Poisson Process Bandits

Sequential models and algorithms for maximising the detection of point process data

James Andrew Grant, B.Sc. (Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at Lancaster University.

July 2019

Abstract

In numerous settings in areas as diverse as security, ecology, astronomy, and logistics, it is desirable to optimally deploy a limited resource to observe events, which may be modelled as point data arising according to a Non-homogeneous Poisson process. Increasingly, thanks to developments in mobile and adaptive technologies, it is possible to update a deployment of such resource and gather feedback on the quality of multiple actions. Such a capability presents the opportunity to learn, and with it a classic problem in operations research and machine learning - the exploration-exploitation dilemma. To perform optimally, how should investigative choices which explore the value of poorly understood actions and optimising choices which choose actions known to be of a high value be balanced? Effective techniques exist to resolve this dilemma in simpler settings, but the Poisson process data brings new challenges.

In this thesis, effective solution methods for the problem of sequentially deploying resource are developed, via a combination of efficient inference schemes, bespoke optimisation approaches, and advanced sequential decision-making strategies. Furthermore, extensive theoretical work provides strong guarantees on the performance of the proposed solution methods and an understanding of the challenges of this problem and more complex extensions.

In particular, Upper Confidence Bound and Thompson Sampling (TS) approaches are derived for combinatorial and continuum-armed bandit versions of the problem, with accompanying analysis displaying that the regret of the approaches is of optimal order. A broader understanding of the performance of TS based on non-parametric models for smooth reward functions is developed, and new posterior contraction results for the Gaussian Cox Process, a popular Bayesian non-parametric model of point data, are derived. These results point to effective strategies for more challenging variants of the event detection problem, and more generally advance the understanding of bandit decision-making with complex data structures.

Acknowledgements

Some of the people acknowledged here helped me write my thesis or develop mathematical ideas, some worked alongside me and let me use them as a sounding board for ideas, some are friends and family. But I'm lucky to be able to say that *all* of them have helped me to keep a work-life balance and (rightly) put my mental health before my research at times when, left to my own devices, I might not have otherwise. There is an unfortunately high rate of mental health issues among postgraduate students and I am of the belief that one of the most positive contributions one can make towards the eventual solution of this issue is to be transparent about one's own experiences. During my PhD, as many unfortunately do, I experienced depression, anxiety, and grief, but this has been made easier, and I have come to the point of writing this thesis, thanks to the patience, support, and understanding of friends, colleagues, mentors, and family. My gratitude for this is the largest.

My first specific thanks are to my academic supervisors David Leslie, Kevin Glazebrook, and Roberto Szechtman. David, I am so grateful to have had your support over the last four years. Your generosity with your time, knowledge, and networks has made my PhD experience rich with interesting research, and exciting challenges. Kevin, I thank you for your confident and assured mentorship, your frank and helpful advice, and lively discussions. It has been a privilege to work with and learn from you. Roberto, thank you for enjoyable collaboration, useful insights, and for hosting me at the Naval Postgraduate School - a trip during my first year of the PhD that I remember

fondly as a moment when I realised how much I enjoyed research. To each of you, I am proud to have been your student.

The opportunity to travel to the Naval Postgraduate School is one of many wonderful opportunities I have had as a student in the excellently-run STOR-i Centre for Doctoral Training. The centre and the broader departments of Mathematics and Statistics, and Management Science have been fantastic and stimulating environments in which to conduct my research. I am grateful to the EPSRC, Lancaster University and industry partners for funding the centre and allowing me to take advantage of a rich variety of events and meet many intelligent and friendly peers. Jon, Idris, Kevin, Kim, Jen, Wendy, Rosemary, and Lucy are all hugely deserving of thanks for their work in running the centre. I look forward to continuing to contribute to the STOR-i community as a Research Associate in the coming year, and continuing to learn from its brilliant students and staff.

I am grateful also to friends and colleagues at Prowler.io, who supported me during an internship where part of the research that comprises my thesis was completed. In particular I am grateful to Alexis and Enrique for their mentorship and support.

My friends in my PhD cohort, Aaron, Ciara, Chrissy, Dan, Emma, Emma, Jack, Lucy, Ollie, Rebecca, and Stephen, have been wonderful people to work through the PhD alongside. Each of them are talented scientists, and more importantly genuinely lovely folk. I'm also grateful to the members of the Lancaster University Theatre Group and of the Sunday Boys choir. Both groups have helped me find some much needed creative outlet and been filled with supportive friends who in turn all instantly regretted asking what my PhD was about.

I turn now to those closest of all to me. My Mum, Shirley, and Dad, John, who passed suddenly last year, have always done everything they can to support me and encouraged me every time I came home and told them I'm thinking of doing *another* degree. I've been hugely lucky to have them as parents. My brother, Allan, grandparents, Sheena and Jim, and wider family, Elaine, Allan, Emily, Phoebe, Jacquie, Shane, and of course Dougal, have all also been such a support to me and

I am so grateful to have such a brilliant, kind family. Finally, I owe an enormous thank you to my wonderful partner Scott, for his love, support, provision of caffeine, and ability to tolerate me.

Declaration

I declare that the work in this thesis has been done by myself (unless clearly indicated otherwise and in which case included for completeness of the discussion) and has not been submitted elsewhere for the award of any other degree.

A version of Chapter 4 has been published as Grant, J.A., Leslie, D.S., Glazebrook, K., Szechtman, R., and Letchford, A. (2020). **Adaptive Policies for Perimeter Surveillance Problems**. *European Journal of Operational Research*. 283 (1): pages 265-278

A version of Chapter 5 has been published as Grant, J.A., Boukouvalas, A., Griffiths, R., Leslie, D.S., Vakili, S., and Munoz de Cote, E. (2019). **Adaptive Sensor Placement for Continuous Spaces**. *In Proceedings of 36th International Conference on Machine Learning*. PMLR 97: pages 2385–2393.

A version of Chapter 6 has been submitted for publication as Grant, J.A., and Leslie, D.S. (2019). **Posterior Contraction Rates for Gaussian Cox Processes with Non-identically Distributed Data**.

An updated version of Chapter 7 has been accepted for publication, to appear as Grant J.A., and Leslie, D.S. (2020). **On Thompson Sampling for Smoother-than-Lipschitz Bandits.** *In Proceedings of AISTATS 2020*.

James Andrew Grant

Contents

Chapter 1

Introduction

With today's advanced sensor, drone, and satellite technology we are increasingly able to detect the time and location of interesting events. This capability is useful in many applications such as ecology, defence, astronomy, logistics, and telecommunications.

Camera traps are used by ecologists to record sightings of endangered species, or document the locations of unusual behaviours. Astronomers use satellite technology to detect signals that are indicative of cosmic activity or unknown planets. Supply chain managers are reliant on sensor technology to detect disruptions in the flow of goods. Military and law enforcement officers use drone technology to monitor criminal activity such as smuggling or illegal fishing, and to covertly gather intelligence.

In such settings, efficiency and accuracy of data collection are often paramount. It can be extremely costly to deploy additional unnecessary resource, and equally every event not detected can represent a substantial financial loss, information loss, or security risk. Therefore the case for making optimal decisions with regards to event detection is a strong one.

One may naturally ask "What is an *optimal* decision?" That will of course vary depending on one's objectives and available actions. The ecologist may be happy to observe a subset of the population so long as the observations they make are of high quality, while the military objective may be to minimise the probability of failing to detect an adversary. The supply-chain manager may choose how many sensors to deploy on different links of a supply network, while the astronomer's decision may be on what hours of the day to turn on their expensive technology. In any of the applications the decision may be a one-off or an adaptable choice that can be revised every hour or day as more data is collected. Each combination of these factors may lead to an interesting problem requiring its own bespoke solution.

In this thesis, we will focus on a particular class of problems in optimal decision making for event detection which poses interesting, new, statistical and mathematical challenges and is relevant in many of the aforementioned applications.

1.1 Event Detection Maximisation

We consider event detection from the point of view of a decision-maker who has an objective of maximising (a function of) the number of events detected in a region $R \subset \mathbb{R}^d$ over some time window $[0, T]$. In formulating this objective we will assume that the detection of an event is a binary outcome - an event which occurs is either detected or not and there is no notion of quality of a detection.

We will suppose that the decision-maker is tasked with choosing an optimal allocation of *search resources* to maximise event detection. These resources may take numerous forms, for instance camera-equipped drones, mobile sensors, or human agents. Equally the *events* may represent the locations where endangered species cross a camera trap, sites where goods are smuggled over a border, times when calls arrive to a call centre etc.. The problem formulation is general and deliberately agnostic to the specific application.

Events will arise stochastically, and be represented as point data on the observable region

R. We will assume that the distribution of events follows a *non-homogeneous Poisson process*, parametrised by a non-negative rate function $\lambda : R \to \mathbb{R}_+$. The key consequence of this is that in any subregion $S \subset R$, the number of events occurring per unit time follows a Poisson distribution with mean $\int_S \lambda(x)dx$ - the integral of the rate over that subregion.

The decision-maker will have access to an action set $\mathcal{A}$. This defines the possible allocations of resource the decision-maker may select. An action $a \in \mathcal{A}$ indexes a subset S_a of S, and choosing action a is interpreted as deploying resource to search for events in the subset S_a. An action a will have value, given by its *expected reward* per unit time $r(a)$, where $r : \mathcal{A} \to \mathbb{R}$ is called a *reward function*.

The reward function r may take numerous forms, and at this stage will not assume any particular one, however some examples are useful for exposition. In the model of Chapter 5, r is specified as

$$r(a) = \int_{S_a} \lambda(x)dx - \int_{S_a} Cdx$$

for some *cost* $C > 0$. Here the reward of an action a is the expected number of events occurring in S_a minus a cost of C per unit area of S_a. In Chapter 4, r is modelled as

$$r(a) = \varphi(a) \int_{S_a} \lambda(x)dx$$

where $\varphi : \mathcal{A} \to [0, 1]$ is a decreasing function of the size of S_a and captures a phenomenon where searching a larger area decreases the probability of successfully detecting events.

If λ, r, and $\mathcal{A}$ are known by the decision-maker, they have full knowledge of the expected reward of all their available actions. Determining an optimal action is then a matter of solving the optimisation problem

$$a^* \in \operatorname*{argmax}_{a \in \mathcal{A}} r(a).$$

An action a^* should then be deployed throughout the time window $[0, T]$ to maximise reward. Commonly however, the decision-maker will have some uncertainty about λ, since their understanding of λ is likely derived from some finite data. This makes the event detection maximisation task much more challenging.

1.2 Sequential Event Detection Maximisation

When there is uncertainty about the form of λ, the optimal action a^* may not be obvious. The approach the decision-maker takes in the face of this uncertainty will depend on their flexibility to alter their selected action.

If the decision-maker is bound to selecting a single deployment of resource and using that throughout $[0, T]$, then they have two options broadly speaking. They may choose to gamble on their current information about λ (previously observed data, expert opinions etc.), and use it to form an estimate of the function λ and choose an action which is optimal with respect to this estimate. This approach maximises the expected reward with respect to the information available, but may risk choosing a highly suboptimal action if the uncertainty is large. Alternatively they may prefer a *robust* decision which is perhaps not optimal for any likely form of λ but performs reasonably well across many of the possible forms. Such an approach insures against the uncertainty but does not attempt to maximise the expected reward. Due to the uncertainty, both methods are unlikely to identify an optimal action, and will therefore incur some gap between their reward and the best possible reward, which is proportional to T.

If the decision-maker has the capacity to change their action as data is observed, they have the opportunity to improve their selected action over time and minimize the gap between their reward and the best possible. This is because the incoming data - locations of detected events - allow the decision-maker to increase their confidence in the form of λ and thus reduce their uncertainty in which actions are optimal. In other words, through the *feedback* on their actions, the decision-maker may *learn* about the value of actions.

All actions are not, however, created equal, and the success of the decision-maker will depend on which actions are selected in what sequence. The decision-maker will gain the most information by playing a mixture of actions, exploring the rate function across the region R and reducing uncertainty across the entire reward function. However, such an approach is unlikely to maximise reward, since many of the actions which are being trialled may be highly suboptimal - substantially increasing the gap between the obtained reward the best possible. The decision-maker must strike a balance between actions which contribute to learning the reward function (helping to identify actions with high reward and rule out those with low rewards) and actions which contribute to the maximisation of reward. This trade-off arises in many sequential problems and is commonly referred to as an exploration-exploitation trade-off.

It is the sequential version of the event detection maximisation problem that we will consider in this thesis. Specifically, we consider a round based set-up where at time points $t \in \{1, \dots, T\}$ the decision-maker selects an action $a_t \in \mathcal{A}$, they then observe event locations and receive a reward. The event locations can be used to update the decision-maker's beliefs about λ before they select an action at the next time step.

This sequential variant of the problem therefore poses challenges to the decision-maker in a number of dimensions. A strategy to solve the sequential event detection maximisation problem consists of three main components:

- **Inference Scheme:** A stochastic model of the generating process for events and detection probabilities under given actions, coupled with a statistical method for estimating the parameters of this model given data. Consideration should be given to the efficiency of the method, its statistical properties (such as bias, variance etc.), and how readily one can quantify uncertainty in the estimates.

- **Optimisation Approach:** An approach which can identify an optimal action a^* given either full knowledge of λ or an estimate. The accuracy and expected complexity of the optimisation are important considerations, as any method for the sequential problem may require frequent use of the optimisation approach.

- **Approach to balancing exploration and exploitation:** A decision-making rule, which utilises the inference and optimisation techniques to evaluate actions in terms of their potential for information gain and reward maximisation and choose an action which strikes the appropriate balance of exploration and exploitation at a given time.

Designing and analysing solution strategies which effectively combine these three components has been the principal research aim of this PhD project.

The contributions of the research in this thesis are threefold. Firstly, we have provided (to the best of our knowledge) the first concrete models for sequential event detection problems, which arise in numerous contexts. Secondly, and perhaps most importantly, we have proposed effective algorithms for sequential event detection problems, accompanied with theoretical and empirical evidence of their efficacy. Finally, through deriving theoretical results for the sequential event detection problem we have contributed to the multi-armed bandit and Bayesian nonparametric research communities. We have advanced the understanding of the efficacy of popular bandit algorithms on complex problems, and derived new results on the finite time properties of Bayesian models of the Poisson process.

1.3 Thesis Outline

The thesis considers the analysis of algorithms for the sequential event detection maximisation problem under various assumptions and the development of tools to aid in this analysis. The main material is presented in the following six chapters, which contain a review of related literature (Chapters 2 and 3), new research which has been published or submitted for publication (Chapters

4, 5, and 7), and results to be developed into a research publication in the future (Chapter 6). Finally Chapter 8 concludes with a summary and discussion of future work. Each main chapter is briefly summarised below. We will discuss the contributions of Chapters 4, 5, 6, and 7 in more detail in Section 3.3 once we have introduced further relevant concepts in Chapters 2 and 3.

Chapter 2: Poisson Processes

The Nonhomogeneous Poisson Process is the assumed underlying stochastic model throughout the thesis. This chapter introduces the model and discusses the practicalities of inference thereupon. We introduce simple piecewise constant estimators, frequentist and Bayesian, relevant to Chapters 4 and 5; and the Gaussian Cox Process family of models which is the topic of Chapter 7.

Chapter 3: Multi-armed Bandits and Online Learning

This chapter gives a review of multi-armed bandits and related sequential decision making problems. We discuss a number of increasingly complex problems, the best achievable performance for these problems and introduce families of solution approaches. The problems considered in Chapters 4, and 5 are variants of multi-armed bandits with Poisson (process) rewards and we will draw on the solution methods described in this chapter to tackle them.

Chapter 4: Combinatorial Multi Armed Bandit Model of Sequential Event Detection

A version of this chapter has been submitted for publication with co-authors David S. Leslie, Kevin Glazebrook, Roberto Szechtman, and Adam Letchford.

This chapter considers the sequential event detection maximisation problem with a discretised action set defined over a grid of cells along a line. Here, the number of cells searched within a round affects the probability of successfully detecting events. We formulate the problem as a combinatorial multi-armed bandit and present upper confidence bound and Thompson sampling solution approaches to the problem. We provide a theoretical analysis of the upper confidence bound algorithm, demonstrating it to achieve performance of optimal order, and show the efficacy of both approaches in an empirical study.

Chapter 5: Continuum Armed Bandit Model of Sequential Event Detection

A version of this chapter has been published in 2019 in the Proceedings of the 36th International Conference on Machine Learning. It was written with co-authors Alexis Boukouvalas, Ryan-Rhys Griffiths, David S. Leslie, Sattar Vakili, and Enrique Munoz de Cote.

This chapter considers the sequential event detection maximisation problem in a continuous action space. We propose a progressive discretisation approach where the size of the action space and the complexity of the inference model are increased as more data becomes available. We design a novel Thompson Sampling approach for this problem and derive theoretical guarantees in terms of Bayesian performance measures.

Chapter 6: Posterior Contraction Rates for Gaussian Cox Processes with Non-Identically Distributed Data

A version of this chapter has been submitted for publication with co-author David S. Leslie.

This chapter considers Gaussian Cox processes which are doubly stochastic versions of the Poisson process. Gaussian Cox processes are a useful and flexible model for non-parametric Poisson process inference. In particular, we are concerned with the rate of contraction of posterior distributions under these models. The work in this chapter provides the first bounds on these posterior distributions under non-identically distributed observations. In particular, this gives results on the rate of contraction of posteriors when different subsets of an observable region are sampled at different frequencies.

Chapter 7: Thompson Sampling for Lipschitz Bandits

In this chapter we analyse the performance of Thompson Sampling for more general sequential decision making problems with smooth reward functions. We provide performance guarantees (in the form of upper bounds on Bayesian regret) for the case where the reward function may have any number of Lipschitz smooth derivatives. These results apply to a wide range of problems, and are useful benchmarks for analysing the most advanced approaches for the sequential decision making problem.

Chapter 2

Poisson Processes

This section is devoted to a discussion of the Poisson process model, which is the assumed underlying stochastic model for point data throughout the thesis. In Section 2.1 we introduce the model and some basic schemes for inference. We devote Section 2.2 to a discussion of Gaussian Cox Processes, a nonparametric Bayesian version of Poisson processes with a Gaussian Process prior on the functional parameter. Gaussian Cox Processes are important in this thesis as the focus of Chapter 6 is the posterior contraction of these models.

2.1 Nonhomogeneous Poisson Process

The Nonhomogeneous Poisson Process (NHPP) (see e.g. Kingman (2005)) is a stochastic model of point process data. It is parameterised by a non-negative *rate function* or *intensity function* $\lambda : \mathcal{X} \to \mathbb{R}_+$ on some observable region $\mathcal{X} \subseteq \mathbb{R}^d$ for $d \in \mathbb{N}$. A special case of the NHPP is that with constant rate function λ, called a *homogeneous Poisson process* (HPP). A realisation of an NHPP is a counting measure N on $\mathcal{X}$ with the following two properties.

1. For any compact set $B \subseteq \mathcal{X}$ the number of points, $N(B)$, falling in the set B is Poisson

distributed with mean $\int_B \lambda(x)dx$.

2. The random variables $N(A), N(B)$ are independent for any disjoint sets $A, B \subset \mathcal{X}$.

Commonly, a realisation of an NHPP is represented simply by the points where the counting process is incremented. For the NHPP with rate λ on $\mathcal{X}$, the likelihood for any set of points $\{x_k\}_{k=1}^K \in \mathcal{X}$ is given as

$$L(\{x_k\}_{k=1}^K|\lambda) = \exp\left(-\int_{\mathcal{X}} dx\lambda(x)\right)\prod_{k=1}^K \lambda(x_k). \tag{2.1.1}$$

In Figure 2.1.1 a realisation of an NHPP on the unit interval is visualised. The intensity function is given by the red curve (in this case $\lambda(x) = 30(4x^5 - 3x^4 + x^3 - 2x^2 + 1)$ has been used) and the black dots represent the location of events in a single realisation of the process. Observe that there are more events occurring where the intensity function is large, and fewer where its value is lower.

Poisson process models are widely used in numerous applications (such as ecology (Heikkinen and Arjas, 1999), queueing theory (Saaty, 1961), epidemiology (Diggle et al., 2013), etc.), because of their flexibility and attractive properties such as the *superposition* property that the union of events from multiple independent Poisson processes is also distributed according to a Poisson process. There is a large literature surrounding methods for inference on the rate function λ. In the remainder of this chapter we review some of the most important approaches.

2.1.1 Parametric Inference Methods

Without assumptions on λ besides the minimal requirement that it is a non-negative, real-valued function on $\mathcal{X}$, inference can be understandably challenging. A popular approach therefore is to assume that λ lies in some smaller class of parametric functions and reduce the task of performing inference on λ to performing inference on the parameters by which it is assumed to be defined.

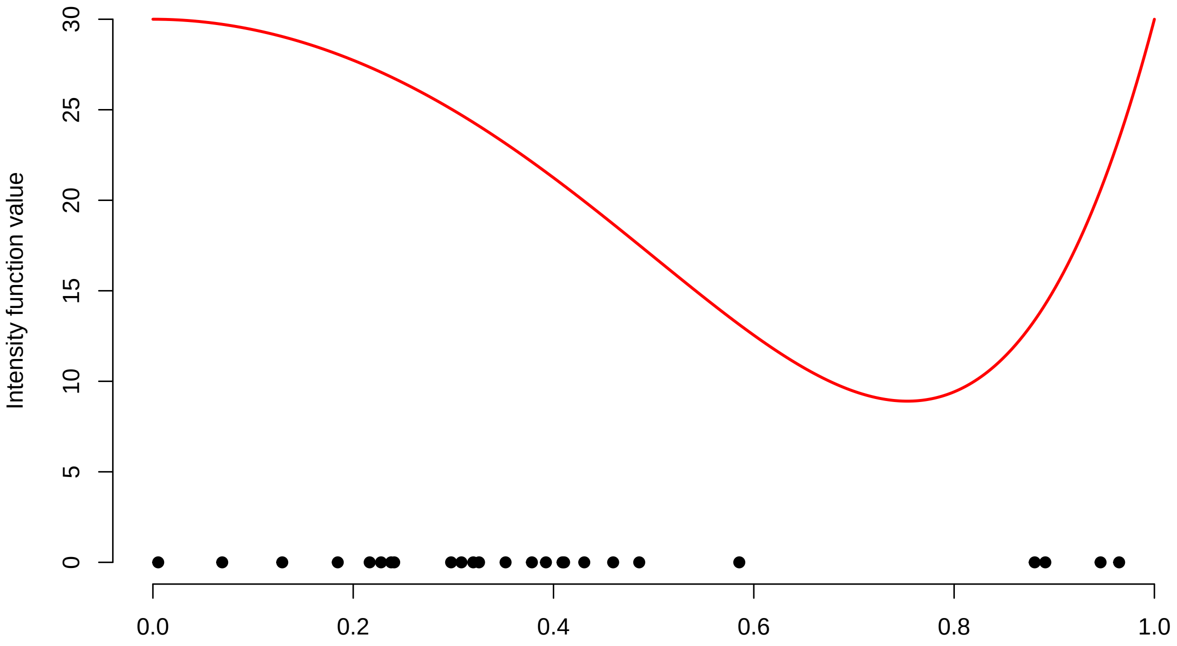


Figure 2.1.1: Example of a Non-homogeneous Poisson process in one-dimension. The intensity function is given in red, and black dots represent the location of events in a single sample.

Parametric methods are generally centred around the use of *Exponential Polynomial* or *Exponential Polynomial Trigonometric* models. An Exponential Polynomial (EP) rate function (Lewis and Shedler, 1976) is one of the form (2.1.2), while an Exponential Polynomial Trigonometric (EPT) rate function (Kuhl et al., 1997) includes the extra terms in expression (2.1.3):

$$\lambda(x) = \exp\left(\sum_{m=1}^{r} \alpha_m x^m\right), \tag{2.1.2}$$

$$\lambda(x) = \exp\left(\sum_{m=1}^{r} \alpha_m x^m + \sum_{k=1}^{p} \gamma_k \sin(\omega_k x + \phi_k)\right). \tag{2.1.3}$$

Both of these classes of model are popular because the EP and EPT functions are convenient means of modelling any continuous function arbitrarily closely, similarly to a Fourier or wavelet transformation.

The principal issue with these models, however, is that parameter estimation is neither a fast nor automatic process. For the more flexible, EPT rate function, the number of trigonometric components must be determined either from prior information on the system or by spectral analysis. The degree of the polynomial component in both models must typically be determined by Likelihood Ratio testing and then the Maximum Likelihood Estimates are determined by Newton-Raphson search. This Newton-Raphson search will also only be successful if the initial estimates are suitably close to the values giving the optimal fit. Kuhl and Wilson (2000) offer a method to fit via Ordinary or Weighted Least Squares which offers some saving on computation, but the process is still far from automatic, and uncertainty quantification is not straightforward.

2.1.2 Non-parametric Inference Methods

Simpler non-parametric methods are often popular choices. A common approach is to model λ as a piecewise combination of very simple functions, with the inbuilt assumption that the behaviour may change abruptly at certain given points or *knots* (so named because the different functions are tied together at these points). For instance, an assumption that λ is piecewise constant,

$$\lambda(x) = \sum_{m=1}^{M} C_m \mathbb{I}\{k_{m-1} < x \leq k_m\}, \quad x \in \mathcal{X},$$

for some constant levels $C_m > 0$, $m \leq M$, knot locations $k_0, \ldots, k_M$, or piecewise linear,

$$\lambda(x) = \sum_{m=1}^{M} (C_m + D_m x) \mathbb{I}\{k_{m-1} < x \leq k_m\}, \quad x \in \mathcal{X},$$

for suitable $C_m, D_m \in \mathbb{R}$, $m \leq M$ or piecewise polynomial (Kao and Chang, 1988),

$$\lambda(x) = \sum_{m=1}^{M} \sum_{a=0}^{A} C_{m,a} x^a \mathbb{I}\{k_{m-1} < x \leq k_m\}, \quad x \in \mathcal{X}$$

for suitable $C_{m,a} \in \mathbb{R}$, $m \leq M, a \leq A$, may be made. An issue with these models however is that if λ does not truly fit the assumed piecewise form, there will be an unavoidable bias to the estimation.

Gugushvili et al. (2018) propose an adaptive Bayesian form of the piecewise constant model, where the number of knots is allowed to slowly increase as the number of observations increases. The unavoidable bias due to enforcing a piecewise constant structure will then decrease since the model becomes gradually more flexible. The simplest form of their model assumes independence across the piecewise sections and is specified as follows for $\mathcal{X} = [0, 1]$.

Consider $t \in \mathbb{N}$ realisations of an NHPP having been observed, consisting for m_j points $\{X_{l,j}\}_{l=1}^{m_l}$, $l \leq t$, and partition $\mathcal{X}$ into K_t bins of equal width K_t^{-1}. For $k \in \{1, \ldots, K_t\}$, let

$$B_{k,t} = \left[\frac{k-1}{K_t}, \frac{k}{K_t}\right)$$

refer to the k^{th} bin. We then model λ as being of the form

$$\lambda_t(x) = \sum_{k=1}^{K_t} \mathbb{I}\{x \in B_{k,t}\}\psi_{k,t}, \text{ with}$$
$$\psi_{k,t} \sim Gamma(\alpha + H_{k,t}, \beta + t/K_t) \ \forall k \in \{1, \dots K_t\}$$

where the ψ parameters all are independent, and $H_{k,t} = \sum_{j=1}^{t} \sum_{l=1}^{m_j} \mathbb{I}\{X_{j,l} \in B_{k,t}\}$ gives the number of events observed over the t realisations in a bin k, and α, β are positive hyperparameters of the conjugate Gamma prior. In Chapter 5, we utilise a version of this model where the Gamma prior (and thus posterior) is replaced with a Truncated Gamma prior. This is an assumption which permits theoretical analysis. Gugushvili et al. prove that if $K_t : \mathbb{N} \to \mathbb{N}$ is defined to be $o(t^{1/(2h+1)})$ and λ is a h-Hölder continuous function, then the posterior distribution will contract around λ at the optimal rate.

A second version of the model is also proposed, where the parameters $\psi_{k,t}$ are not independent, but jointly are a realisation of a Gamma Markov Chain. While, this second version is useful for capturing the realistic scenario where $\lambda(x)$ has some spatial structure, it currently lacks the theoretical guarantees of the independent model.

In Figure 2.1.2 we illustrate the progression of this model over various values of t. A Gamma prior with shape parameter $\alpha = 30$ and scale parameter $\beta = 1$ has been used for each ψ parameter. We see that with few observations the prior dominates, but as the number of observed realisations increases, the posterior begins to concentrate around the true rate function.

In Chapter 5 we make use of this model in an approach to the sequential event detection problem. It is particularly suitable because of its computational efficiency, easy quantification of uncertainty through the tractable posterior distributions, and the simplicity of the model lends itself to tractable analysis of the performance of the resulting algorithm.

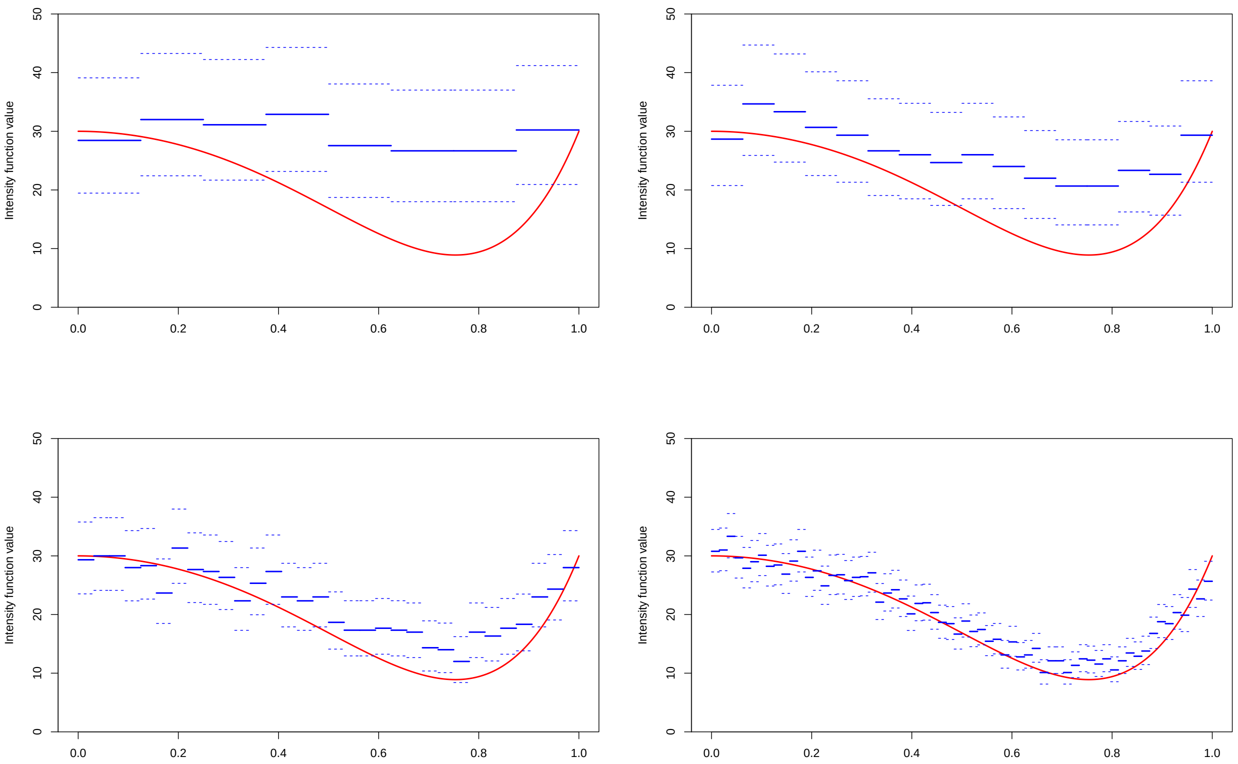


Figure 2.1.2: Evolution of the estimate under the model of Gugushvili et al. (2018) as the number of observed realisations increases. The red line plots the true intensity function, solid blue lines show the posterior mean and dashed blue lines indicate a 95% credible interval. Upper left: $t = 1$ and 8 bins. Upper right: $t = 8$ and 16 bins. Lower left: $t = 64$ and 32 bins. Lower right: $t = 512$ and 64 bins.

2.2 Gaussian Cox Processes

The Cox Process (Cox, 1955) is a doubly stochastic model of the Nonhomogeneous Poisson process, where the rate function, λ, is modelled as another stochastic process. Of particular interest is the Gaussian Cox Process (GCP) family of models where λ is modelled a priori as sample from a transformation of a *Gaussian Process* (GP) (see e.g. Williams and Rasmussen (2006)). A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. A GP on $\mathcal{X}$ can then be specified by a mean function $m : \mathcal{X} \to \mathbb{R}$ and covariance kernel $k : \mathcal{X}^2 \to \mathbb{R}_+$. We may then model a real process $g(x)$ as a GP with mean function $m(x) = \mathbb{E}(g(x))$ and covariance function $k(x, x') = \mathbb{E}((g(x) - m(x))(g(x') - m(x')))$. We write

$$g \sim \mathcal{GP}(m, k),$$

to represent this model. It is a common choice to fix the mean function to zero, and let all features of the function be captured through the covariance function.

Three examples of GCPs have been studied extensively. The Log Gaussian Cox Process (LGCP) (Rathbun and Cressie, 1994) and (Møller et al., 1998) involves modelling λ as the exponential transformation of a zero-mean GP, g,

$$\lambda(x) = \exp(g(x)), \quad x \in \mathcal{X}.$$

The Sigmoidal Gaussian Cox Process (SGCP) (Adams et al., 2009) involves modelling λ as a logistic transformation of g

$$\lambda(x) = \lambda^* \sigma(g(x)) = \lambda^* (1 + e^{g(x)})^{-1}, \quad x \in \mathcal{X},$$

where $\lambda^* > 0$ is an additional hyperparameter modelling the maximum of λ. Finally, under the

Quadratic Gaussian Cox Process (QGCP) (Lloyd et al., 2015) λ is modelled as the square of g

$$\lambda(x) = (g(x))^2, \quad x \in \mathcal{X}.$$

Other variants are possible, but the key factor is that the transformation of the GP must ensure λ only returns values in $\mathbb{R}_+$, otherwise it would not be interpretable as an NHPP intensity function.

Combining a GP prior and link function τ (such that we model $\lambda(x) = \tau(g(x))$) with the likelihood (2.1.1) of the NHPP, gives rise to following posterior on g, given observed events $\{x_k\}_{k=1}^K$

$$\pi(g|\{x_k\}_{k=1}^K) = \frac{\mathcal{GP}(g)\exp\Big(-\int_{\mathcal{X}}\tau(g(x))\mathrm{d}x\Big)\prod_{k=1}^K\tau(g(x_k)))}{\int\mathcal{GP}(g)\exp\Big(-\int_{\mathcal{X}}\tau(g(x))\mathrm{d}x\Big)\prod_{k=1}^K\tau(g(x_k)))\mathrm{d}g}.$$

This posterior distribution is said to be *doubly intractable* (Murray et al., 2006) due to the presence of the intractable integral over the observable region $\mathcal{X}$ on the numerator and over the GP g on the denominator. This poses a particular challenge to inference with GCPs.

