Contextually Adaptive Algorithms for Gaussian Process Bandit Optimization under Heavy-tailed Noise

Hyeonjun Park[a] **and Kyungjae Lee** [b,*]

[a]the Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea
[b]the Department of Statistics, Korea University, Seoul 02841, Republic of Korea
ORCID (Kyungjae Lee): https://orcid.org/0000-0003-0147-2715

Abstract. We consider a Gaussian process (GP) bandit optimization problem when the objective function lives in a reproducing kernel Hilbert space (RKHS), assuming that the payoffs follow a heavy-tailed distribution with a bounded $(1+\epsilon)$-th moment for some $\epsilon \in (0, 1]$. Existing algorithms for this setting face practical challenges due to their significant computational demands and inconsistent theoretical guarantee to translation of noise distribution. To address these issues, we introduce two robust algorithms. The first algorithm utilizes a truncation estimator, achieving the same regret bound as that of the existing algorithm up to logarithmic terms with reduced time complexity. The second algorithm employs a median-of-means estimator and achieves more stable regret bound to alteration of noise distribution with lower time and space complexities compared to existing methods. Finally, we empirically validate the performance of our proposed algorithms against previous methods in both synthetic and real-world datasets.

1 Introduction

Black-box optimization is the problem of finding an optimum of an unknown target function with expensive and noisy evaluations, which generally arises in numerous real-world applications, such as experimental design [32], hyper-parameter optimization [31], material discovery [12], and robotics [20]. *Gaussian process (GP) bandit optimization*, which is one of the most popular method to solve this problem, leverages a surrogate model in a reproducing kernel Hilbert space (RKHS), evaluates the target function to update the surrogate model, selects the next evaluation points by using some specific sampling strategy, and repeat this process towards reaching the optimum of the target function. Many GP bandit algorithms generally aim to design an efficient sampling strategy that allows reaching the optimum with the minimal evaluations.

To design an efficient sampling strategy, algorithms in GP bandit optimization often make assumptions on noise distribution that represents inherent uncertainty and randomness present in real-world applications. Previous studies in GP bandit optimization have commonly assumed sub-Gaussian noise characterized by a *light-tailed* distribution [32, 22, 8, 9, 19]. However, the sub-Gaussian noise assumption may not entirely encompass the range of real-world scenarios encountered in practice, where heavy-tailed noises often occur. In addition, it is known that algorithms designed for the sub-Gaussian noise often suffer from a decrease in convergence speed or a failure to find an optimum when applied in heavy-tailed noise settings [6]. This issue has been addressed by a significant body of researches in bandit literature from multi-armed bandits to GP bandit optimization [6, 26, 30, 37, 23, 25, 2]. Pioneeringly, Chowdhury and Gopalan [10] were at the forefront of addressing this challenge within the context of GP bandit optimization. Chowdhury and Gopalan [10] proposed two robust algorithms designed to handle heavy-tailed noise: the truncated GP-UCB (TGP-UCB) and the adaptively truncated approximate GP-UCB (ATA-GP-UCB). These algorithms employ strategic truncation of extreme observations contaminated by heavy-tailed noise through carefully calibrate truncation thresholds, which allows heavy-tailed observations to be treated like sub-Gaussian noise.

The robust algorithms in [10] have reached a sub-linear regret bound with respect to total trials T, which ensures convergence to the optimum. Despite this progress, each algorithm still suffers from its own drawbacks. First, the regret bound of TGP-UCB, denoted as $\tilde{O}\left(\bar{\nu}^{\frac{1}{1+\epsilon}}\gamma_T T^{\frac{2+\epsilon}{2(1+\epsilon)}}\right)$, leaves a theoretical gap with the lower bound suggested in [10]. Here, γ_T is a maximum information gain, $\bar{\nu}$ is a raw moment of noise distribution, and $\epsilon \in (0, 1]$. To eliminate this gap, Chowdhury and Gopalan [10] introduced the ATA-GP-UCB whose regret bound, denoted as $\tilde{O}(\bar{\nu}^{\frac{1}{1+\epsilon}}\gamma_T T^{\frac{1}{1+\epsilon}})$, matches the lower bounds in terms of T. Despite matching the optimal regret bound, ATA-GP-UCB incurs substantial computational costs due to its approximate truncation technique using Nyström approximation [38]. Unlike TGP-UCB, which truncates raw observations directly, ATA-GP-UCB adaptively trims weighted observations using truncation weights. This delicate approach allows ATA-GP-UCB to achieve the tight regret bound. However, when defining the truncation weights, ATA-GP-UCB utilizes features which represent a mapping from input space to high-dimensional space (in this case, RKHS). Then, ATA-GP-UCB truncates all historical weighted observations across all dimensions of the features. This technique presents a challenge in RKHS where the feature space can be infinite-dimensional. Thus ATA-GP-UCB necessitates an additional step to embed the infinite-dimensional features into a finite-dimensional space for calculating weights, which induces extra computations and the Nyström approximation is used for this feature embedding step. Moreover, the regret bounds of both TGP-UCB and ATA-GP-UCB depend on the raw moments of the noise distribution $\bar{\nu}$, making the algorithms sensitive and unstable to translations in the noise distribution, potentially impairing their performance. These constraints have motivated us to devise algorithms with lower computational com-

* Corresponding author: Kyungjae Lee (e-mails: dlxhrl@korea.ac.kr)

plexity and a more stable theoretical guarantee.

In this paper, we introduce two GP bandit optimization algorithms for handling heavy-tailed noise. The first algorithm, named contextual adaptive truncated GP-UCB (CA-TGP-UCB), deals with heavy-tailed noise using a truncation method [6]. Importantly, it has been shown that adaptively truncating weighted rewards leads to a better regret bound compared to directly truncating raw rewards [30, 37]. While both CA-TGP-UCB and ATA-GP-UCB employ the adaptive truncation method, the primary distinction lies in the definition of the truncation weights. In particular, the truncation weights of CA-TGP-UCB can be directly computed by using a kernel function. The basic idea is to adjust them differently for each input along with the truncation threshold. By doing so, it is possible to define more refined weights and truncation thresholds for different inputs, achieving a similar effect to the weights in ATA-GP-UCB with more reduced time complexity. However, although CA-TGP-UCB achieves the tight regret bound of $O(\bar{\nu}\gamma_T T^{\frac{1}{1+\epsilon}})$, it still scales with the raw moment of the noise, similar to other truncation algorithms.

The second algorithm, median-of-means GP-UCB (MoM-GP-UCB), which employs the median-of-means estimator [6], addresses this problem. Importantly, the median-of-means estimator has been widely used in various fields such as heavy-tailed bandits [6, 26, 30, 37], regression [16, 17], and robust kernel density estimation [18], but it has not been utilized in GP bandit optimization previously. The basic idea is to divide all trials into multiple episodes, repeatedly perform a fixed input within each episode to obtain means corresponding that input, and then update the surrogate model for the target function based on the median of the obtained means. When using the median-of-means (MoM) estimator, the performance of algorithms depends on how means are defined. For example, in cases of stochastic bandits [6] and linear bandits [26] where the MoM estimator is used, the mean is defined as the empirical mean. While this formulation ensures robustness against heavy-tailed noise, it fails to yield a tight regret bound. Similar to truncation-based methods, we set mean of MoM-GP-UCB as weighted mean where the weight has a dependency on input. This technique allows us to achieve a tighter regret bound. Furthermore, MoM-GP-UCB requires updating the surrogate only once per episode, resulting in lower computational complexity compared to truncation-based algorithms that need updates at every time step. Additionally, in contrast to truncation-based algorithms where the regret bound grows with raw moments, the regret bound of MoM-GP-UCB increases with central moments of noise distribution. The complete paper with supplementary and corresponding code are available at this reference [1]. Now, we highlight our contributions as follows:

- We introduce CA-TGP-UCB, a robust GP bandit algorithm with a truncation estimator. It achieves a cumulative regret bound of $\tilde{O}(\bar{\nu}\gamma_T T^{\frac{1}{1+\epsilon}})$, where $\epsilon \in (0, 1]$ and $\bar{\nu}$ represents a raw moment of heavy-tailed noise. This algorithm reduces time complexity compared to previous methods while keeping the same regret bound, up to logarithmic terms.
- We develop MoM-GP-UCB, a robust GP bandit optimization algorithm that uses the median-of-means estimator. MoM-GP-UCB achieves a regret bound of $\tilde{O}(\nu^{\frac{1}{1+\epsilon}}\gamma_T T^{\frac{1}{1+\epsilon}})$, where ν represents a central moment of heavy-tailed noise. Additionally, MoM-GP-UCB improves time and space complexities compared to truncation-based algorithms.
- We experimentally demonstrate that CA-TGP-UCB and MoM-GP-UCB perform better than TGP-UCB and ATA-GP-UCB under heavy-tailed noise, both in synthetic and real-world datasets, with reduced execution times.

2 Related Work

Gaussian process (GP) optimization can be approached in two ways: the Bayesian setting, with functions sampled from a GP prior, and the frequentist setting, assuming functions lie in a reproducing kernel Hilbert space (RKHS). While this paper primarily focuses on the frequentist setting, our methods can be easily extended to the Bayesian setting using similar techniques as in [32]. This section introduces research on the frequentist setting.

GP optimization involves iteratively selecting points and optimizing an unknown objective function based on the feedback obtained at those points. Thus the classification of GP optimization depends on the form in which feedback is provided during this process, such as Thompson sampling [9, 34], expected improvement [7, 15], and upper confidence bound (UCB) [32, 9]. In particular, the formulation of GP optimization with UCB-style bandit feedback was first introduced by Srinivas et al. [32]. They proposed a GP bandit optimization algorithm called GP-UCB and analyzed it in both Bayesian and frequentist settings, establishing $\tilde{O}(\sqrt{T\gamma_T})$ and $\tilde{O}(\sqrt{T}\gamma_T)$ regret bounds where $\tilde{O}$ ignores dimension-independent logarithmic terms, γ_T is a maximum information gain and T denotes total trials. Building upon this research, the subsequent studies have focused on improving the regret bound of GP-UCB [36, 9, 19] or extending its application to more general settings; e.g., contextual GP bandits [22], corruption-tolerant GP bandits [5], misspecified GP bandits[4], and more on. In particular, Chowdhury and Gopalan [9] designed two GP bandit optimization algorithms, named IGP-UCB and GP-TS which are designed under frequentist and Bayesian settings, respectively. The regret bounds of both algorithms are improved from that of GP-UCB by a factor of $O(\ln T)$. Following this work, Janz et al. [19] introduced a GP bandit algorithm that ensures a sublinear regret bound for the Matérn kernel with specific kernel parameters, although its practicality is compromised by a large constant factor multiplied to the regret bound. In discrete action spaces, Valko et al. [36] introduced SupKernelUCB, a sublinear GP bandit algorithm based on SupLinUCB [11]. Recent practical methods in GP optimization, such as tree-based domain shrinking [28], pure exploration [33], and batched pure exploration [24, 35], have achieved regret bounds of $\tilde{O}(\sqrt{\gamma_T T})$. Additionally, Scarlett et al. [29] and Cai and Scarlett [8] proposed kernel-specific lower bounds for GP bandit algorithms under sub-Gaussian noise.

We emphasize that attempts to improve the regret bound of GP-UCB typically rely on the sub-Gaussian noise assumption. However, in GP bandit optimization, Chowdhury and Gopalan [10] recently proposed robust algorithms for heavy-tailed payoffs, a topic less explored in GP bandits compared to other bandit fields. Chowdhury and Gopalan [10] adapted GP-UCB using truncation estimator [6], originally employed in stochastic bandits to deal with heavy-tailed noise, to develop two robust GP bandit algorithm: TGP-UCB and ATA-GP-UCB. These algorithms achieve the cumulative regret bounds $\tilde{O}(\bar{\nu}^{\frac{1}{1+\epsilon}}\gamma_T T^{\frac{2+\epsilon}{2(1+\epsilon)}})$ and $\tilde{O}(\bar{\nu}^{\frac{1}{1+\epsilon}}\gamma_T T^{\frac{1}{1+\epsilon}})$, respectively. However, TGP-UCB shows a gap in regret bounds compared to kernel-specific lower bounds [10], and, while ATA-GP-UCB closes this gap, it demands significant computational resources.

3 Problem Formulation

We now introduce Gaussian process (GP) bandit optimization with bandit feedback under heavy-tailed payoffs. Let us define a com-

pact set $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ and let $f : \mathcal{X} \to \mathbb{R}$ be the objective function. Then we define a positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space. In particular, the Hilbert space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) associated to a kernel k if $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ holds for all $x, x' \in \mathcal{X}$ and the reproducing property $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ is satisfied for all $f \in \mathcal{H}$ where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is an inner product on $\mathcal{H}$. Throughout this paper, we assume that the objective function f lies in a RKHS associated with squared exponential (SE) kernel and Matérn kernel, which are widely-used ones in GP optimization [21] and defined by

$$k_{\text{SE}}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2l^2}\right) \tag{1}$$

$$k_{\text{Matérn}}(x, x') = \frac{2^{1-\omega}}{\Gamma(\omega)} \left(\frac{\sqrt{2\omega}\|x - x'\|}{l}\right)^{\omega} J_{\omega}\left(\frac{\sqrt{2\omega}\|x - x'\|}{l}\right). \tag{2}$$

Here, l denotes the characteristic length scale, ω is a smoothness parameter, and J_{ω} is a modified Bessel function. We now present the standard assumptions in GP bandit optimization [32] and heavy-tailed bandits [6].

