Periodic-GP: Learning Periodic World with Gaussian Process Bandits

Hengrui Cai[1], Zhihao Cen[2], Ling Leng[3] and Rui Song[1]

[1]North Carolina State University, [2]INRIA Saclay, [3]Coupang.

IJCAI 2021 @ RL4ITS Workshop

Madrid Traffic Pollution

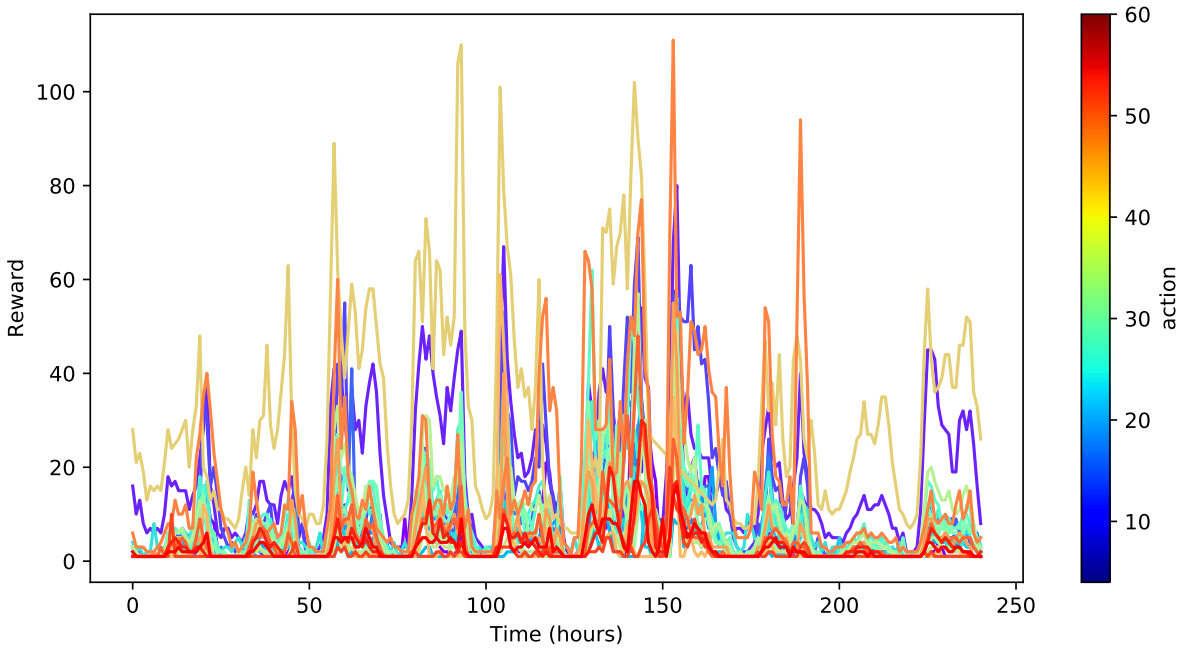

Figure 1: Madrid traffic condition of different sensors across the city over time, quantified by the nitric oxide level (NO, measured in $\mu g/m^3$), that is a highly corrosive gas produced by motor vehicles and fuel burning processes. Data source: Kaggle.

Question: How to identify locations of heaviest traffic over time, in case that the sensors are not available?

Modeling from the Madrid data

- By tracking NO level at different locations (i.e. actions) over time, a rapidly changing environment can be observed with strong seasonality (daily pattern in Madrid example).
- We encode the above environment that periodically repeats with *some non-stationary reward functions* as periodic stationary.

Seasonal Environment is General

In transport system

- Daily pattern in drivers (driver)/customer (demand) in ride-sharing;
- Weekday/weekend pattern traffic;
- Yearly pattern of airline traffic due to holiday / vacation season.

Motivation

Quite often, we have domain knowledge on *seasonality*. Our study aims to leverage it in bandit modeling. Could we do better than considering them as contextual variable?

Continuum Action Space

In transport system

The decision space could be location in the map (continuous space).

