Recovering Bandits

Ciara Pike-Burke
STOR-i CDT
Lancaster University

c.pike-burke1@lancaster.ac.uk

Steffen Grünewälder
Mathematics and Statistics
Lancaster University

s.grunewalder@lancaster.ac.uk

Abstract

We study a variant of the non-stationary stochastic $K$-armed bandit problem which we call recovering bandits. In this problem, the expected reward of each arm changes depending on the time since the arm was last played according to some unknown recovery function. This problem arises in many settings, for example in product recommendation when after a user makes a purchase, we wish to wait before suggesting the same product again. In the recovering bandits problem, the reward at time $t$ depends on all previous actions, so finding the optimal sequence of $T$ actions would be infeasible. In this paper, we discuss alternative strategies which perform well theoretically and experimentally. Specifically, we assume the recovery functions take the form of a Gaussian process and present UCB and Thompson Sampling algorithms which achieve high instantaneous reward (reward from the played arm) and lookahead reward (total reward from the next $d$ arms).

1. Introduction

The (stochastic) multi-armed bandit problem is a classical sequential decision problem modeling the exploration vs exploitation dilemma (Auer et al., 2002; Thompson, 1933). In its standard form, it consists of sequentially selecting arms $J_t$ and receiving rewards $Y_t$ generated from the underlying reward distribution of the arms. The aim is to maximize the total reward over $T$ rounds. In recent years, bandit algorithms have become ubiquitous in many settings such as web advertising and product recommendation. For example, the problem of suggesting items to a user on an internet shopping platform is typically modeled as a bandit problem where each product (or group of products) is an arm. Over time, the bandit algorithm will learn to suggest only good items to the user and once the algorithm learns that an item (eg. a television) is good, it will continue to suggest it. However, if the user buys the television, the benefit of continuing to show them televisions is immediately diminished (but may increase again as the product reaches the end of its lifetime). To improve customer experience and profit, it would be beneficial for the algorithm to learn not to recommend the same product again immediately, but to wait an appropriate amount of time until the reward from that product has ‘recovered’. The recovering bandits framework presented here provides a natural extension of the stochastic bandit problem to capture this phenomenon.

In the recovering bandits problem, we assume that the expected reward of each arm can be modeled as a function of the rounds since it was last played. Since this changes for the played arm (it resets to 0) and also for the unplayed arms (it increases by 1), the expected reward of every arm changes both when it is and is not played, and it depends on...
the sequence of past actions we have taken. Even when the reward functions are known, selecting the best sequence of $T$ arms for this problem is intractable. Hence, we cannot expect an algorithm for the recovering bandits problem to maximize the reward over the complete horizon $T$ and so we consider other success criterion. One option is to select the action that maximizes the *instantaneous* reward, that is, given the sequence of past plays, the algorithm knows how long it has been since each arm was played, and selects the arm that maximizes the expected reward at that time, without considering future decisions. This still poses quite a challenge compared to the standard $K$-armed bandit problem as instead of just learning the expected reward of each arm, here we must learn entire recovery functions. Sometimes maximizing the instantaneous reward is not appropriate. Recall the earlier internet shopping example. If a user has recently purchased a television, the expected reward of suggesting another one would be low, but it still could be higher than that of another product. In this case, it would be better to show the alternative product and wait until the reward of the television recovers. Thus, although selecting an entire sequence of arms is infeasible, it is natural to consider selecting a sequence of $d \geq 1$ arms in order to maximize the total reward in the next $d$ plays.

In order to develop an algorithm for the recovering bandits problem, we must make some assumptions about the recovery functions. We assume that for each arm $j$, the function $f_j(z)$ specifying the expected reward from playing arm $j$ when it has not been played for $z$ steps is smooth enough to be modeled by a Gaussian process (GP). This captures a wide variety of recovery functions, including those which increase until a certain point then plateau, or that increase at first but then decrease (e.g., in product recommendation, a product will become less attractive if a newer model comes out, or the user has bought it elsewhere).

2. Related Work

In the famous *rested* bandits problem, the reward distribution of an arm only changes when it is played, whereas in the *restless* bandits problem, the reward distributions can change at any time independent of the arm played (Whittle, 1988). In recovering bandits, the reward distributions change in every round and also depend on the past actions.

Slivkins and Upfal (2008); Garivier and Moulines (2008); Besbes et al. (2014) and others studied particular restless bandits problems. Levine et al. (2017); Cortes et al. (2017); Bouneffouf and Feraud (2016) focus on rested bandits with rewards that vary with the number of plays of an arm. In the Rogue Bandits problem of Mintz et al. (2017), the expected reward of arm $j$ at time $t$ depends on an underlying state $x_{j,t}$ via some parametric function and the states evolve according to known dynamics. Their expected frequentest regret is $O(\sum_j \log(T)/\delta_j^2)$ where $\delta_j$ can be arbitrarily small. Converting this into a problem independent regret bound via the standard worst case analysis gives a problem independent regret bound of $O^*(T^{2/3}K^{1/3})$.\(^1\) Our algorithms achieve $O^*(\sqrt{KT})$ Bayesian regret while requiring less knowledge of the recovery functions. In experiments, Mintz et al. (2017) also provide an algorithm based on asymptotics which has no theoretical guarantees. In Section 6, we show that our algorithm outperforms this algorithm experimentally. Note that most of the aforementioned works bound the frequentest regret. In this work, since we consider the recovery functions to be samples from a Gaussian process, we consider Bayesian regret.

\(^1\) We use the notation $O^*$ to suppress log factors.
There has been considerable work on the Gaussian process bandit problem, where the aim is to minimize the regret of the actions taken with respect to the maximum of a single GP \( f \). Of particular relevance is the GP-UCB algorithm of Srinivas et al. (2009). They use the Gaussianity of the posterior to construct upper confidence bounds on each \( f(x) \) at each time step. Bogunovic et al. (2016) and Krause and Ong (2011) extended this to slowly drifting and contextual reward functions. However, these contexts and drifts do not depend on the previous actions. Furthermore, in all these works, only instantaneous regret is considered.

3. Problem Definition

We have \( K \) arms and play the bandit game over \( T \) rounds (note that our algorithms do not need to know \( T \)). For each arm \( j \in \{1, \ldots, K\} \) and round \( t \in \{1, \ldots, T\} \), denote by \( Z_{j,t} \) the number of rounds since arm \( j \) was last played, where \( Z_{j,t} \in \mathbb{Z} \subset \mathbb{N} \) and \(|Z| \leq T/K\). If we play arm \( J_t \) at time \( t \), then, at time \( t + 1 \),

\[
Z_{j,t+1} = \begin{cases} 
0 & \text{if } J_t = j \\
Z_{j,t} + 1 & \text{if } J_t \neq j.
\end{cases}
\]

(1)

Note that \( Z_{j,t} \) are random variables since they depend on our past actions. Since \(|Z|\) is finite, there will be some maximal value, \( z_{\text{max}} \), and if \( Z_{j,t} = z_{\text{max}} \), arm \( j \) must be played. We assume that \( Z_1 \) is initialized (either by the environment or we include an initialization step in our algorithm to ensure this) such that \( Z_{j,1} \neq Z_{j',1} \) for \( j \neq j' \) and that they are all strictly less that \( z_{\text{max}} \). We assume that the expected reward of arm \( j \) at time \( t \) is \( f_j(Z_{j,t}) \) where the recovery functions, \( f_j \), are independent samples from Gaussian processes with mean 0 and known kernel. At round \( t \), we observe \( Z_t = (Z_{1,t}, \ldots, Z_{K,t}) \) and play an arm \( J_t \). We receive a noisy observation \( Y_{j,t} = f_j(Z_{j,t}) + \epsilon_t \) where \( \epsilon_t \sim \mathcal{N}(0, \sigma^2) \) are iid and \( \sigma \) is known.

A comprehensive introduction to Gaussian Processes (GPs) is given in Rasmussen and Williams (2006). A GP represents a distribution over a function, \( f \), when for every finite set \( z_1, \ldots, z_N \) of covariates, the distribution of \( f(z_1), \ldots, f(z_N) \) is multivariate Gaussian. A GP is defined by a mean function \( \mu(z) = \mathbb{E}[f(z)] \) and kernel function \( k(z, z') = \mathbb{E}[(f(z) - \mu(z))(f(z') - \mu(z'))] \). After observing \( Y_N = (Y_1, \ldots, Y_N)^T \) at \( z_N = (z_1, \ldots, z_N)^T \) where \( Y_n = f(z_n) + \epsilon_n \) and \( \epsilon_n \sim \mathcal{N}(0, \sigma^2) \) are iid, the posterior distribution of \( f \) is \( \mathcal{GP}(\mu(z; N), k(z, z'; N)) \) where for \( k_N(z) = (k(z_1, z), \ldots, k(z_N, z))^T \) and \( K_N = [k(z_i, z_j)]_{i,j=1}^N \), \( \mu(z; N) = k_N(z)^T(K_N + \sigma^2 I)^{-1}Y_N \), and \( k(z, z'; N) = k(z, z') - k_N(z)^T(K_N + \sigma^2 I)^{-1}k_N(z') \) so \( \sigma^2(z; N) = k(z, z; N) \). Then, the posterior distribution of \( f(z) \) is \( \mathcal{N}(\mu(z; N), \sigma^2(z; N)) \) for all \( z \). For recovering bandits, we have a GP for each arm, so let \( \mu_j(z; n) \) and \( \sigma^2_j(z; n) \) denote the posterior mean and variance of \( f_j(z) \) after \( n \) plays of arm \( j \). Let \( N_j(t) \) be the (random) number of times arm \( j \) has been played in \( t \) time steps. Then, the posterior kernel at time \( t \) is \( k_t(z, z') = k(z, z'|N_j(t-1)) \) and the posterior mean and variance of arm \( j \) at time \( t \) are,

\[
\mu_t(j) = \mu_j(Z_{j,t}; N_j(t-1)) \quad \text{and} \quad \sigma^2_t(j) = \sigma^2_j(Z_{j,t}; N_j(t-1)).
\]