Assumption 1 (Boundedness)**.** *There exists a positive constant B such that $\|f\|_{\mathcal{H}} \leq B$ for all $f \in \mathcal{H}$.*

Note that this assumption implies $\sqrt{k(x, x)} \leq B$ since $\|k(x, \cdot)\|_{\mathcal{H}} = \sqrt{k(x, x)}$.

Assumption 2 (Heavy-tailed noise)**.** *Let $\mathcal{F}_t$ be a filtration generated by $\{x_i\}_{i=1}^{t} \cup \{\eta_i\}_{i=1}^{t}$, where x_i is an action played and η_i is a noise at round $i \in [t]$, respectively. Then x_t and η_t are $\mathcal{F}_t$-measurable, respectively. Suppose that $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$ holds and there exists some constant $\nu < \infty$ such that $\mathbb{E}[|\eta_t|^{1+\epsilon} | \mathcal{F}_{t-1}] \leq \nu$ for any $\epsilon \in (0, 1]$. In addition, define $\bar{\nu} = (\nu + B)^{1+\epsilon}$. Then, $\mathbb{E}[|y_t|^{1+\epsilon} | \mathcal{F}_{t-1}] \leq \bar{\nu}$ holds for all $t \in \mathbb{N}$.*

Assumption 2 represents the $(1 + \epsilon)$-th moment assumption on noisy observations, where $\epsilon \in (0, 1]$. It is worth noting that Assumption 2 encompasses the sub-Gaussian noise assumption posed in [32], especially when $p = 2$. Based on Assumptions 1 and 2, GP bandit optimization is formulated as follows: at each time step $t \in [T]$, the learner (i.e., bandit algorithm) selects an action $x_t \in \mathcal{X}$ and observes a noisy reward $y_t = f(x_t) + \eta_t$ where $f \in \mathcal{H}$ is an (fixed) unknown objective function and noise η_t follows Assumption 2. Then the goal of the learner is to minimize a cumulative regret over total rounds T, $\mathcal{R}_T \coloneqq \sum_{t=1}^{T} f(x_*) - f(x_t)$, where x_* is an (not necessarily unique) optimal point of f. We note that minimizing the cumulative regret is equivalent to maximize the expected cumulative reward $\sum_{t=1}^{T} f(x_t)$. For any $n \in \mathbb{N}$, let $[n] \coloneqq \{1, 2, \ldots, n\}$. Further, we will denote by $\|\cdot\|_p$ the p-th norm for any $p \in \mathbb{N}$. $\tilde{O}$ indicates big O notation that hides logarithmic terms.

4 Algorithms

In this section, we propose two Gaussian process (GP) bandit optimization algorithms under heavy-tailed payoffs. Both algorithms employ upper confidence bound (UCB) strategy [3] to select an action and utilize a Gaussian likelihood as the surrogate model for the objective function. In this context, the objective function is sampled from GP prior $\mathcal{GP}(0, k)$ and specified by posterior mean μ and covariance σ functions where k is a kernel function. Note that the GP surrogate model is only used for algorithm design and we still assume that the fixed objective function f lies in RKHS. Particularly, previous GP bandit algorithms in sub-Gaussian noise setting select an action based on the following rule:

$$x_t = \arg\max_{x \in \mathcal{X}} \mu_{t-1}(x) + \alpha_t \sigma_{t-1}(x) \tag{3}$$

where $\mu_t(x) = \mathbf{k}_t(x)^{\intercal}(K_t + \lambda I_t)^{-1} Y_t$ is the posterior mean, $\sigma_t^2(x) = k(x, x) - \mathbf{k}_t^{\intercal}(x)(K_t + \lambda I_t)^{-1}\mathbf{k}_t(x)$ is the posterior variance when t observations are given, and α_t is a confidence parameter. Here, we denote by the kernel matrix $K_t = [k(x_i, x_j)]_{i,j \in [t]}$, stacked kernel functions $\mathbf{k}_t(x) = [k(x, x_i)]_{i \in [t]}^{\intercal}$, and a vector containing observations $Y_t = [y_1, \ldots, y_t]^{\intercal}$, respectively. In addition, we formally define the maximum information gain that is used to represent the regret bounds in GP bandit optimization as follows.

Definition 1 (Maximum Information Gain [32])**.** *Let f be a function sampled from Gaussian process prior $\mathcal{GP}(0, k)$ adding with i.i.d. Guassian noises $\mathcal{N}(0, \lambda)$ where k is a kernel function and $\lambda > 0$. Then the maximum information gain is defined as*

$$\gamma_T \coloneqq \max_{A \subset D : |A| = T} \frac{1}{2} \ln |I + \lambda^{-1} K_A| \tag{4}$$

where $D \subset \mathbb{R}^d$ is compact and convex set.

Contextual Adaptive Truncated GP-UCB. The first algorithm, named Contextual Adaptive Truncated GP-UCB (Algorithm 1) leverages the truncation estimator [6]. The intuitive idea underlying this method is to trim extreme-valued rewards. Therefore, determining which term to trim and what threshold to use becomes a crucial factor in the performance of the algorithm. Specifically, CA-TGP-UCB truncates heavy-tailed rewards as follows:

$$\hat{y}_t = y_t \mathbb{1}_{\{|b_t y_t| \leq h(x_t)\}} \tag{5}$$

where $\mathbb{1}$ is the indicator function, b_t is a truncation weight, and $h(x_t)$ is truncation threshold defined as $h(x_t) = \|\mathbf{k}_t(x)^{\intercal}(K_t + \lambda I_t)^{-1}\|_{1+\epsilon}$ with $\epsilon \in (0, 1]$. Then, by using the truncated rewards, CA-TGP-UCB select an action, at every round $t \in [T]$, with the following selection rule:

$$x_t = \arg\max_{x \in \mathcal{X}} \hat{\mu}_{t-1}(x) + \alpha_t \sigma_{t-1}(x) \tag{6}$$

where $\hat{\mu}_t(x) \coloneqq \mathbf{k}_t(x)^{\intercal}(K_t + \lambda I_t)^{-1}\hat{Y}_t$ is the truncated posterior mean with truncated observations $\hat{Y}_t = [\hat{y}_1, \ldots, \hat{y}_t]^{\intercal}$ and α_t is a confidence parameter defined in Lemma 3.

In the case of TGP-UCB [10], raw observations are truncated directly, such as $\hat{y}_t = y_t \mathbb{1}_{\{|y_t| \leq g_t\}}$. Here, g_t is a truncation threshold

Algorithm 1 Contextual Adaptive Truncated GP-UCB (CA-TGP-UCB)

Input: T, $\mathcal{X}$, $\alpha_t \in \mathbb{R}^+$, $\lambda > 0$, k, $\epsilon \in (0, 1]$, $\delta \in (0, 1]$, B
Initialization: Set $\hat{\mu}_0 = 0$ and $\sigma^2(x) = k(x, x)$ for all $x \in \mathcal{X}$
for $t = 1, 2, \ldots, T$ **do**
  Set $\alpha_t \coloneqq B + \lambda^{-1/2} t^{\frac{1-\epsilon}{2(1+\epsilon)}} \left(2\lambda^{-1/2}\sqrt{2\left(\gamma_t + \ln\left(\frac{1}{\delta}\right)\right)} + \bar{\nu}\right)$
  Play action $x_t = \arg\max_{x \in \mathcal{X}} \hat{\mu}_{t-1}(x) + \alpha_t \sigma_{t-1}(x)$ and observe payoff y_t
  Set truncation threshold $h(x_t) = \|\mathbf{k}_t(x)^{\intercal}(K_t + \lambda I_t)^{-1}\|_{1+\epsilon}$
  Compute $\hat{y}_t = y_t \mathbb{1}_{\{|b_t y_t| \leq h(x_t)\}}$ where b_t is the t-th element of $\mathbf{k}_t(x)(K_t + \lambda I_t)^{-1}$ and set $\hat{Y}_t = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_t]$
  Compute $\hat{\mu}_t(x) = \mathbf{k}_t(x)^{\intercal}(K_t + \lambda I_t)^{-1}\hat{Y}_t$ and $\sigma_t(x) = \sqrt{k(x, x) - \mathbf{k}_t(x)^{\intercal}(K_t + \lambda I_t)^{-1}\mathbf{k}_t(x)}$
end for

defined as $g_t = \bar{\nu}^{\frac{1}{1+\epsilon}} t^{\frac{1}{2(1+\epsilon)}}$ where $\bar{\nu}$ is a raw moment of heavy-tailed rewards and $\epsilon \in (0, 1]$. This approach leads to a loose regret bound in terms of T, $\tilde{O}(\bar{\nu}^{\frac{1}{1+\epsilon}} \gamma_T T^{\frac{2+\epsilon}{2(1+\epsilon)}})$, leaving a gap with kernel-specific regret lower bounds for squared exponential (SE) and Matérn kernels. To overcome this problem, Chowdhury and Gopalan [10] incorporated a dimension-wise truncation method, which was given by Shao et al. [30] in linear stochastic bandits, into GP bandits and introduce a novel GP bandit algorithm ATA-GP-UCB. In constrast to TGP-UCB, ATA-GP-UCB trims rewards in each feature dimension considering all historical rewards. Note that, in GP bandit setting, where a feature $\phi(x)$ can have an infinite-dimension, such truncation technique cannot be directly applied. Therefore, Chowdhury and Gopalan [10] employed Nyström [38] and QFF projection techniques [27] to embed infinite-dimensional features into finite-dimensions before the applying truncation. More precisely, let $[u_1^\intercal, \dots, u_{m_t}^\intercal]$ be the rows of $\tilde{V}_t^{-1/2} \tilde{\Phi}_t^\intercal$ where $\tilde{\phi}_t$ denotes the embedded feature, m_t is the dimension of $\tilde{\phi}_t$, $\tilde{\Phi}_t^\intercal = [\tilde{\phi}_t(x_1), \dots, \tilde{\phi}_t(x_t)]$, and $\tilde{V}_t = \tilde{\Phi}_t^\intercal \tilde{\Phi}_t + \lambda I_{m_t}$. Then the truncated reward is calculated as follows:

$$\hat{r}_i = \sum_{\tau=1}^{t} u_{i,\tau} y_\tau \mathbb{1}_{|u_{i,\tau} y_\tau| \le a_t}. \tag{7}$$

where $a_t := (\bar{\nu}/\ln(4m_t T/\delta)^{\frac{1}{1+\epsilon}} t^{\frac{1-\epsilon}{2(1+\epsilon)}})$ is a truncation threshold. While ATA-GP-UCB achieves a tighter regret bound $\tilde{O}(\bar{\nu}^{\frac{1}{1+\epsilon}} \gamma_T T^{\frac{1}{1+\epsilon}})$ using the dimension-wise truncation method, this approach comes with extra computational costs. It requires embedding feature vectors and calculating the weight matrix $\tilde{V}_t^{-1/2}$, which cannot be updated online using the Sherman-Morrison formula [14].

We point out that utilizing an adaptive truncation technique (5) for features can reduce the computational cost of ATA-GP-UCB, while still achieving a favorable regret bound $\tilde{O}(\bar{\nu}\gamma_T T^{\frac{1}{1+\epsilon}})$. In contrast to TGP-UCB and ATA-GP-UCB, the weighted truncated term $b_t y_t$ and truncation threshold $h(x_t)$ of CA-TGP-UCB depend on the action played (i.e., the feature $\phi(x_t)$), which helps us to obtain the tight regret bound in terms of T. Specifically, employing an adaptive threshold for features results in a narrow confidence interval.

Lemma 3 (Confidence interval of Algorithm 1)**.** *Suppose that Assumptions 1 and 2 hold. Let $\delta \in (0, 1]$ and $\epsilon \in (0, 1]$, and let $\alpha_t := B + \lambda^{-1/2} t^{\frac{1-\epsilon}{2(1+\epsilon)}} (2\lambda^{-1/2}\sqrt{2(\gamma_t + \ln(1/\delta))} + \bar{\nu})$. Then, the confidence interval of Algorithm 1 is as follows, with probability at least $1 - \delta$, uniformly over all $t \ge 1$,*

$$|\hat{\mu}_t(x) - f(x)| \le \alpha_t \sigma_t(x). \tag{8}$$

The complete proof is deferred to the supplementary. It is worth noting that in GP bandits, the regret bound takes the form of $\tilde{O}(\alpha_t \sum_{t=1}^{T} \sigma_{t-1}(x_t))$, where σ_{t-1} represents the posterior standard deviation and α_t serves as a confidence parameter. Notably, the summation of the posterior standard deviations over T rounds can be bounded by $\tilde{O}(\sqrt{\gamma_T T})$. Therefore, the order of α_t determines the regret bound of GP bandits. For instance, TGP-UCB has a confidence parameter $\alpha_{t,\text{TGP}} = \tilde{O}(t^{\frac{1}{2(1+\epsilon)}}\sqrt{\gamma_T})$ which induces a suboptimal regret bound $\tilde{O}(T^{\frac{2+\epsilon}{2(1+\epsilon)}}\gamma_T)$ in terms of T. Additionally, while ATA-GP-UCB obtains the same order of confidence parameter as ours in terms of t, $\alpha_{t,\text{ATA}} = \tilde{O}(t^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\gamma_T})$, it requires additional feature embedding steps for ensuring the desired order, resulting in additional computational costs. We argue that the feature adaptive threshold can simply yield the confidence parameter $\alpha_{t,\text{CA}} = \tilde{O}(t^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\gamma_T})$ which deduces the favorable regret bound, $\tilde{O}(T^{\frac{1}{1+\epsilon}}\gamma_T)$, in terms of T and can be obtained without extra computational costs. For the sake of completeness, we present a proof sketch of Lemma 3.

