Adaptive Sensor Placement for Continuous Spaces

James A. Grant [1 2]   **Alexis Boukouvalas** [1]   **Ryan-Rhys Griffiths** [3]   **David S. Leslie** [1 4]   **Sattar Vakili** [1]
Enrique Munoz de Cote [1]

Abstract

We consider the problem of adaptively placing sensors along an interval to detect stochastically-generated events. We present a new formulation of the problem as a continuum-armed bandit problem with feedback in the form of partial observations of realisations of an inhomogeneous Poisson process. We design a solution method by combining Thompson sampling with nonparametric inference via increasingly granular Bayesian histograms and derive an $\tilde{O}(T^{2/3})$ bound on the Bayesian regret in T rounds. This is coupled with the design of an efficent optimisation approach to select actions in polynomial time. In simulations we demonstrate our approach to have substantially lower and less variable regret than competitor algorithms.

1. Introduction

In this paper we consider the problem of adaptively placing sensors to detect events occurring stochastically according to a inhomogeneous Poisson process. This is a problem arising in numerous applications including ecology (Heikkinen & Arjas, 1999), and astronomy (Gregory & Loredo, 1992). Adaptive sequential decision-making that learns an optimal placement of sensors in response to observations can lead to detecting many more events than fixed policies based on an assumed Poisson process rate function. We study the problem under a simple abstract framework which encompasses many possible practical scenarios, including choosing which hours to operate to maximise customer engagement, or choosing placement of mobile base stations to service as many requests as possible, as well as the classical

sensing applications.

Suppose that a decision-maker is tasked with placing a finite number of sensors along an interval. The decision-maker's objective is to maximise, through time, a reward function which trades off the number of events detected with the cost of sensing. At each step, each sensor is tasked with sensing a subinterval, with the cost of sensing depending on the length of the subinterval. Only the events that occur in a sensed subinterval are detected. The decision-maker may update the placement of sensors at regular intervals creating a sequential problem where the decision-maker iteratively places sensors and receives feedback on where events occurred.

The decision-maker therefore faces a classic exploration-exploitation dilemma. In each round they will gather information on what was detected in the sensed regions, and will receive a reward. The most informative action is to sense the entire interval, but this may not be the reward-maximising action due to the cost of sensing. Hence the decision-maker must choose sensor placements to trade off learning about regions where information is insufficient, while also capitalising on information they already have to generate large rewards. This paper develops an algorithm to tackle this problem with the aim of minimising Bayesian regret, the difference between the expected reward achieved by constantly selecting an optimal action and the expected reward of actions actually taken, where the expectation is taken with respect to the prior over the reward-generating parameters.

Multi-armed bandits provide models for sequential decision problems, and our problem most closely resembles the continuum-armed or $\mathcal{X}$-armed bandit problem (Agrawal, 1995). In a continuum-armed bandit (CAB) problem a decision-maker sequentially selects points in some d-dimensional continuous space and receives reward in the form of a noisy realisation of some unknown (usually Lipschitz smooth) function on the space. Our sensor placement problem can map to this framework by considering that the placement of sensors can be represented by the set of end-points of the sensors' subintervals. Note, however, that the noise and feedback models in the sensor placement problem are more complex than in previous treatments of CAB

---

models, which have focused on simple numerical reward observations with bounded or sub-Gaussian noise (e.g. Bubeck et al., 2011). In this paper, we handle the added complexities of observing event locations and the heavier-tailed noise of the Poisson distribution.

Our proposed method performs fast Bayesian inference on the rate function, by means of a Bayesian histogram approach (Gugushvili et al., 2018), and makes decisions to trade off exploration and exploitation using Thompson sampling (TS) (e.g. Russo et al., 2018). Gugushvili et al.'s approach to nonparametric inference on the continuous action space imposes a mesh structure over the interval, splitting it into a finite number of bins, with the mesh becoming finer as time increases. Inference is then performed over the rate of event occurrence in each bin. TS methods select an action in a given round according to the posterior probability that it is optimal. In our approach, this is implemented by sampling bin rates from the simple posterior distributions of Gugushvili et al.'s model and selecting an optimal action for these sampled rates via an efficient optimisation algorithm described in Section 3.4.

We analyse the Bayesian regret of the TS algorithm in this setting using similar techniques to those of Russo & Van Roy (2014). This allows us to derive an $\tilde{O}(T^{2/3})$ upper bound on the Bayesian regret that holds across all possible rate functions with a bounded maximum, and has minimal dependency on the prior used by the TS algorithm. The CAB problem with Poisson noise and event data as feedback is to the best of our knowledge unstudied, however our regret upper bound is encouragingly close to the $\Omega(T^{2/3})$ lower bound on simpler CAB models of Kleinberg (2005).

The remainder of the paper is structured as follows. We review related work in Section 2, formalise our model and algorithm in Sections 3, present the regret analysis in Section 4, and conclude with simulation experiments in Section 5.

1.1. Principal Contributions

The principal contributions of this work are: (i) formulation of a new widely applicable model of sequential sensor placement as a CAB; (ii) the first study of CABs with Poisson process feedback, and use of a new progressive discretisation technique as an approximation to the continuous action space; (iii) an efficient optimisation routine for sensor placement given known event rate; (iv) analysis of the Bayesian regret of a TS approach, resulting in a $\tilde{O}(T^{2/3})$ upper bound; (v) numerical validation of the efficacy of the TS method, and its favourable performance relative to upper confidence bound and ϵ-greedy approaches.

2. Related Work

The problem of allocating searchers in a continuous space has been studied by Carlsson et al. (2016) under the assumption that the rate of arrivals is known. A first attempt to solve a version of the problem in which the rate must be learned is presented in Grant et al. (2018), in which the space is discretised to a fixed grid for all time. The objective of our paper is to present the first learning version of the problem for the fully continuous space.

The fixed discretisation version of the problem maps directly to Combinatorial Multi-Armed Bandits (CMAB) (Cesa-Bianchi & Lugosi, 2012; Chen et al., 2016). This is a class of problems wherein the decision-maker may pull multiple arms among a discrete set and receives a reward which is a function of observations from individual arms. In the discretised sensor-placement problem, the individual arms correspond to cells of the grid. The model remains relevant for the continuous version of the problem, as by using an increasingly fine mesh, we approximate the problem with a series of increasingly many armed CMABs.

The continuum-armed bandit (CAB) model (Agrawal, 1995) is an infinitely-many armed extension of the classic multi-armed bandit (MAB) problem. There are two main classes of algorithm for CAB problems: discretisation-based approaches which select from a discrete subset of the continuous action space at each iteration, and approaches which make decisions directly on the whole action space. Our proposed method belongs to the former class. Early discretisation-based approaches focused on fixed discretisation (Kleinberg, 2005; Auer et al., 2007), with more recent approaches typically using adaptive discretisations such as a "zooming" approach (Kleinberg et al., 2008) or a tree-based structure (Bubeck et al., 2011; Bull, 2015; Grill et al., 2015) to manage the exploration. Authors who handle the full continuous action space typically use Gaussian process models to capture uncertainty in the unknown continuous function and balance exploration-exploitation in light of this (Srinivas et al., 2009; Chowdhury & Gopalan, 2017; Basu & Ghosh, 2017). As mentioned in Section 1, our problem can map into a CAB, but since our information structure is more complex, our action space has dimension greater than 1, and the stochastic components have heavier tails than usual, standard algorithms and results do not apply.

Thompson sampling (TS) is a particularly convenient, and generally effective, method for trading off exploration and exploitation. The critical ideas can be traced as far back as Thompson (1933), although the first proofs of its asymptotic optimality came much later (May et al., 2012; Agrawal & Goyal, 2012; Kaufmann et al., 2012). Later, similar results were derived for MABs with rewards from univariate exponential families (Korda et al., 2013) and in multiple play bandits (Komiyama et al., 2015; Luedtke et al., 2016). More

recently, TS has been studied in the CMAB framework by Wang & Chen (2018) and Huyuk & Tekin (2018) under slightly differing models, but both with bounded reward noise. Both papers demonstrate the asymptotic optimality of TS with respect to the frequentist regret, and we anticipate that these results could be extended to univariate exponential families. However, in both of these works, the leading order coefficients can be highly suboptimal. Therefore, rather than attempt to extend these ideas to CABs, we favour an alternative analysis of the Bayesian regret to get bounds that are of slightly suboptimal order but are more meaningful because of their (relatively) small coefficients. The Bayesian regret is less extensively studied than the frequentist regret. However the bounds that have been derived for the Bayesian regret of TS (Russo & Van Roy, 2014; Bubeck & Liu, 2013) are powerful as they do not depend on a specific parameterisation of the reward functions.

3. Model and solution

We now formally present our model and solution method.

3.1. Reward and regret

In each of a series of rounds $t \in \mathbb{N}$, $m_t \geq 0$ events of interest arise at locations $X_{t,1}, ..., X_{t,m_t} \in [0, 1]$ according to a non-homogeneous Poisson process with rate $\lambda : [0, 1] \to \mathbb{R}_+$. U sensors are deployed in each round with each sensor observing a distinct subinterval of $[0, 1]$; the action space $\mathcal{A}$ consists of the sets of at most U disjoint intervals of $[0, 1]$. Let $A_t \subseteq [0, 1]$ be the union of the subintervals covered by the sensors in round t. An event $X_{t,i}$ is detected if it lies in A_t. The system objective is to maximise the number of detected events while penalised by a cost of operating the sensors. The expected reward for playing action A is therefore

$$r(A) = \int_A (\lambda(x) - C) \, \mathrm{d}x,$$

where C is the cost per unit length of sensing. We define the Bayesian regret of an algorithm to be the expected difference (with respect to the prior on λ) between the reward achieved when playing the optimal action in each of T rounds and the actions taken by the algorithm:

$$BReg(T) = \sum_{t=1}^{T} \mathbb{E}\left(r(A^*) - r(A_t)\right)$$

where $A^* = \arg\max_{A \subseteq \mathcal{A}} r(A)$ is the optimal action on the continuous interval.

3.2. Inference

With the Poisson process rate being defined on the continuum $[0, 1]$, nonparametric estimation is preferable to a parametric form. We use the increasingly granular histogram

approach of Gugushvili et al. (2018), since it provides us with fast inference and a concentration rate. At the beginning of each round t a piecewise-constant estimation of λ is considered by counting the number of events to have been observed in each of K_t bins. The number of bins will be gradually increased as rounds proceed. To maintain simplicity in the inference and analysis we choose all bins to be of a constant width $\Delta_t = K_t^{-1}$.

We introduce the notation

$$B_{k,t} \equiv \left[\frac{k-1}{K_t}, \frac{k}{K_t}\right) \quad \forall \ k \in \{1, \ldots, K_t\}, \ \forall \ t \in \mathbb{N},$$

to refer to the kth histogram bin at iteration t (the index t is needed to uniquely index a bin since the number of bins changes as t increases). The number of events in bin $B_{k,t}$ in a single observation of the Poisson process is a Poisson random variable with parameter $\int_{B_{k,t}} \lambda(x) \, \mathrm{d}x$. Since this depends on the width of the bin, we instead estimate the average rate function in a bin, defined as

$$\psi_{k,t} = K_t \int_{B_{k,t}} \lambda(x) \, \mathrm{d}x.$$

We place independent truncated Gamma (TG) priors on each of the $\psi_{k,t}$ parameters, with shape and scale parameters α and β and support on $[0, \lambda_{\max}]$ where $\lambda_{\max}$ is some known upper bound on the maximum of rate functions. (The $\mathrm{TG}(\alpha, \beta, 0, \lambda_{\max})$ distribution has a density proportional to a $\mathrm{Gamma}(\alpha, \beta)$ distribution, but with truncated support $[0, \lambda_{\max}]$.) In practice the $\lambda_{\max}$ parameter may be chosen very conservatively; setting $\lambda_{\max}$ to be too large does not affect the action selection; however it is important to include an upper limit on the prior support to permit tractable regret analysis, and the chosen $\lambda_{\max}$ appears in the regret bound in Theorem 2.

The consequence of this formulation is that, conditional on actions and observations in the first t rounds, we have a posterior distribution over λ at time t which is piecewise constant. A λ_t sampled from this posterior takes the form

$$\lambda_t(x) = \sum_{k=1}^{K_t} \mathbb{I}\{x \in B_{k,t}\} \tilde{\psi}_{k,t}, \quad \text{with}$$

$$\tilde{\psi}_{k,t} \sim \mathrm{TG}(\alpha + H_{k,t}(t), \beta + \Delta_t N_{k,t}(t), 0, \lambda_{\max}), \quad (1)$$

where $H_{k,t}(s) = \sum_{j=1}^{s} \sum_{l=1}^{m_j} \mathbb{I}\{B_{k,t} \subseteq A_j\} \mathbb{I}\{X_{j,l} \in B_{k,t}\}$ gives the number events observed up to iteration s in bin $B_{k,t}$, and $N_{k,t}(s) = \sum_{j=1}^{s} \mathbb{I}\{B_{k,t} \subseteq A_j\}$ gives the number of times to iteration s that bin $B_{k,t}$ has been sensed (see Section 3.3).

Gugushvili et al. (2018) demonstrate that, with a full observation at each iteration, this posterior contracts to the truth at the optimal rate for any h-Hölder continuous rate

function λ. In particular,

$$\mathbb{E}(||\lambda_t - \lambda||_2) \leq t^{\frac{-2h}{2h+1}}$$

if $N_{k,t}(t) = t$ for all $k \in [K_t]$ and $K_t = O(t^{1/3})$. We describe in the next sub-section how the same choice of K_t gives favourable performance in our sequential decision problem, even when we only observe subintervals of $[0, 1]$.

3.3. Thompson sampling

In order to make action selection feasible, and to facilitate the inference using histograms, we constrain the action set of the TS approach using the same (increasingly fine-meshed) grid that the inference is performed over. In particular, in round t, the action A_t is constrained to lie in the set of available actions $\mathcal{A}_t$, consisting of those intervals and unions of intervals where only entire bins (no fractions of bins) are covered and the action consists of at most U subintervals. Recall U is the number of sensors, and the restriction to at most U intervals ensures that each sensor can be allocated a single contiguous subinterval. We allow the number of bins K_t to increase at rate $O(t^{1/3})$ by doubling the number of bins in line with the growth of $t^{1/3}$.

Our TS approach is described in Algorithm 1. In each round t, for each bin $k \in \{1, \ldots, K_t\}$, a rate $\tilde{\psi}_{k,t}$ is sampled according to (1), and then an action is selected that would be optimal if the true rate function were the piecewise-constant combination of these rates. As each bin rate is sampled from the current posterior and the action selected is the optimal action for this set of sampled rates, the selected action is chosen according to the posterior probability that it is the optimal one available. The optimal action conditional on a given sampled rate can be determined efficiently and exactly using the approach described in Section 3.4.

Algorithm 1 Thompson Sampling

Inputs: Gamma prior parameters $\alpha, \beta > 0$, upper truncation point $\lambda_{\max}$
Iterative Phase: For $t \geq 1$

- For each $k \in \{1, \ldots, K_t\}$, evaluate $H_{k,t}(t-1)$ and $N_{k,t}(t-1)$ and sample an index

$$\tilde{\psi}_{k,t} \sim \text{TG}(\alpha + H_{k,t}(t-1), \beta + \Delta_t N_{k,t}(t-1), 0, \lambda_{\max})$$

- Choose an action $A_t \in \mathcal{A}_t$ that maximises $r(A)$ conditional on the true rate being given by the sampled $\tilde{\psi}_{k,t}$ values, and observe the events in A_t

---

3.4. Action selection by iterative merging (AS-IM)

In this section we describe a routine, called action selection by iterative merging (AS-IM), for efficiently determining the optimal action conditional on a given sampled rate function. For the piecewise constant λ_t functions sampled by the TS approach, the above optimization problem can be formulated as an integer program in which each bin $B_{k,t}$ is either searched or not. Grant et al. (2018) solve this program (albeit for more general cost functions and fixed discretisation) using traditional integer programming methods, with exponentially high computation complexities in K_t and U. We instead introduce an efficient optimal action selection policy with polynomial sample complexity.

Firstly, we introduce additional notation that will be useful for explaining the algorithm. Throughout this section we take λ as fixed and piecewise constant on bins $B_{k,t}$, and provide a method to find A^* for this λ. An action $A \in \mathcal{A}$ can be written as the union of disjoint intervals: $A = \cup_{u=1}^{U} I_u$ and $I_u \cap I_{u'} = \emptyset$ for all $1 \leq u, u' \leq U$. Define the *weight* of an interval $I \in [0, 1]$ as $w(I) = \int_I (\lambda(x) - C) \mathrm{d}x$. Thus, we may write the optimal action as

$$A^* = \underset{\{I_u\}_{u=1}^{U}}{\text{argmax}} \sum_{u=1}^{U} w(I_u).$$

AS-IM creates an initial set of candidate intervals $\mathcal{I} = \{I_n\}_{n=1}^{N}$ such that each I_n is the union of a number of adjacent $B_{k,t}$, and for $k = 2, \ldots, K_t$, $B_{k,t}$ and $B_{k-1,t}$ belong to the same I_n if and only if $w(B_{k,t})$ and $w(B_{k-1,t})$ have the same sign. Notice that, by construction, the weights of adjacent intervals have opposite signs. If the number of intervals in $\mathcal{I}$ with positive weight is not bigger than U, AS-IM returns all such intervals as the optimal action. Otherwise, AS-IM proceeds to the next step.

AS-IM iteratively reduces the number of intervals with positive weights by merging the intervals. Specifically, let $M = \{n \in \{2, \ldots, N-1\} : |w(I_n)| \leq |w(I_{n-1})|, |w(I_n)| \leq |w(I_{n+1})|\}$ be the set of intervals that should be considered for merging. If M is empty, no further merging should take place. If M is nonempty let $n = \text{argmin}_M |w(I_n)|$ be the label in M with the smallest absolute weight; AS-IM merges I_n with its two neighbour intervals I_{n+1} and I_{n-1} into one interval and updates the set of intervals $\mathcal{I}$. The merging procedure repeats until either M is empty or the number of intervals with positive weight equals U. At this point AS-IM returns the U intervals with the largest weights as $I_1^*, I_2^*, \ldots, I_U^*$.

We have the following result on AS-IM guaranteeing its optimality and efficiency. The proof is given in the supplementary material via an induction argument.

Theorem 1. *The AS-IM policy returns the optimal action and its sample complexity is not bigger than $O(K_t \log K_t)$.*

4. Regret Bound

In this section, we present our main theoretical contribution: an upper bound on the Bayesian regret of the TS approach. There is an inevitable minimum contribution to regret due to the optimal action likely not being in our discretised action set. But by allowing the mesh to become more fine as more observations are made, we will gradually reduce this discretisation regret and permit a closer approximation to the true underlying rate function.

For the analysis that follows it will be useful to define $A_t^* = \arg\max_{A \in \mathcal{A}_t} r(A)$ as the optimal action available in round t. We then define for any $A \in \mathcal{A}_t$ and $t \in \mathbb{N}$:

$$\delta(A) = r(A^*) - r(A)$$
$$\delta_t(A) = r(A_t^*) - r(A)$$

as the *single-round regret* of the action A with respect to the optimal continuous action and the optimal action available to the algorithm in round t respectively. The difference between $\delta(A)$ and $\delta_t(A)$ is that the "discretisation regret" by choosing actions only from $\mathcal{A}_t$ is present only in $\delta(A)$. Minimising the true regret $\delta(A)$ requires balancing out estimation accuracy (requiring a coarse grid) versus discretisation regret (requiring a finer grid). We find below that choosing the number of bins K_t to be order $O(t^{1/3})$ provides the best theoretical performance guarantees. This coincides with the optimal posterior contraction rate findings in Gugushvili et al. (2018). We verify this numerically in Section 5 and find that this rebinning rate is superior to a faster linear rate of rebinning.

Theorem 2. *Consider the setup of Section 3, with U sensors, and cost of sensing C. Suppose we choose K_t such that there exist positive constants $\underline{K}, \overline{K}$ such that $\underline{K} t^{1/3} \leq K_t \leq \overline{K} t^{1/3}$. Then the Bayesian regret of Algorithm 1 satisfies*

$$BReg(T) \leq 4\overline{K}\big(\log(T+1)\log(T) + 2\lambda_{\max}\big)T^{1/3}$$
$$+ \big(CU\underline{K}^{-1} + \sqrt{24\overline{K}\lambda_{\max}\log(T)}\big)T^{2/3}.$$

This main result is that we have a $O(T^{2/3}\log^{1/2}(T))$ bound on the Bayesian regret. A lower bound for the problem is not currently available. The closest result available is that of Kleinberg (2005) for CABs with bounded Lipschitz smooth reward function and bounded noise. The bound holds only for a one-dimensional action space and is of order $\Omega(T^{2/3})$. The material differences in our setting are that the observation noise is unbounded (with Poisson tails), our reward function is defined on higher dimension (the unrestricted action space of the underlying CAB is of dimension $2U$), and that we observe additional information in the form of event locations. In the context of the nearest related results therefore, Theorem 2 suggests that the TS approach is a strongly performing policy.

Proof of Theorem 2. The Bayesian regret can be decomposed as the sum of the regret due to discretisation and the regret due to selecting suboptimal actions in $\mathcal{A}_t$, as follows

$$BReg(T) = \mathbb{E}\bigg(\sum_{t=1}^T \delta(A_t^*)\bigg) + \mathbb{E}\bigg(\sum_{t=1}^T \delta_t(A_t)\bigg)$$

The expectation in the first term only averages over λ functions, not over action selection, and the sum can be upper bounded uniformly over all λ's by considering the rate of rebinning. In particular we have the following lemma, proved in the supplementary material.

Lemma 1. *The regret due to discretisation is bounded by*

$$\sum_{t=1}^T \delta(A_t^*) \leq CU\underline{K}^{-1}T^{2/3},$$

uniformly over all rates λ.

To handle the stochastic part of the regret we use a decomposition from Proposition 1 of Russo & Van Roy (2014). For all T, for all $1 \leq t \leq T$ and for all $A \in \mathcal{A}_t$, let $L_{t,T}(A)$ and $U_{t,T}(A)$ satisfy $-C|A| \leq L_{t,T}(A) \leq U_{t,T}(A)$ (see below for a judicious choice of these variables). Then, for any T,

$$\mathbb{E}\left[\sum_{t=1}^T \delta_t(A_t)\right] = \mathbb{E}\left[\sum_{t=1}^T r(A_t^*) - r(A_t)\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^T U_{t,T}(A_t) - r(A_t)\right] + \mathbb{E}\left[\sum_{t=1}^T r(A_t^*) - U_{t,T}(A_t^*)\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^T U_{t,T}(A_t) - L_{t,T}(A_t)\right] + \lambda_{\max} \times$$
$$\left[\sum_{t=1}^T P(r(A_t^*) > U_{t,T}(A_t^*)) + \sum_{t=1}^T P(r(A_t) < L_{t,T}(A_t))\right]$$

The key step here is the second equality, which holds for TS because the distribution of $U_t(A_t)$ is precisely the distribution of $U_t(A_t^*)$ due to the method of selecting A_t. The final step follows by noting that, for any A,

$$\mathbb{E}[r(A) - U_{t,T}(A)]$$
$$\leq \mathbb{E}\left[(r(A) - U_{t,T}(A))\mathbb{I}_{\{r(A)-U_{t,T}(A)>0\}}\right]$$
$$\leq \lambda_{\max} P(r(A) > U_{t,T}(A)),$$

and similarly for $E[L_{t,T}(A) - r_t(A)]$. The $\lambda_{\max}$ term arises from $r(A) \leq \lambda_{\max} - C|A|$ and $U_{t,T}(A) \geq -C|A|$ for all $A \in \mathcal{A}_t$.

We will choose $L_{t,T}$ and $U_{t,T}$ so that each sum converges. In particular, the confidence bounds derived in Grant et al.

(2018) for Poisson random variables inspire the definition of

$$D_{k,T}(t-1) = \frac{2\log(t)}{\Delta_T N_{k,T}(t-1)} + \sqrt{\frac{6\lambda_{\max}\log(t)}{\Delta_T N_{k,T}(t-1)}}$$

for all $k \in [K_T]$, with upper and lower confidence bounds on the reward of an action $A \in \mathcal{A}_t$ at time $t \in \mathbb{N}$ as follows:

$$U_{t,T}(A) = \Delta_T \sum_{k:B_{k,T} \subseteq A} \hat{\psi}_{k,T}(t-1) + D_{k,T}(t-1) - C|A|,$$

$$L_{t,T}(A) = \Delta_T \sum_{k:B_{k,T} \subseteq A} \hat{\psi}_{k,T}(t-1) - D_{k,T}(t-1) - C|A|,$$

where $\hat{\psi}_{k,T}(t) = \frac{H_{k,T}(t)}{\Delta_T N_{k,T}(t)}$ gives the empirical mean in bin $B_{k,T}$ after t rounds. It is in the definition of $U_{t,T}$ and $L_{t,T}$ that we see the need for a T-dependence in our choice of upper and lower confidence bounds—we need to count the number times actions A_t for $t < T$ selected the bin $B_{k,T}$ defined for time T.

In the supplementary material we prove the following lemmas, which when combined are sufficient to complete the proof of Theorem 2.

Lemma 2. *For $U_{t,T}$ and $L_{t,T}$ as defined above, we have*

$$\sum_{t=1}^{T} U_{t,T}(A_t) - L_{t,T}(A_t) \leq$$

$$4\overline{K}\log(T)\log(T+1)T^{1/3} + \sqrt{24\overline{K}\lambda_{\max}\log(T)}T^{2/3}$$

Lemma 3. *The deviation probabilities can be bounded*

$$P\left(r(A_t) \notin [L_{t,T}(A_t), U_{t,T}(A_t)]\right) \leq 2K_T t^{-2}$$

Combining these results we have:

$$BReg(T) \leq CU\underline{K}^{-1}T^{2/3} + 4\overline{K}\log(T)\log(T+1)T^{1/3}$$
$$+ \sqrt{24\overline{K}\lambda_{\max}\log(T)}T^{2/3} + 2K_T\lambda_{\max}\sum_{t=1}^{T}2t^{-2}$$

which gives the required result as $\sum_{t=1}^{\infty} t^{-2} \leq \frac{\pi^2}{6}$. □

5. Simulations

In this section, we provide simulation examples on the performance of the Thompson sampling approach presented in Section 3.3. We first examine the effect of the rebinning rate on the regret and then investigate the performance of the Thompson sampling approach in relation to other algorithms.
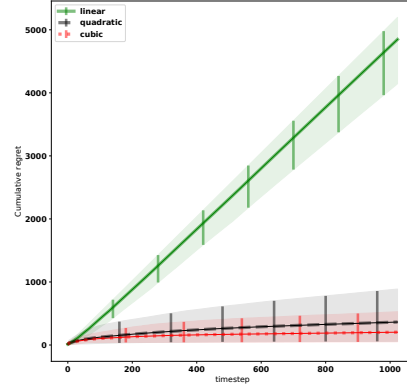

*Figure 1.* Cumulative regret comparing different rebinning rates.

5.1. Effect of rebinning rate

Firstly we examine the effect of different rebinning rates in a simple unimodal setting with $\lambda(x) = \frac{1000}{21}(x-x^2), C = 10$, and $U = 1$ sensor. This setting is chosen such that the optimal action can be calculated as $A^* = [0.3, 0.7]$. Here, and throughout our experiments, we set the prior parameters for Thompson sampling to be $\alpha = 0.5$ and $\beta = 0.5/C$, where scaling by cost C makes the prior relevant to the expected scale of costs in the problem. We also set the truncation $\lambda_{\max}$ to be ten times the true maximal value of λ; $\lambda_{\max}$ is an inconvenient parameter that is only needed for the theory, so we set it to a conservative large value that should have no influence on the real behaviour of the algorithm. The experiment is run 10 times for $T = 1024$ timesteps starting with $K_0 = 4$ bins.

We compare linear, square root and cube root rebinning rates: the number of bins K_t is doubled in rounds where t (in the linear case), $t^{1/2}$ (square root case) or $t^{1/3}$ (cube root case) is twice its value at the last rebinning time. Actions are selected using the TS method of Algorithm 1 and Fig. 1 shows that the cumulative regret is consistently lower under the cube root rate. While under the linear rebinning rate, actions with reward close to that of A^* become available more quickly, reducing the discretisation regret, the issue is that the majority of bins contain very little data and the posterior inference is heavily dependent on the prior. Under the cube root (and indeed square root) rebinning rate the action set grows more slowly but the unavoidable discretisation regret is balanced by better action selection. The square root case is surprisingly similar to the cube root case despite a weaker theoretical rate in this case. We demonstrate the shrinking of the discretisation regret in the supplementary material.

We also show, in Fig. 2, the posterior inference under the linear and cube root settings at the last time step of one run

of the experiment. The posterior under the linear rebinning is highly unconcentrated with simply insufficient numbers of observations in almost all bins. The cube root rate on the other hand results in a posterior which is much more concentrated about the truth in the region where it matters.
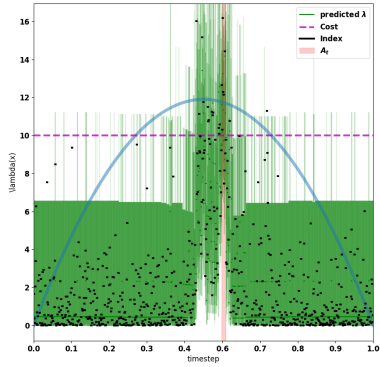
(a) Linear
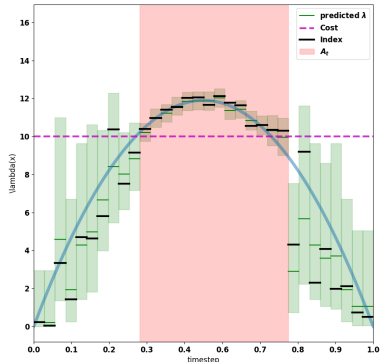
(b) Cube Root

Figure 2. Posterior under the linear and cube root rebinning rates at round $T = 1024$. We show the true rate function (blue) and cost (pink), the posterior credible interval (light green) and mean (dark green) per bin. Thompson samples are shown in black, and the selected interval, A_T, is the (red) vertical bar. The initial number of bins is 4 in both cases and the final number of bins, K_T, is 2048 for the linear rebinning schedule and 32 bins for the cube root schedule.

5.2. Comparison to Baselines

We now compare different baseline policies solely using the cube root rebinning schedule. Experiments with the unimodal rate of Section 5.1 were not informative since the problem is an easy one. We instead use a bimodal rate $\lambda(x) = \max\left(0.001, \frac{15\sin(10x)}{\sqrt{(10x+1)+x}}\right)$ with $C = 2$ and $U = 2$ sensors. Each experiment was run 10 times for $T = 1000$

time steps, starting with $K_0 = 16$ bins and terminating with $K_T = 128$ bins. In addition to the Thompson sampling approach described in Section 3.3, we consider three other algorithms, which are summarised here and described precisely in the supplementary material. (i) An upper confidence bound (UCB) approach, in which the decision-maker chooses what would be an optimal action if the true rates were $U_{t,t}$ (as defined in the proof of Theorem 1); this is essentially the FP-CUCB algorithm of Grant et al. (2018), albeit with a changing mesh, and requires the specification of an upper bound $\lambda_{\max}$ on the rate in order to define the action selection. In our experiments we fix this $\lambda_{\max}$ to the correct value; in practise a conservative estimate is usually available, but for this algorithm the choice of $\lambda_{\max}$ strongly affects the actions selected, in contrast with the TS algorithm, and we choose the most favourable $\lambda_{\max}$ for this algorithm. (ii) A modified-UCB approach (mUCB) where the empirical mean for each histogram bin $\hat{\psi}_k$ is used in place of the overall maximum rate $\lambda_{\max}$. Note this modification invalidates the concentration results used in Grant et al. (2018), but appears to improve performance in practice. (iii) An ϵ-Greedy approach where the intervals are selected according to the empirical mean for each bin $\hat{\psi}_k$ but occasionally an explorative randomisation step occurs in which the algorithm samples, for each bin, a draw from the prior. The randomisation step is taken with probability $\epsilon = 0.01$.
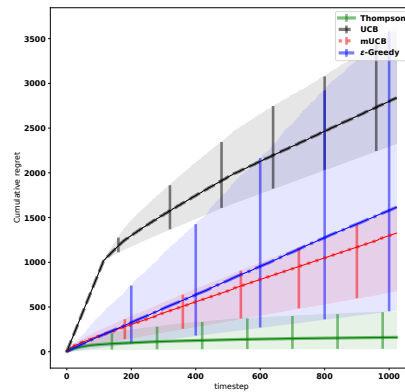

Figure 3. Cumulative regret plot for the bimodal rate functions. The experiments are repeated 10 times and the mean and 95% empirical confidence interval is shown for each policy.

The cumulative regret for each policy is shown in Figure 3. The worst performing policy is the UCB approach, despite its theoretical properties. The poor performance of the UCB policy is due to the overestimation of the true rate as can be seen in the illustrative example shown in Figure 4(d). Even after 900 iterations, the UCB values (in black) are close to the cost threshold even in the regions where the

true rate is low and there is little uncertainty. In contrast the modified-UCB values, that do not depend on $\lambda_{\max}$, are less inflated where the uncertainty is low (Figure 4(c)) resulting in more often choosing a better action. In Fig. 3 the ϵ-Greedy achieves similar mean regret to modified-UCB but with a higher variance. The ϵ-Greedy approach has the highest variance due to the greediness of the algorithm. A higher value of ϵ would reduce variance but would increase the exploration cost. The TS approach consistently outperforms all other policies.

Further intuition can also be gained from the posterior examples shown in Figure 4. These were selected at time step $T = 900$ from one of the experimental runs. The TS approach has selected an action close to optimal. Further, the posterior variance outside the optimal interval is significantly higher that in the selected regions as only a small number of observations were taken in those regions demonstrating the high efficiency of the method. In contrast both UCB approaches have uniformly low posterior variance in the entirety of the domain reflecting the large number of observations taken incurring a high exploration cost. In contrast, the ϵ-Greedy approach selects smaller than optimal intervals with high posterior variance outside these regions. This reflects an under-exploration of the greedy approach which is only able to escape bad local minima when the randomisation step is used.

In summary, the TS approach outperforms all the other approaches we have considered and is able to efficiently trade-off exploration penalty and exploitation reward.

6. Conclusion

We have presented a continuum-armed bandit model of sequential sensor placement. This model introduces the complexities of point process data and heavy-tailed reward distributions to continuum-armed bandits for the first time through its Poisson process observations. We proposed a Thompson sampling approach to make decisions based on fast non-parametric Bayesian inference and an increasingly granular action set, and derived an upper bound on the Bayesian regret of the policy which is independent of the choice of prior distribution.

In our simulation study we have studied two aspects of our approach. Firstly we examined the effect of the rebinning rate on posterior inference and regret. The theoretically-optimal cube root rate resulted in more accurate posterior inference than a linear or square root rebinning rate. This effect was also evident in a lower regret for the cube root rate.

Our empirical study also contrasted our Thompson sampling approach to alternative approaches like UCB or ϵ-greedy policies. In both the cases we examined, we found the other

(a) TS, $A_t =$
$[0.007, 0.281], [0.687, 0.882]$

(b) ϵ-greedy, $A_t =$
$[0.062, 0.312], [0.718, 0.812]$

(c) mUCB, $A_t =$
$[0., 0.320], [0.640, 0.937]$
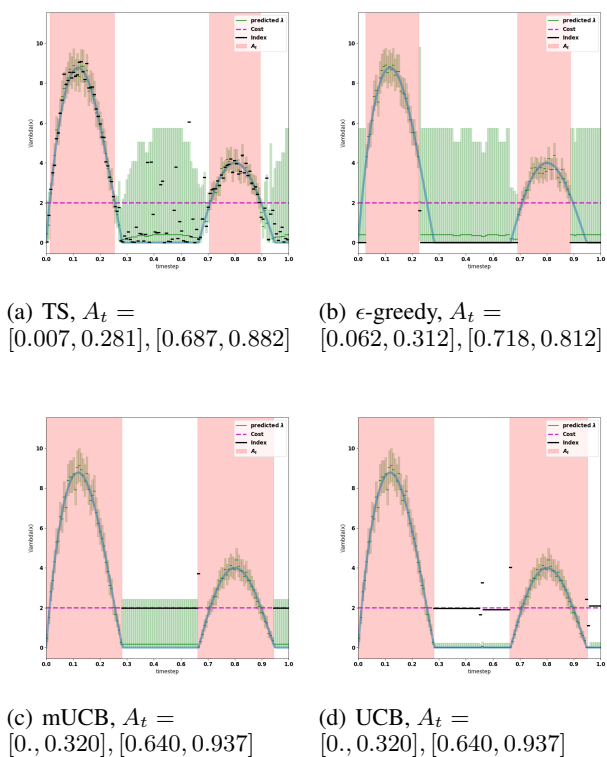
(d) UCB, $A_t =$
$[0., 0.320], [0.640, 0.937]$

Figure 4. Posterior under different action selection strategies for the bimodal test function. The true rate function (orange), posterior mean (blue) and 95% confidence interval (green infill) is shown. Rate samples for each method are shown in black for each bin and the cost threshold is the (magenta) horizontal dashed line. The optimal action is to select two intervals $A^* = [0.013, 0.280], [0.675, 0.882]$.

methods either over-explored (e.g. UCB) or over-exploited (e.g. ϵ-greedy). The TS approach achieved the best trade-off between the two and consistently achieved the lowest regret.

The observation model and rebinning strategies we have presented here are straightforward; it would be interesting to extend the algorithm and analysis to account for imperfect observations and to allow for heterogeneous bin widths, letting us capture more detail of the rate function in areas where we have made many observations and adopt a smoother estimate in others.

An alternative to the discretisation approach we have followed is to employ a continuous model such as a Cox process for which efficient approximate inference methods exist (John & Hensman, 2018). Action selection under the additive cost model would still be possible via a continuous action space extension of the AS-IM routine. The regret analysis in this setting would be more involved although recent concentration results (e.g. Kirichenko & Van Zanten, 2015) suggest possible approaches.

References

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

Agrawal, S. and Goyal, N. Analysis of Thompson Sampling for the multi-armed bandit problem. In *COLT*, 2012.

Auer, P., Ortner, R., and Szepesvári, C. Improved rates for the stochastic continuum-armed bandit problem. In *COLT*, pp. 454–468, 2007.

Basu, K. and Ghosh, S. Analysis of Thompson sampling for Gaussian process optimization in the bandit setting, 2017. arXiv:1705.06808.

Bubeck, S. and Liu, C. Prior-free and prior-dependent regret bounds for thompson sampling. In *NeurIPS*, pp. 638–646, 2013.

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

Bull, A. Adaptive-treed bandits. *Bernoulli*, 21:2289–2307, 2015.

Carlsson, J., Carlsson, E., and Devulapalli, R. Shadow prices in territory division. *Netw. Spat. Econ.*, 16:893–931, 2016.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *J. Comput. Syst. Sci.*, 78:1404–1422, 2012.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *J. Mach. Learn. Res.*, 17:1746–1778, 2016.

Chowdhury, S. and Gopalan, A. On kernelized multi-armed bandits, 2017. arXiv:1704.00445.

Grant, J., Leslie, D., Glazebrook, K., Szechtman, R., and Letchford, A. Adaptive policies for perimeter surveillance problems, 2018. arXiv:1810.02176.

Gregory, P. and Loredo, T. A new method for the detection of a periodic signal of unknown shape and period. *Astrophys. J.*, 398:146–168, 1992.

Grill, J., Valko, M., and Munos, R. Black-box optimization of noisy functions with unknown smoothness. In *NeurIPS*, pp. 667–675, 2015.

Gugushvili, S., van der Meulen, F., Schauer, M., and Spreij, P. Fast and scalable non-parametric bayesian inference for poisson point processes, 2018. arxiv:1804.03616.

Heikkinen, J. and Arjas, E. Modeling a Poisson forest in variable elevations: a nonparametric bayesian approach. *Biometrics*, 55:738–745, 1999.

Huyuk, A. and Tekin, C. Thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms, 2018. arXiv:1809.02707.

John, S. T. and Hensman, J. Large-scale Cox process inference using variational Fourier features, 2018. arXiv:1804.01016.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*, pp. 199–213, 2012.

Kirichenko, A. and Van Zanten, J. Optimality of poisson process intensity learning with gaussian processes. *J. Mach. Learn. Res.*, 16:2909–2919, 2015.

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

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

Komiyama, J., Honda, J., and Nakagawa, H. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays, 2015. arXiv:1506.00779.

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

Luedtke, A., Kaufmann, E., and Chambaz, A. Asymptotically optimal algorithms for multiple play bandits with partial feedback, 2016. arXiv:1606.09388v1.

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

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

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

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design, 2009. arXiv:0912.3995.

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

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits, 2018. arXiv:1803.04623.