A Gaussian process approach for contextual bandit-based battery replacement

Tianshi Zhou*

New York University Shanghai, New York University, 200124, Shanghai, China

Abstract. The sharing economy has recently become a distinctive business model in China, offering advantages such as low prices and ease of use. Shared electric vehicles have also become an essential part of urban transportation. This research aims to assist electric vehicle providers in optimizing battery replacement schedules, with the objective of minimizing operating losses while meeting evening peak demand. Efficient resource allocation is crucial to remain competitive with the established public transport system. This study proposes a multi-armed bandit (MAB) approach to identify optimal periods for inspection and battery replacement in new city launches, even without prior knowledge of user behavior patterns. Modifications are made to the traditional MAB algorithm, incorporating the lower confidence bound (LCB), contextual features, kernelization, and the Gaussian Process (GP) to enhance the Upper Confidence Bound (UCB) MAB in solving this problem. Unlike deep learning techniques, the MAB model offers a lightweight, efficient, and easy-to-deploy solution that adapts to dynamic scenarios even with limited training data. Results indicate that this method performs stably in cumulative regret and in selecting the optimal choice within a short timeframe. Adaptable to seasonal and weekend fluctuations, this optimized approach shows potential for enhancing operational strategies not only in shared transportation but also across other sectors of the sharing economy.

1 Introduction

1.1 Background

As the sharing economy develops, shared bicycles have become an essential mode of transportation in urban areas. In recent years, electronic bicycle sharing has become a new trend. Not only Chinese companies like Hellobike and Meituan deploy electronic bikes, but some Western companies like Uber have also started to provide electronic scooters. However, battery life and operating costs have been essential challenges in shared electric vehicle services. There are several ways to supply electricity nowadays, such as charging stations and changing cabinets. However, for small cities in China or cities that will soon need to introduce shared electric battery vehicles, this site-based battery change method has the

* Corresponding author: tz2321@nyu.edu

challenge of construction, time, and maintenance costs. Hence, for companies such as Hellobike, the primary method remains to deploy a team to change the batteries on-site. Since the evening peak is also the demand peak, suppliers will carry out power replacement and inspection before the evening peak to ensure supply. This method also has weaknesses; approximately 15 percent of users are affected by battery replacement. Because of this, the goal of this research is to find the best period for the battery change team. This could be challenging wildly when the demand for users fluctuates primarily due to the weather, weekends, and seasons.

1.2 Related work

Numerous studies have demonstrated GP-MAB's ability to perform well in some areas. For example, under non-stationary conditions, GP-MAB has stable performance, even if the reward function is nonlinear [1]. Different kernels can be used to deal with different problems, such as the periodic kernel [2]. Secondly, this model can make efficient selections under soft constraints [3]. These features of GP-MAB fit this paper's scenario where the reward depends on a lot of factors. It could overcome weather and seasonal variations and significant usage differences between weekends and weekdays.

GP-MAB has already been applied to many fields. The media industry uses it to personalize advertisement recommendations, which boosts the Click-Through rate (CTR) [4]. The medical industry also uses it to decide on personal treatment plans, for example, Diabetes Management [5]. Çelik et al. propose a GP-UCB variant with volatile arms that takes into account the patient's condition as well as acceptable treatments when recommending new therapies. The practice in many fields reflects the strong universality of this model.

Research in the transportation field related to this study also uses GP-MAB. The daily periodicity of traffic conditions and the regularity of the seasons allow GP-MAB to use context to make better decisions. Cai et al. proposed Periodic-GP, which focuses on learning the stochastic periodic world by leveraging this seasonal law. As mentioned above, the periodic kernel is a strong tool in this situation [2].

The present study on battery replacements focuses more on optimization than strengthening the ability to operate in different environments. Zhou et al. use a Markov chain-based model to optimize battery exchange and rebalancing, emphasizing overall operational efficiency. Zhou et al. 's integrated model optimizes battery replacement and vehicle rebalancing based on historical data but is better suited for long-term planning. In contrast, the MAB model in this study performed well in real-time optimization, especially during periods of surging or uncertain demand [6]. This study provides a lightweight, flexible solution tailored to a specific time window using a MAB approach with GP and UCB. This can quickly adapt to changing requirements, making it ideal for new systems or data-sparse environments.