2.2.1 Inference with Gaussian Cox Processes

Early approaches to inference with GCPs relied on making approximations to the true posterior. Diggle (1985) uses kernel densities, and Rathbun and Cressie (1994) and Møller et al. (1998) propose approximations with finite-dimensional proxy distributions for which inference *is* tractable.

Adams et al. (2009) introduced the first approach that allowed Bayesian inference to be performed on the exact posterior, specifically in the SGCP setting. They achieved this by designing a Markov Chain Monte Carlo scheme operating on an augemented version of the data. Here, the augmentation consists of adding additional events to the observed data, to create a sample representative of a homogeneous Poisson process, for which inference is tractable. The method has since been improved by Teh and Rao (2011) and Gunter et al. (2014) but carries a large computational complexity.

Recently, the literature has turned back towards approximate methods. Variational inference or Variational Bayes (see e.g. Blei et al. (2017)) is a method designed to offer substantial speed-up for complex Bayesian inference procedures at the cost of introducing some approximation bias. Instead of maximising the log-likelihood, which may be too expensive for complex posteriors, some lower bound (for which inference is more tractable) is maximised. Hensman et al. (2015) explore variational inference for the LGCP model, Lloyd et al. (2015) and John and Hensman (2018) provide variational methods for the QGCP, and Donner and Opper (2018) and Aglietti et al. (2019) provide methods for the SGCP model.

While variational inference methods are attractive since their speed-ups can make inference on large datasets feasible, a major disadvantage is the lack of theoretical understanding around the quality of approximations provided by these approaches. Recently some progress has been made to understand these methods (see e.g. Alquier et al. (2016)) but the knowledge around variational inference is yet to catch up with the understanding of exact inference. A further open question is whether the approximation error of the variational approaches is sufficiently smaller (or indeed is smaller at all) than the approximation error of the simpler non-Cox process methods, to justify using these more computationally efficient methods.

Chapter 3

Multi-armed Bandits and Online Learning

The multi-armed bandit (MAB) problem, first proposed by Thompson (1933) and later popularised by Robbins (1952), is a simple but powerful model of sequential decision making problems. The problem and a range of more complex variants were mainly studied in the Operations Research and Statistics communities throughout the latter half of the 20th century. More recently they have enjoyed a surge in interest from the Computer Science and Machine Learning fields, due to their applications in online advertising. This section consists of a discussion of these problems and the solution methods that have been developed for them.

3.1 Bandit Problems

3.1.1 K-armed Bandits

The K-armed bandit problem is the most useful starting point for our review of online learning problems. In this problem, which can be originally attributed to Thompson (1933), a decision-maker is faced with a set of K potential actions (or *arms* referring to arms of a slot machine or "bandit" which inspired the problem's name). In each of a sequence of $T \in \mathbb{N}$ rounds indexed by $t = 1, \dots, T$, the decision-maker must choose an action $A_t \in \{1, ..., K\} \equiv [K]$ to take or "play". The choice of action $A_t = k$ in round $t \in [T]$ grants the decision maker a stochastic reward $X_t = X_{k,t} \in \mathbb{R}$.

The decision-maker's aim is to maximise their cumulative reward

$$Reward(T) = \sum_{t=1}^{T} X_t$$

over T rounds, in expectation, by optimising their choice of actions. This task is complicated by uncertainty. Each action $k \in [K]$ is associated with a distribution ν_k with mean μ_k. For each action, rewards $X_{k,t}$, $t \in [T]$ are independent identically distributed samples from the respective ν_k. The decision-maker knows neither the distributions $\boldsymbol{\nu} = (\nu_1, ...\nu_K)$ nor their expected values $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$.

As such, in order to make any serious attempt at maximising expected reward, the decision-maker must choose actions strategically, balancing those which contribute to learning the unknown distributions and those which contribute to the collection of large rewards. We often refer to these two types of action as *exploratory* and *exploitative* actions, and say that the K-armed bandit model is an example of an *exploration-exploitation dilemma*. The dilemma being to decide how much exploration and how much exploitation to undertake.

Should the decision-maker spend too much of their time exploring, they will have gained a lot of information but not maximised their reward as they spent too much time learning about suboptimal actions. Should the decision-maker spend too much of their time exploiting, they may fail to maximise their rewards due to focussing on actions which are not truly optimal, since they lacked the information to realise this.

Beyond the formulation of this simple model, much of the literature on multi-armed bandits has focussed on the design and analysis of *policies* or *algorithms*. A policy (or algorithm) is a rule for selecting actions. Given a history of actions and observed rewards $\mathbf{H}_{t-1} = \{A_1, X_1, \dots, A_{t-1}, X_{t-1}\}$

over $t-1$ rounds, a policy prescribes an action A_t to be selected in the following timestep t. This mapping from $\mathbf{H}_{t-1}$ to A_t may be deterministic or stochastic. In Section 3.2 we outline popular policies for K-armed and more complex bandit problems.

While the decision-maker's aim is to maximise their expected cumulative reward, an equivalent performance measure is typically considered within the literature. The expected pseudoregret, typically referred to simply as *regret*, of a policy quantifies the difference between the expected reward obtained by an oracle policy that repeatedly plays the optimal arm, and the expected reward obtained by the policy. It may be written as follows:

$$Reg(T) = T\mu^* - \sum_{t=1}^{T} \mathbb{E}(X_{A_t,t}), \tag{3.1.1}$$

where $\mu^* = \max_{k \in [K]} \mu_k$ is the maximal expectation among the arms. Notice, that minimising regret is equivalent to maximising expected cumulative reward. Theoretical analysis of regret is typically more feasible than of reward, which is the primary reason we consider regret as our performance measure.

Theoretical assessment of the quality of an algorithm can be conducted by considering the rate at which its regret grows. As the formula for regret (3.1.1) involves many complicated stochastic dependencies (since A_t is dependent on $\mathbf{H}_{t-1}$ for $t > 1$) it is typically infeasible to compute any closed form value for regret, but upper and lower bounds on regret can be very informative.

The performance of different policies will vary, and typically our understanding of this performance will be expressed through an upper bound on regret - see Section 3.2 for more details. However, all problems have a best achievable performance (in terms of the asymptotic scaling of regret) which can be derived independently of particular algorithms. For a particular instance of the K-armed bandit problem, given by univariate reward distributions $\boldsymbol{\nu}$ with expectations $\boldsymbol{\mu}$, Lai and Robbins (1985) demonstrated that the best achievable expected regret is bounded below by an expression that is logarithmic in the horizon of the problem. They have the following asymptotic result,

$$\lim_{T\to\infty} \inf \frac{Reg(T)}{\log(T)} \geq \sum_{k:\mu_k<\mu^*} \frac{\mu^* - \mu_k}{\mathcal{K}_{\text{inf}}(\nu_k, \mu^*)} \tag{3.1.2}$$

which holds for (so-called) *consistent policies*. A consistent policy (sometimes called a *uniformly good* policy) is one that satisfies $\lim_{T\to\infty} Reg(T)/T^{\alpha} = 0$ for all $\alpha > 0$. The quantity

$$K_{\text{inf}}(\nu, \mu) = \inf\{KL(\nu, \nu') : \nu' \in \mathcal{D} \text{ and } \mathbb{E}(\nu') > \mu\} \tag{3.1.3}$$

captures the difficulty of the problem. The quantity $K_{inf}(\nu, \mu)$ is the minimum KL-divergence between an arm distribution ν and distributions with expectation greater than μ in a distributional family $\mathcal{D}$ to which all reward distributions are assumed to belong. The result in (**??**) was later extended to multi-parameter reward distributions by Burnetas and Katehakis (1997). An algorithm is said to be *asymptotically order-optimal* if it can be shown that its regret satisfies

$$\lim_{T\to\infty} \inf \frac{Reg(T)}{\log(T)} \leq C$$

for some constant $C > 0$ and *asymptotically optimal* if the constant C is $\sum_{k:\mu_k<\mu^*} \frac{\mu^*-\mu_k}{\mathcal{K}_{\text{inf}}(\nu_k,\mu^*)}$, the constant in the lower bound (3.1.2). In Section 3.2 we will discuss certain asymptotically order-optimal and optimal policies.

Stochastic bandits have applications in many real world problems where decision-makers wish to learn the optimal action among several options. In clinical trials, bandit arms may model potential treatments, with rewards being the success or failure of the treatment (Berry, 1978; Berry and Eick, 1995; Villar et al., 2015; Williamson et al., 2017). In website optimisation, web designers may model different content or aesthetic choices as bandit arms and model the problem of adapting these to maximise the clickthrough rate as a bandit problem (Hauser et al., 2009). Similarly, advertisers, news sites or search engines may model the problem of which recommendations to present to users as a bandit problem (Li et al., 2010; Lu et al., 2010; Li et al., 2016). There are further applications in queueing control, optimal patrolling, resource planning, inventory routing, optimal exploration, and a growing list of other applications - the references above are illustrative but certainly not exhaustive.

In the remainder of this section we introduce more complex learning problems which have arisen as extensions of the K-armed bandit model.

3.1.2 Combinatorial Multi-armed Bandits

The Combinatorial Multi-armed Bandit (CMAB) problem is an extension of the K-armed bandit to the setting where multiple actions can be selected simultaneously. Study of this variant of the problem can be traced back to Anantharam et al. (1987). A more general version of the problem is formalised by Chen et al. (2013).

An instance of the CMAB problem is specified by K reward distributions $\boldsymbol{\nu}$ (as in the K-armed bandit case), an *action set* $\mathcal{S} \subset \mathcal{P}([K])$, and a stochastic *reward function* $R : \mathcal{S} \to \mathbb{R}$. Here $\mathcal{P}([K])$ denotes the power set of $[K]$. The action set contains the combinations of arms that may be played simultaneously in a single round. The reward function maps from the observations from individual arms to an overall reward obtained by the decision-maker for a single round.

The problem setup is as follows. In each round $t \in [T]$, the decision-maker selects a set of arms $S_t \in \mathcal{S}$. Observations $X_{k,t}$ are generated for all $k \in S_t$ and a reward $R(S_t)$ is received. In a setting called *semi-bandit feedback* the decision-maker sees the values $X_{k,t}$ for all $k \in S_t$ and $R(S_t)$. In an alternative setting called *bandit feedback* the decision-maker sees only the reward value $R(S_t)$. The bandit feedback version of the CMAB problem is naturally more challenging.

As in the K-armed bandit problem, the decision-maker's objective is to minimise regret. Let $r(S) = \mathbb{E}(R(S))$ denote the expected reward of an action $S \in \mathcal{S}$. The regret for a CMAB problem can then be written

$$Reg(T) = T \max_{S \in \mathcal{S}} r(S) - \sum_{t=1}^{T} \mathbb{E}(R(S_t)).$$

A special case of the CMAB is the *multiple play bandit* where the reward function is the sum of the base arm rewards, $R(S_t) = \sum_{k \in S_t} X_{k,t}$. If reward distributions are supported only on $\mathbb{R}_+$ then the problem is trivial if is possible to play all the arms simultaneously. In such a case the optimal action is clearly to play all arms simultaneously. For this reason, the multiple play bandit is typically studied under the constraint that at most $m < K$ arms can be played at once, i.e. $|S| \leq m$ $\forall S \in \mathcal{S}$. Combes et al. (2015) demonstrate that regret also has a logarithmic order lower bound for the multiple play bandit, specifically,

$$\lim_{T \to \infty} \inf \frac{Reg(T)}{\log(T)} \geq c(\boldsymbol{\mu}) \tag{3.1.4}$$

where $c(\boldsymbol{\mu})$ is defined as the solution to an optimisation problem, involving the mean parameters and, again the KL-divergence function. Kveton et al. (2014) show a similar result for a CMAB with linear reward function. The problem of determining a non-trivial lower bound (regret of zero is a always a trivial lower bound) for more general reward functions is technically-speaking still open, however since there exist algorithms with logarithmic order upper bounds it is generally understood that the lower bound is logarithmic order for these problems also.

The CMAB problem has applications in many of the same areas as the K-armed bandit problem. For instance in web advertising, decisions may involve selecting multiple adverts to show at once, or optimising over several aesthetic features of a website simultaneously. In clincial trials, the MAB model where a single arm is selected for each patient may be inappropriate for modelling combination therapies where several drugs are used simultaneously. In Chapter 4 we use a CMAB problem to model the sequential event detection problem, letting arms represent subsets of the observable region and allowing the decision-maker to select several at once.

Some authors have considered a variant of the CMAB where playing a subset of arms S may *trigger* a play from further arms not in S, and an additional contribution to the reward from these arms. This variant, called the CMAB with *probabilistically triggered arms* is studied by Chen et al. (2016b), and Wang and Chen (2018). This version of the CMAB is used as a model of problems in *influence maximisation*, where advertisers select which members of a social network to target with information to best spread it through a group.

3.1.3 Continuum armed Bandits

The continuum-armed bandit (CAB) problem (also known as the $\mathcal{X}$-armed bandit or infinitely many armed bandit) is relevant to Chapter 5. In a continuum armed bandit problem, the set of available actions is generalised to some compact set $\mathcal{A} \subset \mathbb{R}^d$ of (potentially infinitely many) arms. Often, results are presented for $\mathcal{A} = [0,1]^d$, with $d \in \mathbb{N}$.

The decision-maker sequentially selects single actions $a_t \in \mathcal{A}$ over rounds $t \in [T]$. The reward $r(a_t)$ for selecting action $a_t \in \mathcal{A}$ in round t is a random pertubation of some fixed reward function $r : \mathcal{A} \to \mathbb{R}$. Typically, sub-Gaussian noise is assumed via a model $R_t = r(a_t) + \epsilon_t$ where ϵ_t is a zero-mean sub-gaussian random variable. As in the K-armed and combinatorial bandit problems, the objective remains to minimize regret, which can be written as

$$Reg(T) = T \max_{a \in \mathcal{A}} r(a) - \sum_{t=1}^{T} \mathbb{E}(R(a_t)).$$

Without further assumptions on the smoothness and domain r, this problem is arbitrarily difficult, as there may exist reward functions such that no algorithm can be expected to randomly chance upon the an optimal action in finite-time. For instance, the reward function could contain an atom with the optimal reward value, which an algorithm would fail to discover in finite-time almost surely. For this reason, the CAB problem is studied under the assumption that the reward function belongs to some well-behaved class. For instance we may assume that r is α-Hölder smooth for some $\alpha > 0$. This assumption says that there exists a constant $L > 0$ such that

$$|r(a) - r(a')| \leq L||a - a'||^{\alpha}, \quad a, a' \in \mathcal{A}. \tag{3.1.5}$$

Commonly $||\cdot||$ will be the Euclidean distance in $\mathbb{R}^d$, although this may be generalised to give other notions of smoothness. Note that $\alpha = 1$ implies r is Lipschitz smooth with Lipschitz constant L, and any $\alpha > 1$ implies r is constant. Particular attention has been devoted to the Lipschitz case (sometimes referred to as a *Lipschitz bandit*).

As the CAB problem is clearly more challenging than the problems mentioned previously, the lower bounds on regret are of a higher order. In the setting with solely the assumptions described so far, Kleinberg (2005) showed that the best achievable regret is of order $\Omega(T^{2/3})$. With further (relatively complex) assumptions on the smoothness and convexity of the reward function Bubeck et al. (2011) showed that $\Omega(\sqrt{T})$ is achievable for certain problems. Upper bounds on the regret tend to be specified as worst case results for any r in a certain set or class, rather than being tied to particular parameters (or being "problem dependent") as is typical in the K-armed bandit and CMAB cases.

3.1.4 Further Variants

Many further bandit-type problems have been introduced and studied by varying and relaxing the modelling assumptions of the above problems. We provide a brief discussion of a few of these below, but refer the interested reader to the reviews of Bubeck and Cesa-Bianchi (2012), Lattimore and Szepesvári (2018), and Slivkins (2019) and their references for a wider picture of this substantial field.

Linear Bandits

The linear bandit problem (Auer, 2002; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013) is a variant where the reward function r is linear in a d-vector of unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$. The decision maker must choose a d-dimensional action $\mathbf{x}_t$ in each round and receives a reward $r_{\boldsymbol{\theta}}(\mathbf{x}_t) = \boldsymbol{\theta}^T \cdot \mathbf{x}_t + \eta$, where η is a noise term. Generally, it is assumed that the noise term η is sub-Gaussian and as a result the convergence of least squares estimators of $\boldsymbol{\theta}$ is well-understood, allowing authors to derive closed form results. Linear bandits have applications in fields such as online advertising (Li et al., 2010) and personalisation of health interventions (Tewari and Murphy, 2017) where the different dimensions of $\mathbf{x}_t$ represent components of some complex action.

Non-stationary bandits

A key component of all the preceding (and succeeding) results on regret is the assumption of stationarity of the reward distributions. In many, if not all, of the previously mentioned applications of bandit problems the reward distributions may change through time. For instance, in online advertising, customer preferences may be seasonal or the appeal of a product may diminish as it becomes outdated. This poses a challenge for the learner, as information they have previously gathered may become uninformative. Applying standard approaches for stationary problems can be highly suboptimal as they concentrate decisions on actions that are presumed to be well understood and profitable over time - but they will not necessarily detect changes in the underlying parameters and therefore regret.

Recent treatments of this problem have considered variants where the reward distributions change abruptly at a bounded number of *changepoints* (Kocsis and Szepesvári, 2006; Garivier and Moulines, 2011), where the reward parameters change smoothly (e.g. through Brownian motion) (Slivkins and Upfal, 2008), or where the parameters may change arbitrarily but subject to a bound on the overall variation over some horizon (Besbes et al., 2014).

Best Arm Identification

Minimisation of cumulative regret is not always a decision-maker's aim. Numerous works consider the setting where the decision maker wishes to maximise their probability of identifying the optimal arm after T rounds or minimise their instantaneous regret in the final of T rounds by selecting an action as close to optimality as possible. Jamieson and Nowak (2014) provide a survey of some popular methods in this large literature. These problems are often called *Pure Exploration* problems, because the exploitation of high-reward actions to maximise cumulative reward is not present. Chen et al. (2014) study a CMAB version of the problem and Valko et al. (2013) study a CAB version of the problem. The best arm identification problem has links to other problems in Statistics and Operational Research such as Ranking and Selection (Kim and Nelson, 2007) and Ordinal Optimization (Ho et al., 2008). The CAB version of this problem has links to the field of Bayesian Optimisation (Shahriari et al., 2016).

Non-stochastic Bandits

All of the problems we have described so far where rewards are generated stochastically according to stationary reward distributions (or that change in a stochastic manner) can be considered under an alternative paradigm where this is not the case. In non-stochastic or *adversarial* bandit problems (Auer et al., 1995) worst-case performance is typically of interest, as decision-makers wish to design algorithms that will perform well even when reward sequences are designed to minimize the reward obtained (i.e. by an adversary). The algorithms and theoretical analyses in these problems are quite different to those used in stochastic settings, and are beyond the scope of this thesis. The interested reader is referred to Lattimore and Szepesvári (2018) and references within for a discussion of non-stochastic problems.

3.2 Solution Methods

Having introduced a range of learning problems, and outlined the challenges involved we will now describe families of solution approaches. We will generally introduce these ideas in the context of the K-armed bandit problem and highlight where they may be extended to more complex variants.

3.2.1 Exact Approaches

It is possible to "solve" certain bandit problems exactly - i.e. for certain bandit problems, there exist particular *known* policies which achieve the global maximum (in expectation) of particular objectives. While such known policies are not available for the problems considered in this thesis, they represent an important part of the sequential decision making literature, and we include a short discussion of them here. A much more detailed overview is given in Gittins et al. (2011).

In the regret minimization framework, rewards of the same magnitude are valued equally regardless of when they are received. An alternative view is that rewards received sooner are more valuable, and that the value of a reward decreases the further in to the future it is obtained. This idea can be captured by *discounting* the reward sequence, according to a discount parameter $\beta \in (0, 1)$. Under this "discounted reward maximisation" framework (Bellman, 1956) we take the view that there is a Markov chain $\{X_t^i\}_{t=0}^{\infty}$ called a *bandit process* associated with each arm $i \in [K]$, taking states in a space $\mathcal{X}$. The decision-maker's objective is to choose a policy $g = \{g_t\}_{t=0}^{\infty}$ maximising the expected infinite discounted sum of rewards:

$$\mathbb{E}^g \left(\sum_{t=0}^{\infty} \beta^t \sum_{i=1}^{K} r^i(X_t) \mathbb{I}\{A_t = i\} | \mathbf{X}_0 = \mathbf{x}_0 \right),$$

where $\mathbf{x}_0 = (x_0^1, \ldots, x_0^K) \in \mathcal{X}^K$ is an initial state.

One approach to such a problem is to formulate it as a Markov decision process and derive a solution using Markov decision theory (Puterman, 2014), however this approach does not scale well. An important observation of (Gittins and Jones, 1974) was that this problem (and others) can be solved exactly by an *index approach*. Such an approach for this problem (Gittins, 1979; Gittins and Jones, 1979) yields the Gittins' index for each arm i, ν^i, defined by

$$\nu^i(x^i) = \max_{\tau > 0} \frac{\mathbb{E}\left(\sum_{t=0}^{t} \beta^t r^i(X_t^i) | X_0^i = x^i \right)}{\mathbb{E}\left(\sum_{t=0}^{\tau} \beta^t | X_0^i = x^i \right)}$$

where τ is a $\{\sigma(X_1^i, \ldots, X_t^i)\}_{t=1}^{\infty}$-measurable stopping time. The β-discounted maximal reward is achieved by choosing at each decision epoch t an arm with maximal Gittins' index. This result provides an analysis of classical multi-armed bandits for which the optimisation of a Bayes β-discounted reward over an infinite horizon is the objective. Gittins' index based policies have found applications in many of the previously mention areas such as queuing control, optimal patrolling, and resource planning.

A number of different proofs of the optimality of Gittins' Indices exist (Whittle, 1980; Weber, 1992; Tsitsiklis, 1994; Bertsimas and Niño-Mora, 1996). Much of the subsequent literature focusses on efficient schemes for calculating Gittins' indices for various distributional assumptions and for more complex bandit models. Some recent works have looked at the idea of approximating the Gittins index, to permit its application in problems with long horizons (Gutin and Farias, 2016) or in the regret minimization framework (Lattimore, 2016).

NB: In the remainder of this chapter we return to thinking about the regret minimization framework.

3.2.2 Upper Confidence Bound Algorithms

Upper confidence bound (UCB) algorithms are not exactly optimal solutions to bandit problems. Rather they are myopic heuristic approaches. Here, myopic means that, in contrast to Gittins' indices, they do not explicitly "look-ahead" when considering the value of an action in the current round. However, the major advantage of UCBs and the rest of the methods described in the remainder of this chapter is that they are usually more readily implementable than Gittins' Index-type approaches which are infeasible computationally challenging in many contexts.

UCB algorithms apply the principle of *optimism in the face of uncertainty*. The basic idea is to select actions based on optimistic estimates of their mean rewards. This is usually achieved by creating an index for each action which is the upper limit of some confidence interval on the true mean reward of that action. An action with the maximal index is then selected.

Intuitively speaking, this approach is sensible because it is likely to choose actions falling into two categories - those with high uncertainty and those with high estimated mean. The upper limit of a confidence interval will either be large because of a high variance, indicating that the action to which the confidence interval pertained has high exploratory value, or because the mean estimate is large, indicating it is an action worth exploiting.

The choice of method used to construct such a confidence interval is an important one. Generally speaking the confidence intervals should be designed to contract quickly enough to ensure the algorithm shifts from exploratory actions to exploitative ones as time progresses, but also not so quickly that insufficient exploration is performed. A number of methods for designing appropriate UCB indices have been proposed and we will explore these in the remainder of this subsection.

(Frequentist) UCB

The initial idea of UCB algorithms can be attributed to Lai (1987), however the first widely-used algorithm accompanied with finite-time regret guarantees is from Auer (2002). The algorithm applies the optimism in the face of uncertainty principle using simple high probability confidence intervals.

Consider a K-armed bandit problem with reward distributions $\boldsymbol{\nu}$ whose support is on the interval $[0,1]$. Let A_t denote the action selected by an algorithm in round t, $X_{k,t}$ be the reward received from arm k in round t (let this be 0 if $A_t \neq k$), and define $N_{k,t} = \sum_{s=1}^{t} \mathbb{I}\{A_s = k\}$ as the number of plays of action k in t rounds. We also use $\hat{\mu}_{k,t} = N_{k,t}^{-1} \sum_{s=1}^{t} X_{k,t}$ as the empirical estimate of μ_k after t rounds.

Auer et al. (2002) propose the UCB1 algorithm for this problem, which is given as Algorithm 1. The key component of this algorithm is the calculation of indices $\bar{\mu}_{k,t}^{UCB1}$ (3.2.6) which consist of the current estimate of the mean value μ_k plus an *inflation term* which is decreasing in $N_{k,t}$. The logarithmic component of the inflation term ensures that the index for each arm will always eventually become large enough to force a play of that arm. However, the inflation terms will become dominated by their $N_{k,t-1}$ components as t becomes large, ensuring that the UCB1 starts to make exploitative actions based on the empirical means once the variance in these estimates becomes small.

```
Algorithm 1: UCB1 (Auer et al., 2002)

Initialisation Phase: For $t \in [K]$
    • Select action $A_t = t$
Iterative Phase: For $t = K+1, K+2, ...$
    • Calculate indices
          $\bar{\mu}_{k,t}^{UCB1} = \hat{\mu}_{k,t-1} + \sqrt{\frac{2\ln(t)}{N_{k,t-1}}}$          (3.2.6)
    • Select an action $A_t = \mathrm{argmax}_{k \in [K]} \bar{\mu}_{k,t}^{UCB1}$.
```

Auer (2002) demonstrates that the regret of UCB1 is of logarithmic order. Specifically they have the result that there exists a known constant $C > 0$ such that

$$Reg^{UCB1}(T) \leq \sum_{k \notin k^*} \frac{8 \log(T)}{(\mu^* - \mu_k)} + C \tag{3.2.7}$$

for a specific instance of the K-armed bandit problem with mean rewards $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$, where $k^* = \operatorname{argmax} \mu_k$, and reward distributions have bounded support on $[0, 1]$. Recall that as shown by Lai and Robbins (1985), $\log(T)$ is the optimal order for problem specific regret scaling, but note that the coefficient on the logarithmic term of this guarantee is suboptimal. Nevertheless as the first readily implementable policy with provably asymptotically order optimal regret, the UCB1 algorithm was a landmark development in the MAB literature.

In particular, this bound is achievable because the inflation terms are chosen based on the following property of the empirical means for $s \leq t$:

$$\mathbb{P}\left(|\mu_k - \hat{\mu}_{k,t-1}| > \sqrt{\frac{2 \log(t)}{s}} \;\middle|\; N_{k,t-1} = s \right) \leq 2t^{-3}.$$

This result is a consequence of Hoeffding's inequality. The basic idea of the proof of (3.2.7) is that UCB1 selects a sub-optimal arm $k \notin k^*$ for one of three broad reasons: 1) that the UCB inflation term of arm k is large enough to make it seem like the best arm, 2) that the over-estimation of μ_k is sufficiently large (due to the noise in the observed data) to make k seem like the best arm, or 3) that the under-estimation of μ^* is sufficiently large (due to the noise in the observed data) to make k seem like the best arm. The expected number of times these three events happen can be bounded: 1) since the inflation term is a deterministic function of the number of plays, and once N_k reaches a certain number its impact will be negligible, and 2) and 3) because the over/under-estimation can be bounded by Hoeffding's inequality. The inflation terms are chosen carefully with reference to Hoeffding's inequality, so as to balance forced exploration with exploration due to chance. The bound follows from combining this intuition with the observation that regret can be expressed in terms of the number of times sub-optimal arms are selected.

By extending this idea of choosing inflation terms such that the probability of the UCB indices being far from the true means is small, other authors have been able to extend the UCB principle to K-armed bandits without less restrictive assumptions on the reward distributions (other than their being bounded in $[0, 1]]$), and provably logarithmic order regret. In particular, Cowan et al. (2017) present versions for sub-Gaussian reward distributions, Bubeck et al. (2013) give a version for distributions where the second moment is bounded and Lattimore (2017) gives a version for distributions where the fourth moment is bounded. Where the UCB1 indices are derived by an inversion of Hoeffding's inequality (which holds for sub-Gaussian distributions), these methods derive their indices by inverting alternative concentration results (which hold for heavier-tailed distributions).

The UCB algorithm's principle has been extended to a version for CMABs in the so-named CUCB algorithm (Gai et al., 2012; Chen et al., 2013). The CUCB algorithm, given as Algorithm 2 uses the same underlying indices as the UCB1 algorithm (where $N_{k,t}$ is now calculated as $N_{k,t} = \sum_{s=1}^{t} \mathbb{I}\{k \in S_s\}$, but for action selection it passes these to a combinatorial optimisation algorithm. This then selects the best action with respect to optimism on the mean rewards of all arms. Forming optimistic estimates at the base arm level and passing these to the combinatorial optimisation algorithm should be much more efficient than forming optimistic estimates of the reward of each $S \in \mathcal{S}$ separately - providing the optimisation algorithm is efficient. Like the UCB1 algorithm for MAB problems, the CUCB algorithm achieves logarithmic order regret when applied to CMAB problems and is therefore asymptotically order optimal.

The UCB principle can also be extended to CABs. One example is the Hierarchical Optimistic Optimization (HOO) algorithm of Bubeck et al. (2009). In the HOO algorithm, the action space, $\mathcal{A}$, is partitioned into disjoint regions, each with an associated UCB, whose form is similar to the UCB1 index. In each round an action in the region with the largest UCB is selected and that

```
Algorithm 2: CUCB (Gai et al., 2012)
  Initialisation Phase: For $t \in [K]$
      • Select a random action $S_t \in \mathcal{S}$ such that $t \in S_t$
  Iterative Phase: For $t = K + 1, K + 2, ...$
      • Calculate indices
            $\bar{\mu}_{k,t}^{UCB1} = \hat{\mu}_{k,t-1} + \sqrt{\frac{2\ln(t)}{N_{k,t-1}}}, \quad k \in [K]$
      • Select an action $S_t = \mathrm{argmax}_{S \in \mathcal{S}}\, r_{\bar{\boldsymbol{\mu}}_t^{UCB1}}(S)$.
```

region is further discretised into two disjoint halves each with their own UCB for the next round. The HOO algorithm is designed to be applied to a variant of the CAB where the reward function satisfies particular smoothness and convexity properties. Under this assumption HOO can be shown to have $O(\sqrt{T})$ regret in the problem horizon T, which is asymptotically order optimal for such a problem.

KL-UCB

An alternative method, originally proposed by Lai (1987) and presented with the first finite-time analysis by Garivier and Cappé (2011), is the so-called Kullback-Leibler UCB (KL-UCB) algorithm. The indices of the KL-UCB algorithm take the form of maximisers of a function of empirical KL-divergence. They can also be thought of as an inversion of a Chernoff bound, rather than of Hoeffding's inequality. We explain the indices in the context of a K-armed bandit below.

Consider a K-armed bandit whose reward distributions are bounded in $[0, 1]$ with mean μ_k for each $k \in [K]$. Let $N_{k,t}$ and $\hat{\mu}_{k,t}$ be as defined in the previous sub-section. For $p, q \in [0, 1]^2$ let

$$d(p, q) = p \log\left(\frac{p}{q}\right) + (1 - p) \log\left(\frac{1 - p}{1 - q}\right)$$

denote the Bernoulli KL-divergence. The KL-UCB index for an arm k in round t is then calculated as

$$\bar{\mu}^{KL}_{k,t} = \max\left\{q \in [0,1] : N_{k,t-1} d(\hat{\mu}_{k,t-1}, q) \leq \log(t) + c\log(\log(t))\right\},$$

where c is a variable parameter chosen equal to 3 in the theoretical analysis. The KL-UCB algorithm then has the same form as UCB1, except the indices $\bar{\mu}^{UCB1}$ are replaced with $\bar{\mu}^{KL}$ indices.

The KL-UCB algorithm has stronger regret guarantees than the UCB1 algorithm (Algorithm 1). Indeed, for any reward distributions $\boldsymbol{\nu}$ supported on $[0,1]$ with expectations $\boldsymbol{\mu}$ the KL-UCB algorithm has

$$\lim_{T\to\infty} \sup \frac{Reg(T)}{\log(T)} \leq \sum_{k:\mu_k<\mu^*} \frac{\mu^* - \mu_k}{d(\mu_k, \mu^*)}, \tag{3.2.8}$$

which matches the lower bound of Lai and Robbins (1985) - meaning it is asymptotically optimal. The proof of this result (Cappé et al., 2013) is rather more complex than that of the regret bound for UCB1, but again ultimately relies on bounding the number of plays of suboptimal arms.

Combes et al. (2015) consider the extension of the KL-UCB to multiple play bandits, and show that logarithmic order regret can be achieved. Again the asymptotic performance is superior to that of the UCB1 based algorithm. However, the drawback of their algorithm is its computational complexity as they form a KL-UCB index on every possible *combination* of base arms rather than each base arm individually as in CUCB.

Bayes-UCB

As its name suggests, Bayes-UCB (Kaufmann et al., 2012a; Kaufmann, 2016) takes a Bayesian approach to calculating UCB indices. In this algorithm the decision making indices are quantiles of posterior distributions on the mean rewards. We will illustrate the algorithm for the class of K-armed bandit problems with one-parameter exponential family reward distributions. This class includes problems with Bernoulli rewards, Poisson rewards, and Gaussian rewards with known variance.

For reward models in such a class, let $\pi_{k,n,\hat{\mu}}$ denote the posterior distribution on μ_k, given n observations with empirical mean reward $\hat{\mu}$, for $k \in [K]$. Defining $Q(a, \pi)$ as the a quantile of the distribution π, for $a \in [0, 1]$, the Bayes-UCB indices can be written

$$\bar{q}_{k,t}^{B-UCB} = Q\left(1 - \frac{1}{t(\log(t))^c}, \pi_{k,N_k(t-1),\bar{\mu}_k(t-1)}\right),$$

where $c \geq 7$ is a real parameter chosen in such a range to guarantee theoretical results, and $\bar{\mu}_k(t-1)$ is the restriction of $\hat{\mu}_k(t-1)$ to a range $[\mu_-, \mu_+]$, which again is a requirement to obtain theoretical results. The Bayes-UCB algorithm then has the same form as UCB1 except the indices $\bar{\mu}^{UCB1}$ are replaced with $\bar{q}^{B-UCB}$ indices.

