On Thompson Sampling for Smoother-than-Lipschitz Bandits

James A. Grant
STOR-i Centre for Doctoral Training,
Lancaster University.

David S. Leslie
Department of Mathematics and Statistics,
Lancaster University and PROWLER.io.

Abstract

Thompson Sampling is a well established approach to bandit and reinforcement learning problems. However its use in continuum armed bandit problems has received relatively little attention. We provide the first bounds on the regret of Thompson Sampling for continuum armed bandits under weak conditions on the function class containing the true function and sub-exponential observation noise. Our bounds are realised by analysis of the eluder dimension, a recently proposed measure of the complexity of a function class, which has been demonstrated to be useful in bounding the Bayesian regret of Thompson Sampling for simpler bandit problems under sub-Gaussian observation noise. We derive a new bound on the eluder dimension for classes of functions with Lipschitz derivatives, and generalise previous analyses in multiple regards.

1 Introduction

Thompson Sampling (TS) (Thompson, 1933; Russo et al., 2018) is a Bayesian approach to sequential decision making problems that has been widely applied and found to have both strong empirical performance and desirable theoretical properties. A major advantage of TS is it can typically be extended to new problems in a straightforward manner, with empirical success and without a need to tune parameters or rely on detailed theory to design an algorithmic structure. Two of its shortcomings, however, are that it may be more challenging to analyse theoretically than related approaches, and that for complex problems it may often only be implemented approximately, since it relies on draws from the distribution on the reward function. As a result of these challenges, theoretical guarantees on TS are mostly limited to parametric bandit problems.

Russo and Van Roy (2014) introduced a general analytical technique, based on a measure of problem complexity called the eluder dimension, and applied it to analyse the performance of TS on a family of parametric bandit problems. In this paper we show how this eluder-dimension-based analysis can be generalised substantially. We provide new order-optimal performance guarantees for TS on non-parametric continuum-armed bandit problems whose reward functions have a number of Lipschitz derivatives. These guarantees provide insights into the performance of exact TS which significantly advance current understanding, and also serve as empirical benchmarks and analytical tools for future analyses of approximate TS.

1.1 Bandit Problems

Multi-armed bandit (MAB) problems (Lattimore and Szepesvári, 2018) are classic models of exploration-exploitation dilemmas in sequential decision making problems. Among the most general of these is the stochastic Continuum-Armed Bandit (CAB) problem (Agrawal, 1995). The CAB models a scenario in which a decision-maker repeatedly selects *actions*, represented by elements a of an *action set* $\mathcal{A} \subseteq \mathbb{R}^d$. Taking an action grants the decision-maker a *reward* which is a noisy perturbation of some function $f : \mathcal{A} \to \mathbb{R}$, called the *reward function*, at the selected action a. The decision-maker's objective is to maximise the sum of the rewards they receive over some finite number of actions, without knowledge of f.

Effective strategies toward realising this objective will exhibit an appropriate balance between selecting 'exploratory' actions, which aim to learn the function f across $\mathcal{A}$ to gain confidence in the location of its maximum, and 'exploitative' actions, which target regions where f is empirically suggested to take large values in order to maximise the sum of rewards. This need to balance between exploration and exploitation

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

is present in simpler bandit problems (e.g. those where the set $\mathcal{A}$ is finite, or where the function f is known to have a simple parametric form). However in the more general CAB setting, where we have limited assumptions on f, realising this balance has historically been more challenging.

1.2 Thompson Sampling

Thompson Sampling (TS), also referred to as *posterior sampling*, is a Bayesian approach to sequential decision making problems which aims to achieve an appropriate balance between exploration and exploitation through randomisation (Thompson, 1933; Russo et al., 2018).

Over a sequence of rounds $t \in \mathbb{N}$, the decision-maker utilising TS selects actions by sampling a function $\tilde{f}_t$ from their current posterior belief on the form of the true reward function f, and then selecting an action $a_t \in \mathcal{A}$ which maximises $\tilde{f}_t$ - i.e. an action that would be expected to contribute optimally to the cumulative reward if $\tilde{f}_t$ were the true reward function. Figure 1 illustrates a single step of TS on a CAB.

TS therefore encourages exploration since the posterior distribution has more uncertainty on the value of f in regions of $\mathcal{A}$ where few actions have been selected. TS gradually favours exploitation as the posterior distribution naturally contracts around f and sampled functions have maxima in similar locations to f.

TS has been shown empirically to be highly effective in a wide range of bandit problems (Chapelle and Li, 2011; Russo et al., 2018), and theoretical results (May et al., 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2012; Russo and Van Roy, 2014, etc.) have confirmed this in numerous settings where the reward function may be written in terms of a finite set of parameters. The idea of TS extends readily to nonparametric reward functions, but has received little attention in the literature. We believe that this is, in part, due to the challenges of theoretical analysis and precise inference in complex Bayesian models.

Recently, tools have been developed that mean these challenges are not as insurmountable as they once were. Algorithms for approximate Bayesian inference, such as sequential Monte Carlo and variational inference, have become increasingly sophisticated in recent years, to the point that high quality approximations to TS are now feasible (Lu and Van Roy, 2017; Urteaga and Wiggins, 2018a,b).

On the theoretical side, Russo and Van Roy (2014) introduce a general analytical approach for deriving performance guarantees for TS in bandit problems. This method is based on characterising the entropy of the function class in which possible reward functions are contained, via a quantity called the *eluder dimension*. In Russo and Van Roy (2014) this technique was successfully used to analyse the performance of TS on bandit problems with (generalised) linear reward functions.