Proof sketch of Lemma 3. By the definition of $\hat{\mu}_t$ and triangle inequality, we have

$$|\hat{\mu}_t(x) - f(x)| \le |\phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \hat{N}_t| - |f(x) - \zeta_t(x)| \tag{9}$$

where $\zeta_t(x) := \mathbf{k}_t(x)^\intercal (K_t + \lambda I_t)^{-1} f_t$ for all $x \in \mathcal{X}$ and $f_t = [f(x_1), f(x_2), \dots, f(x_t)]^\intercal$ is a vector of evaluations of f up to time step t. Then, the second term of RHS of (9) bounded by $B\sigma_t(x)$ where B is some constant such that $\|f\|_{\mathcal{H}} \le B$ (Assumption 1). The first term of (9) can be decomposed by

$$\begin{aligned} &|\phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \hat{N}_t| \\ &\le \left| \sum_{i=1}^{t} b_t^i \hat{\eta}_t^i - \mathbb{E}\left[\sum_{i=1}^{t} b_t^i \hat{\eta}_t^i | \mathcal{G}_{t,i-1} \right] \right| + \left| \mathbb{E}\left[\sum_{i=1}^{t} b_t^i \hat{\eta}_t^i | \mathcal{G}_{t,i-1} \right] \right| \end{aligned} \tag{10}$$

where b_t^i is the i-th element of $\mathbf{k}_t(x)^\intercal (K_t + \lambda I_t)^{-1}$, $\hat{\eta}_t^i$ is a truncated noise, and $\mathcal{G}_{t,\tau} := \sigma(\{x_1, x_2, \dots, x_t\} \cup \{y_1, y_2, \dots, y_\tau\})$ is a σ-algebra with $\tau \in [t]$. To bound the first term of RHS of (10), we employ the self-normalized inequality [13], which yields $\tilde{O}(h(x_t)\sqrt{\gamma_t})$ regret upper bound where $h(x_t)$ is a truncation threshold for the played action x_t. Furthermore, the second term of (10) is bounded by $\bar{\nu} h(x_t)$ where $\bar{\nu}$ is the $(1+\epsilon)$-th moment of noisy observations as defined in Assumption 2. Recall that the truncation threshold is defined as $h(x_t) = \|\mathbf{k}_x(x)^\intercal (K_t + \lambda I_t)^{-1}\|_{1+\epsilon}$. Then we have that $h(x_t) \le \lambda^{-1/2} t^{\frac{1-\epsilon}{2(1+\epsilon)}} \sigma_t(x)$ for all $t \in [T]$. By using this property of $h(x_t)$ and combining the inequalities (9) and (10) with their upper bounds, the lemma is established. □

Intuitively, truncating rewards introduces bias and this can be controlled mathematically by (10). Since both TGP-UCB and ATA-GP-UCB employ the truncation technique, similar to our approach, they also have a form of (10) in their proofs. However, the method of bounding this inequality varies depending on the definition of the truncation threshold and the term to be truncated, thereby altering the order of t in confidence interval. For instance, TGP-UCB bounds the second term of RHS of (10) by using the property of their truncation threshold $g_t = \bar{\nu}^{\frac{1}{1+\epsilon}} t^{\frac{1}{2(1+\epsilon)}}$ where $\bar{\nu}$ is the $(1+\epsilon)$-th moment of noise (Lemma 8 in [10]) and the fact that raw reward observation y_t is truncated. In their proof, that term is bounded by $\tilde{O}(t^{\frac{1}{2(1+\epsilon)}}\sqrt{\gamma_t})$, which is less tighter than our bound $\tilde{O}(t^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\gamma_t})$.

On the other hand, ATA-GP-UCB shares the same bound as ours up to logarithmic terms. The bound of the inequality in ATA-GP-UCB is $\tilde{O}(a_t^{-\epsilon}\|u_i\|_{1+\epsilon}^{1+\epsilon})$, where u_i is a truncation weight such that $\|u_i\|_{1+\epsilon}^{1+\epsilon} \le t^{\frac{1-\epsilon}{2(1+\epsilon)}}$ and $a_t = \tilde{O}(t^{\frac{1-\epsilon}{2(1+\epsilon)}})$ is a truncation threshold. Thus the order of truncation threshold is same as ours and the truncation weight u_i plays a role of the context adaptive truncation weight in our algorithm. However, as earlier mentioned, the weight u_i is defined as the i-th column of $\tilde{V}_t^{-1/2} \tilde{\Phi}_t^\intercal$ where $\tilde{\Phi}_t^\intercal = [\tilde{\phi}_t(x_1), \dots, \tilde{\phi}_t(x_t)]$, $\tilde{V}_t = \tilde{\Phi}_t^\intercal \tilde{\Phi}_t + \lambda I_{m_t}$, and m_t is the dimension of embedded feature $\tilde{\phi}_t(x)$. Therefore, an additional step is necessary for embedding all arms, which requires $\tilde{O}(m_t^3 + m_t^2|\mathcal{X}|)$ time complexity. Additionally, constructing $\tilde{V}_t$ requires $\tilde{O}(m_t^2 t)$ time, and computing $\tilde{V}_t^{-1/2}$ takes $\tilde{O}(m_t^3)$ time. Hence, the inclusion of embedding steps to achieve a tighter bound increases the time complexity of ATA-GP-UCB. In contrast, by employing context-adaptive weights

which serve as the embedded weight in ATA-GP-UCB, we can attain the same regret bound with much less time complexity as follows.

Theorem 4 (Cumulative regret bound of Algorithm 1)**.** *Suppose that Assumptions 1 and 2 hold. Let $\delta \in (0, 1]$ and $\epsilon \in (0, 1]$. Then, for any $T \in \mathbb{N}$, the cumulative regret bound of Algorithm 1 after T rounds is*

$$R_T \leq O\left((B\gamma_T + \bar{\nu}\sqrt{\gamma_T})T^{\frac{1}{1+\epsilon}}\right) \tag{11}$$

with probability at least $1 - \delta$.

The complete proof is deferred to the supplementary. This theorem shows the cumulative regret bound of CA-TGP-UCB. Ignoring logarithmic and constant terms that are independent of dimension d, we can write the regret bound as $\tilde{O}(\gamma_T T^{\frac{1}{1+\epsilon}})$. This regret bound is same as that of ATA-GP-UCB up to logarithmic terms in terms of T. In addition, when variance is finite ($\epsilon = 1$), the regret bound becomes $\tilde{O}(\gamma_T\sqrt{T})$ which matches the previous best result in GP bandits [32, 9] under sub-Gaussian noise. The kernel specific lower bounds for GP bandits under heavy-tailed payoffs have been proposed by Chowdhury and Gopalan [10], particularly for SE and Matérn kernels. For the SE kernel, the regret bound of CA-TGP-UCB is $\tilde{O}((\ln T)^d T^{\frac{1}{1+\epsilon}})$ which closes the gap between the lower bound $\tilde{\Omega}((\ln T)^{\frac{d\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}})$ up to $(\ln T)^{\frac{\epsilon}{1+\epsilon}}$. For the Matérn kernel, the regret bound of CA-TGP-UCB is $\tilde{O}(T^{\frac{1}{1+\epsilon}+\frac{d}{2\omega+d}}(\ln T)^{\frac{2\omega}{2\omega+d}})$, which is sublinear when $\frac{d}{2\epsilon} < \omega$ is satisfied. Compared to the lower bound $\tilde{\Omega}(T^{\frac{1}{1+\epsilon}+\frac{d\epsilon^2}{\omega(1+\epsilon)^2+d\epsilon(1+\epsilon)}})$, it still has a gap.

Computational complexity of CA-TGP-UCB. Recall that the truncation threshold of CA-TGP-UCB is defined as $\|\mathbf{k}_t(x)(K_t + \lambda I_t)^{-1}\|_{1+\epsilon}$, where $\epsilon \in (0, 1]$. The truncation threshold can be computed in $O(t^3 + t|\mathcal{X}|)$, where $|\mathcal{X}|$ denotes the cardinality of the input space $\mathcal{X}$. Using the computed value $\mathbf{k}_t(x)(K_t + \lambda I_t)^{-1}$, we can estimate the mean and posterior variance in $O(t)$ and $O(t|\mathcal{X}|)$ time, respectively. Consequently, the per-step time complexity is $O(t^3 + t|\mathcal{X}|)$. Since we need to store $\mathbf{k}_t(x)$, K_t and $\mathbf{k}_t(x)(K_t + \lambda I_t)^{-1}$ for all $x \in \mathcal{X}$, the per-step space complexity is $O(t^2 + t|\mathcal{X}|)$. Therefore the total time and space complexities over T rounds are $O(T^4 + T|\mathcal{X}|)$ and $O(T^3 + T|\mathcal{X}|)$, respectively. In comparison, the total time complexity of ATA-GP-UCB is $O(m_t^2(T^2 + T|\mathcal{X}|))$, where m_t is the dimension of the embedded feature vector $\tilde{\phi}_t(x)$. It is important to note that ATA-GP-UCB requires both feature embedding and the construction of a weight matrix $\tilde{V}_t$ that defines the truncation weight, which is a row vector of $\tilde{V}_t$. These steps lead to an additional time complexity of $O(m_t^3 + m_t^2|\mathcal{X}| + m_t^2 t)$. Moreover, since ATA-GP-UCB needs to store the square inversion of weight matrix $\tilde{V}_t^{-1/2}$ and feature embedding $\tilde{\phi}_t(x)$ for all $x \in \mathcal{X}$, it incurs additional per-step space complexity $O(m_t(m_t + |\mathcal{X}|))$.

If m_t is much smaller than T, the overall time complexity is lower than ours in terms of big-O sense. However, when m_t is too small, performance of ATA-GP-UCB degrades because the embedded feature space fails to adequately capture the information from the original feature space. As a result, ATA-GP-UCB needs to keep m_t sufficiently large, which incurs significant additional computational costs, particularly when $m_t \approx T$. This demonstrates that the time complexity of CA-TGP-UCB is an improvement over ATA-GP-UCB, as empirically shown in experiment section 5.

Median-of-means GP-UCB. Another way to handle heavy-tailed payoffs is to use the median-of-means estimator [6] and the second algorithm (Algorithm 4), median-of-means (MoM) GP-UCB, takes this approach. For a given total time step T, MoM-GP-UCB first

Algorithm 2 Median-of-means (MoM) GP-UCB

Input: T, $\mathcal{X}$, $\alpha_t \in \mathbb{R}^+$, $\lambda > 0$, k, $\epsilon \in (0, 1]$, $\delta, \delta' \in (0, 1]$, B
Initialization: Set $\ell = 8\ln(\frac{2T}{\delta'})$, $N = \lfloor \frac{T}{\ell} \rfloor$, $\tilde{\mu}_0 = 0$
for $n = 1, 2, \ldots, N$ **do**
  Set $\alpha_n = n^{\frac{1-\epsilon}{2(1+\epsilon)}}(4\nu)^{\frac{1}{1+\epsilon}}\left(2B\lambda^{-\frac{1}{2}}\sqrt{\gamma_n + \ln(\frac{1}{\delta})} + \frac{1}{4}\right) + B$
  Select action $x_n = \arg\max_{x\in\mathcal{X}} \tilde{\mu}_{n-1}(x) + \alpha_n\sigma_{n-1}(x)$
  Play x_n with ℓ times and observe payoffs $y_{n,1}, y_{n,2}, \ldots, y_{n,\ell}$
  Compute $\mathbf{k}_n(x)^\intercal(K_n + \lambda I_n)^{-1}Y_n^j$ for each $j \in [\ell]$
  Compute $\tilde{\mu}_n(x) = \text{median}\{\mathbf{k}_n(x)(K_n + \lambda I_n)Y_n^j\}_{j=1}^{\ell}$
  Set $\sigma_n(x) = \mathbf{k}_n(x)^\intercal(K_n + \lambda I_n)^{-1}\mathbf{k}_n(x)$
end for

divides T into $N = \lfloor \frac{T}{\ell} \rfloor$ episodes where ℓ is the length of each episode. Then, for each episode $n \in [N]$, MoM-GP-UCB plays the chosen arm ℓ times and observes ℓ rewards. After that, MoM-GP-UCB finds the median of means as follows:

$$\tilde{\mu}_n(x) = \text{median}\{\mathbf{k}_n(x)^\intercal(K_n + \lambda I_n)^{-1}Y_n^j\}_{j=1}^{\ell} \tag{12}$$

where $Y_n^j \coloneqq \{y_{1,j}, y_{2,j}, \ldots, y_{n,j}\}$ for any $j \in [\ell]$. It is worth noting that the definition of means, $\mathbf{k}_n(x)^\intercal(K_n + \lambda I_n)^{-1}Y_n^j$, is used to derive a tight confidence interval. In other words, similar to the truncation method, adaptively considering raw rewards within the context, rather than directly, helps in achieving a tight regret bound as follows.

