Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

How to gamble with non-stationary $\mathcal{X}$-armed bandits and have no regrets

Valeriy Avanesov

submitted: February 6, 2020

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: valeriy.avanesov@wias-berlin.de

No. 2686
Berlin 2020

2010 *Mathematics Subject Classification.* 62M10, 62H15.

Key words and phrases. $\mathcal{X}$-armed bandits, kernel methods, non-stationary bandits, change-point detection, regret.

The research of “Project Approximative Bayesian inference and model selection for stochastic differential equations (SDEs)” has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 “Data Assimilation”, “Project Approximative Bayesian inference and model selection for stochastic differential equations (SDEs)”. Further, we would like to thank Vladimir Spokoiny, Alexandra Carpentier and Manfred Opper for the discussions which have greatly improved the manuscript.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

How to gamble with non-stationary $\mathcal{X}$-armed bandits and have no regrets

Valeriy Avanesov

Abstract

In $\mathcal{X}$-armed bandit problem an agent sequentially interacts with environment which yields a reward based on the vector input the agent provides. The agent's goal is to maximise the sum of these rewards across some number of time steps. The problem and its variations have been a subject of numerous studies, suggesting sub-linear and sometimes optimal strategies. The given paper introduces a new variation of the problem. We consider an environment, which can abruptly change its behaviour an unknown number of times. To that end we propose a novel strategy and prove it attains sub-linear cumulative regret. Moreover, the obtained regret bound matches the best known bound for GP-UCB for a stationary case, and approaches the minimax lower bound in case of highly smooth relation between an action and the corresponding reward. The theoretical result is supported by experimental study.

1 Introduction

Numerous studies consider variations of an $\mathcal{X}$-armed bandit problem, a problem where an agent sequentially interacts with the environment, supplying a vector X_t at each time step t and receiving a reward y_t, depending (presumably) on X_t. The reward is immediately made known to the agent, so the subsequent actions can be based on it. Typically, the reward is obfuscated by some noise.

This model finds its use in applications such as clinical trials, pricing, finance, logistics, advertisement and recommendation [6, 23, 5, 4, 3, 15, 22] and has (along with stochastic optimization in general) unsurprisingly attracted immense attention in the recent decades. Initially, though, the research was focused on *multi-armed bandits*, where only the actions X_t from a finite set are feasible [13]. Later researchers also turned to consideration of $\mathcal{X}$-armed bandits (with continuous feasible set $\mathcal{X}$) [18, 7].

Currently due to the ever-changing nature of our world, the studies on multi-armed bandits are increasingly more concerned with the environments changing their behavior over the course of time [14, 8, 19]. For instance, an active line of work considers *restless* bandits, referring to a class of problems where the environment switches between the internal states according to a known stochastic law [20]. Also, there are settings where no such knowledge is available, like in [6], presuming a bound on the total variation of the mean reward associated with each arm, or in [10, 2] implying no such bound, yet presuming these switches to be relatively rare. The latter is the setting we extend to the case of $\mathcal{X}$-armed bandits. Particularly, we let the underlying relationship between the action X_t and the reward y_t to abruptly change. The agent has no knowledge of when to expect such a change, neither any information on the nature of the change is revealed. As usual, in a bandit setting the goal is to minimize the *cumulative regret* – a discrepancy between the received reward and the largest possible one.

Formally, consider a number of stationary periods $K \in \mathbb{N}$ and a sequence of functions $\mathcal{F} \coloneqq \{f_1, f_2, ..., f_K\}$ mapping from the convex compact $\mathcal{X}$ to $\mathbb{R}$. Further, we introduce a sequence of

DOI 10.20347/WIAS.PREPRINT.2686 Berlin 2020

time points (being called *change-points*) $0 = \tau_0 < \tau_1 < \tau_2 < ... < \tau_K = T$, when the environment switches between the functions f_i, and denote the lengths of the stationary intervals as $T_i := \tau_i - \tau_{i-1}$. Also consider a piecewise constant map $\varkappa : \mathbb{N} \to \{1, 2, ..., K\}$, such that $\varkappa(1) = 1$, $\varkappa(T) = K$, $\varkappa(i+1) = \varkappa(i) + 1$ if $\exists j \in \mathbb{N} : i = \tau_j$ and $\varkappa(i+1) = \varkappa(i)$ otherwise.

At time points $t = \{1, 2, ..., T\}$ the agent consequently interacts with the environment, providing a vector $X_t \in \mathcal{X}$ and receives a reward

$$y_t = f_{\varkappa(t)}(X_t) + \varepsilon_t,$$

where ε_t denotes i.i.d. centered noise.

Neither the functions f_i nor the change-points τ_k nor the number K is known to the agent. The goal is to minimize the *cumulative regret* $\mathbb{E}[R_T]$, where

$$R_T := \sum_{t=1}^{T} \max_{x \in \mathcal{X}} \left[f_{\varkappa(t)}(x) \right] - f_{\varkappa(t)}(X_t).$$

Here we emphasize the crucial difference between the problem we consider and the adversarial setting. We assess the performance of the agent in comparison to an oracle, choosing an optimal strategy for each stationary period, while the literature on adversarial environments competes with oracles whose choice of input remains the same during all T steps, or allows for some number of switches, contributing linearly to the cumulative regret [1].

The algorithm is said to be *no-regret*, if $\lim_{T \to +\infty} \mathbb{E}[R_T]/T = 0$, which is exactly what we achieve. Specifically, our contribution is outlined as follows

- We propose a novel approach for a non-stationary $\mathcal{X}$-armed bandit problem.
- We establish a sub-linear bound on the cumulative regret under mild assumptions.
- The upper bound matches such for GP-UCB (see [11, 18]) for the stationary case, which implies the transition to the non-stationary setting came free of (asymptotic) charge.
- The upper bound approaches the minimax lower bound in case of highly smooth functions f_i up to slowly growing factors, implying near-optimality of our approach.
- The theoretical findings are verified empirically. A comparative study has also been conducted.
- Our approach is adaptive to the number of change-points (unlike in [10], where K is deemed known). This adaptiveness is not achieved via keeping track of the number of number of change-points detected (as in [2]), yielding consistent behavior across all stationary intervals.
- The algorithm is not adaptive to the horizon T, nevertheless, it is robust to misspecification of this parameter. Namely, the approach is driven by a small power of T, being more tolerant to an incorrect choice of horizon, than the approaches suggested in [10, 2], relying on the specification of $\sqrt{T}$.
- We propose an algorithm, generally suited for detection of a change-point in regression, not only in a bandit setting.

The paper is organized as follows. Section 2 introduces the suggested approach along with the necessary background and is followed by a rigorous theoretical study given in Section 3. The theoretical results are put to a test in Section 4 describing the empirical study. We conclude the paper with Section 5 outlining the directions for future research.

2 The proposed strategy

This section presents the proposed algorithm in sub-section 2.3. We also develop a novel change-point detection algorithm as its necessary building block and describe it in sub-section 2.2. Both of these algorithms rely on Gaussian Process Regression, therefore we open the section with a brief description of this well-known approach.

2.1 Background: Gaussian Process Regression

In the given study we rely on a well known black-box non-parametric approach known as Gaussian Process Regression [16]. Formally, we model the noise with a normal distribution and impose the zero-mean Gaussian Process prior with covariance function $k(\cdot, \cdot)$ on the regression function. Then for a sequence of covariates $X_1, X_2, \ldots, X_n$ we have

$$\begin{aligned} f &\sim \mathcal{GP}\left(0, \rho k(\cdot, \cdot)\right), \\ y_j &\sim \mathcal{N}(f(X_j), \sigma^2) \text{ for } j \in 1..n, \end{aligned}$$

where n is the number of covariate-response pairs under consideration and ρ is a regularization parameter.

For a given covariate X^* the predictive distribution is also Gaussian with mean

$$\mu_* = k^* \mathcal{K}^{-1} y$$

and variance

$$\sigma_*^2 = k(X^*, X^*) - \langle k^* \mathcal{K}^{-1}, k^* \rangle,$$

where $y = [y_i]_{i=1..n}$, $\mathcal{K} = [\rho k(X_i, X_j) + \sigma^2 \delta_{ij}]_{i,j=1..n}$ and $k^* = [k(X^*, X_i)]_{i=1..n}$.

2.2 Change-point detection procedure

Our approach requires a change-point detection procedure as its crucial building block. To that end we suggest Algorithm 1. Given a sequence of covariate and response pairs $\{(X_t, y_t)\}_{t=1}^{2n}$, we train Gaussian Process Regression twice – using the first and the second half of the given data respectively. This way we obtain two predictive functions μ_1 and μ_2. Next, we make predictions for all the provided covariates and calculate the discrepancy between these predictions

$$\hat{\Delta}^2 \coloneqq \frac{1}{n} \sum_{t=1}^{2n} \left(\mu_1(X_t) - \mu_2(X_t)\right)^2.$$

Finally, we compare the discrepancy against some predetermined threshold θ_n. Intuitively, if the covariate-response pairs were generated with the same functional relationship, $\hat{\Delta}^2$ should be small,

while violation of this assumption should lead to larger values.

Algorithm 1: CPD

Data: Covariate-response pairs $\{(X_t, y_t)\}_{t=1}^{2n}$, threshold θ_n, regularization parameter ρ_{CP}
Result: `True` if a break is detected, `False` otherwise

1 $\mu_1(\cdot) \leftarrow$ train GPR on $\{(X_t, y_t)\}_{t=1}^{n}$ with ρ_{CP}
2 $\mu_2(\cdot) \leftarrow$ train GPR on $\{(X_t, y_t)\}_{t=n+1}^{2n}$ with ρ_{CP}
3 $\hat{\Delta}^2 \leftarrow \frac{1}{n}\sum_{t=1}^{2n}\left(\mu_1(X_t) - \mu_2(X_t)\right)^2$
4 **return** $\hat{\Delta}^2 > \theta_n$

2.3 GP-UCB-CPD algorithm

A well known approach called GP-UCB was proven [18] to attain sub-linear regret in a stationary setting. The main idea behind the algorithm is to train Gaussian Process Regression at each time point t, using the history of rewards. Denote the obtained predictive mean $\mu_t(\cdot)$ and predictive variance $\sigma_t^2(\cdot)$. The next input vector is chosen using the optimistic rule

$$X_t = \arg\max_{X \in \mathcal{X}} \mu_t(X) + \sqrt{\beta_t}\sigma_t(X), \tag{2.1}$$

obtaining an exploration-exploitation trade-off, where β_t are hyperparameters.

In a non-stationary setting we cannot hope for good performance of GP-UCB anymore, as non-stationarity of the underlying distribution violates assumptions of GPR consistency results. To that end we suggest to use Algorithm 1 in order to detect a change and abandon the history acquired so far. Unfortunately, we cannot use the history acquired by (2.1), as the chosen vectors might concentrate in the vicinity of $\arg\max_{X \in \mathcal{X}} \mathscr{f}_i(X)$, which will be the case after some number of break-free iterations. Therefore, we dedicate some portion of steps to uniform exploration (line 12). It is chosen adaptively with no access to the number of change-points K (see line 11). At each iteration we check whether we have accumulated enough $(2n)$ uniformly sampled points (`uniformlySampled`). If so, we apply Algorithm 1 to them (line 6). If a change-point is detected, we abandon all the data we have accumulated so far (lines 7-8). Further, it is decided (line 11) whether this step should be dedicated to uniform exploration. If so, an input X_t is chosen uniformly from $\mathcal{X}$. Otherwise, the rule (2.1) is used. In both cases the reward is received and stored along with the input in `history` and if it was an exploration step, also in `uniformlySampled`.

Remark 2.1. *The idea behind the rule choosing the number of the uniform exploration steps (line 11) can be back-ported into an earlier method suggested by [10] for multi-armed bandits, which relies on the knowledge of the number of change-points K.*

Algorithm 2: GP-UCB-CPD

Data: Convex compact $\mathcal{X}$, natural n, threshold θ_n, horizont T, sequence of positive real numbers $\{\beta_t\}$, real parameter $\xi > 0$, regularization parameters ρ_{UCB} and ρ_{CP}

```
history ← [ ] // empty list
uniformlySampled ← [ ] // empty list
for t ∈ 1, 2, ..., T do
    if |uniformlySampled| ≥ 2n then
        tail ← uniformlySampled[−2n :] // take the last 2n elements
        if CPD(tail, θ_n, ρ_CP) then
            uniformlySampled ← [ ]
            history ← [ ]
        end
    end
    if |uniformlySampled| ≤ ξ√|history| then
        Play X_t ~ U[X]
        Receive y_t ← f_{κ(t)}(X_t) + ε_t
        Append (X_t, y_t) to uniformlySampled
    else
        t' = |history|
        μ_{t'}(·), σ²_{t'}(·) ← train GPR on history with ρ_UCB
        Play X_t ← arg max_{X∈X} μ_{t'}(X) + √β_{t'} σ_{t'}(X)
        Receive y_t ← f_{κ(t)}(X_t) + ε_t
    end
    Append (X_t, y_t) to history
end
```

3 Theoretical analysis of GP-UCB-CPD

First of all, we assume, the noise ε_t has light tails. Formally, we presume them to be sub-Gaussian.

Definition 3.1 (Sub-Gaussianity)**.** *We say, a centered random variable x is sub-Gaussian with $\mathfrak{g}^2$ if*

$$\mathbb{E}\left[\exp(sx)\right] \leq \exp\left(\mathfrak{g}^2 s^2/2\right), \quad \forall s \in \mathbb{R}.$$

We say, a centered random vector X is sub-Gaussian with $\mathfrak{g}^2$ if for all unit vectors u the product $\langle u, X\rangle$ is sub-Gaussian with $\mathfrak{g}^2$.

For the sake of simplicity we restrict ourselves to $\mathcal{X} \subset \mathbb{R}$ and Matérn covariance function

$$k(x, x') \coloneqq 2^{1-\alpha} r^{\alpha} B_{\alpha}(r)/\Gamma(\alpha),$$

where $r = \sqrt{2\alpha}\,\|x - x'\|/l$, $\alpha > 1$ controls smoothness, l is the lengthscale and B_{α} denotes modified Bessel function of the second kind. Similar results can be established for the multivariate case (yet, not for a high-dimensional one, which is left for the future research) as well as for other classes of kernels (e. g. squared exponential).

Clearly, in order to quantify the difficulty of change-point detection we have to introduce a measure of discrepancy between the functions f_i. To that end we employ $\mathcal{L}_2$-norm:

$$\Delta^2 \coloneqq \min_{i=1..K-1} \int_{X\in\mathcal{X}} \left(f_i(X) - f_{i+1}(X)\right)^2 dX.$$

In the theoretical part of the paper we use $\|\cdot\|$ to denote the Euclidean norm, $\|\cdot\|_\infty$ denotes the sup-norm, while $\|\cdot\|_k$ stands for the norm of the reproducing kernel Hilbert space.

Theorem 3.1. *Let ε_t be sub-Gaussian with $\mathfrak{g}^2$, and let there exist a positive F such that $\sup_i \max\{\|f_i\|_\infty, \|f_i\|_k\} \leq F$. Choose some positive ξ,*

$$\begin{gathered}\sigma^2 = \mathfrak{g}^2,\\ n = (\mathrm{c}\Delta)^{-\frac{2\alpha}{\alpha-1/2}} T^{3/(20\alpha)},\\ \rho_{\mathrm{UCB}} = (6\log T)^{-1},\ \rho_{\mathrm{CP}} = B^2/\log n,\\ \beta_t = \mathrm{D}t^{1/(\alpha+1)}\log^3(tT),\\ \theta_n = 2\mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/\alpha}\end{gathered} \tag{3.1}$$

for some D, C and c depending only on F, $\mathbb{V}\mathrm{ar}\,[\varepsilon_1]$ and α. Finally, assume, there is enough space between the change points

$$n\left(\frac{2}{\xi}+\frac{1}{\xi^2}\right)\sqrt{2\max_i T_i} \leq \min_i T_i.$$

Then

$$\begin{aligned}\mathbb{E}\left[R_T\right] &= O\left(\left(\log^4 T + \Delta^{-\frac{2\alpha}{\alpha-1/2}}\right)\sum_{i=1}^{K} T_i^{\frac{\alpha+3}{2\alpha+2}}\right)\\ &= O\left(\left(\log^4 T + \Delta^{-\frac{2\alpha}{\alpha-1/2}}\right) K^{\frac{\alpha-1}{2\alpha+2}} T^{\frac{\alpha+3}{2\alpha+2}}\right).\end{aligned}$$

We defer the proof to Appendix A. Let us compare the obtained bound against the known results. For the sake of clarity we will use $O^*(\cdot)$ notation omitting the polylog factors. As demonstrated in [11, 18], in a stationary case GP-UCB accumulates the regret of at most $O^*\left(T^{\frac{\alpha+3}{2\alpha+2}}\right)$. Now consider a non-stationary setting and assume, the change-point locations $\tau_1, \tau_2, \ldots \tau_K$ have been made known to the agent. In such a case we can obviously bound the regret as $O^*\left(\sum_i T_i^{\frac{\alpha+3}{2\alpha+2}}\right)$. This is exactly the bound we obtained for GP-UCB-CPD under fixed Δ[1] in the realistic setting of unknown change-point locations. Therefore we conclude, the change-point detection comes with no asymptotic overhead. Next, consider the lower bound obtained in [17] for the stationary case $\Omega\left(T^{\frac{\alpha+1}{2\alpha+1}}\right)$. Clearly, for K stationary periods the lower bound is $\Omega\left(\sum_i T_i^{\frac{\alpha+1}{2\alpha+1}}\right)$. This indicates GP-UCB-CPD does not achieve minimax optimality, yet the obtained rate is considered (see [9]) to be closely following the lower bound. Moreover, in case of highly smooth functions ($\alpha \gg 1$) the method is nearly optimal.

Further, in [10] the length of stationary periods is presumed to be at least $\sim \sqrt{T}$. The assumption we make is notably weaker, presuming much slower dependence on T.

The suggested choice of parameters indicates a need for T to be known in advance. Yet, the choice of n, exhibiting the highest sensitivity to specification of T, is driven by a small power of T, approaching 0 for large α. Other parameters depend only on $\log T$. These observations imply robustness of the algorithm to misspecification of T, exceeding such of the approaches suggested in [10, 2], explicitly depending on $\sqrt{T}$.

[1] As examination of our proof reveals, we can also allow Δ to approach 0 at some polynomial rate, still matching the GP-UCB bound. The detail is omitted for brevity.

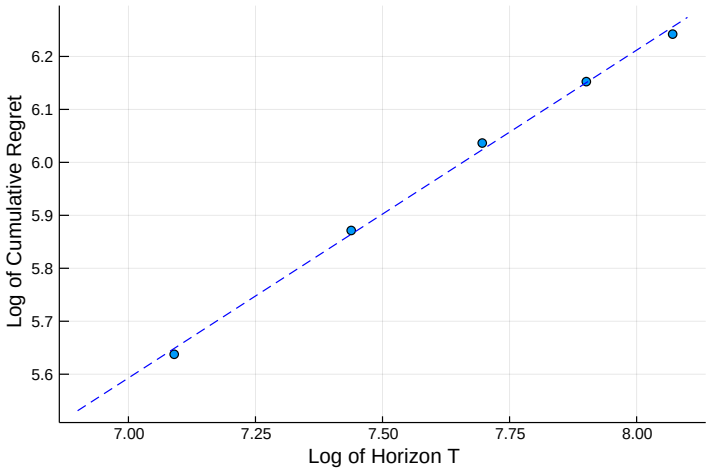

Figure 1: Here we present the dependence of cumulative regret R_T on the horizon T under fixed number of stationary periods $K = 4$. The dashed line depicts the fitted curve $3.525T^{0.619}$. Both axes are in log scale.

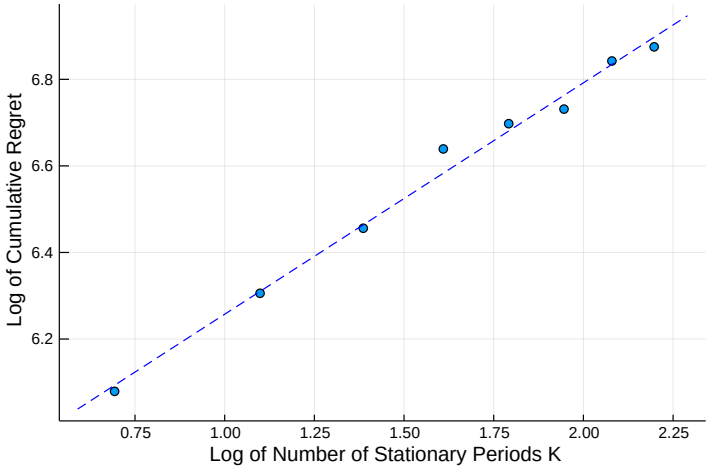

Figure 2: Cumulative regret R_T for the number of stationary periods running from 2 to 9 under fixed horizon T. The fitted curve $305.7K^{0.5346}$ is shown with a dashed line. Both axes are in log scale.

4 Experimental study

In this section we support the theoretical results experimentally and present a comparative study. We consider Matérn kernel with smoothness index $\alpha = 5/2$ and lengthscale $l = 1$, denoting it $k_0(\cdot, \cdot)$. The functions f_i are drawn independently from $\mathcal{GP}(0, k_0(\cdot, \cdot))$. The noise ε_t is independent, centered and Gaussian, its standard deviation is 0.05. The chosen domain $\mathcal{X} = [0, 5]$ is discretized into 1000 evenly spaced points. The change-points are chosen to be evenly spaced, as this is obviously the most hostile setting maximizing the theoretical lower bound. For all the experiments we choose the parameter controlling the portion of the steps, dedicated to the uniform sampling $\xi = \sqrt{3}$, the covariance function and the standard deviation of the noise the Gaussian Process Regression uses is $k_0(\cdot, \cdot)$ and $\sigma = 0.05$. ρ_{UCB} and ρ_{CP} are chosen to be equal to 1 as their variation prescribed by the theorem induces only marginal change to the overall performance. $\mathrm{D} = 0.02$, threshold of change-point detection algorithm $\theta_n = 0.2$, while its window size $n = 20$. The experiments are repeated 100 times and the results are averaged.

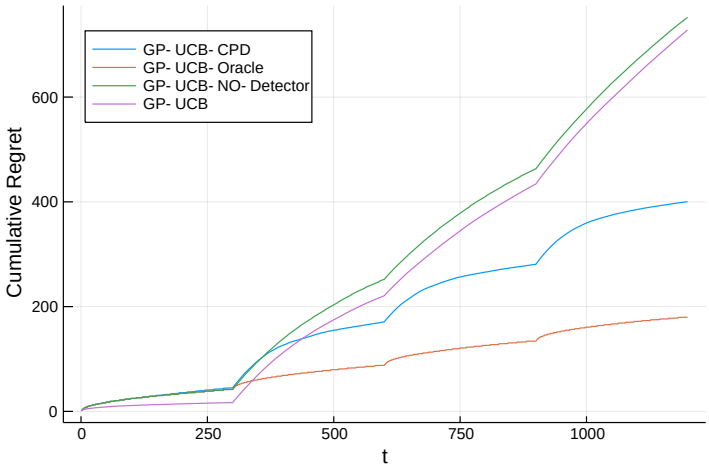

Figure 3: The plot demonstrates averaged cumulative regret of several algorithms interacting with an environment changing its behaviour every 300 points. The compared algorithms are Algorithm 2 (denoted GP-UCB-CPD), Algorithm 2 using an oracle change-point detector (GP-UCB-Oracle), Algorithm 2 using a change-point detector, which never detects a change point (Algorithm 2 with $\theta_n = +\infty$, denoted GP-UCB-NO-Detector) and finally, GP-UCB, suggested in [18] (Algorithm 2 with $\xi = 0$).

In the first experiment we examine dependence of the cumulative regret R_T on horizon T under fixed number of stationary periods K. Namely, we run Algorithm 2 for $K = 4$ and $T \in \{1200, 1700, 2200, 2700, 3200\}$. The results are shown in Figure 1. We also fit a parametric curve CT^c and the optimal power coefficient is $c = 0.619$ with 95% confidence interval being $(0.56, 0.69)$, demonstrating the dependence is clearly sub-linear. As long as Theorem 3.1 suggests $c = 0.786$ based on the smoothness index $\alpha = 5/2$, implying a significantly higher regret than the one demonstrated in the experiment. On the other hand, the authors of [17] conjecture the regret accumulated by GP-UCB can be in fact bounded by $O^*\left(T^{0.64}\right)$ for $d = 1$ and $\alpha = 5/2$, which corresponds to the rate exhibited in the experiment.

In the next step we keep the horizon fixed $T = 2700$ and assess $K = 2, 3, ..., 9$. The results are reported in Figure 2 along with the fitted curve $305.7K^{0.5346}$. The 95% confidence interval for the power coefficient is $(0.481, 0.588)$. Again, the dependence is evidently sub-linear, although the regret is accumulated faster than the theoretical result suggests. This effect is caused by the theoretical result being pessimistic in terms of T and consequently optimistic in terms of K.

We conclude the section with a comparative study. Here we choose $T = 1200$ and $K = 4$. The findings are presented in Figure 3. Algorithm 2 is denoted as GP-UCB-CPD. For the sake of comparison we also consider a version of Algorithm 2, equipped with an oracle change-point detector and is referred to as GP-UCB-Oracle. The other two algorithms we compare against are an algorithm equipped with a change-point detector, never detecting a change-point and the algorithm suggested in [18], abandoning the change-point detection (and the uniform sampling) altogether (via the choice $\xi = 0$). We call these approaches GP-UCB-NO-Detector and GP-UCB.

As we can see on the Figure 3, before the first change-point GP-UCB performs best, which is not surprising as other algorithms have accumulated regret during uniform sampling. The fact that GP-UCB-Oracle and GP-UCB-CPD accumulate approximately equal regret during this period implies low probability of false positive decision by the change-point detection algorithm. After the first break GP-UCB-CPD performs notably worse, than GP-UCB-Oracle for some period of time, which is due to an unavoidable delay of change point detection. This behaviour is repeated when the subsequent

change points happen. In spite of the stellar score before the first change-point, performance of GP-UCB greatly deteriorates after the first break. Moreover, it turns out to perform only marginally better than GP-UCB-NO-Detector.

Overall, unsurprisingly the lowest average regret is achieved by GP-UCB-Oracle, the method aware of the location of the change-point. The second best is GP-UCB-CPD.

5 Future work

In the study we have considered a realistic setting of an $\mathcal{X}$-armed bandit problem and suggested a strategy achieving sub-linear cumulative regret and near-optimality for highly smooth functions. This conclusion follows from both theoretical and empirical studies. Yet, many questions remain unanswered. The lines of our future work can be foreseen as follows

- As long as our approach relies on Gaussian Process Regression, whose performance deteriorates in high dimension, the suggested methodology is only effective in a low-dimensional setting. High-dimensional $\mathcal{X}$-armed bandits have already attracted researchers' interest in the past [12], but the non-stationary setting is yet to be analysed.
- Switching from GPR-based to tree-based approaches (see [7]) can yield an approach attaining nearly-optimal performance for wider classes of functions f_i.
- A different setting can also be considered (akin to the one suggested in [6]), where we allow the environment to change its behaviour at every step t, yet impose a bound on the total variation
- Gaussian Process Regression is notorious for its cubic time complexity, which renders it ineffective on large samples of data which are common nowadays. Thankfully, numerous linear-time approximate approaches have been developed (see [16]) and can be used to alleviate the issue. Moreover, as we have to deal with ever-growing datasets, suggesting a distributed approach is another worthy step.

A Proof of Theorem 3.1

Proof of Theorem 3.1. Choose $\omega = 10\alpha$. Choose β_t as prescribed by Lemma A.1. Now it applies here, since (C.2) is implied by (3.1) and $\alpha > 1$, the other assumptions are explicitly required to hold. Hence,

$$R_T \leq \left(\sqrt{\frac{8 \log^8 T}{\log(1+\sigma^{-2})}} + \xi + 2(c\Delta)^{-\frac{2\alpha}{\alpha-1/2}} \left(\frac{2}{\xi} + \frac{1}{\xi^2} \right) \right) \sum_{i=1}^{K} T_i^{\frac{\alpha+3}{2\alpha+2}}$$

on a set of probability at least $1 - T^{-1/2} - 2T^{-1}$, where η_T is defined by (B.1). Using the fact that $R_T \leq FT$ we have,

$$\mathbb{E}\left[R_T\right] \leq \left(\sqrt{\frac{8 \log^8 T}{\log(1+\sigma^{-2})}} + \xi + 2(c\Delta)^{-\frac{2\alpha}{\alpha-1/2}} \left(\frac{2}{\xi} + \frac{1}{\xi^2} \right) \right) \sum_{i=1}^{K} T_i^{\frac{\alpha+3}{2\alpha+2}} + F\left(\sqrt{T}+2\right)$$

or asymptotically

$$\mathbb{E}\left[R_T\right] = O\left(\left(\log^4 T + \Delta^{-\frac{2\alpha}{\alpha-1/2}}\right)\sum_{i=1}^{K} T_i^{\frac{\alpha+3}{2\alpha+2}}\right).$$

Optimization over T_i under constraint $\sum_i T_i = T$ yields the second line of the claim. Finally, use Lemma B.8 to bound η_t entering β_t. □

Lemma A.1. *Let ε_t be sub-Gaussian with $\mathfrak{g}^2$, denote*

$$F \coloneqq \max_{i=1..K}\{\max\{\left\|\not{\ell}_i\right\|_k, \left\|\not{\ell}_i\right\|_\infty\}\}$$

and choose some positive ξ and ω,

$$\sigma^2 = \mathfrak{g}^2,$$
$$\rho_{\mathrm{UCB}} = (6\log T)^{-1},$$
$$\beta_t = 14400(\eta_t \log^3 t + F^2)\log^3 T, \text{ where } \eta_t \text{ is defined by (B.1)},$$
$$\theta_n = 2\mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/\alpha}$$

and n such that

$$(\mathrm{c}\Delta)^{\frac{2\alpha}{\alpha-1/2}} \geq \frac{\log n}{n}, \tag{A.1}$$
$$n\left(\frac{2}{\xi}+\frac{1}{\xi^2}\right)\sqrt{2\max_i T_i} \leq \min_i T_i,$$

for some C and c depending only on F, $\mathbb{V}\mathrm{ar}\left[\varepsilon_1\right]$ and ω. Then on a set of probability at least $1 - 3T/n^\omega - 2T^{-1}$

$$R_T \leq \sqrt{\frac{8\log^8 T}{\log(1+\sigma^{-2})}}\sum_{i=1}^{K} T_i^{\frac{\alpha+3}{2\alpha+2}} + \left(\xi + 2n\left(\frac{2}{\xi}+\frac{1}{\xi^2}\right)\right)\sum_{i=1}^{K}\sqrt{T_i}.$$

Proof. We apply Lemma C.3 to each instance of usage of Algorithm 1. The statement of the lemma holds for all of them on a set of probability at least $1-3T/n^\omega$. The rest of the argument is conditioned on this set. Denote the regret accumulated between τ_i and τ_{i+1} in the line 19 (we exclude uniform sampling from consideration for now just like the iterations when the change has happened, but was not detected yet) as R_i. Now we apply Lemma B.7 for each interval between the changes. Its statement holds on a set of probability at least $1 - 2K/T^{-2}$, but as long as $K \leq T$, the probability is at least $1 - 2T^{-1}$.

$$\begin{aligned}\sum_{i=1}^{K} R_i &\leq \sum_{i=1}^{K}\sqrt{8T_i\beta_{T_i}\eta_{T_i}/\log(1+\sigma^{-2})} \\ &\leq \sqrt{\frac{8\log^3 T}{\log(1+\sigma^{-2})}}\sum_{i=1}^{K} T_i^{\frac{\alpha+3}{2\alpha+2}},\end{aligned} \tag{A.2}$$

where we have also used Lemma B.8 bounding η_t. The bound (A.2) does not take into account the regret accumulated during the uniform sampling and the periods when the change has happened, yet was remaining undetected. In order to estimate the delay of detection of i-th change point, consider an equation

$$\sqrt{T_i + T_{i+1} + \gamma} - \sqrt{T_i + T_{i+1}} = 1/\xi,$$

characterizing the maximal number of iterations γ between two consecutive uniform sampling steps. Clearly,

$$\gamma = \left(2\sqrt{T_i + T_{i+1}} + 1/\xi\right)/\xi \leq \left(\frac{2}{\xi} + \frac{1}{\xi^2}\right)\sqrt{T_i + T_{i+1}}.$$

Hence, the total delay of detection is at most

$$\begin{aligned}\sum_i n\left(\frac{2}{\xi} + \frac{1}{\xi^2}\right)\sqrt{T_i + T_{i+1}} &\leq \sum_i n\left(\frac{2}{\xi} + \frac{1}{\xi^2}\right)\left(\sqrt{T_i} + \sqrt{T_{i+1}}\right) \\ &\leq 2\sum_i n\left(\frac{2}{\xi} + \frac{1}{\xi^2}\right)\sqrt{T_i}.\end{aligned}$$

Also we note, the regret of $\xi\sum_i \sqrt{T_i}$ is accumulated uring the uniform sampling. Incorporating these observations with (A.2) constitutes the claim. □

B Analysis of UCB rule

This section adapts the regret bound for UCB obtained in [18]. In this section we assume the environment is stationary, i. e. for $t = 1, 2, ..., T$

$$y_t = \not{\ell}(X_t) + \varepsilon_t.$$

Denote the set of time-steps when condition in the line 11 of Algorithm 2 computes to `True` as $\mathcal{T}$, its complement as $\bar{\mathcal{T}}$ and $\mathbb{T} \coloneqq \{1, 2, ..., T\}$. Further, for a sequence of real values $\{a_i\}$ and a set of indexes E we write $a_E \coloneqq [a_i]_{i\in E}$. If not said otherwise, in this section we choose $\rho_{\mathrm{UCB}} = 1$.

Here we employ the concept of information gain, defined as mutual information between the function $f \sim GP(0, k(\cdot,\cdot))$ and the observations $y_{\bar{\mathcal{T}}}$

$$\mathbb{I}(y_{\bar{\mathcal{T}}}; f) \coloneqq H(y_{\bar{\mathcal{T}}}) - H(y_{\bar{\mathcal{T}}}|f),$$

where $H(\cdot)$ denotes entropy and in our case

$$\mathbb{I}(y_{\bar{\mathcal{T}}}; f) = \frac{1}{2}\log\det\left(I_{|\bar{\mathcal{T}}|} + \sigma^{-2}\mathcal{K}_{\bar{\mathcal{T}}}\right),$$

where $\mathcal{K}_{\bar{\mathcal{T}}} = [k(X_t, X_{t'})]_{t,t'\in\bar{\mathcal{T}}}$. In order to extend the results by [18] for the case allowing for uniform sampling (see line 12) we prove the following trivial lemma.

Lemma B.1.

$$\mathbb{I}(y_{\bar{\mathcal{T}}}; f) \leq \mathbb{I}(y_{\mathbb{T}}; f)$$

Proof. The claim follows from the fact that the eigenvalues of $I_{|\bar{\mathcal{T}}|} + \sigma^{-2}\mathcal{K}_{\bar{\mathcal{T}}}$ and $I_{|\mathbb{T}|} + \sigma^{-2}\mathcal{K}_{\mathbb{T}}$ are larger or equal to 1, while $\left|\bar{\mathcal{T}}\right| \leq |\mathbb{T}|$. □

The next result connects information gain and predictive variance of GPR.

Lemma B.2 (Lemma 5.3 by [18])**.**

$$\mathbb{I}(y_{\bar{\mathcal{T}}}; f) = \frac{1}{2}\sum_{t\in\bar{\mathcal{T}}}\left(1 + \sigma^{-2}\sigma_t^2(X_t)\right).$$

Now we are ready to assess the properties of the highest information gain possible from n observations

$$\eta_n = \sup_{\{X_i\}_{i=1}^n} \mathbb{I}([y_i]_{i=1..n}; f). \tag{B.1}$$

Lemma B.3 (extension of Lemma 7.1 by [18])**.** *For an arbitrary positive* ζ

$$\frac{1}{2}\sum_{t\in\bar{\mathcal{T}}} \max\{\sigma^{-2}\sigma_t^2(X_t), \zeta\} \leq \frac{2\zeta}{\log(1+\zeta)}\eta_T.$$

Proof. The proof consists in combining Lemma B.2, Lemma B.1 and the fact that $\min\{r, \zeta\} \leq \zeta\log(1+r)/\log(1+\zeta)$ for all positive r. □

Next we extend GPR consistency result for the case of sub-Gaussian noise.

Lemma B.4 (Theorem 6 in [18])**.** *Let* $\delta \in (0,1)$, $\sup_{t\in\mathbb{T}}|\varepsilon_t| \leq \sigma$ *and choose*

$$\beta_t = 2\|f\|_k^2 + 300\eta_t \ln^3(t/\delta).$$

Then on a set of probability at least $1-\delta$ *for all* $x \in \mathcal{X}$ *for all* t *for the predictive mean* $\mu_t(\cdot)$ *obtained based on* t *observations*

$$|\mu_t(x) - f(x)| \leq \beta_t^{1/2}\sigma_t(x).$$

Lemma B.5. *Let* ε_t *be sub-Gaussian with* $\mathfrak{g}^2$, $\delta \in (0,1)$ *and* $\mathrm{u} > 0$. *Choose*

$$\sigma = \mathfrak{g}\sqrt{2\left(\mathrm{u} + \log T\right)}$$

and

$$\beta_t = 2\|f\|_k^2 + 300\eta_t \log^3(t/\delta).$$

Then on a set of probability at least $1-\delta-\exp(\mathrm{u})$ *for all* $x \in \mathcal{X}$ *for all* t *for the predictive mean* $\mu_t(\cdot)$ *obtained based on* t *observations*

$$|\mu_t(x) - f(x)| \leq \beta_t^{1/2}\sigma_t(x).$$

Proof. Due to sub-Gaussianity for all t for any positive x

$$\mathbb{P}\left\{|\varepsilon_t| > \mathrm{x}\right\} \leq 2\exp\left(-\frac{\mathrm{x}^2}{2\mathfrak{g}^2}\right)$$

and uniformly

$$\mathbb{P}\left\{\sup_{t\leq T}|\varepsilon_t| > \mathrm{x}\right\} \leq 2T\exp\left(-\frac{\mathrm{x}^2}{2\mathfrak{g}^2}\right).$$

Change of variables yields for any positive u

$$\mathbb{P}\left\{\sup_{t\leq T}|\varepsilon_t| > \mathfrak{g}\sqrt{2\left(\mathrm{u}+\log T\right)}\right\} \leq \exp(-\mathrm{u}).$$

Finally, choose $\sigma = \mathfrak{g}\sqrt{2\left(\mathrm{u}+\log T\right)}$ and apply Lemma B.4. □

Using Lemma B.5 instead of Theorem 6 by [18] we extend Theorem 3 by [18] in the desired way bounding the regret of GP-UCB-CPD in the absence of change-points.

Lemma B.6. *Let ε_t be sub-Gaussian with $\mathfrak{g}^2$, choose*

$$\sigma^2 = 2\mathfrak{g}^2(\mathrm{u} + \log T)$$

and

$$\beta_t = 2\left\|\not{f}\right\|_k^2 + 300\eta_t \log^3(t/\delta).$$

Then on a set of probability at least $1 - \delta - \exp(-\mathrm{u})$

$$R_T \leq \sqrt{8T\beta_T\eta_T / \log(1+\sigma^{-2})}.$$

As one can see, Lemma B.6 suggests a choice of σ^2 depending on the horizon T. For the sake of uniformity we note, this scheme is equivalent to the one with σ^2 independent of T at the cost of non-trivial regularization parameter ρ. Really, Lemma B.6 suggests $k(\cdot,\cdot) + 2\mathfrak{g}^2(\mathrm{u}+\log T)\delta(\cdot,\cdot)$ as a covariance function of responses. Clearly, one can use $k(\cdot,\cdot)/(2(\mathrm{u}+\log T)) + \mathfrak{g}^2\delta(\cdot,\cdot)$ instead. By the means of simple algebra one verifies, the posterior mean remains the same, while the posterior variance shall be $2(\mathrm{u}+\log T)$ times smaller, and hence, β_t shall be adjusted accordingly. The next lemma summarizes this observation and changes the variables driving the probability of the set its statement is conditioned on.

Lemma B.7. *Let ε_t be sub-Gaussian with $\mathfrak{g}^2$, and $\omega > 0$, $\left\|\not{f}\right\|_k < +\infty$ and choose*

$$\sigma^2 = \mathfrak{g}^2,$$

$$\rho_{\mathrm{UCB}} = (6\log T)^{-1}$$

and

$$\beta_t = 14400(\eta_t \log^3 t + F^2)\log^3 T.$$

Then on a set of probability at least $1 - 2T^{-2}$

$$R_T \leq \sqrt{8T\beta_T\eta_T / \log(1+\sigma^{-2})}.$$

Proof. As seen above, the desired bound can be established on a set of probability at least $1 - \delta - \exp(-\mathrm{u})$ under the choice

$$\sigma^2 = 2\mathfrak{g}^2,$$

$$\rho_{\mathrm{UCB}} = (2(\mathrm{u}+\log T))^{-1},$$

and

$$\beta_t = 2(\mathrm{u}+\log T)\left(2\left\|\not{f}\right\|_k^2 + 300\eta_t \log^3(tT^2)\right).$$

Now we choose $\delta = T^{-2}$ and $\mathrm{u} = 2\log T$. Substitution yields the claim. □

In conclusion, we cite a result bounding the information gain.

Lemma B.8 (Theorem 5 by [18])**.** *Let $k(\cdot,\cdot)$ be Matérn covariance function with smoothness index α.*

$$\eta_T = O\left(T^{1/(\alpha+1)}\log T\right).$$

C Formal treatment of Algorithm 1

In this section we establish two theoretical results regarding our change-point detection procedure. Namely, Lemma C.1 provides an upper bound on $\hat{\Delta}^2$ in the absence of a change-point, while Lemma C.2 gives its lower bound. These two results combined induce a proper choice of the threshold θ_n.

First, assume $\{X_t\}_{t=1}^{2n} \overset{\text{iid}}{\sim} U(\mathcal{X})$ and let

$$y_t = f(X_t) + \varepsilon_t,$$

where ε_t denotes i.i.d. centered noise. Further, we consider a broad class of smooth functions.

Definition C.1 (Sobolev class)**.** *Consider an orthonormal basis $\{\psi_j\}$ in $L_2(\mathbb{R}^p)$ and a function $f = \sum_j f_j\psi_j \in L_2(\mathbb{R}^p)$. We call it α-smooth Sobolev for $\alpha > 1/2$ if*

$$\exists B : \sum_{j=1}^{\infty} j^{2\alpha} f_j^2 \leq B^2.$$

Lemma C.1. *Let ε_i be sub-Gaussian, $k(\cdot,\cdot)$ satisfy Assumption D.1 and Assumption D.2 and f be Sobolev with α and B. Choose $\rho_{\text{CP}} = B^2/\log n$. Then for any $\omega > 0$, some constant C, which depends only on $\mathbb{V}\mathrm{ar}\,[\varepsilon_1]$, ω and B, with probability at least $1 - 2n^{-\omega}$*

$$\hat{\Delta}^2 \leq \mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/\alpha}.$$

Proof. The proof consists in applying Lemma D.1 twice, yielding concentrations of μ_1 and μ_2 around f and a piece of straightforward algebra.

$$\begin{aligned}
\hat{\Delta}^2 &= \frac{1}{n}\sum_{t=1}^{2n}\left(\mu_1(X_t) - \mu_2(X_t)\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{2n}\left(\mu_1(X_t) - f(X_t) + f(X_t) - \mu_2(X_t)\right)^2 \\
&\leq 4\Delta_f^2 \\
&\lesssim B^{1/\alpha}\left(\frac{\sigma^2\log n}{n}\right)^{(\alpha-1/2)/\alpha},
\end{aligned}$$

where Δ_f comes from Lemma D.1. □

On the other hand, let $\{X_t\}_{t=1}^{2n} \overset{\text{iid}}{\sim} U(\mathcal{X})$ as before and let there be two functions f_1 and f_2 such that

$$y_t = f_1(X_t) + \varepsilon_t \text{ for } t \leq n$$

and

$$y_t = f_2(X_t) + \varepsilon_t \text{ for } t > n.$$

Needless to say, an ability of the algorithm to detect a change-point depends on some measure of discrepancy between the two functions. We suggest to consider $\mathcal{L}_2$-norm.

$$\Delta^2 \coloneqq \int_{X\in\mathcal{X}} \left(f_1(X) - f_2(X)\right)^2 dX.$$

Lemma C.2. *Let the functions f_1 and f_2 be bounded: $F \coloneqq \max\{\|f_1\|_\infty, \|f_2\|_\infty\}$. Choose $\rho_{\mathrm{CP}} = B^2/\log n$. Then for any positive v, ω and a positive constant C depending only on F, B, $\mathbb{V}\mathrm{ar}[\varepsilon_1]$ and ω*

$$\mathbb{P}\left\{\hat{\Delta} \geq \Delta - \sqrt{\frac{\mathrm{v}}{2n}} - \mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/(2\alpha)}\right\} \geq 1 - 2\exp(-\mathrm{v}) - 2n^{-\omega}.$$

Proof. First, consider

$$\tilde{\Delta}^2 \coloneqq \frac{1}{n}\sum_{t=1}^{2n}\left(f_1(X_t) - f_2(X_t)\right)^2.$$

Clearly, $\mathbb{E}\left[\tilde{\Delta}^2\right] = \Delta^2$, the summands are i.i.d. and bounded, hence Hoeffding's inequality applies here and yields for any positive v

$$\mathbb{P}\left\{\left|\Delta^2 - \tilde{\Delta}^2\right| > F\sqrt{\frac{\mathrm{v}}{n}}\right\} \leq 2\exp(-\mathrm{v})$$

and using $\left|\Delta - \tilde{\Delta}\right| \leq 2F$ we have

$$\mathbb{P}\left\{\left|\Delta - \tilde{\Delta}\right| > \sqrt{\frac{\mathrm{v}}{2n}}\right\} \leq 2\exp(-\mathrm{v}). \tag{C.1}$$

In the next piece of algebra for a function ϕ defined on $\mathcal{X}$ we write $\phi^{1..2n}$ to refer an element-wise application $\phi^{1..2n} \coloneqq (\phi(X_1), \phi(X_2), ..., \phi(X_{2n}))^T$.

$$\begin{aligned}
\hat{\Delta} &= \frac{1}{\sqrt{n}}\left\|\mu_1^{1..2n} - \mu_2^{1..2n}\right\| \\
&= \frac{1}{\sqrt{n}}\left\|\mu_1^{1..2n} - f_1{}^{1..2n} + \left(f_1{}^{1..2n} - f_2{}^{1..2n}\right) + f_2{}^{1..2n} - \mu_2^{1..2n}\right\| \\
&\geq \tilde{\Delta} - \frac{1}{n}\left(\left\|\mu_1^{1..2n} - f_1{}^{1..2n}\right\| + \left\|\mu_2^{1..2n} - f_2{}^{1..2n}\right\|\right).
\end{aligned}$$

Now we make use of Lemma D.1 (defining Δ_f) and (C.1) and obtain

$$\hat{\Delta} \geq \Delta - \sqrt{\frac{\mathrm{v}}{2n}} - \mathrm{C}\Delta_f$$

on a set of probability at least $1 - 2\exp(-\mathrm{v}) - 2n^{-\omega}$. □

Finally, we are ready to describe the behavior of Algorithm 1.

Lemma C.3. *Let f, f_1 and f_2 be Sobolev with α and B, further let these functions be bounded: $F \coloneqq \max\{\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\} < +\infty$. Choose $\rho_{\mathrm{CP}} = B^2/\log n$ and an arbitrary positive ω. There exist positive constants c and C depending only on F, B, $\mathbb{V}\mathrm{ar}[\varepsilon_1]$ and ω, such that the choice of the threshold*

$$\theta_n = 2\mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/\alpha}$$

and the length of the sample n satisfying

$$(\mathrm{c}\Delta)^{\frac{2\alpha}{\alpha-1/2}} \geq \frac{\log n}{n} \tag{C.2}$$

implies no false alarm if the data are not subject to a change and guaranties detection of a change if such is present with probability at least $1 - 3n^{-\omega}$.

Proof. Lemma C.1 bounds $\hat{\Delta}^2$ in the absence of a break:

$$\hat{\Delta}^2 \leq \mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/\alpha}.$$

Hence, the choice of the threshold $\theta_n = 2\mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/\alpha}$ allows for at most $2n^{-\omega}$ first type error rate. Now, using Lemma C.2 we see, it is sufficient to show, that

$$\Delta - \sqrt{\frac{\mathrm{v}}{2n}} - \mathrm{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/(2\alpha)} \geq \theta_n. \tag{C.3}$$

Below we use $\mathcal{C}$ to denote a generalized constant, which depends only on F, B, $\mathbb{V}\mathrm{ar}\left[\varepsilon_1\right]$, ω its value might change from line to line. (C.3) is equivalent to

$$\Delta \geq \mathcal{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/(2\alpha)} + \mathcal{C}\sqrt{\frac{\mathrm{v}}{2n}}.$$

Under the choice of $\mathrm{v} = \omega \log n$ due to $(\alpha - 1/2)/(2\alpha) < 1/2$ it is sufficient to obtain

$$\Delta \geq \mathcal{C}\left(\frac{\log n}{n}\right)^{(\alpha-1/2)/(2\alpha)},$$

which follows from (C.2). □

D Consistency of Gaussian Process Regression by [21]

In this section we quote a consistency result for predictions of Gaussian Process Regression. It imposes the following two assumptions on the covariance function $k(\cdot,\cdot)$.

Assumption D.1. *Let there exist C_ψ and L_ψ s.t. for eigenfunctions $\{\psi_j(\cdot)\}_{j=1}^\infty$ of covariance function $k(\cdot,\cdot)$*

$$\max_j \|\psi_j\|_\infty \leq C_\psi$$

and for all $x_1, x_2 \in \mathcal{X}$

$$|\psi_j(x_1) - \psi_j(x_2)| \leq jL_\psi \|x_1 - x_2\|.$$

Assumption D.2. *Let for the eigenvalues $\{\lambda_j\}_{j=1}^\infty$ of covariance function $k(\cdot,\cdot)$ exist positive c and C s.t. $cj^{-2\alpha} \leq \lambda_j \leq Cj^{-2\alpha}$ for $\alpha > 1/2$.*

Matérn kernel with smoothness index α satisfies these assumptions. In [21] the authors claim, their results also hold for kernels with non-polynomially decaying eigenvalues, like RBF and polynomial kernels.

Lemma D.1 (Corollary 2.1 in [21])**.** *Assume ε_i are sub-Gaussian and f is α-smooth Sobolev. Assume, $X_i \sim U[\mathcal{X}]$. Further, let $k(\cdot,\cdot)$ satisfy Assumption D.1 and Assumption D.2 and choose*

$$\rho = \frac{B^2}{\log n},$$

where n is the size of the training sample. Then, with probability at least $1 - n^{-\omega}$ for any positive ω we have

$$\|f - \mu\|_\infty \leq \Delta_f \coloneqq A(\omega)B^{1/2\alpha}\left(\frac{\mathbb{V}\mathrm{ar}\left[\varepsilon_1\right]\log n}{n}\right)^{(\alpha-1/2)/(2\alpha)},$$

where μ denotes the predictive function and $A(\omega)$ is a constant depending only on ω.

This result is demonstrated in [21] for $\omega = 10$, however this choice is purely arbitrary.

References

[1] Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: From regret to policy regret. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2:1503–1510, 2012.

[2] Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 138–158, Phoenix, USA, 25–28 Jun 2019. PMLR.

[3] Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 45–53, New York, NY, USA, 2004. ACM.

[4] Dirk Bergemann and Ulrich Hege. The financing of innovation : learning and stopping. *The RAND Journal of Economics*, 36, 02 2001.

[5] Dirk Bergemann and Juuso Välimäki. Learning and strategic pricing. *Econometrica*, 64(5):1125–49, 1996.

[6] Omar Besbes. Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. pages 1–9.

[7] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *J. Mach. Learn. Res.*, 12:1655–1695, July 2011.

[8] J C. Gittens and Michael Dempster. Bandit processes and dynamic allocation indices [with discussion]. *Journal of the Royal Statistical Society. Series B: Methodological*, 41:148–177, 02 1979.

[9] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian Process Optimization with Adaptive Sketching: Scalable and No Regret. 99:1–25, 2019.

[10] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly Optimal Adaptive Procedure for Piecewise-Stationary Bandit: a Change-Point Detection Approach. 2018.

[11] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. *34th International Conference on Machine Learning, ICML 2017*, 2:1397–1422, 2017.

[12] Josip Djolonga, Andreas Krause, and Volkan Cevher. High-Dimensional Gaussian Process Bandits. *Advances in Neural Information Processing Systems 26*, pages 1025–1033, 2013.

[13] J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley, Chichester, NY, 1989.

[14] J. C. Gittins and D. M. Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

[15] R. Kleinberg and T. Leighton. The value of knowing a demand curve: bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605, Oct 2003.

[16] Rasmussen and Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[17] Jonathan Scarlett, Ilijia Bogunovic, and Volkan Cevher. Lower Bounds on Regret for Noisy Gaussian Process Bandit Optimization. pages 1–20, 2017.

[18] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. 2009.

[19] P. Whittle. Arm-acquiring bandits. *Ann. Probab.*, 9(2):284–292, 04 1981.

[20] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.

[21] Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. pages 1–43, 2017.

[22] Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2483–2491. Curran Associates, Inc., 2011.

[23] M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146, 1969.