Russo and Van Roy's technique can be applied much more widely. In this paper, we show that the method for deriving performance guarantees in terms of the eluder dimension can be extended to CABs whose reward functions are members of non-parametric function classes. We show that TS achieves order optimal performance subject to sufficient conditions on the smoothness of these functions (that they have infinitely many Lipschitz derivatives). We further formalise the framework in which this is achievable in the following subsection.

1.3 Model

We specify a general CAB problem as a tuple $(\mathcal{A}, f_0, p_\eta)$, where $\mathcal{A}$ is the set of available actions, $f_0 : \mathcal{A} \to \mathbb{R}$ is the unknown reward function, and p_η is the distribution of the reward noise. We model f_0 as being a sample from p_0, a non-parametric prior on a function class $\mathcal{F}$ whose nature we will specify later.

In a sequence of rounds $t \in [T] \subseteq \mathbb{N}$, the decision-maker selects an action $a_t \in \mathcal{A}$ and receives a reward $R_t = f_0(a_t) + \eta_t$, which is a noisy perturbation of the reward function at a_t with noise terms η_t distributed according to p_η. Let $\mathcal{H}_t = \sigma(a_1, R_1, \ldots, a_t, R_t)$ be the σ-algebra induced by the history of the first t actions and rewards. We assume that for $t \in [T]$, η_t is (σ^2, b)-sub-exponential conditioned on a_t, meaning

$$\mathbb{E}\big(e^{\lambda \eta_t} | \mathcal{H}_{t-1}, a_t\big) \leq e^{\frac{\lambda^2 \sigma^2}{2}}, \quad \forall \ |\lambda| \leq \frac{1}{b}. \tag{1}$$

The noise terms η_t are also assumed to be conditionally independent given the actions a_t, $t \in [T]$.

We are interested in the performance of TS as a policy to select actions a_t for $t \in [T]$. Let p_t denote the posterior distribution on f_0 conditioned on $\mathcal{H}_t$ and let $\tilde{f}_t$ be a sample from p_t. The TS approach is the one which chooses an action $a_t \in \operatorname{argmax}_{a \in \mathcal{A}} \tilde{f}_{t-1}(a)$ in round t, breaking ties arbitrarily if the maximiser is non-unique.

We principally concern ourselves with the Bayesian regret of TS in T rounds, given as

$$BR(T) = \mathbb{E}_{p_0}\bigg(\sum_{t=1}^{T} \max_{a \in \mathcal{A}} f_0(a) - f_0(a_t)\bigg), \tag{2}$$

where $\mathbb{E}_{p_0}$ denotes expectation with respect to the prior p_0. In particular, we are interested in bounding the Bayesian regret as a function of T for particular

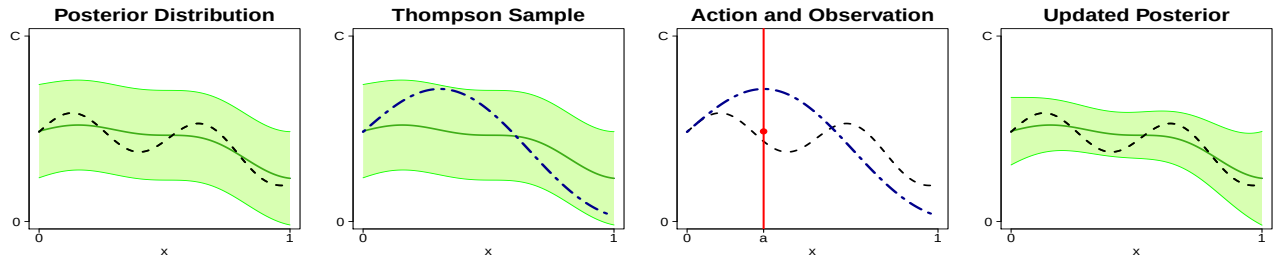

Figure 1: Illustrative example of Thompson Sampling in a round $t \in \mathbb{N}$. The first (leftmost) pane displays a credible interval of a posterior p_{t-1} on $\mathcal{F}$ in green, and a true reward function $f_0 \in \mathcal{F}$ in black. In the second pane the blue curve represents a function $\tilde{f}_t$ sampled from p_{t-1}. In the third pane, the choice of action $a_t \in \text{argmax}_{a\in[0,1]} \tilde{f}_t(a)$ is highlighted in red, along with a reward observation $R(a_t)$ highlighted as a red dot. Finally, the fourth pane displays the posterior π_t updated on the basis of $(a_t, R(a_t))$.

$\mathcal{A}$ and $\mathcal{F}$, and the order with respect to T that such bounds possess. We will also derive lower bounds on the frequentist regret of any algorithm. The frequentist regret,

$$Reg(T) = T \max_{a\in\mathcal{A}} f_0(a) - \sum_{t=1}^{T} \mathbb{E}\big(f(a_t) \,|\, f = f_0\big),$$

is similar to the Bayesian regret (2), the only difference being that the expectation is (as the name suggests) a frequentist expectation conditioned on a fixed reward function f_0, whereas the Bayesian regret additionally takes the expectation of the frequentist regret with respect to this reward function f_0. Frequentist regret bounds which are available for any function f_0 may then be seen as uniform bounds on the Bayesian regret for any prior. We will assess the performance of TS by considering the gap (if any) between the order of the upper and lower bounds. We note that while the analytical tools to upper bound the frequentist regret of non-parametric TS are not currently available, the Bayesian regret is still a useful, and indeed natural, metric to consider in the Bayesian framework.