- Lime decides where to place scooter/bike, and how many.
- Ride-sharing / taxi company guides drivers to high demand locations in real time.

Solutions

- Discretization action space. Usually poor performance numerically and high complexity.
- *Gaussian Process (GP) - Upper Confidence Bound(UCB)* (Srinivas et al., 2009). GP is a Bayesian tool to approximate function over continuous space.

GP (Rasmussen, 2003)

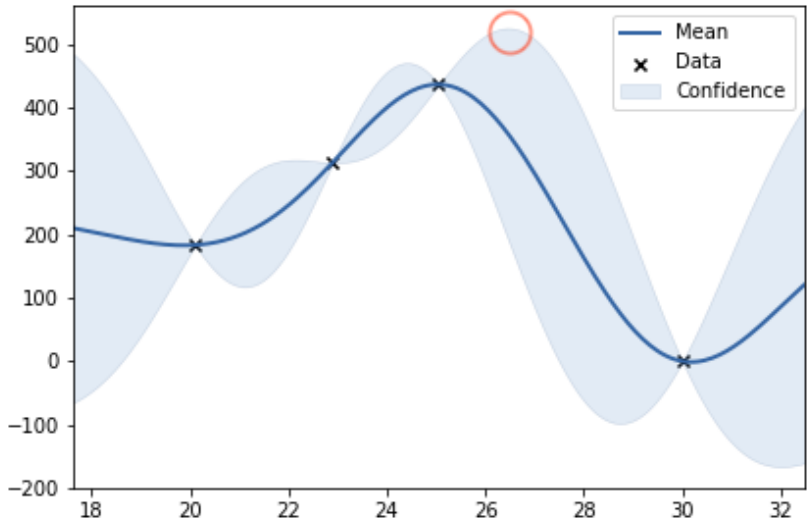

Figure 2: Model reward function by $\mathcal{GP}$.

GP Learning

Given prior $f \sim \mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}))$, after observing $(\boldsymbol{x}, y)_{\leq t}$, the posterior distribution of f is a $\mathcal{GP}$ that $f(x) \sim N(\boldsymbol{\mu}_t(x), \boldsymbol{\sigma}_t^2(x))$:

$$\begin{aligned} \boldsymbol{\mu}_t(\boldsymbol{x}) =& \boldsymbol{k}_{\leq t}(\boldsymbol{x})^\top (\boldsymbol{K}_{\leq t} + \sigma^2 \boldsymbol{I}_t)^{-1} \boldsymbol{y}_{\leq t} \\ \boldsymbol{\sigma}_t^2(\boldsymbol{x}) =& k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_{\leq t}(\boldsymbol{x})^\top (\boldsymbol{K}_{\leq t} + \sigma^2 \boldsymbol{I}_t)^{-1} \boldsymbol{k}_{\leq t}(\boldsymbol{x}), \end{aligned} \tag{1}$$

where $\boldsymbol{k}_{\leq t}(\boldsymbol{x}) = [k(\boldsymbol{x}_i, \boldsymbol{x})]_{i \leq t}$, and $\boldsymbol{K}_{\leq t} = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j \leq t}$.

GP-UCB (Srinivas et al., 2009)