The asymptotic regret of Bayes-UCB matches that of KL-UCB, i.e. equation (3.2.8) holds for Bayes-UCB also, as shown by Kaufmann (2016). Intuitively, the approach works because the quantiles of the posterior distribution are high probability upper limits on the true mean. As more data is collected, the posteriors will contract, and eventually even when high quantiles are taken, the optimal arms will be preferred, more often than not.

GP-UCB

In the CAB problem, when the unknown parameter may be infinite-dimensional, one cannot apply the previous parametric methods. An alternative which captures the same principles is the GP-UCB method (Srinivas et al., 2010). In this Bayesian algorithm the reward function is modelled a priori as a Gaussian Process (GP) (Williams and Rasmussen, 2006). An upper confidence bound on the entire reward function is generated by computing a function of the posterior mean and variance of the GP model of reward. The properties of the GP mean that the variance function will take larger values in regions where fewer actions have been taken.

The GP-UCB algorithm is given as Algorithm 3. Actions are selected according to the rule

$$a_t \in \operatorname*{argmax}_{a \in \mathcal{A}} \mu_{t-1}(a) + \sqrt{\beta_t}\sigma_{t-1}(a),$$

where $\mu_s(a)$ and $\sigma_s(a)$ denote the posterior mean and standard deviation functions at location $a \in \mathcal{A}$ of the GP after s rounds. The values $\{\beta_t\}_{t\in\mathbb{N}}$ are a slowly increasing sequence of constants chosen with reference to the covariance kernel of the GP to minimise the regret. The maximisation step is typically approximate and performed by evaluating the index on a fine grid of values.

```
Algorithm 3: GP-UCB (Srinivas et al., 2010)
  Iterative Phase: For $t = 1, 2, \dots$
    • Select an action $a_t \in \operatorname{argmax}_{a\in\mathcal{A}} \mu_{t-1}(a) + \sqrt{\beta_t}\sigma_{t-1}(a)$.
    • Observe $y_t = r(a_t) + \eta_t$
    • Perform Bayesian update to obtain $\mu_t$ and $\sigma_t$
```

Srinivas et al. (2012) provide high-probability bounds on the regret of GP-UCB using information theoretic arguments. In particular they consider a setting where $\mathcal{A} \subset [0, r]^d$ is compact and convex with $r > 0$ and $d \in \mathbb{N}$. They show that with sub-Gaussian reward noise, the regret is $\tilde{O}(\sqrt{dT\gamma_T})$ with high-probability where γ_T is a the *maximum information gain* - a concept from information theory (see e.g. Cover and Thomas (2012)) which quantifies the mutual information between the reward function and observed rewards. The maximum information gain may be bounded depending on the kernel function of the GP and adds only logarithmic terms to the regret for common choices such as the Matern covariance (Williams and Rasmussen, 2006).

3.2.3 Thompson Sampling Algorithms

Thompson Sampling (TS) is a simple and widely applicable, but effective heuristic approach to exploration-exploitation problems. It is a Bayesian method, specified by a prior over the reward generating distributions for available actions. Actions are selected according to the posterior probability that they are optimal, with the posterior distribution being repeatedly updated as new data is observed.

Typically, one can avoid explicitly calculating the posterior probability of each action being optimal. Action selection according to the TS principle can be achieved by sampling reward generating parameters from the posterior and selecting actions that are optimal with respect to these parameters. In the K-armed bandit case, this corresponds to drawing one sample from the posterior on the reward distribution of each arm and playing the arm with the largest sample. We give the TS approach for a K-armed Beta-Bernoulli bandit as Algorithm 4.

```
Algorithm 4: Thompson Sampling ($K$-armed Bernoulli Bandit)
Iterative Phase: For $t = 1, 2, ...$
    • Sample indices
          $\bar{\mu}^{TS}_{k,t} \sim Beta(\alpha_0 + \sum_{s=1}^{t-1} X_{k,t}, \beta_0 + N_k(t-1))$
    • Select an action $A_t = \operatorname{argmax}_{k \in [K]} \bar{\mu}^{TS}_{k,t}$.
```

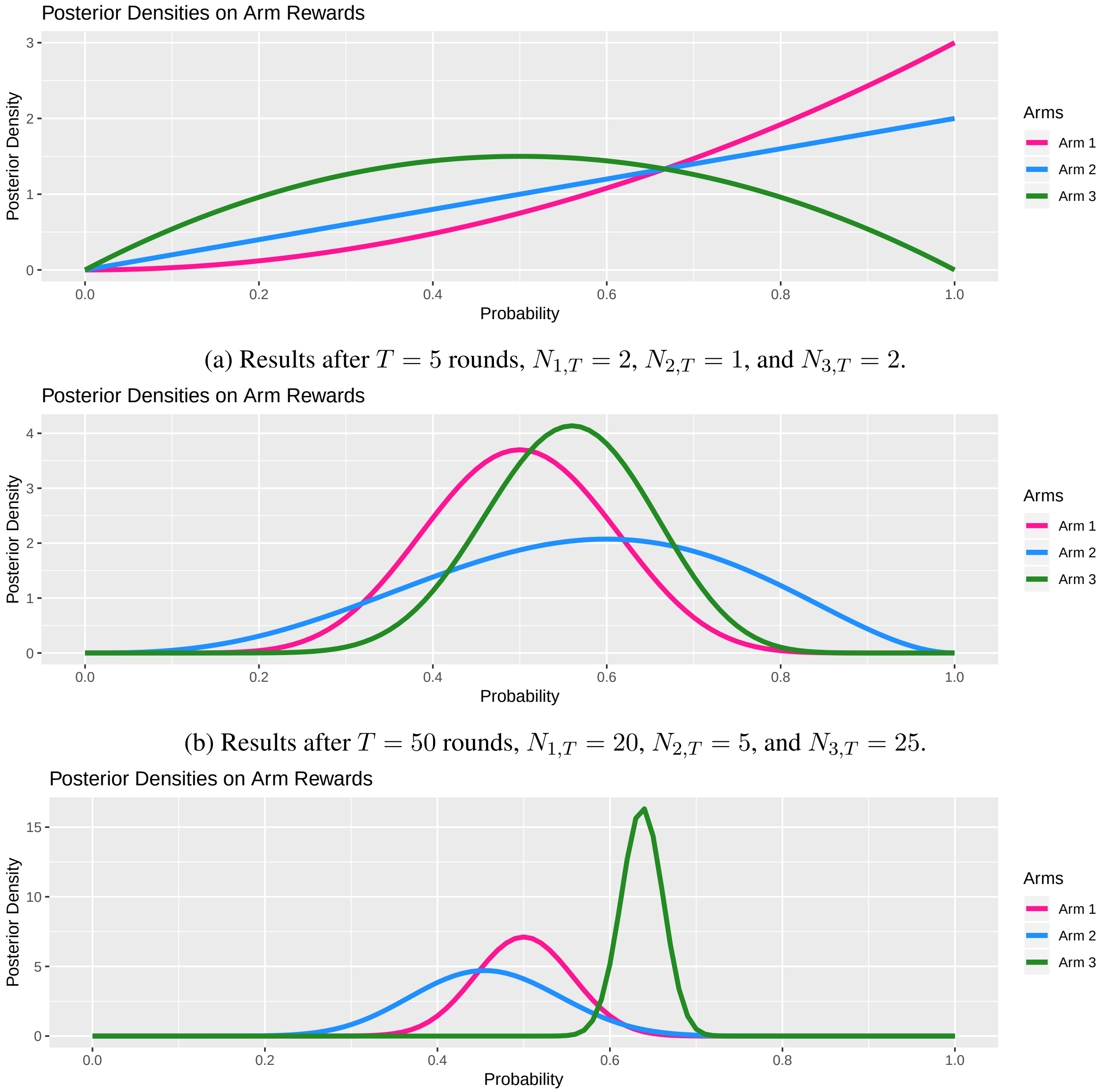
In Figures 3.2.1a, 3.2.1b, and 3.2.1c we display how the posterior distributions on arm rewards evolve as the TS algorithm progresses on a 3-armed Bernoulli bandit problem. Here the Bernoulli distributions for arms 1, 2, and 3, have parameters 0.55, 0.5 and 0.65 respectively and independent $Beta(1, 1)$ priors are used for each unknown parameter. Figure 3.2.1a displays the posterior densities after 5 rounds, there has been little exploration so far and the densities are all (relatively speaking) quite flat. In Figure 3.2.1b 50 rounds have passed and the posterior distributions are becoming more concentrated around the true parameter values. There is an increasing chance that the sampled index from arm 3 will be larger than arm 2, and since there is little information on arm 2, there is still a lot of variability in its sampled index. Finally, in Figure 3.2.1c, the posteriors following 500 rounds are displayed. As the posteriors become more concentrated it is apparent that the probability of a sample from the posterior on the mean associated with arms 1 or 2 being larger than a sample from that associated with arm 3 becomes increasingly small. Intuitively, this is why TS works. Exploration occurs initially because the flat posteriors will produce variable samples. Then as the algorithm progresses, the posteriors will contract and more often the arms with large expected rewards will be favoured.

TS was originally proposed by Thompson (1933) but received little academic interest until fairly recent empirical studies such as that of Chapelle and Li (2011) demonstrated its effectiveness and May et al. (2012) proved its asymptotic consistency. Unlike UCB approaches, whose indices are deterministic functions of the observed data (except in the case of ties), TS is a randomised algorithm. The additional stochasticity brought in by the action selection typically makes the regret of TS harder to analyse than that of UCB algorithms. Often it is easier to analyse a Bayesian version of the regret.

Consider any bandit problem with reward function r_θ, parameterised by θ and action set $\mathcal{A}$. Let A_t be the action selected at time $t \in [T]$. The Bayesian regret is then defined as

$$BReg(T) = \sum_{t=1}^{T} \mathbb{E}_{\theta_0}\left(\max_{a \in \mathcal{A}} r_\theta(a) - r_\theta(A_t) \right) \tag{3.2.9}$$

where the $\mathbb{E}_{\theta_0}$ denotes that expectation is taken with respect to the prior $\pi_0(\theta)$ on the parameters θ of the reward distribution. Published results on TS are a mixture of those on Bayesian regret and



(a) Results after $T = 5$ rounds, $N_{1,T} = 2$, $N_{2,T} = 1$, and $N_{3,T} = 2$.

(b) Results after $T = 50$ rounds, $N_{1,T} = 20$, $N_{2,T} = 5$, and $N_{3,T} = 25$.

(c) Results after $T = 500$ rounds, $N_{1,T} = 78$, $N_{2,T} = 33$, and $N_{3,T} = 389$.

Figure 3.2.1: Posterior distributions on parameters associated with reward distributions in a 3-armed Bernoulli bandit. Arm 1 has $\mu_1 = 0.55$, arm 2 has $\mu_2 = 0.5$, and arm 3 has $\mu_3 = 0.65$.

the standard (frequentist) regret, given in this setting as

$$Reg(T, \theta) = \sum_{t=1}^{T} \mathbb{E}\bigg(\max_{a \in \mathcal{A}} r_\theta(a) - r_\theta(A_t) \;\bigg|\; \theta \bigg),$$

where θ is fixed and the expectation is with respect to the reward noise and stochasticity in the action selection only.

Results on the frequentist regret of TS were found first, and relatively recently compared to those on UCB algorithms. Agrawal and Goyal (2012) and Kaufmann et al. (2012b) demonstrated TS to be an asymptotically optimal approach for a Bernoulli K-armed bandit with Beta prior distributions, in a similar fashion to the proof of asymptotic optimality for KL-UCB. Later, Korda et al. (2013) extended these results to the K-armed bandit with one-parameter exponential family reward distributions. Beyond these simple distributions, the answer to the question of asymptotic optimality is not clear-cut. Honda and Takemura (2014), for instance, demonstrated that for Gaussian distributions with unknown mean and variance (a two-parameter exponential family model) certain priors yield asymptotic optimality and others do not. For distributions that are well understood, the analysis has been extended to demonstrate the order-optimality of TS for multiple play bandits (Komiyama et al., 2015) and CMABs (Wang and Chen, 2018).

Contrastingly, results on the Bayesian regret can be obtained with almost no assumptions on the prior, thanks to one special property of TS (Russo and Van Roy, 2014). Conditional on a given history of actions and observations over $t-1$ rounds, $\mathbf{H}_{t-1}$, the posterior distributions of the optimal action and the action selected in round t by the TS approach are the same. Define, $A^* = \arg\max_{A \in \mathcal{A}} r_\theta(A)$, as the optimal action, and let A_t^{TS} be the action chosen by TS. The key result is that

$$\mathbb{P}(A^* = \cdot \mid \mathbf{H}_{t-1}) = \mathbb{P}(A_t^{TS} = \cdot \mid \mathbf{H}_{t-1}). \tag{3.2.10}$$

As a result of (3.2.10) the Bayesian regret, as defined in (3.2.9), may be decomposed as follows, for any sequence of deterministic functions $\{U_t : \mathcal{A} \to \mathbb{R}\}_{t=1}^T$,

$$\begin{aligned}
BReg(T) &= \sum_{t=1}^{T} \mathbb{E}\Big(r_\theta(A^*) - r_\theta(A_t^{TS}) \Big) \\
&= \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{E}\Big(r_\theta(A^*) - r_\theta(A_t^{TS}) \;\Big|\; \mathbf{H}_{t-1} \Big) \Bigg] \\
&= \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{E}\Big(r_\theta(A^*) - U_t(A_t^{TS}) + U_t(A_t^{TS}) - r_\theta(A_t^{TS}) \;\Big|\; \mathbf{H}_{t-1} \Big) \Bigg] \\
&= \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{E}\Big(r_\theta(A^*) - U_t(A^*) + U_t(A_t^{TS}) - r_\theta(A_t^{TS}) \;\Big|\; \mathbf{H}_{t-1} \Big) \Bigg] \\
&= \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{E}\Big(r_\theta(A^*) - U_t(A^*) \;\Big|\; \mathbf{H}_{t-1} \Big) + \mathbb{E}\Big(U_t(A_t^{TS}) - r_\theta(A_t^{TS}) \;\Big|\; \mathbf{H}_{t-1} \Big) \Bigg] \\
&= \mathbb{E}\Bigg[\sum_{t=1}^{T} \Big(r_\theta(A^*) - U_t(A^*) \Big) + \Big(U_t(A_t^{TS}) - r_\theta(A_t^{TS}) \Big) \Bigg], \qquad (3.2.11)
\end{aligned}$$

where the fourth equality is a result of (3.2.10) and the final equality uses the tower rule for expectation.

The consequence of (3.2.11), is that the task of bounding the Bayesian regret can be reduced to finding a sequence of functions such that $\sum_{t=1}^{T}(r_\theta(A) - U_t(A))$ is bounded with high probability for any $A \in \mathcal{A}$. Functions with this property are upper confidence functions, of the type used to form indices in the previous section. Notice, however, that the TS algorithm itself does not actually utilise this upper confidence bound sequence it is only introduced for the analysis. Notice also that the decomposition above holds for many choices of these functions. TS may therefore enjoy performance determined by the best performing upper confidence bound sequence without having to specify said sequence in advance.

In Russo and Van Roy (2014) this technique is deployed to derive order-optimal instance-independent bounds on the Bayesian regret of TS for MAB problems, linear bandits and CAB problems where the reward function can be modelled as a realisation of a GP. Note that in each of these analyses it is assumed that the reward noise has a sub-Gaussian distribution, but it is possible to extend beyond this setting, as we will demonstrate in Chapter 6. The general technique has since been deployed in other settings, such as bandit recommender systems (Kawale et al., 2015), and reinforcement learning (Osband et al., 2013), and we also utilise it to prove the strong performance of our TS approach for a CAB in Chapter 5.

3.2.4 Further Variants

Many further approaches can be constructed. Popular simple choices are ϵ-greedy approaches (derived from reinforcement learning (Sutton and Barto, 1998)), which choose an action randomly with small probability $\epsilon \in [0,1]$ and the action currently having highest expected reward otherwise, and *explore-then-commit* strategies (which can be traced back as far as Robbins (1952) and Anscombe (1963)), which perform a fixed amount of exploration and then settle on a single action for the remaining rounds. The *Knowledge Gradient* technique (Frazier et al., 2008; Ryzhov et al., 2012) is a compromise of sorts between UCB and Gittins indices which assigns indices to arms based on a one-step look ahead, rather than the full horizon calculation used in Gittins indices. Some weaknesses in the performance of Knowledge Gradient methods for MABs are exposed in Edwards et al. (2017)

3.3 Extensions to Poisson Process Bandits

Having introduced Poisson processes and a range of bandit problems, we are in a position to describe the challenges of the sequential event detection problem more fully. The sequential event detection problem, as described in the Introduction, is a CAB problem.

The main difference between our problem and existing treatments of the CAB problem is that our information structure is richer. In the "traditional" CAB problem, the feedback in a given round $t \in \mathbb{N}$ is in the form of a scalar reward observation - which is a noisy evaluation of the reward function r at a selected action a_t. In our setting of sequential event detection, we receive feedback in the form of a noisy evaluation of the reward function, but we will typically also receive information on the locations of any observed events. This information is useful to the decision-maker, as it will ultimately allow λ to be estimated more accurately more quickly. However, quantifying the expected reduction in uncertainty is much more complicated than in the setting where only R_t is observed. As a result, standard techniques for analysing the performance of algorithms for CAB problems do not readily apply.

A further point to note is that the Poisson distribution is not sub-Gaussian. A very common assumption in the bandit literature is that the noisy reward observations follow a sub-Gaussian distribution (or are bounded which implies the same concentration properties). Guarantees on the performance of bandit algorithms typically depend on tight concentration or martingale inequalities, and those which hold for sub-Gaussian distributions typically do not hold for the Poisson distribution. Therefore, we must consider the additional factor of incorporating alternative concentration results when adapting existing results to our problem.

Throughout the remainder of the thesis we investigate solution methods for the sequential event detection problem, and derive theoretical results related to the performance of these methods. In Chapters 4 and 5 we consider specific models of the problem and propose and analyse solution methods based on relatively simple inference schemes. In Chapters 6 and 7 we derive theoretical results to support the analysis of algorithms incorporating more sophisticated inference, which ultimately is necessary to make the best use of the event location data and potential spatial structure in the rate function.

Specifically, in Chapter 4 we study the sequential event detection problem under the imposition of a fixed discretisation on the action set and inference scheme. That is to say, that for ease of inference, action selection and the design of an appropriate sequential decision-making policy, we split the observable domain into bins. We then model the rate function as being constant over these bins and only consider actions which ensure each bin is either fully covered or not covered at all. This choice may introduce an unavoidable contribution to regret thanks to this discretisation, since the true optimal action may not be one which coincides with the chosen cell endpoints. However, it is also a choice that may often be made in practice as decisions may be made to some level of rounding. The simplicity of the model means that we can also study the effects of *filtering*, where events in a region that we choose to observe are detected according to a probability which is a function of the size of the region we choose to observe. This introduces an additional stochastic component making the analyses more involved. We show that a UCB approach specialised to the problem achieves $O(log(T))$ regret with respect to the optimal action in the discretised action set.

In Chapter 5, we tackle the CAB version of the problem, through *progressive discretisation*. As in Chapter 4 we discretise the observable region into bins and thus introduce an unavoidable contribution to regret. However, by allowing the number of bins to increase with the number of rounds, this unavoidable regret per round shrinks as the algorithm progresses. By choosing the rate at which the number of bins increases carefully, we can show that a TS algorithm, based on the Bayesian inference model of Gugushvili et al. (2018) has $\tilde{O}(T^{2/3})$ Bayesian regret.

The solution methods proposed in Chapters 4 and 5 are both based on simple inference schemes which assume independence across bins. In many situations, the rate function λ will have some smooth form, implying some spatial structure to the function, which we could exploit to improve an algorithm's performance. In Chapter 2 we mentioned that Gaussian Cox Processes are a family of nonparametric Bayesian models which can often successfully model smooth rate functions. In Section 3.2.3 we described how the TS approach can be readily deployed in scenarios where a Bayesian model is tractable. A TS approach built upon a GCP model is therefore a sensible proposition to tackle the CAB version of the sequential event detection problem while exploiting assumed spatial structure in the rate function.

To analyse the performance of such an approach, we need to consider the rate of concentration of the inference model. In Chapter 6, we consider the contraction of GCP models based on partial realisations of an NHPP. Kirichenko and Van Zanten (2015) derive a result showing the SGCP model is asymptotically consistent and the model's posterior mass contracts around the true intensity function at the optimal rate, given independent, full realisations of a fixed NHPP. This result was an important contribution to the understanding of asymptotic properties of nonparametric NHPP models, but does not extend so far as to be usable in deriving finite-time performance guarantees for sequential decision making using GCP models. In Chapter 6 we move further towards this aim, deriving finite-time results on the posterior contraction of the SGCP and QGCP models given partial realisations of an NHPP.

An alternative, less direct approach towards the analysis of a TS algorithm based on a GCP model is to utilise the idea that TS inherits the performance of the best UCB approach. In Russo and Van Roy (2014), it is observed that the least squares estimator has known concentration properties which can be generalised across many function classes. In Chapter 7 we present ongoing work which investigates results on TS for Lipschitz bandits with sub-exponential rewards (a broader class that the sequential event detection problem can be seen as fitting in to).

Chapter 4

Combinatorial Multi Armed Bandit model of Sequential Event Detection

A version of this chapter has been published as Grant, J.A., Leslie, D.S., Glazebrook, K., Szechtman, R., and Letchford, A. (2020). Adaptive Policies for Perimeter Surveillance Problems. *European Journal of Operational Research*. 283 (1): pages 265-278.

The proof of Theorem 4.7.1 is thanks to Adam Letchford and is included for completeness, rather than under the assertion that it is my own original work.

In this chapter we introduce our first formal model of the sequential event detection problem as a combinatorial multi-armed bandit. We then propose and analyse the performance of an upper confidence bound approach.

4.1 Introduction

Many common surveillance tasks concern the detection of activity along a border or perimeter. Monitoring the movements of endangered or migratory species through crossings using camera traps, covertly tracking illegal fishing in territorial waters via adaptive satellite technology, and quantifying traffic across a border using drone technology are a few among many examples of important potential aims in this domain. Equally, a number of common scheduling challenges involve events arising through time. For instance, scheduling call center staff to meet random arrivals, or deciding what times traffic cameras should be in operation to catch speeding drivers.

Approaches to the optimal design of observation strategies are invaluable not only at the operational level, but also at the strategic level because they can inform decision makers about expected outcomes for different budget scenarios and policies. In each of these tasks the notion of optimality can be equated to maximising the rate of detection of events, or equivalently, detecting as many events as possible over some fixed time horizon.

These surveillance problems coincide with the sequential event detection problem as described in Chapter 1. In this chapter we will derive a precise model of a version of the problem and propose and analyse solution approaches.

We consider a scenario where observations are made by a team of *searchers* (representing cameras, sensors, human searchers, etc.), coordinated by a central agent referred to as the *controller* who chooses which segments of a line segment each searcher will observe. As the line segment may be thought of as indexing space or time, the formulation captures a wide range of examples (we will discuss the spatial problem in what follows for ease of exposition). We will assume that events arise along the line segment according to a Poisson process and the likelihood of an event being detected depends on the allocation of resource chosen by the controller. We will, of course, be interested in a scenario where the rate of the Poisson process is unknown and the controller may update the allocation of searchers as information is gathered.

To permit analysis of this problem we shall assume two discretisations to simplify the controller's action set. We will consider that opportunities to update the allocation of searchers occur only at particular time points $t \in \mathbb{N}$. Thus, the problem can be thought of as taking place over a series of rounds. We will also suppose that the search space has been divided into a number of cells such that each searcher is allocated a connected set of cells in which to patrol, disjoint from those allocated to other searchers. Imposing this discrete structure on the problem is useful as it allows us to draw on a large literature concerning *combinatorial multi-armed bandit* (CMAB) problems, as introduced in Chapter 3, when designing and analysing solutions to the problem.

CMAB problems are relevant to this sequential resource allocation problem because they provide a framework for studying *exploration-exploitation dilemmas*, which is the principal challenge faced by the controller here. In order to reliably make optimal actions, data must be collected from all cells to accurately estimate the expected number of detections associated with an action - i.e. the action space should be *explored*. However, data is being collected on a *live* problem - real events are passing undetected when sub-optimal actions are played. As such there is a pressure to *exploit* information that has been collected and select actions which are believed to yield high detection rates over those with more exploratory value. A balance must be struck. One may suppose that this is a trivial issue which can be resolved by simply searching in all cells in all rounds. However, searching more cells will not necessarily lead to more accurate information or a higher detection rate. Under the formulation in this chapter searchers become less effective at detecting events the more cells they are allocated, because events may be undetected if a searcher is aiming to detect over too large a region. Indeed, an optimal action may well be to assign each searcher to just a single cell.

4.1.1 Related Work

We select a Poisson process, as introduced in Chapter 2, as the data generating model for our problem. We recall that there is a large literature on inference for Poisson processes, which has led to a variety of sophisticated techniques, such as those involving Gaussian processes (Adams et al., 2009; John and Hensman, 2018) or kernel-based smoothing (Diggle, 1985). However, we also recall that the theoretical properties of the more complex methods are typically only understood asymptotically (Helmers et al., 2005; Kirichenko and Van Zanten, 2015; Gugushvili et al., 2018) and therefore in the interest of developing tight guarantees on the performance of sequential decision making algorithms, we favour a simple piecewise-constant frequentist model for the Poisson process rate in this chapter. More complex inference schemes will be considered in later chapters.

Search theory has its origins in WWII with the study of barrier patrols during the Battle of the Atlantic (Koopman (1946)). The works of Stone (1976) and Washburn (2002) present a much broader and more contemporary range of applications in search theory and detection, and are by now the classic references on the subject. More closely related to our work is Szechtman et al. (2008), who study the perimeter protection problem when the parameters of the arrival process are fully known, for mobile and fixed searchers. Carlsson et al. (2016) study the problem of optimally partitioning a space in $\mathbb{R}^2$ to maximise a function of an intensity of events over the space. Their problem bears resemblance to the full information version of our problem though our solution method is quite different due to our discretisation of the problem. Our work is, to the best of our knowledge, the first to tackle the learning aspect of such a problem.

The CMAB problem models a framework where the decision-maker may select multiple actions in each round and the reward is a function of the observations from the underlying distributions associated with the selected actions. Chen et al. (2013) consider a setting where this function may be non-linear. As described in Chapter 3, a number of alternative models, such as the multiple-play bandit (Luedtke et al., 2016) and CMAB with probabilistically triggered arms (Chen et al., 2016b), have been studied since, however the model of Chen et al. (2013) is the work closest to ours as the later developments model features that are not present in our setting. The fundamental differences between our model and theirs are that we consider heavy tailed rewards and a setting where reward distributions depend on the selected action.

For bandit-type problems, it has famously been shown that under certain assumptions optimal policies can be derived by formulating the problem as a Markov Decision Process and using an index approach (Gittins et al., 2011). In CMAB problems however, these approaches are inappropriate, not least, since the combinatorial action sets induce dependencies between rewards generated by distinct actions which invalidates Gittins' theory. See also Remark 1 in Section 4.2. We will therefore focus on upper confidence bound (UCB) algorithms. We recall that these are heuristic methods which balance exploration and exploitation by selecting actions based on optimistic estimates of the associated expected rewards and can be applied to a range of bandit problems.

Auer (2002) originally proposed a UCB approach for multi-armed bandits (MAB) with underlying distributions whose support lies entirely within $[0, 1]$. Chen et al. (2013) extended the principles of this algorithm to a version suitable for CMAB problems with nonlinear rewards. Broader classes of unbounded distributions have been considered by other authors. Cowan et al. (2017), Bubeck and Cesa-Bianchi (2012), Bubeck et al. (2013), and Lattimore (2017) give UCB algorithms suitable for use with unbounded distributions, studying distributions that are Gaussian, have sub-Gaussian tails, known variance and known kurtosis respectively. Luedtke et al. (2016) have studied multiple-play bandits (a special case of CMABs) with exponential family distributions. However for CMAB problems with non-linear reward functions attention has focussed on the $[0, 1]$ case. Accompanying each of these proposals of UCB algorithms is a corresponding proof which demonstrates the performance of that algorithm achieves the optimal order, albeit with a sub-optimal coefficient.

Stronger performance guarantees (i.e. those with improved leading-order coefficients) have been obtained in MAB problems using Thompson Sampling (TS) type approaches (Kaufmann et al., 2012b; Agrawal and Goyal, 2012; Russo and Van Roy, 2016; Wang and Chen, 2018) and approaches which utilise the KL divergence of the reward distributions (Cappé et al., 2013; Kaufmann, 2016). Combes et al. (2015) have successfully extended the KL divergence based results to multiple play bandits with bounded rewards. However extending these results to the framework of our problem presents a significant analytical challenge, and therefore in this work we focus on the theoretical analysis of a more standard UCB approach. A TS alternative is presented and evaluated numerically in Section 4.5.

4.1.2 Key Contributions

This chapter makes a number of contributions to the theory of multi-armed bandits and broader online optimisation. Simultaneously, we give a practically useful solution to a real problem encountered in many applications. We summarise the headline contributions below:

- Introduction of a formal model for sequential event detection problems and an efficient integer programming solution to the full-information version of the problem;
- Introduction of the *filtered feedback* model for combinatorial multi-armed bandits;
- Development of a bespoke treatment of combinatorial bandits with *Poisson* rewards, leading to a new martingale inequality for filtered Poisson data and an accompanying UCB approach;
- Regret analysis yielding an optimal order upper bound on finite time regret of the UCB algorithm and a problem-specific lower bound on asymptotic regret for any uniformly good algorithm.

We also present extensive numerical work which displays the robustness of the UCB approach in contrast to its competitors.

4.1.3 Chapter Outline

The remainder of the chapter is structured as follows. Section 4.2 introduces a model of the sequential problem. In Section 4.3 we solve the full information problem (the non-sequential resource allocation problem where the rate function of the arrival process is known). The proposed integer programming solution forms the backbone of the proposed solution methods for the sequential problem. In Section 4.4 we introduce a solution method, the *Filtered Poisson Combinatorial Upper Confidence Bound* algorithm, for the sequential resource allocation problem, and derive a performance guarantees in the form of upper bounds on expected regret of the policy. Here, we also derive a lower bound on the expected regret possible for any policy and thus show that our algorithm has a bound of the correct order. We conclude in Sections 4.5 and 4.6 with numerical experiments and a discussion respectively.

4.2 The Model

Before introducing solution methods we give a mathematical model of the problem. Throughout the paper, for a positive integer W let the notation $[W]$ represent the set $\{1, 2, ..., W\}$.

The observation domain (line segment) comprises K cells which can be searched by U searchers. We write

$$a_k = u, \quad k \in [K], \;\; u \in [U]$$

to denote the deployment of searcher u to cell k, while

$$a_k = 0, \quad k \in [K]$$

is used when cell k goes unsearched. An *action* $\mathbf{a} := (a_1, a_2, ..., a_K) \in \{0, 1, ...U\}^K$ describes a deployment of the searchers across the line. We impose the requirement that $\mathbf{a} \in \mathcal{A}$, the *action set*, where

$$\mathcal{A} = \{\mathbf{a} : \;\; a_i = a_j = u \Rightarrow a_k = u, \;\; \forall i, j, k \in [K] : i \leq k \leq j, \;\; i < j, \text{ and } \forall u \in [U]\}.$$

This definition of $\mathcal{A}$ ensures that under any action $\mathbf{a} \in \mathcal{A}$ searchers are assigned to disjoint connected sub-regions of the perimeter. The actions are uniquely defined by indicator variables $a_{iju} \in \{0,1\}$ for $i,j \in [K]$, $i<j$ and $u \in [U]$ such that

$$a_{iju} = 1 \Leftrightarrow \text{agent } u \text{ is assigned to the cells } \{i, i+1, ..., j\} \text{ only,}$$

which will be useful for the specification of the optimisation problem in the following section.

Each action $\mathbf{a} \in \mathcal{A}$ gives rise to a certain detection probability $\gamma_k(\mathbf{a}) \in [0,1]$ in all cells $k \in [K]$. The detection probabilities capture the effectiveness of each searcher in observing an event in a specific cell. We write $\boldsymbol{\gamma}(\mathbf{a})$ for the K-vector whose k^{th} component is $\gamma_k(\mathbf{a})$. The detection probabilities are structured such that for any $\mathbf{a}, \mathbf{b} \in \mathcal{A}$ and $i \leq j$,

$$a_{iju} = b_{iju} = 1 \Rightarrow \gamma_k(\mathbf{a}) = \gamma_k(\mathbf{b}), \ \ \forall k \text{ such that } i \leq k \leq j.$$

Hence, the detection probability in a cell depends only on the sub-region assigned to the single agent searching that cell and is unaffected by the sub-regions assigned to other searchers. We assume that if a cell is searched there will be some non-zero probability of detecting events that occur. That is to say for any $k \in [K]$, $\gamma_k(\mathbf{a}) > 0$ for any $\mathbf{a} \in \mathcal{A}$ such that $a_k \neq 0$.

We consider two cases with respect to knowledge of the detection probabilities:

(I) The detection probabilities $\boldsymbol{\gamma}(\mathbf{a})$ are known for all $\mathbf{a} \in \mathcal{A}$. This scenario occurs when the controller knows $\boldsymbol{\gamma}(\mathbf{a})$ from the past.

(II) The functions $\boldsymbol{\gamma}$ have a particular known parametric form but unknown parameter values. This case is realistic when properties of the detection probabilities are dictated by physical considerations, such as the searchers' speed, the visibility in particular locations or the time for which an event is observable.

Our sequential decision problem may now be described as follows:

1. At each time $t \in \mathbb{N}$ an action $\mathbf{a}_t = (a_{1t}, \ldots, a_{Kt} \in \mathcal{A}$ is taken, inducing a detection probability $\gamma_k(\mathbf{a}_t)$ in each cell $k \in [K]$;

2. Events are generated by K independent Poisson processes, one for each cell. We use X_k to denote the number of events in cell k (whether observed or not) occurring during the period of a single search. We have

   $$X_k \sim Pois(\lambda_k), \;\; k \in [K]$$

   where the rates $\lambda_k \in \mathbb{R}_+$ are unknown, and write $\lambda_{max} \geq \max_{k \in [K]} \lambda_k$ for a known upper bound on the arrival rates. We use X_{kt} for the number of events generated in cell k during search t.

3. Should action $\mathbf{a}_t$ be taken at time t, a random vector of events $\mathbf{Y}_t = \{Y_{1t}, Y_{2t}, ..., Y_{Kt}\} \in \mathbb{N}^K$ is observed. Events in the underlying X-process are observed or not independently of each other. We write

   $$Y_{kt} | X_{kt}, \mathbf{a}_t \sim Bin(X_{kt}, \gamma_k(\mathbf{a}_t)), \;\; k \in [K].$$

   It follows from standard theory that

   $$Y_{kt} | \mathbf{a}_t \sim Pois(\lambda_k \gamma_k(\mathbf{a}_t)), \;\; k \in [K],$$

   and are independent random variables. It follows that the mean number of events observed under action $\mathbf{a}$ is given by

   $$r_{\boldsymbol{\lambda}, \boldsymbol{\gamma}}(\mathbf{a}) := \boldsymbol{\gamma}(\mathbf{a})^\top \boldsymbol{\lambda},$$

   where $\top$ denotes vector transposition and $\boldsymbol{\lambda}$ is the K-vector whose k^{th} component is λ_k.