1.3 Motivation and framework

First of all, the static method of battery replacement cannot adapt to the environment change. Secondly, electronic bikes are a new traffic choice, but not much information is provided. That means the model needs a balance between exploration and exploitation, which is a perfect scenario for MAB. There are also some challenges, the reward distribution is dynamic and periodic. The problem also focuses more on loss than reward.

This study aims to construct a GP contextual MAB model to overcome these difficulties. The main contribution is first to use the GP to model the battery demand and citizen's traffic style, making it contextual. Secondly, it designs a dynamic scheduling strategy based on the LCB model to balance short-term gains and long-term exploration. Lastly, the simulations

are used to test the performance of different kernels and compare the GP-MAB model to other methods.

2 Methodology

2.1 Overview

The flowchart of the experimental structure is shown in Figure 1. All of the models in this study are MAB models. MAB is often used in situations where decisions are made with incomplete information. The goal is to maximize the cumulative reward in the long term. MAB can balance exploration and exploitation without prior knowledge. The basic form of MAB is to treat different schemes as arms, and then gradually explore the reward distribution of each arm. Every selection would update the distribution, which can help the model make the next decision. The model evaluation will be based on cumulative regret. MAB models are mostly used in recommendation systems, advertising, pricing strategy selection, etc. [7].

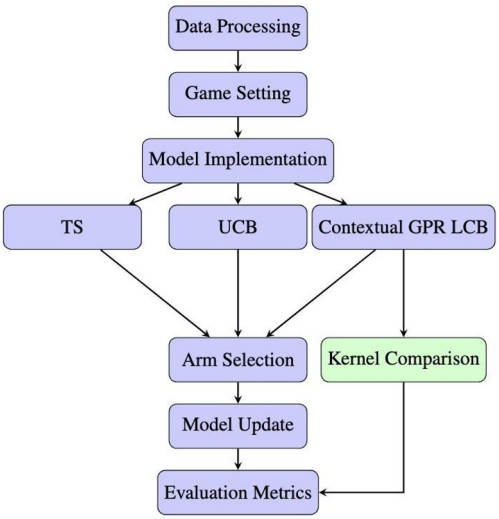

Fig. 1. Flowchart of the experiment with model and kernel comparisons

2.2 Data processing

2.2.1 Data import and conversion

The Citi bike trip dataset from 2013 to 2017 is used. This dataset contains the date, start time, end time, duration, and site information of each trip. The date, month, and hour information are extracted for the time period analysis.

2.2.2 Data cleaning

Duplicate records were identified by matching identical station id, start time of trips and end time of trips, then they are removed. Outliers in trip durations were detected using the interquartile range (IQR) method, where trip durations exceeding 1.5 times the IQR were excluded. Missing values in trip duration were imputed using the median duration of trips at the respective station to preserve data consistency.

2.2.3 Data screening

The four hours from 2 p.m. to 6 p.m. before the evening peak were selected. They are divided into four equal periods for data classification.

2.2.4 Data conversion and encoding

Calculate revenue using duration as the basis of the current duration pricing model and identify whether it is a weekend or not based on the date. The class feature 0,1 is given on weekends. The month feature retains its cyclic characteristics through polar coordinates.

2.2.5 Sampling

The study samples according to the month to simulate the seasons of a year, that is, 30 rounds are one month, and each round randomly samples a row of data in the month of any year. Twelve months is a cycle.

2.2.6 Comparison

The final dataset has been significantly condensed, reducing the data volume from 650,000 records to a few thousand rows. This reduction was achieved through careful aggregation and cleaning processes. The resulting dataset is compact, focusing on rewards-related value such as the revenue of a specific hour, and total revenue. It also includes feature variables (month, weekend), making it highly efficient for downstream modeling tasks such as MAB algorithms according to Figure 2and Figure 3.

