Delayed Feedback in Kernel Bandits

Sattar Vakili [1] **Danyal Ahmed** [1] **Alberto Bernacchia** [1] **Ciara Pike-Burke** [2]

Abstract

Black box optimisation of an unknown function from expensive and noisy evaluations is a ubiquitous problem in machine learning, academic research and industrial production. An abstraction of the problem can be formulated as a kernel based bandit problem (also known as Bayesian optimisation), where a learner aims at optimising a kernelized function through sequential noisy observations. The existing work predominantly assumes feedback is immediately available; an assumption which fails in many real world situations, including recommendation systems, clinical trials and hyperparameter tuning. We consider a kernel bandit problem under stochastically delayed feedback, and propose an algorithm with $\tilde{\mathcal{O}}(\sqrt{\Gamma_k(T)T}+\mathbb{E}[\tau])$ regret, where T is the number of time steps, $\Gamma_k(T)$ is the maximum information gain of the kernel with T observations, and τ is the delay random variable. This represents a significant improvement over the state of the art regret bound of $\tilde{\mathcal{O}}(\Gamma_k(T)\sqrt{T}+\mathbb{E}[\tau]\Gamma_k(T))$ reported in Verma et al. (2022). In particular, for very non-smooth kernels, the information gain grows almost linearly in time, trivializing the existing results. We also validate our theoretical results with simulations.

1. Introduction

The kernel bandit problem is a flexible framework which captures the problem of learning to optimise an unknown function through successive input queries. Typically, the game proceeds in rounds where in each round the learner selects an input point to query and then immediately receives a noisy observation of the function at that point. This observation can be used immediately to improve the learners decision of which point to query next. Due to its generality, the kernel bandit problem has become very popular in practice. In particular, it enables us to sequentially learn to optimise a variety of different functions without needing to know many details about the functional form.

However, in many settings where we may want to use kernel bandits, we also have to deal with delayed observations. For example, consider using kernel bandits to sequentially learn to select the optimal conditions for a chemical experiment. The chemical reactions may not be instantaneous, but instead take place at some random time in the future. If we start running a sequence of experiments, we can start new experiments before receiving the stochastically delayed feedback from the previous ones. However, in this situation we have to update the conditions for future experiments before receiving all the feedback from previous experiments. Similar situations arise in recommendation systems, clinical trials and hyperparameter tuning, so it is of practical relevance that we design kernel bandit algorithms that are able to deal with delayed feedback. Moreover, a big challenge for existing kernel bandit algorithms is the computational complexity. Generally speaking, in each round t, algorithms for kernel bandits require fitting a kernel model to the t observed data points which can have an $\mathcal{O}(t^3)$ complexity. To reduce the complexity, there has been a recent interest in considering batch versions of kernel bandit algorithms. These algorithms select τ input values using the same model then update the model after receiving all observations in the batch. This corresponds to a delay of at most τ in receiving each observation, and thus can be thought of as an instance of delayed feedback in kernel bandits.

In this paper, we study the kernel bandit problem with stochastically delayed feedback. We propose Batch Pure Exploration with Delays (BPE-Delay) —an adaptation of the Batch Pure Exploration algorithm (Li and Scarlett, 2022) to the delayed feedback setting—, and show that, under mild assumptions on the unknown delay distribution, BPE-Delay achieves near optimal regret. Indeed, we prove that BPE-Delay enjoys regret scaling as $\tilde{\mathcal{O}}(\sqrt{\Gamma_k(T)T}+\mathbb{E}[\tau])$,[1] where T is the number of time steps, $\Gamma_k(T)$ is the maximum information gain of the kernel k with T observations (see Section 2), and τ is the delay random variable. This result

[1]MediaTek Research [2]Imperial College London. Correspondence to: Sattar Vakili <sattar.vakili@mtkresearch.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]We use the notation $\mathcal{O}$ and $\tilde{\mathcal{O}}$ to denote order and order up to logarithmic factors, respectively.

essentially shows that the impact of delayed feedback on the regret of this algorithm is independent of the horizon T and the problem parameter $\Gamma_k(T)$. This desirable property means that as T or $\Gamma_k(T)$ increase, the impact of the delay remains the same.

We note that for linear models, $\Gamma_k(T) = \mathcal{O}(d\log(T))$, where d is the input dimension. In the case of very smooth kernels such as Squared Exponential (SE), $\Gamma_k(T) = \text{poly}\log(T)$. However $\Gamma_k(T)$ can become arbitrary close to linear in T for less smooth kernels. For example, in the case of a Matérn kernel with smoothness parameter ν, we have $\Gamma_k(T) = \tilde{\mathcal{O}}(T^{\frac{d}{2\nu+d}})$ (Vakili et al., 2021c). Therefore, our results represent a significant theoretical improvement over the existing work on kernel bandits with delayed feedback, where an $\tilde{\mathcal{O}}(\Gamma_k(T)\sqrt{T} + \mathbb{E}[\tau]\Gamma_k(T))$ regret bound was shown for an algorithm based on upper confidence bound (UCB) acquisition of observation points (Verma et al., 2022). In particular, for non-smooth kernels, our result reduces the delay penalty from possibly near $\tilde{\mathcal{O}}(\mathbb{E}[\tau]T)$ to just $\tilde{\mathcal{O}}(\mathbb{E}[\tau])$. We demonstrate in Section 7 that these theoretical improvements are also visible experimentally. In addition, when applied to the special case of linear bandits (a kernel bandit problem with linear kernel), our results improve the dependency of the delay related term in the regret bound on the input dimension by a $d^{3/2}$ factor compared to the state of the art in Howson et al. (2022). A detailed comparison with related work is provided in Section 2.

2. Background and Related Work

In this section, we give an overview of the background on kernel bandits with immediate feedback and delayed feedback in simpler stochastic bandit formulations (namely in K-armed and linear bandits). We then provide a detailed comparison with the most related work on kernel bandits with delayed feedback.

Kernel bandits with immediate feedback. In the typical kernel bandit setting with immediate feedback, classic algorithms such as selecting the query points based on upper confidence bounds (GP-UCB) (Srinivas et al., 2010) and Thompson Sampling (Chowdhury and Gopalan, 2017) achieve a regret bound of $\tilde{\mathcal{O}}(\Gamma_k(T)\sqrt{T})$. This regret bound is suboptimal (see, Vakili et al., 2021e, for details), and can be improved to $\tilde{\mathcal{O}}(\sqrt{\Gamma_k(T)T})$ using more recent work. In particular, Batch Pure Exploration (BPE) introduced in Li and Scarlett (2022) and Threshold Domain Shrinking (GP-ThreDS) proposed in Salgia et al. (2021) achieve this improved regret bound. Here, $\Gamma_k(T)$ is a kernel specific complexity term, which can be interpreted as the *information gain* or the *effective dimension*.

Information gain and effective dimension. While the feature space representation of common kernels is infinite dimensional, with a finite data set, only a finite number of features have a significant effect on the kernel based regression. That leads to the definition of the effective dimension. In particular, consider a finite set $\mathbf{X}_T = \{X_1, \dots, X_T\}$ of observation points and a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathbf{K}_{\mathbf{X}_T} = [k(X_t, X_{t'})]_{t,t'=1}^T$ be the kernel matrix resulting from the pairwise kernel values between the observation points. The effective dimension for a given kernel and observation set is often defined as (Zhang, 2005; Valko et al., 2013)

$$\tilde{d}_k(T) = \text{tr}\left(\mathbf{K}_{\mathbf{X}_T}(\mathbf{K}_{\mathbf{X}_T} + \lambda^2\mathbf{I}_T)^{-1}\right), \tag{1}$$

where $\lambda > 0$ is a free parameter and $\mathbf{I}_T$ denotes an identity matrix of dimension T.