1.4 Related Work

Numerous authors have studied the frequentist regret of TS in bandit problems, with varying assumptions on the action set, feedback mechanism and reward noise distribution (May et al., 2012; Agrawal and Goyal, 2012; Kaufmann et al., 2012; Korda et al., 2013; Komiyama et al., 2015; Wang and Chen, 2018). None of these works address the fully nonparametric CAB.

Study of the Bayesian regret of TS originated with Russo and Van Roy (2014). Using the eluder dimension measure of the complexity of the reward function class they derived a bound on the Bayesian regret of TS for general action sets and parametric reward function classes. They specialise this to bandit problems with (generalised) linear reward functions. Quadratic functions and applications in model-based reinforcement learning are considered by Osband and Van Roy (2014). Our paper considers a more substantial extension of this technique to reward functions with Lipschitz derivatives.

As already noted, one challenge in deploying Thompson sampling is that sampling from the requisite posterior distributions can only be carried out approximately, thus rendering the theoretical results obsolete. Recently, Phan et al. (2019) have studied the regret of approximations of TS, demonstrating a link between the (assumed to be fixed) error of the approximation and the regret of TS for K-armed bandit problems. We do not address this aspect of the theory in the current article.

The main alternatives to TS in CAB problems, are upper confidence bound (UCB) approaches. These methods, which follow from ideas in Lai and Robbins (1985) and Auer et al. (2002) for simpler K-armed bandits, encourage exploration by making decisions with respect to optimistic estimates of the reward function. Certain UCB methods have been shown to have order-optimal regret bounds in certain CAB problems. These approaches typically employ an *adaptive discretisation* structure, where the action space available at time t is limited to some $\mathcal{A}_t \subset \mathcal{A}$ to force an appropriate level of exploration.

In particular, the 'zooming algorithm' of Kleinberg et al. (2008) maintains a finite set of 'active arms' in $\mathcal{A}$ and only selects actions from within this set. The size of this set is gradually increased by adding arms with high exploitative or exploratory value. The frequentist regret of the zooming algorithm can be shown to be bounded as $O(T^{2/3})$ for the CAB with Lipschtiz reward function and sub-Gaussian noise. Lu et al. (2019) extend these results to heavy-tailed reward noise distributions. This rate is known to be optimal, as Kleinberg (2005) demonstrate that the best achievable re-

gret is $\Omega(T^{2/3})$ across all possible problem instances.

A similar approach is the Hierarchical Online Optimisation (HOO) algorithm of Bubeck et al. (2011a), which discretises the action space according to a tree-based algorithm. In Bubeck et al. (2011a) a yet more general bandit problem is studied where the action set may be any appropriate metric space. HOO is shown to have frequentist regret bounded with order $O(T^{(d'+1)/(d'+2)})$ where $d' > 0$ is a parameter related to the covering number of the metric space, and nature of the possible reward functions. Recent works of Slivkins (2019) and Kleinberg et al. (2019) provide more extensive summaries of bandits on metric spaces.

Apart from this, the special case of a CAB problem with sub-Gaussian noise whose reward function is a sample from a Gaussian process (GP), sometimes referred to as *GP optimisation*, has received particular attention. This setting is more restrictive than ours, but is popular because of its intersection with common modelling assumptions in Bayesian optimisation (Shahriari et al., 2016). The GP-UCB approach of Srinivas et al. (2010, 2012) exploits the closed-form of the GP posterior to calculate an upper confidence function (a combination of the mean and variance of the posterior GP) at each round which is optimised to select actions and enjoys optimal order regret. In this setting both GP-UCB and a GP-based variant of TS can be shown to have $O(\sqrt{T\log(T)})$ Bayesian regret (Srinivas et al., 2012; Russo and Van Roy, 2014), which is optimal for the problem up to a logarithmic factor.

1.5 Key Contributions and Structure

Our main contribution is a bound on the Bayesian regret of Thompson Sampling applied to Continuum-armed Bandits where the reward function is a sample from a prior distribution on the class of bounded functions functions with $M \in \mathbb{N}$ Lipschitz smooth derivatives and the reward noise is sub-exponentially distributed. As far as we are aware this is the first analysis of the performance of TS based on nonparametric inference that considers such a general framework. We derive a $O(T^{(2M^2+11M+10)/(4M^2+14M+12)})$ Bayesian regret bound, which approaches $O(\sqrt{T})$ as $M \to \infty$.

In the process of proving this result we give the first bound on the ϵ-eluder dimension of Lipschitz function classes, and we extend bounds on the Bayesian regret of Thompson Sampling for bandit problems with (generalised) linear reward function to the sub-exponential reward noise setting.

Furthermore we derive an $\Omega(T^{(M+2)/(2M+3)})$ lower bound on regret. There is thus an $O(T^{(3M+2)/(4M^2+14M+12)})$ gap between the lower and upper bounds, which is small for large M. It is an open question as to whether this gap is due to TS being suboptimal, or whether the upper (or lower) bounds we have derived are not tight.

The remainder of the material is organised as follows. In Section 2 we present an extension of Russo and Van Roy (2014)'s general bound on the Bayesian regret. We specialise this to problems where the reward function class has Lipschitz derivatives in Section 3, and conclude with a discussion in Section 4. Proofs are relegated to the Appendices.