1	date	arm_0_star	arm_1_star	arm_2_star	arm_3_star	total_start_	arm_0_tota	arm_1_tota	arm_2_tota	arm_3_tota	total_reven	month	weekend
2	2017/1/28	68	57	46	36	207	334	267.5	144	150	895.5	1	1
3	2014/1/26	36	29	18	26	109	164	118.5	55	111	448.5	1	1
4	2016/1/9	102	108	36	30	276	277	406	74	69	826	1	1
5	2017/1/25	21	12	18	27	78	71.5	33	36	52.5	193	1	0
6	2015/1/25	94	71	85	50	300	366	193.5	282.5	205	1047	1	1
7	2014/1/26	36	29	18	26	109	164	118.5	55	111	448.5	1	1
8	2014/1/15	1	9	5	10	25	1.5	28.5	21.5	20	71.5	1	0
9	2015/1/4	17	22	18	7	64	40.5	65	44	16.5	166	1	1

Fig. 2. Table of the Raw Data

1	arm_0_tota	arm_1_tota	arm_2_tota	arm_3_tota	total_reven	month	weekend
2	57.5	86	69.5	110	323	1	0
3	14.5	10.5	21	7.5	53.5	1	0
4	10.5	9.5	14.5	27.5	62	1	0
5	271	312.5	343	198	1124.5	1	0
6	21	41	64	35.5	161.5	1	0
7	34	53.5	65	107	259.5	1	0

Fig. 3. Table of the Processed Data

2.3 Experimental setup and notation

Here are the game settings and notations:

- Arms are the four periods before the evening peak.
- The instant optimal arm $\hat{a}_t$ is the arm that has the lowest loss at round t.

- The chosen arm at round t is a_t^*.
- The overall optimal arm a_0. It is defined as the arm that has the lowest mean loss.
- Every round, the game sets 15 percent of users to be denied by the system so that the model will get a loss, which is the sum of their revenue.
- Regret is defined as the loss difference between a_t^* and a_0.
- Optimal Frequency is defined as the number of times that the model selects the instant optimal arms divided by the total round.

2.4 Baseline model

2.4.1 Thompson sampling

Thompson Sampling (TS) is based on Bayesian probability. Ts samples the reward distribution for each arm and selects the arm with the highest sampling value. The specific steps are as follows:

- Define Prior Distributions: For each arm a, establish a prior distribution, such as a Beta distribution Beta(α_a, β_a).
- Sampling: Draw a sample value θ_a from the posterior distribution of each arm a.
- Select Arm: Choose the arm a^* with the highest sampled value θ_a.
- Observe and Update: Pull arm a^* and get the reward r_{a^*}. Then it updates the posterior distribution of the arm a^* based on the observed reward.

2.4.2 UCB model

Its core idea is to balance Exploration and Exploitation by calculating the UCB of each arm. Specifically, the model selects the arm with the highest UCB value at each round, as the equation (1):

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\ln t}{N_a(t)}} \tag{1}$$

- $\hat{\mu}_a(t)$ is the average reward estimation of arm a at timestep t.
- $N_a(t)$ is the count of selected times of arm a at timestep t.
- $\sqrt{\frac{2\ln t}{N_a(t)}}$ is the width of UCB

The UCB model achieves the goal by selecting the arm with the highest UCB. Here, the study modified it to LCB, whose details are shown below.

2.5 Contextual_GPR_LCB model

This study proposes the Contextual_GPR_LCB model to address the Contextual Multi-Armed Bandit problem. GPR is used to model the relationship between contextual features and expected losses for each arm. Since the main task here is to minimize the loss, the lower bound is used instead of the upper bound.

2.5.1 GPR basics and principle

GPR is a non-parametric, Bayesian approach to regression. It operates by defining a distribution over functions directly rather than fitting specific model parameters. Given a set of input features or contexts $\mathbf{C}$ and corresponding target values or losses $\mathbf{y}$, GPR models the relationship as a joint Gaussian distribution. The goal of GPR is to predict the expected value and uncertainty (variance) for new contexts based on previously observed data, making it particularly suitable for problems involving uncertainty and exploration, such as multi-armed bandits.

The key advantage of GPR lies in its flexibility in modeling non-linear relationships through the choice of kernel function. It provides a posterior distribution over possible functions, enabling predictions with not only a mean estimate but also a measure of uncertainty, denoted as $\mu_a(\mathbf{C})$ and $\sigma_a(\mathbf{C})$, respectively [8].