Lemma 5 (Confidence interval of median-of-means estimator)**.** *Suppose that Assumptions 1 and 2 hold. Let $\delta, \delta' \in (0, 1]$, and $\epsilon \in (0, 1]$. Let us denote the median of means estimator in the n-th episode of Algorithm 4 by $\tilde{\mu}_n$ for any $n \in [N]$. Then the following holds for all $x \in \mathcal{X}$ and uniformly over all $n \geq 1$,*

$$\mathbb{P}\{|\tilde{\mu}_n(x) - f(x)| \leq \alpha_n\sigma_n(x)\} \geq 1 - \delta'/T \tag{13}$$

where $\alpha_n \coloneqq n^{\frac{1-\epsilon}{2(1+\epsilon)}}(4\nu)^{\frac{1}{1+\epsilon}}\left(2B\lambda^{-\frac{1}{2}}\sqrt{\gamma_n + \ln(\frac{1}{\delta})} + \frac{1}{4}\right) + B.$

The complete proof is deferred to the supplementary. Note that the order of confidence parameter $\alpha_n = \tilde{O}(n^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\gamma_n})$ is same as that of CA-TGP-UCB (Lemma 3), resulting in the same order of regret bound with respect to T (Theorem 6). However, unlike CA-TGP-UCB, where the regret bound scales with the $(1 + \epsilon)$-th raw moment of noise $\bar{\nu}$, MoM-GP-UCB has the regret bound scaled with the $(1 + \epsilon)$ central moment of noise ν. Due to this property, MoM-GP-UCB obtains a regret bound that is invariant to the translation of the noise distribution. Now, we present a proof sketch of Lemma 5.

Proof sketch of Lemma 5. Before proving the confidence interval of median-of-means, we need to establish the confidence interval of the means. In MoM-GP-UCB, the mean represents a weighted observation $\mu_{n,j}(x) \coloneqq \mathbf{k}_n(x)^\intercal(K_n + \lambda I_n)^{-1}Y_n^j$, i.e., the posterior mean for each $j \in [\ell]$ evaluations. Then, by the same argument in the proof of Lemma 3, we have for all $x \in \mathcal{X}$,

$$|\mu_{n,j}(x) - f(x)| \leq |\phi(x)^\intercal W_n^{-1}\Phi_n^\intercal\tilde{N}_n^j| + |f(x) - \zeta_n(x)| \tag{14}$$

where $\tilde{N}_n^j \coloneqq [\tilde{\eta}_{1,j}, \tilde{\eta}_{2,j} \ldots, \tilde{\eta}_{n,j}]^\intercal$ with $\tilde{\eta}_{n,j} = y_{n,j} - f(x_{n,j})$ for all $n \in [N]$ and $j \in [\ell]$, and $\zeta_n(x) \coloneqq \mathbf{k}_n(x)^\intercal(K_n + \lambda I_n)^{-1}f_n$ with $f_n = [f(x_1), f(x_2), \ldots, f(x_n)]^\intercal$. Again, with Assumption 1, the second term of RHS of (14) can be bounded by $B\sigma_{n,j}(x)$ where $\sigma_{n,j}(x)$ is a posterior standard deviation of the j-th mean in the n-th episode. The proof strategy of bounding the first term of (14) differs

from that of truncation method. Note that for some $c \in \mathbb{R}$

$$|\phi(x)^\intercal W_n^{-1}\Phi_n^\intercal \tilde{N}_n^j| = \left|\sum_{i=1}^n b_i\tilde{\eta}_{i,j}\right|$$
$$\leq \left|\sum_{i=1}^n b_i\tilde{\eta}_{i,j}\mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}|\leq c\}} - \mathbb{E}[b_i\tilde{\eta}_{i,j}\mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}|\leq c\}}|\mathcal{G}_{i,j}]\right|$$
$$+ \left|\sum_{i=1}^n \mathbb{E}[b_i\tilde{\eta}_{i,j}\mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}|\leq c\}}|\mathcal{G}_{i,j}]\right| + \left|\sum_{i=1}^n \mathbb{E}[b_i\tilde{\eta}_{i,j}\mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}|> c\}}|\mathcal{G}_{i,j}]\right| \tag{15}$$

where $\mathcal{G}_{n,j} \coloneqq \sigma(\{x_1,\dots,x_n\})\cup\{\tilde{\eta}_{1,j},\dots,\tilde{\eta}_{n-1,j}\})$ with $n \in [N]$ and $j \in [\ell]$. Then, by using Cauchy-schwartz inequality and the self-normalized inequality, the first term of RHS of (15) can be bounded by $\tilde{O}(2c\sqrt{\gamma_n})$ with probability at least $1-\delta$. By applying Hölder's inequality and Markov's inequality, we can bound the second term by $h(x_n)^{1+\epsilon}\nu c^{-\epsilon}$ where $h(x_n) \coloneqq (\sum_{i=1}^n |b_{i,j}|^{1+\epsilon})^{\frac{1}{1+\epsilon}}$, $\epsilon \in (0,1]$ and ν is the $(1+\epsilon)$-th central moment of noise distribution. For the third term of RHS of (15), we show that the following holds:

$$\mathbb{P}\{\exists i \in [n] \text{ s.t. } |b_i\tilde{\eta}_{i,j}| > c|\mathcal{G}_{i,j}\} \leq 1/4. \tag{16}$$

This implies that we can bound the third term of RHS of (15) by 0 with probability at least $1-\frac{1}{4}$. In particular, by using union bound and Markov's inequality, we have

$$\mathbb{P}\{\exists i \in [n] \text{ s.t. } |b_i\tilde{\eta}_{i,j}| > c|\mathcal{G}_{i,j}\} \leq \frac{\nu h(x_n)^{1+\epsilon}}{c^{1+\epsilon}}. \tag{17}$$

Then, by setting $c = (4\nu)^{\frac{1}{1+\epsilon}}h(x_n)$, the inequality (16) is satisfied. Combining the inequalities (14) and (15) with their bounds, we have, with probability at least $1-\frac{1}{4}-\delta$,

$$|\mu_{n,j}(x) - f(x)| \leq \alpha_n\sigma_{n,j}(x) \tag{18}$$

where $\alpha_n \coloneqq n^{\frac{1-\epsilon}{2(1+\epsilon)}}(4\nu)^{\frac{1}{1+\epsilon}}\left(2B\lambda^{-1/2}\sqrt{\gamma_n + \ln(1/\delta)} + 4^{-1}\right) + B$. Now, we define the indicator function $X_{n,j} \coloneqq \mathbb{1}_{\{|\mu_{n,j}-f(x)|>\alpha_n\sigma_{n,j}(x)\}}$ and $p_{n,j} \coloneqq \mathbb{P}\{X_{n,j}=1\}$. Note that $p_{n,j} \leq \frac{1}{4}+\delta$ holds by (18). Then applying Azuma-Hoeffding's inequality and union bound yields for all $x \in \mathcal{X}$ and uniformly over all $n \geq 1$,

$$\mathbb{P}\{|\tilde{\mu}_n(x) - f(x)| \leq \alpha_n\sigma_n(x)\} \leq 1-\delta'/T \tag{19}$$

with $\delta' \in (0,1]$. □

Theorem 6 (Cumulative regret bound of Algorithm 4)**.** *Suppose that Assumptions 1 and 2 hold. Let $\delta, \delta' \in (0,1]$ and let $\epsilon \in (0,1]$. For a given total rounds T, Algorithm 4 proceeds $N \coloneqq \lfloor \frac{T}{\ell} \rfloor$ episodes where $\ell \coloneqq (\frac{8\delta}{5})\ln(2T/\delta')$ denotes the length of each episodes. Then, for any $N \in \mathbb{N}$, the cumulative regret bound of Algorithm 4 over N episodes is bounded by*

$$\mathcal{R}_N \leq O(B\nu^{\frac{1}{1+\epsilon}}\gamma_T T^{\frac{1}{1+\epsilon}}\ln(T)) \tag{20}$$

with probability at least $1-\delta'$.

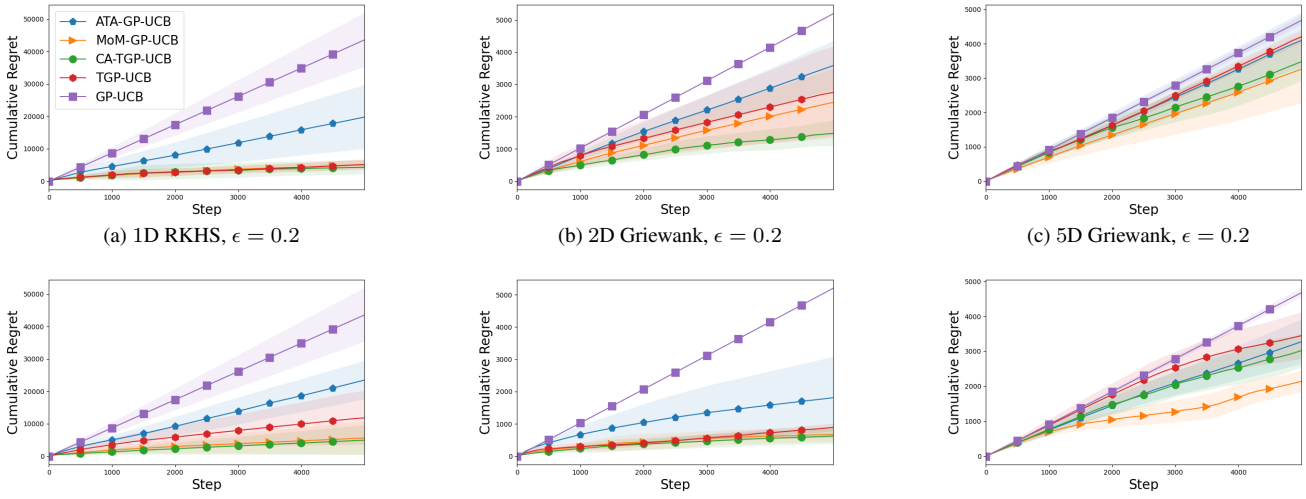
The complete proof is deferred to the supplementary. This theorem presents the cumulative regret bound of MoM-GP-UCB. We would like to note that unlike CP-TGP-UCB, TGP-UCB, and ATA-GP-UCB, whose regret bounds scale with raw moments $\bar{\nu}$, the regret bound of MoM-GP-UCB is scaled with central moments of noise distribution, ν. This attribute makes MoM GP-UCB robust to translations in the noise distribution. Disregarding dimension-independent logarithmic and constant terms, the above bound can be written as $\tilde{O}(T^{\frac{1}{1+\epsilon}}\gamma_T)$. Note that if we set $\epsilon = 1$, the regret bound becomes $\tilde{O}(\sqrt{T}\gamma_T)$ which recovers the regret bound of GP-UCB in sub-Gaussian noise setting [32, 9]. For the Matérn kernel, the bound is $\tilde{O}(T^{\frac{1}{1+\epsilon}+\frac{d}{2\nu+d}}(\ln T)^{\frac{2\nu}{2\nu+d}})$. For the SE kernel, the bound is $\tilde{O}(T^{\frac{1}{1+\epsilon}}(\ln T)^d)$. These are the same results as those of CA-TGP-UCB.

Computational complexity of MoM-GP-UCB. By employing the median-of-means estimator, MoM-GP-UCB achieves better computational complexity than truncation-based algorithms. A key property is that MoM-GP-UCB divides the total rounds T into N episodes. This division reduces the size of the kernel matrix from $O(T^2)$ to $O(N^2)$, resulting in reduced computational costs. In line 7 of Algorithm 4, MoM-GP-UCB computes a posterior mean ℓ times, where ℓ is a length of each episode. Note that $\mathbf{k}_n(x)^\intercal(K_n+\lambda I_n)^{-1}$ need to be computed only once in each episode, requiring $O(n^3+n|\mathcal{X}|)$ time, where $|\mathcal{X}|$ is the cardinality of the input space $\mathcal{X}$. Therefore, the computation of ℓ means has a time complexity of $O(n^3+n|\mathcal{X}|+n\ell)$. Then, in line 8, the median-of-means estimator is obtained in $O(\ell^2)$ time. Finally, in line 9, we can compute $\sigma_n(x)$ for all $x \in \mathcal{X}$ in $O(n|\mathcal{X}|)$ time by using already computed $\mathbf{k}_n(x)^\intercal(K_n+\lambda I_n)^{-1}$ and $\mathbf{k}_n(x)^\intercal$. Since we need to store $\mathbf{k}_n(x)$, K_n, and $\mathbf{k}_n(x)(K_n+\lambda I_n)^{-1}$, the space complexity is $O(n^2)$. Therefore, the total time and space complexities over N episodes are $O(N^4)$ and $O(N^2)$, respectively. This can be rewritten in terms of T as $O(T^4(8\ln(2T/\delta'))^{-4})$ and $O(T^2(8\ln(2T/\delta'))^{-2})$. Importantly, we claim that the term $(8\ln(2T\delta'))^{-4}$ multiplied by T distinctly reduces the actual execution time of MoM-GP-UCB. Additionally, unlike truncation-based algorithms where $|\mathcal{X}|$ is multiplied by T in time and space complexities, in MoM-GP-UCB, they are multiplied by N. This minimizes the impact of $|\mathcal{X}|$ on the execution time of MoM-GP-UCB. We show this experimentally in section 5.