2 General Bound on the Bayesian Regret

We first give a bound on the Bayesian regret for general function classes, $\mathcal{F}$, and action sets, $\mathcal{A}$ - including the CAB whose reward function has Lipschitz derivatives. Our result is similar to, but more general than, Proposition 10 of Russo and Van Roy (2014). Their result holds only under sub-Gaussian noise on the reward observations, and has less flexibility in terms of being able to tune the terms based on the properties of $\mathcal{F}$. Our result has such added flexibility and applies to sub-exponential rewards.

Both our bound and that of Russo and Van Roy (2014) are expressed in terms of measures of the complexity of the function class $\mathcal{F}$. This is natural, since in more complex function classes, it will be more challenging to learn the true function. Specifically, two notions of the complexity of $\mathcal{F}$ are of interest, the *ϵ-eluder dimension*, and *ball-width function*, which we introduce below.

Firstly, to define the ϵ-eluder dimension, we introduce the notion of ϵ-dependence. An action $a \in \mathcal{A}$ is called *ϵ-dependent* of actions $a_{1:n} = \{a_1, \ldots, a_n\} \in \mathcal{A}$ with respect to $\mathcal{F}$ if any pair of functions $f, \tilde{f} \in \mathcal{F}$ satisfying $\sqrt{\sum_{i=1}^{n}(f(a_i) - \tilde{f}(a_i))^2} \leq \epsilon$ also satisfies $f(a) - \tilde{f}(a) \leq \epsilon$ for some $\epsilon > 0$. An action a is *ϵ-independent* of $a_{1:n}$ if it is not ϵ-dependent of $a_{1:n}$. The *ϵ-eluder dimension* $dim_E(\mathcal{F}, \epsilon)$ is the length of the longest sequence of elements in $\mathcal{A}$, such that for some $\epsilon' \geq \epsilon$, every element is ϵ'-independent of its predecessors.

Informally, the eluder dimension is a measure of the 'wigglyness' of the functions in $\mathcal{F}$, as it quantifies how long a sequence of actions may be such that at each action, there exist two functions in $\mathcal{F}$ that take well-separated values, but have similar (enough) values for all actions taken previously. We will later show that the more Lipschitz derivatives the functions in a function class have, the smaller its eluder dimension is.

Second, we introduce a *ball-width function* β_n^*. This

ball-width function defines the size of high-probability confidence sets in the function class $\mathcal{F}$, in terms of n, a number of reward observations. Russo and Van Roy (2014) introduce an analogous function, in their equation (8), for the case of sub-Gaussian noise. The properties of sub-exponential distributions mean that our function is necessarily more complex, but its interpretation is the same. In particular β_n^* depends on $N(\alpha, \mathcal{F}, ||\cdot||_\infty)$, the α-covering number of the function class $\mathcal{F}$ with respect to the uniform norm, $||\cdot||_\infty$. Furthermore it depends on σ^2 and b, the sub-exponential parameters of the reward noise distribution, free parameters $\alpha, \delta > 0$ which will be chosen to optimise the regret bound, and λ, which retains its interpretation as the free parameter in Equation (1).

The ball-width function has the following form:

$$\beta_n^*(\mathcal{F}, \delta, \alpha, \lambda) = \frac{2\alpha}{1 - 2\lambda\sigma^2} \times \left[\frac{\log(N(\alpha, \mathcal{F}, ||\cdot||_\infty)/\delta)}{2\lambda\alpha} + n(4C + \alpha)(1 - \lambda\sigma^2) + \sum_{i \leq \lfloor n_0 \rfloor} \sqrt{2\sigma^2 \log(4i^2/\delta)} + \sum_{i \geq \lceil n_0 \rceil}^{n} 2b \log(4i^2/\delta) \right], \tag{3}$$

where $n_0 = \sqrt{\frac{\delta}{4} \exp \frac{\sigma^2}{2b^2}}$.

Together, the eluder dimension and ball-width function characterise a bound on the Bayesian regret of TS applied to the general bandit problem with reward function drawn from $\mathcal{F}$ and actions selected from $\mathcal{A}$. This bound is given in the following theorem.

Theorem 1. *Consider Thompson sampling with prior p_0 on a function class $\mathcal{F}$ applied to the bandit problem $(\mathcal{A}, f_0, p_\eta)$ where the reward function f_0 is drawn from p_0, all functions $f \in \mathcal{F}$ are $f : \mathcal{A} \to [0, C]$ for some $C > 0$, and the reward noise distribution p_η is (σ^2, b)-sub-exponential. For all problem horizons $T \in \mathbb{N}$, nonincreasing functions $\kappa : \mathbb{N} \to \mathbb{R}_+$, and parameters $\alpha > 0, \delta \leq 1/(2T)$, and $|\lambda| \leq (2Cb)^{-1}$, it is the case that*

$$BR(T) \leq T\kappa(T) + (dim_E(\mathcal{F}, \kappa(T)) + 1)C + 4\sqrt{dim_E(\mathcal{F}, \kappa(T))\beta_T^*(\mathcal{F}, \alpha, \delta, \lambda)T}. \tag{4}$$

The bound (4) is useful because it characterises the regret in terms of the eluder dimension and ball-width function of the function class $\mathcal{F}$. Each of these may be bounded in terms of T based on the properties of $\mathcal{F}$. Through judicious choice of κ, α, and δ as functions of T, we can derive regret bound expressions which are sublinear in T. We will do so in Section 3.