4. Defining the Regret

We compare the reward of our algorithm, \( \pi \), to that of an oracle, \( \pi^* \), that knows the recovery functions, \( f_j \) and either (i) takes the \( Z_i \)'s constructed by \( \pi \) at time \( t \) and selects the best
single arm or (ii) selects the best sequence of $d$ arms starting from $Z_t$. Since the recovery functions may cross, the ‘gap’ may be 0, so we only consider problem independent regret. We also consider the Bayesian regret since the recovery functions are samples from a GP. Also note that if $Z_{j,t} = z_{\text{max}}$, any algorithm will have to play arm $j$ so will get regret 0.

**Instantaneous Regret** We define the instantaneous regret of an algorithm $\pi$ as the cumulative difference in expected reward between the optimal action for a given $Z_t$ and that chosen by $\pi$. The $Z_i$’s in this definition are generated by the algorithm $\pi$ and given to both $\pi$ and $\pi^*$ at time $t$. Let $J_t$ be the action chosen by $\pi$ at time $t$, then,

$$
\mathbb{E}[R_T(\pi)] = \sum_{t=1}^{T} \mathbb{E}[r_t] = \sum_{t=1}^{T} \mathbb{E}\left[ \max_{1 \leq j \leq K} f_j(Z_{j,t}) - f_{J_t}(Z_{J_t,t}) \right]
$$

The frequentist equivalent of this is the definition of regret in most non-stationary bandit problems and in Mintz et al. (2017).

**$d$-step Lookahead Regret** In recovering bandits, even though we cannot consider the whole sequence of future rewards, it may be worth using knowledge of how the $Z_{j,t}$’s evolve to look a few steps ahead and choose a good sequence of arms. For example, if there are two arms $j_1, j_2$ with similar $f_{j_1}(Z_{j_1,t})$ but if we don’t play $j_1$ then in the next round its reward doubles whereas the reward of $j_2$ stays the same, it is better to play $j_2$ first and wait for the reward of $j_1$ to increase. Building on this, we define $d$-step lookahead regret with respect to a sequence of $d$ arms. This can be modeled as a decision tree with nodes corresponding to the $Z$ values and edges representing transitions due to playing arms. Each possible sequence of $d$ actions is a leaf of this tree. For any leaf node $i$ of tree with root $Z$, let $M_i(Z)$ denote the sum of the $f_j$’s along the edges to $i$ at the corresponding $Z_j$ values (where the $Z_j$’s are updated in the tree according to (1)). We define the regret with respect to an oracle which knows $f_j$ and selects the leaf with highest $M_i(Z_i)$. This corresponds to selecting the best sequence of $d$ arms from $Z_t$. Let $I_t$ be the leaf we select at time $t$. We play the arms to $I_t$ for the next $d$ steps, so we select a sequence of arms every $d$ plays. Then,

$$
\mathbb{E}[R_T^{(d)}(\pi)] = \sum_{h=0}^{\lfloor T/d \rfloor} \mathbb{E}\left[ \max_i M_i(Z_{hd+1}) - M_{I_{hd+1}}(Z_{hd+1}) \right].
$$

Generally speaking, the choice of $d$ represents a trade-off between computationally efficiency and accuracy. If $d$ is too big, computing the $d$-step lookahead policy will be difficult, whereas if $d$ is too small, the reward of the optimal $d$-step lookahead policy may not be as high.

5. Gaussian Process Recovery

Since our recovery functions are from a GP, the performance of our algorithms depends on the kernel. As in Srinivas et al. (2009) this manifested in an information theoretic quantity in our regret. For a set $\mathcal{S}$ of covariates and observations $Y_S = [f(z) + \epsilon_z]_{z \in \mathcal{S}}$, we define the information gain, $\mathcal{I}(Y_S; f) = H(Y_S) - H(Y_S|f)$ where $H(\cdot)$ is the entropy. Intuitively, this is the increase in information about $f$ after observing data $Y_S$. Define the maximal information gain form $N$ observations, $\gamma_N = \max_{\mathcal{S} \subset \mathbb{Z}^N} \mathcal{I}(Y_S; f)$.
Algorithm 1 $d$-step lookahead for Recovering Gaussian Process

**Input:** exploration parameter $\alpha_t$ as in (2) (for UCB); inflation parameter $\rho_t$ (for TS).

**Initialization:** Define $\mathcal{T}_d = \{1, d + 1, 2d + 1, \ldots\}$. For arms $j \in A$, set $Z_{j,1} = Z_{\text{max}} - j$.

for $t \in \mathcal{T}_d$ do

Construct the $d$-step lookahead tree (omitting branches with invalid $Z_{j,t}$'s). Then,

If UCB: $I_t = \arg\max_{1 \leq i \leq K^d} \{ \eta_t(i) + \alpha_t \varsigma_t(i) \}$, 
If TS: (i) $\forall j$ sample $\tilde{\eta}_t(i) \sim \mathcal{N}(\eta_t(i), \rho_t \varsigma_t(i))$
(ii) $I_t = \arg\max_{1 \leq i \leq K^d} \{ \tilde{\eta}_t(i) \}$

for $\ell = 1, \ldots, d$ do

Play $\ell$th arm to $I_t$, $J_\ell$, and receive reward $Y_{J_\ell,t+\ell}$.
Set $Z_{J_\ell,t+\ell+1} = 0$. For all $j \neq J_\ell$, set $Z_{j,t+\ell+1} = Z_{j,t+\ell} + 1$.
end for

Update the posterior distributions of all arms $J_1, \ldots, J_d$ played.
end for

5.1 $d$-step Lookahead Algorithms

The $d$-step lookahead UCB ($d$RGP-UCB) and Thompson Sampling ($d$RGP-TS) algorithms are given in Algorithm 1. To begin with, if no values of $Z_t$ are provided by the environment, the algorithm initializes the $Z_{j,t}$ values so that arm $j$ is played at time $t$. Then, every $d$ steps we construct the $d$-step lookahead tree by considering all possible sequences of $d$ arms from $Z_t$ (eliminating leaves with $Z_{j,t+d} \geq z_{\text{max}}$). At time $t$, we select a sequence of arms by choosing a leaf $I_t$ of the tree with root node $Z_t$. We play these $d$ arms for the next $d$ time steps. For any leaf $i$ of a tree with root $Z_t$, define the reward as $M_i(Z_t)$,

$$M_i(Z_t) = \sum_{\ell=0}^{d-1} f_{J_{t+\ell}}(Z_{J_{t+\ell},t+\ell})$$

where $\{J_{t+\ell}\}_{\ell=0}^{d-1}$ and $\{Z_{J_{t+\ell},t+\ell}\}_{\ell=0}^{d-1}$ are the sequences of arms and $z$'s on the path to leaf $i$. Since each $f_j(z)$ is Gaussian, all $M_i(Z_t)$'s are also Gaussian, and $\{M_i(Z_t)\}_{i=1}^{K^d}$ can be viewed as a sample from a GP. At time $t$, we know the posterior distributions of $f_j$. By normality, the posterior means, $\eta_t(i)$, and variances, $\varsigma_t^2(i)$, of $M_i(Z_t)$ at time $t$ are then,

$$\eta_t(i) = \sum_{\ell=0}^{d-1} \mu_t(J_{t+\ell}) \quad \text{and} \quad \varsigma_t^2(i) = \sum_{\ell,q=0}^{d-1} \text{cov}_t(f_{J_{t+\ell}}(Z_{J_{t+\ell},t+\ell}), f_{J_{t+q}}(Z_{J_{t+q},t+q}))$$

where $\text{cov}_t(f_{J_{t+\ell}}(Z_{J_{t+\ell},t+\ell}), f_{J_{t+q}}(Z_{J_{t+q},t+q})) = \mathbb{I}\{J_{t+\ell} = J_{t+q}\} k_t(Z_{J_{t+\ell},t+\ell}, Z_{J_{t+q},t+q})$.

For the $d$RGP-UCB algorithm, we use Gaussianity to construct upper confidence bounds on each $M_i(Z_t)$. We then select the leaf $I_t$ with largest upper confidence bound at time $t$,

$$I_t = \arg\max_{1 \leq i \leq K^d} \{ \eta_t(i) + \alpha_t \varsigma_t(i) \} \quad \text{where} \quad \alpha_t = \sqrt{2 \log(K^d(t + d - 1)^2 | Z|)}.$$ (2)

In $d$RGP-TS, we select a sequence of $d$ arms by sampling $\tilde{\eta}_t(i) \sim \mathcal{N}(\eta_t(i), \rho_t \varsigma_t(i))$ for each leaf node $i$ and choosing the leaf $I_t$ with highest $\tilde{\eta}_t(i)$. Note for the theoretical analysis,
we inflate the variance by $\rho_t = \sqrt{2 \log(K^d(t + d - 1)^2|Z|)}$ to encourage exploration (much in the same way as for Thompson sampling for linear bandits Abeille and Lazarić (2017); Agrawal and Goyal (2013)) but in practice, we will often set $\rho_t = 1$.

**Theorem 1** The $d$-step lookahead regret up to horizon $T \geq K|Z|$ satisfies,

(i) $\mathbb{E}[R_T^{(d)}] \leq O\left(\sqrt{KT \gamma_T \log(K^d T|Z|)}\right)$ for $d$RGP-UCB.

(ii) $\mathbb{E}[R_T^{(d)}] \leq O\left(\sqrt{dKT \gamma_T \log(T|Z|)} + \sqrt{d \log(K)}\right)$ for $d$RGP-TS with $\rho_t = \sqrt{3 \log(K^d(t + d - 1)^2|Z|)}$.

**Proof** A full proof is given in Appendix B for UCB and Appendix C for Thompson Sampling.

5.2 **Instantaneous Algorithms**

Setting $d = 1$ in the above, we recover algorithms for the instantaneous problem satisfying,

**Corollary 2** The instantaneous regret up to horizon $T \geq K|Z|$ satisfies,

(i) $\mathbb{E}[R_T] \leq O\left(\sqrt{KT \gamma_T \log(T K|Z|)}\right)$ for $1$RGP-UCB.

(ii) $\mathbb{E}[R_T] \leq O\left(\sqrt{KT \gamma_T \log(T K|Z|)}\right)$ for $1$RGP-TS with $\rho_t = \sqrt{3 \log(K^2|Z|)}$.

5.3 **Choice of Kernel**

The regret of our algorithm depends on the GP kernel of the recovery functions through the maximal information gain, $\gamma_T$. Assume the kernel family is the same for each arm. Theorem 5 of Srinivas et al. (2009) gives bounds on $\gamma_T$ for some popular kernels. For linear or squared exponential kernels, $\gamma_T = O(\log(T))$ for any lengthscale, and for Matérn kernels with smoothness parameter $\nu$ and any lengthscale, $\gamma_T = O(T^{2/(2\nu+2)} \log(T))$.

6. **Experimental Results**

To test the experimental performance of our algorithms, we used 10 arms, $z_{\text{max}} = 30$ and noise variance $\sigma^2 = 0.01$. We ran each algorithm to horizon 1000 and averaged the results over 100 replications. We used the GPy package (GPy, 2012–) to fit the Gaussian process. For $d$RGP-TS, we set $\rho_t = 1$.

Firstly, we sampled the recovery functions from a GP with squared exponential kernel with lengthscale $l = 5$. Figure 1, illustrates that in the instantaneous case, RGP-UCB
accurately estimates the recovery functions and learns to play each arm where the reward is high (the same is true for RGP-TS and \( d > 1 \)). This shows that even when trying to minimize the instantaneous regret, our algorithm learns to play in good regions and so obtains good reward. We then compared our algorithms to RogueUCB-Tuned (Mintz et al., 2017) in two settings with parametric recovery functions. As in Mintz et al. (2017), we only considered instantaneous reward and regret. The first recovery function was a 3 parameter logistic function, \( f(z) = \frac{\theta_0}{1 + \exp\{-\theta_2(z - \theta_3)\}} \). The second was a modified gamma, \( f(z) = \theta_0 C \exp\{-\theta_1 z\} z^{\theta_2} \) where \( C \) is a normalizer. We used squared exponential kernels in 1RGP-UCB and 1RGP-TS, with lengthscale \( l = 5 \) (dots), \( l = 7.5 \) (dashes) and \( l = 10 \) (solid line). The cumulative regret (regret here is calculated with respect to the fixed functions) and reward are in Figures 2a–2d. Despite RogueUCB-Tuned requiring much more knowledge of the recovery functions, 1RGP-UCB achieves higher reward and lower regret. Interestingly, 1RGP-TS achieves the highest reward but also high regret. This is most likely due to the fact that the definition of regret depends on the current state \( Z_t \) generated by the algorithm which will be different for the different algorithms.

7. Conclusion

In this paper we have studied the recovering bandits problem with Gaussian process recovery functions and shown that UCB and Thompson Sampling algorithms can achieve regret \( O^*(\sqrt{KT}) \) in the instantaneous case and \( O^*(\sqrt{KdT}) \) in the \( d \)-step lookahead case. We have also shown good experimental performance of our approaches. However, it remains an open problem to provide explicit lower bounds and investigate the choice of \( d \).
References


Appendix

Appendix A. Preliminaries

Define the filtration \( \{ \mathcal{F}_t \}_{t=0}^{\infty} \) as \( \mathcal{F}_0 = \emptyset \) and
\[
\mathcal{F}_t = \sigma(J_1, \ldots, J_t, Y_1, \ldots, Y_t, Z_1, \ldots, Z_t)
\]
where \( Z_t = [Z_{1,t}, \ldots, Z_{K,t}] \). It is important to note that \( \mu_t(j), \sigma_t(j), J_t \) and \( Z_t \) are \( \mathcal{F}_{t-1} \)-measurable.

Recall that in both dRGP-UCB and dRGP-TS, we select a sequence of arms to play at time \( t \) by building a \( d \)-step lookahead tree with root \( Z_t \) and selecting the leaf node \( i \) with highest upper confidence bound on \( M_i \), the cumulative reward from playing all arms in that policy,
\[
M_i(Z_t) = \sum_{\ell=0}^{d-1} f_{J_t+\ell}(Z_{J_t+\ell,t+\ell})
\]
where \( \{J_{t+\ell}\}_{\ell=0}^{d-1} \) are the sequence of arms played on the path to leaf \( i \) and \( \{Z_{J_{t+\ell},t+\ell}\}_{\ell=0}^{d-1} \) the corresponding \( z \) values. Denote the posterior mean and variance of \( M_i(Z_t) \) at time \( t \) as \( \eta_t(i) \) and \( \varsigma_t(i) \), then, conditional on the history \( \mathcal{F}_{t-1} \), \( M_i(Z_t) \sim \mathcal{N}(\eta_t(i), \varsigma_t^2(i)) \). When each arm can be played multiple times, there are interaction terms in the variance of the \( M_i(Z_t) \)'s and thus we suffer some additional cost for not updating after every play. For each leaf node \( i \), we can calculate
\[
\varsigma_t^2(i) = \sum_{\ell=0}^{d-1} \sigma_t^2(J_{t+\ell}) + \sum_{\ell \neq q, \ell, q=0}^{d-1} \text{cov}_t(f_{J_t+\ell}(Z_{J_t+\ell,t+\ell}), f_{J_{t+q}}(Z_{J_{t+q},t+q}))
\]
where \( \text{cov}_t(f_{j_t}(Z_{j_t}), f_{j_q}(Z_{j_q})) = 0 \) if \( j_t \neq j_q \) and \( k_{j_t}(Z_{j_t}, Z_{j_t}'; N_{j_t}(t-1)) \) for \( j_t = j_q \).

Before providing the proofs of the regret bounds, we need the following lemmas,

**Lemma 3** Let \( X_1, \ldots, X_n \) be Gaussian random variables such that \( \max_{1 \leq i \leq n} \mathbb{V}(X_i) \leq \zeta^2 \). Then,
\[
\mathbb{E} \left[ \max_{1 \leq i \leq n} X_i \right] \leq \zeta \sqrt{2 \log(n)}.
\]

**Proof** See for example, Lemma 2.2 in Devroye and Lugosi (2001).

**Lemma 4** (Borel-TIS Inequality) Let \( X_i \) be a separable centered Gaussian process with index set \( i \in I \) and let \( \zeta^2 = \max_i \mathbb{V}(X_i) \). Then, if \( \mathbb{P}(\max_i X_i < \infty) > 0 \),
\[
\mathbb{P}(\max_i X_i - \mathbb{E}[\max_i X_i] \geq u) \leq \exp \left\{ -\frac{u^2}{2\zeta^2} \right\}
\]

**Proof** See for example, Giné and Nickl (2015), Theorem 2.5.8.

Then, note that since \( K \) and \( |Z| \) are both finite, and each \( f_j \sim \mathcal{G}\mathcal{P}(0, k(z, z')) \), the joint Gaussian Process over all arms \( j \) and covariates \( z \) satisfies the conditions of Lemma 4.
Lemma 5 Let \( S_f = \sqrt{2\log(K|Z|)} + \sqrt{2\log(T)} \) and assume that \( T \) is such that \( T \geq K|Z| \).
Then,
\[
\mathbb{E} \left[ \max_{j,z} f_j(z) I\{\max_{j,z} f_j(z) \geq S_f\} \right] \leq \frac{2}{T}.
\]

Proof Let \( h_x(s) \) be the probability density function of the random variable \( \max_{j,z} f_j(z) \) and recall that for each arm \( j \), \( f_j \) is sampled from \( \mathcal{GP}(0, k(z, z')) \) independently of other arms and \( \sigma_j^2(z) \leq 1 \) for all arms \( j \) and covariates \( z \). Hence, the collection \( \{f_j(z)\}_{z,j} \) is a set of \( K|Z| \) zero mean Gaussian random variables, and so we can use the results of Lemmas 3 and 4 for \( \max_{j,z} f_j(z) \). We bound,
\[
\mathbb{E} \left[ \max_{j,z} f_j(z) I\{\max_{j,z} f_j(z) \geq S_f\} \right] = \int_{S_f}^{\infty} x h_x(x) \, dx
\]
\[
\leq \int_{S_f}^{\infty} x \mathbb{P}\left( \max_{j,z} f_j(z) \geq x \right) \, dx
\]
\[
= \int_{S_f}^{\infty} x \mathbb{P}\left( \max_{j,z} f_j(z) - \mathbb{E}[\max_{j,z} f_j(z)] \geq x - \mathbb{E}[\max_{j,z} f_j(z)] \right) \, dx
\]
\[
\leq \int_{S_f}^{\infty} x \mathbb{P}\left( \max_{j,z} f_j(z) - \mathbb{E}[\max_{j,z} f_j(z)] \geq x - \sqrt{2\log(K|Z|)} \right) \, dx
\]
\[
= \int_{\sqrt{2\log(T)}}^{\infty} (u + \sqrt{2\log(K|Z|)}) \mathbb{P}\left( \max_{j,z} f_j(z) - \mathbb{E}[\max_{j,z} f_j(z)] \geq u \right) \, du
\]
\[
\leq \int_{\sqrt{2\log(T)}}^{\infty} (u + \sqrt{2\log(T)}) \exp\left\{ - \frac{u^2}{2 \max_{j,z} \sigma_j^2(z)} \right\} \, du
\]
\[
\leq 2 \int_{\sqrt{2\log(T)}}^{\infty} u \exp\left\{ -u^2/2 \right\} \, du
\]
\[
\leq 2 \left[ -\exp\left\{ -u^2/2 \right\} \right]^{\infty}_{\sqrt{2\log(T)}} \leq \frac{2}{T}
\]
where in (4) we have used Lemma 3, in (5) we have used Lemma 4 and our assumption that \( T \geq K|Z| \), and in (6), we have used \( \sigma_j^2(z) \leq 1 \) for all arms \( j \) and covariates \( z \).

Lemma 6 For \( S_f \) defined as in Lemma 5,
\[
\max_{j,z} f_j(z) - \min_{j,z} f_j(z) \leq 2S_f + L_f + L_g
\]
where \( L_f = \max_{j,z} f_j(z) I\{\max_{j,z} f_j(z) \geq S_f\} \), \( L_g = \max_{j,z} g_j(z) I\{\max_{j,z} g_j(z) \geq S_f\} \) for \( g_j(z) = -f_j(z) \), and by symmetry \( \mathbb{E}[L_f] = \mathbb{E}[L_g] \).

Proof
\[
\max_{j,z} f_j(z) - \min_{j,z} f_j(z) \leq \max_{j,z} f_j(z) + \max_{j,z} g_j(z)
\]
\[
= \max_{j,z} f_j(z) (Z_t) I\{\max_{j,z} f_j(z) < S_f\} + \max_{j,z} f_j(z) (Z_t) I\{\max_{j,z} f_j(z) \geq S_f\}
\]
\[
+ \max_{j,z} g_j(z)(Z_t) \mathbb{I}\{\max_{j,z} g_j(z) < S_f\} + \max_{j,z} g_j(z)(Z_t) \mathbb{I}\{\max_{j,z} g_j(z) \geq S_f\}
\leq 2S_f + \max_{j,z} f_j(z) \mathbb{I}\{\max_{j,z} f_j(z) \geq S_f\} + \max_{j,z} g_j(z) \mathbb{I}\{\max_{j,z} g_j(z) \geq S_f\}
\leq 2S_f + L_f + L g.
\]

We also need the following lemma relating the posterior variances to the maximal information gain.

**Lemma 7**

\[
\sum_{t=1}^{T} \sum_{j=1}^{K} \sigma_j^2(J_t) \mathbb{I}\{J_t = j\} \leq C_1 K \gamma T.
\]

where \(C_1 = 1/\log(1 + \sigma^{-2})\).

**Proof** Using the results of Lemma 5.4 of Srinivas et al. (2009) and the fact that the maximal information gain is increasing in the number of data points, it follows that

\[
\sum_{t=1}^{T} \sum_{j=1}^{K} \sigma_j^2(J_t) \mathbb{I}\{J_t = j\} = \sum_{j=1}^{K} \sum_{n=1}^{N_j(T)} \sigma_j^2(Z_j^{(n)}; n - 1)
\leq T \sum_{j=1}^{K} C_1 I(y_j, N_j(T); f_j, N_j(T)) \leq C_1 \sum_{j=1}^{K} \gamma N_j(T) \leq C_1 K \gamma T.
\]

The following lemmas bound the amount of information we lose by only updating the posterior every \(d\) steps.

**Lemma 8** For any \(z \in Z\) arm \(j\) and \(n \in \mathbb{N}, n \geq 1\), let \(Z^{(n)}\) be the \(z\) value when arm \(j\) is played for the \(n\)th time. Then,

\[
\sigma_j^2(z; n - 1) - \sigma_j^2(z; n) = \frac{k_j^2(Z_j^{(n)}; z; n - 1)}{\sigma_j^2(Z_j^{(n)}; n - 1)} + \sigma^2 \leq \frac{\sigma_j^2(Z_j^{(n)}; n - 1)}{\sigma_j^2}.
\]

**Proof** We include the proof, even though it is probably well known. For convenience, we drop the \(j\) notation and let \(k_n(z) = [k(Z^{(1)}, z), \ldots, k(Z^{(n)}, z)]^T\) and \(K_n = [k(Z^{(i)}, Z^{(j)})]_{i,j=1}^{n}\). Then,

\[
\sigma^2(z; n - 1) - \sigma^2(z; n)
= k(z, z) - k_{n-1}(z)^T(K_{n-1} + \sigma^2 I)^{-1}k_{n-1}(z) - k(z, z) + k_n(z)^T(K_n + \sigma^2 I)^{-1}k_n(z)
= k_n(z)^T(K_n + \sigma^2 I)^{-1}k_n(z) - k_{n-1}(z)^T(K_{n-1} + \sigma^2 I)^{-1}k_{n-1}(z)
\]
We write,
\[ k_n(z) = \begin{bmatrix} k_{n-1}(z) \\ k(Z^{(n)}, z) \end{bmatrix}, \quad K_n + \sigma^2 I = \begin{pmatrix} K_{n-1} + \sigma^2 I & k_{n-1}(z) \\ k(Z^{(n)}, z) & k(Z^{(n)}, Z^{(n)}) + \sigma^2 \end{pmatrix} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}. \]

Then, by the block matrix inversion formula,
\[
(K_n + \sigma^2 I)^{-1} = \begin{pmatrix} A^{-1} + A^{-1} B (C - B^T A^{-1} B)^{-1} B^T A^{-1} & -A^{-1} B (C - B^T A^{-1} B)^{-1} \\ -C - B^T A^{-1} B)^{-1} B^T A^{-1} & (C - B^T A^{-1} B)^{-1} \end{pmatrix}
\]

Hence,
\[
k_n(z)^T (K_n + \sigma^2 I)^{-1} k_n(z) = [k_{n-1}(z)^T, k(Z^{(n)}, z)] (K_n + \sigma^2 I)^{-1} \begin{bmatrix} k_{n-1}(z) \\ k(Z^{(n)}, z) \end{bmatrix}
\]

\[
= k_{n-1}(z)^T (A^{-1} + A^{-1} B (C - B^T A^{-1} B)^{-1} B^T A^{-1}) k_{n-1}(z)
- k(Z^{(n)}, z) (C - B^T A^{-1} B)^{-1} B^T A^{-1} k_{n-1}(z)
- k_{n-1}(z)^T A^{-1} B (C - B^T A^{-1} B)^{-1} k(Z^{(n)}, z)
+ k(Z^{(n)}, z) (C - B^T A^{-1} B)^{-1} k(Z^{(n)}, z)
\]

\[
= k_{n-1}(z)^T A^{-1} k_{n-1}(z)
+ k_{n-1}(z)^T (A^{-1} B (C - B^T A^{-1} B)^{-1} B^T A^{-1} k_{n-1}(z) - k(Z^{(n)}, z))
+ (k(Z^{(n)}, z) - k_{n-1}(z)^T A^{-1} B) (C - B^T A^{-1} B)^{-1} k(Z^{(n)}, z)
= k_{n-1}(z)^T A^{-1} k_{n-1}(z)
+ (k(Z^{(n)}, z) - k_{n-1}(z)^T A^{-1} B) (C - B^T A^{-1} B)^{-1} (k(Z^{(n)}, z) - (k_{n-1}(z)^T A^{-1} B)^T)
\]