It is also useful to define a related notion of information gain. For this, assume f is a centered Gaussian Process (GP) on the domain $\mathcal{X}$ with kernel k. Information gain then refers to the mutual information $\mathcal{I}(\mathbf{y}_T; f)$ between the data values $\mathbf{y}_T = [y_t = f(X_t) + \epsilon_t]_{t=1}^T$ and f (assuming the surrogate GP distribution and a zero mean Gaussian noise ϵ_t with variance λ^2). From the closed form expression of mutual information between two multivariate Gaussian distributions (Cover, 1999), it follows that $\mathcal{I}(\mathbf{y}_T; f) = \frac{1}{2}\log\det\left(\mathbf{I}_T + \frac{1}{\lambda^2}\mathbf{K}_{\mathbf{X}_T}\right)$. We then define the data-independent and kernel-specific *maximum information gain* as follows (Srinivas et al., 2010):

$$\Gamma_k(T) = \sup_{\mathbf{X}_T \subset \mathcal{X}} \mathcal{I}(\mathbf{y}_T; f). \tag{2}$$

It is known that the information gain and the effective dimension are the same up to logarithmic factors. Specifically, we have $\tilde{d}_k(T) \leq \mathcal{I}(\mathbf{y}_T; f)$, and $\mathcal{I}(\mathbf{y}_T; f) = \mathcal{O}(\tilde{d}_k(T)\log(T))$ (Calandriello et al., 2019). For specific kernels, explicit bounds on $\Gamma_k(T)$ are derived in Srinivas et al. (2010); Vakili et al. (2021b;c). We report our regret bounds in terms of information gain that can be easily replaced by the effective dimension, up to logarithmic factors.

Regret lower bounds. For commonly used SE and Matérn kernels, Scarlett et al. (2017) derived lower bounds on regret (in the immediate feedback setting). In particular, they showed $\Omega(\sqrt{T(\log(T))^{d/2}})$ and $\Omega(T^{\frac{\nu+d}{2\nu+d}})$ lower bounds on regret, in the case of SE and Matérn kernels, respectively. These bounds are matched by the regret bounds for GP-ThreDS (Salgia et al., 2021) and BPE (Li and Scarlett, 2022), up to logarithmic factors.

Delayed feedback in stochastic bandits. Delayed feedback has been studied significantly in the stochastic bandit problem where there are K independent arms. Joulani et al. (2013) provided a queue based wrapper algorithm which,

when applied with any bandit algorithm, leads to an additive penalty of $\mathbb{E}[\tau]$ in the regret compared to the non-delayed setting. They also showed that using a UCB algorithm (Auer et al., 2002) just with the observed data would lead to a similar regret bound of $\tilde{\mathcal{O}}(\sqrt{KT} + \mathbb{E}[\tau])$. Mandel et al. (2015) also provided a queue based algorithm with the same regret bound. Vernade et al. (2017) developed a UCB algorithm to deal with delayed feedback where some observations are censored if their delay exceeds a threshold. We note that we cannot directly extend these queue based approaches to the kernel bandit problem where we have a large number of dependent arms. However we show that the same additive penalty can be maintained even in our more difficult setting. Additionally, while it is possible to extend the UCB based algorithms to our setting, in Section 7, we show that our proposed algorithm performs better empirically than using a delayed version of GP-UCB.

There has also been work studying the impact of delayed feedback in generalised linear bandits. Zhou et al. (2019) and Howson et al. (2022) provided adaptations of optimistic (UCB) algorithms to account for delayed feedback with sub-exponential delays. Zhou et al. (2019) obtained a regret bound scaling with $\tilde{\mathcal{O}}(d\sqrt{T} + \sqrt{dT\mathbb{E}[\tau]})$, while Howson et al. (2022) obtained an improved regret bound of $\tilde{\mathcal{O}}(d\sqrt{T} + d^{\frac{3}{2}}\mathbb{E}[\tau])$, although here the delay penalty still depends on the dimension which is not the case for the delayed stochastic K-armed bandit problem. When applied to this setting, we show that our proposed algorithm removes this interaction between the dimension and the delay. Specifically, our results improve the delay related term in the regret bound with a $d^{\frac{3}{2}}$ factor in the special case of linear bandits (a kernel bandit problem with linear kernel). We also note that Vernade et al. (2020) extended their work on delayed feedback with censoring to the contextual linear bandit setting, and Dudik et al. (2011) studied constant delays in the the contextual bandit setting, although these settings are not directly comparable to ours.

Delayed feedback in kernel bandits. The most relevant work to ours is Verma et al. (2022) where the kernel bandit problem with stochastically delayed feedback was also considered. Verma et al. (2022) proposed algorithms based on GP-UCB (Srinivas et al., 2010) and Thompson sampling (Chowdhury and Gopalan, 2017) in the delayed feedback setting. In these algorithms, referred to as GP-UCB-SDF and GP-TS-SDF (where SDF stands for Stochastic Delayed Feedback), the unavailable feedback due to delay are replaced by minimum function value (assuming it is known). They provided a regret bound for this algorithm of $\tilde{\mathcal{O}}(\Gamma_k(T)\sqrt{T} + \mathbb{E}[\tau]\Gamma_k(T))$. This is an improvement over a naive application of the existing algorithms (which would lead to a regret bound of $\mathcal{O}(\Gamma_k(T)\sqrt{T\mathbb{E}[\tau]})$), but still suffers from a scaling of the term involving the delay by $\Gamma_k(T)$. In comparison, our algorithm does not require the additional knowledge of the minimum function value (we simply disregard the unavailable observations). Furthermore, our results significantly improve upon Verma et al. (2022), by completely decoupling the impact of the delay and the problem parameters. Our regret bounds are also order optimal and cannot be further improved for the cases where a lower bound on regret is known, in particular, the bounds given in Scarlett et al. (2017) for the SE and Matérn kernels (under the immediate feedback setting).

Kernel bandits with batch observations can be considered as a special case of our delayed feedback framework, with constant (non-stochastic) delays. Specifically, the observations for the points in a batch are available with a fixed delay that is at most the length of the batch. Notable examples are Desautels et al. (2014); Vakili et al. (2021d); Daxberger and Low (2017); Chowdhury and Gopalan (2019) which are based on the *hallucination* technique introduced in Desautels et al. (2014). In this technique, the unavailable observations are *hallucinated* to be the kernel based prediction using only the available feedback. This is equivalent to keeping the prediction unaffected by the unavailable observations, while updating the uncertainty estimate using all selected observation points, including the ones with delayed observations (Chowdhury and Gopalan, 2019). In contrast, Verma et al. (2022) set the unavailable observations to the minimum function value (assuming it is known). In our algorithm, we simply disregard the unavailable observations in the prediction, while using both variations of uncertainty estimate (one updated using all observation points, and one updated using only the observation points with available feedback). Details are given in Section 5. Our regret bounds also offer a significant improvement over the batch setting, despite this being a significantly simpler setting due to the fixed and known delays. In particular, in the batch setting, the best known regret bounds by Chowdhury and Gopalan (2019) are $\tilde{\mathcal{O}}(\Gamma_k(T)\sqrt{T} + \tau\sqrt{T\Gamma_k(T)})$, where with an abuse of notation to enable comparison with our results, we use τ for the batch size. Theorem 6.1 improves upon both terms in this regret bound.

3. Problem Definition

Consider a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ supported on a compact d-dimensional set $\mathcal{X} \subset \mathbb{R}^d$. A Hilbert space $\mathcal{H}_k$ of functions on $\mathcal{X}$ equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is called a reproducing kernel Hilbert space (RKHS) with reproducing kernel k if the following is satisfied. For all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}_k$, and for all $x \in \mathcal{X}$ and $f \in \mathcal{H}_k$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$ (reproducing property).

We consider a kernel bandit problem. We assume there exists an unknown objective function $f : \mathcal{X} \to \mathbb{R}$ of interest in the RKHS of a known kernel k. This is a very gen-

eral assumption, since the RKHS of common kernels can approximate almost all continuous functions on compact subsets of the Euclidean space (Srinivas et al., 2010). Our aim is to find an input x^* which maximises the unknown function f, i.e. $x^* \in \arg\max_{x\in\mathcal{X}} f(x)$. In order to do this, we can sequentially query f at a chosen sequence of observation points $X_t \in \mathcal{X}$, $t = 1, 2, \ldots$, and receive the noisy feedback $y_t = f(X_t) + \epsilon_t$ where ϵ_t is a sequence of independently distributed sub-Gaussian noise with zero mean.