As mentioned previously, Proposition 10 of Russo and Van Roy (2014) constructs a similar bound to (4). The material difference between the bounds is that in Russo and Van Roy (2014) $\kappa(T)$ is effectively fixed to T^{-1}, which unnecessarily constrains the results which can be obtained for specific function classes. By allowing for other choices of $\kappa(T)$ we have greater flexibility and can achieve tighter bounds.

In the supplementary material we provide a proof of Theorem 1. Central to the proof is a decomposition of the Bayesian regret of TS in terms of the widths of a sequence of high probability confidence sets for f_0. These sets are centred on a least squares estimator of the reward function. Crucially, their widths can be written in terms of the ball-width function and eluder dimension regardless of whether the estimator itself has a convenient analytical form.

We proceed, in the following section, to specify the bound (4) in the settings where $\mathcal{F}$ is the class of functions with $M \in \mathbb{N}$ Lipschitz derivatives. In Russo and Van Roy (2014), the analogue of (4) is extended only to (generalised) linear function classes. Our results are therefore substantially more general, since we consider non-parametric function classes, which include the (generalised) linear classes as special cases. Nevertheless, in the supplementary material, we demonstrate that our results for sub-exponential noise can explicitly be extended to these (generalised) linear function classes, with no increase in the order of the regret bound.

3 Bounds for Smoother-than-Lipschitz Function Classes

In this section we consider the specification of the general result to classes of functions with Lipschitz derivatives. For any $C, L > 0$ and $M \in \mathbb{N}$, we define $\mathcal{F}_{C,M,L}$ as the class of C-bounded functions, $f : [0, 1] \to [0, C]$, with M L-Lipschitz smooth derivatives. Functions in $\mathcal{F}_{C,M,L}$ satisfy

$$|f^{(m)}(a) - f^{(m)}(a')| \leq L|a - a'|, \ \ \forall a, a' \in [0, 1],$$

for each $m \leq M$. Note that when $M = 0$ this is simply the class of bounded Lipschitz functions.

For larger M, including $M = \infty$, all polynomial functions are trivially included within an $\mathcal{F}_{C,M,L}$, as are appropriately weighted combinations of sufficiently smooth basis functions. Functions sampled from GPs with smooth kernels can also be shown to be members of $\mathcal{F}_{C,M,L}$, since the derivative of a GP is also a GP (Williams and Rasmussen, 2006, Section 9.4). We note also that each $\mathcal{F}_{C,M,L}$ may be represented as a ball within a corresponding Sobolev space, and some readers may find it instructive to think of this interpretation.

3.1 Regret Upper Bound

Our main result, below, is a bound on the Bayesian regret of TS applied where f_0 is drawn from a prior on $\mathcal{F}_{C,M,L}$.

Theorem 2. *Consider Thompson sampling with prior p_0 applied to the bandit problem $([0,1], f_0, p_\eta)$ where f_0 is drawn from a prior p_0 on $\mathcal{F}_{C,M,L}$ and p_η is sub-exponential. For all problem horizons $T \in \mathbb{N}$, we have that the Bayesian regret is bounded as*

$$BR(T) = O(T^{(2M^2+11M+10)/(4M^2+14M+12)}). \quad (5)$$

The consequence of this result is more transparent when we consider particular values of M. We have Bayesian regret of order $O(T^{5/6})$ when the reward function is Lipschitz and of order $O(T^{23/30})$ when it has a Lipschitz first derivative. As the number of Lipschitz derivatives $M \to \infty$ the order of the Bayesian regret approaches $O(\sqrt{T})$. We discuss these results in relation to lower bounds in Section 3.3

Proof of Theorem 2: The proof of Theorem 2 relies on bounding the eluder dimension and ball-width function for the function class $\mathcal{F}_{C,M,L}$. The following theorem provides the necessary bound on the eluder dimension of Lipschitz function classes.

Theorem 3. *For $M \in \mathbb{N}$, and $C, L, \epsilon > 0$ the ϵ-eluder dimension of $\mathcal{F}_{C,M,L}$ is bounded as follows,*

$$\dim_E(\mathcal{F}_{C,M,L}, \epsilon) = o((\epsilon/L)^{-1/(M+1)}). \quad (6)$$

This result is a non-trivial extension of the existing bounds on the eluder dimension of simpler function classes, and is the first bound on the eluder dimension of a non-parametric class of functions. A sketch of the proof of this theorem is given in Section 3.2, and the full proof is given in the supplementary material.

To use Theorem 3 within Theorem 1 we will be considering $\dim_E(\mathcal{F}_{C,M,L}, \kappa(T))$ for a nonincreasing function κ. The effect of M in (6) demonstrates that, for large M, the influence on the regret of κ through the eluder dimension is minimal.

Bounding the ball-width function relies in turn on a bound on the covering number of the Lipschitz function class. The covering numbers of Lipschitz function classes were amongst the first to be discovered (Kolmogorov and Tikhomirov, 1961). Specifically, for $M \in \mathbb{N}$ and $\mathcal{F}_{C,M,L}$ as defined previously, the following is known,

$$\log N(\alpha, \mathcal{F}_{C,M,L}, ||\cdot||_\infty) = \Theta(\alpha^{-\frac{1}{M+1}}).$$

We wish to select α as a function of T to minimise the order of $\beta_T^*(\mathcal{F}_{C,M,L}, \delta, \alpha(T), \lambda)$ with respect to T. Choosing $\alpha(T) = T^{-(M+1)/(M+2)}$ we have,

$$\beta_T^*(\mathcal{F}_{C,M,L}, \delta, T^{-\frac{M+1}{M+2}}, \lambda) = O(T^{1/(M+2)}) \quad (7)$$

as the best available result.