Then, substituting back \( A = K_{n-1} + \sigma^2 I, B = k_{n-1}(z), C = k(Z^{(n)}, z(n)) + \sigma^2 \) gives,
\[
k_n(z)^T (K_n + \sigma^2 I)^{-1} k_n(z) = k_{n-1}(z)^T (K_{n-1} + \sigma^2 I)^{-1} k_{n-1}(z)
+ (k(Z^{(n)}, z) - k_{n-1}(z)^T (K_{n-1} + \sigma^2 I)^{-1} k_{n-1}(Z^{(n)}))
+ (k(Z^{(n)}, Z^{(n)}) - k_{n-1}(z(n))^T (K_{n-1} + \sigma^2 I)^{-1} k_{n-1}(Z^{(n)}) + \sigma^2)^{-1}
+ (k(Z^{(n)}, z) - (k_{n-1}(z)^T (K_{n-1} + \sigma^2 I)^{-1} k_{n-1}(Z^{(n)}))^T)
\]

\[
= k_{n-1}(z)^T (K_{n-1} + \sigma^2 I)^{-1} k_{n-1}(z) + \frac{k^2(Z^{(n)}, z; n - 1)}{\sigma^2(Z^{(n)}; n - 1) + \sigma^2}
\]

Hence, substituting into (7) gives,
\[
\sigma^2(z; n - 1) - \sigma^2(z; n) = \frac{k^2(Z^{(n)}, z; n - 1)}{\sigma^2(Z^{(n)}; n - 1) + \sigma^2}.
\]

Then, since the covariance matrix is positive semi-definite, for any \( z, z' \) and \( m \in \mathbb{N} \),
\( k(z, z'; m) \leq \sqrt{\sigma^2(z; m) \sigma^2(z'; m)} \) and so
\[
\sigma^2(z; n - 1) - \sigma^2(z; n) \leq \frac{\sigma^2(Z^{(n)}; n - 1) \sigma^2(z; n - 1)}{\sigma^2(Z^{(n)}; n - 1) + \sigma^2} \leq \frac{\sigma^2(Z^{(n)}; n - 1)}{\sigma^2}.
\]
since for any \( z \in \mathcal{Z} \) and \( m \in \mathbb{N} \), \( 0 \leq \sigma^2(z; m) \leq 1 \). This concludes the proof. 

We then use this result in the following lemma,

**Lemma 9** For any leaf node \( i \) of the \( d \)-step look ahead tree constructed at time \( t \),

\[
\zeta_t^2(i) \leq 3 \sum_{j=1}^{K} \frac{N_j(t+d)}{\sum_{m=N_j(t)+1} \sigma^2(z^{(m)}; m-1)} = \zeta_t^2
\]

and \( \zeta_t \) is \( \mathcal{F}_{t-1} \) measurable.

**Proof** First note that since the posterior covariance matrix of \( f_j \) is positive semi-definite, for any \( z_1, z_2 \) and number of samples, \( n-1, k_j(z_1, z_2; n-1) \leq 1/2(\sigma_j^2(z_1; n-1) + \sigma_j^2(z_2; n-1)). \) Hence,

\[
\zeta_t(i) \leq 3 \sum_{\ell=0}^{d-1} \sigma_j^2(J_{t+\ell}).
\]

Now consider arm \( j \) and assume it appears \( s \leq d \) times in the \( d \)-step look ahead policy selected at time \( t \). Then, the contribution of arm \( j \) (which for ease of notation we assume has been played \( n-1 \) times previously) to \( \zeta_t^2(i) \) is, given below where we use the notation \( \sigma_j^2(z^{(i)}; n-1) \) to denote the posterior variance of the \( i \)th play of arm \( j \) given \( n-1 \) observations of arm \( j \).

\[
\sum_{m=n}^{n+s-1} \sigma_j^2(Z_j^{(m)}; n-1) = \sigma_j^2(z^{(n)}; n-1) + \cdots + \sigma_j^2(z^{(n+s-1)}; n-1)
\]

\[
= \sigma_j^2(z^{(n)}; n-1) + \sigma_j^2(z^{(n+1)}; n) + \left( \sigma_j^2(z^{(n+1)}; n-1) - \sigma_j^2(z^{(n+1)}; n) \right) + \cdots
\]

\[
+ \sigma_j^2(z^{(n+s-1)}; n+s-2) + \left( \sigma_j^2(z^{(n+s-1)}; n+s-3) - \sigma_j^2(z^{(n+s-1)}; n+s-2) \right) + \cdots
\]

\[
\leq \sigma_j^2(z^{(n)}; n-1) + \sigma_j^2(z^{(n+1)}; n) + \frac{\sigma_j^2(z^{(n)}; n-1)}{\sigma^2} + \cdots
\]

\[
+ \sigma_j^2(z^{(n+s-1)}; n+s-2) + \cdots + \frac{\sigma_j^2(z^{(n+1)}; n)}{\sigma^2} + \frac{\sigma_j^2(z^{(n)}; n-1)}{\sigma^2}
\]

\[
= \sum_{q=0}^{s-1} \left( 1 + \frac{s-q-1}{\sigma^2} \right) \sigma_j^2(z^{(n+q)}; n+q-1)
\]

\[
\leq \sum_{q=0}^{s-1} \frac{s-q}{\sigma^2} \sigma_j^2(z^{(n+q)}; n+q-1)
\]

which follows by recursively applying Lemma 8. Then, summing over all arms \( j \) gives,

\[
\zeta_t^2(i) \leq 3 \sum_{j=1}^{K} \left( \sum_{m=N_j(t)+1}^{N_j(t+d)} \sigma_j^2(z^{(m)}; N_j(t)) \right)
\]
Which follows since, for a given root node, we first prove the following lemma,

\[ \text{Lemma 10} \]

Let \( t \) be the per step regret at time \( t \) where we have played arm according to the choice of \( f \) using the unknown \( 1 \)-s, \( M_I^*(Z_t) = \max_{j_1, \ldots, j_d} \sum_{\ell} f_{j_{t+\ell}}(z_{j_{t+\ell}, t+\ell}). \)

Let \( r_t \) be the per step regret at time \( t \). We now bound the expected regret from time steps \( t, t+1, \ldots, t+d-1 \) where we have played arm according to the choice of \( I_t \) by our algorithm,

\[ \text{Lemma 11} \]

Let \( S_f = \sqrt{2 \log(K |Z|)} + \sqrt{2 \log(T)} \), then

\[
\sum_{s=t}^{t+d-1} \mathbb{E}[r_s | F_{t-1}] \leq \frac{4dS_f}{(t+d-1)^2} + 2\alpha_t \sqrt{\frac{\zeta^2(I_t)}{2} + \frac{2d\mathbb{E}}{\max_{j,z} f_j(z)\{\max_{j,z} f_j(z) \geq S_f\} |F_{t-1}|}}
\]
**Proof** Define $I_t$ to be the leaf of the path chosen at time $t$ and let

$$\lambda_t(I_t) = \alpha_t \sqrt{\varsigma^2_t(I_t)}$$

Then,

$$\sum_{s=t}^{t+d-1} \mathbb{E}[r_s | \mathcal{F}_{t-1}] = \mathbb{E}[M_{I_t^*}(Z_t) - M_{I_t}(Z_t) | \mathcal{F}_{t-1}]$$

$$= \mathbb{E}[(M_{I_t^*}(Z_t) - M_{I_t}(Z_t)) \mathbb{I}\{M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1}]$$

$$+ \mathbb{E}[(M_{I_t^*}(Z_t) - M_{I_t}(Z_t)) \mathbb{I}\{M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \leq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1}]$$

$$\leq \mathbb{E}[(M_{I_t^*}(Z_t) - M_{I_t}(Z_t)) \mathbb{I}\{M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1}] + \mathbb{E}[2\lambda_t(I_t) | \mathcal{F}_{t-1}]$$

$$= \mathbb{E}[(M_{I_t^*}(Z_t) - M_{I_t}(Z_t)) \mathbb{I}\{M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1}] + 2\alpha_t \sqrt{\varsigma^2_t(I_t)}$$

(8)

where the last line follows since $I_{t+\ell}$ is $\mathcal{F}_{t-1}$ measurable for all $\ell = 0, \ldots, d-1$ since at time $t$ we select a sequence of arms using only the information up to time $t-1$, and $\sigma_t(j)$ is $\mathcal{F}_{t-1}$-measurable for all arms $j$. Hence, all that remains is to bound the first term of (8). Using Lemma 6,

$$\mathbb{E}[(M_{I_t^*}(Z_t) - M_{I_t}(Z_t)) \mathbb{I}\{M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1}]$$

$$\leq \mathbb{E} \left[ \frac{d(m f_j(z) - \min_{j,z} f_j(z)) \mathbb{I}\{M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1} \right]$$

$$\leq \mathbb{E} \left[ 2d S_f + L_f + L_g \mathbb{I}\{M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1} \right]$$

$$\leq 2d \mathbb{E} \left[ \max_{j,z} f_j(z) \mathbb{I}\{\max_{j,z} f_j(z) \geq S_f\} | \mathcal{F}_{t-1} \right]$$

$$+ 2d S_f \mathbb{E} \left[ \mathbb{I}\{I_t \neq I_t^*\}, M_{I_t^*}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)\} | \mathcal{F}_{t-1} \right].$$