We measure the performance of any procedure for this problem in terms of its *regret*. The regret is defined as

$$\mathcal{R}(T) = \sum_{t=1}^{T} f(x^*) - f(X_t) \tag{3}$$

and represents the cumulative amount that we loose by querying f at $X_1, \ldots, X_T$ rather than at an unknown maxima $x^* \in \arg\max_{x\in\mathcal{X}} f(x)$.

In order to make the problem tractable, we make the following assumptions which are common in the literature.

Assumption 3.1. The RKHS norm of the objective function f is bounded, i.e., $\|f\|_{\mathcal{H}_k} \leq C_k$, for some $C_k > 0$, where the notation $\|\cdot\|^2_{\mathcal{H}_k} = \langle\cdot,\cdot\rangle_{\mathcal{H}_k}$ is used for the RKHS norm.

Assumption 3.2. The observation noise terms ϵ_t are independent σ-sub-Gaussian random variables with zero mean. That is, for all $t \in \mathbb{N}$, for all $\eta \in \mathbb{R}$, and for some $\sigma > 0$, the moment generating function of ϵ_t satisfies $\mathbb{E}[\exp(\eta\epsilon_t)] \leq \exp(\frac{\eta^2\sigma^2}{2})$.

Delayed feedback. In this work, we are interested in the kernel bandit problem in a setting with stochastically delayed feedback. In this setting, at each time t, as well as generating the observation y_t, the environment also independently generates a stochastic delay $\tau_t \geq 0$. The learner receives the feedback for the decision made at time t at the observation time $t+\tau_t$. We assume that $\tau_1, \ldots, \tau_T$ represent a sequence of independent and identically distributed sub-exponential random variables. For simplicity, we assume that the τ_t random variables are discrete, although it is easy to extend our results to continuous delays by considering the events that $\tau_t + t \leq s$ for all $s \geq t$. Our assumption on the delay distribution is formally stated below, and is the same as that commonly made in the literature (Zhou et al., 2019; Howson et al., 2022).

Assumption 3.3. The delays τ_t are i.i.d. sub-exponential random variables. That is, for all $t \in \mathbb{N}$, for some $\xi, b > 0$, for all $|\eta| \leq \frac{1}{b}$, the moment generating function of τ_t satisfies $\mathbb{E}\left[\exp\left(\eta(\tau_t - \mathbb{E}[\tau_t])\right)\right] \leq \exp(\frac{\eta^2\xi^2}{2})$.

4. Confidence Bounds

Kernel based regression provides powerful predictors and uncertainty estimates. Those could be used to form confidence intervals for the unknown objective function f, that are a crucial building block in developing algorithms for the kernel based bandit problem. Consider a set of observation points $\mathbf{X}_t = \{X_1, \ldots, X_t\}$, and the corresponding vector of observation values $\mathbf{y}_t = [y_1, \ldots, y_t]^\top$. Recall $\mathbf{K}_{\mathbf{X}_t} = [k(X_s, X_{s'})]^t_{s,s'=1}$ the positive definite kernel matrix resulting from the pairwise kernel values between the observation points. We have the following predictor and uncertainty estimate for f, which can be interpreted as the mean and variance of a surrogate GP model for f with kernel k, respectively (e.g., see, Rasmussen and Williams, 2006; Kanagawa et al., 2018)

$$\begin{aligned}\mu_{\mathbf{X}_t,\mathbf{y}_t}(x) &= \mathbf{k}^\top_{\mathbf{X}_t}(x)(\mathbf{K}_{\mathbf{X}_t} + \lambda^2\mathbf{I}_t)^{-1}\mathbf{y}_t \\ \sigma^2_{\mathbf{X}_t}(x) &= k(x,x) \\ &\quad - \mathbf{k}^\top_{\mathbf{X}_t}(x)(\mathbf{K}_{\mathbf{X}_t} + \lambda^2\mathbf{I}_t)^{-1}\mathbf{k}_{\mathbf{X}_t}(x),\end{aligned} \tag{4}$$

where $\mathbf{k}_{\mathbf{X}_t}(x) = [k(x_1, x), \ldots, k(x_t, x)]^\top$ is the kernel values vector between observation points and the point of interest and $\lambda > 0$ is a free parameter.

Equipped with the above predictor and uncertainty estimate, we have the following confidence bounds.

Lemma 4.1. *[(Vakili et al., 2021a)] Consider the predictor and uncertainty estimate given in (4). Assume the observation set $\mathbf{X}_t$ is selected independent of the observation noise ϵ_s, $s = 1, \ldots, t$. Under Assumptions 3.1 and 3.2, the following inequalities each hold with probability $1 - \delta$ for any fixed $x \in \mathcal{X}$,*

$$\begin{aligned} f(x) &\leq \mu_{\mathbf{X}_t,\mathbf{y}_t}(x) + \beta(\delta)\sigma_{\mathbf{X}_t}(x), \\ f(x) &\geq \mu_{\mathbf{X}_t,\mathbf{y}_t}(x) - \beta(\delta)\sigma_{\mathbf{X}_t}(x), \end{aligned}$$

where $\beta(\delta) = C_k + \frac{\sigma}{\lambda}\sqrt{2\log(\frac{1}{\delta})}$, C_k and σ are the parameters given in Assumptions 3.1 and 3.2, and λ is the parameter in (4).

If the domain $\mathcal{X}$ is finite, then uniform confidence bounds readily follow from this result via a union bound, and $\frac{\delta}{|\mathcal{X}|}$ can be substituted for δ. For continuous domains, a technical discretization argument can be used to extend Lemma 4.1 to a uniform bound under the following continuity assumption.

Assumption 4.2. For each $t \in \mathbb{N}$, there exists a discretization $\mathbb{X}$ of $\mathcal{X}$ such that, for any $f \in \mathcal{H}_k$ with $\|f\|_{\mathcal{H}_k} \leq C_k$, we have $f(x) - f([x]) \leq \frac{1}{t}$, where $[x] = \mathrm{argmin}_{x'\in\mathbb{X}}||x' - x||_{l^2}$ is the closest point in $\mathbb{X}$ to x, and $|\mathbb{X}| \leq cC_k^d t^d$, where c is a constant independent of t and C_k.

Assumption 4.2 is a mild and technical assumption that holds for typical kernels such as SE and Matérn with $\nu >$

1 (Srinivas et al., 2010; Chowdhury and Gopalan, 2017; Vakili et al., 2021a).

Lemma 4.3 (A special case of Corollary 3.7 in Vakili et al. (2022))**.** *Under the setting and assumptions of Lemma 4.1, under Assumption 4.2, the following inequalities each hold uniformly in $x \in \mathcal{X}$ with probability at least $1-\delta$,*

$$f(x) \leq \mu_{\mathbf{X}_t,\mathbf{y}_t}(x) + \frac{2}{t} + \beta'_\delta(t)(\sigma_{\mathbf{X}_t}(x) + \frac{2}{\sqrt{t}}),$$
$$f(x) \geq \mu_{\mathbf{X}_t,\mathbf{y}_t}(x) - \frac{2}{t} - \beta'_\delta(t)(\sigma_{\mathbf{X}_t}(x) + \frac{2}{\sqrt{t}}),$$

where $\beta'_\delta(t) = \beta(\frac{\delta}{2\Gamma_t})$, $\Gamma_t = c(\tilde{C}_k(\frac{\delta}{2}))^d t^d$, $\tilde{C}_k(\delta) = C_k + \frac{\max\{\sigma, k_{\max}\}\sqrt{t}}{\lambda}\sqrt{2\log(\frac{2t}{\delta})}$, $k_{\max} = \max_{x\in\mathcal{X}} k(x,x)$.