4. We write

   $$\mathbf{H}_t = \{\mathbf{a}_1, \mathbf{Y}_1, ..., \mathbf{a}_{t-1}, \mathbf{Y}_{t-1}\}$$

for the *history* (of actions taken and events observed) available to the decision-maker at time $t \in \mathbb{N}$. A *policy* is a rule for decision-making and is determined by some collection of functions $\big\{\pi_t : \mathbf{H}_t \to \mathcal{A}, t \in \mathbb{N}\big\}$ adapted to the filtration induced by $\mathbf{H}_t$. In practice a policy will be determined by some algorithm A. We will use the terms policy and algorithm interchangeably in what follows.

The goal of analysis is the elucidation of policies whose performance (as measured by the mean number of events observed) is strong uniformly over $\boldsymbol{\lambda}, \boldsymbol{\gamma}$ and over partial horizons $\{1, 2, ..., n\} \subseteq \mathbb{N}$. We write

$$\mathbb{E}_A\bigg(\sum_{t=1}^{n} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}_t)\bigg)$$

for the mean number of events observed up to time $n \in \mathbb{N}$ under algorithm A. If we write

$$\text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} := \max_{\mathbf{a} \in \mathcal{A}} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}),$$

then it is plain that, for any choice of A

$$n \cdot \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} \geq \mathbb{E}_A\bigg(\sum_{t=1}^{n} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}_t)\bigg),$$

with achievement of the left hand side dependent on knowledge of $\boldsymbol{\lambda}$. Assessment of algorithms will be based on the associated *regret function*, the expected reward lost through ignorance of $\boldsymbol{\lambda}$, given for algorithm A and horizon n by

$$Reg^{A}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) := n \cdot \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} - \mathbb{E}_A\bigg(\sum_{t=1}^{n} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}_t)\bigg), \tag{4.2.1}$$

which is necessarily positive and nondecreasing in n, for any fixed A. In related bandit-type problems the regret of the best algorithms typically grows at $O(\log(n))$ uniformly across all $\boldsymbol{\lambda}$. We will demonstrate both that this is also the case for the algorithms we propose and that the best achievable growth for this problem is also $O(\log(n))$.

Remark 1: An alternative, indeed classical, formulation uses Bayes sequential decision theory. Here the goal of analysis is the determination of an algorithm A to maximise

$$\mathbb{E}_\rho \left[\mathbb{E}_A \left(\sum_{t=1}^{n} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}_t) \right) \right]$$

where the outer expectation is taken over some prior distribution ρ for the unknown $\boldsymbol{\lambda}$. A standard approach would formulate this as a Markov Decision Process (MDP) with an informational state at time t taken to be some sufficient statistic for $\boldsymbol{\lambda}$. The objections to this approach in this context are many. First, any serious attempt to derive such a formulation which is likely tractable will require strong assumptions on the prior ρ including, for example, independence of the components of $\boldsymbol{\lambda}$. These would each typically have a conjugate gamma prior. Even then the resulting dynamic program would be computationally intractable for any reasonable choices of K and n. Second, the realities of our problem (and, indeed, many others) are such that specification of any reasonably informed prior is impractical. Confidence in the analysis would inevitably require robustness of the performance of any proposed algorithm to specification of the prior. Indeed, our formulation centred on regret simply seeks robustness of performance with respect to values of the unknown $\boldsymbol{\lambda}$. Third, the MDP approach would require up front specification of the decision horizon n. This is practically undesirable for our problem. Moreover, the value of n is not unimportant. It will determine the nature of good policies in important ways. For example, the “last” decision at time n is guaranteed to be optimally “greedy” since there is no further need to learn about $\boldsymbol{\lambda}$ at that point.

4.3 The Full Information Problem

In order to develop strongly performing policies, it is critical that we are able to solve the *full information* optimisation problem

$$\text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} := \max_{\mathbf{a}\in\mathcal{A}} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a})$$

for any pre-specified $\boldsymbol{\lambda} \in (\mathbb{R}_{+})^{K}$. A naive proposal for a policy addressing the problem outlined in the previous section would choose an action $\mathbf{a}_t$ at time t to solve the full information problem for some estimate $\boldsymbol{\lambda}_t$ of the unknown $\boldsymbol{\lambda}$ available at time t. While such a proposal would fail to adequately address the challenge of learning about $\boldsymbol{\lambda}$, we will in the succeeding sections develop effective algorithms which choose allocations determined by solutions of full information problems for carefully chosen $\boldsymbol{\lambda}$-values.

A challenge to the solution of the full information problem is the non-linearity in $\mathbf{a}$ of the objective $r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a})$ inherited from the non-linearity of the detection mechanism $\boldsymbol{\gamma}(\mathbf{a})$. To develop efficient solution approaches we produce a formulation as a linear integer program (IP) in which this non-linearity is removed by precomputing key quantities. In particular we write

$$q_{\boldsymbol{\lambda},\boldsymbol{\gamma},iju} = \sum_{k=i}^{j} \gamma_k(\mathbf{a}_{iju})\lambda_k$$

for the mean number of events detected when agent u is allocated to the subregion $\{i, i+1, ..., j\}$ where $\mathbf{a}_{iju}$ is any $\mathbf{a} \in \mathcal{A}$ such that $a_{iju} = 1$. Efficient solution of the full information problem relies on precomputing these $q_{\boldsymbol{\lambda},\gamma,iju}$ for all $1 \leq i \leq j \leq K$, and $u \in [U]$. We now have that

$$\text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} = \max_{\{a_{iju}, 1\leq i\leq j\leq K, u\in[U]\}} \sum_{i=1}^{K}\sum_{j=i}^{K}\sum_{u=1}^{U} q_{\boldsymbol{\lambda},\boldsymbol{\gamma},iju}a_{iju} \tag{4.3.2}$$

$$\begin{aligned}\text{such that} \quad \sum_{i=1}^{K}\sum_{j=i}^{K} a_{iju} &\leq 1, \;\; u \in [U]\\ \sum_{i=1}^{k}\sum_{j=k}^{K}\sum_{u=1}^{U} a_{iju} &\leq 1, \;\; k \in [K]\\ a_{iju} &\in \{0,1\}, \;\; 1 \leq i \leq j \leq K, \;\; u \in [U].\end{aligned}$$

The first constraint above guarantees that each searcher u is assigned to at most one sub-region while the second constraint guarantees that each cell k is searched by at most one searcher. We view the solution of (4.3.2) as the optimal allocation strategy and the optimal value function as the best achievable performance for an agent with perfect knowledge of $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$.

When we require solutions to the full information problem for the implementation of algorithms for the problem described in the preceding section, we solve an appropriate version of the above IP (ie, for suitably chosen $\boldsymbol{\lambda}$) by means of branch and bound. While it can be shown that the IP (4.3.2) belongs to a class of problems which is $\mathcal{NP}$-hard (see Section 4.7.1) we find that the solution of this IP is very efficient in practice. We believe that this is because the solution of the Linear-Programming-relaxation of (4.3.2) often coincides with the exact solution of the IP. Indeed, in empirical tests this occurred more than 90% of the time and in the remaining instances the gap between the two solutions was always less that 1%. For all problem sizes considered in this paper the pre-processing and solution steps can be completed in less than a second using basic linear program solvers in the statistical programming language `R` on a single laptop.

4.4 Sequential Problem

In the sequential problem, the controller's objective is to minimise regret (4.2.1) over a sequence of rounds. To do so the controller must construct a strategy which balances exploring all cells to accurately estimate the underlying rate parameters $\boldsymbol{\lambda}$, while also exploiting the information gained to detect as many events as possible. In this section we introduce and analyse two upper confidence bound (UCB) algorithms as policies for the case of fully known detection probabilities (case (I)) and the case where only the nature of the scaling of detection probabilities is known (case (II)).

The model we introduced in Section 4.2 is closely related to the *Combinatorial Multi Armed Bandit problem* (CMAB) model of Chen et al. (2013). We recall that the CMAB problem models a scenario where a decision-maker is faced with a set of K basic actions (or *arms*) each associated with a random variable of unknown probability distribution. In each round $t \in \mathbb{N}$, the decision-maker may select a subset of basic actions to take (or *arms to pull*) and receives a *reward* which is a (possibly randomised) function of realisations of the random variables associated with the selected basic actions. The decision-maker's aim is to maximise the cumulative reward over a given horizon. Chen et al. study a CMAB problem where the decision-maker receives *semibandit feedback* on the selected actions, meaning both the overall reward and realisations of the random variables associated with the selected arms are observed. Realisations of the random variables are identically distributed for a given arm and independent both across time and arms.

In our sequential event detection problem, electing to search a cell k in a round t, i.e. setting $a_{kt} \neq 0$, is the analogue of pulling an arm k. The total number of events detected in a round is the analogue of reward. The fundamental, and non-trivial difference between our model and that of Chen et al. lies in the feedback mechanism. Our framework is more complex in two important regards. Firstly, we do not by default observe independent identically distributed (i.i.d.) realisations of the underlying random variable of interest X_{kt} each time we elect to search a cell. We observe a *filtered observation* Y_{kt} whose distribution depends on the action $\mathbf{a}_t$ selected in that round. This introduces complex dependencies within the sequence of rewards meaning standard concentration results for independent observations do not apply. Secondly, because of the U possibly heterogeneous searchers, we can have multiple ways of searching the same collection of cells. While this is implicitly permitted within the framework of Chen et al., it is not explicitly acknowledged nor to the best of our knowledge are any problems with such a structure explored in related work .

Our analytical challenge is to extend earlier work in order to meet these novel features. Specifically we will propose a UCB algorithm for both cases of our problem and derive upper bounds on the expected regret of these policies. UCB algorithms apply the principle of *optimism in the face of uncertainty* to sequential decision problems. Such an algorithm calculates an *index* for each action in each round which is the upper limit of a high probability confidence interval on the expected reward of that action and then selects the action with the highest index. In this way the algorithm will select actions which either have high indices due to a large mean estimate - leading it to exploit what has been profitable so far - or due to a large uncertainty in the empirical mean - leading it to explore actions which are currently poorly understood. As the rounds proceed, the confidence intervals will concentrate on the true means and fewer exploratory actions will be selected in favour of more exploitative ones.

4.4.1 Case (I): Known detection probabilities

In our first version of the problem, case (I), the only unknowns are the underlying rate parameters $\boldsymbol{\lambda}$. We assume that detection probability vectors $\boldsymbol{\gamma}(\mathbf{a})$ are known for all $\mathbf{a} \in \mathcal{A}$. Therefore we do not need to explicitly form UCB indices for every action separately. It will suffice to form a UCB index on each unknown λ_k for $k \in [K]$. Optimistic estimates of the value of each action will then arise by calculating the $q_{\boldsymbol{\lambda},\gamma,iju}$ quantities with the optimistic estimate of $\boldsymbol{\lambda}$ in place of known $\boldsymbol{\lambda}$.

Our proposed approach to the sequential search problem in case (I), the FP-CUCB algorithm (Filtered Poisson - Combinatorial Upper Confidence Bound), is given as Algorithm 5. The algorithm consists of an initialisation phase of length K where allocations are selected such that every cell is searched in some capacity at least once. Then in every subsequent round $t > K$, a UCB index

$$\bar{\lambda}_{k,t} = \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{6\max(1, \sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}}, \tag{4.4.3}$$

is calculated for each cell k, where $\gamma_{k,s}$ is the filtering probability applied to cell k in round s. This $\bar{\lambda}_{k,t}$ gives an upper bound for λ_k with high probability and is derived from the theory of martingale concentration. A full derivation of this term is given in the proof of the following theorem.

The algorithm then selects an action which is optimal with respect to the K-vector of inflated rates $\bar{\boldsymbol{\lambda}}_t = (\bar{\lambda}_{1,t}, ..., \bar{\lambda}_{K,t})$ by solving the IP (4.3.2) with $\bar{\boldsymbol{\lambda}}_t$ in place of $\boldsymbol{\lambda}$. The inflation terms involve a parameter $\lambda_{max} \geq \max_{k\in[K]} \lambda_k$. This is necessary to construct UCBs which concentrate at a rate that matches the concentration of Poisson random variables, which is defined by the mean parameter.

```
Algorithm 5: FP-CUCB (case (I))

Inputs: Upper bound $\lambda_{max} \geq \lambda_k, \ \ k \in [K]$.
Initialisation Phase: For $t \in [K]$
    • Select an arbitrary allocation $\mathbf{a} \in \mathcal{A}$ such that $a_t \neq 0$
Iterative Phase: For $t = K+1, K+2, ...$
    • Calculate indices
        $\bar{\lambda}_{k,t} = \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{6\max(1, \sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}}, \ \ k \in [K]$
    • Select an allocation $\mathbf{a}^*_{\bar{\boldsymbol{\lambda}}_t}$ such that $r_{\bar{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a}^*_{\bar{\boldsymbol{\lambda}}_t}) = \max_{\mathbf{a}\in\mathcal{A}} r_{\bar{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a})$.
```

To analyse the regret of this algorithm we must first introduce some additional notation for *optimality gaps*, the differences in expected reward between optimal and suboptimal actions. For

$k \in [K]$ define,

$$\Delta^k_{max} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} - \min_{\mathbf{a}\in\mathcal{A}_k} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}),$$
$$\Delta^k_{min} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} - \max_{\mathbf{a}\in\mathcal{A}_k} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}),$$

where $\mathcal{A}_k = \{\mathbf{a} \in \mathcal{A} : a_k \neq 0, \mathbf{a} \notin \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}\}$ for $k \in [K]$, and $\Delta_{max} = \max_{k\in[K]} \Delta^k_{max}$, and $\Delta_{min} = \min_{k:\Delta^k_{min}>0} \Delta^k_{min}$. The quantity Δ_{max} is then the difference in expected reward between an optimal allocation of searchers and the worst possible allocation, while Δ_{min} is the difference in expected reward between an optimal allocation and the closest to optimal suboptimal allocation. The quantities Δ^k_{max} and Δ^k_{min} are the largest and smallest gaps between the expected reward of an optimal allocation and allocations where cell k is searched in some capacity. All Δ terms depend on $\boldsymbol{\lambda}, \boldsymbol{\gamma}$ but we drop this dependence in the notation for simplicity.

Upper bound on regret

Now, in Theorem 4.4.1 we provide an analytical bound on the expected regret of the FP-CUCB algorithm in n rounds.

Theorem 4.4.1. *The regret of the FP-CUCB algorithm with λ_{max} applied to the sequential surveillance problem with known $\boldsymbol{\gamma}$ satisfies*

$$Reg^{\text{FP-CUCB}}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) \leq \sum_{k:\Delta^k_{min}>0} \frac{24K^2\lambda_{max,1}c_{0,k}}{\gamma_{k,min}\Delta^k_{min}} \log(n) + 5K\Delta_{max}, \tag{4.4.4}$$

where $\lambda_{max,1} = \max(1, \lambda_{max})$, $\gamma_{k,min} = \min_{\mathbf{a}:a_k\neq 0} \gamma_k(\mathbf{a})$, and $c_{0,k}$ are known constants depending on K, Δ^k_{max}, and Δ^k_{min}.

The full expression for each $c_{0,k}$ may be found within the proof, but the expression above captures the main dependencies of the bound. We notice that small detections probabilities may lead to a large regret bound, since the algorithm may take longer to learn. Further, because the bound is derived via bounding the number of sub-optimal actions, small optimality gaps Δ^k_{min} may also lead to a large regret bound since the algorithm may take a very long time to differentiate between the optimal actions and those that have only slightly lower reward.

To give a proof of this theorem we must introduce a new way of thinking about the action space. Consider that while we have previously (for ease of exposition) defined actions in terms of allocations of searchers to cells, $\mathbf{a} \in \mathcal{A}$, the real impact on reward comes from the vectors of detection probabilities, $\boldsymbol{\gamma}(\mathbf{a})$, which arise from these allocations. As multiple allocations may give rise to the same vector of detection probabilities (if, for instance, two searchers have identical capabilities, then switching their assignments would have no impact on the quality of the search) the set $\mathcal{G} = \{\boldsymbol{\gamma}(\mathbf{a}),\ \forall \mathbf{a} \in \mathcal{A}\}$ of possible detection probability vectors most parsimoniously describes the set of possible actions in this problem.

For an element $\mathbf{g} = (g_1, ..., g_K) = \mathcal{G}$ we then have expected reward $\mathbf{g}^T \cdot \boldsymbol{\lambda}$ and optimality gap $\Delta_{\mathbf{g}} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} - \mathbf{g}^T \cdot \boldsymbol{\lambda}$. Let $\mathcal{G}_k$ be the set of vectors $\mathbf{g}$ with $g_k > 0$ and $\mathcal{G}_{k,B} \subseteq \mathcal{G}_k$ be the set of vectors in $\mathcal{G}_k$ with sub-optimal expected reward - i.e. $\mathcal{G}_{k,B} = \{\mathbf{g} \in \mathcal{G}_k : \Delta_{\mathbf{g}} > 0\}$. Let $B_k = |\mathcal{G}_{k,B}|$ and label the vectors in $\mathcal{G}_{k,B}$ as $\mathbf{g}^1_{k,B}, \mathbf{g}^2_{k,B}, ..., \mathbf{g}^{B_k}_{k,B}$ in increasing order of expected reward. We use the following notation for optimality gaps with respect to these ordered vectors

$$\Delta^{k,j} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} - (\mathbf{g}^j_{k,B})^T \cdot \boldsymbol{\lambda} \quad j \in [B_k],\ \ k \in [K] \tag{4.4.5}$$

and thus the gaps defined previously can be expressed as $\Delta^k_{max} = \Delta^{k,1}$ and $\Delta^k_{min} = \Delta^{k,B_k}$. We introduce counters $D_{k,t} = \sum_{s=1}^t g_{k,s}$ for $k \in [K]$, $t \in \mathbb{N}$ where $\mathbf{g}_s$ is the detection probability vector selected in round s. These allow us to keep track of the total detection probability *applied* to a cell up to the end of round t.

The central idea in proving Theorem 4.4.1 is that if for a certain sub-optimal action $\mathbf{g} : \Delta_{\mathbf{g}} > 0$, all the cells k with $g_k > 0$ have been sampled sufficiently, the mean estimates ought to be accurate enough that the probability of selecting that sub-optimal action again before horizon n is small. We show that this sufficient sampling level is $O(\log(n))$ and the "small" probabilities of selecting the sub-optimal action after sufficient sampling are so small as to converge to a constant. Thus by re-expressing expected regret as a function of the number of plays of sub-optimal actions, we can bound it from above as the sum of a $O(\log(n))$ term derived from the sufficient sampling level and a constant independent of n.

To count the plays of sub-optimal actions we maintain counters $N_{k,t}$, which collectively count the number of suboptimal plays. We update them as follows. Firstly, after the K initialisation rounds we set $N_{k,K} = 1$ for $k \in [K]$. Thereafter, in each round $t > K$, let $k' = \arg\min_{j:g_{j,t}>0} N_{j,t-1}$, where if k' is non-unique, we choose a single value randomly from the minimising set. If $\mathbf{g}_t^T \cdot \boldsymbol{\lambda} \neq \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}$ then we increment $N_{k'}$ by one, i.e. set $N_{k',t} = N_{k',t-1} + 1$. The key consequences of these updating rules are that $\sum_{k=1}^{K} N_{k,t}$ provides an upper bound on the number of suboptimal plays in t rounds, and $D_{k,t} \geq \gamma_{k,min} N_{k,t}$ for all k and t.

Proof of Theorem 4.4.1: We prove the theorem by decomposing regret into a function of the number of plays of suboptimal arms, up to and after some sufficient sampling level. We then introduce two propositions which give bounds for quantities in the decomposition which are then combined to give the bound in (4.4.4). The proofs of these propositions are reserved for Section 4.7.3.

Let $N_{k,t}^{l,suf}, N_{k,t}^{l,und}$ for $l \in [B_k]$ be counters associated with elements of $\mathcal{G}_{k,B}$ for $k \in [K]$. These counters are defined as follows:

$$N_{k,n}^{l,suf} = \sum_{t=K+1}^{n} \mathbb{I}\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{k,t-1} > h_{k,n}(\Delta^{k,l})\}, \tag{4.4.6}$$

$$N_{k,n}^{l,und} = \sum_{t=K+1}^{n} \mathbb{I}\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{k,t-1} \leq h_{k,n}(\Delta^{k,l})\}, \tag{4.4.7}$$

where $h_{k,n}(\Delta) = 12b(\Delta)\frac{\log(n)K^2}{\gamma_{k,min}\Delta^2}$. A cell k is said to be *sufficiently* sampled with respect to a choice of detection probabilities $\mathrm{g}^{l}_{k,B}$ if $N_{k,t-1} > h_{k,n}(\Delta^{k,l})$, and thus $N^{l,und}_{k,n}$, $N^{l,suf}_{k,n}$ count the suboptimal plays leading to incrementing $N^{l}_{k,n}$ up to and after the sufficient level, respectively.

From the definitions (4.4.6) and (4.4.7) we have $N_{k,n} = 1+\sum_{l=1}^{B_k}(N^{l,suf}_{k,n}+N^{und}_{k,n})$. The expected regret at time horizon n can also be bounded above using this notation as

$$Reg_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) \leq \mathbb{E}\left[\sum_{k=1}^{K}\left(\Delta^{k,1} + \sum_{l=1}^{B_k}(N^{l,suf}_{k,n} + N^{l,und}_{k,n}) \cdot \Delta^{k,l}\right)\right] \tag{4.4.8}$$

where $\Delta^{k,1}$ arises as a worst case view of the initialisation. We can derive an analytical bound on regret by bounding the expectations of the random variables in (4.4.8).

Firstly, for the beyond sufficiency counter we have

Proposition 4.4.2. *For any time horizon $n > K$,*

$$\mathbb{E}\left(\sum_{k=1}^{K}\sum_{l=1}^{B_k} N^{l,suf}_{k,n}\right) \leq \frac{\pi^2}{3} \cdot K. \tag{4.4.9}$$

The full proof of Proposition 4.4.2 is given in Section 4.7.3, but it depends in particular on the following Lemma describing the concentration of filtered Poisson data. The derivation of the concentration result for the observations $Y_1, ..., Y_t$ requires careful treatment as the parameters of these distributions, and therefore the observations themselves, are not independent. The stochastic dependencies between the sequence of random variables $\gamma_1, ..., \gamma_s$ may be highly complex, so rather than attempt to quantify these relationships exactly, we appeal to martingale theory which allows us to derive the concentration result without assuming independence. We provide the necessary concentration result in the lemma below.

Lemma 4.4.3. *Let $Y_1, ..., Y_s$ be any sequence of Poisson random variables with means $\gamma_1\lambda, ...\gamma_s\lambda$ respectively, such that the sequence $\{Z_j\}_{j=1}^{s}$ is a martingale where $Z_j = \sum_{i=1}^{j}(Y_i - \mathbb{E}(Y_i|Y_{i-1}, ..., Y_1))$.*

Then, given parameters $t \geq s$ and $\lambda_{max} \geq \lambda$ the following holds:

$$\mathbb{P}\left(\left|\frac{\sum_{j=1}^{s} Y_j}{\sum_{j=1}^{s}\gamma_j} - \lambda\right| \geq \frac{6\max(1,\sqrt{\lambda_{max}})\log(t)}{\sum_{j=1}^{s}\gamma_j} + \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{j=1}^{s}\gamma_j}}\right) \leq 2t^{-3}. \tag{4.4.10}$$

The proof of this Lemma is given in Section 4.7.2. The consequence of this Lemma is that the UCB indices (4.4.3) are of the correct form to guarantee that the probability of making suboptimal plays beyond the sufficient sampling level is small.

For the under sufficiency counter we have the following proposition, also proved in Section 4.7.3,

Proposition 4.4.4. *For any time horizon $n > K$ and $k : \Delta^k_{min} > 0$,*

$$\sum_{l=1}^{B_k} N^{l,und}_{k,n}\Delta^{k,l} \leq h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \int_{\Delta^{k,B_k}}^{\Delta^{k,1}} h_{k,n}(x)dx. \tag{4.4.11}$$

Combining the decomposition (4.4.8), with the bounds (4.4.9) and (4.4.11) we have

$$\begin{aligned}
Reg_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) &\leq \mathbb{E}\bigg(\sum_{k=1}^{K}\Big(\Delta^{k,1} + \sum_{l=1}^{B_k}(N^{l,suf}_{k,n} + N^{l,und}_{k,n})\Delta^{k,l}\Big)\bigg) \\
&= \mathbb{E}\bigg(\sum_{k=1}^{K}\Big(\Delta^{k,1} + \sum_{l=1}^{B_k}N^{l,suf}_{k,n}\Delta^{k,l}\Big)\bigg) + \mathbb{E}\bigg(\sum_{k=1}^{K}\sum_{l=1}^{B_k}N^{l,und}_{k,n}\Delta^{k,l}\bigg) \\
&\leq K\Delta_{max} + \mathbb{E}\bigg(\sum_{k=1}^{K}\sum_{l=1}^{B_k}N^{l,suf}_{k,n}\Delta^{k,l}\bigg) \\
&\qquad + \sum_{k:\Delta^k_{min}>0}\bigg(h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \int_{\Delta^{k,B_k}}^{\Delta^{k,1}} h_{k,n}(x)dx\bigg) \\
&\leq \Big(\frac{\pi^2}{3} + 1\Big)K\Delta_{max} + \sum_{k:\Delta^k_{min}>0}\bigg(h_{k,n}(\Delta^k_{min})\Delta^k_{min} + \int_{\Delta^k_{min}}^{\Delta^k_{max}} h_{k,n}(x)dx\bigg)
\end{aligned}$$

$$= \sum_{k:\Delta^k_{min}>0} \frac{12K^2}{\gamma_{k,min}} \left[\frac{b(\Delta^k_{min})}{\Delta^k_{min}} + \int_{\Delta^k_{min}}^{\Delta^k_{max}} \frac{b(x)}{x^2} dx \right] \log(n) + \left(\frac{\pi^2}{3} + 1 \right) K\Delta_{max}. \quad \square$$

In the remainder of this section we show that the bound obtained in Theorem 4.4.1 is of optimal order, by deriving a lower bound on the expected regret of the best possible policies. We also proceed to show a second upper bound of sub-optimal order with respect to n but that has the advantage of holding for any problem instance, and therefore does not depend on the optimality gaps, Δ^k_{min} and Δ^k_{max}, $\forall k \in [K]$.

Lower Bound on Regret

To analyse the performance of the best possible policies, we introduce the notion of a *uniformly good policy*. A uniformly good policy (Lai and Robbins, 1985) is one where

$$\mathbb{E}\left(\sum_{t=1}^{n} \mathbb{I}\{\mathbf{g}_t = \mathbf{g}\} \right) = o(n^\alpha) \quad \forall \ \alpha > 0$$

for every g : $\Delta_{\mathbf{g}} > 0$ and every $\boldsymbol{\lambda} \in \mathbb{R}^K_+$. Clearly, then all uniformly good policies must eventually favour optimal actions over suboptimal ones - with the suboptimal actions being necessary to accurately estimate $\boldsymbol{\lambda}$. For a given rate vector $\boldsymbol{\lambda}$ we define the set of optimal actions as

$$J(\boldsymbol{\lambda}) = \{\mathbf{g} \in \mathcal{G} : \ \ \mathbf{g}^T \cdot \boldsymbol{\lambda} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}\}.$$

We write $S(\boldsymbol{\lambda}) = \mathcal{G} \setminus J(\boldsymbol{\lambda})$ to be the set of suboptimal actions. The difficulty of a particular problem depends on the particular combination of $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$. We define

$$\mathcal{I}(\boldsymbol{\lambda}) = \{k : \ \exists\, \mathbf{g} \in J(\boldsymbol{\lambda}) \text{ s.t. } g_k > 0\}$$

as the set of arms which are played in at least one optimal action and

$$B(\boldsymbol{\lambda}) = \{\boldsymbol{\theta} \in \mathbb{R}_+^K : \ \ \mathbf{g}^T \cdot \boldsymbol{\theta} < \text{opt}_{\boldsymbol{\theta},\gamma} \ \ \forall \mathbf{g} \in J(\boldsymbol{\lambda}) \ \ \text{and} \ \ \theta_k = \lambda_k \ \ \forall k \in \mathcal{I}(\boldsymbol{\lambda})\}$$

as the set of mean vectors such that all actions in $J(\boldsymbol{\lambda})$ are suboptimal but this cannot be discerned by playing only actions in $J(\boldsymbol{\lambda})$. The larger the set $B(\boldsymbol{\lambda})$, the more challenging the problem is. If $\mathcal{I}(\boldsymbol{\lambda}) = [K]$, then one can simultaneously play optimal actions and gather the information necessary to affirm that these actions are optimal. In such a case the lower bound on expected regret is simply 0.

We have the following lower bound on regret for any uniformly good policy. A key consequence of this result is the assertion that policies with $O(\log(n))$ regret are indeed of optimal order and thus that the regret induced by the FP-CUCB algorithm in case (I) grows at the lowest achievable rate. This result is analogous to results in other classes of bandit problem as shown by Lai and Robbins (1985) and Burnetas and Katehakis (1996).

Theorem 4.4.5. *For any $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ such that $B(\boldsymbol{\lambda}) \neq \emptyset$, and for any uniformly good policy π for the sequential surveillance problem with known $\boldsymbol{\gamma}$, we have*

$$\liminf_{n\to\infty} \frac{Reg^{\pi}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n)}{\log(n)} \geq c(\boldsymbol{\lambda}) \tag{4.4.12}$$

where $c(\boldsymbol{\lambda})$ is the optimal value of the following optimisation problem

$$\inf_{\mathbf{d}\geq\mathbf{0}} \sum_{\mathbf{g}\in S(\boldsymbol{\lambda})} d_{\mathbf{g}} \Delta_{\mathbf{g}} \tag{4.4.13}$$

$$\textit{such that} \ \inf_{\boldsymbol{\theta}\in B(\boldsymbol{\lambda})} \sum_{\mathbf{g}\in S(\boldsymbol{\lambda})} d_{\mathbf{g}} \sum_{k=1}^{K} g_k kl(\lambda_k, \theta_k) \geq 1. \tag{4.4.14}$$

and $kl(\lambda, \theta) = \lambda \log(\frac{\lambda}{\theta}) + \theta - \lambda$ is the Kullback Leibler divergence between two Poisson distributions *with mean parameters λ, θ respectively.*

We prove this theorem fully in Section 4.7.4, but here note that a key step of its proof is to invoke Theorem 1 of Graves and Lai (1997), which is a similar result for a more general class of controlled Markov Chains. It is possible to derive an analytical expression giving a lower bound on $c(\boldsymbol{\lambda})$ by following steps similar to those in the proof of Theorem 2 in Combes et al. (2015). However we omit this here in the interests of succinctness as it is not an especially useful or elegant expression.

Gap-free upper bound on regret

The logarithmic order bounds are useful as they establish the order-optimality of the UCB algorithm. However, the coefficients may be very large in problem instances where the Δ^k_{min} terms are very small. Additionally, in absence of knowledge of the $\Delta_{k,min}$ terms they do little to inform one of expected performance of the algorithm. For these reasons, we also present the following upper bound on regret, which is order-suboptimal, being of order $O(K\sqrt{n\log(n)})$, but holds uniformly across any choice of $\boldsymbol{\lambda} \in [0, \lambda_{max}]^K$ and does not depend on the optimality gaps.

Theorem 4.4.6. *The regret of the FP-CUCB algorithm with λ_{max} applied to the sequential surveillance problem with known $\boldsymbol{\gamma}$ satisfies*

$$Reg^{FP-CUCB}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) \leq \sqrt{\frac{92K^2\lambda_{max}}{\gamma_{min}} n\log(n)} + \frac{24K\sqrt{\lambda_{max,1}}}{\gamma_{min}} \log^2(n) + \frac{5K\lambda_{max}}{2} \tag{4.4.15}$$

where $\lambda_{max,1} = \max(1, \lambda_{max})$.

We note that although the dependence on the optimality gaps may be removed, the dependence on the minimal detection probability remains. This has an important relationship with performance in our problem because it controls the information gained per action.

Proof of Theorem 4.4.6: We first consider the following decomposition of regret, which works by adding and subtracting the 'reward' with respect to inflated rates $\bar{\boldsymbol{\lambda}}_t$ of the action $\mathbf{a}_t$ selected in round t, and then using that by definition $r_{\bar{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a}_t) \geq r_{\bar{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a}^*)$,

$$\begin{aligned}
Reg^{FP-CUCB}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) &= \mathbb{E}\bigg(\sum_{t=1}^{n} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}^*) - r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}_t)\bigg) \\
&= \mathbb{E}\bigg(\sum_{t=1}^{n} r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}^*) - r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}_t) + r_{\bar{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a}_t) - r_{\bar{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a}_t)\bigg) \\
&= \mathbb{E}\bigg(\sum_{t=1}^{n}\sum_{k=1}^{K} \lambda_k g_k^* - \bar{\lambda}_{kt} g_{kt} + \bar{\lambda}_{kt} g_{kt} - \lambda_k g_{kt}\bigg) \\
&\leq \mathbb{E}\bigg(\sum_{t=1}^{n}\sum_{k=1}^{K} (\lambda_k - \bar{\lambda}_{kt}) g_k^* + (\bar{\lambda}_{kt} - \lambda_k) g_{kt}\bigg) \\
&= \sum_{t=1}^{n}\sum_{k=1}^{K} \mathbb{E}\bigg((\lambda_k - \bar{\lambda}_{kt})\bigg) g_k^* + \sum_{t=1}^{n}\sum_{k=1}^{K} \mathbb{E}\bigg((\bar{\lambda}_{kt} - \lambda_k) g_{kt}\bigg)
\end{aligned} \tag{4.4.16}$$

The terms of the first sum in (4.4.16) are very unlikely to be positive, increasingly so as more data is collected. If we upper bound by ignoring the case of negative terms we have:

$$\begin{aligned}
&\sum_{t=1}^{n}\sum_{k=1}^{K} \mathbb{E}\bigg((\lambda_k - \bar{\lambda}_{kt})\bigg) g_k^* \\
&\leq \sum_{t=1}^{n}\sum_{k=1}^{K} g_k^* \mathbb{P}(\lambda_k > \bar{\lambda}_{kt}) \mathbb{E}(\lambda_k - \bar{\lambda}_{kt} | \lambda_k > \bar{\lambda}_{kt}) \\
&\leq \sum_{t=1}^{n} \lambda_{max} \sum_{k=1}^{K} \mathbb{P}(\lambda_k > \bar{\lambda}_{kt}) \\
&= K\lambda_{max} \sum_{t=1}^{n} \mathbb{P}\bigg(\sum_{k=1}^{K} \lambda_k > \sum_{k=1}^{K} \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{6\max(1, \sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}}\bigg) \\
&\leq K\lambda_{max} \sum_{t=1}^{n} t^{-3} \leq \frac{5K\lambda_{max}}{4}
\end{aligned} \tag{4.4.17}$$

where the second inequality holds since $g_K^*\mathbb{E}(\lambda_k - \bar{\lambda}_{kt}) \leq \lambda_{max}\ \forall k$, and where the penultimate inequality is due to an application of Lemma 4.4.3.