The second term of this can then be bounded by considering the probability that we play to leaf $I_t$ when it is clearly sub-optimal for a given root $Z_t$. Note that if $I_t \neq I_t^*$, this means $\eta_t(I_t) + \lambda_t(I_t) \geq \eta_t(I_t^*) + \lambda_t(I_t^*)$. Then, if $M_{I_t^*}(Z_t) - M_{I_t}(Z_t) > 2\lambda_t(I_t)$, this can only happen if one of the following occur:

(i) $M_{I_t^*}(Z_t) \geq \eta_t(I_t^*) + \lambda_t(I_t^*)$

(ii) $\eta_t(I_t) - \lambda_t(I_t) \geq M_{I_t}(Z_t)$

since if neither occur,

$$\eta_t(I_t) + \lambda_t(I_t) < \eta_t(I_t) + (M_{I_t^*}(Z_t) - M_{I_t}(Z_t))/2 \leq M_{I_t}(Z_t) + \lambda_t(I_t) + (M_{I_t^*}(Z_t) - M_{I_t}(Z_t))/2$$

$$< M_{I_t}(Z_t) + (M_{I_t^*}(Z_t) - M_{I_t}(Z_t)) = M_{I_t^*}(Z_t) \leq \eta_t(I_t^*) + \lambda_t(I_t^*)$$

16
and we would not play arm $J_t$. Hence, using Lemma 10,

$$
\mathbb{P}(I_t \neq I^*_t, M_{I_t}(Z_t) - M_{I_t}(Z_t) \geq 2\lambda_t(I_t)|F_{t-1})
$$

$$
\leq \mathbb{P}(M_{I_t}(Z_t) \geq \eta_t(I_t)^* + \lambda_t(I_t)|F_{t-1}) + \mathbb{P}(\eta_t(I_t) - \lambda_t(I_t) \geq M_{I_t}(Z_t)|F_{t-1})
$$

\[ \leq \frac{4dS_f}{(t + d - 1)^2} + 2\sqrt{2 \log(K^d T|Z|)} \sqrt{\varsigma^2_f(I_t)} + 2d \mathbb{E} \left[ \max_{j,z} f_j(z) \mathbb{I}\{\max_{j,z} f_j(z) \geq S_f\}|F_{t-1}) \right] \]

Hence, combining the above,

\[ \sum_{s=t}^{t+d-1} \mathbb{E}[r_s|F_{t-1}] \leq \frac{4dS_f}{(t + d - 1)^2} + 2\sqrt{2 \log(K^d T|Z|)} \sqrt{\varsigma^2_f(I_t)} + 2d \mathbb{E} \left[ \max_{j,z} f_j(z) \mathbb{I}\{\max_{j,z} f_j(z) \geq S_f\}|F_{t-1}) \right] \]

We now prove the regret bound,

**Theorem 12** The regret of $d$RGP-UCB with respect to the optimal $d$-step look ahead policy satisfies,

$$\mathbb{E}[\mathcal{R}_T] \leq O \left( \sqrt{KT\gamma T \log(K^d T|Z|)} \right).$$

**Proof** First note that

$$\mathbb{E}[\mathcal{R}_T] = \sum_{h=0}^{\lfloor T/d \rfloor} \sum_{s=hd+1}^{(h+1)d} \mathbb{E}[r_s|F_{hd}].$$

Then, let $t$ be a time step when a $d$-step look ahead policy is started and define $I_t$ to be the leaf of the path chosen at time $t$ and let $\lambda_t(I_t) = \sqrt{2 \log(K^d T|Z|)} \sqrt{\varsigma^2_f(I_t)}$ and $S_f = \sqrt{2 \log(K|Z|) + \sqrt{2 \log(T)}}$. So, by Lemma 11,

$$\sum_{s=t}^{t+d-1} \mathbb{E}[r_s|F_{t-1}] \leq \frac{4dS_f}{(t + d - 1)^2} + 2\alpha_t \sqrt{\varsigma^2_f(I_t)} + 2d \mathbb{E} \left[ \max_{j,z} f_j(z) \mathbb{I}\{\max_{j,z} f_j(z) \geq S_f\}|F_{t-1}) \right]$$

where we have used the fact that $\varsigma_t(I_t)$ is $F_{t-1}$ measurable. Then, note that from Lemma 9, it follows that

$$\varsigma_t^2(i) \leq 3 \sum_{j=1}^{K} \sum_{m=N_j(t+d) + 1}^{N_j(t+d) - m + 1} \frac{N_j(t+d) - m + 1}{\sigma^2} \sigma_f^2(z(m); m-1) \leq \frac{3d}{\sigma^2} \sum_{j=1}^{K} \sum_{m=N_j(t+d) + 1}^{N_j(t+d)} \sigma_f^2(z(m); m-1).$$

Hence, taking the expectation and summing over all time points where we start a $d$-step look ahead policy, it follows that,

$$\mathbb{E}[\mathcal{R}_T] = \sum_{h=0}^{\lfloor T/d \rfloor - 1} \mathbb{E} \left[ \sum_{s=hd+1}^{(h+1)d} \mathbb{E}[r_s|F_{hd}] \right]$$
Pike-Burke and Grünewälder

\[ \leq \sum_{h=0}^{\left\lceil \frac{T}{d} \right\rceil} \mathbb{E} \left[ \frac{4dS_f}{(h+1)^2d^2} + 2\alpha_{hd+1}\sqrt{\text{shd+1}(I_{hd+1})} + d\mathbb{E} \left[ \max_{j,z} f_j(z)I_{\max f_j(z) \geq S_f} | \mathcal{F}_{hd} \right] \right] \]

\[ \leq \frac{4S_f}{d} \sum_{h=1}^{\left\lceil \frac{T}{d} \right\rceil} \frac{1}{h^2} + 2 \sum_{h=0}^{\left\lceil \frac{T}{d} \right\rceil} \sqrt{2 \log(K^d|Z|(h+1)^2d^2)} \mathbb{E} \left[ \sqrt{\sum_{j=1}^{K} \sum_{m=N_j(dh)+1}^{N_j(d(h+1))} \sigma_j^2(z^{(m)}; m-1)} \right] \]

\[ \quad + 2d \sum_{h=0}^{\left\lceil \frac{T}{d} \right\rceil} \mathbb{E} \left[ \max_{j,z} f_j(z)I_{\max f_j(z) \geq S_f} \right] \]

\[ \leq \frac{2S_f\pi^2}{3d} + 2\sqrt{\frac{6d}{\sigma^2} \log(K^d|Z|)|T/d| + 1} \mathbb{E} \left[ \sqrt{\sum_{h=0}^{\left\lceil \frac{T}{d} \right\rceil} \sum_{j=1}^{K} \sum_{m=N_j(dh)+1}^{N_j(d(h+1))} \sigma_j^2(z^{(m)}; m-1)} \right] \]

Then, from Lemma 7 and the fact that \( \gamma_n \) is increasing in \( n \),

\[ \sqrt{\sum_{h=0}^{\left\lceil \frac{T}{d} \right\rceil} \sum_{j=1}^{K} \sum_{m=N_j(dh)+1}^{N_j(d(h+1))} \sigma_j^2(z^{(m)}; m-1)} = \sqrt{\sum_{j=1}^{K} \sum_{m=1}^{\gamma_{N_j(T)}} \sigma_j^2(z^{(m)}; m-1)} \]

\[ \leq \sum_{j=1}^{K} C_1 \gamma_{N_j(T)} \leq \sqrt{C_1 K^\gamma_T} \]

for \( C_1 = (1 + \log(\sigma^{-2}))^{-1} \). Hence,

\[ \mathbb{E}[R_T] \leq \frac{2S_f\pi^2}{3d} + 2 \frac{d}{T} + 2\sqrt{\frac{6d}{\sigma^2} \log(K^d|Z|)|T/d| + 1} \sqrt{C_1 K^\gamma_T} \]

and so the result follows. \( \blacksquare \)

Appendix C. Theoretical Results for \( dRGP-TS \)

Before we prove the regret bound for the Thompson sampling algorithm, we need the following result (see for example Abramowitz and Stegun (1964))

Lemma 13 For a Gaussian random variable \( W \) with mean \( \mu \) and variance \( \sigma^2 \), for any \( \lambda \geq 1 \),

\[ \frac{1}{2\sqrt{\pi \lambda}} \exp\{-\lambda^2/2\} \leq \mathbb{P}(|W - \mu| \geq \lambda \sigma) \leq \frac{1}{\sqrt{\pi \lambda}} \exp\{-\lambda^2/2\} \]
Theorem 14 The regret of dRGP-TS with respect to the optimal d-step lookahead policy satisfies,

**Proof** The proof is similar to Agrawal and Goyal (2013). Let,

\[
\alpha_t = \sqrt{2 \log((t + d - 1)^2 K^d |Z|)} \quad \beta_t = \sqrt{2 \log((t + d - 1)^2 K^d |Z|)}
\]

and define the inflation factor by which we inflate the posterior variance when we sample \( \bar{\eta}_t(i) \), \( \rho_t \),

\[
\rho_t = \sqrt{3 \log((t + d - 1)^2 K |Z|)}.
\]