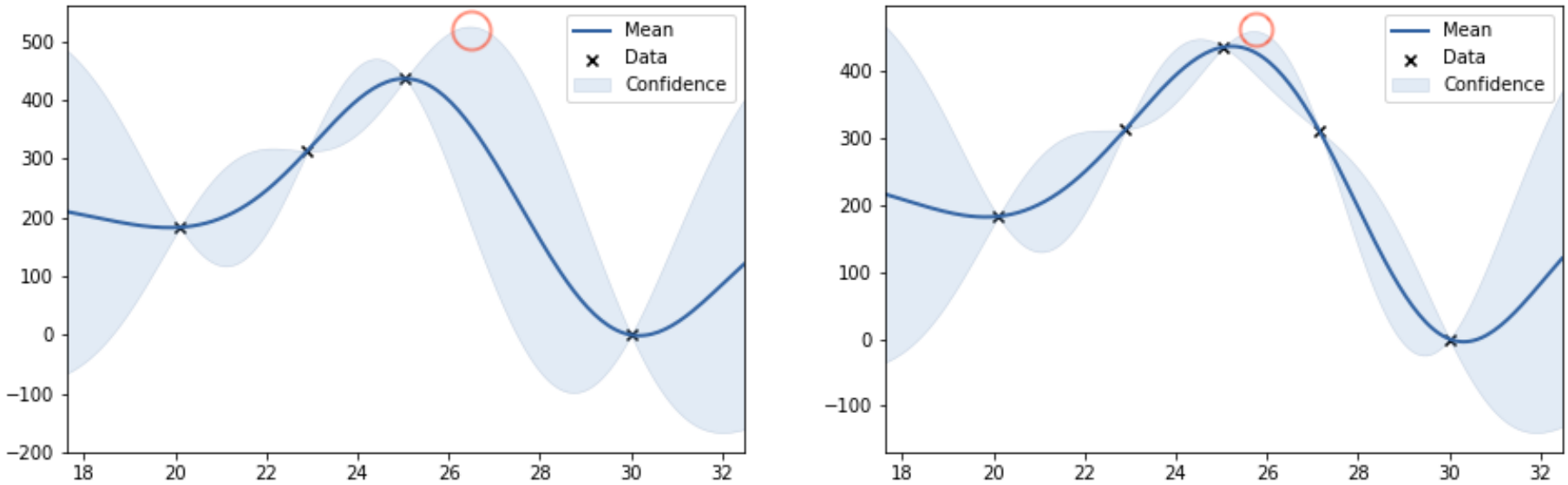

Figure 3: Left panel: choose the current best action that maximizes the upper confidence bound of reward function; Right panel: observe the reward and update the confidence bound.

Periodic Bandit Framework

- Consider an environment with reward function $f : \mathcal{X} = \mathcal{A} \times \mathcal{T} \to \mathbb{R}$ over an (potentially infinite) action space $\mathcal{A} \subset \mathbb{R}^d$ and time space $\mathcal{T} = \{1, 2, \cdots\}$.
- At each time step t, we choose an action $\boldsymbol{a}_t \in \mathcal{A}$ and receive an immediate reward y_t.
- **Periodicity assumption**: The reward function f has seasonal property of a fixed *known* period τ:

$$f(\boldsymbol{a}, t) = f(\boldsymbol{a}, t \bmod \tau) + \epsilon_t, \quad \forall \boldsymbol{a} \in \mathcal{A}, t = 1, 2, \cdots . \tag{2}$$

where $\|\epsilon_t\| \ll \|f(\cdot, t \bmod \tau)\|$.

Periodic Gaussian Process (Periodic-GP)

- Model the periodic environment by a periodic kernel over time:

$$k_{\mathcal{T}}(t, t') = \exp\left[-\frac{2}{l_{\mathcal{T}}^2}\sin^2\left(\frac{\pi|t-t'|}{\tau}\right)\right], \tag{3}$$

  where $l_{\mathcal{T}}$ is the length scale.
- With a kernel on the action space as $k_{\mathcal{A}}(\cdot,\cdot) : \mathcal{A} \to \mathbb{R}$, we define the kernel function over $\mathcal{X}$ as:

$$k(\boldsymbol{x}, \boldsymbol{x}') = k\{(\boldsymbol{a}, t), (\boldsymbol{a}', t')\} = k_{\mathcal{A}}(\boldsymbol{a}, \boldsymbol{a}') \times k_{\mathcal{T}}(t, t'). \tag{4}$$

- In our experiments, we apply the RBF Kernel on $k_{\mathcal{A}}$,

$$k_{\mathcal{A}}(\boldsymbol{a}, \boldsymbol{a}') = \exp\left(-\frac{\|\boldsymbol{a}-\boldsymbol{a}'\|^2}{2l_{\mathcal{A}}^2}\right).$$

Algorithm

Algorithm 1 Periodic GP-UCB

Require: a pre-specified τ;

Require: hyper-parameters in $\mathcal{GP}$ kernels $k(\boldsymbol{x}, \boldsymbol{x}')$;

1: **for** $t = 1, ..., T$ **do**

2: Update β_t following specific rule (detail in paper);

3: $\boldsymbol{a}_t \leftarrow \arg\max_{\boldsymbol{a} \in \mathcal{A}} \left[\boldsymbol{\mu}_{t-1}(\boldsymbol{a}, t) + \beta_t^{-\frac{1}{2}} \boldsymbol{\sigma}_{t-1}(\boldsymbol{a}, t) \right]$; ▷ UCB

4: Receive reward y_t;

5: Update $\mathcal{GP}$ posterior $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ based following (1).

Regret Bound of Periodic-GP-UCB

Theorem 1

Let $\delta \in (0, 1)$ and τ is a fixed constant. Under some assumption over $\mathcal{A}$ and β_t (see Paper for detail), we have the regret bound for Periodic-GP-UCB as $\mathcal{O}\left(\sqrt{T\beta_T\gamma_T^{\mathcal{S}}}\right)$ with probability at least $1 - \delta$. Or equivalently, we have:

$$Pr\left\{R_T \leq \sqrt{c_3 T\beta_T\gamma_T^{\mathcal{S}}} + \pi^2/6,\ \forall T \geq 1\right\} \geq 1 - \delta, \tag{5}$$

where $c_3 = 8/\log(1 + \sigma^{-2})$.

Experiments

Methods in Comparison

- **GP-UCB** Srinivas et al. (2009): stationary environment;
- **C-GP-UCB** Krause & Ong (2011): stationary environment with *contextual* information;
- **R-GP-UCB** Bogunovic et al. (2016): non-stationary environment using resetting techniques;
- **TV-GP-UCB** Bogunovic et al. (2016): non-stationary environment using decaying techniques;
- **Periodic-GP-UCB** (ours): proposed method for periodic scenario.

Synthetic Data

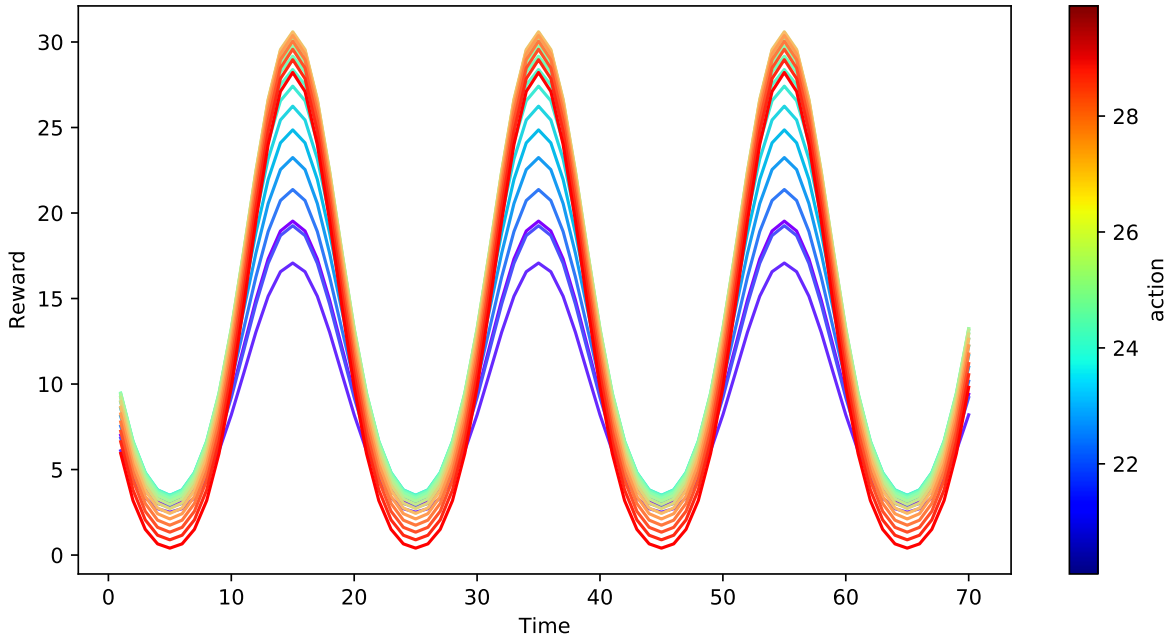

Figure 4: The reward function under different actions over time for synthetic data.

Result on Synthetic Data

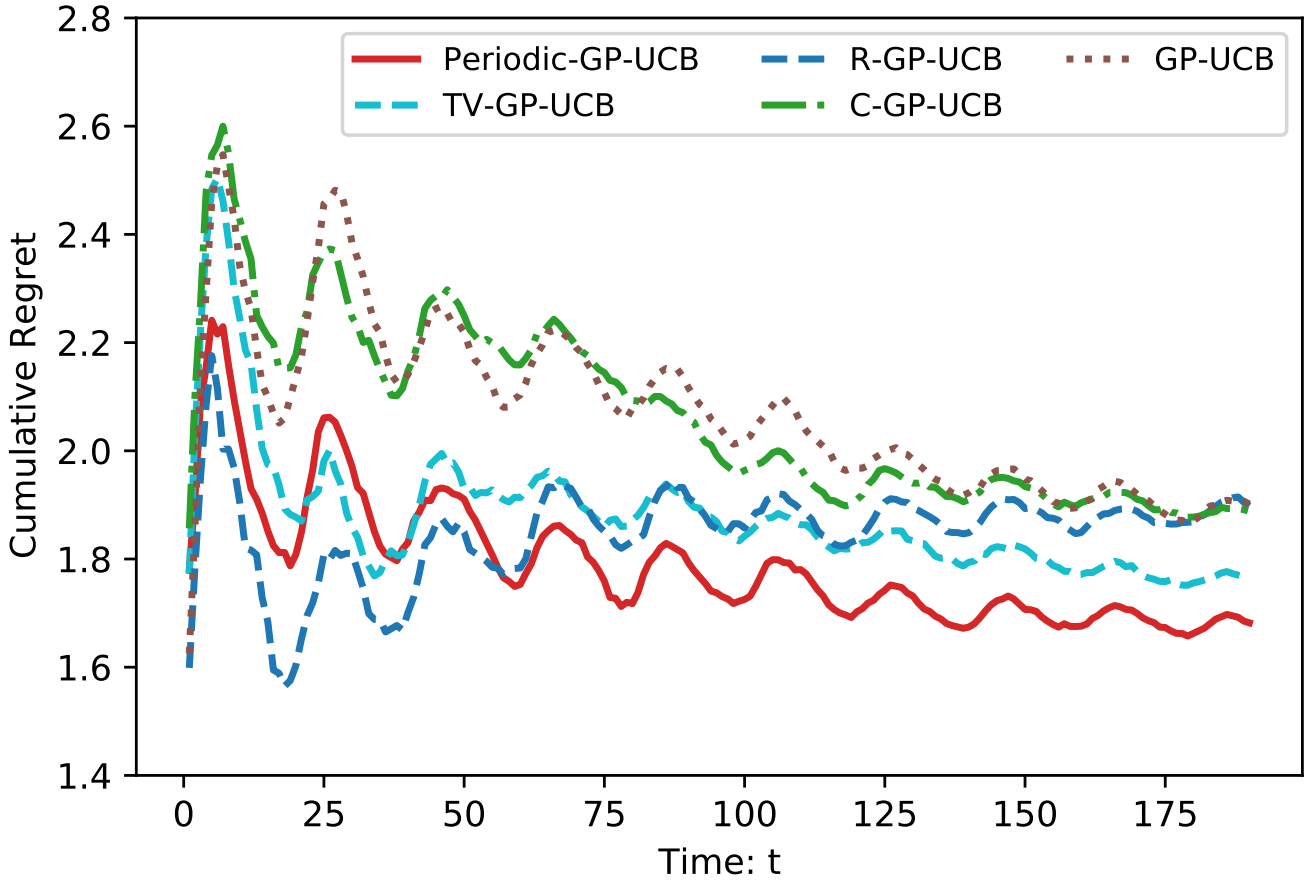

Figure 5: The mean cumulative regret over time under different methods for synthetic data.

Madrid Traffic Pollution

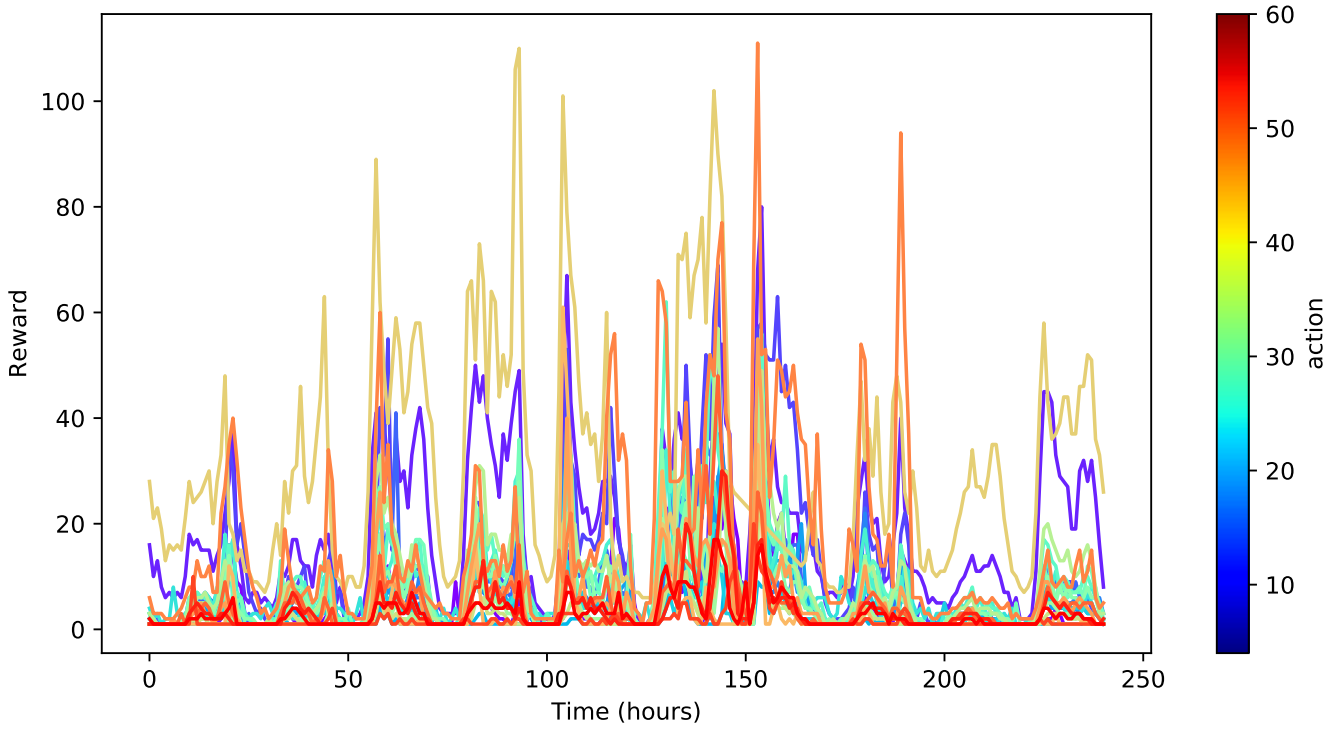

Figure 6: Madrid traffic pollution dataset. Data source: Kaggle.

Result on Madrid Data

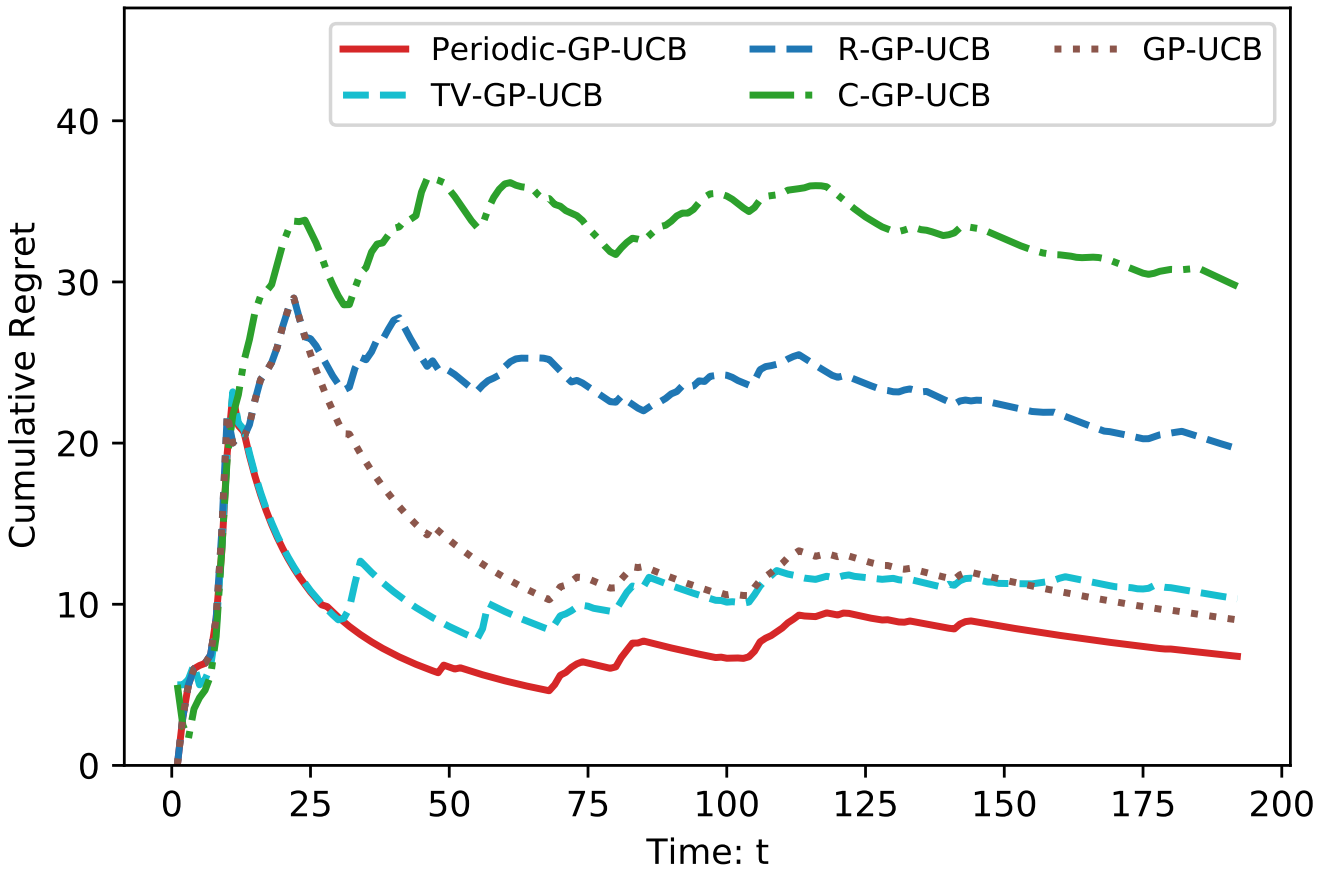

Figure 7: The cumulative regret under different methods for the Madrid traffic pollution data.

Thanks!

Q&A...

References I

Bogunovic, I., Scarlett, J., and Cevher, V. Time-varying gaussian process bandit optimization. In *Artificial Intelligence and Statistics*, pp. 314–323, 2016.

Krause, A. and Ong, C. S. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*, pp. 2447–2455, 2011.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer, 2003.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.