Lemma 4.3 uses a high probability bound $\tilde{C}_k(\delta)$ on the RKHS norm of $\mu_{\mathbf{X}_t,\mathbf{y}_t}$ (that is a random variable due to the random noise), together with applying the continuity assumptions to $\mu_{\mathbf{X}_t,\mathbf{y}_t}$ and $\sigma_{\mathbf{X}_t}$ to construct the confidence bounds. Effectively, the multiplier of the width of the confidence intervals, $\beta'_\delta(t)$, scales as

$$\beta'_\delta(t) = \mathcal{O}\left(C_k + \frac{\sigma}{\lambda}\sqrt{d\log(\frac{tC_k}{\delta})}\right). \tag{5}$$

In the kernel bandit with delayed feedback setting, some observations may not be available due to delayed feedback. Let $\tilde{\mathbf{X}}_t = \{X_s \in \mathbf{X}_t : s + \tau_s \leq t\}$ be the set of observation points for which the feedback has arrived by time t, and note that this is a random set of observations due to the stochastic delays. Simplifying the notation, we use μ_t, σ_t for the predictor and uncertainty estimate using $\mathbf{X}_t, \mathbf{y}_t$, and $\tilde{\mu}_t, \tilde{\sigma}_t$ for the predictor and uncertainty estimate using $\tilde{\mathbf{X}}_t, \tilde{\mathbf{y}}_t$, where $\tilde{\mathbf{y}}_t = [f(X_s) + \epsilon_s]^\top_{X_s \in \tilde{\mathbf{X}}_t}$ is the vector of available observations. We note that in the delayed feedback setting, we can compute σ_t in addition to $\tilde{\mu}_t, \tilde{\sigma}_t$ since this only depends on the chosen inputs, not the observations. On the other hand, μ_t cannot be computed since it depends on some of the missing feedback. All three available statistics $\tilde{\mu}_t, \tilde{\sigma}_t, \sigma_t$ are used in designing our algorithm, BPE-Delay.

5. Batch Pure Exploration with Delays

In this section, we describe our proposed algorithm: Batch Pure Exploration with Delays (BPE-Delay). The algorithm proceeds in rounds $r = 1, 2, \ldots, R$, where R is the total number of rounds. Each round consists of t_r time steps so that $\sum_{r=1}^R t_r = T$. During each round, the observation points are selected based on a maximum uncertainty acquisition function (defined in (7)). At the end of each round, confidence intervals are used to remove the points which are unlikely to be the maximiser of f.

The length t_r of round r is increasing in r and is chosen carefully to balance the exploration needed in each round, taking into account the stochastically delayed feedback, and the number of rounds. In particular we set $t_r = \lceil q_r + u_T(\delta)\rceil$ where $u_T(\delta)$ is a $1-\delta$ upper bound on the delay random variable, and $q_r = \sqrt{q_{r-1}T}$ is determined recursively, for $r \geq 1$, initialized at $q_0 = 1$. The delay related quantity is set to $u_T(\delta) = \mathbb{E}[\tau] + \psi_T(\frac{\delta}{2})$ with

$$\psi_t(\delta) = \min\left\{\sqrt{2\xi^2\log(\frac{3t}{2\delta})},\ 2b\log(\frac{3t}{2\delta})\right\}, \tag{6}$$

where ξ and b are the parameters specified in Assumption 3.3. In the analysis in Section 6, we will see that $u_T(\delta)$ is a $1-\frac{\delta}{2}$ upper confidence bound on the delay random variable. It turns out that this choice of t_r depending on $u_T(\delta)$ is crucial in enabling us to improve the delay dependence of our algorithm. If we know there is no delay in the observations, we can set $u_T(\delta)$ to zero. The length of the last round can also be easily adjusted to ensure $\sum_{t=1}^R t_r = T$.

BPE-Delay maintains a set $\mathcal{X}_r$ of potential maximisers of f. This set is recursively pruned from $\mathcal{X}_{r-1}$ using confidence intervals around each $x \in \mathcal{X}_{r-1}$ to remove those that are sub-optimal with high probability. We start with the full input space, $\mathcal{X}_0 = \mathcal{X}$. During each round r, the k-th observation point $X_{k,r}$ is selected from $\mathcal{X}_r$ based on a maximum uncertainty acquisition:

$$X_{k,r} \leftarrow \arg\max_{x\in\mathcal{X}_r} \sigma_{k-1,r}(x), \tag{7}$$

where $\sigma_{k,r}(\cdot)$ denotes the uncertainty estimate in (4) calculated using the points $\mathbf{X}_{k,r} = \{X_{1,r}, X_{2,r}, \ldots, X_{k,r}\}$. BPE-Delay uses only the inputs chosen in round r so far, to calculate the acquisition function and choose the remainder of the points in round r based on the uncertainty estimates. While using entire past observations may be effective, establishing corresponding regret bounds becomes difficult. The reason is that creating such dependency among observation points invalidates tight confidence intervals given in Lemma 4.1. Nonetheless, Li and Scarlett (2022) showed in experiments that discarding the data collected in previous rounds, although may seem wasteful, does not significantly harm the performance. From the definition of the uncertainty estimate given in (4), we can see that it does not depend on the observation values. Thus, $\sigma^2_{k,r}(\cdot)$ is well defined, despite not all observation values being available due to delay. We use double index $r = 1, 2, \ldots$, $k = 1, 2, \ldots, t_r$ for clarity of presentation. The observation points however can be indexed using $t = 1, 2, \ldots$ by concatenating the observation points in all rounds (see Algorithm 1).

To contrast the statistics derived from the entire set of observations in round r —ignoring the delay—, and the ones using available observations —considering the delay—, we recall the notations $\tilde{\mu}_{k,r}$ and $\tilde{\sigma}_{k,r}$ for the statistics using just the available observations. Specifically, $\tilde{\mu}_{k,r}$ and $\tilde{\sigma}_{k,r}$

Algorithm 1 Batch Pure Exploration with Delays (BPE-Delay)

Input: Action set $\mathcal{X}$, number of steps T;
Initialize: $\mathcal{X}_1 \leftarrow \mathcal{X}$, $q_0 = 1$, $t = 1$;
# Algorithm proceeds in rounds $r = 1, 2, \ldots$.
for $r = 1, 2, \ldots,$ until $t \leq T$ **do**
  $q_r = \lceil \sqrt{T q_{r-1}} \rceil$ # q_r is used to determine t_r
  $t_r = \lceil q_r + u_T(\delta) \rceil$ # t_r is the length of round r
  for $k = 1, \ldots, t_r$ **do**
    # the points in each round are chosen by a maximum uncertainty acquisition function
    $X_{k,r} \leftarrow \arg\max_{x \in \mathcal{X}_r} \sigma_{k-1,r}(x)$
    $X_t \leftarrow X_{k,r}$
    $t = t + 1$
  end for
  # Remove the input points which are unlikely to be the maximiser of f, using confidence intervals
  $\mathcal{X}_{r+1} \leftarrow \{x \in \mathcal{X}_r : U_{r,\delta}(x) \geq L_{r,\delta}(z), \forall z \in \mathcal{X}_r\}$
end for

are the prediction and uncertainty estimate after selecting k observation points in round r using the set $\tilde{\mathbf{X}}_{k,r} = \{X_{s,r} \in \mathbf{X}_{k,r} : s + \tau_{s,r} \leq k\}$ of points selected in round r, with available observations $\tilde{\mathbf{y}}_{k,r} = [y = f(X_{s,r}) + \epsilon_{s,r}]_{X_{s,r} \in \tilde{\mathbf{X}}_{k,r}}$ in round r.

At the end of each round r, using the available observations in round r, we create upper and lower confidence bounds on $f(\cdot)$, $U_r(\cdot)$ and $L_r(\cdot)$ respectively. These are used to remove points which are unlikely to be a maximiser of f. In particular, if there exists a point $z \in \mathcal{X}_r$ for which the lower confidence bound is larger than the upper confidence bound for $x \in \mathcal{X}_r$, then x is unlikely to be the maximiser of f so we can remove it. The confidence interval widths are selected in a way that all the confidence intervals used in the algorithm hold true with high probability (see Theorems 6.1 and 6.2 for details).

