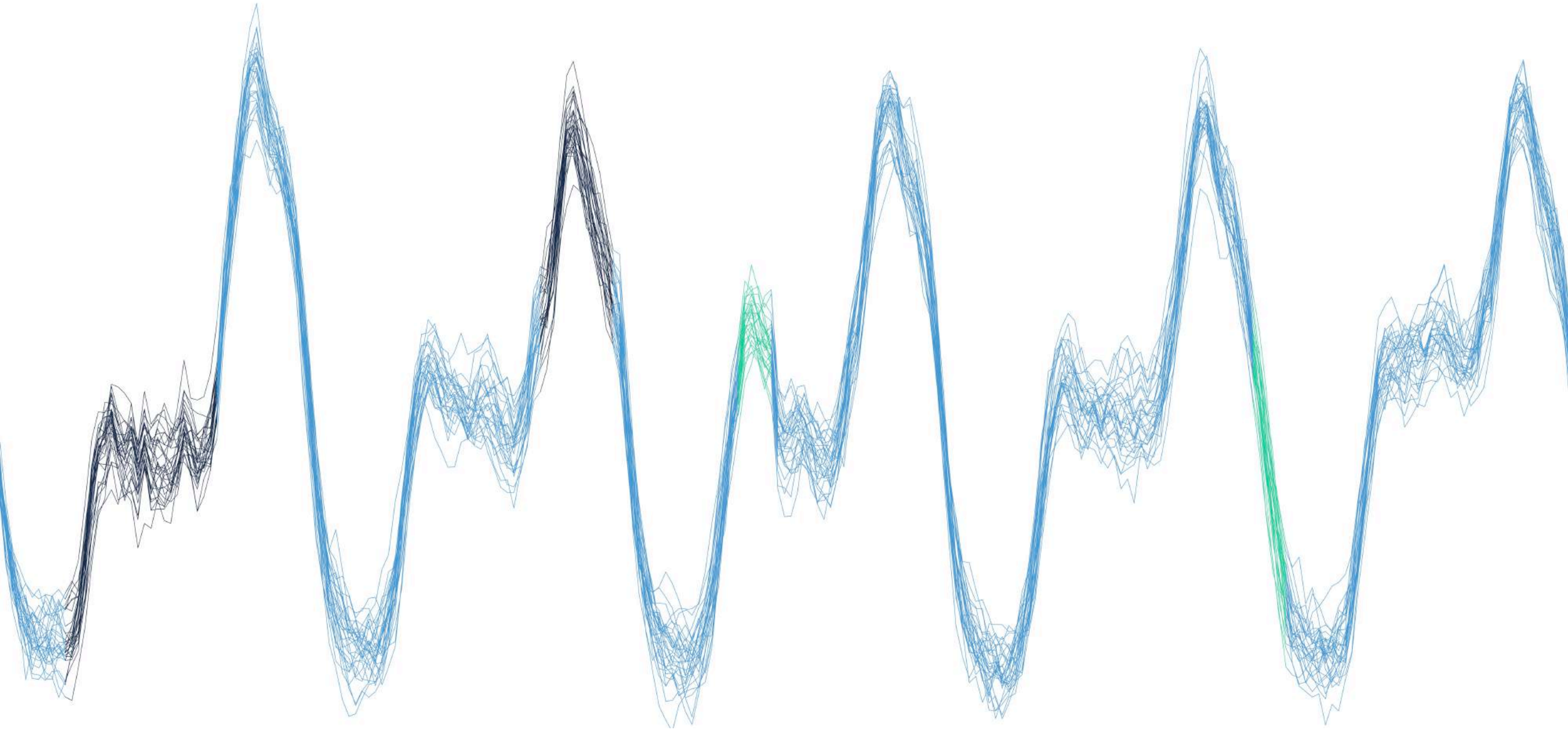
A single tariff - Low (L), Normal (N) or High (H) - is offered to all the population for each half hour interval.

- Select customers with more than 95% of data available (980 clients) and consider their mean consumption.