5 Experiment

In this section, we present the experimental results of the proposed algorithms, CA-TGP-UCB and MoM-GP-UCB. The experiments were conducted in both synthetic and real-world datasets. The comparison algorithms include GP-UCB [32] proposed in the sub-Gaussian setting, and TGP-UCB and ATA-GP-UCB [10] proposed in the heavy-tailed setting. For both synthetic and real-world datasets, we generate a heavy-tailed noise by using a Pareto random variable z_t with the moment parameter α_z and the scale parameter λ_z. In addition, we define a Redemacher random variable ζ_t with a $1/2$ probability of being 1 and a $1/2$ probability of being -1. Then, the synthetic noise η_t is defined as $\zeta_t(z_t - \mathbb{E}[z_t])$, whose support is $(-\infty,\infty)$ and mean is zero. We set $\alpha_z \coloneqq (1+\epsilon)+0.01$ to make the $(1+\epsilon)$-th moment of η_t is bounded. In our experiments, we test all algorithms for $\epsilon = 0.2$ and $\epsilon = 0.8$. For 1d function, we generate the target function $f \in \mathcal{H}$ on $[-10,10]$, partitioned into 1000 equally spaced points. Especially, $f \coloneqq \sum_{i=1}^{10} a_i k(\cdot, c_i)$ is randomly generated by sampling a_i and c_i from the normal distributions, $10(-1+2\mathcal{N}(0,1))$ and $\mathcal{N}(0,3)$, respectively. For a d-dimensional, we define the Griewank function as $f(x) \coloneqq 1 + \frac{1}{4000}\sum_{i=1}^d x_i^2 - \prod_{i=1}^d \cos\left(\frac{x_i}{\sqrt{i}}\right)$. When $d=2$, we set the input space $\mathcal{X}$ as $[-5,5]^2$, where the interval is partitioned into 400 evenly space points along both the x and y axes. In addition, for 5d function, we define $\mathcal{X}$ as 5000 points randomly sampled from $\mathcal{N}(0,1)$ with 5 dimensions.

Fig. 1 presents the experimental results on the synthetic datasets. Overall, the algorithms CA-TGP-UCB and MoM-GP-UCB outperform TGP-UCB and ATA-GP-UCB across 1d, 2d, and 5d function



(a) 1D RKHS, $\epsilon = 0.2$ (b) 2D Griewank, $\epsilon = 0.2$ (c) 5D Griewank, $\epsilon = 0.2$

(d) 1D RKHS, $\epsilon = 0.8$ (e) 2D Griewank, $\epsilon = 0.8$ (f) 5D Griewank, $\epsilon = 0.8$

Figure 1: Cumulative regret for synthetic datasets.

Table 1: Comparison of execution time of each algorithm (in seconds) when $\epsilon = 0.2$. The average runtime across 10 seeds is reported, with standard deviations shown in parentheses.

Problem	**GP-UCB**	**TGP-UCB**	**ATA-GP-UCB**	**CA-TGP-UCB**	**MoM-GP-UCB**
1D RKHS func.	132.24s ($\pm$87.61)	73.04s ($\pm$8.86)	7745.75s ($\pm$2206.46)	107.24s ($\pm$26.50)	70.09s ($\pm$7.13)
2D Griewank	98.26s ($\pm$35.93)	32.49s ($\pm$10.45)	1535.51s ($\pm$1285.80)	10.12s ($\pm$1.71)	7.63s ($\pm$0.45)
5D Griewank	9564.76s ($\pm$7898.22)	9573.52s ($\pm$7746.40)	76103.21s ($\pm$41673.41)	3598.58s ($\pm$155.02)	1893.95s ($\pm$40.50)

settings, regardless of the moments of noise distribution. Significantly, the cumulative regret of GP-UCB increases linearly across all problem settings due to its inability to handle heavy-tailed noise. It is noteworthy that CA-TGP-UCB shows the best performance among the truncation-based algorithms in all setups. The improved performance of CA-TGP-UCB over TGP-UCB aligns with theoretical results presented in Theorem 4. However, ATA-GP-UCB shows worse performance than TGP-UCB in 1d and 2d function settings while it marginally outperforms TGP-UCB in 5d function setting. This suggests that the approximation truncation technique does not consistently outperform naïve truncation and performance can vary depending on the problem setting. Importantly, MoM-GP-UCB consistently shows better performance than truncation-based algorithms while it demands much less time and space complexities.

Table 1 provides the execution time of each algorithm when they are conducted on the synthetic datasets with $\epsilon = 0.2$. Across all scenarios, ATA-GP-UCB has the slowest execution time due to the need for additional feature embedding steps. GP-UCB and TGP-UCB share similar execution speeds because the only additional step in TGP-UCB compared to GP-UCB is a naïve truncation with $O(t)$ time complexity. MoM-GP-UCB shows the fastest execution speed across all scenarios, attributed to updating the kernel matrix per episode. This is in contrast to truncation-based algorithms, which require updating the kernel matrix at every time step $t \in [T]$. All algorithms have the longest execution time at 5d function and the shortest execution speed at 2d function. This difference arises from the discretization method of the input space in each problem setup. In the 5d function setting, the domain has 5000 points with 5 dimensions, while in the 2d function setting, the domain is partitioned into 400 points with 2 dimensions. Considering that the cardinality of the domain $\mathcal{X}$ is included in time complexities of all algorithms, it is natural that MoM-GP-UCB, where N is multiplied by $|\mathcal{X}|$, has a shorter execution time compared to truncation-based algorithms where T is multiplied. Table 2 shows the experimental results on real-world stock dataset. Similar to synthetic data settings, CA-TGP-UCB performs the best among the truncation-based algorithms and MoM-GP-UCB outperforms other truncation-based algorithms. The problem setting of real-world dataset is deferred to the supplementary.

Table 2: Experimental results on real-world stock dataset.

Algorithm	**Cumulative regret ($\pm$ std.)**
GP-UCB	29.27 ($\pm$0.088)
TGP-UCB	28.42 ($\pm$0.055)
ATA-GP-UCB	29.04 ($\pm$0.026)
CA-TGP-UCB	27.79 ($\pm$0.002)
MoM-GP-UCB	26.88 ($\pm$0.029)

6 Conclusion

In this paper, we have proposed two Gaussian process (GP) bandit optimization algorithm under heavy-tailed noise. The first algorithm, CA-TGP-UCB, utilizes a truncation estimator and achieves the same regret bound as that of the best existing algorithm up to logarithmic terms with reduced computational complexity. The second algorithm, MoM-GP-UCB, is the first to utilize the median-of-means estimator in GP bandit optimization. Unlike truncation-based algorithms, where the regret bound is expressed in terms of raw moments, the regret bound of MoM-GP-UCB is formulated based on central moments of noise distribution. This characteristic improves the robustness of the algorithm against shifts in noise distribution. In addition, we theoretically demonstrate that MoM-GP-UCB shows better computational complexity than all truncation-based algorithms. We support our theoretical findings through experimental results validating the performance of the proposed algorithms.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00211357).

References

[1] *Contextually Adaptive Algorithms for Gaussian Process Bandit Optimization under Heavy-tailed Noise*, Aug. 2024. Zenodo. doi: 10.5281/zenodo.13345239. URL https://doi.org/10.5281/zenodo.13345239.

[2] S. Agrawal, S. K. Juneja, and W. M. Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR, 2021.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 05 2002. doi: 10.1023/A:1013689704352.

[4] I. Bogunovic and A. Krause. Misspecified gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 34: 3004–3015, 2021.

[5] I. Bogunovic, Z. Li, A. Krause, and J. Scarlett. A robust phased elimination algorithm for corruption-tolerant gaussian process bandits. *arXiv preprint arXiv:2202.01850*, 2022.

[6] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[7] A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(88):2879–2904, 2011.

[8] X. Cai and J. Scarlett. On lower bounds for standard and robust gaussian process bandit optimization. In *International Conference on Machine Learning*, pages 1216–1226. PMLR, 2021.

[9] S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17. JMLR.org, 2017.

[10] S. R. Chowdhury and A. Gopalan. Bayesian optimization under heavy-tailed payoffs. In *Annual Conference on Neural Information Processing Systems*, pages 13790–13801, 2019.

[11] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR Proceedings*, pages 208–214. JMLR.org, April 2011.

[12] R. Dehghannasiri, D. Xue, P. V. Balachandran, M. R. Yousefi, L. A. Dalton, T. Lookman, and E. R. Dougherty. Optimal experimental design for materials discovery. *Computational Materials Science*, 129:311–322, 2017.

[13] A. Durand, O.-A. Maillard, and J. Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *Journal of Machine Learning Research*, 19, 08 2017.

[14] G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.

[15] S. Gupta, S. Rana, S. Venkatesh, et al. Regret bounds for expected improvement algorithms in gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 8715–8737. PMLR, 2022.

[16] D. J. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32, pages 37–45. JMLR.org, June 2014.

[17] D. J. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1): 543–582, 2016.

[18] P. Humbert, B. Le Bars, and L. Minvielle. Robust kernel density estimation with median-of-means principle. In *International Conference on Machine Learning*, pages 9444–9465. PMLR, 2022.

[19] D. Janz, D. Burt, and J. Gonzalez. Bandit optimisation of functions in the matérn kernel rkhs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2486–2495. PMLR, 26–28 Aug 2020.

[20] N. Jaquier, V. Borovitskiy, A. Smolensky, A. Terenin, T. Asfour, and L. D. Rozo. Geometry-aware bayesian optimization in robotics using riemannian matérn kernels. In *Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 794–805. PMLR, November 2021.

[21] 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.

[22] A. Krause and C. Ong. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[23] K. Lee and S. Lim. Minimax optimal bandits for heavy tail rewards. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[24] Z. Li and J. Scarlett. Gaussian process bandit optimization with few batches. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 92–107. PMLR, 28–30 Mar 2022.

[25] S. Lu, G. Wang, Y. Hu, and L. Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4154–4163. PMLR, 2019.

[26] A. M. Medina and S. Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48, pages 1642–1650. JMLR.org, June 2016.

[27] M. Mutny and A. Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[28] S. Salgia, S. Vakili, and Q. Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28836–28847. Curran Associates, Inc., 2021.

[29] J. Scarlett, I. Bogunovic, and V. Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1723–1742. PMLR, 07–10 Jul 2017.

[30] H. Shao, X. Yu, I. King, and M. R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Annual Conference on Neural Information Processing Systems*, pages 8430–8439, December 2018.

[31] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25*, pages 2960–2968, December 2012.

[32] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. of the 27th International Conference on Machine Learning*, pages 1015–1022. Omnipress, 2010.

[33] S. Vakili, N. Bouziani, S. Jalali, A. Bernacchia, and D. shan Shiu. Optimal order simple regret for gaussian process bandits. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[34] S. Vakili, H. Moss, A. Artemev, V. Dutordoir, and V. Picheny. Scalable thompson sampling using sparse gaussian process models. In *Advances in Neural Information Processing Systems*, volume 34, pages 5631–5643. Curran Associates, Inc., 2021.

[35] S. Vakili, J. Scarlett, D. Shiu, and A. Bernacchia. Improved convergence rates for sparse approximation methods in kernel-based learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 21960–21983. PMLR, 2022.

[36] M. Valko, N. Korda, R. Munos, I. N. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2013.

[37] B. Xue, G. Wang, Y. Wang, and L. Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2936–2942. ijcai.org, 2020.