We then complete the proof by using the general bound of (4). We choose $\kappa(T) = T^{-\frac{1}{2}\frac{2M^2+3M+2}{2M^2+7M+6}}$, and bound the eluder dimension as in (6) and ball-width function as in (7) to achieve the stated result. □

3.2 Eluder Dimension Bound

In this section we sketch the proof of the eluder dimension bound given as Theorem 3. To aid in this we first define a related function class:

$$\mathcal{G}_{C,M,L} = \left\{ g = f - f', \forall f, f' \in \mathcal{F}_{C,M,L} \right\},$$

which is the class of absolute difference functions for all pairs of functions in $\mathcal{F}_{C,M,L}$. As the eluder dimension is defined in terms of difference of functions $f, f' \in \mathcal{F}_{C,M,L}$, considering the behaviour of functions in $\mathcal{G}_{C,M,L}$ will allow us to bound the eluder dimension. Functions $g \in \mathcal{G}_{C,M,L}$ also possess M Lipschitz derivatives. Specifically, we have the following result, which has its proof in the supplementary material.

Proposition 1. *All functions $g \in \mathcal{G}_{C,M,L}$ are $[-C, C]$-bounded and possess M $2L$-Lipschitz smooth derivatives.*

We may also define the eluder dimension in terms of $\mathcal{G}_{C,M,L}$, which will be useful for the proof of Theorem 3. Let $a_{1:k} \in [0,1]^k$ denote a sequence of actions $(a_1, \ldots, a_k)$ and define

$$w_k(a_{1:k}, \epsilon') = \sup_{g \in \mathcal{G}_{C,M,L}} \left\{ g(a_k) : \sqrt{\textstyle\sum_{i=1}^{k-1} (g(a_i))^2} \leq \epsilon' \right\}.$$

We then define the ϵ-eluder dimension as follows:

$$dim_E(\mathcal{F}_{C,M,L}, \epsilon) = \max_{\tau \in \mathbb{N}, \epsilon' > \epsilon} \Big\{ \tau : \exists\ a_{1:\tau} \in [0,1]^\tau \text{ with } w_k(a_{1:k}, \epsilon') > \epsilon' \text{ for every } k \leq \tau \Big\}.$$

Based on this definition we will sketch the proof of Theorem 3 in the remainder of this section. The full proof is reserved for the supplementary material.

Sketch of Proof of Theorem 3: The proof relies on the observation that $w_k(a_{1:k}, \epsilon') > \epsilon'$ may only be satisfied if there exists a function $g \in \mathcal{G}_{C,M,L}$ which takes a relatively large value at a_k, i.e. with $g(a_k) > \epsilon'$, but changes rapidly enough to have relatively small absolute value at previous elements of the sequence, i.e. $\sum_{i=1}^{k-1} (g(a_i))^2 \leq (\epsilon')^2$.

Any smooth function g with $g(a) > \epsilon'$ at some $a \in [0,1]$ must have an associated region, of non-zero size, which we call $B(g) \subseteq [0,1]$ where $|g(x)| > \epsilon'/3$. The smoother g is, the larger the region $B(g)$ must be. A necessary condition for satisfying $w_k(a_{1:k}, \epsilon') > \epsilon'$ is that there exists a function $g \in \mathcal{G}_{C,M,L}$ with $g(a_k) > \epsilon'$ such that there are not too many among the points $a_{1:k-1}$ within $B(g)$, specifically fewer than nine (since $\sqrt{9 \times (\epsilon'/3)^2} = \epsilon'$).

It follows that a necessary condition for the ϵ-eluder dimension of $\mathcal{F}_{C,M,L}$ to take value at least τ is that there exists a sequence $a_{1:\tau} \in [0,1]$ and a sequence of functions $g_1, \ldots, g_\tau \in \mathcal{G}_{C,M,L}$ with $g_i(a_i) > \epsilon'$, $i \leq \tau$, such that $\sum_{i=1}^{k-1} \mathbb{I}\{a_i \in B(g_k)\} < 9$ for all $k \leq \tau$. We derive upper bounds on the eluder dimension by bounding the value of τ for which this necessary condition may be satisfied. This is feasible, as the size of the region $B(g)$ for any $g \in \mathcal{G}_{C,M,L}$ and $a \in [0,1]$ such that $g(a) > \epsilon'$ may be related to the smoothness of the class $\mathcal{G}_{C,M,L}$ and the largest value of τ such that the necessary condition can be satisfied may be related to the size of the B regions.