- Build a realistic simulator (based on Generalized Additive Model) assuming homogeneous customers

**Context + Price level → Global consumption**

# Simulator



# Design of the experiment

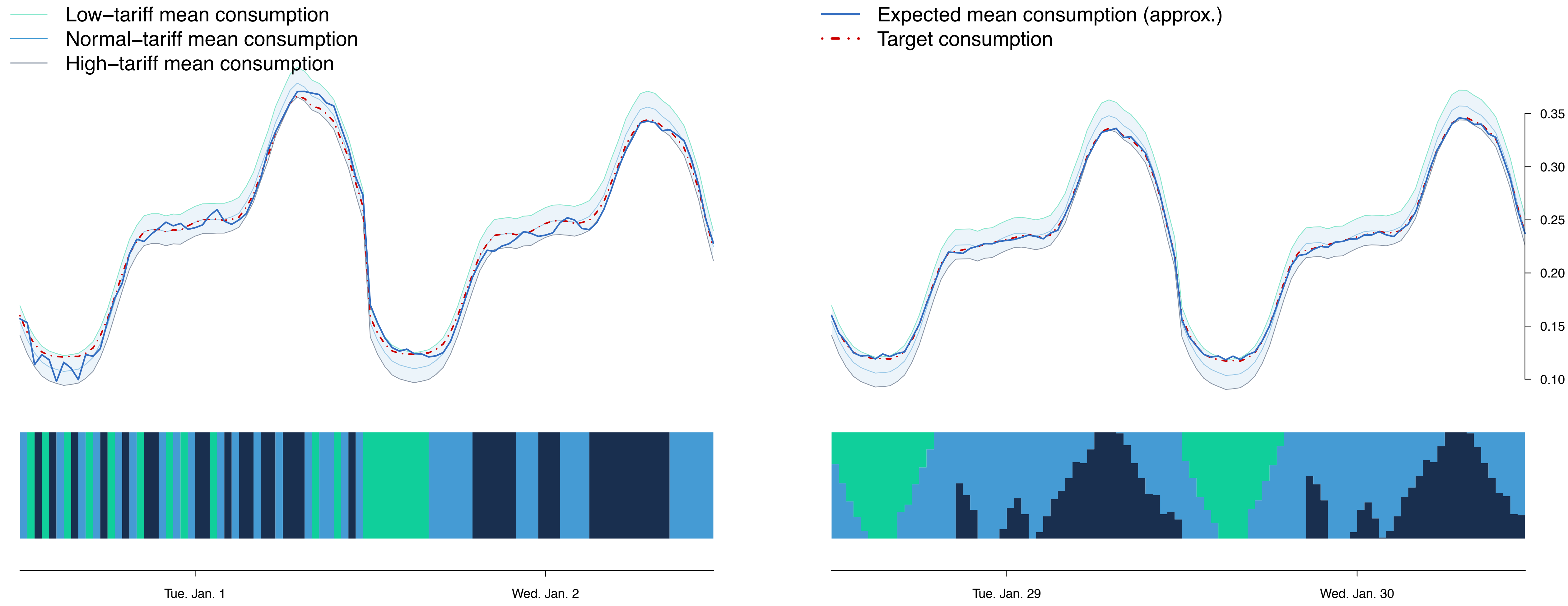
- **Target creation:** attainable targets which stay in the convex envelope of the mean consumption associated to the High and Low tariffs

- $\mathcal{P}$ **restriction** (to a grid): electricity provider cannot send Low and High tariffs at the same round and the population can be split in 100 equal parts