Now consider the second sum in (4.4.16)

$$\begin{aligned}
&\sum_{t=1}^{n}\sum_{k=1}^{K}\mathbb{E}\Big((\bar{\lambda}_{kt}-\lambda_k)g_{kt}\Big)\\
&=\sum_{t=1}^{n}\mathbb{E}\Big(\sum_{k=1}^{K}\gamma_{k,t}\Big(\frac{\sum_{s=1}^{t-1}Y_{k,s}}{\sum_{s=1}^{t-1}\gamma_{k,s}}+\frac{6\max(1,\sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}+\sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}}-\lambda_k\Big)\Big)\\
&\leq\sum_{t=1}^{n}K\lambda_{max}\mathbb{P}\Big(\sum_{k=1}^{K}\frac{\sum_{s=1}^{t-1}Y_{k,s}}{\sum_{s=1}^{t-1}\gamma_{k,s}}-\lambda_k>\sum_{k=1}^{K}\frac{6\max(1,\sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}+\sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Big)\\
&\quad+\sum_{t=1}^{n}\mathbb{E}\Big(\sum_{k=1}^{K}2\gamma_{k,t}\Big(\frac{6\max(1,\sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}+\sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Big)\Big)\\
&\leq\frac{5K\lambda_{max}}{4}+\mathbb{E}\Big(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{12\gamma_{k,t}\max(1,\sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}\Big)+\mathbb{E}\Big(\sum_{t=1}^{n}\sum_{k=1}^{K}\gamma_{k,t}\sqrt{\frac{24\lambda_{max}\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Big)\\
&\leq\frac{5K\lambda_{max}}{4}+12\max(1,\sqrt{\lambda_{max}})\log(n)\mathbb{E}\Big(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{\gamma_{k,t}}{\sum_{s=1}^{t-1}\gamma_{k,s}}\Big)\\
&\quad+\sqrt{24\lambda_{max}\log(n)}\mathbb{E}\Big(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{\gamma_{k,t}}{\sqrt{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Big).
\end{aligned}\tag{4.4.18}$$

Here, we have again used the property that $\gamma_{k,t}(\bar{\lambda}_{kt}-\lambda_k)$ is bounded for all k,t and applied Lemma 4.4.3 in the penultimate inequality. The two expectations remaining in (4.4.18) can be bounded in terms of n. Considering the expectation in the final term, we have,

$$\mathbb{E}\Big(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{\gamma_{k,t}}{\sqrt{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Big)\leq\sum_{k=1}^{K}\sum_{t=2}^{n}\frac{1}{\sqrt{(t-1)\gamma_{min}}}\leq\frac{2K}{\sqrt{\gamma_{min}}}\sqrt{n}.\tag{4.4.19}$$

Similarly for the other expectation, we have

$$\mathbb{E}\bigg(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{\gamma_{k,t}}{\sum_{s=1}^{t-1}\gamma_{k,s}}\bigg) \leq \sum_{k=1}^{K}\sum_{t=2}^{n}\frac{1}{(t-1)\gamma_{min}} \leq \frac{K}{\gamma_{min}}(1+\log(n)). \tag{4.4.20}$$

Finally, combining (4.4.17), (4.4.18), (4.4.19), and (4.4.20) we have the following gap-free bound on regret:

$$\begin{aligned} Reg_{\boldsymbol{\lambda},\boldsymbol{\gamma}}^{FP-CUCB}(n) \leq \frac{5K\lambda_{max}}{2} + \frac{12K\max(1,\sqrt{\lambda_{max}})}{\gamma_{min}}\log(n)(1+\log(n)) \\ + \sqrt{\frac{92K^2\lambda_{max}\log(n)n}{\gamma_{min}}}, \end{aligned}$$

as stated in Theorem 4.4.6. □

4.4.2 Case (II): Known scaling of detection probabilities

In the second case we suppose that we do not know exactly what probability of successful detection each searcher has in each cell, but that we have some idea of how these detection probabilities change as the searchers are assigned more cells to search. If, for example, the searcher is moving back-and-forth over l cells at a constant speed s, then the time between successive visits to a cell is $2l/s$, suggesting that the detection probability may decay like $s/(2l)$ with the number of cells l.

In order to be precise about this case we suppose that detection probabilities have the form

$$\gamma_k(\mathbf{a}) = \sum_{u=1}^{U}\phi_u(\mathbf{a})\omega_{ku}\mathbb{I}\{a_k = u\}, \;\; k \in [K], \tag{4.4.21}$$

where $\phi_u : \mathcal{A} \to [0,1]$ are known *scaling functions*, and $\omega_{ku} \in (0,1]$ $\forall k \in [K], u \in [U]$ are unknown *baseline detection probabilities* - the probability of searcher u detecting events in cell k given that it is the only cell they are assigned to search. Functions ϕ_u are assumed to be decreasing in the number of cells searcher u must search. For instance, and as suggested in the preceding paragraph, one suitable function may be $\phi_u(\mathbf{a}) = (\sum_{k=1}^{K} \mathbb{I}\{a_k = u\})^{-1}$, the reciprocal of the number of cells the searcher u is assigned. Searcher effectiveness may however decay more slowly as the number of cells assigned grows if for instance events are visible for an extended period of time.

In case (II) the action set and observed rewards remain entirely the same as for case (I), it is the information initially available to the controller that differs. Here, both $\boldsymbol{\lambda}$, the K-vector of rate parameters, and $\boldsymbol{\omega} = (\omega_{1,1}, ..., \omega_{1,U}, \omega_{2,1}..., \omega_{K,U})$, the KU-vector of baseline detection probabilities are unknown as opposed to solely $\boldsymbol{\lambda}$ in case (I). Due to nonidentifiability we cannot make direct inference on $\boldsymbol{\lambda}$ or $\boldsymbol{\omega}$. However, simply estimating the products of certain components is sufficient for optimal decision making as estimating the expected reward does not depend on having separate estimates of each parameter. Therefore we can simply consider KU unknowns $\boldsymbol{\tau} = (\omega_{1,1}\lambda_1, ..., \omega_{1,U}\lambda_1, \omega_{2,1}\lambda_2, ..., \omega_{K,U}\lambda_K)$ when referring to the unknown parameters. For $k \in [K], u \in [U]$ and $s \in [T]$, $\phi_{ku,s}$ will refer to the detection probability applied by searcher u to cell k in round s.

As such this second case of the sequential search problem can also be modelled as a CMAB problem with filtered feedback. The set of arms is given by searcher-cell pairs $ku \in [K] \times [U]$. Each arm ku is associated with a Poisson distribution with unknown parameter $\tau_{ku} = \omega_{k,u}\lambda_k$. We continue to use $\mathcal{A}$ to specify the action set and filtering is governed by scaling function vectors $\boldsymbol{\phi}(\mathbf{a}) = (\phi_1(\mathbf{a}), ..., \phi_U(\mathbf{a}))$. Let $\phi_{ku,t}$ denote the filtering probability associated with the searcher-cell pair ku in round t. It is 0 if $a_{k,t} \neq u$ and $\phi_u(\mathbf{a}_t)$ if $a_{k,t} = u$.

Let reward in this setting be defined

$$r_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{a}) = \tilde{r}_{\boldsymbol{\tau},\boldsymbol{\phi}}(\mathbf{a}) = \sum_{u=1}^{U} \phi_u(\mathbf{a}) \sum_{k=1}^{K} \tau_{ku}\mathbb{I}\{a_k = u\}$$

and define optimality gaps in this setting for $ku \in [K] \times [U]$ as

$$\Delta^{ku}_{max} = \text{opt}_{\boldsymbol{\lambda},\gamma} - \min_{\mathbf{a}\in\mathcal{A}}\{r_{\boldsymbol{\lambda},\gamma}(\mathbf{a}) \mid r_{\boldsymbol{\lambda},\gamma}(\mathbf{a}) \neq \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}, a_k = u\}$$
$$\Delta^{ku}_{min} = \text{opt}_{\boldsymbol{\lambda},\gamma} - \max_{\mathbf{a}\in\mathcal{A}}\{r_{\boldsymbol{\lambda},\gamma}(\mathbf{a}) \mid r_{\boldsymbol{\lambda},\gamma}(\mathbf{a}) \neq \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}, a_k = u\}.$$

The appropriate FP-CUCB algorithm for case (II) then calculates upper confidence bounds for each τ_{ku} parameter instead of λ_k and as in the FP-CUCB algorithm for case (I) this induces an optimistic estimate of the value of every $\mathbf{a} \in \mathcal{A}$. We describe this second variant in Algorithm 6.

```
Algorithm 6: FP-CUCB (case (II))

    Inputs: Upper bound $\tau_{max} \geq \tau_{ku},\ \ k \in [K]$ and $u \in [U]$.
    Initialisation Phase: For $t \in [KU]$
        • Select an arbitrary allocation $\mathbf{a} \in \mathcal{A}$ such that $a_t \neq 0$
    Iterative Phase: For $t = KU + 1, KU + 2, ...$
        • Calculate indices
              $\bar{\tau}_{ku,t} = \frac{\sum_{s=1}^{t-1} Y_{ku,s}}{\sum_{s=1}^{t-1} \phi_{ku,s}} + \frac{6\max(1,\sqrt{\tau_{max}})\log(t)}{\sum_{s=1}^{t-1} \phi_{ku,s}} + \sqrt{\frac{6\tau_{max}\log(t)}{\sum_{s=1}^{t-1} \phi_{ku,s}}},\ \ ku \in [K] \times [U]$
        • Select an allocation $\mathbf{a}^*_{\bar{\boldsymbol{\lambda}}_t}$ such that $\tilde{r}_{\bar{\boldsymbol{\tau}}_t,\boldsymbol{\phi}}(\mathbf{a}^*_{\bar{\boldsymbol{\lambda}}_t}) = \max_{\mathbf{a}\in\mathcal{A}} \tilde{r}_{\bar{\boldsymbol{\tau}}_t,\boldsymbol{\phi}}(\mathbf{a})$.
```

Since our CMAB model in case (II) and second variant of FP-CUCB are of the same form as in case (I), the analogous results to Theorems 4.4.1 and 4.4.5 can be derived. Specifically we have a regret upper bound for FP-CUCB in Corollary 4.4.7 and a lower bound for regret of any uniformly good algorithm in Corollary 4.4.8.

Corollary 4.4.7. *The regret of the FP-CUCB algorithm in case (b) defined by τ_{max} applied to the* *sequential search problem as defined previously satisfies*

$$Reg^{\textit{FP-CUCB}}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) \leq \sum_{ku:\Delta^{ku}_{min}>0} \frac{24(KU)^2\tau_{max,1}c_{1,ku}}{\phi_{ku,min}\Delta^{ku}_{min}} \log(n) + 5KU\Delta_{max},$$

where $\tau_{max,1} = \max(1, \tau_{max})$, $\phi_{ku,min} = \min_{\mathbf{a}:a_k=u} \phi_u(\mathbf{a})$, and $c_{1,k}$ are known constants depending on K, U, Δ^{ku}_{max} and Δ^{ku}_{min}.

Corollary 4.4.8. *For any $\boldsymbol{\tau} \in \mathbb{R}^{KU}_{+}$ such that $\tilde{B}(\boldsymbol{\tau}) \neq \emptyset$, and for any uniformly good policy π for the sequential surveillance problem with known $\boldsymbol{\phi}$, we have*

$$\lim \inf_{n\to\infty} \frac{Reg^{\pi}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n)}{\log(n)} \geq \tilde{c}(\boldsymbol{\tau})$$

where $\tilde{c}(\boldsymbol{\tau})$ is the solution of an optimisation problem analogous to (4.4.13).

Precise specification of $\tilde{c}(\boldsymbol{\tau})$ requires redefining notation from Section 4.4.1 in the context of case (II) and produces an entirely unsurprising analogue. In the interests of brevity we omit this. The techniques used in proving Theorems 4.4.1 and 4.4.5 can be easily extended to prove Corollaries 1 and 2.

4.5 Numerical Experiments

We now numerically evaluate the performance of the FP-CUCB algorithm in comparison to a greedy approach and Thompson Sampling (TS). The greedy approach is one which always selects the action currently believed to be best (following an initialisation period, where each cell is searched at least once). As such it is a fully exploitative policy which fails to recognise the benefit of the information gain inherent in exploration. TS is a randomised, Bayesian approach where an action is selected with the current posterior probability that it is the best one. This is achieved by sampling indices from a posterior distribution on each arm and passing these samples to the optimisation algorithm. We define these algorithms in the setting of known detection probabilities (case (I)) in Algorithms 7 and 8 respectively.

```
Algorithm 7: Greedy

  Initialisation Phase: For $t \in [K]$
    • Select an arbitrary allocation $\mathbf{a} \in \mathcal{A}$ such that $a_t \neq 0$
  Iterative Phase: For $t = K+1, K+2, ...$
    • For each $k \in [K]$ calculate $\hat{\lambda}_{k,t} = \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}}$
    • Select an allocation $\mathbf{a}^*_{\hat{\boldsymbol{\lambda}}_t}$ such that $r_{\hat{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a}^*_{\hat{\boldsymbol{\lambda}}_t}) = \max_{\mathbf{a} \in \mathcal{A}} r_{\hat{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a})$.
```

```
Algorithm 8: Thompson Sampling (TS)

  Inputs: Gamma prior parameters $\alpha, \beta$
  Iterative Phase: For $t = 1, 2, ...$
    • For each $k \in [K]$ sample $\tilde{\lambda}_{k,t}$ from a Gamma($\alpha + \sum_{s=1}^{t-1} Y_{k,s}, \beta + \sum_{s=1}^{t-1} \gamma_k(\mathbf{a}_s)$).
    • Select an allocation $\mathbf{a}^*_{\tilde{\boldsymbol{\lambda}}_t}$ such that $r_{\tilde{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a}^*_{\tilde{\boldsymbol{\lambda}}_t}) = \max_{\mathbf{a} \in \mathcal{A}} r_{\tilde{\boldsymbol{\lambda}}_t,\boldsymbol{\gamma}}(\mathbf{a})$.
```

We compare the FP-CUCB, Greedy and TS algorithms by randomly sampling $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ values which define problem instances. We then test our algorithms' performance on data generated from the models of these problem instances. We assume that detection probabilities have the form given in (4.4.21) but we know both the ϕ functions and ω values.

Specifically, we conduct four tests encompassing a range of different problem sizes and parameter values to display the efficacy of our proposed approach uniformly across problem instances. In each test 50 $(\boldsymbol{\lambda}, \boldsymbol{\omega})$ pairs are sampled and functions $\boldsymbol{\phi}$ are selected. For each $(\boldsymbol{\lambda}, \boldsymbol{\omega})$ pair 5 datasets are sampled giving underlying counts of intrusion events in each cell in each round up to a horizon of $n = 2000$. This gives us 250 simulations for each test framework. Parameters are simulated as below:

(i) $K = 15$ cells and $U = 5$ searchers. Cell means λ_k are sampled from a Uniform$(10, 20)$ distribution for $k \in [K]$. Baseline detection probabilities ω_{ku} are sampled from Beta$(u, 2)$ distributions for $u \in [U]$, $k \in [K]$. Scaling functions are $\phi_u(\mathbf{a}) = (\sum_{k=1}^{K} \mathbb{I}\{a_k = u\})^{-1}$ for $u \in [U]$, $\mathbf{a} \in \mathcal{A}$.

(ii) $K = 50$ cells and $U = 3$ searchers. Cell means λ_k are sampled from Uniform distributions on the intervals $[k, k+10]$ for $k = 1, ..., 10$, $[20-k, 30-k]$ for $k = 11, ..., 20$, $[k-20, k-10]$ for $k = 21, ..., 30$, $[40-k, 50-k]$ for $k = 31, ..., 40$, and $[k-40, k-30]$ for $k = 41, ..., 50$. Baseline detection probabilities ω_{ku} are sampled from Beta$(u+2, 2)$ distributions for $u \in [U]$, $k \in [K]$. Scaling functions are $\phi_u(\mathbf{a}) = (0.5 + 0.5 \sum_{k=1}^{K} \mathbb{I}\{a_k = u\})^{-1}$ for $u \in [U]$, $\mathbf{a} \in \mathcal{A}$.

(iii) $K = 25$ cells and $U = 10$ searchers. Cell means λ_k are sampled from a Uniform$(90, 100)$ distribution for $k \in [K]$. Baseline detection probabilities ω_{ku} are sampled from a Beta$(30, 5)$ distribution for $u \in [U]$, $k \in [K]$. Scaling functions are $\phi_u(\mathbf{a}) = (\sum_{k=1}^{K} \mathbb{I}\{a_k = u\})^{-1}$ for $u \in [U]$, $\mathbf{a} \in \mathcal{A}$.

(iv) $K = 25$ cells and $U = 5$ searchers. Cell means λ_k are sampled from a Uniform$(0.4, 1)$ distribution for $k \in [K]$. Baseline detection probabilities ω_{ku} are sampled from a Beta$(1, 1)$ distribution for $u \in [U]$, $k \in [K]$. Scaling functions are $\phi_u(\mathbf{a}) = (0.5 + 0.5 \sum_{k=1}^{K} \mathbb{I}\{a_k = u\})^{-1}$ for $u \in [U]$, $\mathbf{a} \in \mathcal{A}$.

We test a variety of parametrisations of FP-CUCB (in terms of λ_{max}) and TS (in terms of the prior mean and variance - from which particular α and β values can be uniquely found) in each test. In each case we use λ_{max} values which are both larger and smaller than the true maximal rate. Similarly we investigate TS with prior mean larger and smaller than the true maximal rate and with several different levels of variance. It is not always fully realistic to assume knowledge of λ_{max}

will be perfect and therefore it is of interest to investigate the effects of varying it. Also, the choice of priors in TS is a potentially subjective one and it is important to understand its impact.

We measure the performance of our algorithms by calculating the expected regret incurred by their actions, rescaled by the expected reward of a single optimal action. For an algorithm A and particular history $\mathbf{H}_n$ we write

$$ScaleReg^{A}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{H}_n) = \frac{\sum_{t=1}^{n} \Delta_{\mathbf{a}_t}}{\text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}}.$$

We calculate this value for all algorithms, all 250 datasets and rounds $1 \leq n \leq 2000$. We choose to rescale our regret to make a fairer comparison across the 50 different problem instances in each test (i)-(iv) which will all have different optimal expected rewards.

In Figure 4.5.1 we illustrate how regret evolves over time by plotting the median scaled regret across the 250 runs of each algorithm in all rounds of test (i). The rate of growth shown in these plots is typical of the results in the other three tests. An immediate observation is that the greedy algorithm does very poorly on average and its full median regret over the 2000 rounds cannot be included in the graphs without obscuring differences between the other algorithms. We see also that the performance of both FP-CUCB and TS is strongly linked to the chosen parameters. For the FP-CUCB algorithm it seems in Figure 1 that the larger the parameter λ_{max} is the larger the cumulative regret becomes. For TS, larger prior variances seem to induce lower regret, the relationship with the prior mean is more complex. Accurate specification of the prior mean seems to ensure good performance, but underestimation and overestimation of the mean can lead to poor performance (particularly when the variance is small).

We analyse these behaviours further in Figures 4.5.2 and 4.5.3. Here we calculate a scaled regret at time $n = 2000$ for all 250 runs of each algorithm and plot the empirical distribution of these values for each parameterisation of each algorithm. The results for tests (i) and (ii) are given in Figure 4.5.2 and for tests (iii) and (iv) in Figure 4.5.3. We omit the greedy algorithm's performance from these figures as the values are so large. In Section 4.7.5 we provide median values and lower and upper quantiles of the scaled regret for each algorithm. We see from these values that the greedy algorithm performs substantially worse than the FP-CUCB and TS algorithms which better address the exploration-exploitation dilemma.

Examining Figures 2 and 3 we see that the FP-CUCB algorithm enjoys greater robustness to parameter choice than the TS approach. In particular in the results of test (iii) we see that many parametrisations of TS give rise to a long tailed distribution of round 2000 regret - meaning the performance of TS is highly variable and often poor. This variability of performance does seem to coincide with underestimation of the mean, however FP-CUCB manages to maintain strong performance even when the λ_{max} parameter is far from the true maximal rate. When the prior variance is sufficiently large and the prior mean is close to the true λ_{max} TS seems to do the best job of balancing exploration and exploitation and incurs the smallest regret.

4.6 Discussion

In this chapter we have considered the problem of adaptively assigning multiple searchers to cells along a line (in space or time) in order to detect the maximum number of events occurring along the line. We have modelled the problem, and proposed and analysed solution methods. The challenge at the heart of this problem is to correctly balance exploration and exploitation, in the face of initial ignorance as to the arrival process of events.

We formulated our sequential decision problem as a combinatorial multi-armed bandit with Poisson rewards and a novel filtered feedback mechanism. To design quality policies for this problem we first derived an efficient solution method to the full information problem. This IP forms the backbone of all policies for the sequential problem, as it allows us to quickly identify an optimal solution given some estimate of the arrival process' rate parameters.

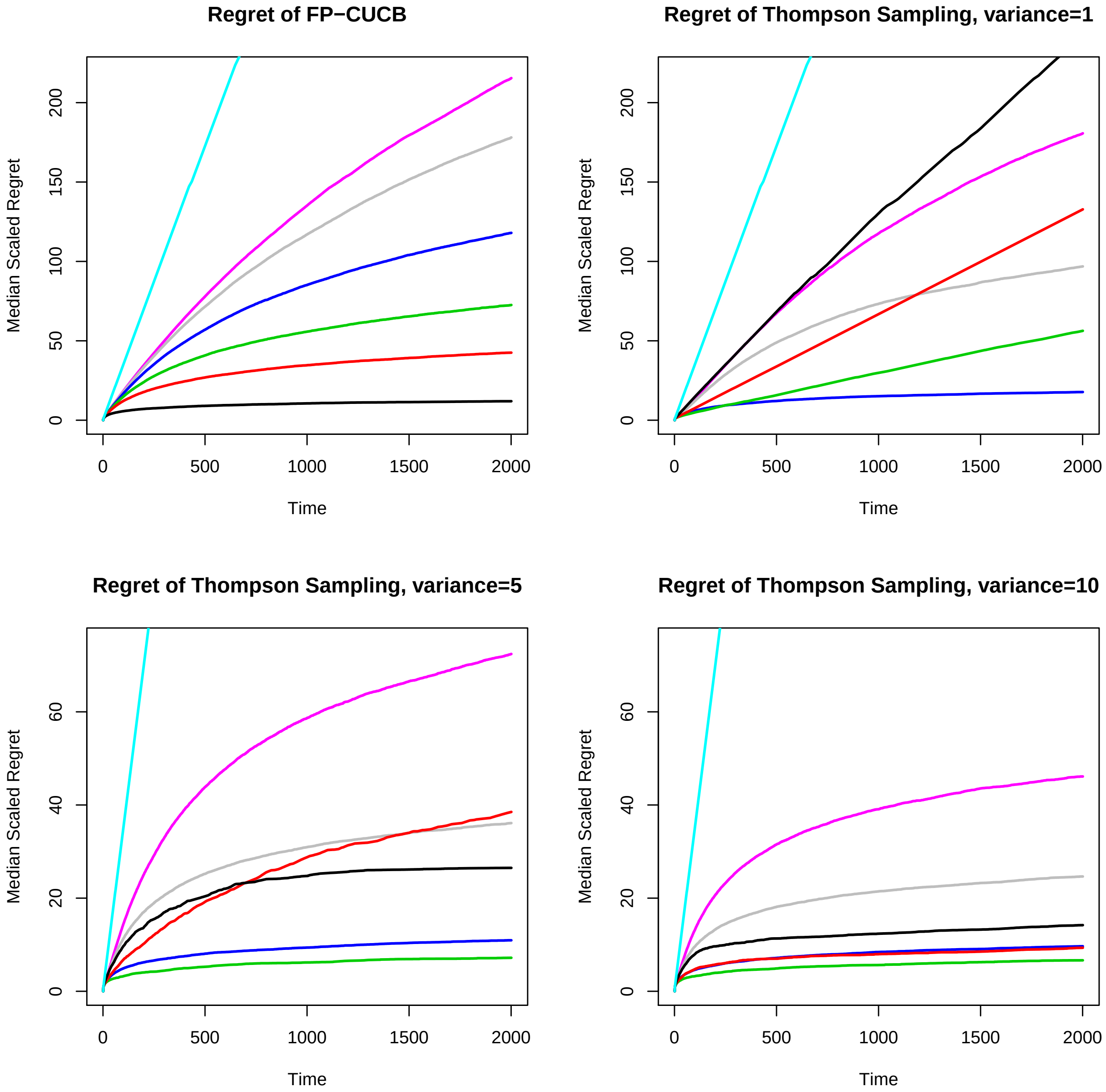


Figure 4.5.1: Cumulative Regret histories for Test (i). Upper left: FP-CUCB, upper right: TS with a prior variance of 1, lower left: TS with a prior variance of 5, lower right: TS with prior variance of 10. In each case the plotted lines are the median values of scaled regret calculated at each time point from 1 to 2000. Black lines represent $\lambda_{max} = 1$ or a prior mean of 1, red represents the same parameters taking the value 5, green 10, blue 20, grey 40, and pink 60. In all sub-figures the teal line represents regret of the greedy algorithm. Note that the vertical axis scales differ between the top and bottom rows.

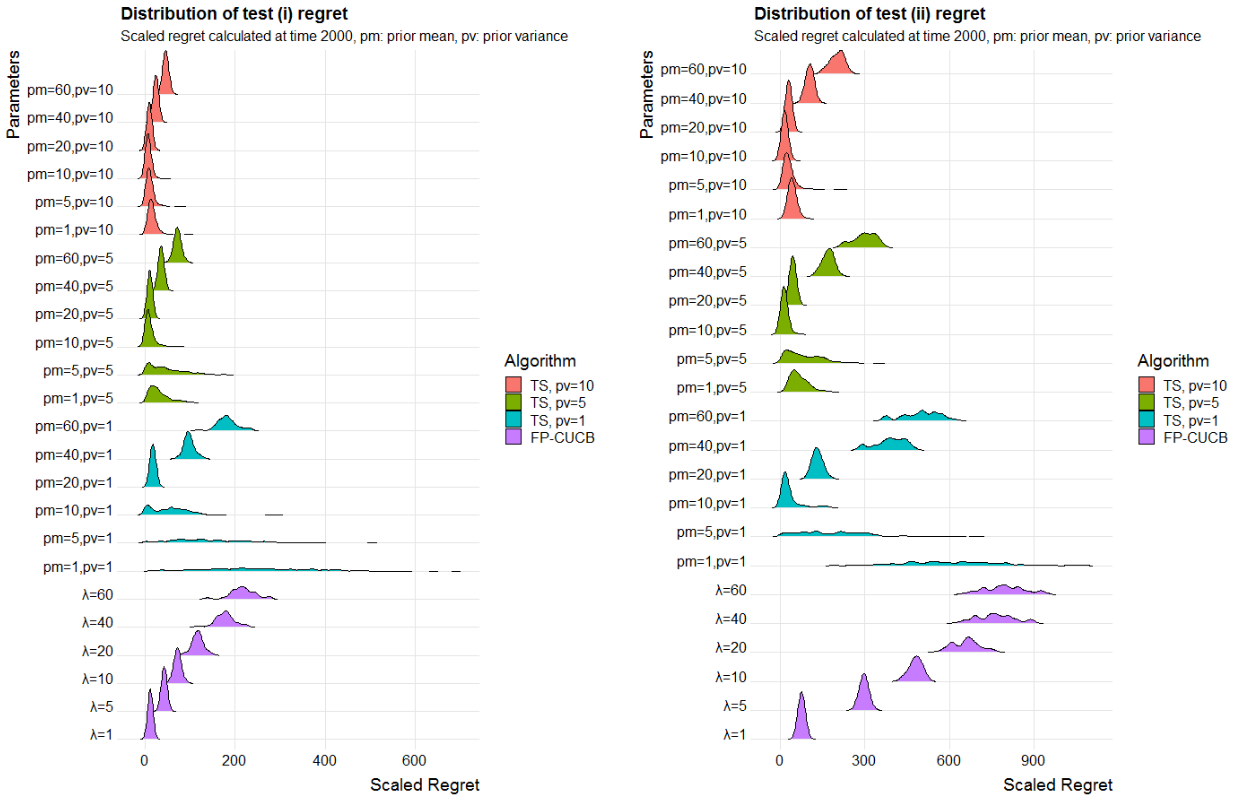


Figure 4.5.2: Scaled regret distributions in tests (i) and (ii). In both tests we have a true largest rate of 20.

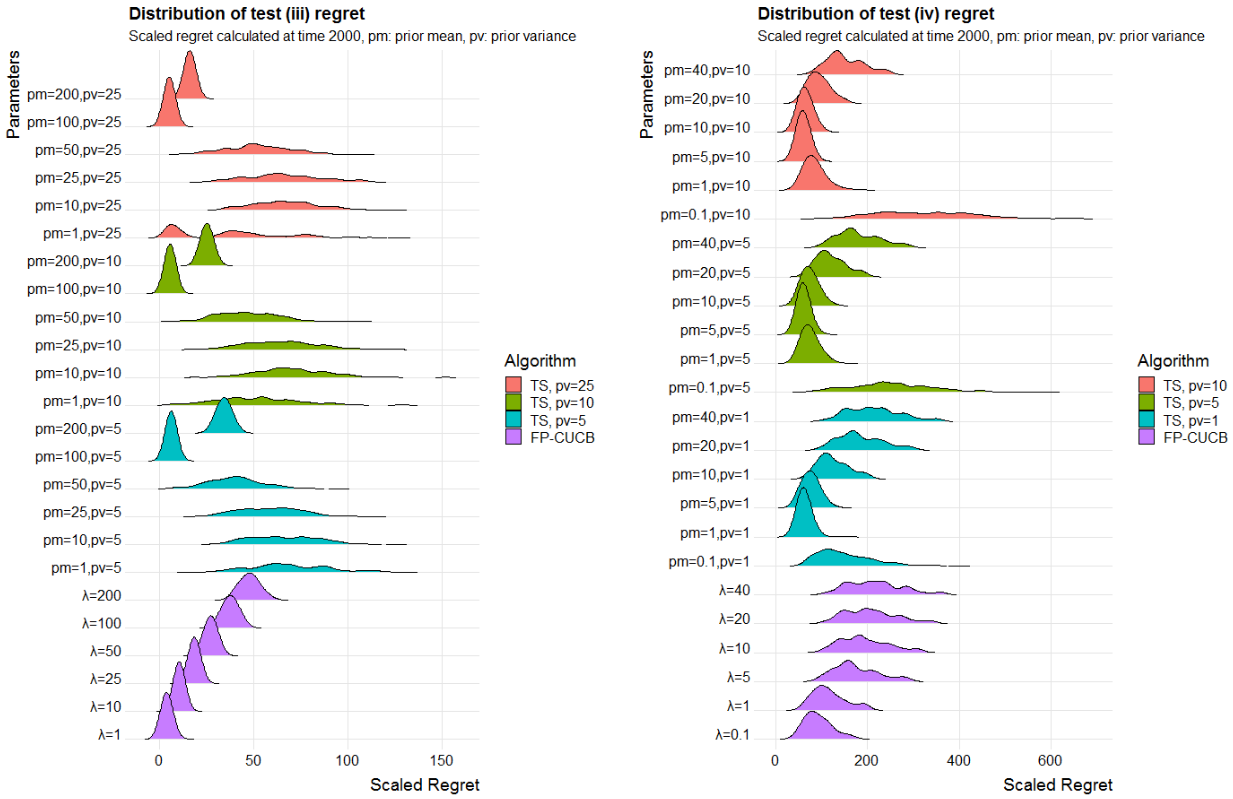


Figure 4.5.3: Scaled regret distributions in tests (iii) and (iv). In test (iii) the true largest rate is 100, and in test (iv) the true largest rate is 1.

We considered the sequential problem in two informational scenarios - firstly where the probability of detecting events is known, and secondly where these probabilities are unknown but one knows how they scale as the number of cells searched increases. For both of these cases we proposed an upper confidence bound approach. We derived lower bounds on the regret of all uniformly good algorithms under this our new feedback mechanism and upper bounds on the regret of our proposed approach.

In addition to the advantage of theoretical guarantees, the FP-CUCB algorithm is somewhat more reliable than TS. It is clear from the results of Section 4.5 that TS outperforms FP-CUCB for certain parametrisations (commonly larger choices of variance and mean close to the true arrival rates). However, we see that TS is particularly vulnerable to poor performance when the mean of the prior underestimates the true rate parameters. Even though our theoretical results for FP-CUCB depend on $\lambda_{max} \geq \lambda_k, k \in [K]$ we see that it is robust to underestimating this parameter. The reason FP-CUCB still performs well even when a key assumption does not hold is likely due to the fact that de la Peña's inequality does not give the tightest possible bound on Poisson tail probabilities (and therefore the rate of concentration of the mean). However, in order to construct the algorithm we required a symmetric tail bound for which an inflation term giving the type of concentration in Lemma 1 could be identified. Other bounds may be tighter but lack these properties.

The variability of TS most likely arises due to the potential for the Gamma conjugate prior to be dominated by a small number of observations and create a scenario where TS behaves similarly to a greedy policy - sometimes fixing on good actions, but sometimes on poor ones. This phenomenon of variability of regret is understudied in multi-armed bandits, not least because it is much more challenging to analyse theoretically. However, in practical scenarios (where of course the learning and regret minimisation process will only occur once) this is a risk of TS. We note that both algorithms comfortably outperform the greedy algorithm in almost all examples, which speaks to the benefit of making some attempt to balance exploration and exploitation.

An alternative treatment of bandit decision making is the *non-stochastic* or *adversarial* bandit (Auer et al., 1995). Under such a model, the assumptions that rewards are drawn i.i.d. from a fixed distribution are dropped, and may instead be any arbitrary sequence. Adversarial bandits necessitate a randomised strategy to guarantee good performance across any chosen reward sequence. Such methods have been developed in the MAB and CMAB settings (Auer et al., 1995; Cesa-Bianchi and Lugosi, 2012). As further work the problem could be studied under a non-stochastic, or even a fully game-theoretic framework, relaxing some of our assumptions. This would however require a markedly different set of algorithmic and analytical tools. Within application domains, variants of the problem exist all along the spectrum from purely stochastic to fully game-theoretic. Our work has considered the stochastic setting in detail and in doing so provided a solution to many real-world problems.

4.7 Additional proofs and results

4.7.1 Proof of NP-hardness of the IP (4.3.2)

Theorem 4.7.1. *Integer Linear Programs of the following type are $\mathcal{NP}$-hard in the strong sense:*

$$\begin{aligned}
\max_{a_{iju}, 1\leq i\leq j\leq K, u\in[U]} \sum_{i=1}^{K}\sum_{j=i}^{K}\sum_{u=1}^{U} q_{iju}a_{iju} \\
\textit{such that } \sum_{i=1}^{K}\sum_{j=i}^{K} a_{iju} &\leq 1, \quad u \in [U] \\
\sum_{i=1}^{K}\sum_{j=i}^{K}\sum_{u=1}^{U} a_{iju} &\leq 1, \quad k \in [K] \\
a_{iju} &\in \{0,1\}, \quad 1 \leq i \leq j \leq K, u \in [U].
\end{aligned}$$

Proof of Theorem 4.7.1:

The following problem is known to be $\mathcal{NP}$-complete in the strong sense (Garey and Johnson, 1979):

3-PARTITION: Given positive integers $w_1, ..., w_{3n}$ and a positive integer "target" t, does there exist a partition of $\{1, ..., 3n\}$ into subsets $S_1, ..., S_n$ such that $|S_i| = 3$ and $\sum_{j \in S_i} w_j = t$ for $i = 1, ..., n$?

We reduce this to an IP of the given type as follows. First, we assume without loss of generality that $\sum_{j=1}^{3n} w_j = nt$, since otherwise the answer to 3-PARTITION is trivially "no". Let $K = nt$ and $U = 3n$. For $k = 1, ..., 3n$, set $q_{iju} = w_u$ if $j - i = w_u$ and the half-open interval $[i, j)$ does not include a multiple of t. Set all other q_{iju} to zero. Then the answer to 3-PARTITION is "yes" if and only if there is a solution to the IP with profit equal to nt. □

4.7.2 Lemma 4.4.3 Proof: Concentration of filtered Poisson estimator

By definition $Z_j = \sum_{i=1}^{j}(Y_i - \mathbb{E}(Y_i))$, the sum of the accumulated noise to round t is a martingale. Therefore, $W_j = Z_j - Z_{j-1} = \sum_{i=1}^{j}(Y_i - \mathbb{E}(Y_i)) - \sum_{i=1}^{j-1}(Y_i - \mathbb{E}(Y_i)) = Y_j - \mathbb{E}(Y_j)$ is a martingale difference sequence. We will utilise the following concentration result for martingale difference sequences due to de la Peña (1999):

Theorem 4.7.2 (de la Peña's inequality). *Let $\{d_i, \mathcal{F}_i\}$ be a martingale difference sequence with $\mathbb{E}(d_j|\mathcal{F}_{j-1}) = 0$, $\mathbb{E}(d_j^2|\mathcal{F}_{j-1}) = \sigma_j^2$, $V_n^2 = \sum_{j=1}^{n} \sigma_j^2$. Furthermore assume that there exists c, $0 < c < \infty$ such that $\mathbb{E}(|d_j|^k|\mathcal{F}_{j-1}) \leq \frac{k!}{2}\sigma_j^2 c^{k-2}$ for $k \geq 2$. Then, for all $x, y > 0$*

$$\mathbb{P}(\sum_{i=1}^{n} d_i \geq x, V_n^2 \leq y \textit{ for some } n) \leq \exp\left(\frac{-x^2}{2(y+cx)}\right). \tag{4.7.22}$$

Plainly, $\mathbb{E}(W_j|\cdot) = 0$ and $\mathbb{E}(W_j^2|\cdot) = \gamma_j\lambda$. The proof of the condition on higher order moments is more involved. Firstly we define μ_k to be the k^{th} central moment of a Poisson distribution with parameter λ. Riordan (1937) gives us the following second order recurrence relationship for the central moments μ of the Poisson distribution

$$\mu_k = \lambda\bigg(\frac{d\mu_{k-1}}{d\lambda} + (k-1)\mu_{k-2}\bigg), \quad k = 2, 3, ...$$

We first demonstrate a bound on the order (with respect to λ) of μ_k.

Lemma 4.7.3. *For $k \geq 2$, $\mu_k = o(\lambda^{k/2})$.*

Proof of Lemma 4.7.3 We can prove Lemma 4.7.3 via an induction argument. Note that $\mu_1 = 0$ and $\mu_2 = \lambda$. Assume for some $r > 3$ that $\mu_r = o(\lambda^{r/2})$ and $\mu_{r-1} = o(\lambda^{(r-1)/2})$. Then consider $\mu_{r+1} = \lambda\frac{d\mu_r}{d\lambda} + r\lambda\mu_{r-1}$. For the first term we have $\frac{d\mu_r}{d\lambda} = o(\lambda^{r/2-1})$ and thus $\lambda\frac{d\mu_r}{d\lambda} = o(\lambda^{r/2})$. The second term is plainly of order $o(\lambda^{(r+1)/2})$ and thus $\mu_{r+1} = o(\lambda^{(r+1)/2})$, completing the induction argument and proving Lemma 4.7.3 $\square$.

Now introduce $\nu_k = \frac{k!}{2}\lambda \max(1, \sqrt{\lambda})^{k-2}$ for $k \geq 2$. The following lemma will be sufficient to demonstrate that the condition on higher order moments holds.

Lemma 4.7.4. *For $k \geq 2$, $\nu_k \geq \mu_k$.*

Proof of Lemma 4.7.4 Firstly we write ν_k as a recurrence relationship

$$\nu_k = k \max(1, \sqrt{\lambda})\nu_{k-1} = k(k-1)\max(1, \sqrt{\lambda})^2\nu_{k-2}, \quad k = 2, 3, ...$$

We also prove this Lemma via an induction argument, which proceeds as follows. For μ_k we have the following initial values $\mu_2 = \lambda$, $\mu_3 = \lambda$, $\mu_4 = 4\lambda^2$ and for ν_k we have $\nu_2 = \lambda$, $\nu_3 = 3\lambda\max(1, \sqrt{\lambda})$, $\nu_4 = 12\lambda\max(1, \sqrt{\lambda})^2$. Clearly, these initial values satisfy $\nu_k \geq \mu_k$. Now assume that for some $p > 5$, we have $\nu_p \geq \mu_p$ and $\nu_{p-1} \geq \mu_{p-1}$. Then consider μ_{p+1} as follows:

$$\begin{aligned} \mu_{p+1} &= \lambda\frac{d\mu_p}{d\lambda} + p\lambda\mu_{p-1} \\ &\leq \lambda\frac{d\mu_p}{d\lambda} + p\lambda\nu_{p-1} \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{p}{2}\mu_p + p\lambda\nu_{p-1}\\
&\leq \frac{p}{2}\nu_p + p\lambda\nu_{p-1}\\
&= \frac{p^2}{2}\max(1,\sqrt{\lambda})\nu_{p-1} + p\lambda\nu_{p-1}\\
&\leq \max(1,\sqrt{\lambda})^2(\frac{1}{2}p^2 + p)\nu_{p-1}\\
&\leq \max(1,\sqrt{\lambda})^2(p+1)p\nu_{p-1} = \nu_{p+1},
\end{aligned}
$$

completing the proof by induction. The first and third inequalities are due to the assumed relationships for p and $p-1$, the second inequality is a consequence of Lemma 4.7.3 and the differentiation of a polynomial. □.

The martingale difference sequence W_j therefore satisfies the conditions of de la Peña's inequality with $c = \max(1,\sqrt{\lambda})$ and we have

$$\mathbb{P}\bigg(\sum_{j=1}^{s} W_j \geq x, \sum_{j=1}^{s}\mathbb{E}(W_j^2|\cdot) \leq y\bigg) \leq \exp\left(\frac{-x^2}{2y + 2\max(1,\sqrt{\lambda})x}\right).$$

We have that $\sum_{j=1}^{s}\mathbb{E}(W_j^2|\cdot) \leq \sum_{j=1}^{s}\gamma_j\lambda$ with probability 1, so we may use the simplified result

$$\mathbb{P}\bigg(\sum_{j=1}^{s} W_j \geq x\bigg) \leq \exp\left(\frac{-x^2}{2\lambda\sum_{j=1}^{s}\gamma_j + 2\max(1,\sqrt{\lambda})x}\right).$$

Then if $x = 6\max(1,\sqrt{\lambda_{max}})\log(t) + \sqrt{6\lambda_{max}\sum_{j=1}^{s}\gamma_j\log(t)}$, and introducing the shorthand $\lambda_m = \lambda_{max}$ to save space, we have,

$$P\bigg(\sum_{i=1}^{n}(Y_i - \mathbb{E}(Y_i)) > x\bigg)$$

$$
\begin{aligned}
&\leq \exp\left(- \frac{36 \max(1, \lambda_m) \log^2(t) + 12 \max(1, \sqrt{\lambda_m}) \log(t) \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)} + 6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)}{2\lambda \sum_{j=1}^{s} \gamma_j + 12 \max(1, \sqrt{\lambda \cdot \lambda_m}) \log(t) + 2 \max(1, \sqrt{\lambda}) \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)}} \right) \\
&= \exp\left(- \log(t) \frac{36 \max(1, \lambda_m) \log(t) + 12 \max(1, \sqrt{\lambda_m}) \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)} + 6\lambda_m \sum_{j=1}^{s} \gamma_j}{12 \max(1, \sqrt{\lambda \cdot \lambda_m}) \log(t) + 2 \max(1, \sqrt{\lambda}) \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)} + 2\lambda \sum_{j=1}^{s} \gamma_j} \right) \\
&= \exp\left(- 3\log(t) \frac{12 \max(1, \lambda_m) \log(t) + 4 \max(1, \sqrt{\lambda_m}) \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)} + 2\lambda_m \sum_{j=1}^{s} \gamma_j}{12 \max(1, \sqrt{\lambda \cdot \lambda_m}) \log(t) + 2 \max(1, \sqrt{\lambda}) \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)} + 2\lambda \sum_{j=1}^{s} \gamma_j} \right) \\
&\leq t^{-3}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&P\Bigg(\sum_{i=1}^{n} (Y_i - \mathbb{E}(Y_i)) > 6 \max(1, \sqrt{\lambda_m}) \log(t) + \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)} \Bigg) \\
&= P\Bigg(\sum_{i=1}^{n} Y_i - \lambda \sum_{i=1}^{n} \gamma_i > 6 \max(1, \sqrt{\lambda_m}) \log(t) + \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)} \Bigg) \\
&= P\Bigg(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \gamma_i} - \lambda > \frac{6 \max(1, \sqrt{\lambda_{max}}) \log(t) + \sqrt{6\lambda_m \sum_{j=1}^{s} \gamma_j \log(t)}}{\sum_{i=1}^{n} \gamma_i} \Bigg) \leq t^{-3}.
\end{aligned}
$$

Finally, note that $\bar{Z}_j = \sum_{i=1}^{j} (\mathbb{E}(Y_i) - Y_i) = -Z_j$ is also a martingale whose difference series satisfies the conditions of de la Peña's inequality and thus we can achieve the same bound for deviations on the left, and introduce achieve the required result. □

4.7.3 Theorem 4.4.1 Proof: Expected regret of FP-CUCB

To complete the proof of Theorem 4.4.1 provided in the main text, we separately prove Propositions 4.4.2 and 4.4.4.

Proof of Proposition 4.4.2:

Here we prove a bound on the expected number of plays of an arm *after* it has reached its sufficient sampling level. Define the event

$$\mathcal{N}_t = \left\{ \left| \frac{\sum_{j=1}^{t-1} Y_{k,j}}{D_{k,t-1}} - \lambda_k \right| < \frac{6\max(1, \sqrt{\lambda_{max}})\log(t)}{D_{k,t-1}} + \sqrt{\frac{6\lambda_{max}\log(t)}{D_{k,t-1}}} \;\; \forall k \in [K] \right\}.$$

Define random variables $\Lambda_{k,t} = \frac{6\max(1,\sqrt{\lambda_{max}})\log(t)}{D_{k,t-1}} + \sqrt{\frac{6\lambda_{max}\log(t)}{D_{k,t-1}}}$ for $k \in [K]$ and $\Lambda_t = \max_{k:g_{k,t}>0}(\Lambda_{k,t})$. Define $\Lambda_t^{k,l} = \frac{6\max(1,\sqrt{\lambda_{max}})\log(t)}{\gamma_{k,min}h_{k,n}(\Delta^{k,l})} + \sqrt{\frac{6\lambda_{max}\log(t)}{\gamma_{k,min}h_{k,n}(\Delta^{k,l})}}$ for $l \in [B_k]$, $k \in [K]$, $t \in [n]$, which are not random variables. By these definitions and the definition of UCB indices $\bar{\lambda}_{k,t}$ we have the following properties.

$$\mathcal{N}_t \Rightarrow \bar{\lambda}_{k,t} - \lambda_k > 0 \;\; \forall k \in [K]$$

$$\mathcal{N}_t \Rightarrow \bar{\lambda}_{k,t} - \lambda_k < 2\Lambda_t \;\; \forall k : g_{k,t} > 0$$

$$\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{s,t-1} > h_{k,n}(\Delta^{k,l}) \;\; \forall s : g_{s,t} > 0\} \Rightarrow \Lambda_t^{k,l} > \Lambda_t \;\; \forall k \in [K], \forall l \in [B_k]$$

For any particular $k \in [K]$ and $l \in [B_k]$ if $\{\mathcal{N}_t, \mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{s,t-1} > h_{k,n}(\Delta^{k,l}) \;\; \forall s : g_{s,t} > 0\}$ holds at time t the following is implied

$$\mathbf{g}_t^T \cdot \boldsymbol{\lambda} + 2K\Lambda_t^{k,l} > \mathbf{g}_t^T \cdot \boldsymbol{\lambda} + 2K\Lambda_t \geq \mathbf{g}_t^T \cdot \bar{\boldsymbol{\lambda}}_t \geq (\mathbf{g}_{\boldsymbol{\lambda}}^*)^T \cdot \bar{\boldsymbol{\lambda}}_t \geq (\mathbf{g}_{\boldsymbol{\lambda}}^*)^T \cdot \boldsymbol{\lambda} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} \tag{4.7.23}$$

where $\mathbf{g}_{\boldsymbol{\lambda}}^*$ is an action that is optimal with respect to rate vector $\boldsymbol{\lambda}$. However, by definition $2K\Lambda_t^{k,l} \geq \Delta^{k,l}$ and therefore (4.7.23) is a contradiction of the definition of $\Delta^{k,l} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} - \mathbf{g}_{k,B}^l \cdot \boldsymbol{\lambda}$. Therefore

$$\mathbb{P}(\mathcal{N}_t, \mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, \forall s : g_{s,t} > 0, N_{s,t-1} > h_{k,n}(\Delta^{k,l})) = 0 \;\; \forall k \in [K], \;\; \forall l \in [B_k]$$

and

$$\sum_{k=1}^{K}\sum_{l=1}^{B_k}\mathbb{P}(\mathbf{g}_t=\mathbf{g}_{k,B}^{l}, N_{k,t}>N_{k,t-1}, N_{s,t-1}>h_{k,n}(\Delta^{k,l})\ \ \forall s: g_{s,t}>0)\leq\mathbb{P}(\neg\mathcal{N}_t)\leq 2Kt^{-2}.$$

The bound on $\mathbb{P}(\neg\mathcal{N}_t)$ comes from applying Lemma 1 and is sufficient to prove Proposition 4.4.2 since

$$\begin{aligned}\mathbb{E}\left(\sum_{k=1}^{K}\sum_{l=1}^{B_k}N_{k,n}^{l,suf}\right)&=\mathbb{E}\left(\sum_{t=K+1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{B_k}\mathbb{I}\{\mathbf{g}_t=\mathbf{g}_{k,B}^{l}, N_{k,t}>N_{k,t-1}, N_{k,t-1}>h_{k,n}(\Delta^{k,l})\}\right)\\&\leq\sum_{t=K+1}^{n}2Kt^{-2}\leq\frac{\pi^2}{3}\cdot K.\quad\square\end{aligned}$$

Proof of Proposition 2

Now consider the number of plays made prior to reaching the sufficient sampling level. Firstly set $h_{k,n}(\Delta^{k,0})=0$ to simplify notation and consider the following steps. Then for any cell k in $\{j\in[K]|\Delta_{min}^{j}>0\}$

$$\begin{aligned}&\sum_{l=1}^{B_k}N_{k,n}^{l,und}\cdot\Delta^{k,l}\\&=\sum_{t=K+1}^{n}\sum_{l=1}^{B_k}\mathbb{I}\Big\{\mathbf{g}_t=\mathbf{g}_{k,B}^{l}, N_{k,t}>N_{k,t-1}, N_{k,t-1}\leq h_{k,n}(\Delta^{k,l})\Big\}\Delta^{k,l}\\&=\sum_{t=K+1}^{n}\sum_{l=1}^{B_k}\sum_{j=1}^{l}\mathbb{I}\Big\{\mathbf{g}_t=\mathbf{g}_{k,B}^{l}, N_{k,t}>N_{k,t-1}, N_{k,t-1}\in\Big(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j})\Big)\Big\}\Delta^{k,l}\\&\leq\sum_{t=K+1}^{n}\sum_{l=1}^{B_k}\sum_{j=1}^{l}\mathbb{I}\Big\{\mathbf{g}_t=\mathbf{g}_{k,B}^{l}, N_{k,t}>N_{k,t-1}, N_{k,t-1}\in\Big(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j})\Big)\Big\}\Delta^{k,j}\end{aligned}$$

as $\Delta^{k,1} \geq \Delta^{k,2} \geq ... \geq \Delta^{k,B_k}$,

$$
\begin{aligned}
&\leq \sum_{t=K+1}^{n} \sum_{l=1}^{B_k} \sum_{j=1}^{B_k} \mathbb{I}\Big\{ \mathbf{g}_t = \mathbf{g}^l_{k,B}, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \Big(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j}) \Big) \Big\} \Delta^{k,j} \\
&= \sum_{t=K+1}^{n} \sum_{j=1}^{B_k} \mathbb{I}\Big\{ \mathbf{g}_t \in \mathcal{G}_{k,B}, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \Big(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j}) \Big) \Big\} \Delta^{k,j} \\
&= \sum_{j=1}^{B_k} \sum_{t=K+1}^{n} \mathbb{I}\Big\{ \mathbf{g}_t \in \mathcal{G}_{k,B}, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \Big(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j}) \Big) \Big\} \Delta^{k,j} \\
&\leq \sum_{j=1}^{B_k} \Big(h_{k,n}(\Delta^{k,j}) - h_{k,n}(\Delta^{k,j-1}) \Big) \Delta^{k,j}
\end{aligned}
$$

since N_k can only be incremented a maximum of $h_{k,n}(\Delta^{k,j}) - h_{k,n}(\Delta^{k,j-1})$ times while remaining in this range

$$
\begin{aligned}
&= h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \sum_{j=1}^{B_k - 1} h_{k,n}(\Delta^{k,j}) \cdot (\Delta^{k,j} - \Delta^{k,j+1}) \\
&\leq h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \int_{\Delta^{k,B_k}}^{\Delta^{k,1}} h_{k,n}(x) dx.
\end{aligned}
$$

The last inequality holds since $h_{k,n}(x)$ are decreasing functions. □

4.7.4 Theorem 4.4.5 Proof: Lower bound on regret

To prove Theorem 4.4.5, we must define the additional quantities necessary to apply Theorem 1 of Graves and Lai (1997) and frame the problem accordingly.

We consider the reward history $(\mathbf{Y}_t)_{t=1}^n$ to be a realisation of a controlled Markov Chain moving on the state space $\mathbb{N}^K$ where the controls are the detection probability vectors selected in each round. Each control g $\in \mathcal{G}$ then has an associated set of $\boldsymbol{\lambda}$ parameter vectors under which it is an optimal control $\Lambda_{\mathbf{g}} = \{\boldsymbol{\lambda} \in \mathbb{R}_+^K : \mathbf{g}^T \cdot \boldsymbol{\lambda} = \text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}\}$, which may be the empty set. For any states $\mathbf{y}, \mathbf{z} \in \mathbb{N}^K$ transition probabilities are straightforward Poisson probabilities due to independence across rounds:

$$p(y, z; \boldsymbol{\lambda}, \mathbf{g}) = p(z; \boldsymbol{\lambda}, \mathbf{g}) = \prod_{k=1}^{K} \frac{(g_k \lambda_k)^{z_k} e^{-g_k \lambda_k}}{z_k!}.$$

These transition probabilities define the Kullback Leibler Information number for any control $\mathbf{g} \in \mathcal{G}$:

$$I^{\mathbf{g}}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \log\left(\frac{p(z_k; \boldsymbol{\lambda}, \mathbf{g})}{p(z_k; \boldsymbol{\theta}, \mathbf{g})}\right) p(z_k; \boldsymbol{\lambda}, \mathbf{g}) = \sum_{k=1}^{K} \text{kl}(g_k \lambda_k, \gamma_k \theta_k) = \sum_{k=1}^{K} g_k \text{kl}(\lambda_k, \theta_k).$$

With these quantities and those defined in Section 4.4.1 we can apply Theorem 1 of Graves and Lai (1997) to reach the following result for any uniformly good policy π

$$\liminf_{n \to \infty} \sum_{\mathbf{g} \in \mathcal{J} \setminus J(\boldsymbol{\lambda})} \frac{I^{\mathbf{g}}(\boldsymbol{\lambda}, \boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\lambda}}(\sum_{t=1}^{n} \mathbb{I}\{\mathbf{g}_t = \mathbf{g}\})}{\log(n)} \geq 1 \text{ for every } \boldsymbol{\theta} \in B(\boldsymbol{\lambda}).$$

Since $Reg^{\pi}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) = \sum_{\mathbf{g} \in \mathcal{J} \setminus J(\boldsymbol{\lambda})} \Delta_{\mathbf{g}} \mathbb{E}_{\boldsymbol{\lambda}}(\sum_{t=1}^{n} \mathbb{I}\{\mathbf{g}_t = \mathbf{g}\})$ the required result follows. □

4.7.5 Numerical Results

Algorithm	Parameters	0.025 Quantile	Median	0.975 Quantile
FP-CUCB	$\lambda_{max} = 1$	9.52	11.96	15.89
	$\lambda_{max} = 5$	36.42	42.53	50.03
	$\lambda_{max} = 10$	57.44	72.57	88.70
	$\lambda_{max} = 20$	89.07	117.97	143.95
	$\lambda_{max} = 40$	123.23	178.07	223.81
	$\lambda_{max} = 60$	143.87	215.46	276.25
Thompson Sampling	Mean=1, Variance=1	38.44	242.39	508.93
	Mean=5, Variance=1	1.95	132.79	358.15
	Mean=10, Variance=1	1.44	56.30	134.12
	Mean=20, Variance=1	11.66	17.76	25.88
	Mean=40, Variance=1	75.24	96.87	124.57
	Mean=60, Variance=1	122.72	180.67	233.25
	Mean=1, Variance=5	5.69	26.49	90.89
	Mean=5, Variance=5	2.32	38.51	134.07
	Mean=10, Variance=5	2.18	7.19	43.90
	Mean=20, Variance=5	7.17	10.95	15.80
	Mean=40, Variance=5	30.00	36.11	43.23
	Mean=60, Variance=5	57.61	72.42	87.30
	Mean=1, Variance=10	6.31	14.21	36.57
	Mean=5, Variance=10	3.60	9.35	35.87
	Mean=10, Variance=10	3.28	6.65	18.41
	Mean=20, Variance=10	6.55	9.67	15.97
	Mean=40, Variance=10	20.15	24.65	30.25
	Mean=60, Variance=10	40.17	46.12	55.09
Greedy		79.77	679.76	1657.52

Table 4.7.1: Quantiles of scaled regret at horizon $n = 2000$ for algorithms applied to Test (i) data

Algorithm	Parameters	0.025 Quantile	Median	0.975 Quantile
FP-CUCB	$\lambda_{max} = 1$	65.82	75.92	89.31
	$\lambda_{max} = 5$	269.77	297.74	323.55
	$\lambda_{max} = 10$	433.92	480.78	517.22
	$\lambda_{max} = 20$	577.74	661.08	754.25
	$\lambda_{max} = 40$	643.61	759.90	891.13
	$\lambda_{max} = 60$	665.45	794.38	931.36
Thompson Sampling	Mean=1, Variance=1	286.11	603.51	969.56
	Mean=5, Variance=1	7.94	184.48	568.05
	Mean=10, Variance=1	8.12	21.05	159.10
	Mean=20, Variance=1	102.40	132.00	174.17
	Mean=40, Variance=1	286.61	395.04	472.38
	Mean=60, Variance=1	371.53	504.06	609.47
	Mean=1, Variance=5	26.95	61.18	153.86
	Mean=5, Variance=5	9.56	70.19	224.13
	Mean=10, Variance=5	6.55	13.80	40.48
	Mean=20, Variance=5	36.57	45.19	56.15
	Mean=40, Variance=5	128.60	172.27	208.23
	Mean=60, Variance=5	222.33	303.67	361.93
	Mean=1, Variance=10	25.38	41.44	69.92
	Mean=5, Variance=10	12.61	26.23	100.81
	Mean=10, Variance=10	10.22	15.79	32.32
	Mean=20, Variance=10	24.28	30.60	39.17
	Mean=40, Variance=10	84.45	106.17	122.09
	Mean=60, Variance=10	151.68	206.13	244.60
Greedy		296.46	720.45	1163.15

Table 4.7.2: Quantiles of scaled regret at horizon $n = 2000$ for algorithms applied to Test (ii) data

Algorithm	Parameter	0.025 Quantile	Median	0.975 Quantile
FP-CUCB	$\lambda_{max} = 1$	2.17	3.37	7.20
	$\lambda_{max} = 10$	9.19	10.34	11.78
	$\lambda_{max} = 25$	15.92	18.45	21.33
	$\lambda_{max} = 50$	22.57	27.39	31.58
	$\lambda_{max} = 100$	30.41	37.59	44.85
	$\lambda_{max} = 200$	38.90	48.07	57.96
Thompson Sampling	Mean=1, Variance=5	30.47	66.95	115.65
	Mean=10, Variance=5	36.78	64.41	98.24
	Mean=25, Variance=5	29.06	58.44	95.57
	Mean=50, Variance=5	10.82	39.65	71.71
	Mean=100, Variance=5	4.83	6.05	7.71
	Mean=200, Variance=5	28.24	34.20	40.37
	Mean=1, Variance=10	12.61	52.06	97.08
	Mean=10, Variance=10	33.99	68.30	109.44
	Mean=25, Variance=10	30.97	64.55	105.03
	Mean=50, Variance=10	17.32	46.39	80.35
	Mean=100, Variance=10	4.26	5.52	7.09
	Mean=200, Variance=10	21.37	25.06	29.00
	Mean=1, Variance=25	3.87	37.19	102.98
	Mean=10, Variance=25	36.51	66.12	107.72
	Mean=25, Variance=25	30.87	64.73	106.71
	Mean=50, Variance=25	20.21	51.32	86.70
	Mean=100, Variance=25	3.86	5.09	6.79
	Mean=200, Variance=25	14.08	15.92	18.09
Greedy		21.57	49.20	95.89

Table 4.7.3: Quantiles of scaled regret at horizon $n = 2000$ for algorithms applied to Test (iii) data

Algorithm	Parameters	0.025 Quantile	Median	0.975 Quantile
FP-CUCB	$\lambda_{max} = 0.1$	47.56	87.07	162.36
	$\lambda_{max} = 1$	62.60	108.48	195.80
	$\lambda_{max} = 5$	98.70	163.13	279.62
	$\lambda_{max} = 10$	109.59	184.01	311.37
	$\lambda_{max} = 20$	116.40	200.99	336.25
	$\lambda_{max} = 40$	120.39	210.65	356.25
Thompson Sampling	Mean=0.1, Variance=1	70.68	136.84	284.14
	Mean=1, Variance=1	42.78	61.44	91.98
	Mean=5, Variance=1	43.96	75.38	119.21
	Mean=10, Variance=1	75.45	118.86	197.36
	Mean=20, Variance=1	104.58	174.02	291.32
	Mean=40, Variance=1	119.72	207.46	349.74
	Mean=0.1, Variance=5	94.23	246.71	467.06
	Mean=1, Variance=5	43.48	73.41	119.94
	Mean=5, Variance=5	41.71	60.07	88.64
	Mean=10, Variance=5	45.15	72.69	119.42
	Mean=20, Variance=5	69.43	113.12	191.90
	Mean=40, Variance=5	102.60	169.98	281.94
	Mean=0.1, Variance=10	134.60	320.63	588.63
	Mean=1, Variance=10	48.26	81.35	146.95
	Mean=5, Variance=10	41.43	58.66	84.74
	Mean=10, Variance=10	40.78	62.10	99.55
	Mean=20, Variance=10	55.42	89.68	146.88
	Mean=40, Variance=10	86.98	141.99	239.18
Greedy		664.28	1825.61	1999.89

Table 4.7.4: Quantiles of scaled regret at horizon $n = 2000$ for algorithms applied to Test (iv) data

Chapter 5

Continuum Armed Bandit Model of Sequential Event Detection

A version of this chapter has been published as Grant, J.A., Boukouvalas, A., Griffiths, R., Leslie, D.S., Vakili, S., and Munoz de Cote, E. (2019). Adaptive Sensor Placement for Continuous Spaces. *In Proceedings of 36th International Conference on Machine Learning*.

This work was completed in collaboration with research staff at PROWLER.io, and the code to produce the experiments and figures was produced by software engineers at PROWLER.io, with my instruction on the desired output. The optimisation algorithm Action Selection by Iterative Merging (AS-IM) and the proof of the bound on its sample complexity in Theorem 5.2.1 are thanks to Sattar Vakili and are included for completeness, rather than under the assertion that they are my own original work.

5.1 Introduction

In this chapter we study the sequential event detection problem as a continuum-armed bandit problem, with the application of sensor placement in mind. The model of reward and event detection is slightly different to that of the previous chapter. We 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. All the events that occur in a sensed subinterval are detected, but none which occur outside a sensed subinterval will be detected. 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. In this chapter we consider 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.

Our model most closely resembles the continuum-armed or $\mathcal{X}$-armed bandit problem Agrawal (1995). We recall that 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 a vector of endpoints 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 scalar reward observations with bounded or sub-Gaussian noise (e.g. Bubeck et al., 2011). In tackling the sequential event detection problem, we handle the added complexities of observing event locations and the heavier-tailed noise of the Poisson distribution.

The method we will propose performs fast Bayesian inference on the rate function, by means of the Bayesian histogram approach Gugushvili et al. (2018), and makes decisions to trade off exploration and exploitation using Thompson sampling (TS) (see 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 5.2.4.

We analyse the Bayesian regret of the TS algorithm in this setting using similar techniques to those of Russo and 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).

5.1.1 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. In Chapter 4, we presented the first attempt to solve a version of the problem in which the rate must be learned, in which the space is discretised to a fixed grid for all time. The objective of this chapter 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 the Combinatorial Multi-Armed Bandit (CMAB) problem (Cesa-Bianchi and Lugosi, 2012; Chen et al., 2016a). We recall that 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., 2010; Chowdhury and Gopalan, 2017; Basu and Ghosh, 2017). As mentioned in Section 5.1, 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 and Goyal, 2012; Kaufmann et al., 2012a). 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 and Chen (2018) and Huyuk and Tekin (2019) 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 and Van Roy, 2014; Bubeck and Liu, 2013) are powerful as they do not depend on a specific parameterisation of the reward functions.

5.1.2 Key Contributions

Similarly to Chapter 4, this chapter makes a number of contributions to bandit theory and again provides a practically useful solution to a real problem. We summarise the principal contributions below:

- Formulation of a new widely applicable model of sequential sensor placement as a CAB;
- 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;
- An efficient optimisation routine for sensor placement given known event rate;
- Analysis of the Bayesian regret of a TS approach, resulting in a $\tilde{O}(T^{2/3})$ upper bound;

- Numerical validation of the efficacy of the TS method, and its favourable performance relative to upper confidence bound and ϵ-greedy approaches.

5.1.3 Chapter Outline

The remainder of the chapter is structured as follows. In Section 5.2 we formalise our model and algorithm. The algorithm has three main components, the Bayesian histogram approach to inference, a bespoke optimisation routine to select actions, and the TS component to balance exploration and exploitation. In Section 5.3 we present theoretical analysis of the Bayesian regret, leading to an $\tilde{O}(T^{2/3})$ upper bound. We conclude in sections 5.4 and 5.5 with numerical experiments and a discussion respectively.

5.2 Model and solution

We now formally present our model and solution method.

5.2.1 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 Lipschitz smooth 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.

5.2.2 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

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

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 TG$(\alpha, \beta, 0, \lambda_{\max})$ distribution has a density proportional to a Gamma(α, β) 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 5.3.1.

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

$$\begin{aligned} \lambda_t(x) &= \sum_{k=1}^{K_t} \mathbb{I}\{x \in B_{k,t}\} \tilde{\psi}_{k,t}, \quad \text{with} \\ \tilde{\psi}_{k,t} &\sim \mathrm{TG}(\alpha + H_{k,t}(t), \beta + \Delta_t N_{k,t}(t), 0, \lambda_{\max}), \end{aligned} \tag{5.2.1}$$

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 5.2.3).

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,

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

if $N_{k,t}(t) = t$ for all $k \in [K_t]$ and $K_t = O(t^{h/2h+1})$. 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 $[0, 1]$.

5.2.3 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.

Our TS approach is described in Algorithm 9. In each round t, for each bin $k \in \{1, \dots, K_t\}$, a rate $\tilde{\psi}_{k,t}$ is sampled according to (5.2.1), 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 action 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 5.2.4.

5.2.4 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. In Chapter

Algorithm 9: Thompson Sampling

```
Inputs: Gamma prior parameters $\alpha, \beta > 0$, upper truncation point $\lambda_{\max}$
Iterative Phase: For $t \geq 1$
    • For each $k \in \{1, \dots, K_t\}$, 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 $A_t$
```

4 we solved 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. Here, instead, we 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 $w(I) = \int_I (\lambda(x) - C)dx$. 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 $k = 2, ..., K_t$, $B_{k,t}$ 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 $M = \{n \in \{2, \dots, N-1\} : |w(I_n)| \leq |w(I_{n-1})|, |w(I_n)| \leq |w(I_{n+1})|\}$ 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 $n = \mathrm{argmin}_M |w(I_n)|$ 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 $I_1^*, I_2^*, ..., I_U^*$.

We have the following result on AS-IM guaranteeing its optimality and efficiency. The proof is given in the Section 5.6.2 via an induction argument.

Theorem 5.2.1. *The AS-IM policy returns the optimal action and its sample complexity is not bigger than* $O(K_t \log K_t)$.

5.3 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 finer 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" incurred 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 5.4 and find that this rebinning rate is superior to a faster linear rate of rebinning.

Theorem 5.3.1. *Consider the setup of Section 5.2, 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 9 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. Nevertheless, we have encouraging evidence that our Thompson Sampling approach is a strongly performing policy.

Proof of Theorem 5.3.1. 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

$$BReg(T) = \mathrm{E}\left(\sum_{t=1}^{T} \delta(A_t^*)\right) + \mathrm{E}\left(\sum_{t=1}^{T} \delta_t(A_t)\right)$$

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 re-binning. In particular we have the following lemma, proved in the Section 5.6.1.

Lemma 5.3.2. *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 and 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,

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

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,

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

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:

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

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 Section 5.6.1 we prove the following lemmas, which when combined are sufficient to complete the proof of Theorem 5.3.1.

Lemma 5.3.3. *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 5.3.4. *The deviation probabilities can be bounded*

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

Combining these results we have:

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

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

5.4 Simulations

In this section, we provide simulation examples on the performance of the Thompson sampling approach presented in Section 5.2.3. 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.

5.4.1 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

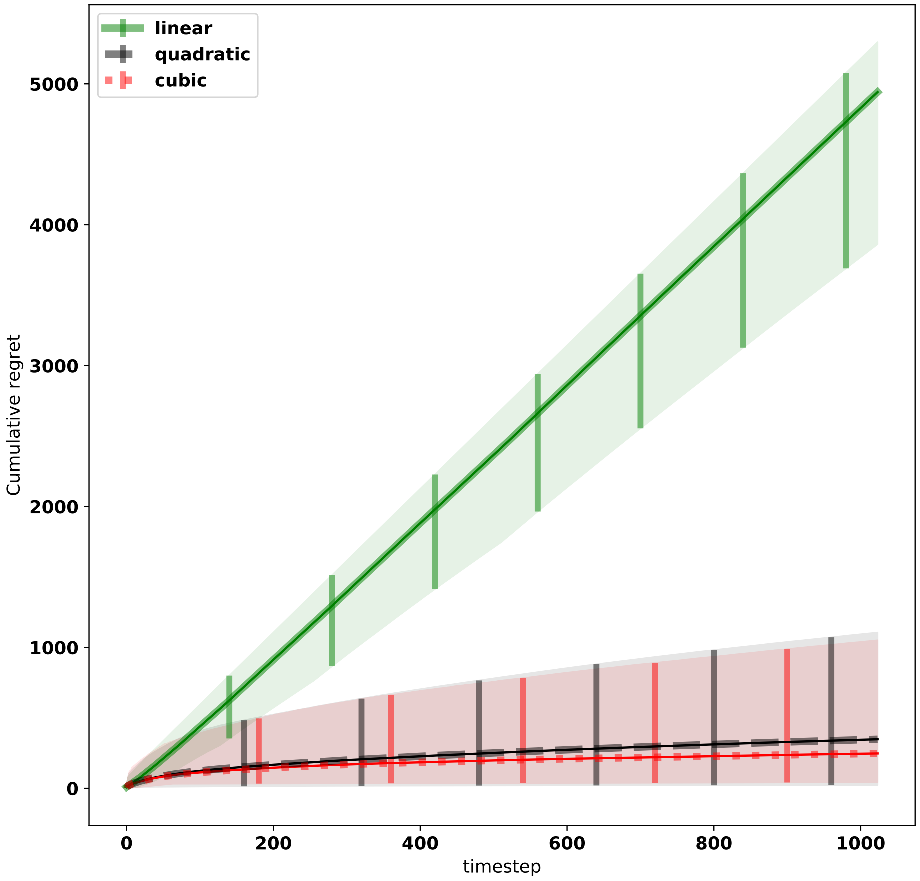


Figure 5.4.1: Cumulative regret comparing different rebinning rates. Green line denotes the average regret under linear rebinning, black dashed line denotes the average regret under square root order rebinning, and red dot-dashed line denotes the average regret under cube root order rebinning. Shaded areas illustrate empirical 95% confidence intervals.

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 9 and Fig. 5.4.1 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 Section 5.6.3.

We also show, in Fig. 5.4.2, 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.

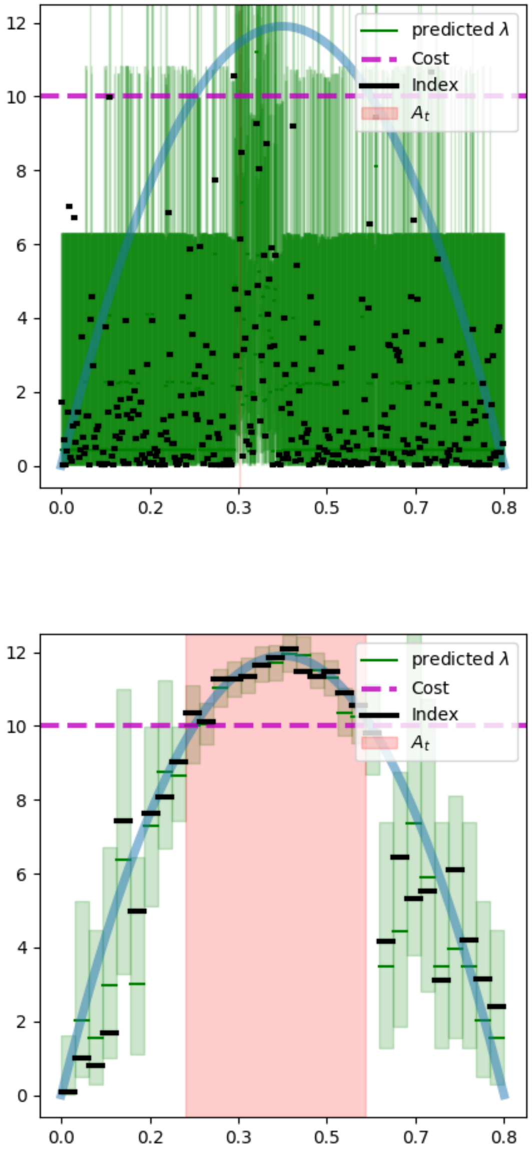


Figure 5.4.2: Posterior under the linear and cube root rebinning rates at round $T = 1024$. We show the true rate function (blue) and cost (pink), the posterior credible interval (light green) and mean (dark green) per bin. Thompson samples are shown in black, and the selected interval, A_T, is the (red) 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.

5.4.2 Comparison to Baselines

We now compare different baseline policies solely using the cube root rebinning schedule. Experiments with the unimodal rate of Section 5.4.1 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 5.2.3, 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) (a paper which is an early version of Chapter 4), 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 a 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 $\epsilon = 0.01$.

The cumulative regret for each policy is shown in Figure 5.4.3. 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 5.4.4(d). Even after 900 iterations, the UCB values (in black) are close to the cost threshold even

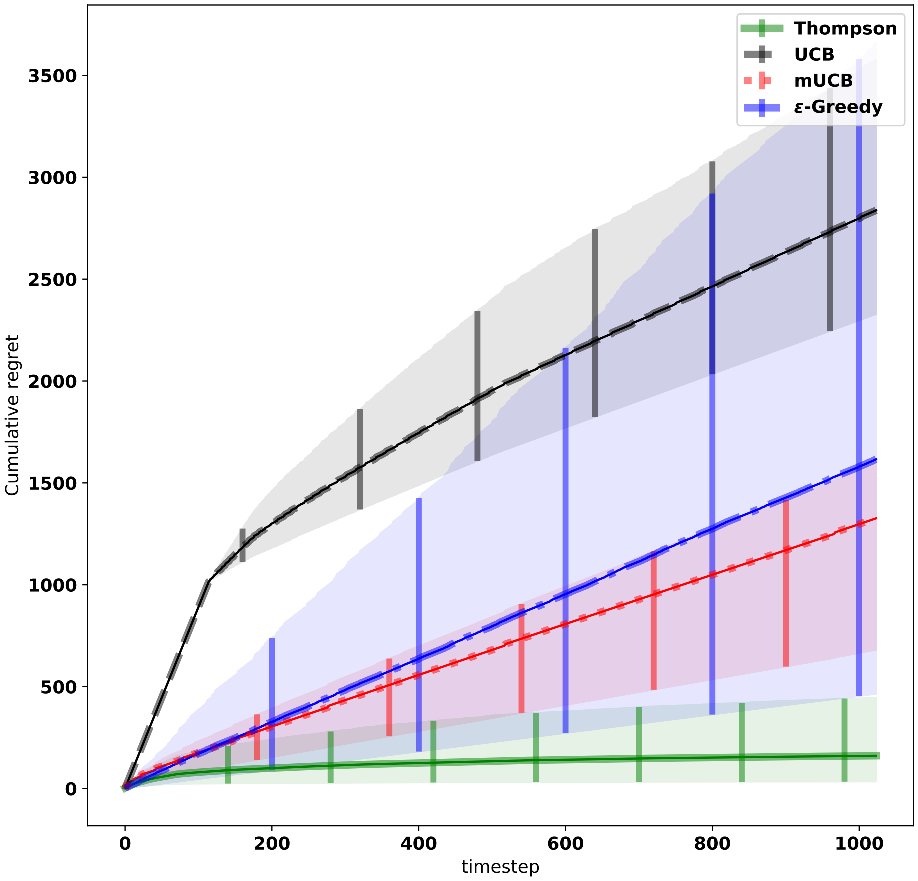


Figure 5.4.3: Cumulative regret plot for the bimodal rate functions. Green line denotes the average regret for Thompson Sampling, black dashed line denotes the average regret for the UCB algorithm, red dotted line denotes the average regret for the modified UCB algorithm, and blue dot-dashed line denotes the average regret for the greedy algorithm. Shaded areas illustrate empirical 95% confidence intervals.

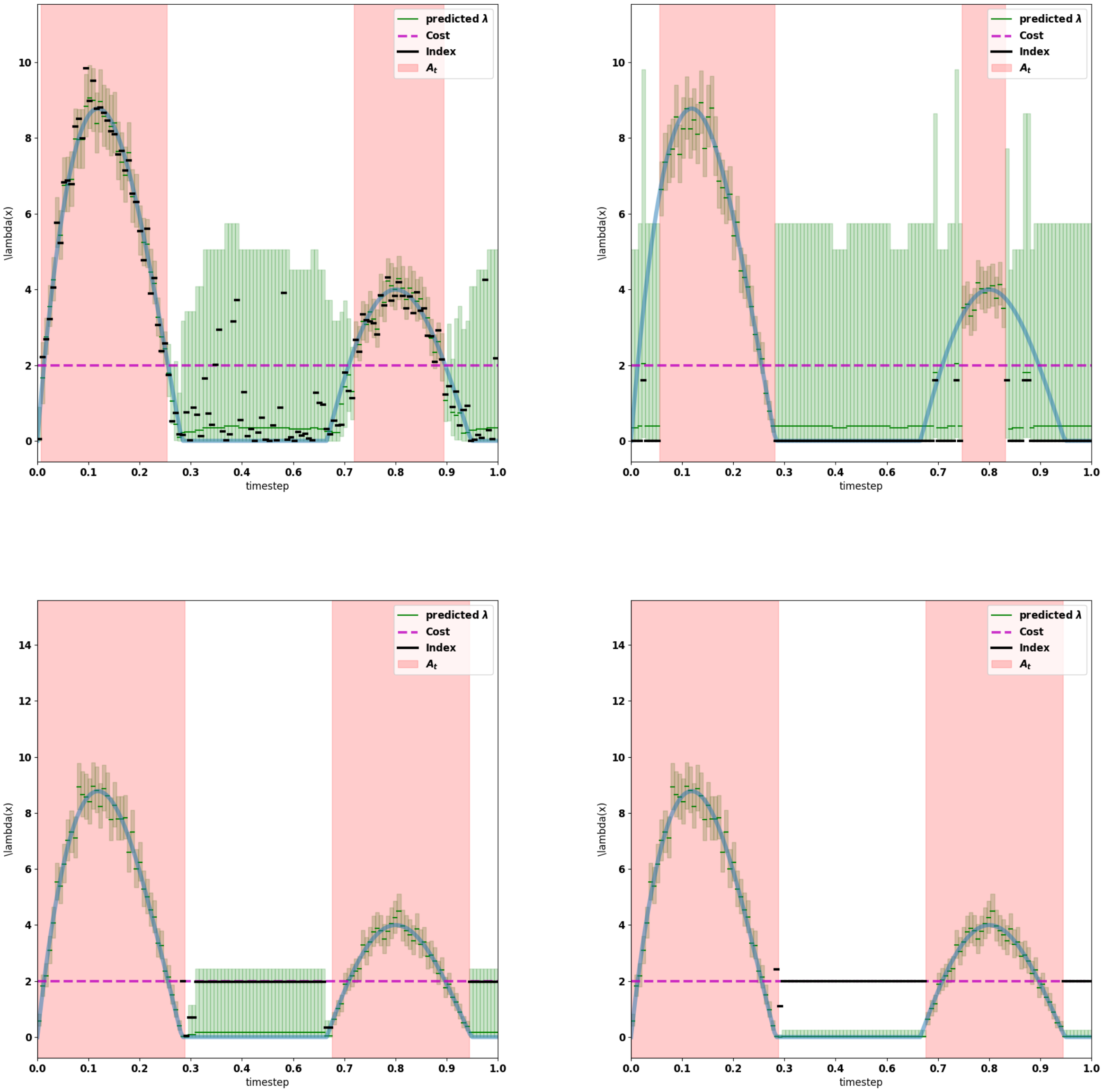


Figure 5.4.4: Posterior under different action selection strategies for the bimodal test function. The true rate function (orange), posterior mean (blue) and 95% confidence interval (green infill) is shown. Rate samples for each method are shown in black for each bin and the cost threshold is the (magenta) horizontal dashed line. The optimal action is to select two intervals $A^* = [0.013, 0.280], [0.675, 0.882]$.

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 5.4.4(c)) resulting in more often choosing a better action. In Fig. 5.4.3 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 5.4.4. 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.

5.5 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 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 and 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 and Van Zanten, 2015) suggest possible approaches. In the next chapter we shall look at extending such concentration results to meet the features of the data arising from making decisions in sequence.

5.6 Further proofs and algorithms

5.6.1 Regret bound proofs

Proof of Lemma 1

Define $A_{\mathrm{min},t} = \bigcap_{A \in \mathcal{A}_t \,:\, A^* \subseteq A} A$ as the smallest interval (or union of intervals) in $\mathcal{A}_t$ containing the optimal interval (or union of intervals). It will be easier to bound the regret of $A_{\mathrm{min},t}$ than A_t^* wrt A^*. We have, for $t \in \mathbb{N}$,

$$\begin{aligned}
\delta(A_t^*) &= r(A^*) - r(A_t^*) \\
&\leq r(A^*) - r(A_{\mathrm{min},t}) \\
&= \int_{A^*} (\lambda(x) - C)\,\mathrm{d}x - \int_{A_{\mathrm{min},t}} (\lambda(x) - C)\,\mathrm{d}x \\
&= C|A_{\mathrm{min},t} \setminus A^*| - \int_{A_{min,t} \setminus A^*} \lambda(x) dx \\
&\leq 2CU\Delta_t.
\end{aligned}$$

Here, the final inequality holds since $2\Delta_t$ bounds the difference between the lengths of subintervals of $A_{min,t}$ and A_t^*, and there are U such subintervals. Since $\Delta_t = K_t^{-1} \leq \underline{K}^{-1} T^{-1/3}$ the result follows immediately.

Proof of Lemma 5.3.2

Consider the term inside the expectation

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

$$
\begin{aligned}
&= 2\Delta_T \sum_{t=1}^{T} \sum_{k:B_{k,T}\subseteq A_t} D_{k,T}(t-1) \\
&= 2\Delta_T \sum_{t=1}^{T} \sum_{k:B_{k,T}\subseteq A_t} \frac{2\log(t)}{\Delta_T N_{k,T}(t-1)} + \sqrt{\frac{6\lambda_{\max}\log(t)}{\Delta_T N_{k,T}(t-1)}} \\
&= 2\Delta_T \sum_{t=1}^{T} \sum_{k=1}^{K_T} \mathbb{I}\{B_{k,T}\subseteq A_t\}\Bigg(\frac{2\log(t)}{\Delta_T \sum_{s=1}^{t-1}\mathbb{I}\{B_{k,T}\subseteq A_s\}} + \sqrt{\frac{6\lambda_{\max}\log(t)}{\Delta_T \sum_{s=1}^{t-1}\mathbb{I}\{B_{k,T}\subseteq A_s\}}}\Bigg) \\
&\le 2\Delta_T \sum_{k=1}^{K_T} \sum_{j=1}^{N_{k,T}} \frac{2\log(T)}{j\Delta_T} + \sqrt{\frac{6\lambda_{\max}\log(T)}{j\Delta_T}} \\
&\le 2\Delta_T K_T \Bigg(\sum_{j=1}^{T} \frac{2\log(T)}{j\Delta_T} + \sum_{j=1}^{T} \sqrt{\frac{6\lambda_{\max}\log(T)}{j\Delta_T}} \Bigg) \\
&= 4K_T \log(T)\log(T+1) + \sqrt{24\lambda_{\max} K_T \log(T)}\, T^{1/2} \\
&\le 4\overline{K}\log(T)\log(T+1)T^{1/3} + \sqrt{24\overline{K}\lambda_{\max}\log(T)}\, T^{2/3}
\end{aligned}
$$

where the penultimate line is due to $\Delta_T = K_T^{-1}$, and the final inequality is because $K_T \le \overline{K}T^{1/3}$.

Proof of Lemma 3

We have the following, which holds for any round t

$$
\begin{aligned}
&P\Big(r(A_t) \notin [L_{t,T}(A_t), U_{t,T}(A_t)]\Big) \\
&\le P\Big(r(A_t) \le L_{t,T}(A_t)\Big) + P\Big(r(A_t) \ge U_{t,T}(A_t)\Big) \\
&= P\Bigg(\sum_{k:B_{k,T}\subseteq A_t} \psi_{k,T} \le \sum_{k:B_{k,T}\subseteq A_t} \Big[\hat{\psi}_{k,T}(t-1) - D_{k,T}(t-1)\Big] \Bigg) \\
&\quad + P\Bigg(\sum_{k:B_{k,T}\subseteq A_t} \psi_{k,T} \ge \sum_{k:B_{k,T}\subseteq A_t} \Big[\hat{\psi}_{k,T}(t-1) + D_{k,T}(t-1)\Big] \Bigg)
\end{aligned}
$$

$$\begin{aligned}
&\leq \sum_{k:B_{k,T}\subseteq A_t} \left[P\Big(\psi_{k,T} - \hat{\psi}_{k,T}(t-1) \leq -D_{k,T}(t-1) \Big) + P\Big(\psi_{k,T} - \hat{\psi}_{k,T}(t-1) \geq D_{k,T}(t-1) \Big) \right] \\
&\leq \sum_{k=1}^{K_T} P\left(|\psi_{k,T} - \hat{\psi}_{k,T}(t-1)| \geq \frac{2\log(t)}{\Delta_T N_{k,T}(t-1)} + \sqrt{\frac{6\lambda_{\max}\log(t)}{\Delta_T N_{k,T}(t-1)}} \right) \\
&\leq \sum_{k=1}^{K_T} \sum_{s=1}^{t-1} P\left(|\psi_{k,T} - \hat{\psi}_{k,T}(t-1)| \geq \frac{2\log(t)}{\Delta_T N_{k,T}(t-1)} + \sqrt{\frac{6\lambda_{\max}\log(t)}{\Delta_T N_{k,T}(t-1)}} \;\middle|\; N_{k,T}(t-1) = s \right) \\
&\leq 2K_T t^{-2}.
\end{aligned}$$

The final inequality is a direct application of Lemma 1 of Grant et al. (2018) which in turn exploits Bernstein's Inequality for independent Poisson random variables.

5.6.2 Proof of optimality and efficiency of AS-IM

Proof of Theorem 1

Recall that the reward of an action is the sum of the weights of the intervals that comprise that action.

We prove the theorem by induction. Assume at least one initial I_n has a positive weight (otherwise the optimal action is to do no sensing). For $N = 1$ initial interval, which therefore has a positive weight, AS-IM simply returns this interval, which is optimal. For $N = 2$ initial intervals, with one positive weight, AS-IM returns the postitively-weighted interval, which is the optimal action. Now, assuming AS-IM returns the optimal action for $N \geq 1$, we prove that AS-IM returns the optimal action for $N + 2$ initial intervals. The result follows by induction.

Given $\mathcal{I} = \{I_n\}_{n=1}^{N+2}$, if the number of intervals in $\mathcal{I}$ with positive weight is not bigger than U, AS-IM returns all such intervals. This is the optimal action since all bins with positive reward can be covered without incurring the cost of any bins with negative reward; any other action either omits a positive-reward bin, or includes a negative-reward bin.

Similarly, consider the situation in which no interval satisfies the merging condition. Suppose that the optimal action A^* places a sensor on a sequence of intervals $I_m \cup \cdots \cup I_n$ with $n > m$. Clearly we must have $w(I_m) > 0$ and $w(I_n) > 0$ since otherwise the total weight could be increased by omitting the negatively-weighted end interval. But the fact that no interval can be merged implies that either $|w(I_{m+1})| > |w(I_m)|$ or $|w(I_{n-1})| > |w(I_n)|$. Hence removing either $I_m \cup I_{m+1}$ or $I_{n-1} \cup I_n$ from the sensor will improve the total weight. It follows that, under A^*, each sensor is allocated to a single interval, and allocating to the U highest-weight intervals, as specified by AS-IM, maximises the reward.

Now, assume that at least one interval is merged in AS-IM. Let I_n be the interval which minimises $|w(I_n)|$ and so is the first interval which is merged with its neighbours in AS-IM into a single interval $\tilde{I}_n = I_{n-1} \cup I_n \cup I_{n+1}$. Let $\tilde{A}^*$ be AS-IM's solution for the set of intervals $\tilde{\mathcal{I}} = \{I_1, \cdots, I_{n-2}, \tilde{I}_n, I_{n+2}, \cdots, I_{N+2}\}$. By induction, $\tilde{A}^*$ is optimal for $\tilde{I}$. We prove that A^*, the optimal solution for $\mathcal{I}$, is equal to $\tilde{A}^*$. To prove this, we consider different cases based on the sign of $w(I_n)$.

Case 1: $w(I_n) < 0$**.** First note that the optimal solution cannot include only one neighbour of I_n. If I_{n-1} were included but I_{n+1} were not, we could add both I_n and I_{n+1} and increase the overall weight (since I_n has the smallest absolute weight). Similarly, A^* can not include both I_{n-1} and I_{n+1} but not I_n; if so then A^* could be improved by (i) using a single sensor in place of the two that cover I_{n-1} and I_{n+1}, adding I_n to A^*, and (ii) redeploying the sensor we have saved to either split one existing sensor by removing a negative-weight I_m with $|w(I_m)| > |w(I_n)|$, or adding a new positive-weight I_m with $|w(I_m)| > |w(I_n)|$. The net outcome is an improved total weight. We have shown that A^* includes either all or none of $I_{n-1} \cup I_n \cup I_{n+1}$. Since A^* is optimal for $\mathcal{I}$, and the restriction to $\tilde{\mathcal{I}}$ does not prevent AS-IM from finding this optimal A^*, it follows that $\tilde{A}^* = A^*$.

Case 2: $w(I_n) > 0$. Under the optimal solution A^*, a sensor cannot have a negative-weighted interval as an end interval, since dropping the negative-weight interval only increases the total weight. Furthermore, a sensor cannot include I_n as an end interval of a series of intervals, since then the total weight could be improved by stopping sensing both I_n and its sensed neighbour. Thus if I_n is included in A^* then either a sensor is observing only I_n, or a single sensor observes all of $I_{n-2} \cup I_{n-1} \cup I_n \cup I_{n+1} \cup I_{n+2}$. As in Case 1, if a sensor is observing only I_n we can improve on A^* by redeploying this sensor to either sense a better interval, or stop sensing an interval which has a higher negative weight than is lost by stopping sensing I_n. So again, under A^*, I_n is either sensed with all its neighbours, or none of them are sensed. The same logic as in Case 1 ensures $\tilde{A}^* = A^*$.

Complexity: AS-IM requires sorting the N initial intervals. Noticing that there are at most N mergings, and assuming constant complexity for each merging, AS-IM offers an $O(N \log N)$ sample complexity. Since $N \leq K_t$, AS-IM has a sample complexity not bigger than $O(K_t \log K_t)$.

5.6.3 Discretisation error under linear and cubic root rates

The effect of the different rates on the unavoidable discretisation error is depicted in Figure 5.6.5. The regret for the linear rate is reduced at a faster rate than for the cubic root rate as the number of bins is increased at a much faster rate. However as we show in the main paper (Section 5.1) the other part of the regret due to error in action selection from the model forecast is much higher under the linear regret rate.

5.6.4 Baselines used in the empirical study

In the paper we have compared the TS approach other approaches which we now describe in more details.

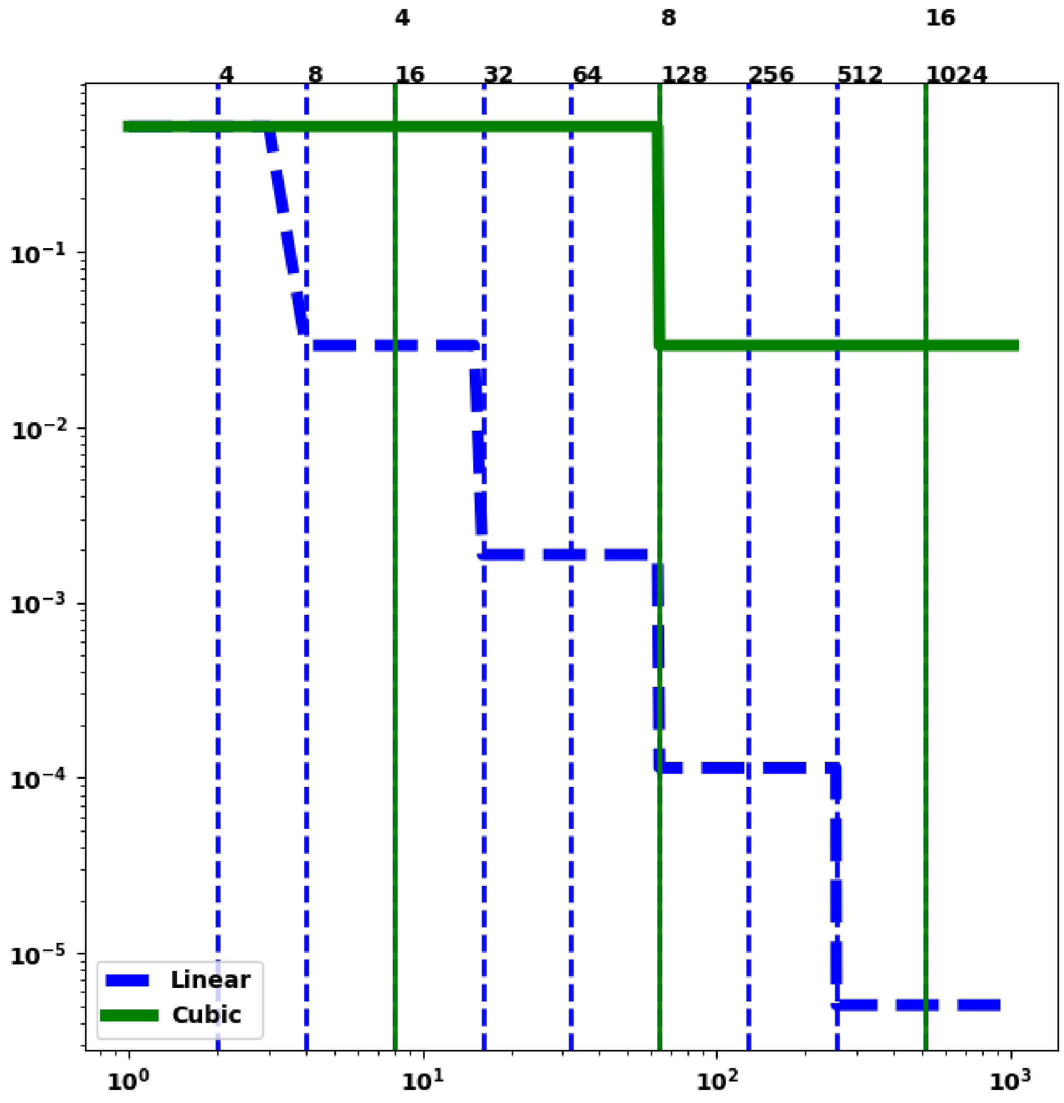


Figure 5.6.5: Instantaneous regret comparing linear and cube root rebinning rates. The vertical lines depict the rebinning times for the two different rate schedules. The time step (horizontal axis) and the regret (vertical axis) are both on a log scale. The number of bins for each rebinning rate are shown on the top horizontal axis.

1. *UCB* approach, which is based on the FP-CUCB algorithm of Grant et al. (2018) and requires the specification of an upper bound on the rate which we fix to the correct value in our experiments; in practise a conservative estimate is usually available. This is described in Algorithm 10.

Algorithm 10: UCB

```
Inputs: Upper bound $\lambda_{\max} \geq \max_{x \in [0,1]} \lambda(x)$
Initialisation Phase: For $t = 1$
    • Select $A = [0, 1]$
Iterative Phase: For $t \geq 2$
    • For each $k \in \{1, \dots, K_t\}$, evaluate $H_{k,t}(t-1)$ and $N_{k,t}(t-1)$ and calculate an index
          $\bar{\psi}_{k,t} = \frac{H_{k,t}(t-1)}{\Delta_t N_{k,t}(t-1)} + \frac{2\log(t)}{\Delta_t N_{k,t}(t-1)} + \sqrt{\frac{6\lambda_{\max}\log(t)}{\Delta_t N_{k,t}(t-1)}}.$
    • Choose an action $A_t$ that maximises $r(A)$ conditional on the true rate being given by the $\bar{\psi}_{k,t}$ values
    • Observe the events in $A_t$
```

2. A modified-UCB approach (*mUCB*) which has the same form as Algorithm 1 except $\lambda_{\max}$ is replaced with the empirical mean. Note this modification breaks the upper bound regret guarantee. The indices are :

$$\bar{\psi}_{k,t} = \hat{\psi}_{k,t}(t-1) + \frac{2\log(t)}{\Delta_t N_{k,t}(t-1)} + \sqrt{\frac{6\hat{\psi}_{k,t}(t-1)\log(t)}{\Delta_t N_{k,t}(t-1)}}, \quad k \in [K_t]$$

where $\hat{\psi}_{k,t}(t-1) = \frac{H_{k,t}(t-1)}{\Delta_t N_{k,t}(t-1)}$.

3. An ϵ-*Greedy* approach where with probability $1 - p_\epsilon$ an action A_t is selected that maximises $r(A)$ conditional on the rate being given by the empirical mean values $\hat{\psi}_{k,t}$. With probabil-

ity p_ϵ, the action is instead chosen by sampling rates $\tilde{\psi}_{k,t}$ from independent $Gamma(\alpha, \beta)$ priors. In our experiments we fix $p_\epsilon = 0.01$.

Chapter 6

Posterior Contraction Rates for Gaussian Cox Processes with Non-identically Distributed Data

A version of this Chapter has been submitted for publication as Grant, J.A., and Leslie, D.S. (2019). Posterior Contraction Rates for Gaussian Cox Processes with Non-identically Distributed Data.

6.1 On the Structure of the Remaining Material

In Chapters 4 and 5 we presented bandit algorithms for the sequential event detection problem. These algorithms were based on inference schemes which approximate the Nonhomogeneous Poisson process (NHPP) rate function λ with piecewise constant functions. In many cases the rate function may in fact be smooth. While the approach in Chapter 5 can capture smooth functions asymptotically, if we believe λ is likely to be smooth, then it is reasonable to suggest that our probabilistic model should capture this from the outset.

In Chapter 2 we described the Gaussian Cox Process (GCP) family of models which are popular models for the NHPP with smooth λ. As a GCP is a Bayesian model, a natural bandit algorithm based on its inference is a Thompson Sampling (TS) approach, which we term Gaussian Cox Process based Thompson Sampling (GCP-TS) and give as Algorithm 11. GCP-TS follows the canonical TS structure. In each round a sample is drawn from the current posterior distribution on the unknown parameter (in this case the function λ) and an action is chosen which would be optimal if that sample were the true parameter (function). The posterior belief is then updated based on the chosen action and observed data, and the process iterates.

Algorithm 11: Gaussian Cox Process based Thompson Sampling (GCP-TS)

Inputs: GCP prior distribution on λ, $\pi_0(\lambda)$, action set $\mathcal{A}$, reward function $r : \mathcal{A} \to \mathbb{R}$
Iterative Phase: For $t = 1, 2, ...$

- Sample a rate function $\tilde{\lambda}_t$ from the posterior distribution $\pi_{t-1}(\lambda)$
- Select an allocation $\mathbf{a}^*_{\bar{\boldsymbol{\lambda}}_t}$ such that $r_{\bar{\boldsymbol{\lambda}}_t}(\mathbf{a}^*_{\bar{\boldsymbol{\lambda}}_t}) = \max_{\mathbf{a}\in\mathcal{A}} r_{\bar{\boldsymbol{\lambda}}_t}(\mathbf{a})$, i.e. one maximising reward with respect to $\tilde{\lambda}_t$.
- Observe reward R_t and event locations $\{X_t^1, \ldots, X_t^{m_t}\}$, and update the posterior based on this data to $\pi_t(\lambda)$.

While the design of this approach is relatively straightforward, the task of producing a tight analysis of its (Bayesian) regret is a challenging one. Existing theoretical results on the performance of TS do not apply to the GCP-TS approach. The action space is more complex than those considered in related work, and the posterior distribution is doubly intractable, meaning that quantifying the contraction of the posterior under bandit feedback is difficult. Furthermore, for reasons of efficiency, it is likely that practical implementations of this approach would utilise variational inference - meaning samples used for decision-making would not be from the exact posterior. This is another challenging aspect which is not covered by existing results, which typically assume sampling from the exact posterior. Producing a (tight) analysis of the performance of GCP-TS would require extensions to existing work in a number of dimensions, and the results required to do this are not, in our opinion, currently known.

Therefore, in the material that follows, rather that presenting a direct analysis of the regret of GCP-TS (Algorithm 11) we derive theoretical results which give insights in to the performance of algorithms for simpler related problems and of GCP inference schemes. In this Chapter we consider the contraction of the GCP posterior using analytical tools from Bayesian nonparametrics. In Chapter 7 we extend the understanding of the performance of TS to cover its application to continuum-armed bandits (CABs) with smooth reward functions and sub-exponential reward noise. The developments of both chapters increase our understanding of the properties of decision-making and inference relevant to GCP-TS, and contribute to the broader understanding of TS and GCPs.

6.2 Introduction

This chapter differs from those preceding in that we are not developing or considering the performance of a sequential decision making algorithm. Instead we focus on the concentration properties of sophisticated inference methods. Specifically we focus GCP models introduced in Chapter 2, and derive results on their posterior contraction - which could then be used in the design and analysis of more sophisticated sequential decision making algorithms than those we have previously considered.