For each choice of M and an $a \in [0,1]$ we can identify a function $h_{M,a} \in \mathcal{G}_{C,M,L}$ which satisfies $h_{M,a}(a) > \epsilon'$ but minimises the size of B_a, i.e.

$$h_{M,a} \in \operatorname*{argmin}_{h \in \mathcal{G}_{C,M,L} : h(a) > \epsilon'} \int \mathbb{I}\{|h(x)| \geq \epsilon'/3\} dx,$$

and the minimising values

$$B^*_{M,a} = \min_{h \in \mathcal{G}_{C,M,L} : h(a) > \epsilon'} \int \mathbb{I}\{|h(x)| \geq \epsilon'/3\} dx.$$

The functions $h_{M,a}$ can be shown to be characterised by having zeros of their derivatives at specific locations. In particular, odd ordered derivatives should have zeros at a and the points where $h_{M,a}(x) = -\epsilon/3$ and even ordered derivatives should have zeros at points where $h_{M,a}(x) = \epsilon/3$. Allowing the highest order derivative to be linear subject to these conditions ensures the region $B_a(h_{M,a})$ is as small as possible. Figure 2 illustrates functions $h_{a,0}, h_{a,1}, h_{a,2}$ and their first derivatives. We can see the increasing width of $B^*_{a,M}$ as M increases.

The minimising values $B^*_{M,a}$ are shown to be $o((\epsilon/L)^{1/(M+1)})$. In turn, this means that if there is a sequence of τ points $a_{1:\tau}$ with $\tau = o((\epsilon/L)^{-1/(M+1)})$ placed in $[0,1]$, it is impossible to satisfy $w_k(a_{1:k}, \epsilon')$ for every $k \leq \tau$. By definition the eluder-dimension may then be bounded as $o((\epsilon/L)^{1/(M+1)})$.

3.3 Regret Lower Bounds

The following theorem, a restatement of Theorem 1 of Bubeck et al. (2011b), gives a lower bound on the regret of any algorithm for the CAB with a Lipschitz reward function. It is an adaptation of the stronger results in Kleinberg (2005); Kleinberg et al. (2008); Bubeck et al. (2011a) which apply to bandits on metric spaces. For ease of exposition, and following convention, we will assume in the remainder, without loss of generality, that the bounding constant is $C = 1$.

Theorem 4. *Let `ALG` be any algorithm for Lipschitz continuum armed bandits with time horizon T, and Lipschitz constant L. Let $M = 0$, i.e. the Lipschitz condition apply only to the reward function, not its derivatives. There exists a problem instance $\mathcal{I} = \mathcal{I}(x^*, \epsilon)$ for some $x^* \in [0,1]$ and $\epsilon > 0$ such that*

$$\mathbb{E}(R(T)|\mathcal{I}) \geq \Omega(L^{1/3} T^{2/3}).$$

The proof of the Theorem relies on the construction of a particularly challenging CAB instance $\mathcal{I}(x^*, \delta)$ with reward function μ where

$$\mu(x) = \begin{cases} 0.5, & \text{for } x : |x - x^*| > \delta/L, \\ 0.5 + \delta - L|x - x^*|, & \text{otherwise.} \end{cases} \tag{8}$$

Theorem 4 does not apply for $M > 0$. This is because the reward function μ defined as in (8) used to define the worst-case problem instance, does not have a Lipschitz first derivative and thus is not a valid reward function for the problem class being considered.

In the theorem below, we give an M-dependent lower bound on regret, for CABs whose reward functions have $M \geq 0$ Lipschitz derivatives.

Theorem 5. *Let `ALG` be any algorithm for the CAB problem with reward function in $\mathcal{F}_{C,M,L}$. There exists a problem instance $\mathcal{I} = \mathcal{I}(x^*, \delta)$ for some $x^* \in [0,1]$ and $\delta > 0$ such that*

$$\mathbb{E}(R(T)|\mathcal{I}) \geq \Omega(T^{(M+2)/(2M+3)}).$$

The proof of this theorem is provided in the supplementary material.

3.4 Comparing Upper and Lower Bounds

Firstly, we notice that for $M = \infty$, the upper and lower bounds match up to a constant, in that they are both order $\sqrt{T}$. This implies that exact TS is an order-optimal algorithm for CAB problems with reward function drawn from a prior on (any subset of) $\mathcal{F}_{C,\infty,L}$. This is a more general result than those presented in Russo and Van Roy (2014), as they had similar results only for special cases within $\mathcal{F}_{C,\infty,L}$ - namely (generalised) linear reward functions and reward functions modelled as samples from Gaussian processes. Further, we even present a marginal improvement in those cases, as we remove a mutliplicative $\log(T)$ factor from the upper bounds.

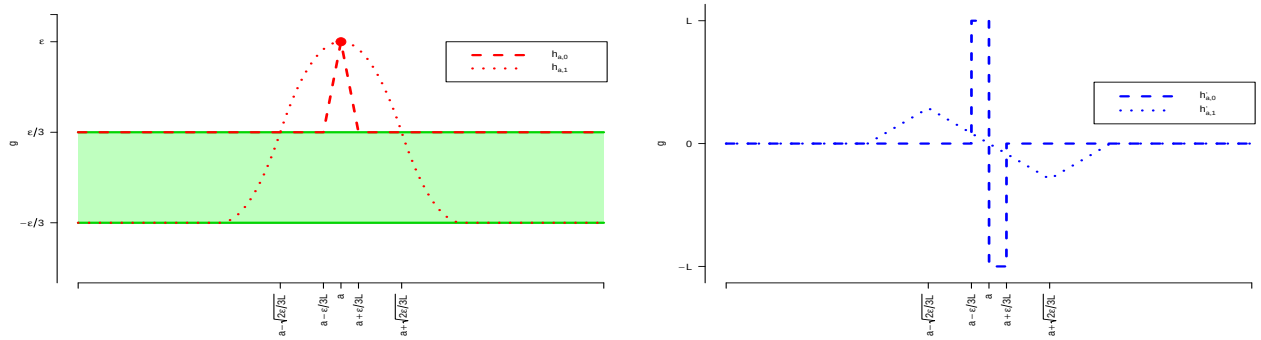

Figure 2: This figure displays functions $g \in \mathcal{G}^*_{C,M,L}(a)$ for $M = 0$ and $M = 1$. These functions take value greater than ϵ at a, which is well separated from 0 and 1. The functions then decrease on the left and right in to the interval $[-\epsilon/3, \epsilon/3]$ at the quickest rate possible for functions in $\mathcal{G}_{C,M,L}$.