We emphasize that while we use $\sigma_{k,r}$ for selecting the observation points within each round, $\tilde{\mu}_{k,r}$ and $\tilde{\sigma}_{k,r}$ are used for forming the confidence intervals at the end of each round. Using $\sigma_{k,r}$ avoids selecting repetitive points due to unavailable feedback, while the valid confidence intervals can only be formed using $\tilde{\mu}_{k,r}$ and $\tilde{\sigma}_{k,r}$.

The complete pseudo-code for the BPE-Delay algorithm is given in Algorithm 1.

6. Analysis

In this section, we provide the analysis of the BPE-Delay algorithm. Specifically, we prove a high probability $\tilde{\mathcal{O}}(\sqrt{T\Gamma_k(T)} + \mathbb{E}[\tau])$ bound on its regret. This regret bound completely decouples the effect of the delay from the dominant time dependent regret term. We note that the additional regret due to delay in our result is significantly smaller than the existing work and matches what is seen in the simpler K-armed bandit problem (Joulani et al., 2013).

We consider two cases of finite and continuous $\mathcal{X}$ separately, due to minor differences in the confidence intervals which lead to mild differences in the regret bounds. In particular, when $\mathcal{X}$ is finite, the regret bound scales with $\mathcal{O}(\sqrt{\log(|\mathcal{X}|)})$. In the case of a continuous and compact $\mathcal{X} \subset \mathbb{R}^d$, under Assumption 4.2, we prove a similar regret bound which scales with an $\mathcal{O}(\sqrt{d})$ factor instead of $\mathcal{O}(\sqrt{\log(|\mathcal{X}|)})$. While in the kernel bandit setting, $\sqrt{d}$ is typically hidden in the $\mathcal{O}$ notation as a constant not growing with T, in the linear bandit setting, the dependency on d should be pronounced and is important in evaluating the algorithms.

Theorem 6.1. *Consider the kernel based bandit problem with delayed feedback described in Section 3. When $\mathcal{X}$ is finite, set the confidence intervals in BPE-Delay (Algorithm 1) to*

$$U_{r,\delta}(x) = \tilde{\mu}_{t_r,r}(x) + \tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x),$$
$$L_{r,\delta}(x) = \tilde{\mu}_{t_r,r}(x) - \tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x), \tag{8}$$

with $\tilde{\beta}(\delta) = C_k + \frac{\sigma}{\lambda}\sqrt{2\log(\frac{4R|\mathcal{X}|}{\delta})}$. Under Assumptions 3.1, 3.2, 3.3, with probability at least $1-\delta$, the regret of BPE-Delay is bounded by

$$\mathcal{R}(T) = \mathcal{O}\left(\tilde{\beta}(\delta)\log\log(T)\sqrt{T\Gamma_k(T)} + \mathbb{E}[\tau]\right). \tag{9}$$

Note that with a finite $\mathcal{X}$, $\tilde{\beta}(\delta) = \mathcal{O}(1)$, not depending on problem parameters such as d and T. In the next theorem, we bound the regret when $\mathcal{X}$ is not finite, but a compact subset of $\mathbb{R}^d$.

Theorem 6.2. *Consider the kernel based bandit problem with delayed feedback described in Section 3, under Assumptions 3.1, 3.2, 3.3 and 4.2. Set the confidence intervals in BPE-Delay (Algorithm 1) to*

$$U_{r,\delta}(x) = \tilde{\mu}_{t_r,r}(x) + \frac{2}{u_r} + \tilde{\beta}'_r(\delta)(\tilde{\sigma}_{t_r,r}(x) + \frac{2}{\sqrt{u_r}}),$$
$$U_{r,\delta}(x) = \tilde{\mu}_{t_r,r}(x) + \frac{2}{u_r} + \tilde{\beta}'_r(\delta)(\tilde{\sigma}_{t_r,r}(x) + \frac{2}{\sqrt{u_r}}), \tag{10}$$

with $\tilde{\beta}'_r(\delta) = \beta'_{u_r}(\frac{\delta}{4R})$ and β'_t given in Lemma 4.3. Then, with probability at least $1-\delta$, the regret of BPE-Delay is bounded by,

$$\mathcal{R}(T) = \mathcal{O}\left(\tilde{\beta}'_R(\delta)\log\log(T)\sqrt{T\Gamma_k(T)} + \mathbb{E}[\tau]\right). \tag{11}$$

From Lemma 4.3, we can see that $\tilde{\beta}'_R(\delta) = \mathcal{O}(C_k + \frac{\sigma}{\lambda}\sqrt{d\log(\frac{TRC_k}{\delta})})$. Thus the regret bound scales with $\mathcal{O}(\sqrt{d})$ that becomes particularly important when applied to linear bandits.

Proof Sketch. By the choice of size t_r of each round, the number R of rounds is bounded as $R = \mathcal{O}(\log\log(T))$. We first show that the uncertainty estimate using t_r observations $\tilde{\sigma}_{t_r,r}$ sufficiently shrinks. Using the confidence intervals, we prove that i) the maximiser x^* is not removed during any round; i.e. $x^* \in \mathcal{X}_r$, for all r, and ii) the instantaneous regret incurred by selecting each point $X_{k,r} \in \mathcal{X}_r$ is bounded. The bound on $f(x^*) - f(X_{k,r})$ is based on the amount of exploration in previous round, taking into account the delay. Using the bound on instantaneous regret and the size t_r of round r we bound the regret in round r. A detailed proof is provided in the appendix. □

Theorem 6.1 can be specialized for many commonly used kernels including Matérn and SE, as well as for the special case of linear kernels.

Corollary 6.3. *Under the setting of Theorem 6.2, the following hold with probability at least* $1-\delta$*,*

- *In the case of Matérn kernel with smoothness parameter* ν*,*

$$\mathcal{R}(T) = \tilde{\mathcal{O}}\left(T^{\frac{\nu+d}{2\nu+d}} + \mathbb{E}[\tau]\right). \tag{12}$$

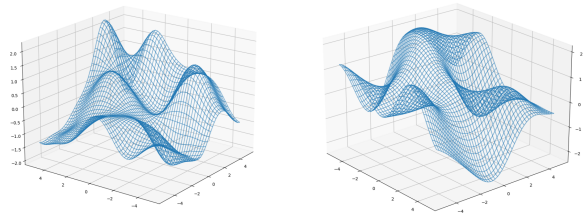
- *In the case of SE kernel,*

$$\mathcal{R}(T) = \mathcal{O}\left(\sqrt{T}(\log(T))^{d+1}\log\log(T) + \mathbb{E}[\tau]\right). \tag{13}$$

- *In the case of linear kernel,*

$$\mathcal{R}(T) = \mathcal{O}\left(d\sqrt{T}\log(T)\log\log(T) + \mathbb{E}[\tau]\right). \tag{14}$$

Comparing to the lower bound in the case of linear bandit (Lattimore and Szepesvári, 2020), and the lower bounds for kernel bandit in the case of SE and Matérn kernels (Scarlett et al., 2017), the regret bounds shown in Corollary 6.3 are order optimal, up to logarithmic factors and the additive delay penalty which is independent of all other problem parameters. We thus reduced the delay related term in the regret bounds to only $\mathbb{E}[\tau]$ without affecting the dominant term in the regret bound. For comparison, in the (generalized) linear setting the best known bound on the delay related regret was $\mathcal{O}(d^{\frac{3}{2}}\mathbb{E}[\tau])$ (Howson et al., 2022); that is improved with a $d^{\frac{3}{2}}$ factor in our results. In the kernel bandit setting, we also improved the $\mathcal{O}(\Gamma_T\mathbb{E}[\tau])$ delay related regret (Verma et al., 2022) to only $\mathcal{O}(\mathbb{E}[\tau])$. This represents a significant improvement considering that Γ_T can become arbitrarily close to linear in T, in the case of a Matérn kernel with a small ν and large d. This is in addition to the improvement in the first term in regret from $\tilde{\mathcal{O}}(\Gamma_T\sqrt{T})$ to $\tilde{\mathcal{O}}(\sqrt{\Gamma_T T})$.