We say a leaf node \( i \) is saturated if \( M_t^i(Z_t) - M_t^i(Z_{t-1}) \geq g_t \varsigma_t(i) \) where \( g_t = \alpha_t + \rho_t \beta_t \) and let \( D(t) \) denote the set of saturated leaf nodes at time \( t \). We now define events that represent the posterior mean of the \( M_t^i(Z_t) \) being close to the true value, and the sampled value \( \tilde{\eta}_t(i) \) being close to expectation of the distribution from which it was sampled,

\[
E^M(t) = \{ \forall i \forall z | \eta_t(i) - M_t(i) | \leq \alpha_t \varsigma_t(i) \} \quad E'(t) = \{ \forall i \forall z | \tilde{\eta}_t(i) - \eta_t(i) | \leq \beta_t \rho_t \varsigma_t(i) \}.
\]

By Lemma 10, \( \mathbb{P}(E^M(t)) \geq 1 - \frac{1}{t^2} \). Then, also,

**Lemma 15** For any time \( t \),

\[
\mathbb{P}(E'(t)|F_{t-1}) \geq 1 - \frac{1}{t^2}
\]

**Proof** We consider the complementary event and note that for leaf nodes \( i \), \( \eta_t(i) \) and \( \varsigma_t(i) \) are \( F_{t-1} \)-measurable and, conditional on \( F_{t-1} \), \( \tilde{\eta}_t(i) \) is sampled from a \( \mathcal{N}(\eta_t(j), \rho_t^2 \varsigma_t^2(j)) \) distribution. Then by Gaussian concentration,

\[
\mathbb{P}(\overline{E'}(t)|F_{t-1}) = \mathbb{P}(\exists i \exists z; | \tilde{\eta}_t(i) - \eta_t(i) | \leq \beta_t \rho_t \varsigma_t(i)|F_{t-1}) \leq 1 \sum_{i=1}^{K^d} \sum_{z \in Z} \mathbb{P}(| \tilde{\eta}_t(i) - \eta_t(i) | \leq \beta_t \rho_t \varsigma_t(i)|F_{t-1}) \leq \sum_{i=1}^{K^d} \sum_{z \in Z} \exp\left\{ -\frac{\beta_t^2}{2} \right\} \leq 1 \frac{1}{t^2}
\]

Then, since \( \tilde{\eta}_t(i) \sim \mathcal{N}(\eta_t(i), \rho_t^2 \varsigma_t^2(i)) \), we can use an anti-concentration argument to get

**Lemma 16** Let \( p = \frac{1}{4\sqrt{\pi e}} \), then,

\[
\mathbb{P}(\tilde{\eta}_t(I^*_t) \geq M_t(I^*_t), E^M(t)|F_{t-1}) \geq p - \frac{1}{t^2}
\]

**Proof**

\[
\mathbb{P}(\tilde{\eta}_t(I^*_t) > M_t(I^*_t), E^M(t)|F_{t-1}) = \mathbb{P}\left( \frac{\tilde{\eta}_t(I^*_t) - \eta_t(I^*_t)}{\nu_t \varsigma_t(I^*_t)} > \frac{M_t(I^*_t) - \eta_t(I^*_t)}{\nu_t \varsigma_t(I^*_t)}, E^M(t)|F_{t-1}\right) \\
\geq \mathbb{P}\left( \frac{\tilde{\eta}_t(I^*_t) - \eta_t(I^*_t)}{\nu_t \varsigma_t(I^*_t)} > \frac{\rho_t \varsigma_t(I^*_t)}{\nu_t \varsigma_t(I^*_t)}, E^M(t)|F_{t-1}\right) \\
\geq \mathbb{P}\left( \frac{\tilde{\eta}_t(I^*_t) - \eta_t(I^*_t)}{\nu_t \varsigma_t(I^*_t)} > \frac{\rho_t}{\nu_t} F_{t-1}\right) - \mathbb{P}(\overline{E^M(t)}|F_{t-1}) \quad (9)
\]
\[
\sum_{j=1}^{K} \mathbb{E} \left[ \left\{ \frac{\hat{\eta}(I_t^*) - \eta(I_t^*)}{\nu_t \sigma_t(I_t^*)} > \frac{\rho_t}{\nu_t}, I_t^* = i \right\} \mid \mathcal{F}_{t-1} \right] - \frac{1}{t^2} = \sum_{j=1}^{K} \mathbb{E} \left[ \left\{ \frac{\hat{\eta}(i) - \eta(i)}{\nu_t \sigma_t(i)} > \frac{\rho_t}{\nu_t}, I_t^* = i \right\} \mid \mathcal{F}_{t-1} \right] - \frac{1}{t^2}
\]

\[
= \sum_{j=1}^{K} \mathbb{P} \left( \frac{\hat{\eta}(i) - \eta(i)}{\nu_t \sigma_t(j)} > \frac{\rho_t}{\nu_t} \mid \mathcal{F}_{t-1} \right) \mathbb{P}(I_t^* = i \mid \mathcal{F}_{t-1}) - \frac{1}{t^2}
\]

\[
\geq \sum_{j=1}^{K} \frac{1}{4\sqrt{\pi} e (\rho_t/\nu_t)} \exp \left\{ -\frac{(\rho_t/\nu_t)^2}{2} \right\} \mathbb{P}(I_t^* = i) - \frac{1}{t^2}
\]

\[
\geq \frac{1}{4\sqrt{\pi} e} - \frac{1}{t^2} = p - \frac{1}{t^2}
\]

where we have used,

\[
\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) \geq \mathbb{P}(A) - \mathbb{P}(B)
\]

in (9), Lemma 10 in (10), the fact that conditional on \( \mathcal{F}_{t-1} \), the sample of \( \hat{\eta}(i) \) being far from its mean is independent of the optimal node in (11), and (12) follows by Lemma 13. The last inequality follows since \( \rho_t/\nu_t \leq 1 \).

It then follows by Lemma 3 in Agrawal and Goyal (2013) that,

\[
\mathbb{P}(I_t \notin D(t), E^M(t) \mid \mathcal{F}_{t-1}) \geq p - \frac{2}{t^2}
\]

We now bound \( \mathbb{E}[(M_t(Z_t) - M_t(Z_t)) \mathbb{I}\{E^M(t) \mid \mathcal{F}_{t-1}\}] \) in terms of \( \zeta_t(i), g_t \) and \( S_f \).

**Lemma 17** Let \( S_f = \sqrt{2 \log(K|Z|)} + \sqrt{2 \log(T)} \) and \( L_f = \max_{j,z} f_j(z) \{ \max_{j,z} f_j(z) \geq S_f \} \). Then,

\[
\mathbb{E}[(M_t(Z_t) - M_t(Z_t)) \mathbb{I}\{E^M(t) \mid \mathcal{F}_{t-1}\}] \leq \frac{3g_t}{p} \mathbb{E}[\zeta_t(I_t) \mathbb{I}\{E^M(t) \mid \mathcal{F}_{t-1}\}] + \frac{8dg_T}{p^2 t^2} + \frac{4dS_f}{t^2} + 2d \mathbb{E}[L_f \mid \mathcal{F}_{t-1}]
\]

**Proof** For ease of notation, let \( M_t(i) = M_t(Z_t) \). Then, note that for \( t \leq 2/p \), the result holds since, \( S_f \leq g_T \), so, by Lemma 6,

\[
\mathbb{E}[(M_t(I_t^*) - M_t(I_t^*)) \mathbb{I}\{E^M(t) \mid \mathcal{F}_{t-1}\}] \leq 2dS_f + 2d \mathbb{E}[L_f \mid \mathcal{F}_{t-1}] \leq 2dS_f + 2d \mathbb{E}[L_f \mid \mathcal{F}_{t-1}]
\]

\[
\leq \frac{8dg_T}{p^2 t^2} + 2d \mathbb{E}[L_f \mid \mathcal{F}_{t-1}]
\]

\[
\leq \frac{3g_t}{p} \mathbb{E}[\zeta_t(I_t) \mathbb{I}\{E^M(t) \mid \mathcal{F}_{t-1}\}] + \frac{8dg_T}{p^2 t^2} + \frac{4dS_f}{t^2} + 2d \mathbb{E}[L_f \mid \mathcal{F}_{t-1}]
\]

20
Hence, for the remainder, we assume \( t > 2/p \). For time \( t > 2/p \), let \( I_t^{\text{min}} \) denote the unsaturated node with smallest posterior standard deviation, \( I_t^{\text{min}} = \arg\min_{i \notin D(t)} \varsigma_t(i) \) and note that \( I_t^{\text{min}} \) is \( \mathcal{F}_{t-1} \) measurable. However, recall that by the sampling procedure in Thompson Sampling, \( I_t \) depends on the sampled values, \( \tilde{\eta}_t(i) \), and so is not \( \mathcal{F}_{t-1} \) measurable. Then, by (13),

\[
\mathbb{E}[\varsigma_t(I_t)\mathbb{I}\{E^t(t)\}|\mathcal{F}_{t-1}] \geq \mathbb{E}[\varsigma_t(I_t)\mathbb{I}\{E^M(t), I_t \notin D(t)\}|\mathcal{F}_{t-1}]
\geq \varsigma_t(I_t^{\text{min}})\mathbb{E}[\mathbb{I}\{E^M(t), I_t \notin D(t)\}|\mathcal{F}_{t-1}] \geq \varsigma_t(I_t^{\text{min}})(p - \frac{2}{t^2})
\]

Then, if both \( E^M(t) \) and \( E^o(t) \) are true,

\[
\tilde{\eta}_t(i) \leq \eta_t(i) + \beta_t \rho_t \varsigma_t(i) \leq M_t(i) + (\alpha_t + \beta_t \rho_t) \varsigma_t(i) = \hat{M}_t(i) + g_t \varsigma_t(i),
\]

and,

\[
M_t(i) \leq \eta_t(i) + \alpha_t \varsigma_t(i) \leq \tilde{\eta}_t(i) + (\alpha_t + \beta_t \rho_t) \varsigma_t(i) = \hat{\eta}_t(i) + g_t \varsigma_t(i).
\]

So, in this case,

\[
M_t(I_t^*) - M_t(I_t) = M_t(I_t^*) - M_t(I_t^{\text{min}}) + M_t(I_t^{\text{min}}) - M_t(I_t)
\leq M_t(I_t^*) - M_t(I_t^{\text{min}}) + \tilde{\eta}_t(I_t) + g_t \varsigma_t(I_t) + g_t \varsigma_t(I_t^{\text{min}})
\leq M_t(I_t^*) - M_t(I_t^{\text{min}}) + g_t \varsigma_t(I_t) + g_t \varsigma_t(I_t^{\text{min}})
\leq 2g_t \varsigma_t(I_t^{\text{min}}) + g_t \varsigma(I_t).
\]

since if we play node \( I_t \) at time \( t \), \( \tilde{\eta}_t(I_t^*) \geq \tilde{\eta}_t(j) \) for all node \( i \) and \( I_t^{\text{min}} \notin D(t) \) so \( M_t(I_t^*) - M_t(I_t^{\text{min}}) \leq g_t \varsigma_t(I_t^{\text{min}}) \). Then,

\[
\mathbb{E}[(M_t(I_t^*) - M_t(I_t))\mathbb{I}\{E^M(t)\}|\mathcal{F}_{t-1})]
\leq \mathbb{E}[(M_t(I_t^*) - M_t(I_t))\mathbb{I}\{E^M(t), E^o(t)\}|\mathcal{F}_{t-1}) + \mathbb{E}[(M_t(I_t^*) - M_t(I_t))\mathbb{I}\{E^M(t), E^o(t)\}|\mathcal{F}_{t-1})]
\leq \mathbb{E}[(2g_t \varsigma_t(I_t^{\text{min}}) + g_t \varsigma_t(I_t))\mathbb{I}\{E^M(t), E^o(t)\}|\mathcal{F}_{t-1})
\quad + \mathbb{E}[d(\max_{j,i} f_z(j) - \min_{j,i} f_z(i))\mathbb{I}\{E^M(t), E^o(t)\}|\mathcal{F}_{t-1})]
\leq \frac{2g_t}{p - 2t^2} \mathbb{E}[\varsigma_t(I_t)\mathbb{I}\{E^M(t)\}|\mathcal{F}_{t-1}) + \frac{g_t}{p} \mathbb{E}[\varsigma_t(I_t)\mathbb{I}\{E^M(t)\}|\mathcal{F}_{t-1}) + 2dS_f \mathbb{E}\{\frac{E^o(t)}{\mathcal{F}_{t-1})}\}
\quad + 2d \mathbb{E}[L_f|\mathcal{F}_{t-1})]
\leq \frac{3g_t}{p} \mathbb{E}[\varsigma_t(I_t)\mathbb{I}\{E^M(t)\}|\mathcal{F}_{t-1}) + \frac{8d g_t}{p^2 t^2} + \frac{4d S_f}{t^2} + 2d \mathbb{E}[L_f|\mathcal{F}_{t-1})]
\leq \frac{3g_t}{p} \mathbb{E}[\varsigma_t(I_t)\mathbb{I}\{E^M(t)\}|\mathcal{F}_{t-1}) + \frac{8d g_t}{p^2 t^2} + \frac{4d S_f}{t^2} + 2d \mathbb{E}[L_f|\mathcal{F}_{t-1})]
\]

where the last inequality has used the fact that \( \varsigma_t(i) \leq d \) and since \( t > 2/p \), \( p - 2/t^2 > p - p^2/2 \geq p/2 \).
Then, using the above, we can bound the expected regret of 1RGP-TS by,

\[
\mathbb{E}[\mathcal{R}_T] = \sum_{h=0}^{[T/d]} \mathbb{E}[(M_{h+1}(\mathbf{Z}_{hd+1}) - M_{h+1}(\mathbf{Z}_{hd+1}))]
\]

\[
= \sum_{h=0}^{[T/d]} \mathbb{E}[(M_{h+1}(\mathbf{Z}_{hd+1}) - M_{h+1}(\mathbf{Z}_{hd+1}))\mathbb{I}\{E^M(hd + 1)\}]
\]

\[
+ \sum_{h=0}^{[T/d]} \mathbb{E}[(M_{h+1}(\mathbf{Z}_{hd+1}) - M_{h+1}(\mathbf{Z}_{hd+1}))\mathbb{I}\{E^M(hd + 1)\}]
\]

\[
= \sum_{h=0}^{[T/d]} \mathbb{E}[\mathbb{E}[\mathbf{E}^M(hd + 1)\mathbb{I}|\mathcal{F}_{hd}]]
\]

\[
+ \sum_{h=0}^{[T/d]} \mathbb{E}[d(\max_{j,z} f_j(z) - \min_{j,z} f_j(z)) | \mathcal{F}_{hd}]
\]

\[
\leq \sum_{h=0}^{[T/d]} \mathbb{E}\left[\frac{3g_{h+1}}{p} \mathbb{E}[\mathbb{I}\{E^M(hd + 1)\} | \mathcal{F}_{hd}] + \frac{8dg_r}{p^2(hd + 1)^2} + \frac{4dS_f}{(hd + 1)^2}\right]
\]

\[
+ 2dS_f \sum_{h=0}^{[T/d]} \mathbb{E}[\mathbb{I}\{E^M(hd + 1)\} | \mathcal{F}_{hd}] + 4d \sum_{h=0}^{[T/d]} \mathbb{E}[\max_{j,z} f_j(z) | \mathbb{I}\{\max_{j,z} f_j(z) \geq S_f\}]
\]

\[
\leq \frac{3g_r}{p} \mathbb{E}\left[\sum_{h=0}^{[T/d]} \mathbb{I}\{E^M(hd + 1)\} + \sum_{h=0}^{[T/d]} \frac{8dg_r}{p^2(h + 1)^2d^2}\right]
\]

\[
+ \sum_{h=0}^{[T/d]} \frac{4dS_f}{(h + 1)^2d^2} + 2dS_f \sum_{h=0}^{[T/d]} \mathbb{E}\left[\frac{1}{(h + 1)^2d^2}\right] + 4d([T/d] + 1)1/T
\]

\[
\leq \frac{3g_r}{p} \mathbb{E}\left[\sum_{h=0}^{[T/d]} \mathbb{I}\{E^M(hd + 1)\} + \frac{64g_r\pi^3e^2}{3d} + \frac{2S_f g_r \pi^2}{3d} + \frac{\pi^2S_f}{3d} + 4 + \frac{4d}{T}\right].
\]

Then, we can use Lemmas 9 and 7 to get,

\[
\frac{[T/d]}{\sum_{h=0}^{[T/d]} \mathbb{I}\{E^M(hd + 1)\}} = \sum_{h=0}^{[T/d]} \sqrt{\frac{3d}{\sigma^2}} \sum_{j=1}^{K} \sum_{m=N_j(hd+1)}^{N_j((h+1)d)} \sigma_j^2(Z_j^{(m)}; m - 1)
\]

\[
\leq \sqrt{\frac{3}{\sigma^2}} \frac{(T + d)}{\sum_{j=1}^{K} \sum_{n=1}^{N_j(T)} \sigma_j^2(Z_j^{(n)}; n - 1)}
\]

\[
\leq \sqrt{\frac{3}{\sigma^2}} C_1 KT \gamma_T
\]
where \( C_1 = (1 + \log(\sigma^{-2}))^{-1} \). Hence,

\[
\mathbb{E}[\mathcal{R}_T] \leq \frac{3g_T}{p} \left[ \sum_{h=0}^{\lfloor T/d \rfloor} \varsigma_{hd+1}(I_{hd+1}) \right] + \frac{64g_T\pi^3 e^2}{3d} + \frac{2S_f g_T \pi^2}{3d} + \frac{\pi^2 S_f}{3d} + 2 + \frac{2d}{T}
\]

\[
\leq \frac{3g_T}{p} \sqrt{\frac{3}{\sigma^2} C_1 KT \gamma_T} + \frac{64d g_T \pi^3 e^2}{3} + \frac{4g_T \pi^2}{6} + \frac{S_f}{d} + 2
\]

and using that \( g_t = O(\log(TK^d|Z|)) \) gives the final result. □