[38] T. Yang, Y.-f. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Table 3: Notations

Symbol	Definition
$\mathcal{X}$	action space
$\mathcal{H}$	Hilbert space
$\phi(x)$	feature mapping from $\mathcal{X}$ to $\mathcal{H}$
x_t	selected action at round t
η_t	noise at round t
y_t	reward at round t
ν	central moment of noise
$\bar{\nu}$	raw moment of noise
f	objective function
B	positive constant such that $\|f\| \leq B$
K_t	kernel matrix at round t
Y_t	vector containing observations y_t defined as $Y_t = [y_1, \ldots, y_t]^\intercal$
$\mathbf{k}_t(x)$	stacked kernel functions
$\mu_t(x)$	posterior mean
$\sigma_t^2(x)$	posterior variance
γ_t	maximum information gain
Φ_t	vector containing features defined as $[\phi_1, \ldots, \phi_t]$
W_t	regularized kernel matrix defined as $(\Phi_t^\intercal \Phi_t + \lambda I_t)$
$\hat{\eta}_t$	truncated noise at round t
$\hat{N}_t$	vector containing truncated noises at round t
λ	regularization parameter
$\mathbb{1}_{\{\cdot\}}$	indicator function

A Proofs of Lemma and Theorem regarding Algorithm 1

A.1 Proof of Lemma 3

Proof. Note that for all $x \in \mathcal{X}$ and any $t \in \mathbb{N}$, the following holds.

$$\begin{aligned}
\hat{\mu}_t(x) - f(x) &= \phi(x)^\intercal \hat{\theta}_t - \phi(x)^\intercal f && (21)\\
&= \phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \hat{Y}_t - \phi(x)^\intercal f && (22)\\
&= \phi(x)^\intercal W_t^{-1} \Phi_t \hat{Y}_t - \phi(x)^\intercal W_t^{-1} (\Phi_t^\intercal \Phi_t + \lambda I_t) f && (23)\\
&= \phi(x)^\intercal W_t^{-1} \Phi_t \hat{Y}_t - \phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \Phi_t f - \lambda \phi(x)^\intercal W_t^{-1} f && (24)\\
&= \phi(x)^\intercal W_t^{-1} \Phi_t \hat{Y}_t - \phi(x)^\intercal W_t^{-1} \Phi_t^\intercal f_t - \lambda \phi(x)^\intercal W_t^{-1} f && (25)\\
&= \phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \hat{N}_t - \lambda \phi(x)^\intercal W_t^{-1} f, && (26)
\end{aligned}$$

where $f_t = [f(x_1), f(x_2), \ldots, f(x_t)]^\intercal$, $\hat{N}_t = [\hat{\eta}_t^1, \hat{\eta}_t^2, \ldots, \hat{\eta}_t^t]^\intercal$ with $\hat{\eta}_t^i = y_i \mathbb{1}_{\{|b_t^i y_i| \leq h_t(x)\}} - f(x_i)$ for all $i \in [t]$. Note that b_t^i and $h_t(x)$ are a weight of the ith observation and truncation threshold, respectively, which will be defined later. By the same argument in the proof of Lemma 8 in [10], we have

$$\lambda \phi(x)^\intercal W_t^{-1} f = f(x) - \zeta_t(x), \tag{27}$$

where $\zeta_t(x) \coloneqq \mathbf{k}_t(x)^\intercal (K_t + \lambda I_t)^{-1} f_t$. Then, by using triangle inequality, we obtain

$$|\hat{\mu}_t(x) - f(x)| \leq |\phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \hat{N}_t| + |f(x) - \zeta_t(x)|. \tag{28}$$

The proof strategy is to upper bound the first and second term in RHS of the inquality (28), respectively. For the second term of the inequality (28), since $f(x) - \zeta_t(x) = \lambda \langle W_t^{-1/2} \phi(x), W_t^{-1/2} f \rangle_H$ holds by reproducing property, we have for all $x \in \mathcal{X}$, by using Cauchy-Schwartz inequality,

$$|f(x) - \zeta_t(x)| \leq \lambda \|\phi(x)\|_{W_t^{-1}} \|f\|_{W_t^{-1}} \leq \lambda^{1/2} \|\phi(x)\|_{W_t^{-1}} \|f\|_{\mathcal{H}} \leq B \sigma_t(x) \tag{29}$$

where $W_t^{-1} \prec \lambda^{-1} I_{\mathcal{H}}$, B is some constant such that $\|f\|_{\mathcal{H}} \leq B$ (Assumption 1), and $\sigma_t(x)$ is a poseterior variance defined as $\sigma_t(x) = \lambda^{1/2} \|\phi(x)\|_{W_t^{-1}}$. For the first term of the inequality (28), note that

$$\begin{aligned}
|\phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \hat{N}_t| &= |\phi(x)^\intercal (\Phi_t^\intercal \Phi_t + \lambda I_{\mathcal{H}})^{-1} \Phi_t^\intercal \hat{N}_t| && (30)\\
&= |\phi(x)^\intercal \Phi_t^\intercal (\Phi_t \Phi_t^\intercal + \lambda I_t)^{-1} \hat{N}_t| && (31)\\
&= |\mathbf{k}_t(x)^\intercal (K_t + \lambda I_t)^{-1} \hat{N}_t|. && (32)
\end{aligned}$$

Now, without loss of generality, let us denote, for a fixed $x \in \mathcal{X}$,

$$\beta_t = \phi(x)^\intercal W_t^{-1} \Phi_t^\intercal = \mathbf{k}_t(x)^\intercal (K_t + \lambda I_t)^{-1} = [b_t^1, b_t^2, \dots, b_t^t], \tag{33}$$

where, now, we set b_t^i as each element of vector β_t. Further, let us define a σ-algebra as $\mathcal{G}_{t,\tau} = \sigma(\{x_1, x_2, \dots, x_t\} \cup \{y_1, y_2, \dots, y_\tau\})$ with $t \in \mathbb{N}$ and $\tau \in [t]$. Then, β_t is $\mathcal{G}_{t,0}$-measurable. Note that the first term of the inequality (28) can be decomposed as follows,

$$\begin{aligned}
\left| \mathbf{k}_t(x)^\intercal (\Phi_t \Phi_t^\intercal + \lambda I_t)^{-1} \hat{N}_t \right| &= \left| \sum_{i=1}^t b_t^i \hat{\eta}_t^i \right| && (34)\\
&= \left| \sum_{i=1}^t b_t^i \hat{\eta}_t^i - \mathbb{E}\left[\sum_{i=1}^t b_t^i \hat{\eta}_t^i \middle| \mathcal{G}_{t,i-1} \right] + \mathbb{E}\left[\sum_{i=1}^t b_t^i \hat{\eta}_t^i \middle| \mathcal{G}_{t,i-1} \right] \right| && (35)\\
&\leq \left| \sum_{i=1}^t b_t^i \hat{\eta}_t^i - \mathbb{E}\left[\sum_{i=1}^t b_t^i \hat{\eta}_t^i \middle| \mathcal{G}_{t,i-1} \right] \right| + \left| \mathbb{E}\left[\sum_{i=1}^t b_t^i \hat{\eta}_t^i \middle| \mathcal{G}_{t,i-1} \right] \right| && (36)
\end{aligned}$$

From the definition of a filtration $\mathcal{G}_{t,\tau}$, $\mathbb{E}[|y_i|^{1+\epsilon} | \mathcal{G}_{t,i-1}] \leq \bar{\nu}$ holds for all $i \in [t]$. Since the sampling processes of noise are independent of the arm played, we can simply write $\mathbb{E}[|y_i|^{1+\epsilon} | \mathcal{G}_{i-1}] \leq \bar{\nu}$, which is the same representation in Assumption 2. Now, consider the second term of the inequality (36) as follows,

$$\begin{aligned}
\left| \mathbb{E}\left[\sum_{i=1}^t b_t^i \hat{\eta}_t^i \middle| \mathcal{G}_{t,i-1} \right] \right| &= \left| \sum_{i=1}^t \mathbb{E}\Big[b_t^i \big(y_i \mathbb{1}_{\{|b_t^i y_i| \leq h_t(x)\}} - f(x_i) \big) \Big| \mathcal{G}_{t,i-1} \Big] \right| && (37)\\
&= \left| \sum_{i=1}^t \mathbb{E}\Big[b_t^i \cdot -y_i \mathbb{1}_{\{|b_t^i y_i| > h_t(x)\}} \Big| \mathcal{G}_{t,i-1} \Big] \right| && (38)\\
&\leq \sum_{i=1}^t \mathbb{E}\Big[|b_t^i y_i| \mathbb{1}_{\{|b_t^i y_i| > h_t(x)\}} | \mathcal{G}_{t,i-1} \Big] && (39)\\
&\leq \sum_{i=1}^t (\mathbb{E}[|b_t^i y_i|^{1+\epsilon} | \mathcal{G}_{t,i-1}])^{\frac{1}{1+\epsilon}} \mathbb{P}(|b_t^i y_i| > h_t(x) | \mathcal{G}_{t,i-1})^{\frac{\epsilon}{1+\epsilon}} && \text{(Hölder's inequality)} \quad (40)\\
&\leq \sum_{i=1}^t |b_t^i| \bar{\nu}^{\frac{1}{1+\epsilon}} \frac{\mathbb{E}[|b_t^i y_i|^{1+\epsilon} | \mathcal{G}_{t,i-1}]^{\frac{\epsilon}{1+\epsilon}}}{h_t(x)^\epsilon} && \text{(Markov's inequality)} \quad (41)\\
&\leq \sum_{i=1}^t |b_t^i|^{1+\epsilon} \frac{\bar{v}}{h_t(x)^\epsilon} && \text{(Assumption 2)} \quad (42)\\
&= \bar{v} h_t(x) && (43)
\end{aligned}$$

where $h_t(x) := \|\phi(x)^\intercal W_t^{-1} \Phi_t^\intercal\|_{1+\epsilon} = (\sum_{i=1}^t |b_t^i|^{1+\epsilon})^{\frac{1}{1+\epsilon}}$. Here, now, we define $h_t(x)$ as the same as in the Algorithm 1. For the first term of the inequality (36), we have

$$\begin{aligned}
\left| \sum_{i=1}^t b_t^i \hat{\eta}_t^i - \mathbb{E}[b_i \hat{\eta}_t^i | \mathcal{G}_{t,i-1}] \right| &= \left| \phi(x)^\intercal W_t^{-1} \Phi_t^\intercal \hat{\xi}_t \right| && (44)\\
&= \left| \phi(x)^\intercal (\Phi_t^\intercal \Phi_t + \lambda I_t)^{-1} \Phi_t^\intercal \hat{\xi}_t \right| && (45)\\
&= \left| \phi(x)^\intercal (\Phi_t^\intercal \Phi_t + \lambda I_t)^{-1} \Phi_t^\intercal \hat{\xi}_t \right| && (46)\\
&= |\langle W_t^{-1/2} \phi(x), W_t^{-1/2} \Phi_t^\intercal \hat{\xi}_t \rangle_{\mathcal{H}}| && (47)
\end{aligned}$$

where $\hat{\xi}_t = [\hat{\xi}_t^1, \hat{\xi}_t^2, \dots, \hat{\xi}_t^t]$ with $\hat{\xi}_t^i := \hat{\eta}_t^i - \mathbb{E}[\hat{\eta}_t^i | \mathcal{G}_{t,i-1}] = \hat{y}_t^i - \mathbb{E}[\hat{y}_t^i | \mathcal{G}_{t,i-1}]$ for all $i \in [t]$ and some fixed t. Then, by Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
\left| \sum_{i=1}^t b_t^i \hat{\eta}_t^i - \mathbb{E}[b_t^i \hat{\eta}_t^i | \mathcal{G}_{t,i-1}] \right| &\leq \|\phi(x)\|_{W_t^{-1}} \|\Phi_t^\intercal \hat{\xi}_t\|_{W_t^{-1}} && (48)\\
&= \lambda^{-1/2} \sigma_t(x) \|\Phi_t^\intercal \hat{\xi}_t\|_{W_t^{-1}} && (49)
\end{aligned}$$

Note that $|\hat{\xi}_t^i| \leq 2h_t(x)$ holds for all $i \in [t]$ and some fixed $t \in \mathbb{N}$. This shows that $\hat{\xi}_t^i$ is $2h_t(x)$-sub-Gaussian random variable conditioned on $\mathcal{F}_{i-1}$ with zero mean. From the definition of a filtration $\mathcal{F}_t$, x_i is $\mathcal{F}_{i-1}$-measurable for all $i \in [t-1]$. In addition, since y_i is $\mathcal{F}_i$-measurable

$\hat{y}_t^i$ is also F_i-measurable. Thus, applying Lemma 8 yields, for any $\delta \in (0, 1]$, uniformly over all $t \geq 1$,

$$\|\Phi_t^\intercal \hat{\xi}_t\|_{W_t^{-1}} \leq 2h_t(x)\sqrt{2\left(\frac{1}{2}\ln|I_{\mathcal{H}} + \lambda^{-1}\Phi_t\Phi_t^\intercal| + \ln(1/\delta)\right)} \tag{50}$$

$$= 2h_t(x)\sqrt{2\left(\frac{1}{2}\ln|I_t + \lambda^{-1}K_t| + \ln(1/\delta)\right)} \tag{51}$$

with probability at least $1 - \delta$. By combining the (in)equalities (36), (49), (51), and (43), we have

$$|\phi(x)^\intercal W_t^{-1}\Phi_t^\intercal \hat{N}_t| \leq 2\lambda^{-1/2}\sigma_t(x)h_t(x)\sqrt{2\left(\frac{1}{2}\ln|I_t + \lambda^{-1}K_t| + \ln(1/\delta)\right)} + \bar{\nu}h_t(x). \tag{52}$$

Note that we have,

$$B \geq \sqrt{k(x,x)} \geq \sigma_t(x) \tag{53}$$

$$= \lambda^{1/2}\|\phi(x)\|_{W_t^{-1}} \tag{54}$$

$$= \lambda^{1/2}\sqrt{\phi(x)^\intercal W_t^{-1}(\Phi_t^\intercal\Phi_t + \lambda I_{\mathcal{H}})W_t^{-1}\phi(x)} \tag{55}$$

$$\geq \lambda^{1/2}\sqrt{\phi(x)^\intercal W_t^{-1}(\Phi_t^\intercal\Phi_t)W_t^{-1}\phi(x)} \tag{56}$$

$$= \lambda^{1/2}\|\phi(x)W_t^{-1}\Phi_t^\intercal\|_2 \tag{57}$$

$$\geq \lambda^{1/2}t^{\frac{\epsilon-1}{2(1+\epsilon)}}\|\phi(x)W_t^{-1}\Phi_t^\intercal\|_{1+\epsilon} \tag{58}$$

where the last inequality follows from Hölder's inequality. Then $h(x_t) \leq \lambda^{-1/2}t^{\frac{1-\epsilon}{2(1+\epsilon)}}\sigma_t(x)$ holds for all $t \in \mathbb{N}$, which implies that

$$|\phi(x)^\intercal W_t^{-1}\Phi_t^\intercal \hat{N}_t| \leq \lambda^{-1/2}\sigma_t(x)t^{\frac{1-\epsilon}{2(1+\epsilon)}}\left(2B\lambda^{-1/2}\sqrt{2\left(\frac{1}{2}\ln|I_t + \lambda^{-1}K_t| + \ln(1/\delta)\right)} + \bar{\nu}\right). \tag{59}$$

By substituting the inequality (59) and (29) to the inequality (28), we have, with probability at least $1 - \delta$,

$$|\hat{\mu}_t(x) - f(x)| \leq \sigma_t(x)\left\{B + \lambda^{-1/2}t^{\frac{1-\epsilon}{2(1+\epsilon)}}\left(2B\lambda^{-1/2}\sqrt{2\left(\frac{1}{2}\ln|I_t + \lambda^{-1}K_t| + \ln(1/\delta)\right)} + \bar{\nu}\right)\right\}. \tag{60}$$

By defining $\alpha_{t+1} \coloneqq \left\{B + \lambda^{-1/2}t^{\frac{1-\epsilon}{2(1+\epsilon)}}\left(2B\lambda^{-1/2}\sqrt{2\left(\frac{1}{2}\ln|I_t + \lambda^{-1}K_t| + \ln(1/\delta)\right)} + \bar{\nu}\right)\right\}$, the lemma is established. □

A.2 Proof of Theorem 4

Proof. By Lemma 3, for any $\delta \in (0, 1]$ and uniformly over all $t \geq 1$, the following holds

$$r_t = f(x_*) - f(x_t) \tag{61}$$

$$\leq \hat{\mu}_{t-1}(x_*) + \alpha_t\sigma_{t-1}(x_*) - f(x_t) \tag{62}$$

$$\leq \hat{\mu}_{t-1}(x_t) + \alpha_t\sigma_{t-1}(x_t) - f(x_t) \qquad \text{(the selection rule of Algorithm 1)} \tag{63}$$

$$\leq 2\alpha_t\sigma_{t-1}(x_t) \tag{64}$$

with probability at least $1 - \delta$. Then, from the definition of γ_t, it is clear that $\alpha_t \leq (B + \lambda^{-1/2}t^{\frac{1-\epsilon}{2(1+\epsilon)}}(2B\lambda^{-1/2}\sqrt{\gamma_t + \ln(1/\delta)}) + \bar{\nu})$. In addition, by Lemma 9 and Cauchy-Schwartz inequality, we have $\sum_{t=1}^T \sigma_{t-1}(x_t) \leq \sqrt{2(1+\lambda)\gamma_t T}$. Hence, for any $\delta \in (0, 1]$, the cumulative regret of Algorithm 1 after total round T is as follows,

$$\sum_{t=1}^T r_t \leq \sum_{t=1}^T 2\alpha_t\sigma_{t-1}(x_t) \tag{65}$$

$$\leq \alpha_T\sum_{t=1}^T \sigma_{t-1}(x_t) \tag{66}$$

$$\leq 2(B + \lambda^{-1/2}T^{\frac{1-\epsilon}{2(1+\epsilon)}}(2B\lambda^{-1/2}\sqrt{\gamma_T + \ln(1/\delta)} + \bar{\nu}))\sqrt{2(1+\lambda)\gamma_T T} \tag{67}$$

$$\leq O\left((B\gamma_T + \bar{\nu}\sqrt{\gamma_T})T^{\frac{1}{1+\epsilon}}\right) \tag{68}$$

with probability at least $1 - \delta$. □

B Proofs of Lemma and Theorem regarding Algorithm 4

Lemma 7 (Confidence interval of the j-th evaluation)**.** *Suppose that Assumptions 1 and 2 hold. Let $\delta \in (0,1]$, $\epsilon \in (0,1]$, $\ell \coloneqq (\frac{8\delta}{5})\ln(2T/(\delta+1/4))$ and $\alpha_n \coloneqq n^{\frac{1-\epsilon}{2(1+\epsilon)}}(4\nu)^{\frac{1}{1+\epsilon}}\left(2B\lambda^{-1/2}\sqrt{\gamma_n + \ln(1/\delta)} + 4^{-1}\right) + B$. Then the j-th mean estimate in the n-th episode $\mu_{n,j}$ satisfies, for all $x \in \mathcal{X}$ and $j \in [\ell]$, and uniformly over all $n \geq 1$,*

$$|\mu_{n,j}(x) - f(x)| \leq \alpha_n \sigma_{n,j}(x) \tag{69}$$

with probability at least $1 - \frac{1}{4} - \delta$. Here, $\sigma_{n,j}(x)$ is a posterior standard deviation of $\mu_{n,j}$.

Proof. Note that, by the same argument in the proof of Lemma 3, we have for all $x \in \mathcal{X}$ and $j \in [\ell]$,

$$|\mu_{n,j}(x) - f(x)| \leq |\phi(x)^\intercal W_n^{-1} \Phi_n^\intercal \tilde{N}_n^j| + |f(x) - \zeta_n(x)| \tag{70}$$

where $\tilde{N}_n^j \coloneqq [\tilde{\eta}_{1,j}, \tilde{\eta}_{2,j} \ldots, \tilde{\eta}_{n,j}]^\intercal$ with $\tilde{\eta}_{n,j} = y_{n,j} - f(x_{n,j})$ for all $n \in [N]$, and $\zeta_n \coloneqq \mathbf{k}_n(x)^\intercal (K_n + \lambda I_n)^{-1} f_n$ with $f_n = [f(x_1), f(x_2), \ldots, f(x_n)]^\intercal$. From the Assumption 1, $|f(x) - \zeta_n(x)|$ is bounded by $B\sigma_{n,j}(x)$ where B is some constant such that $\|f\|_{\mathcal{H}} \leq B$ and $\sigma_{n,j}$ is a posterior standard deviation of the j-th evaluation in the n-th episode. Now, let us consider the first term of RHS of the inequality (70). Choose some $c \in \mathbb{R}$ and denote the i-th element of $\phi(x)^\intercal W_n^{-1} \Phi_n^\intercal$ as b_i for all $i \in [n]$. Then we can decompose the first term as follows:

$$\begin{aligned}
|\phi(x)^\intercal W_n^{-1} \Phi_n^\intercal \tilde{N}_n^j| &= \left|\sum_{i=1}^n b_i \tilde{\eta}_{i,j}\right| && (71)\\
&\leq \left|\sum_{i=1}^n b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| \leq c\}} - \mathbb{E}[b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| \leq c\}} | \mathcal{G}_{i,j}]\right| + \left|\sum_{i=1}^n \mathbb{E}[b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| \leq c\}} | \mathcal{G}_{i,j}]\right| && (72)\\
&+ \left|\sum_{i=1}^n \mathbb{E}[b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| > c\}} | \mathcal{G}_{i,j}]\right| && (73)
\end{aligned}$$

where $\mathcal{G}_{n,j} \coloneqq \sigma(\{x_1, \ldots, x_n\}) \cup \{\tilde{\eta}_{1,j}, \ldots, \tilde{\eta}_{n-1,j}\})$ with $n \in [N]$ and $j \in [\ell]$. By using Cauchy-Schwartz inequality, the first term of the inequality (73) can be bounded as follows:

$$\begin{aligned}
\left|\sum_{i=1}^n b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| \leq c\}} - \mathbb{E}[b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| \leq c\}} | \mathcal{G}_{i,j}]\right| &\leq \left|\sum_{i=1}^n b_i \tilde{\xi}_{i,j}\right| = \left|\phi(x)^\intercal W_n^{-1} \Phi_n^\intercal \tilde{\xi}_n^j\right| && (74)\\
&\leq \lambda^{-1/2} \sigma_{n,j}(x) \|\Phi_n^\intercal \tilde{\xi}_n^j\|_{W_n^{-1}} && (75)
\end{aligned}$$

where $\tilde{\xi}_n^j \coloneqq [\tilde{\xi}_{1,j}, \ldots, \tilde{\xi}_{n,j}]$ and $\tilde{\xi}_{i,j} = \tilde{\eta}_{i,j} - \mathbb{E}[\tilde{\eta}_{i,j} | \mathcal{G}_{i,j}]$ for all $i \in [n]$. Since $|\tilde{\eta}_{i,j}| \leq 2c$ holds for all $i \in [n]$, by a similar argument in the proof of Lemma 3 and applying Lemma 8, we obtain

$$\|\Phi_n^\intercal \tilde{\xi}_n^j\|_{W_n^{-1}} \leq 2c\sqrt{2\left(\frac{1}{2}\ln|I_n + \lambda^{-1}K_n| + \ln\left(\frac{1}{\delta}\right)\right)}. \tag{76}$$

For the second term of the inequality (70), notice that

$$\begin{aligned}
\left|\sum_{i=1}^n \mathbb{E}[b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| \leq c\}} | \mathcal{G}_{i,j}]\right| &= \left|-\sum_{i=1}^n \mathbb{E}[b_i \tilde{\eta}_{i,j} \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| > c\}} | \mathcal{G}_{i,j}]\right| && (77)\\
&\leq \sum_{i=1}^n \mathbb{E}[|b_i \tilde{\eta}_{i,j}| \mathbb{1}_{\{|b_i\tilde{\eta}_{i,j}| > c\}} | \mathcal{G}_{i,j}] && (78)\\
&\leq \sum_{i=1}^n (\mathbb{E}[|b_i \tilde{\eta}_{i,j}|^{1+\epsilon} | \mathcal{G}_{i,j}])^{\frac{1}{1+\epsilon}} \mathbb{P}\{|b_i\tilde{\eta}_{i,j}| > c | \mathcal{G}_{i,j}\}^{\frac{\epsilon}{1+\epsilon}} && (79)\\
&\leq \sum_{i=1}^n (\mathbb{E}[|b_i \tilde{\eta}_{i,j}|^{1+\epsilon} | \mathcal{G}_{i,j}])^{\frac{1}{1+\epsilon}} \frac{\mathbb{E}[|b_i\tilde{\eta}_{i,j}|^{1+\epsilon} | \mathcal{G}_{i,j}]^{\frac{\epsilon}{1+\epsilon}}}{c^\epsilon} && (80)\\
&\leq \sum_{i=1}^n |b_i|^{1+\epsilon} \frac{\nu}{c^\epsilon} = h(x_n)^{1+\epsilon} \nu c^{-\epsilon}. && (81)
\end{aligned}$$

where we define $h(x_n) = (\sum_{i=1}^n |b_i|^{1+\epsilon})^{\frac{1}{1+\epsilon}}$, and the second inequality follows from Hölder's inequality and the third inequality uses Markov's inequality. Now, we claim that the third term of the inequality (73) can be bounded by 0 with probability at least $1 - \frac{1}{4}$. To prove this, we will show that the following inequality holds:

$$\mathbb{P}\{\exists i \in [n] \text{ s.t. } |b_i\tilde{\eta}_{i,j}| > c | \mathcal{G}_{i,j}\} \leq \frac{1}{4}. \tag{82}$$

Note that, by union bound and Markov's inequality, we have

$$\mathbb{P}\{\exists i \in [n] \text{ s.t. } |b_i \tilde{\eta}_{i,j}| > c|\mathcal{G}_{i,j}\} \leq \sum_{i=1}^{n} \mathbb{P}\{|b_i \tilde{\eta}_{i,j}| > c|\mathcal{G}_{i,j}\} \tag{83}$$

$$\leq \sum_{i=1}^{n} \frac{\mathbb{E}[|b_i \tilde{\eta}_{i,j}|^{1+\epsilon}|\mathcal{G}_{i,j}]}{c^{1+\epsilon}} \tag{84}$$

$$\leq \sum_{i=1}^{n} |b_i|^{1+\epsilon} \frac{\nu}{c^{1+\epsilon}} = \frac{\nu h(x_n)^{1+\epsilon}}{c^{1+\epsilon}}. \tag{85}$$