(a) f_1 (b) f_2

Figure 1. Objective functions used in our experiments.

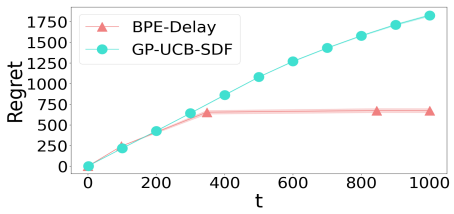
7. Experiments

Following our theoretical analysis, we provide numerical experiments on the performance of BPE-Delay and compare it to the GP-UCB-SDF (Verma et al., 2022). We test the algorithms on the objective functions f_1 and f_2 shown in Figure 1. These functions are generated by fitting a kernel based model to points randomly generated from a multivariate Gaussian. This is a common technique to create RKHS elements (e.g., see Chowdhury and Gopalan, 2017; Vakili et al., 2021a; Li and Scarlett, 2022). We use a SE kernel with a length scale parameter $l = 0.8$ for f_1 and $l = 1.0$ for f_2 in order to generate these objective functions. The learner can then choose from $|\mathcal{X}| = 2500$ points over a uniform 50×50 grid. The sampling noise is zero mean Gaussian with standard deviation $\sigma = 0.02$. The stochastic delay in the feedback is generated from a Poisson distribution with parameter λ. The calculation of ψ_t for BPE-Delay uses $\xi = 9$ and $b = 1$ given in Assumption 4.2. The cumulative regret curves are the average of 10 independent experiments. The error bars indicate half a standard deviation. The code for these experiments is provided in a GitHub repository.[2]

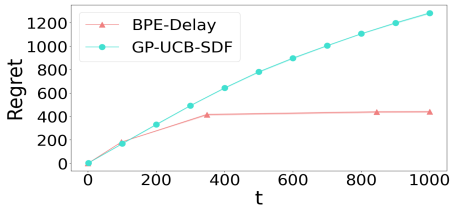
In Figure 2, we compare the regret performance of BPE-Delay with GP-UCB-SDF introduced in the most related work in the same setting as ours (Verma et al., 2022). The figure shows a significant improvement in the regret performance using BPE-Delay in both cases. In this experiment, the delays are Poisson distributed with $\lambda = 50$.

In Figures 3 and 4, we show the effect of delay for these two algorithms. Specifically, we vary the Poisson delay parameter as $\lambda = 0, 25$ and 50, while maintaining the rest of parameters as the previous experiment. BPE-Delay shows a linear excess in the regret with the expected delay. GP-UCB-SDF on the other hand shows dramatic jump in regret

[2] https://github.com/svakili89/delayed_kernel_bandit

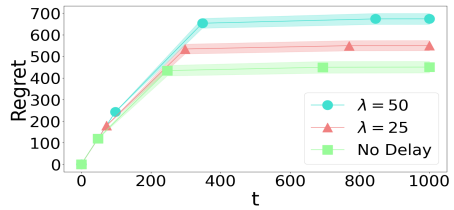


(a) f_1

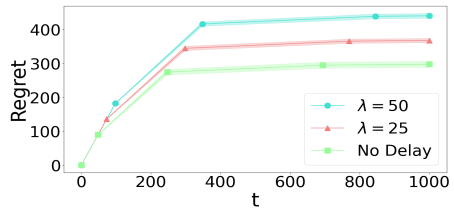


(b) f_2

Figure 2. Comparing regret performance of BPE-Delay and GP-UCB-SDF.

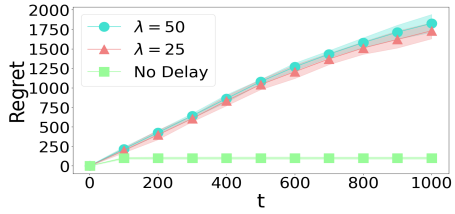


(a) f_1

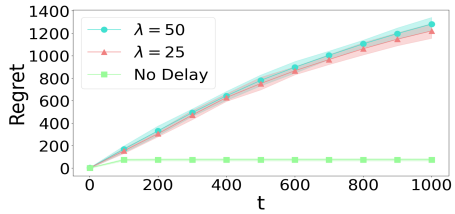


(b) f_2

Figure 3. Regret performance of BPE-Delay for varying delay Poisson parameters.

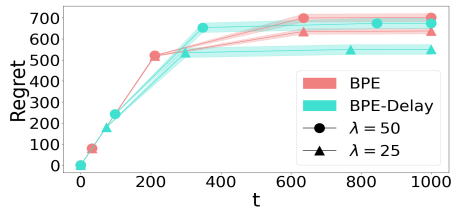


(a) f_1

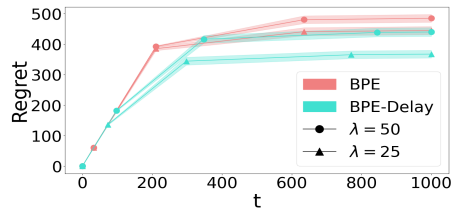


(b) f_2

Figure 4. Regret performance of GP-UCB-SDF for varying delay Poisson parameters.



(a) f_1



(b) f_2

Figure 5. Comparing the regret performance of BPE and BPE-Delay in the delayed feedback setting.

even with moderate delay values. We also compare the performance of the BPE-Delay and BPE (Li and Scarlett, 2022) which was designed for non-delayed settings. As we can see from Figure 5, BPE-Delay naturally performs better than BPE in a delayed setting.

8. Conclusion

There has been a great attention paid to kernel bandits due to their numerous applications in machine learning, and academic and industrial experimental design. In many real world scenarios, e.g., recommender systems, the feedback from the decisions are not immediately available, naturally leading to the formulation of kernel bandits with stochastically delayed feedback. For this setting, we proposed BPE-Delay which is able to efficiently deal with the delayed feedback. We showed an order optimal regret bound for BPE-Delay with a very small excess in regret due to the delayed feedback, significantly improving upon the existing work in the same setting. We also show that these theoretical improvements are maintained in several simulation studies. In addition, when applied to the special case of linear kernels, our theoretical regret bounds improve the delay related penalty by a factor of $d^{\frac{3}{2}}$ compared to the state of the art. In all cases, our results are the first to show that the additive delay penalty only depends on $\mathbb{E}[\tau]$. This extends what is known for the simpler K-armed bandit setting to the much more realistic kernel bandit setting.

References

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

D. Calandriello, L. Carratino, A. Lazaric, M. Valko, and L. Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, 2019.

S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853, 2017.

S. R. Chowdhury and A. Gopalan. On batch bayesian optimization. *arXiv preprint arXiv:1911.01032*, 2019.

T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

E. A. Daxberger and B. K. H. Low. Distributed batch gaussian process optimization. In *International conference on machine learning*, pages 951–960. PMLR, 2017.

T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:3873–3923, 2014.

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178, 2011.

B. Howson, C. Pike-Burke, and S. Filippi. Delayed feedback in generalised linear bandits revisited. *arXiv preprint arXiv:2207.10786*, 2022.

P. Joulani, A. Gyorgy, and C. Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Z. Li and J. Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, pages 92–107. PMLR, 2022.

T. Mandel, Y.-E. Liu, E. Brunskill, and Z. Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

C. E. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

S. Salgia, S. Vakili, and Q. Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. *Advances in Neural Information Processing Systems*, 34:28836–28847, 2021.

J. Scarlett, I. Bogunovic, and V. Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742, 2017.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.

S. Vakili, N. Bouziani, S. Jalali, A. Bernacchia, and D.-s. Shiu. Optimal order simple regret for Gaussian process bandits. In *Conference on Neural Information Processing Systems*, 2021a.

S. Vakili, M. Bromberg, J. Garcia, D.-s. Shiu, and A. Bernacchia. Uniform generalization bounds for overparameterized neural networks. *arXiv preprint arXiv:2109.06099*, 2021b.

S. Vakili, K. Khezeli, and V. Picheny. On information gain and regret bounds in Gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90, 2021c.

S. Vakili, H. Moss, A. Artemev, V. Dutordoir, and V. Picheny. Scalable thompson sampling using sparse gaussian process models. *Advances in Neural Information Processing Systems*, 34:5631–5643, 2021d.

S. Vakili, J. Scarlett, and T. Javidi. Open problem: Tight online confidence intervals for rkhs elements. In *Conference on Learning Theory*, pages 4647–4652. PMLR, 2021e.

S. Vakili, J. Scarlett, D.-s. Shiu, and A. Bernacchia. Improved convergence rates for sparse approximation methods in kernel-based learning. In *International Conference on Machine Learning*. PMLR, 2022.

M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, page 654–663, 2013.

A. Verma, Z. Dai, and B. K. H. Low. Bayesian optimization under stochastic delayed feedback. In *International Conference on Machine Learning*, pages 22145–22167. PMLR, 2022.

C. Vernade, O. Cappé, and V. Perchet. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

C. Vernade, A. Carpentier, T. Lattimore, G. Zappella, B. Ermis, and M. Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR, 2020.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17 (9):2077–2098, 2005.

Z. Zhou, R. Xu, and J. Blanchet. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Proofs

In this section, we provide detailed proofs for Theorems 6.1, 6.2 and Corollary 6.3. We structure the proof of theorems as follows. We first overview the kernel based confidence intervals, the bounds on the delay and the number of rounds. After formalizing these bounds, we proceed with the proof in three steps. In the first step, we bound the maximum uncertainty at any point $x \in \mathcal{X}_{r+1}$ at the end of each round r. In the second step, we bound the instantaneous regret for selecting a point in each round. In the third step, we bound the total regret. At the end of the proof, we discuss Corollary 6.3.

Confidence Intervals. Recall the confidence intervals used in the BPE-Delay algorithm to update the set $\mathcal{X}_r$ of potential maximisers in round r. Using Lemmas 4.1 and 4.3 and a probability union bound over rounds, all confidence intervals used in the algorithm are valid with probability at least $1 - \frac{\delta}{2}$. Specifically, let us define the event

$$\mathcal{E} = \{L_{r,\delta}(x) \leq f(x) \leq U_{r,\delta}(x), \forall r = 1, 2, \ldots, R, \forall x \in \mathcal{X}_r\}. \tag{15}$$

Then, we have $\Pr[\mathcal{E}] \geq 1 - \frac{\delta}{2}$.

Delay. Recall the definition of $\psi_t(\delta)$ given in (6). We have the following concentration for sub-exponential random variables.

Lemma A.1 ((Howson et al., 2022))**.** *Let $\tau_t - \mathbb{E}[\tau_t]$, $t = 1, 2, \ldots,$ be i.i.d. sub-exponential random variables with parameters ξ and b as specified in Assumption 3.3. Then,*

$$\Pr\left[\forall t \geq 1, \tau_t \leq \mathbb{E}[\tau] + \psi_t(\delta)\right] \geq 1 - \delta. \tag{16}$$

Let $\tau_{k,r}$ denote the random delay for the k-th observation in round r of BPE-Delay. Let us define the event

$$\mathcal{E}' = \{\forall r, \forall k \leq q_r : \tau_{k,r} \leq u_T(\delta)\}. \tag{17}$$

Using Lemma A.1 and definition of u_T, we have $\Pr[\mathcal{E}'] \geq 1 - \frac{\delta}{2}$.

We thus have the probability that both $\mathcal{E}$ and $\mathcal{E}'$ hold true $\Pr[\mathcal{E} \cap \mathcal{E}'] \geq 1 - \delta$.

We condition the rest of the proof on $\mathcal{E} \cap \mathcal{E}'$.

The number of rounds. The number R of rounds in BPE-Delay is bounded in the following lemma.

Lemma A.2. *Recall the choice of round lengths t_r in BPE-Delay. We have $R = \mathcal{O}(\log\log(T))$.*

The proof follows from similar steps as in the proof of Proposition 1 in (Li and Scarlett, 2022).

We now proceed to the main steps in the proof of theorems.

Step 1 (Maximum uncertainty at the end of each round). In each round r, the sum of uncertainties using all observations (including the ones with delayed feedback) at the end of q_r observations can be bounded using $\Gamma_k(q_r)$.

Lemma A.3. *[(Srinivas et al., 2010)] Recall definition of $\Gamma_k(t)$ given in (2), for a set of t observation $\mathbf{X}_t$, we have*

$$\sum_{s=1}^{t} \sigma^2_{\mathbf{X}_{s-1}}(X_s) \leq c_1 \Gamma_k(t), \tag{18}$$

where $c_1 = \frac{2}{\log(1+\frac{1}{\lambda^2})}$.

Thus, considering the observation points in round r, we have

$$\sum_{k=1}^{q_r} \sigma^2_{k-1,r}(X_{k,r}) \leq c_1 \Gamma_k(q_r), \tag{19}$$

where c_1 is the constant given in Lemma A.3. By the selection rule of observation points based on maximum uncertainty acquisition (7), we have, for all $x \in \mathcal{X}_r$, $1 \leq k \leq q_r$

$$\sigma_{k-1,r}(x) \leq \sigma_{k-1,r}(X_{k,r}). \tag{20}$$

In addition, by positive definiteness of the kernel matrix, conditioning on a bigger set of observations reduces the uncertainty (see (4)). Thus,

$$\sigma_{q_r,r}(x) \leq \sigma_{k-1,r}(X_{k,r}). \tag{21}$$

Combining the last inequality with (19), we obtain

$$\sum_{k=1}^{q_r} \sigma^2_{q_r,r}(x) \leq c_1 \Gamma_k(q_r), \tag{22}$$

that implies

$$\sigma^2_{q_r,r}(x) \leq \frac{c_1 \Gamma_k(q_r)}{q_r}. \tag{23}$$

Recall the event $\mathcal{E}'$ given in (17). Under $\mathcal{E}'$, $\tilde{\sigma}^2_{t_r,r}$ is conditioned on a larger set of observations compared to $\sigma^2_{q_r,r}$. The positive definiteness of the kernel matrix implies that conditioning on a larger set of observations reduces the uncertainty. Thus, we have $\tilde{\sigma}^2_{t_r,r}(x) \leq \sigma^2_{q_r,r}(x), \forall x \in \mathcal{X}_r$. As a result,

$$\tilde{\sigma}^2_{t_r,r}(x) \leq \frac{c_1 \Gamma_k(q_r)}{q_r}, \quad \forall x \in \mathcal{X}_r. \tag{24}$$

Step 2 (Instantaneous regret). We can use the confidence intervals for RKHS elements given in Lemmas 4.1 and 4.3 to bound the instantaneous regret as follows. Recall the event $\mathcal{E}$ defined in (15).

We first note that for all $x \in \mathcal{X}_r$,

$$\begin{aligned} U_{r,\delta}(x^*) &\geq f(x^*) \\ &\geq f(x) \\ &\geq L_{r,\delta}(x). \end{aligned}$$

Thus, $U_{r,\delta}(x^*) \geq \max_{x \in \mathcal{X}_r} L_{r,\delta}(x)$ for all r, and x^* will not be eliminated during any round of BPE-Delay. Therefore $x^* \in \mathcal{X}_r$, for all r.

We now proceed to bounding the instantaneous regret under two different settings of finite $\mathcal{X}$ and continuous $\mathcal{X}$, separately. The two cases correspond to Theorems 6.1 and 6.2, respectively.

When $\mathcal{X}$ is finite (The setting of Theorem 6.1):

For all $x \in \mathcal{X}_{r+1}$,

$$\begin{aligned} f(x^*) - f(x) &\leq U_{r,\delta}(x^*) - L_{r,\delta}(x) \\ &= \tilde{\mu}_{t_r,r}(x^*) + \tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x^*) - \tilde{\mu}_{t_r,r}(x) - \tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x^*) \\ &= L_{r,\delta}(x^*) - U_{r,\delta}(x) + 2\tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x^*) + 2\tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x) \\ &\leq 2\tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x^*) + 2\tilde{\beta}(\delta)\tilde{\sigma}_{t_r,r}(x). \end{aligned} \tag{25}$$

The first and second equality follow from the choice of the confidence intervals in Theorem 6.1. The last line follows from the choice of $\mathcal{X}_{r+1}$ that ensures for all $x, z \in \mathcal{X}_{r+1}$, $L_{r,\delta}(z) \leq U_{r,\delta}(x)$.