Interestingly, for finite M, the bounds do not match. For instance, with $M = 0$ the upper bound has order $O(T^{5/6})$ and the lower bound has order $\Omega(T^{2/3})$. Generally speaking there is a gap of order $T^{(3M+2)/(4M^2+14M+12)}$ between the bounds for finite M. This raises an interesting open question: are the eluder-dimension based bounds simply not tight for finite M, or is TS inherently suboptimal?

There would seem to be some credence to both arguments. If we consider the nature of algorithms which do achieve order optimal bounds for the Lipschitz bandit problem, such as the Zooming algorithm of Kleinberg (2005), we notice that they generally employ an adaptive discretisation component. That is to say, they limit the actions available to the algorithm to some set $\mathcal{A}_t \subset \mathcal{A}$ in each round $t \in \{1, \ldots, T\}$, and in doing so force a certain level of exploration. It could be that the TS algorithm analysed here which has access to the entire action set $\mathcal{A}$ somehow carries a greater risk of conducting insufficient exploration.

On the other hand it is possible that the true performance of the TS approach analysed here does in fact match the lower bound, and analysis of Russo and Van Roy (2014) which we have adapted to this setting is too loose in this framework. The contribution of the covering number term to the overall order for instance in the $M = 0$ setting is $T^{1/4}$ and the $\sqrt{T}$ factor from the least squares analysis is also unavoidable. Thus, even with a $\kappa(T)$-eluder dimension of $O(1)$ the resulting bound would be suboptimal compared with the $\Omega(T^{2/3})$ lower bound. Inspection of the proof suggests that while this technique is highly versatile, it would not be possible to adapt it to achieve an optimal order bound in CAB problems whose reward function is drawn from $\mathcal{F}_{C,M,L}$, with finite M.

4 Conclusion

This work extends the understanding of Thompson Sampling for stochastic bandit problems. The results are bounds on the Bayesian regret of Thompson Sampling for continuum-armed bandits where the reward function possesses M Lipschitz derivatives and where the reward noise is sub-exponential. We achieved these results by extending the application of the eluder dimension technique of Russo and Van Roy (2014) which allows the Bayesian regret of TS to be bounded in terms of the complexity of the reward function class.

Our results represent a substantial advance on the generality of existing performance guarantees available for TS. While previous results have focussed on d-dimensionally parametrised functions or Gaussian process priors only, our framework captures TS based on non-parametric priors over the reward function class. As such our results are applicable in much broader settings where only limited assumptions about the reward function are possible.

While exact sampling from the posterior distributions on which our analysis is based may be challenging, these fundamental results are useful in two regards. They provide a useful benchmarking tool for subsequent analyses, and generally inform us as to how the smoothness properties of the reward function class are likely to impact the performance of TS.

Finally, our work raises interesting open questions around the analysis of non-parametric TS. Firstly, whether the gap between the upper and lower regret bounds for finite M is a feature of the eluder-dimension based analysis (i.e. it can be improved) or of TS itself (i.e. it is inherent and unavoidable). Secondly, to what extent this performance may be recovered by approximate TS algorithms, which are popular and often necessary for complex problems.

References

Agrawal, R. (1995). The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33:1926–1951.

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011a). $\mathcal{X}$-armed bandits. *J. Mach. Learn. Res.*, 12:1655–1695.

Bubeck, S., Stoltz, G., and Yu, J. Y. (2011b). Lipschitz bandits without the lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer.

Kleinberg, R., Slivkins, A., and Upfal, E. (2019). Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):30.

Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *NeurIPS*, pages 697–704.

Kleinberg, R. D., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proc. 40th Annu. ACM Symp. on Theory of Computing*, pages 681–690.

Kolmogorov, A. N. and Tikhomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364.

Komiyama, J., Honda, J., and Nakagawa, H. (2015). Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays.

Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. In *NeurIPS*, pages 1448–1456.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. *preprint*.

Lu, S., Wang, G., Hu, Y., and Zhang, L. (2019). Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163.

Lu, X. and Van Roy, B. (2017). Ensemble sampling. In *Advances in neural information processing systems*, pages 3258–3266.

May, B. C., Korda, N., Lee, A., and Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. *J. Mach. Learn. Res.*, 13:2069–2106.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474.

Phan, M., Abbasi-Yadkori, Y., and Domke, J. (2019). Thompson sampling and approximate inference. *arXiv preprint arXiv:1908.04970*.

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Math. Oper. Res.*, 39:1221–1243.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on Thompson sampling. *Found. Trends Mach. Learn.*, 11:1–96.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

Slivkins, A. (2019). Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Urteaga, I. and Wiggins, C. (2018a). Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 698–706.

Urteaga, I. and Wiggins, C. H. (2018b). (sequential) importance sampling bandits. *arXiv preprint arXiv:1808.02933*.

Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5101–5109.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA.