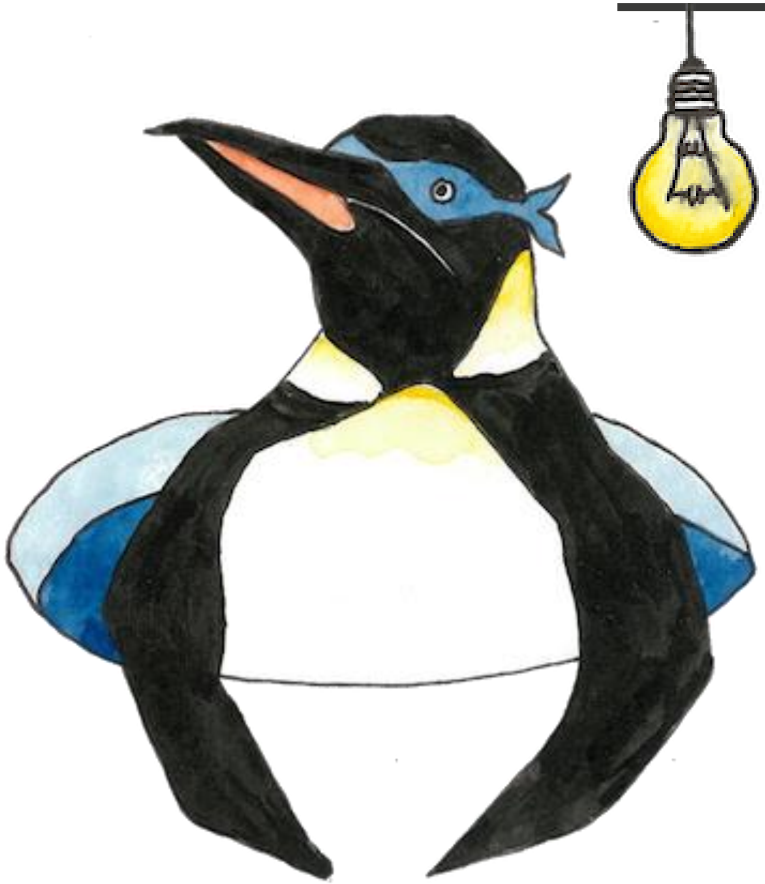
Stochastic Bandit Algorithms for Demand Side Management

Margaux Brégère • Nov. 18, 21 Rencontres chercheur·euse·s et ingénieur·e·s • Phiméca

Under the supervision of Gilles Stoltz, Yannig Goude and Pierre Gaillard



“ Bandit manchot” is the French translation for “one-armed bandit”; however, a word-to-word translation would be “crook penguin”.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

EDF

Inria INVENTEURS DU MONDE NUMÉRIQUE

Introduction - Motivation

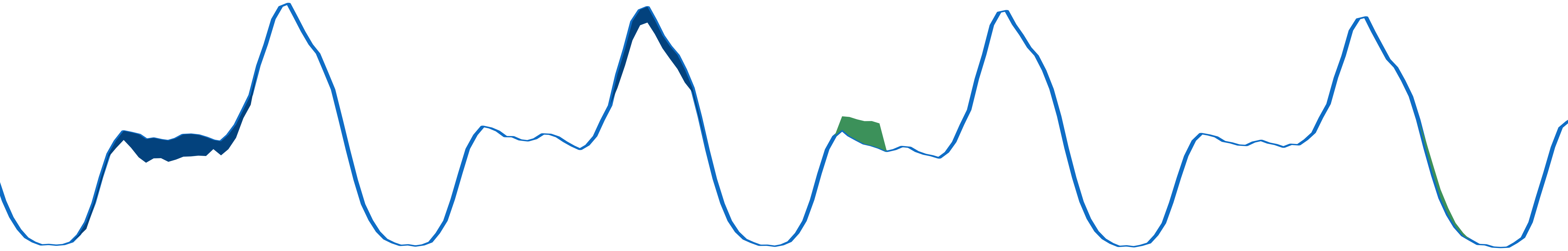
As electricity is hard to store, balance between production and demand must be strictly maintained

Current solution: forecast demand and adapt production accordingly

- With the development of renewable energies, production becomes harder to adjust
- New (smart) meters provide access to data and instantaneous communication

Prospective solution: send incentive signals (electricity tariff variations) to manage demand response

Introduction - Motivation



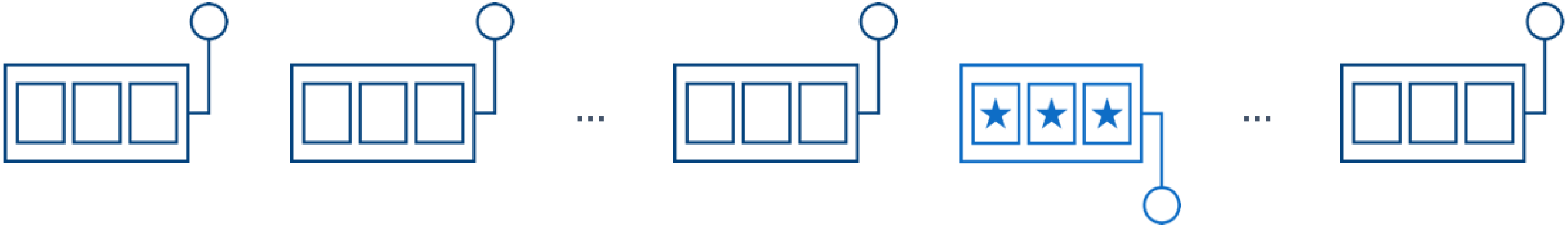
How to develop automatic solutions
to chose incentive signals dynamically?

Exploration: learn consumer behavior

Exploitation: optimize signal sending

Apply mathematical bandit theory to the sequential learning problem of demand side management

Stochastic multiarmed bandit



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 multiarmed bandit

Each arm (slot machine) *k* is defined by an unknown probability distribution ν_k with $\mu_k = \mathrm{E}[\nu_k]$.

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

- Picks a machine $I_t \in \{1, \ldots, K\}$
- Receives a reward Y_t, with $Y_t \mid I_t = k \sim \nu_k$

Maximizing the expected cumulative reward = Minimizing pseudo-regret

$$R_T = \overbrace{T \max_{k=1,\ldots,K} \mu_k}^{\text{Mean reward of the best machine is known}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right]}_{\text{Mean reward of the strategy}}$$

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

Upper Confidence Bound (UCB) algorithm (Lai *et al.* 1985)

- Estimate the expectations μ_k (empirical means) based on past observations:

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

- Build a confidence interval for the expectations μ_k with high probability

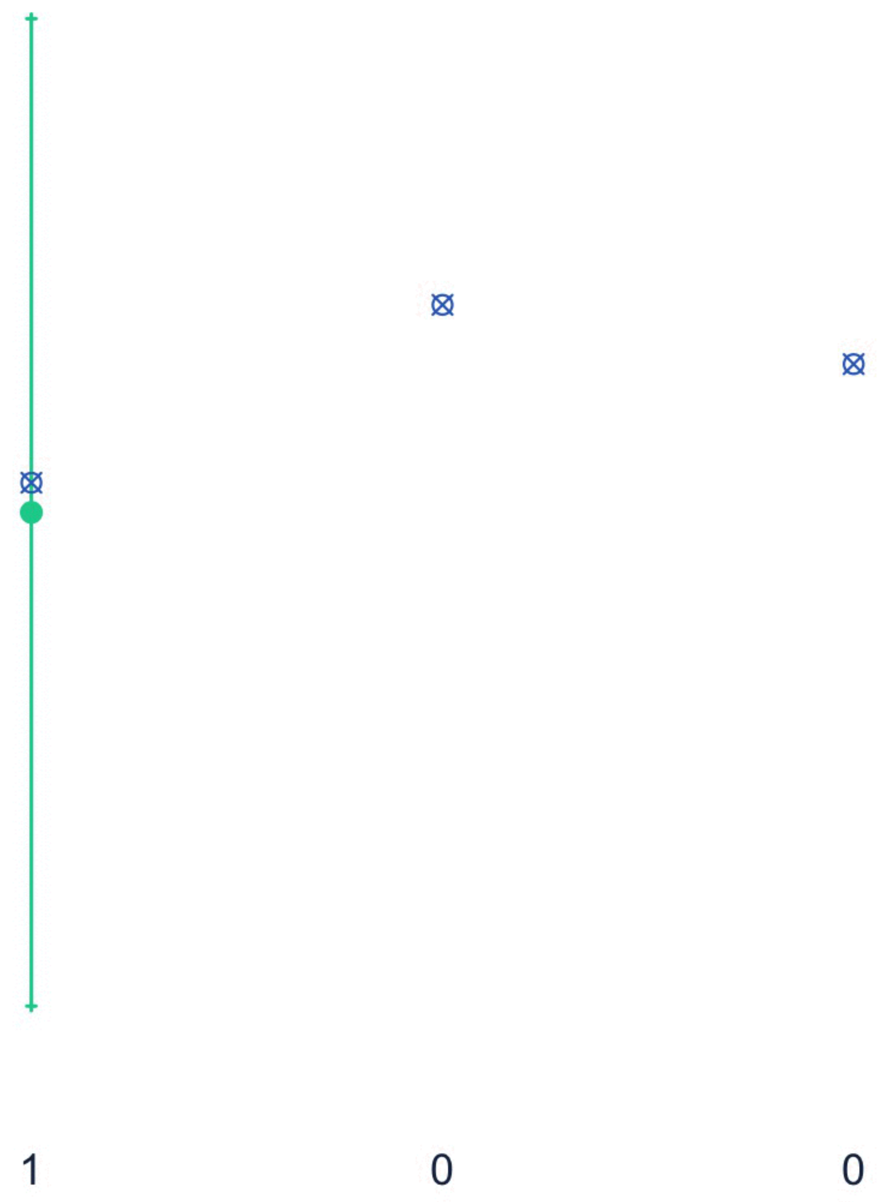
With probability at least $1 - t^{-3}$
(Hoeffding-Azuma Inequality)

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

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

$$I_t = \underset{k \in \{1,\dots,K\}}{\arg\max} \ \hat{\mu}_{t-1,k,} + \alpha_{t,k} \qquad \text{which ensures} \qquad R_T \lesssim \sqrt{T\,K \log T}$$

T = 1



First of all: modeling

How to model electricity demand?

‣ Using classical (for EDF) power consumption forecasting methods

How to formalize the sequential learning?

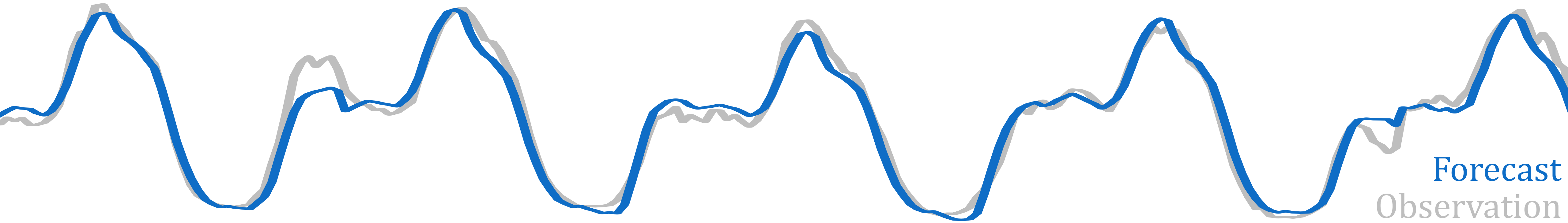
‣ Defining a protocol (under some assumptions)

Generalized additive models for electricity demand

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

- There is a known transfer function ϕ and an unknown parameter θ such that

$$Y_t = \phi(\text{temperature}, \text{position in the year}, \text{hour}, \text{tariff} \dots)^T \theta + \text{noise}$$



Electricity demand modeling

Assumption:

- K tariffs
- Homogenous population

At each round $t = 1, \ldots$

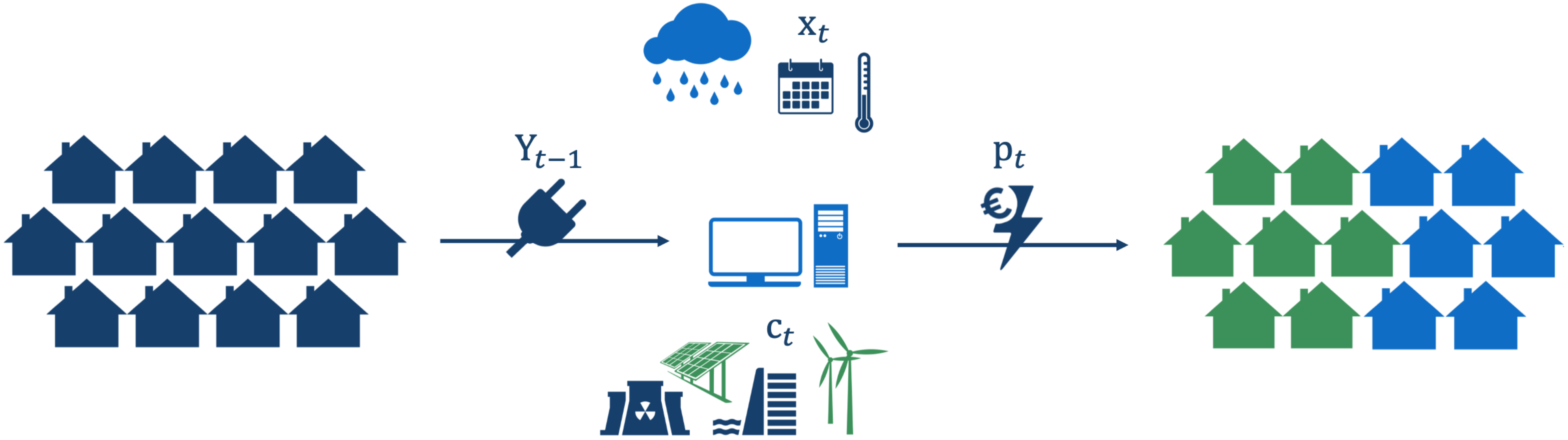
- Observe a context x_t
- Choose price levels p_t
- Observe the electricity demand $Y_t = \phi(x_t, p_t)^T\theta + p_t^T \varepsilon_t$

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

Protocol for target tracking

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

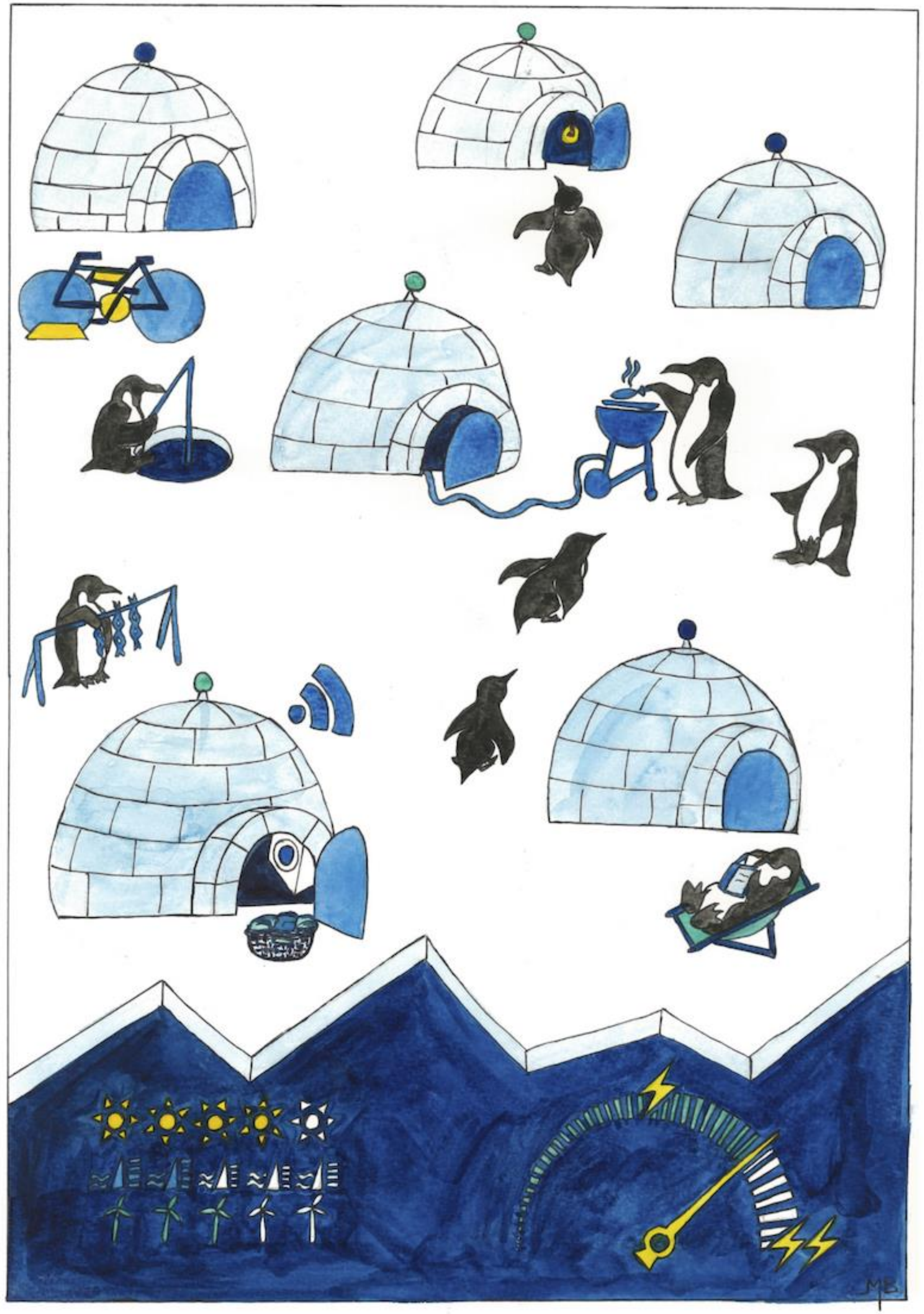
- Observe a context x_t and a target c_t
- Choose price levels p_t
- Observe the resulting demand Y_t and suffer a loss $(Y_t - c_t)^2$



Bandit algorithm for the management of a homogenous population

How to evaluate a target tracking algorithm?
‣ Defining a regret criterion

How to adapt existing bandit theory?
‣ Adapting LinUCB algorithm (Li et al. 2010)



Joint work with Pierre Gaillard, Yannig Goude and Gilles Stoltz, International Conference on Machine Learning, 2019

Protocol: target tracking for contextual bandits

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

- Observe a context x_t and a target c_t
- Choose price levels p_t
- Observe a resulting demand $Y_t = \phi(x_t, p_t)^T \theta + p_t^T \varepsilon_t$ with $\mathbb{V}[\varepsilon_t] = \Sigma$
- Suffer a loss $(Y_t - c_t)^2$ such that

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

Aim: minimize the pseudo-regret

$$R_T = \sum_{t=1}^{T} \left(\phi(x_t, p_t)^T \theta - c_t\right)^2 + p_t^T \Sigma p_t - \sum_{t=1}^{T} \min_p \left(\phi(x_t, p)^T \theta - c_t\right)^2 + p^T \Sigma p$$

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

Optimistic algorithm

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

For $t = 1,2, \dots, \tau$

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

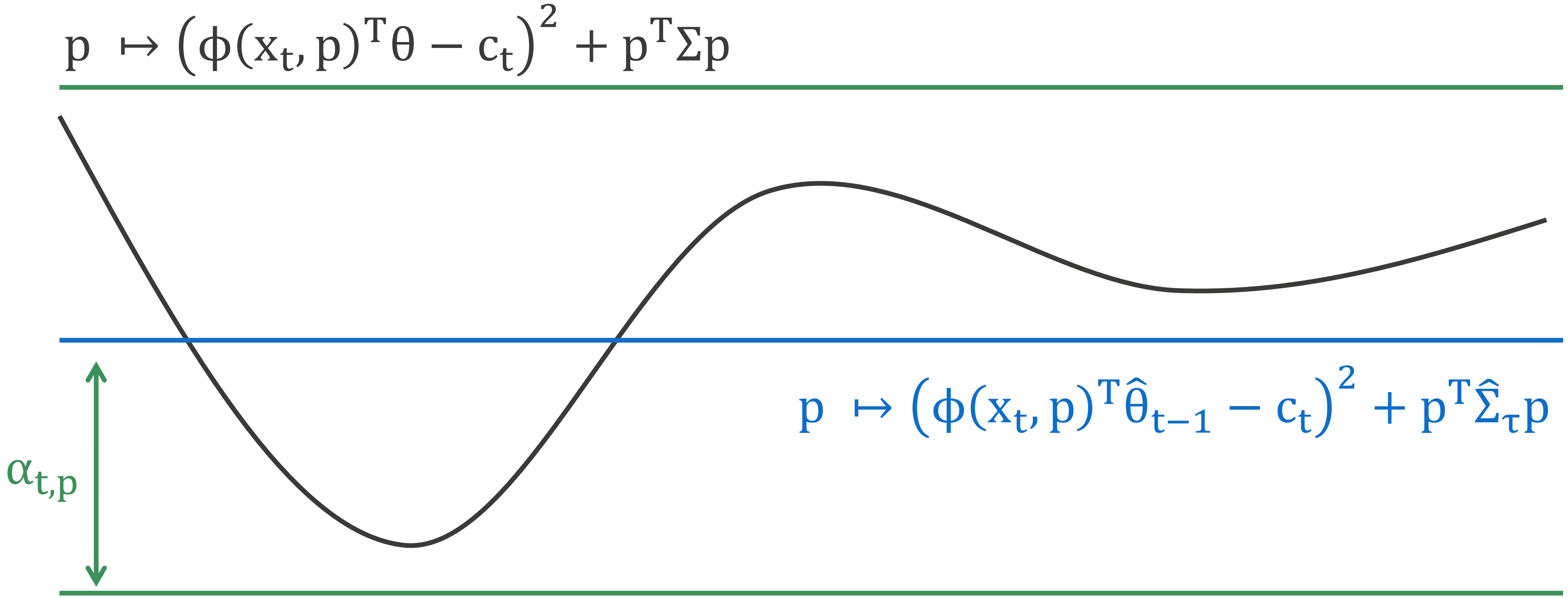
For $t = \tau, \dots, T$

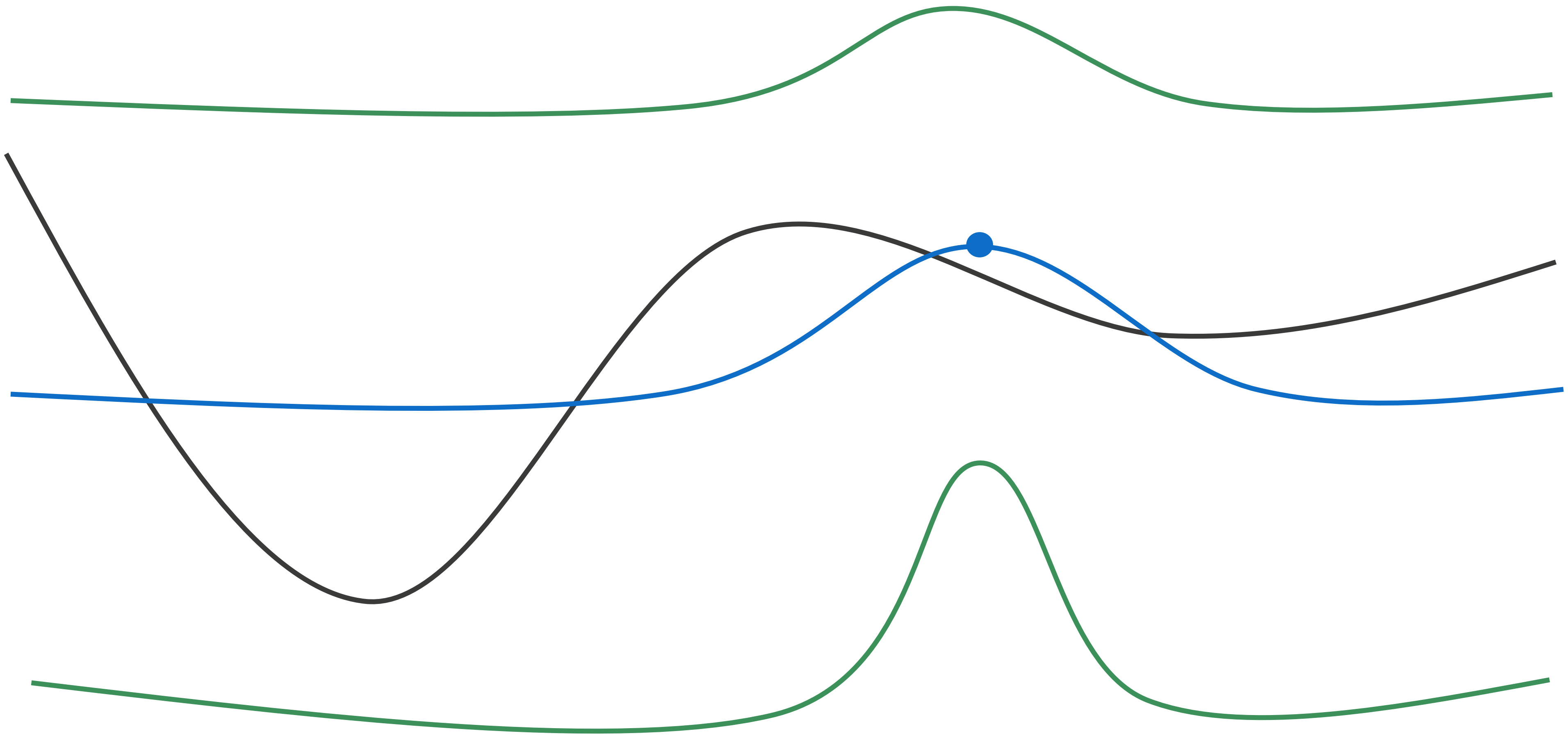
- Estimate θ based on past observations with $\hat{\theta}_{t-1}$ (Ridge regression)
- Estimate the future expected loss for each p: $\left(\phi(x_t, p)^T\hat{\theta}_{t-1} - c_t\right)^2 + p^T\hat{\Sigma}_\tau p$
- Get a confidence bound for each p

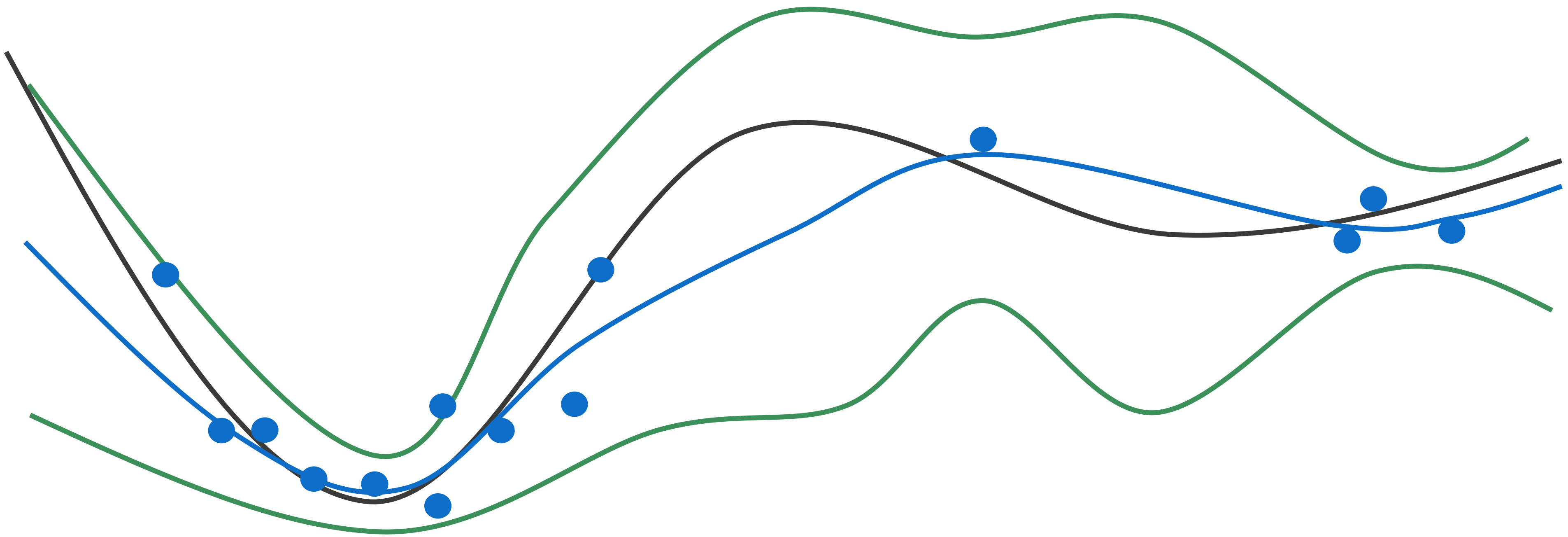
$$\left\| \left(\phi(x_t, p)^T\hat{\theta}_{t-1} - c_t\right)^2 + p^T\hat{\Sigma}_\tau p \; - \left(\phi(x_t, p)^T\theta - c_t\right)^2 + p^T\Sigma p \right\| \leq \alpha_{t,p}$$

- Select price levels optimistically

$$p_t \in \arg\min_p \left\{ \left(\phi(x_t, p)^T\hat{\theta}_{t-1} - c_t\right)^2 + p^T\hat{\Sigma}_\tau p \; - \alpha_{t,p} \right\}$$


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





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

Regret bound

Theorem

For proper choices of confidence levels $\alpha_{t,p}$ and number of exploration rounds τ, with high probability

$$R_T = \sum_{t=1}^{T} \left(\phi(x_t, p_t)^T\theta - c_t\right)^2 + p_t^T \Sigma p_t - \sum_{t=1}^{T} \min_{p\in\mathcal{P}} \left\{\left(\phi(x_t, p)^T\theta - c_t\right)^2 + p^T\Sigma p\right\} \leq \mathcal{O}\left(T^{2/3}\right)$$

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

Elements of proof

- Deviation inequalities on $\hat{\theta}_t$ [1] and on $\hat{\Sigma}_\tau$
- Inspired from LinUCB regret bound analysis [2]

[1] Laplace's method on supermartingales: Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits, 2011

[2] Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions, 2011

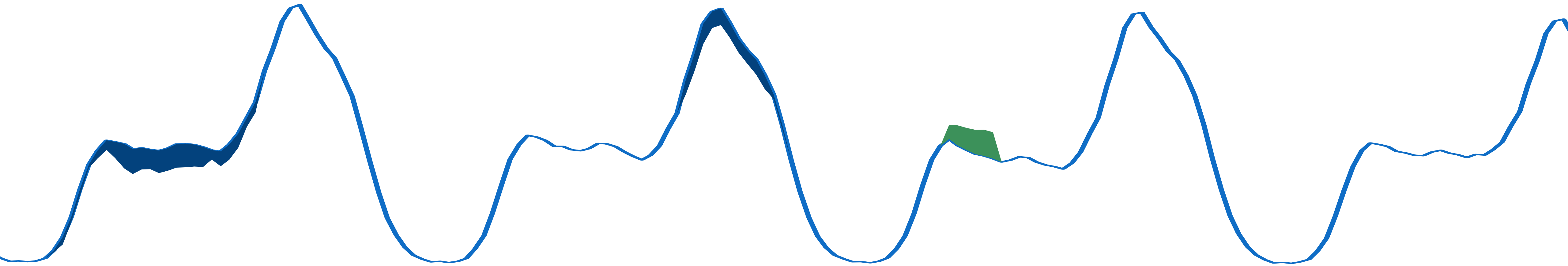
Smart Meter Energy Consumption Data

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

Individual electricity demand at half-an-hour intervals throughout 2013 of

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

Three tariffs: Low, Normal, High



Design of the experiment

- Alternative policies cannot be tested on historical data... How to test bandit algorithms?
  - Simulating data with $Y_t = f(x_t) + p_t^T \begin{bmatrix} \xi_{Low} \\ \xi_{Normal} \\ \xi_{High} \end{bmatrix} + p_t^T \varepsilon_t$ and $\mathbb{V}[\varepsilon_t] = \Sigma$

where $\Sigma = \begin{pmatrix} \sigma_{Low} & 0 & 0 \\ 0 & \sigma_{Normal} & 0 \\ 0 & 0 & \sigma_{High} \end{pmatrix}$

Experiment 1: $\sigma_{Low} = \sigma_{Normal} = \sigma_{High}$

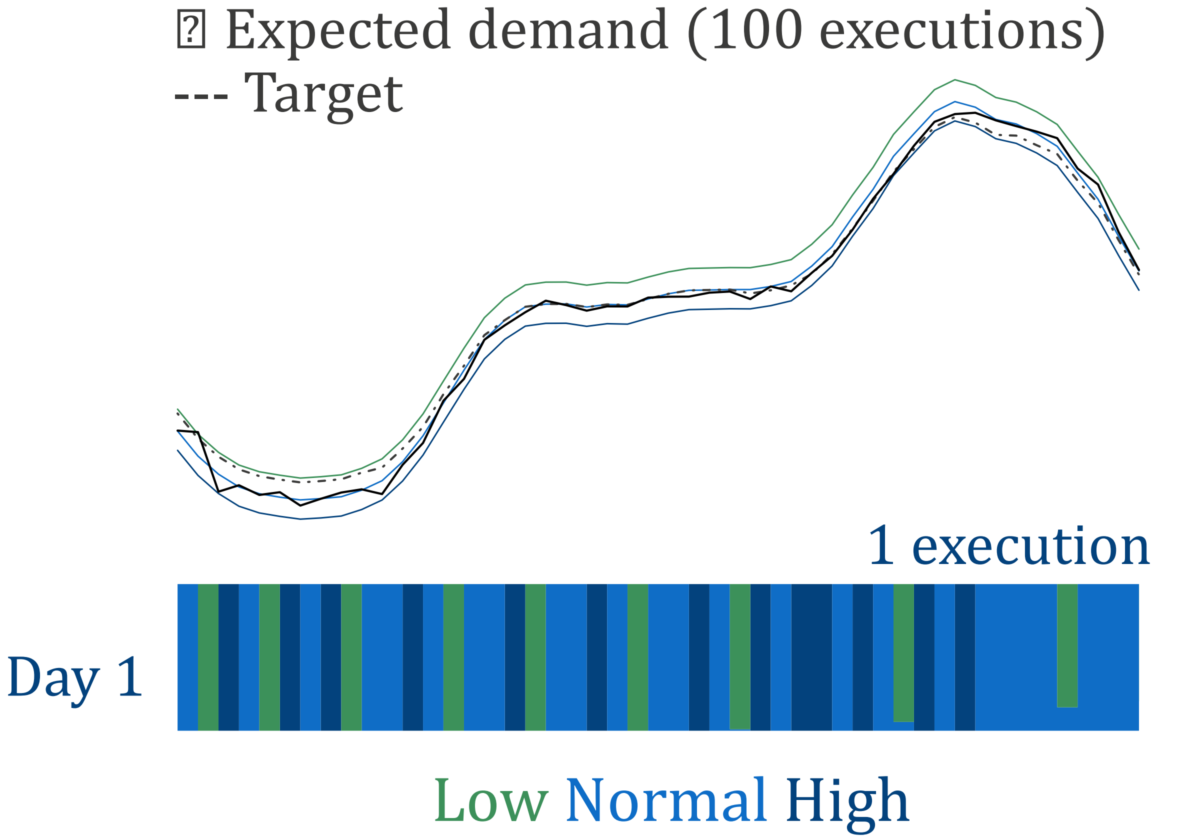
Experiment 2: $\sigma_{Low} > \sigma_{High} > \sigma_{Normal}$

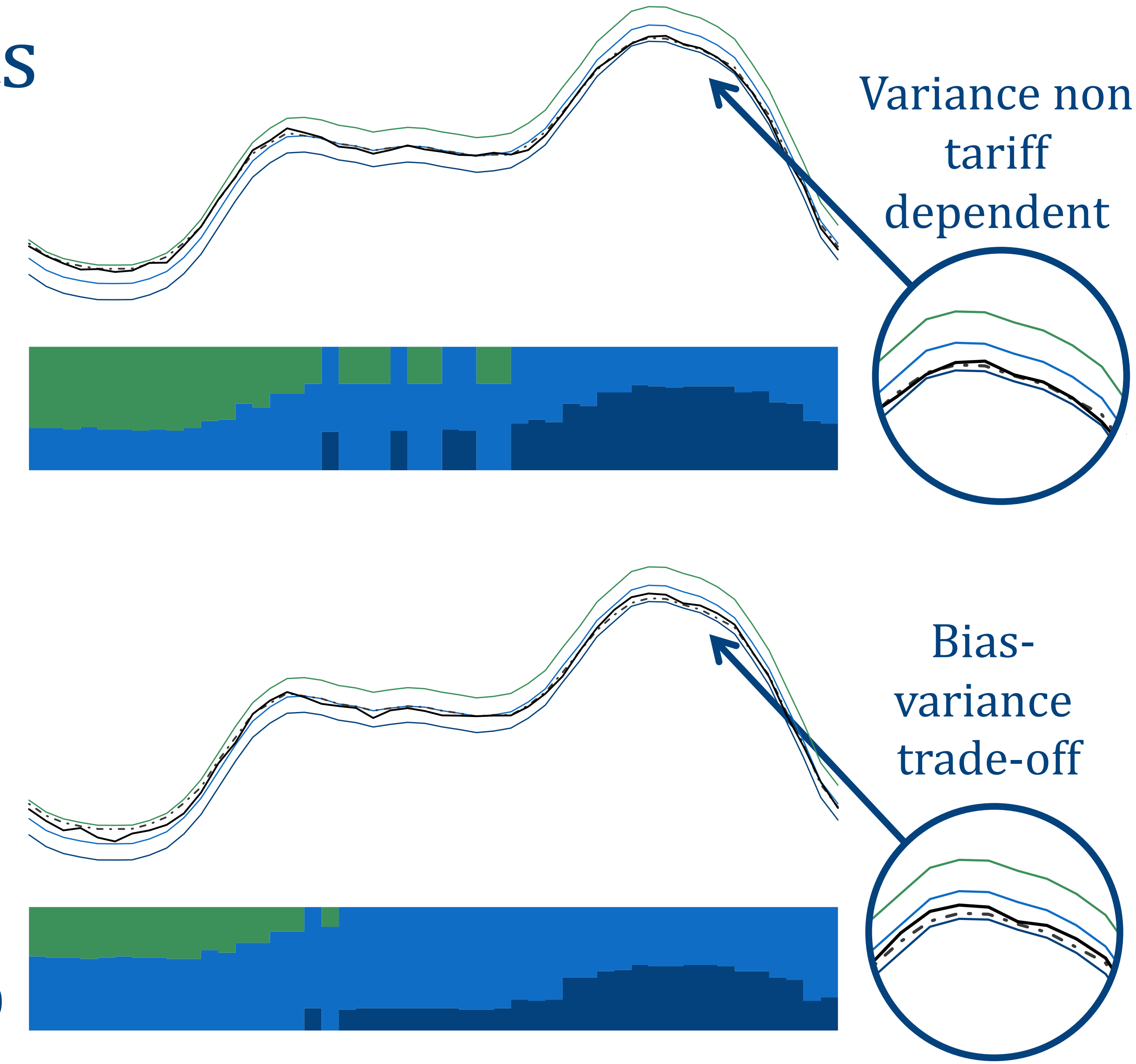
- Which target to choose?
  - Close to average High demand during the evening
  - Close to average Low demand during the night

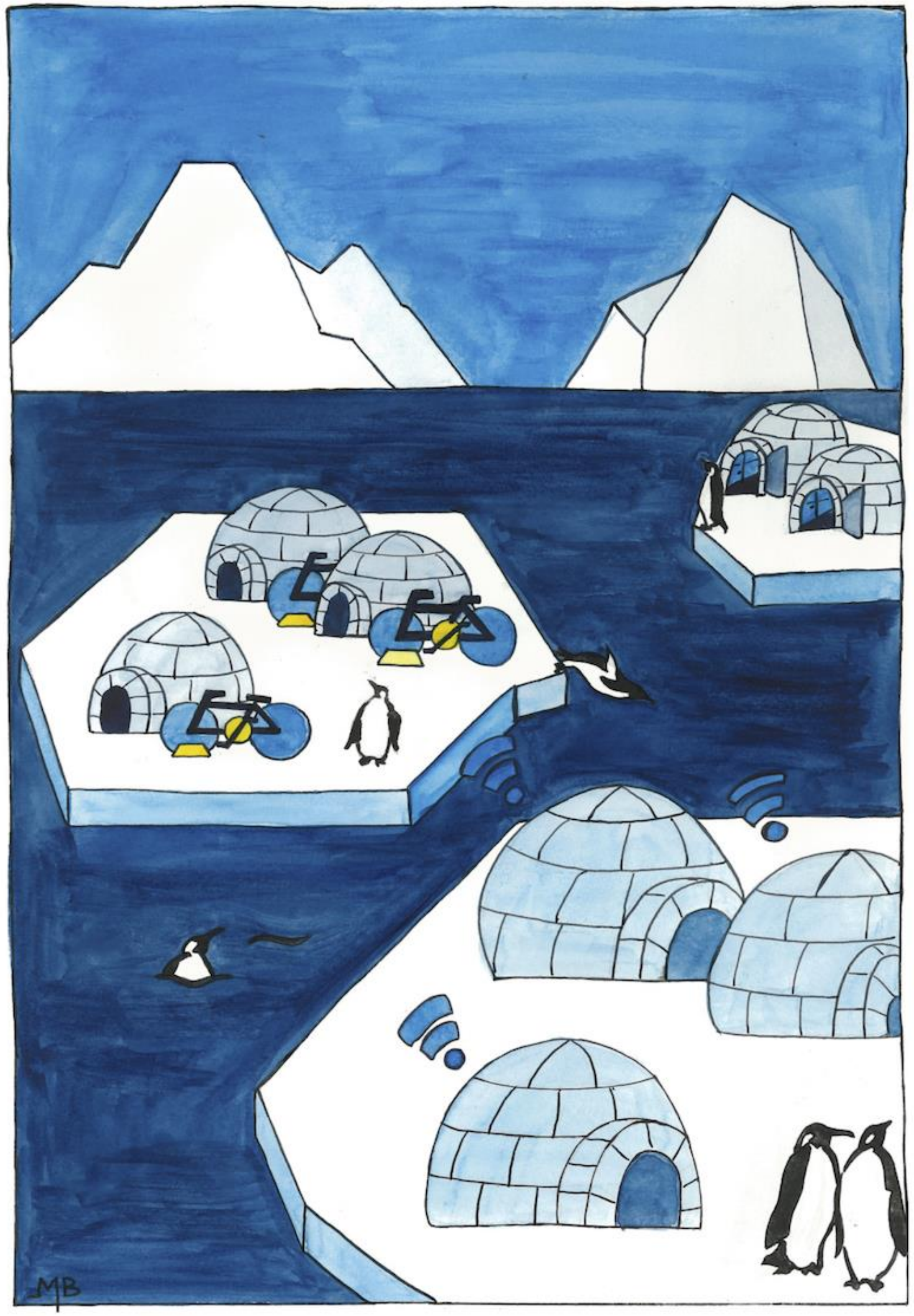
- Which context to choose?
  - Algorithm executed on historical context

- Operational constraints on legible allocations of price levels:
  - Impossible to send Low and High tariffs at the same time
  - Population split in 100 equal subsets

First experimental results







Towards the application of theoretical results

How to drop
the homogenous population assumption?

- Clustering households (or igloos)

How to test bandit algorithms?

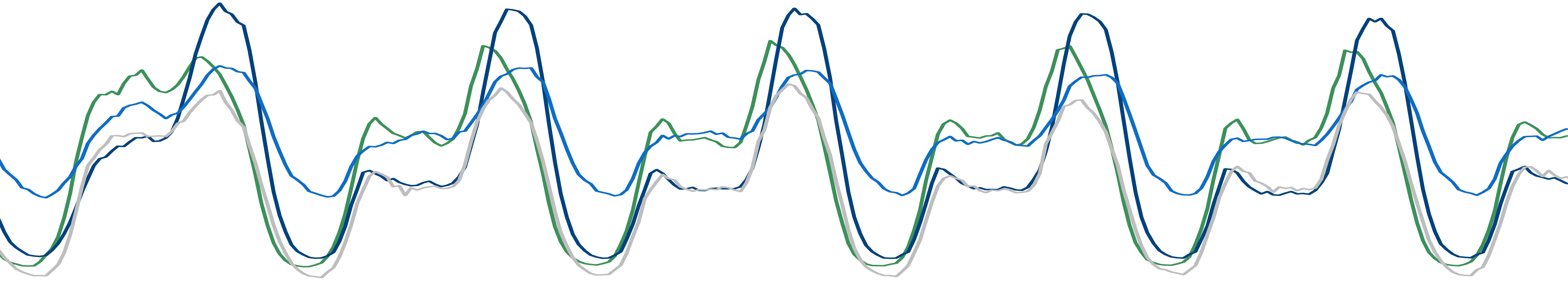
- Using a data simulator

Dropping the homogeneous population assumption

Double segmentation:

- geographical, based on region information
- behavioral:

Historical individual time series → Feature creation using low rank approximation → Clustering algorithm (k-medoid) → k clusters



Simulating electricity demand

- A semi-parametric approach with "generalized additive models + noise"
  - Illustrate the theory

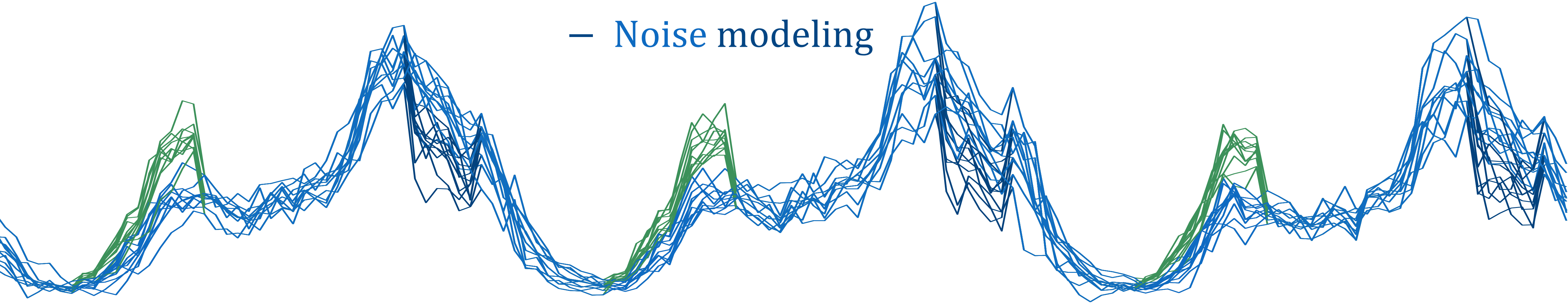
- A black-box approach with conditional variational auto-encoders
  - Test the algorithm robustness

Joint work with Ricardo Jorge Bessa, IEEE access, 2020

Demand generated for different tariff signals

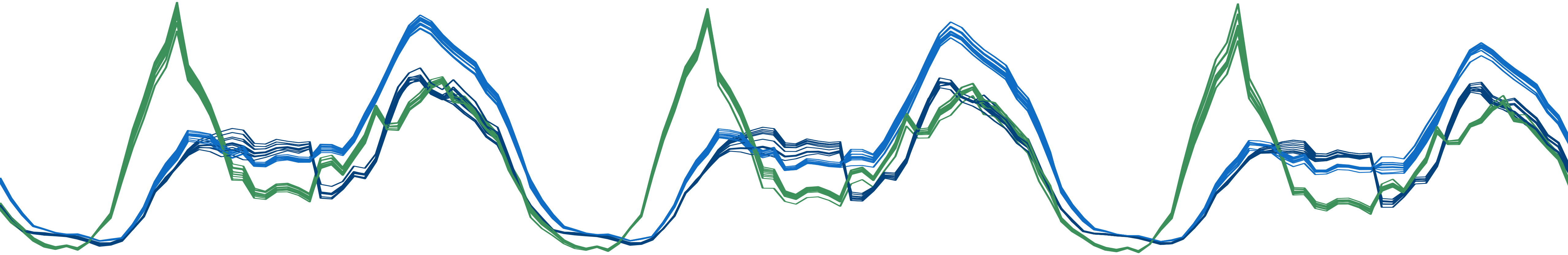
- Semi-parametric generator: + Interpretable

  − Noise modeling



- Black-box generator: + Rebound effect

  − Limited generalization capacity → transfer learning



Synthesis - Operational demand side management

- Personalizing incentive signals according to
    - Local meteorological condition
    - Consumption behavior

- Taking into account operational
    - Network constraints (renewable energies integration)
    - Commercial constraints (electricity supply contract)

Personalized demand side management

Protocol

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

- Observe G contexts $(x_t^i)_{i=1,\ldots,G}$
- Observe some sub-targets, which may correspond to renewable energy production, c_t^g, with $g \in \mathcal{P}(1, \ldots, G)$ and some weights κ_t^g
- For $i = 1, \ldots, G$
  - Choose price levels $p_t^i \in$ {prices allowed by the electricity contract at t}
  - Observe the resulting demand $Y_t^i = \phi^i\left(x_t^i, p_t^i\right)^T \theta^i + p_t^{i\,T} \varepsilon_t^i$, with $\mathbb{V}\left[\varepsilon_t^i\right] = \Sigma^i$
- Suffer a loss

$$\sum_g \kappa_t^g \left(\sum_{i \in g} Y_t^i - c_t^g \right)^2$$

Thank you for your attention!

Prospects

- Improving experiments (by integrating operational constraints, splitting clusters to send several tariffs, testing with various data generators…)

- Integrating online hierarchical forecasting to personalized demand side management bandit algorithm