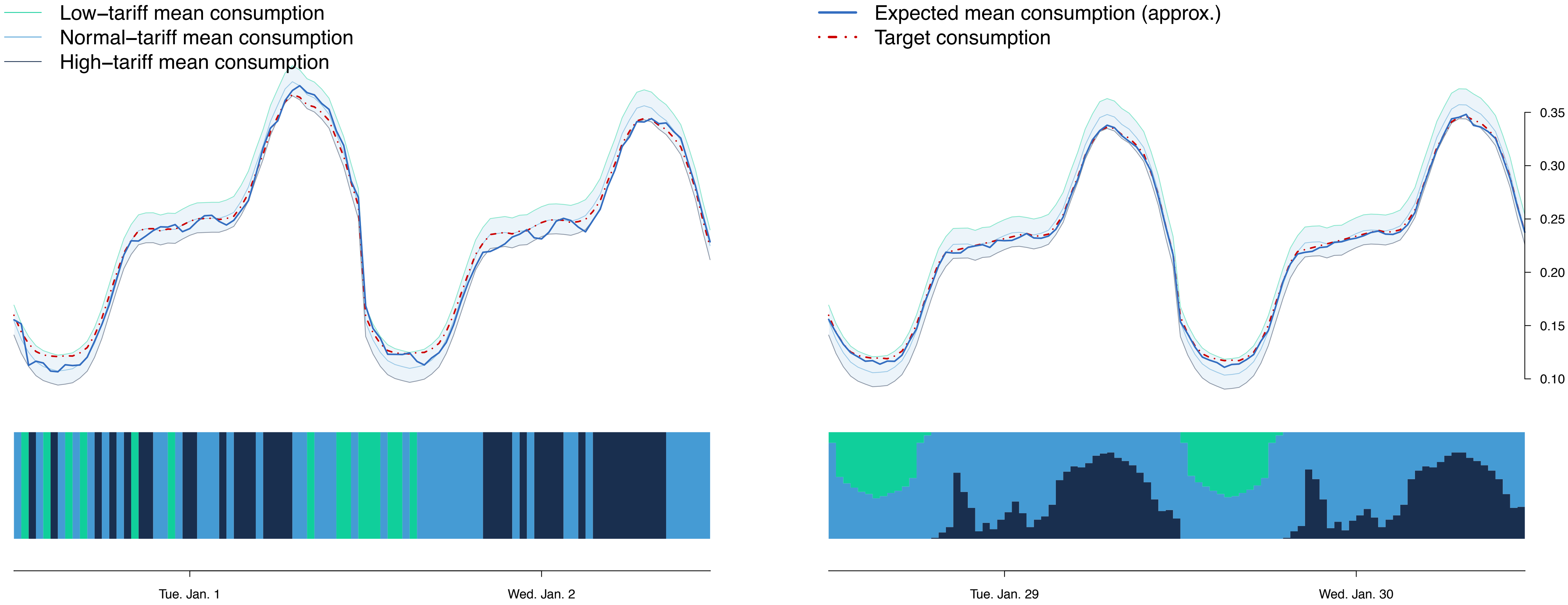
- **Training period:** one year of data using historical contexts and assuming that only Normal tariff is picked

- **Testing period**: for an additional month (based on the historical contexts) tariffs are picked according to the algorithm

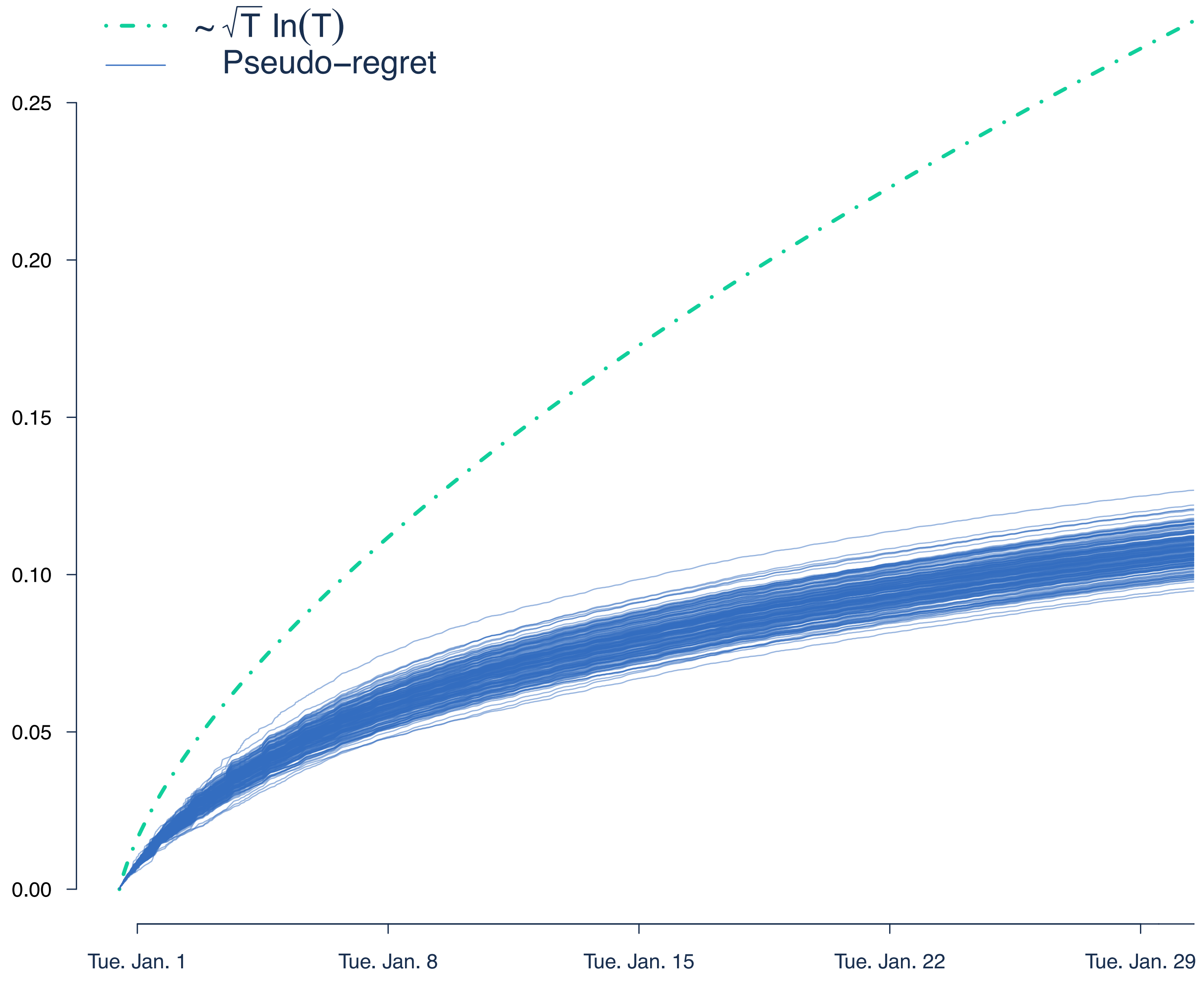
# Results: overlap the target



# Result: bias-variance trade-off



# Result: what about pseudo-regret ?



# Conclusions and perspectives

**Summary**

- Design, implement and test an efficient algorithm with theoretical guaranties to track a target consumption under basic assumptions.

**What's next?**

- More experiments, simulations
- Non homogeneous consumers: create client clusters to send individual signals (device dependent, clients with battery) and improve power consumption control.
- More complex models? Anticipation of future high prices, ...
- Operational constraints

# References

## Thank you!


- Auer, P. et al. (2002). “Finite-time analysis of the multiarmed bandit problem”. Machine learning.
- Brégère, M. et al. (2019). “Target Tracking for Contextual Bandits: Application to Power Consumption Steering”.
- Hastie, T. J. and R. J. Tibshirani (1990). “Generalized additive models.”
- Lai, T. L. and H. Robbins (1985). “Asymptotically efficient adaptive allocation rules”. Advances in applied mathematics.
- Li, L. et al. (2010). “A contextual-bandit approach to personalized news article recommendation”. Proceedings of the 19th International Conference on World Wide Web (WWW’10).
- Pierrot, A. and Y. Goude (2011). “Short-term electricity load forecasting with generalized additive models”. Proceedings of ISAP power.
- Schofield, J. et al. (2014). “Residential consumer responsiveness to time-varying pricing.”