Here, if we set $c = (4\nu)^{\frac{1}{1+\epsilon}} h(x_n)$, then, the probability can be bounded by $\frac{1}{4}$. This implies that the third term of the inequality (70) can be bounded by 0 with probability at least $1 - \frac{1}{4}$. Finally, by combining the inequalities (70), (73), (75), (81), and (82), we have for all $x \in \mathcal{X}$ and $j \in [\ell]$, and uniformly over all $n \geq 1$,

$$|\mu_{n,j}(x) - f(x)| \leq |\phi(x)^\intercal W_n^{-1} \Phi_n^\intercal \tilde{N}_n^j| + |f(x) - \zeta_n(x)| \tag{86}$$

$$\leq \lambda^{-1/2} \sigma_{n,j}(x) 2c \sqrt{2\left(\frac{1}{2}\ln|I_n + \lambda^{-1}K_n| + \ln(1/\delta)\right)} + h(x_n)^{1+\epsilon} c^{-\epsilon} \nu + 0 + B\sigma_{n,j}(x) \tag{87}$$

$$\leq \lambda^{-1/2} B 2c \sqrt{2\left(\frac{1}{2}\ln|I_n + \lambda^{-1}K_n| + \ln(1/\delta)\right)} + h(x_n)^{1+\epsilon} c^{-\epsilon} \nu + B\sigma_{n,j}(x) \tag{88}$$

$$\leq \lambda^{-1/2} B 2(4\nu)^{\frac{1}{1+\epsilon}} h(x_n) \sqrt{2\left(\frac{1}{2}\ln|I_n + \lambda^{-1}K_n| + \ln(1/\delta)\right)} + 4^{-\frac{\epsilon}{1+\epsilon}} \nu^{\frac{1}{1+\epsilon}} h(x_n) + B\sigma_{n,j}(x) \tag{89}$$

$$\leq \sigma_{n,j}(x) \left\{ n^{\frac{1-\epsilon}{2(1+\epsilon)}} (4\nu)^{\frac{1}{1+\epsilon}} \left(2B\lambda^{-1/2}\sqrt{\gamma_n + \ln(1/\delta)} + 4^{-1}\right) + B \right\} \tag{90}$$

with probability at least $1 - \frac{1}{4} - \delta$. Here, the third inequality follows from the fact that $h(x_n) \leq n^{\frac{1-\epsilon}{2(1+\epsilon)}} \lambda^{-1/2} \sigma_{n,j}(x)$ and definition of γ_n. Letting $\alpha_n \coloneqq n^{\frac{1-\epsilon}{2(1+\epsilon)}} (4\nu)^{\frac{1}{1+\epsilon}} \left(2B\lambda^{-1/2}\sqrt{\gamma_n + \ln(1/\delta)} + 4^{-1}\right) + B$ completes the proof. □

B.1 Proof of Lemma 5

Proof. By Lemma 7, we have for all $x \in \mathcal{X}$ and uniformly over all $j \geq 1$ and $n \in [N]$,

$$|\mu_{n,j}(x) - f(x)| \leq \alpha_n \sigma_{n,j}(x) \tag{91}$$

with probability at least $1 - \frac{1}{4} - \delta$. Let us define the indicator function $X_{n,j} \coloneqq \mathbb{1}_{\{|\mu_{n,j}(x) - f(x)| > \alpha_n \sigma_{n,j}(x)\}}$ and let $p_{n,j} \coloneqq \mathbb{P}\{X_{n,j} = 1\}$ for all $n \in [N]$ and $j \in [\ell]$. From Lemma 7, we know that $p_{n,j} \leq \frac{1}{4} + \delta$ holds. Applying Azuma-Hoeffding's inequality yields for any $n \in [N]$,

$$\mathbb{P}\Big\{\sum_{j=1}^{\ell} X_{n,j} \geq \frac{\ell}{2}\Big\} = \mathbb{P}\Big\{\sum_{j=1}^{\ell} X_{n,j} - p_{n,j} \geq \frac{\ell}{4} - \delta\ell\Big\} \tag{92}$$

$$\leq \exp(-\frac{\ell}{8} - \frac{\delta\ell}{2}) \leq \exp(-\frac{\ell}{8}) \tag{93}$$

$$= \frac{\delta'}{2T}. \tag{94}$$

This implies that at least half of the $j \in [\ell]$ evaluations satisfies $\mu_{n,j}(x) - f(x) > \alpha_n \sigma_{n,j}(x)$ with probability at most $\frac{\delta'}{2T}$. Hence, the median of the j evaluations in the n-th episode, denoted as $\tilde{\mu}_n$, satisfies

$$\tilde{\mu}_n(x) - f(x) \leq \alpha_n \sigma_n(x) \tag{95}$$

with probability at least $1 - \frac{\delta'}{2T}$. Here, $\alpha_n \coloneqq n^{\frac{1-\epsilon}{2(1+\epsilon)}} (4\nu)^{\frac{1}{1+\epsilon}} \left(2B\lambda^{-1/2}\sqrt{\gamma_n + \ln(1/\delta)} + 4^{-1}\right) + B$. By a similar argument and union bound, we have

$$\mathbb{P}\{|\tilde{\mu}_n(x) - f(x)| \leq \alpha_n \sigma_n(x)\} \leq 1 - \frac{\delta'}{T} \tag{96}$$

with $\delta' \in (0, 1]$. This completes the proof. □

B.2 Proof of Theorem 6

Proof. We define the length of each episode by $\ell := 8\ln(\frac{2T}{\delta'})$. Then, by Lemma 5, we have that for all $x \in \mathcal{X}$, with probability at least $1-\delta'/T$,

$$|\tilde{\mu}_n(x) - f(x)| \leq \alpha_n \sigma_n(x) \tag{97}$$

where $\tilde{\mu}_n$ denotes the median-of-means estimator in the n-th episode and α_n is a confidence parameter defined in Lemma 5. Note that the instantaneous regret can be bounded by

$$\begin{aligned}
r_n &= f(x_*) - f(x_n) && && (98)\\
&\leq \tilde{\mu}_n(x_*) + \alpha_n\sigma_n(x) - f(x_n) && \text{(Lemma 5)} && (99)\\
&\leq \tilde{\mu}_n(x_n) + \alpha_n\sigma_n(x) - f(x_n) && \text{(the optimistic selection rule)} && (100)\\
&\leq 2\alpha_n\sigma_n(x) && && (101)
\end{aligned}$$

with probability at least $1-\delta'/T$. Here, x_* is an optimal point. Therefore, the cumulative regret bound over N episodes can be bounded by

$$\begin{aligned}
\mathcal{R}_N = \ell\sum_{n=1}^{N} r_n &\leq \ell\sqrt{N\sum_{n=1}^{N} r_n^2} && \text{(Cauchy-Schwartz inequality)} && (102)\\
&\leq 2\ell\alpha_N\sqrt{N\sum_{n=1}^{N}\sigma_n^2(x)} && && (103)\\
&\leq 2\ell\sqrt{N}\alpha_N\sqrt{2(1+\lambda)\gamma_N} && && (104)\\
&\leq 2\ell\sqrt{N}\left\{N^{\frac{1-\epsilon}{2(1+\epsilon)}}(4\nu)^{\frac{1}{1+\epsilon}}(2B\lambda^{-1/2}\sqrt{\gamma_N+\ln(1/\delta)}+4^{-1})+B\right\}\sqrt{2(1+\lambda)\gamma_N} && && (105)\\
&\leq O\left(2\ell N^{\frac{1}{1+\epsilon}}(4\nu)^{\frac{1}{1+\epsilon}}(2B\lambda^{-1/2}\sqrt{(\gamma_N+\ln(1/\delta))}+4^{-1})\sqrt{\gamma_N}+2\ell B\sqrt{N\gamma_N}\right) && && (106)\\
&\leq O(B\nu^{\frac{1}{1+\epsilon}}\gamma_T T^{\frac{1}{1+\epsilon}}\ln(T)) && && (107)
\end{aligned}$$

with probability at least $1-\delta'$, as desired. □

C Technical Lemmas

Lemma 8 (in [13])**.** *Let $\{z_t\}_{t\geq 1}$ be an $\mathbb{R}^d$-valued discrete time stochastic processes such that z_t is predictable with respect to filtration $\{G_t\}_{t\geq 0}$, i.e., z_t is G_{t-1}-measurable for all $t \geq 1$. Let $\{w_t\}_{t\geq 1}$ be a real-valued stochastic proces such that for all $t \geq 1$, w_t is (a) G_t-measurable, and (b) R-sub-Gaussian conditionally on G_{t-1} for some $R > 0$. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$, uniformly over all $t \geq 1$,*

$$\left\|\sum_{\tau=1}^{t} w_\tau\phi(z_\tau)\right\|_{Z_t^{-1}} \leq R\sqrt{2\left(\frac{1}{2}\ln\frac{|Z_t|}{|Z|}+\ln(\frac{1}{\delta})\right)} \tag{108}$$

where $Z_t = Z + \sum_{\tau=1}^{t}\phi(z_\tau)\phi(z_\tau)^\intercal$ and $Z : \mathcal{H}_k(\mathbb{R}^d) \to \mathcal{H}_k(\mathbb{R}^d)$ is a positive definite operator.

Lemma 9 (in [10])**.** *If $k(x,x) \leq 1$ for all $x \in \mathcal{X}$, then $\sum_{s=1}^{t}\sigma_{s-1}^2(x_s) \leq 2(1+\lambda)\gamma_t$.*

D Experimental setting for real-world stock datasets

We analyze stock market data from Apple, Baidu, and Hawaiian Holdings, all of which are listed on the NASDAQ 100 index. The dataset includes the rate of change, l_t, for open (o), high (h), low (l), and close price (c), as well as the trading volume (v) from the time step $t-1$ to t. For each type $i \in \{o,h,l,c,v\}$, the change rate l_t^i is calculated as $l_t^i = \frac{p_t^i - p_{t-1}^i}{p_{t-1}^i}$. We utilize stock market data from January 2010 to January 2022, with the training and test periods defined as detailed in Table 4. The reward at time step t is defined as $r_t := \max_s(l_{s,t+1}^c) - l_{s',t+1}^c$. Here, the first term represents the maximum change rate based on the close price from t to $t+1$ across all stocks s. The second term indicates the change rate for the close price of stock s', which is predicted to have the maximum change rate from t to $t+1$.

Table 4: Train and Test data split with stock market data of Apple (AAPL), Baidu (BIDU), and Hawaiian Holdings (HA) in the NASDAQ 100 index.

Company	Training Period	Test Period
AAPL, BIDU, HA	Jan 2010 - Jan 2018	Jan 2018 - Jan 2022

Algorithm 3 Median-of-means (MoM) GP-UCB

Input: T, $\mathcal{X}$, $\alpha_t \in \mathbb{R}^+$, $\lambda > 0$, k, $\epsilon \in (0, 1]$, $\delta, \delta' \in (0, 1]$, B
Initialization: Set $\ell = 8 \ln(\frac{2T}{\delta'})$, $N = \lfloor \frac{T}{\ell} \rfloor$, $\tilde{\mu}_0 = 0$
for $n = 1, 2, \ldots, N$ **do**
  Set $\alpha_n = n^{\frac{1-\epsilon}{2(1+\epsilon)}} (4\nu)^{\frac{1}{1+\epsilon}} \left(2B\lambda^{-\frac{1}{2}} \sqrt{\gamma_n + \ln(\frac{1}{\delta})} + \frac{1}{4} \right) + B$
  Select action $x_n = \arg\max_{x \in \mathcal{X}} \tilde{\mu}_{n-1}(x) + \alpha_n \sigma_{n-1}(x)$
  Play x_n with ℓ times and observe payoffs $y_{n,1}, y_{n,2}, \ldots, y_{n,\ell}$
  Compute $\mathbf{k}_n(x)^\intercal (K_n + \lambda I_n)^{-1} Y_n^j$ for each $j \in [\ell]$
  Compute $\tilde{\mu}_n(x) = \text{median}\{\mathbf{k}_n(x)(K_n + \lambda I_n) Y_n^j\}_{j=1}^{\ell}$
  Set $\sigma_n(x) = \mathbf{k}_n(x)^\intercal (K_n + \lambda I_n)^{-1} \mathbf{k}_n(x)$
end for

Algorithm 4 Gaussian process bandit algorithm

Input: T, $\mathcal{X}$, $\alpha_t \in \mathbb{R}^+$, $\lambda > 0$, k, $\epsilon \in (0, 1]$, $\delta, \delta' \in (0, 1]$, B
Initialization: Set $\ell = 8 \ln(\frac{2T}{\delta'})$, $N = \lfloor \frac{T}{\ell} \rfloor$, $\tilde{\mu}_0 = 0$
for $n = 1, 2, \ldots, N$ **do**
  Set $\alpha_n = n^{\frac{1-\epsilon}{2(1+\epsilon)}} (4\nu)^{\frac{1}{1+\epsilon}} \left(2B\lambda^{-\frac{1}{2}} \sqrt{\gamma_n + \ln(\frac{1}{\delta})} + \frac{1}{4} \right) + B$
  Select action $x_n = \arg\max_{x \in \mathcal{X}} \tilde{\mu}_{n-1}(x) + \alpha_n \sigma_{n-1}(x)$
  Play x_n with ℓ times and observe payoffs $y_{n,1}, y_{n,2}, \ldots, y_{n,\ell}$
  Compute $\mathbf{k}_n(x)^\intercal (K_n + \lambda I_n)^{-1} Y_n^j$ for each $j \in [\ell]$
  Compute $\tilde{\mu}_n(x) = \text{median}\{\mathbf{k}_n(x)(K_n + \lambda I_n) Y_n^j\}_{j=1}^{\ell}$
  Set $\sigma_n(x) = \mathbf{k}_n(x)^\intercal (K_n + \lambda I_n)^{-1} \mathbf{k}_n(x)$
end for