Combining with the bound on $\tilde{\sigma}_{t_r,r}(x)$ obtained in the previous step, we bound the instantaneous regret for all $x \in \mathcal{X}_{r+1}$ as follows.

$$f(x^*) - f(x) \leq 4\tilde{\beta}(\delta)\sqrt{\frac{c_1 \Gamma_k(q_r)}{q_r}}. \tag{26}$$

When $\mathcal{X}$ is continuous (The setting of Theorem 6.2):

Following similar steps as in the previous case, for all $x \in \mathcal{X}_{r+1}$,

$$\begin{aligned} f(x^*) - f(x) &\leq U_{r,\delta}(x^*) - L_{r,\delta}(x) \\ &\leq \tilde{\mu}_{t_r,r}(x^*) + \frac{2}{q_r} + \tilde{\beta}'_r(\delta)(\tilde{\sigma}_{t_r,r}(x^*) + \frac{2}{\sqrt{q_r}}) \\ &\quad - \tilde{\mu}_{t_r,r}(x) + \frac{2}{q_r} + \tilde{\beta}'_r(\delta)(\tilde{\sigma}_{t_r,r}(x) + \frac{2}{\sqrt{q_r}}) \\ &= L_{r,\delta}(x^*) - U_{r,\delta}(x) \\ &\quad + \frac{8}{q_r} + 2\tilde{\beta}'_r(\delta)\left(\tilde{\sigma}_{t_r,r}(x^*) + \tilde{\sigma}_{t_r,r}(x) + \frac{4}{\sqrt{q_r}}\right) \\ &\leq \frac{8}{q_r} + 2\tilde{\beta}'_r(\delta)\left(\tilde{\sigma}_{t_r,r}(x^*) + \tilde{\sigma}_{t_r,r}(x) + \frac{4}{\sqrt{q_r}}\right). \end{aligned}$$

Combining with the bound on $\tilde{\sigma}_{t_r,r}(x)$ obtained in the previous step, we bound the instantaneous regret for all $x \in \mathcal{X}_{r+1}$ as follows.

$$f(x^*) - f(x) \leq \frac{8}{q_r} + 4\tilde{\beta}'_r(\delta)\left(\sqrt{\frac{c_1\Gamma_k(q_r)}{q_r}} + \frac{2}{\sqrt{q_r}}\right). \tag{27}$$

Step 3 (Bounding cumulative regret): Let $\mathcal{R}_r = \sum_{k=1}^{t_r}(f(x^*) - f(X_{k,r}))$ be the total regret in round r. Then, we have $\mathcal{R}(T) = \sum_{r=1}^{R}\mathcal{R}_r$. Summing up the regret in round r, we obtain the following.

When $\mathcal{X}$ is finite (The setting of Theorem 6.1):

Replacing the bound on instantaneous regret from (26), for $r \geq 2$,

$$\begin{aligned} \mathcal{R}_r &\leq 4t_r\tilde{\beta}(\delta)\sqrt{\frac{c_1\Gamma_k(q_{r-1})}{q_{r-1}}} \\ &\leq 4\lceil\sqrt{Tq_{r-1}} + u_T(\delta)\rceil\tilde{\beta}(\delta)\sqrt{\frac{c_1\Gamma_k(q_{r-1})}{q_{r-1}}} \\ &\leq 4\tilde{\beta}(\delta)\sqrt{c_1T\Gamma_k(q_{r-1})} + (u_T(\delta)+1)\tilde{\beta}(\delta)\sqrt{\frac{c_1\Gamma_k(q_{r-1})}{q_{r-1}}}. \end{aligned}$$

The second inequality is obtained by the choice of t_r.

When $\mathcal{X}$ is continuous (The setting of Theorem 6.2):

Replacing the bound on instantaneous regret from (27), for $r \geq 2$,

$$\begin{aligned} \mathcal{R}_r &\leq t_r\left(\frac{8}{q_{r-1}} + 4\tilde{\beta}'_{r-1}(\delta)\left(\sqrt{\frac{c_1\Gamma_k(q_{r-1})}{q_{r-1}}} + \frac{2}{\sqrt{q_{r-1}}}\right)\right) \\ &\leq \lceil\sqrt{Tq_{r-1}} + u_T(\delta)\rceil\left(\frac{8}{q_{r-1}} + 4\tilde{\beta}'_{r-1}(\delta)\left(\sqrt{\frac{c_1\Gamma_k(q_{r-1})}{q_{r-1}}} + \frac{2}{\sqrt{q_{r-1}}}\right)\right) \\ &\leq 4\tilde{\beta}'_{r-1}(\delta)\sqrt{c_1T\Gamma_k(q_{r-1})} + 8\sqrt{\frac{T}{q_{r-1}}} + 2\sqrt{T} \\ &\quad + (u_T(\delta)+1)\left(\frac{8}{q_{r-1}} + 4\tilde{\beta}'_{r-1}(\delta)\left(\sqrt{\frac{c_1\Gamma_k(q_{r-1})}{q_{r-1}}} + \frac{2}{\sqrt{q_{r-1}}}\right)\right). \end{aligned}$$

We also note that $f(x) = \langle f, k(\cdot, x)\rangle$ (reproducing property). Thus $|f(x)| \leq \|f\|_{\mathcal{H}_k}\|k(\cdot, x)\|$; i,e, $|f(x)| \leq C_k k_{\max}$. Thus, the regret in the first round can be simply bounded using its length,

$$\begin{aligned}\mathcal{R}_1 &\leq t_1 C_k k_{\max} \\ &\leq \lceil\sqrt{T} + u_T(\delta)\rceil C_k k_{\max}.\end{aligned}$$

Recall $u_T(\delta) = \mathcal{O}(\mathbb{E}[\tau] + \log(\frac{T}{\delta}))$ and note that, for $r > 1$,

$$\frac{8}{q_{r-1}} + 4\tilde{\beta}'_{r-1}(\delta)\left(\sqrt{\frac{c_1\Gamma_k(q_{r-1})}{q_{r-1}}} + \frac{2}{\sqrt{q_{r-1}}}\right) = \mathcal{O}(1), \tag{28}$$

since $q_{r-1} \geq \sqrt{T}$ for $r > 1$.

Adding up the regret over all rounds $r = 1, 2, \ldots, R$ with $R = \log\log(T)$, we obtain

$$\mathcal{R}(T) = \mathcal{O}\left(\tilde{\beta}(\delta)\log\log(T)\sqrt{T\Gamma_k(T)} + C_k\mathbb{E}[\tau]\right), \tag{29}$$

and,

$$\mathcal{R}(T) = \mathcal{O}\left(\tilde{\beta}'_R(\delta)\log\log(T)\sqrt{T\Gamma_k(T)} + C_k\mathbb{E}[\tau]\right), \tag{30}$$

under the discrete and continuous $\mathcal{X}$ cases, proving Theorems 6.1 and 6.2, respectively.

Corollary 6.3 immediately follows from replacing the value of Γ_k, in particular, $\Gamma_k(T) = \mathcal{O}(T^{\frac{d}{2\nu+d}}(\log(T))^{\frac{2\nu}{2\nu+d}})$, $\Gamma_k(T) = \mathcal{O}((\log(T))^{d+1})$ and $\Gamma_k(T) = \mathcal{O}(d\log(T))$, in the cases of Matérn, SE and linear kernels respectively (Srinivas et al., 2010; Vakili et al., 2021c). In the case of linear kernel, we emphasize the $\mathcal{O}(\sqrt{d})$ factor in $\tilde{\beta}'_R(\delta)$.