2.5.2 Model architecture

For every arm a, the Contextual_GPR_LCB builds an independent GPR, allowing for tailored modeling based on specific contextual information so that this model could be easily optimized. The definition of the GPR is given in equations (2) and (3) [9,10]:

$$f_a(\mathbf{X}_a) = y_a | X_a \sim \mathcal{GP}\big(m_a(\mathbf{X}_a), k_a(\mathbf{X}_a, \mathbf{X}_a{}')\big) \tag{2}$$

Where $m_a(\mathbf{X}_a)$ is the mean function, typically set to zero, and $k_a(\mathbf{X}_a, \mathbf{X}_a{}')$ is the kernel function, composed of a Constant Kernel C_a and a Radial Basis Function (RBF) Kernel [11]:

$$k_a(\mathbf{X}_a, \mathbf{X}_a{}') = C_a \cdot \exp\left(-\frac{\| \mathbf{X}_a - \mathbf{X}_a{}' \|^2}{2l_a^2}\right) \tag{3}$$

Here, C_a represents the variance parameter, and l_a represents the length scale of the RBF kernel.X_a is the history context matrix of X_a and X'_a is the latest context. They all belong to arm a.

The choice of kernel function is crucial in determining the performance of the GPR model. The study employs the RBF kernel, which is widely used due to its ability to capture non-linear relationships between contexts. Different kernel functions can be chosen here to adapt different features such as periodic kernels, or combined kernels.

The periodic kernel is a kernel function used in GPR to model periodic or repeating patterns in the data. It extends the RBF kernel by incorporating periodicity into the covariance structure, making it suitable for modeling phenomena that repeat over regular intervals, such as daily or seasonal trends. The study also implements it. The periodic kernel is defined in equation (4):

$$k(\mathbf{C}_i, \mathbf{C}_j) = \exp\left[-\frac{2\sin^2\left(\frac{\pi|\mathbf{C}_i - \mathbf{C}_j|}{p}\right)}{l^2}\right] \tag{4}$$

where: $\mathbf{C}_i$ and $\mathbf{C}_j$ are two input context vectors, l is the length scale hyperparameter that controls the smoothness of the function, p is the period of the function. p determines the interval at which the function repeats. $\sin^2\left(\frac{\pi|\mathbf{C}_i - \mathbf{C}_j|}{p}\right)$ captures the periodicity of the kernel by

measuring the sine of the distance between the input points, scaled by the period p[2]. Libraries like scikit-learn provide built-in mechanisms to optimize l during model training.

The periodic kernel ensures that the covariance between two points is a periodic function of their distance, making it particularly useful for tasks involving cyclical data, such as daily temperature patterns, seasonal trends in sales, or repeating user behavior. By adjusting the period p, the kernel can be tailored to specific periodic behaviors in the data.

2.5.3 Arm selection

At each round t, the context $\mathbf{C}_t$ would be observed, and here is the process of GPR for each arm.

- Prediction: the GPR predicts the expected loss $\mu_a(\mathbf{C}_t)$ and the uncertainty $\sigma_a(\mathbf{C}_t)$.
- The model would Compute the LCB by the formula given in equation (5).

$$LCB_a(t) = \mu_a(\mathbf{C}_t) - \beta \cdot \sigma_a(\mathbf{C}_t) \tag{5}$$

  Where β is a hyperparameter. It can be modified to encourage the model to explore more or not. Here are some common values for β. Setting $\beta = 1$ can have a moderate exploration and setting $\beta = 2,3$ can make a more agressive exploration.
- Select Arm: The Model chooses the arm with the smallest LCB given by equation (6).

$$a^* = \underset{a}{\operatorname{argmin}} LCB_a(t) \tag{6}$$

2.5.4 Model update mechanism

After selecting arm a^*, models will get a corresponding loss y_{a^*}, the model updates the GPR of a^*.The updating process is given by equation (7) and (8)

$$\mathbf{X}_{a^*} \leftarrow \mathbf{X}_{a^*} \cup \mathbf{C}_t \tag{7}$$

$$\mathbf{y}_{a^*} \leftarrow \mathbf{y}_{a^*} \cup y_{a^*} \tag{8}$$

The context dataset $\mathbf{X}_{a^*}$ and the loss dataset $\mathbf{y}_{a^*}$ are updated to retrain the GPR. The model sets a limitation of the max train size since the complexity of the GP is $O(n^3)$.

2.6 Evaluation metrics

The study employs two primary metrics to evaluate the performance of the multi-armed bandit algorithm:

Cumulative Regret: This metric is defined as the total difference between the rewards obtained by the overall optimal arm and those obtained by the algorithm over all time steps. It measures how well the algorithm minimizes the loss relative to the best possible strategy. Lower cumulative regret signifies more efficient learning and decision-making.

Frequency of Selecting the Optimal Arm: This metric calculates the proportion of times the algorithm selects the best-performing arm throughout the decision process. It reflects the algorithm's ability to identify correctly and consistently prioritize the most rewarding arm, indicating its effectiveness in optimizing resource allocation.

3 Results and analysis

According to the evaluation, three sets of comparisons are tested. Firstly, the comparison between the baseline model and the GP model is presented in Figure 4 and Figure 5.

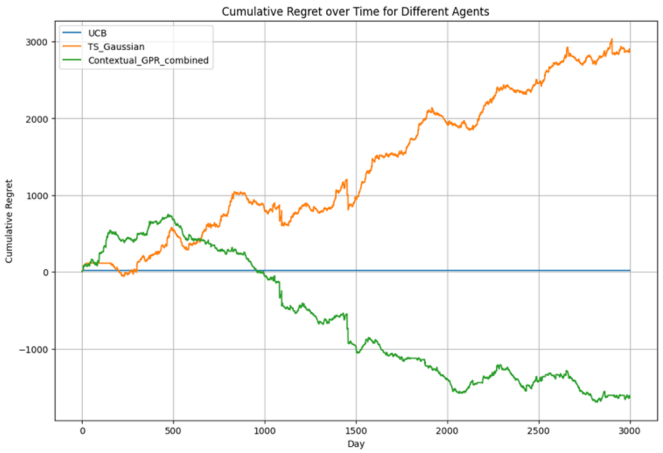

Fig. 4. Cumulative regret comparison (Baseline VS GP)

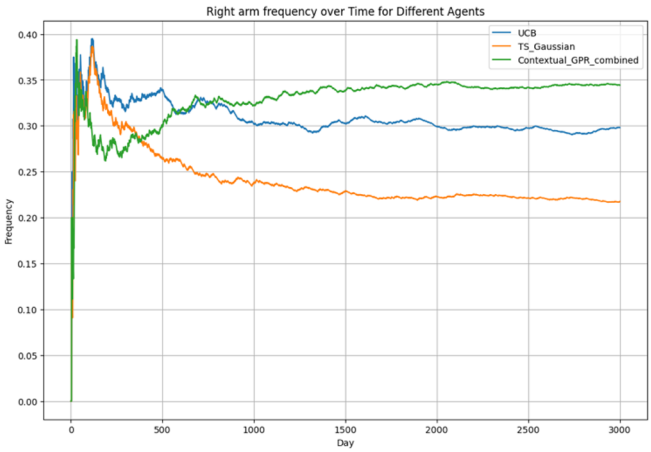

Fig. 5. Frequency comparison (Baseline VS GP)

The results indicate that the contextual GP model performs significantly better than the baseline model after a short-term exploration, in terms of both cumulative regret and frequency.

Secondly, comparisons are made between different kernels. The plots for the periodic kernel and the RBF kernel are shown in Figure 6 and Figure 7. From the plot, the periodic kernel does not perform well on cumulative regret or frequency. It also takes four times longer to run.

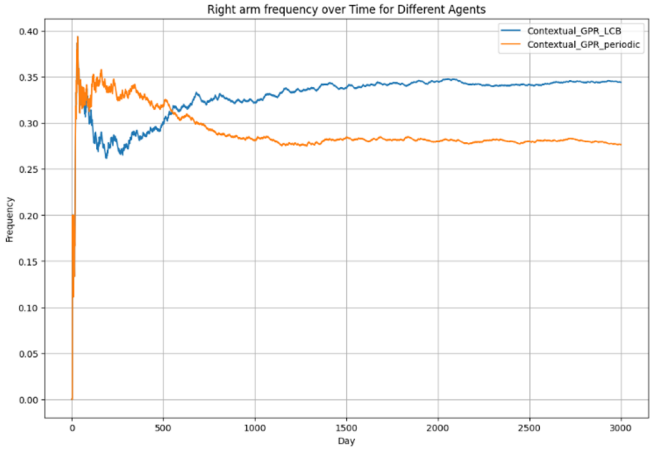

Fig. 6. Frequency comparison (RBF VS Periodic)

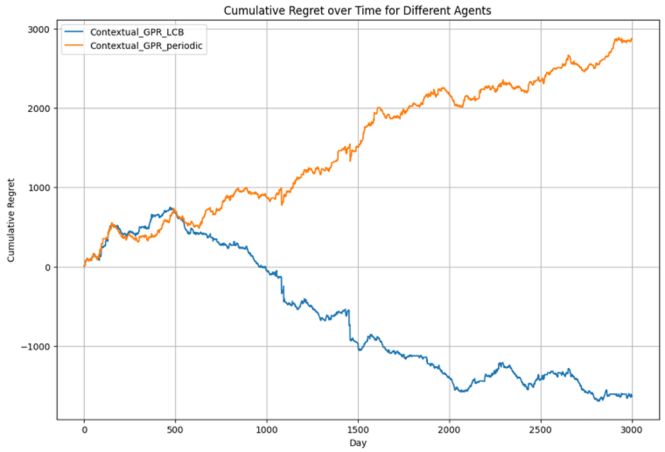

Fig. 7. Cumulative regret comparison (RBF VS Periodic)

To test how will the model perform in the extreme condition, this study doubled the revenue of some months, so that the user behaviour fluctuates greatly. At the same time, the feature of every month exaggerates. In this case, model with GPR still take the lead with cumulative regret which is less than half of the baseline model's. Due to the more obvious month features, periodic kernel makes a better performance than the RBF kernel. The flat line shows that GPR model is very robust and can adapt to fluctuating data and environment. The cumulative regret at a lower level showed that its ability to make relatively correct decisions under extreme circumstances. The results are showed in Figure 8 and Figure 9.

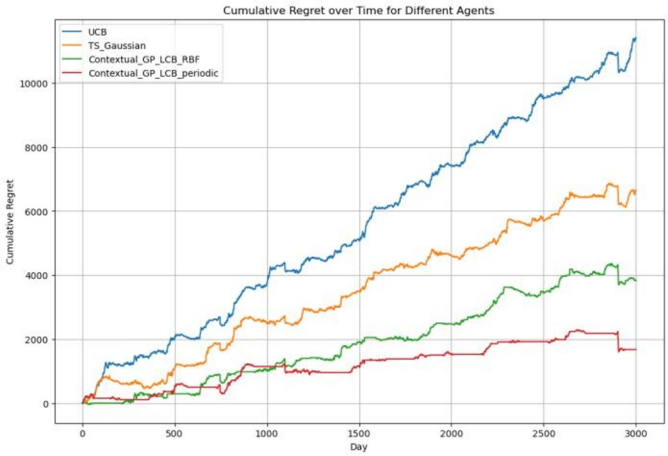

Fig. 8. Cumulative regret comparison in extreme conditions

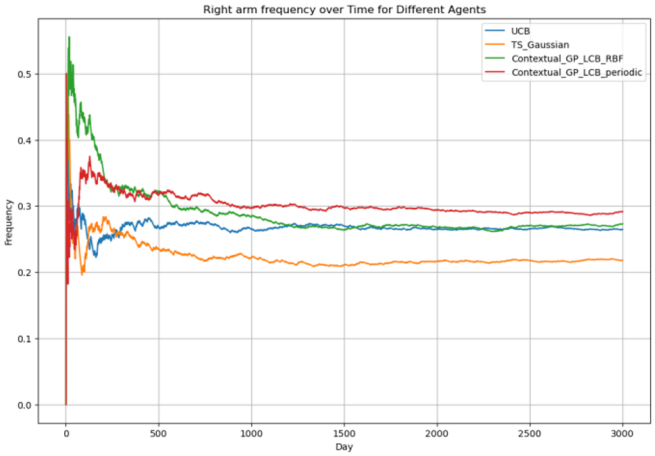

Fig. 9. Frequency comparison in extreme conditions

In sum, every model except the Contextual_GPR_LCB model cannot do better than always choosing the overall optimal arm. The model built in this study also has a higher frequency than any other model, which means it dynamically chooses the instant optimal arm efficiently. After a short exploration of about a year, it began to outperform other models, choosing more right instant optimal than others. This proves that the Contextual_GPR_LCB model perfectly balances exploration and exploitation. It also does great in unstable environments with high adaptability and optimization ability.

4 Limitations and future outlooks

4.1 Limitations

Firstly, the data sampling presents certain limitations. The sampling granularity is carried out by month, and it is not possible to smoothly show the climate change. There are also many features that can be observed but cannot be collected in this study, such as weather and holidays. This study was only able to take into account seasons and weekends. Secondly, for model building, the bound given to the kernel is not fit enough. In the process of study,

different experiments will have different bound ranges for the kernel. Sklearn keeps suggesting that a smaller bound value can be selected, but a smaller bound will take much longer to run the experiment. Sometimes, the best value is infinitely close to zero, so it is hard to tune. At the same time, the study did not set the number of GPR iterations, which may make the model trapped in the critical point while it is not at the extreme value. More iteration times can help the model get out of this situation. A study may find the best hyperparameters if better hardware support is available.

4.2 Outlooks

In terms of features, future research can integrate higher dimensional data. Now, the feature dimension is too small, so the RBF kernel is smoother and better than others. When many features are considered at the same time, processing different features with multi-kernels combined can have a better performance. In terms of model building, as said in the limitations, increasing iteration when computing power allows can help the model escape from non-optimal solutions. There are also different feature processing methods to be tried, such as combining deep learning to process features to improve the adaptability and stability of decision-making. Last but not least, carrying out multi-objective optimization based on this research, adding user satisfaction, and so on to the optimization goal could improve the reward function and realize the balance between economic benefits and user feedback.

5 Conclusion

The results show that the context-based Contextual_GPR_LCB model can effectively identify and prioritize the low loss time period. This optimizes the battery replacement in cities that cannot maintain charging stations and in cities that are just beginning to deploy shared electric vehicles. The growth rate of cumulative regret slowed down significantly, even going to negative. The decision is much better than choosing the overall optimal arm for the whole time. The frequency also converges to a higher value than other models, indicating that the model gradually reached a better balance strategy. The gradual rise of the optimal arm selection frequency verifies the adaptability and accuracy of the algorithm in different contexts. Comprehensive evaluation shows the algorithm has good learning ability and resource optimization effect. This study addresses the emerging issue of battery replacement. Different from the research direction of how to set up the battery changing station and how to allocate and dispatch, this study focuses on the operation mode of most non-developed cities and gives the scheme. It proposes a new idea for the dynamic decision-making of battery replacement problems and expands the practical application of the multi-armed bandit's algorithm in emerging scenarios.

References

1. Y. Deng, X. Zhou, B. Kim, A. Tewari, A. Gupta, N. Shroff, Weighted Gaussian process bandits for non-stationary environments. Proc. Mach. Learn. Res. 151, 6909-6932 (2022).
2. H. Cai, Z. Cen, L. Leng, R. Song, Periodic-GP: learning periodic world with Gaussian process bandits. CoRR, abs/2105.14422 (2021).
3. X. Zhou, B. Ji, On kernelized multi-armed bandits with constraints. Adv. Neural Inf. Process. Syst. 35, 14-26 (2022).

4. S. Gowri, A. Srikanth, G. Gowthaam, G. M, G. D T, L. C R, H. Lavaniya, Dynamic personalized ads recommendation system using contextual bandits. In 2023 Int. Conf. Intell. Syst. Commun. IoT Security (ICISCoIS), 339-344 (2023).
5. A. Çelik, Diabetes management via Gaussian process bandits. Ph.D. thesis, ProQuest (2021).
6. Y. Zhou, Z. Lin, R. Guan, J.-B. Sheu, Dynamic battery swapping and rebalancing strategies for e-bike sharing systems. Transp. Res. Part B: Methodol., 177, 102820 (2023).
7. N. Silva, H. Werneck, T. Silva, A. C. M. Pereira, L. Rocha, Multi-armed bandits in recommendation systems: a survey of the state-of-the-art and future directions. Expert Syst. Appl. 197, 116669 (2022).
8. H. Zenati, A. Bietti, E. Diemert, J. Mairal, M. Martin, P. Gaillard, Efficient kernelized UCB for contextual bandits. Proc. Mach. Learn. Res. 151, 5689-5720 (2022).
9. V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, Gaussian process regression for materials and molecules. Chem. Rev. 121, 10073-10141 (2021).
10. S. Vakili, N. Bouziani, S. Jalali, A. Bernacchia, D. Shiu, Optimal order simple regret for Gaussian process bandits. Adv. Neural Inf. Process. Syst. 34, 21202-21215 (2021).
11. Y. Miyake, R. Watanabe, T. Mine, Online nonstationary and nonlinear bandits with recursive weighted Gaussian process. In 2024 IEEE 48th Ann. Comput. Softw. Appl. Conf. (COMPSAC), 11-20 (2024).