A GCP, as introduced in Chapter 2, is a doubly stochastic model of the Nonhomogeneous Poisson process (NHPP) where λ is modelled as a transformation of a Gaussian process (GP). In this chapter we focus on two classes of GCP, the Sigmoidal GCP (SGCP) of Adams et al. (2009) and the Quadratic GCP (QGCP) of Lloyd et al. (2015). We recall that in the SGCP the rate function is modelled as a multiple of a logistic transformation of a GP. In the QGCP the rate function is modelled as the square of a GP. Here, we are concerned with the quality of posterior inference on λ arising from these models. Specifically we are interested in the rate at which the expected posterior mass the models assign to functions far from the true λ decreases.

The GCP is a model over functions and is defined on some space of non-negative functions Λ. Given a true rate function $\lambda_0 \in \Lambda$, observed data $X_{1:n}$ collected over $n \in \mathbb{N}$ timesteps, and a relevant distance $d_n(\lambda, \lambda')$ defined for all $\lambda, \lambda' \in \Lambda$, we look for results of the form

$$\mathbb{E}_{\lambda_0}\bigg(\Pi\big(\lambda \in \Lambda : d_n(\lambda, \lambda_0) \geq \epsilon_n | X_{1:n}\big)\bigg) \leq f_n \tag{6.2.1}$$

for decreasing sequences ϵ_n, f_n, where $\Pi(\cdot|X_{1:n})$ denotes the posterior probability mass and $\mathbb{E}_{\lambda_0}$ denotes expectation with respect to the probability measure implied by λ_0. The sequences ϵ_n, f_n define the rate of posterior contraction of a model. If such a bound holds for certain $\epsilon_n, f_n \to 0$ as $n \to \infty$ this displays that the model is consistent. However, we are also interested in the order of the sequences and for which n results of the form (6.2.1) can be identified.

Asymptotic consistency results of the form

$$\mathbb{E}_{\lambda_0}\bigg(\Pi\big(\lambda \in \Lambda : d_n(\lambda, \lambda_0) \geq \epsilon_n | X_{1:n}\big)\bigg) \to 0$$

as $n \to \infty$ are prevalent in the Bayesian nonparametrics literature; for example Kirichenko and Van Zanten (2015) gives such a result for i.i.d. $X_{1:n}$ under the SGCP and a broader family of GCPs which have a smooth and bounded link function. Such asymptotic results are undoubtedly useful contributions to the understanding Bayesian models and inference, however they provide limited support to finite-time analyses thereof. We extend beyond existing results in four important regards by

1. Providing results for independent non-identically distributed (i.n.i.d.) data,
2. Providing results for the QGCP model as well as the SGCP,

3. Providing a rate on the shrinkage of the posterior mass f_n (as well as on ϵ_n), and
4. Providing results for finite values of n, not only asymptotically, and relating specific choices of hyperpriors (and parameters) to these results.

Studying i.n.i.d. data places this work in contrast to the majority of previous studies of non-parametric inference on NHPPs. However, ours is an important, practically-relevant setting. Commonly when observing point process data, the detection of events may be imperfect. This may be due to visibility conditions, unreliable signals or the fallability of observation equipment. A result of this is that while events may occur independently and according to a stationary process, the distribution of *observed* events can vary as data is collected. Equally, different subsections of a region of interest may be observed at different rates by design, as in the sequential event detection problem. Another possibility is that data collectors may be more readily able to gather data in a particular region, resources may be too costly to gather the same quality of information everywhere or multiple sub-investigations may be combined to form a joint dataset. As GCP models are typically used to model situations with underlying spatial smoothness and covariance structure, a unified analysis is still desirable, however existing contraction results only handle the setting where an entire region of interest has been observed uniformly. The results we obtain in this paper apply to the setting where (whether through design or imprecision) different rates of observation have been applied at different locations. Therefore, we present results that are more relevant to the practical settings in which GCP models are utilised than those which consider only identically distributed data.

The QGCP model has recently received attention in the literature (Lloyd et al., 2015; John and Hensman, 2018) as a model for NHPP inference, due to the ability to carry out fast and accurate inference. Previously however, there was little theoretical understanding of the model. We provide theoretical foundations for this new variant of the GCP model. This is non-trivial since the link function in the QGCP is not bounded, in contrast with the SGCP. Consequently, we find the rate of contraction to be lower for the QGCP than for the traditional SGCP.

Providing a rate f_n on the shrinkage of the posterior mass for finite values of n is also an important development. A trend in the existing literature is to focus on the asymptotic results and present that if the width of a ball around the true rate is chosen to decrease at the correct rate (with respect to the number of observations) then the probability of lying outside this ball tends to 0 as the number of observations goes to infinity. Such results are typically cleaner, and clearly demonstrate the consistency of a method, while the finite-time result can usually be extracted from the proofs provided for such results if desired. If the rate f_n is explicitly given or can be inferred, it is often only specified as holding for "sufficiently large" n. Inferring the rate f_n and determining the order of n that qualifies as sufficiently large, can be challenging to users of these results. By explicitly giving a form of f_n and quantifying the values of n (in terms of functions of the chosen hyperparameters) for which it is valid, we present a more informative set of results that are useful for end-users of this theory.

Such theory can be used in the analysis and design of sequential decision making algorithms, for the sequential event detection problem and more broadly. Understanding the rate at which the inference model contracts allows one to address an exploration-exploitation dilemma appropriately by allocation sufficiently many actions to exploratory behaviour. The work in earlier chapters has relied on assuming simpler inference models to obtain performance guarantees. Guarantees on the contraction of Cox process posteriors *with rates on the posterior mass* will be important in the design and analysis of more sophisticated approaches to these problems.

Another use for these results is in experimental design and resource planning problems. It is valuable for decision-makers to know the expected level of uncertainty in a rate function given a certain number of observations. They can then appropriately design sampling strategies or deploy resources to collect information in a way that is tailored to achieving a certain level of confidence in the inference.

6.2.1 Chapter Outline

In the remainder of this section, we discuss related work in GCPs and general contraction results for Bayesian models. In Section 6.2 we formally introduce our GCP models and notation. Section 6.3 includes all our main theoretical results and proofs, and in Section 6.4 we conclude with a discussion. Throughout we have aspired to make our assumptions transparent and demonstrate how they can be met. In Section 6.5.8 we verify that all assumed conditions can be satisfied for finite numbers of observations.

6.2.2 Related Work

The Cox process (Cox, 1955) is a class of doubly stochastic process where the rate function of an non-homogeneous Poisson process (see e.g. Moller and Waagepetersen (2003)) is modelled as another stochastic process. The Gaussian Cox process (GCP), as mentioned above, is a particular subset of this class where the rate function of the NHPP is modelled via a transformation of a Gaussian process (Williams and Rasmussen, 2006). Three main transformations have been proposed yielding three main models. Firstly, the Log-Gaussian Cox Process (LGCP) of Rathbun and Cressie (1994) and Møller et al. (1998) where λ is modelled as an exponential transformation of a GP. Secondly, the Sigmoidal-Gaussian Cox Process (SGCP) of Adams et al. (2009) where λ is modelled as a multiple of a logistic transformation of a GP. Finally, the Quadratic-Gaussian Cox Process (QGCP) of Lloyd et al. (2015) where λ is modelled as a quadratic transformation of a GP. We focus on the SGCP and QGCP models, as (for reasons discussed fully in Section 4) the LGCP model requires separate techniques to derive a contraction result.

General results for the contraction of posterior density estimates given i.i.d. data are available thanks to the seminal papers Ghosal et al. (2000) and Ghosal and Van Der Vaart (2001). The link between density estimation and function estimation is exploited in Belitser et al. (2015) to extend this work to show contraction rates for Bayesian Poisson process inference subject to appropriate prior conditions. Furthermore, Belitser et al. (2015) proposes a spline based prior satisfying these conditions. The result of Belitser et al. (2015) and GP concentration results of van der Vaart and van Zanten (2009) are used by Kirichenko and Van Zanten (2015) to show an asymptotic rate of posterior contraction for the SGCP - Kirichenko and Van Zanten (2015) is the existing work most similar to our contribution. However we are able to move beyond i.i.d. data to the independent non-identically distributed (i.n.i.d.) case, thanks to the work of Ghosal and Van Der Vaart (2007) in deriving contraction results for posterior density estimates under such data.

6.3 Model

In this section we introduce the data generating model and two prior models considered in the paper, along with other relevant notation required to understand our main results.

6.3.1 Likelihood

We consider an NHPP with bounded non-negative rate function λ_0 on $[0, 1]^d$. We suppose that n independent realisations of the NHPP $\tilde{X}_1, ..., \tilde{X}_n$ are generated. Each realisation j consists of a collection m_j of points $\{\tilde{X}_j^1, ..., \tilde{X}_j^{m_j}\} \in [0, 1]^d$. We write

$$\tilde{X}_j = \sum_{i=1}^{m_j} \delta_{\tilde{X}_j^i}, \quad j = 1, ..., n$$

where δ_x denotes the Dirac measure at x. By the definition of the NHPP model, each realisation j is distributed such that the number of points in any set $R \subseteq S$, denoted $\tilde{X}_j(R)$ follows a Poisson distribution with mean $\int_B \lambda(s)ds$. Furthermore $\tilde{X}_j(R_1), \tilde{X}_j(R_2)$ are independent if the sets $R_1, R_2 \subseteq S$ are disjoint.

Under our model, the realisations $\tilde{X}_1, ..., \tilde{X}_n$ are not directly observed. Instead, so-called *fil-*

tered realisations $X_{1:n} = X_1, ..., X_n$ are observed. The events in a filtered realisation X_j are a subset of the events in the corresponding *raw realisation* $\tilde{X}_j$. The relationship between $X_{1:n}$ and $\tilde{X}_{1:n}$ is governed by a set of *filtering functions* $\gamma_{1:n} = \gamma_1, ..., \gamma_n$.

Each filtering function $\gamma_j : [0,1]^d \to [0,1]$ evaluated at a point $s \in S$ gives the probability of observing an event in $\tilde{X}_j$ given that it has occurred at location s. Every event that occurs in $\tilde{X}_j$ is observed or not independently according to these probabilities.

By standard results, X_j is distributed according to an NHPP with rate $\gamma_i \lambda_0$. That is to say, the n filtered realisations $X_{i:n}$ are then realisations of *independent, non-identically distributed* NHPPs with rates $\gamma_1 \lambda_0, ..., \gamma_n \lambda_0$ respectively.

It follows that the likelihood of a particular set of observations $X_{1:n} = X_1, ..., X_n$ given a rate function λ and filtering functions $\gamma_{1:n}$ can be written

$$\mathcal{L}(X_{1:n}|\lambda_0, \gamma_{1:n}) = \prod_{j=1}^{n} \exp\left(\int_S \gamma_j(s)\lambda_0(s) dX_j(s) - \int_S (\gamma_j(s)\lambda_0(s) - 1) ds\right),$$

using the law of the realisation X_j as given by Proposition 6.1 of Karr (1986).

We note that the case of i.i.d. data as considered in Kirichenko and Van Zanten (2015) and Gugushvili et al. (2018) is a special case of this model, where $\gamma_j(x) = 1, \forall x \in [0,1]^d, \forall j = 1, \ldots, n$.

6.3.2 Prior Models

In this paper we consider two Bayesian models of the Poisson process where the rate function λ_0 is modelled a priori as a transformation of a Gaussian process. Under the SGCP model (Adams et al., 2009), the true rate function is modelled a priori as

$$\lambda(s) = \lambda^* \sigma(g(s)) = \lambda^* (1 + e^{-g(s)})^{-1} \quad s \in S \tag{6.3.2}$$

where $\lambda^* > 0$ is a scalar hyperparameter endowed with an independent Gamma prior and g is a zero-mean GP. The sigmoidal transformation σ is bounded in $[0, 1]$ so the hyperparameter λ^* models the maximum of the rate function, $||\lambda_0||_\infty$. The QGCP model (Lloyd et al., 2015) uses a more straightforward transformation. The rate function is modelled a priori as

$$\lambda(s) = (g(s))^2 \quad s \in S \tag{6.3.3}$$

where again, g is a GP.

For both models, we specify certain additional properties of the GP to support our subsequent analyses. These conditions are standard in the posterior contraction literature (van der Vaart and van Zanten, 2008, 2009; Kirichenko and Van Zanten, 2015). We require that the covariance kernel f of the GP g, can be given in its spectral form by

$$Ef(s)f(s') = \int e^{-i<\xi, l(s'-s)>} \mu(\xi) d\xi, \quad s, s' \in S. \tag{6.3.4}$$

Here $l > 0$ is an (inverse) length scale parameter and μ is a spectral density on $\mathbb{R}^d$ such that the map $a \mapsto \mu(a\xi)$ on $(0, \infty)$ is decreasing for every $\xi \in \mathbb{R}^d$ and that satisfies

$$\int e^{\delta||\xi||} \mu(d\xi) < \infty$$

for some $\delta > 0$. Condition (6.3.4) is satisfied, for instance, by the squared exponential covariance function

$$Ef(s)f(s') = e^{-l^2||s-s'||^2}, \quad s, s' \in S$$

since it corresponds to a centred Gaussian spectral density.

The length scale parameter should have a prior π_l on $[0, \infty)$ which satisfies

$$C_1 x^{q_1} \exp(-D_1 x^d \log^{q_2} x) \leq \pi_l(x) \leq C_2 x^{q_1} \exp(-D_2 x^d \log^{q_2} x) \tag{6.3.5}$$

for positive constants C_1, C_2, D_1, D_2, non-negative constants q_1, q_2, and every sufficiently large $x > 0$. In particular if l^d is endowed with a Gamma(a, b) prior, then

$$\pi_l(x) = \frac{b^a d}{\Gamma(a)} x^{da-1} \exp(-bx^d)$$

for $x > 0$, and thus (6.3.5) is satisfied with $C_1 = C_2 = \frac{b^a d}{\Gamma(a)}$, $D_1 = D_2 = b$, $q_1 = da - 1$, and $q_2 = 0$. We will assume a Gamma prior on l^d in the remainder of the paper for ease of analysis and presentation, but note that similar results are obtainable for other choices.

Finally, for the SGCP model we assume a positive, continuous prior $p_{\lambda*}$ for λ^* on $[0, \infty)$ satisfying

$$\int_{\lambda'}^{\infty} p_{\lambda^*}(x) dx \leq C_0 e^{-c_0 (\lambda')^\kappa} \tag{6.3.6}$$

for some constants $c_0, C_0, \kappa > 0$ and all $\lambda' > 0$. This condition is satisfied by, for instance, choosing a Gamma prior on λ^*.

6.3.3 Additional Notation

In the following section, we will derive results on the posterior distribution of $\lambda_0 | X_{1:n}$ under the two models. We will denote the prior distributions as $\Pi(\cdot)$ and the posteriors as $\Pi(\cdot | X_{1:n})$. Certain results will be valid for the class of all continuous functions on $[0, 1]^d$, which will be denoted $\mathcal{C}([0, 1]^d)$, and others will hold for the class of all α-Hölder continuous functions on $[0, 1]^d$ denoted $\mathcal{C}^\alpha[0, 1]^d$.

Contraction results will inevitably depend on the particular filtering functions $\gamma_{1:n}$, therefore it is convenient to define versions of standard distances averaged with respect to $\gamma_{1:n}$. We have the averaged uniform norm

$$\Gamma_{n,\infty}(\lambda,\lambda') = \frac{1}{n}\sum_{i=1}^{n} ||\lambda\gamma_i - \lambda'\gamma_i||_\infty = \frac{1}{n}\sum_{i=1}^{n} \sup_{x\in[0,1]^d} |\lambda(x)\gamma_i(x) - \lambda'(x)\gamma_i(x)|,$$

averaged L_2 norm

$$\Gamma_{n,2}(\lambda,\lambda') = \frac{1}{n}\sum_{i=1}^{n} ||\lambda\gamma_i - \lambda'\gamma_i||_2 = \frac{1}{n}\sum_{i=1}^{n} \int_{[0,1]^d} (\lambda(x)\gamma_i(x) - \lambda'(x)\gamma_i(x))^2 dx,$$

and square rooted averaged L_2 norm

$$\Gamma_{n,2}^{1/2}(\lambda,\lambda') = \frac{1}{n}\sum_{i=1}^{n} ||\sqrt{\lambda\gamma_i} - \sqrt{\lambda'\gamma_i}||_2 = \frac{1}{n}\sum_{i=1}^{n} \int_{[0,1]^d} (\sqrt{\lambda(x)\gamma_i(x)} - \sqrt{\lambda'(x)\gamma_i(x)})^2 dx,$$

for rate functions $\lambda, \lambda \in \mathcal{C}([0,1]^d)$. Using these definitions we can guarantee a rate of convergence appropriate to the level of filtering.

Finally let $N(\epsilon, \mathcal{S}, l)$ denote the ϵ-covering number of a set $\mathcal{S}$ with respect to distance l.

6.4 Posterior Contraction Results

In this section we state our results on the finite-time contraction of the posterior of the QGCP and SGCP models. Our results assert that given n realisations of the NHPP, the expected posterior mass concentrated on functions outside a Hellinger-like ball of a given width will not exceed a transformation of the width of the ball. Theorem 6.4.1 gives the result for the QGCP, and Theorem 6.4.2 for the SGCP.

Theorem 6.4.1. *Suppose that $\lambda_0 \in \mathcal{C}^\alpha([0,1]^d)$ for some $\alpha > 0$ and $\lambda_0 : [0,1]^d \to [\lambda_{0,min}, \infty)$. Suppose that the filtering functions $\gamma_{1:n}$ are known. Then for all sufficiently large $M, n > 0$ the* *posterior under the QGCP satisfies*

$$E_{\lambda_0}\Big(\Pi\Big(\lambda : \frac{1}{n}\sum_{i=1}^{n} ||\sqrt{\lambda\gamma_i} - \sqrt{\lambda_0\gamma_i}||_2 \geq \sqrt{2}M\epsilon_n|X_{1:n}\Big)\Big) = \tilde{o}\big(n^{\frac{-d}{4\alpha+d}}\big) \tag{6.4.7}$$

for $\epsilon_n = 2\sqrt{||\lambda_0|_\infty n^{-\alpha/(4\alpha+d)}(\log(n))^{\rho+d+1} + n^{-2\alpha/(4\alpha+d)}(\log(n))^{2\rho+2d+2}}$ *with* $\rho = \frac{1+d}{4+d/\alpha}$.

Theorem 6.4.2. *Suppose that* $\lambda_0 \in \mathcal{C}^\alpha([0,1]^d)$ *for some* $\alpha > 0$ *and* $\lambda_0 : [0,1]^d \to [\lambda_{0,min}, \lambda_{0,max}]$. *Suppose that the filtering functions* $\gamma_{1:n}$ *are known. Then for all sufficiently large* $M, n > 0$ *the posterior under the SGCP satisfies*

$$E_{\lambda_0}\Big(\Pi\Big(\lambda : \frac{1}{n}\sum_{i=1}^{n} ||\sqrt{\lambda\gamma_i} - \sqrt{\lambda_0\gamma_i}||_2 \geq \sqrt{2}M\epsilon_n|X_{1:n}\Big)\Big) = \tilde{o}\big(n^{\frac{-d}{2\alpha+d}}\big) \tag{6.4.8}$$

for $\epsilon_n = n^{-\alpha/(2\alpha+d)}(\log(n))^{\rho+d+1}$ *with* $\rho = \frac{1+d}{2+d/\alpha}$.

In each case analytical results free from "little-o" notation and a specific value for the "sufficiently large" conditions on M and n are given in the proofs in Sections 6.4.2 and 6.4.3.

The key difference between the two results is that for the QGCP we can only guarantee convergence on larger ball widths ϵ_n and at a slower rate f_n. Notice that under the QGCP the ball width is $\tilde{o}(n^{-\alpha/(4\alpha+d)})$ and the contraction rate is $\tilde{o}(n^{-d/(4\alpha+d)})$, whereas for the SGCP the ball width is $\tilde{o}(n^{-\alpha/(2\alpha+d)})$ and the contraction rate is $\tilde{o}(n^{-d/(2\alpha+d)})$.

In the simplest setting where $\lambda_0 \in \mathcal{C}^1([0,1])$ - i.e. where we consider Lipschitz smooth functions on $d = 1$ - this means we have a contraction rate of $\tilde{o}(n^{-1/5})$ on balls of width $\tilde{o}(n^{-1/5})$ for the QGCP and a contraction rate of $\tilde{o}(n^{-1/3})$ on balls of width $\tilde{o}(n^{-1/3})$ for the SGCP. The result on the SGCP is therefore tighter in two senses, we are able to say that the posterior mass shrinks quicker than for the QGCP and on the probability of being in a larger subspace (since the ball width ϵ_n is smaller, the area outside the ball is larger).

The different results arise as a consequence of the different transformation functions. For the posterior to contract at a given rate, we must demonstrate that the prior model satisfies certain properties related to this rate. Both models are built upon a GP g, and by considering the properties of g, we can verify that the SGCP and QGCP prior models meet the necessary conditions.

The results of van der Vaart and van Zanten (2009) demonstrate that for g as described in Section 6.3, relevant properties of g can be shown, i.e. that the prior mass g assigns to certain parts of the function space is bounded by sequences of a known form. It follows that appropriate transformations of these sequences can be used to show that the SGCP and QGCP priors also assign their prior mass across the function space in the required manner. The transformed sequences give rise to our ball widths ϵ_n which in turn influence the contraction rate. Since the SGCP and QGCP involve different transformations of g, we also require different transformations of the sequences for which desirable properties of g hold, and therefore different results are obtained.

More informally, the issue is that by applying a quadratic transformation to the GP over a logistic one, prior mass is dispersed more across the function space and the resulting posterior takes longer to contract around the true λ_0.

6.4.1 Contraction of NHPP models under general priors

Before we prove Theorems 6.4.1 and 6.4.2 we introduce a third result which gives a sufficient set of conditions on prior models to attain posterior contraction at a known rate under i.n.i.d. observations. Theorem 6.4.3 extends Theorem 1 of Ghosal and Van Der Vaart (2007) to apply to for Poisson processes. The extension is in the same manner as the result of Belitser et al. (2015) extends Theorem 2 of Ghosal and Van Der Vaart (2001) for i.i.d. Poisson process realisations. In addition we retain the rate on the shrinkage of the posterior mass, as well as on the ball width, unlike these earlier papers.

Theorem 6.4.3. *Assume that $\lambda_0 : [0, 1]^d \to [\lambda_{0,min}, \infty)$ and that filtering functions $\gamma_{1:n}$ are known. Suppose that for positive sequences δ, $\bar{\delta}_n \to 0$, such that $n \min(\delta_n, \bar{\delta}_n)^2 \to \infty$ as $n \to \infty$, it* *holds that there exist subsets* $\Lambda_n \subset \mathcal{C}(S)$*, some* $n_0 \in \mathbb{N}$*, and constants* $c_1, c_2, c_3 > 0$*,* $c_4 > 1$*, and* $c_5 > c_2 + 2$ *such that*

$$\Pi_n\Big(\lambda : \Gamma_{n,\infty}(\lambda, \lambda_0) \leq \delta_n\Big) \geq c_1 e^{-c_2 n \delta_n^2} \tag{6.4.9}$$

$$\sup_{\delta > \bar{\delta}_n} \log N\left(\frac{\delta}{36\sqrt{2}}, \sqrt{\Lambda_{n,\delta}}, \Gamma_{n,2}\right) \leq c_3 n \bar{\delta}_n^2 \tag{6.4.10}$$

$$\Pi_n(\Lambda \setminus \Lambda_n) \leq c_4 e^{-c_5 n \delta_n^2}. \tag{6.4.11}$$

for all $n \geq n_0$ *where* $\Lambda_{n,\epsilon} = \Big\{\lambda \in \Lambda_n : h_n(p_\lambda, p_{\lambda_0}) \leq \epsilon\Big\}$*, and* $h_n(p_\lambda, p_{\lambda_0})$*, is given by*

$$h_n^2(p_\lambda, p_{\lambda'}) = \frac{1}{n}\sum_{i=1}^{n} 2\left(1 - E_{\lambda\gamma_i}\left(\sqrt{\frac{p(X^{(i)}|\lambda, \gamma_i)}{p(X^{(i)}|\lambda', \gamma_i)}}\right)\right).$$

Then for $\epsilon_n = \max(\delta_n, \bar{\delta}_n)$ *and any* $C > 0, J \geq 1, M \geq 2$*,*

$$E_{\lambda_0}\left[\Pi_n\left(\lambda : \Gamma_{n,2}^{1/2}(\lambda, \lambda_0) \geq \sqrt{2} J M \epsilon_n | X_{1:n}\right)\right] \leq \frac{1}{C^2 n \epsilon_n^2} + e^{-M^2 n \epsilon_n^2/4} + 2e^{-(M^2/2 - c_3) n \epsilon_n^2} + \frac{2}{c_1} e^{-(c_2 M^2 J^2/4 - C - 1) n \epsilon_n^2} \tag{6.4.12}$$

for $n \geq \max(n_0, n_1, n_2, n_3)$ *where* $n_1 = \arg\min\{n : \epsilon_n \leq \lambda_{min}\}$*,* $n_2 = \arg\min\{n : \epsilon_n \leq \frac{1}{\sqrt{2}M}\}$*, and* $n_3 = \arg\min\{n : e^{-n\epsilon_n^2 K M^2/4} \leq 1/2\}$*.*

We prove this theorem in Section 6.4.4. This establishes that given the prior model satisfies certain conditions, the expected posterior mass assigned to rate functions outside an order ϵ_n width ball around λ_0 (measured with respect to an averaged L_2 distance) decreases at rate $o((n\epsilon_n^2)^{-1})$ for sufficiently large n. The conditions on the prior model are standard and are inherited from the conditions of Theorem 4 of Ghosal and Van Der Vaart (2007) required to show posterior contraction in a density estimation setting. Condition (6.4.9), the *prior mass condition*, ensures that a sufficient proportion of the prior mass is assigned to functions close to λ_0. Condition (6.4.10), the *entropy condition*, and condition (6.4.11), the *remaining mass condition*, together prescribe that there exist subsets of the function space such that the entropy of these subsets is not too large, but the probability of lying outside these is also small.

Equipped with this general result, we are now in a position to prove Theorems 6.4.1 and 6.4.2 by demonstrating that the QGCP and SGCP models meet conditions (6.4.9), (6.4.10), and (6.4.11).

6.4.2 Proof of Theorem 6.4.1: Contraction of the QGCP model

To prove Theorem 6.4.1 we verify that the QGCP model described in Section 6.3 meets the conditions of Theorem 6.4.3. The following sections handles each condition in turn. Throughout we have

$$\delta_n = 2\sqrt{||\lambda_0||_\infty} n^{-\alpha/(4\alpha+d)} \log^{\rho}(n) + n^{-2\alpha/(4\alpha+d)} \log^{2\rho}(n), \tag{6.4.13}$$

$$\bar{\delta}_n = 2\sqrt{||\lambda_0||_\infty} n^{-\alpha/(4\alpha+d)} \log^{\rho+d+1}(n) + n^{-2\alpha/(4\alpha+d)} \log^{2\rho+2d+2}(n). \tag{6.4.14}$$

Prior Mass Condition

The first condition, the so-called prior mass condition (6.4.9) does not rely on the existence of particular subsets Λ_n, and can be verified by the following lemma, which we prove in Section 6.6.1.

Lemma 6.4.4. *If $\lambda_0 = g_0^2$ where $g_0 \in \mathcal{C}^\alpha([0,1]^d)$ for some $\alpha > 0$ then under the QGCP model there exist constants $c_1, c_2 > 0$ for δ_n as defined in* (6.4.13) *such that the prior satisfies*

$$\Pi(\lambda : ||\lambda - \lambda_0||_\infty \leq \delta_n) \geq c_1 e^{-c_2 n \delta_n^2}$$

for all $n \geq 3$.

Then consider that since $\gamma_i \in [0,1]$ for all $i = 1, ..., n$,

$$\Gamma_{n,\infty}(\lambda, \lambda_0) = \frac{1}{n}\sum_{i=1}^{n} ||\lambda\gamma_i - \lambda_0\gamma_i||_\infty \leq ||\lambda - \lambda_0||_\infty. \tag{6.4.15}$$

Thus, by Lemma 6.4.4 we have that there exist constants $c_1, c_2 > 0$ such that

$$\Pi_n\Big(\lambda : \Gamma_{n,\infty}(\lambda, \lambda_0) \leq \delta_n\Big) \geq \Pi_n(\lambda : ||\lambda - \lambda_0||_\infty \leq \delta_n) \geq c_1 e^{-c_2 n \delta_n^2},$$

satisfying condition (6.4.9).

Definition of Sieves

We now define the subsets Λ_n for which the QGCP satisfies the constraints of Theorem 6.4.3. Let,

$$\Lambda_n = (\mathcal{G}_n)^2 \tag{6.4.16}$$

where

$$\mathcal{G}_n = \left[\beta_n\sqrt{\frac{\zeta_n}{\chi_n}}\mathbb{H}_1^{\zeta_n} + \kappa_n\mathbb{B}_1\right] \cup \left[\bigcup_{a \leq \chi_n}(\beta_n\mathbb{H}_1^a) + \kappa_n\mathbb{B}_1\right], \tag{6.4.17}$$

$\mathbb{B}_1$ is the unit ball in $\mathcal{C}([0,1]^d)$ with respect to the uniform norm, and $\mathbb{H}_1^l$ is the unit ball of the RKHS $\mathbb{H}^l$ of the GP g with covariance as given in (6.3.4). We define the sequences involved as follows,

$$\begin{aligned}
\zeta_n &= L_2 n^{\frac{1}{d}\frac{2\alpha+d}{4\alpha+d}}(\log(n))^{2\rho/d} + L_3 n^{\frac{1}{d}\frac{\alpha+d}{4\alpha+d}}(\log(n))^{3\rho/d} + L_4 n^{\frac{1}{d}\frac{d}{4\alpha+d}}(\log(n))^{4\rho/d} \\
\beta_n &= L_5 n^{\frac{1}{2}\frac{2\alpha+d}{4\alpha+d}}(\log(n))^{2\rho+\frac{d+1}{2}} + L_6 n^{\frac{1}{2}\frac{\alpha+d}{4\alpha+d}}(\log(n))^{3\rho+\frac{d+1}{2}} + L_7 n^{\frac{1}{2}\frac{d}{4\alpha+d}}(\log(n))^{4\rho+\frac{d+1}{2}} \\
\kappa_n &= \frac{1}{3}\bar{\delta}_n, \qquad\qquad \chi_n = \frac{\bar{\delta}_n}{6\tau\sqrt{d}\beta_n},
\end{aligned}$$

for constants

$$L_2 > (8c_5||\lambda_0||_\infty)/D_1, \qquad L_3 > (8c_5\sqrt{||\lambda_0||_\infty})/D_1, \qquad L_4 > 2c_5/D_1$$

such that $L_2 + L_3 + L_4 > \max(A, e)$ and

$$L_5 \geq \max\left(\sqrt{\frac{16K_5L_2^d\mathcal{K}_1^{1+d}}{\log^{2\rho}(3)}}, \sqrt{32||\lambda_0||_\infty}c_5, L_2^{1/3}\Big(\frac{8\max(1,\sqrt{||\lambda_0||_\infty})}{(3/36\sqrt{2})^{3/2}d^{1/4}\sqrt{2\tau}}\Big)^{2/3}\right)$$
$$L_6 \geq \max\left(\sqrt{\frac{16K_5L_3^d\mathcal{K}_1^{1+d}}{\log^{3\rho}(3)}}, \sqrt{32\sqrt{||\lambda_0||_\infty}c_5}\right), \quad L_7 \geq \max\left(\sqrt{\frac{16K_5L_4^d\mathcal{K}_1^{1+d}}{\log^{4\rho}(3)}}, \sqrt{8c_5}\right),$$

such that $L_5+L_6+L_7 > \frac{4L_1\max(1,\sqrt{||\lambda_0||_\infty})}{3\sqrt{||\mu||}}$, and $L_2L_5^3 > \left(\frac{8\max(1,||g_0||_\infty)}{(3/L_1)^{3/2}d^{1/4}\sqrt{2\tau}}\right)^2$ where $L_1 = 1/(36\sqrt{2})$ and $\mathcal{K}_1 = \log\big(\frac{3(L_2+L_3+L_4)}{\min(1,2\sqrt{||\lambda_0|_\infty})}\big) + \frac{2\alpha d+2\alpha+d}{4\alpha d+d^2} + (4\rho - \rho/d - d - 1)$.

The definition of $\mathcal{G}_n$ and these sequences is important as it allows general GP results of van der Vaart and van Zanten (2009) to be applied. The extensive conditions on the constants are important to ensure that the results hold for finite values of n.

Entropy Condition

The following lemma allows us to verify condition (6.4.10) which stipulates that the log entropy of the subsets Λ_n is not too large. The proof of this lemma is provided in Section 6.6.3. In particular it exploits an existing bound on the covering number of $\mathcal{G}_n$ with respect to the infinity norm from van der Vaart and van Zanten (2009).

Lemma 6.4.5. *For Λ_n defined as in* (6.4.16)*, a constant $L_1 > 0$, and $\bar{\delta}_n$ as defined in* (6.4.14)*, there exists a constant $c_3 > 0$ such that*

$$\log N(L_1\bar{\delta}_n, \sqrt{\Lambda_n}, ||\cdot||_2) \leq c_3 n\bar{\delta}_n^2,$$

for all n such that

$$4||\lambda_0||_\infty \log^{2d+2-2\rho}(n) \geq \frac{m\sum_{i=2}^4 L_i^d}{2^{1+d}}\Bigg(\log\Big(\frac{27\tau\sqrt{d}(\sum_{i=5}^7 L_i)^3\sum_{i=2}^4 L_i}{4||\lambda_0||_\infty^{3/2}}\Big) \\ + \Big(4+\frac{12+d+d^2}{8\alpha d+2d^2}\Big)\log(n)\Bigg)^{1+d}$$

and

$$2\log\left(\frac{6\sqrt{||\mu||}(L_5+L_6+L_7)}{2L_1\sqrt{||\lambda_0||_\infty}+L_1}\right) \leq 4||\lambda_0||_\infty n^{(4\alpha+2d)/(8\alpha+2d)}\log^{2\rho+2d+2}(n) \\ -\log\Bigg(n^{(6\alpha+d)/(4\alpha+d)}\log^{6\rho}(n)\Bigg).$$

To apply Lemma 6.4.5, notice that $\frac{1}{n}\sum_{i=1}^n ||\gamma_i \times \cdot||_2 \leq ||\cdot||_2$ since the functions $\gamma_i \in [0,1]$ for all $i = 1, ..., n$. It follows that

$$N\left(\frac{\delta}{36\sqrt{2}}, \sqrt{\Lambda_{n,\delta}}, \frac{1}{n}\sum_{i=1}^n ||\gamma_i \times \cdot||_2\right) \leq N\left(\frac{\delta}{36\sqrt{2}}, \sqrt{\Lambda_{n,\delta}}, ||\cdot||_2\right) \leq N\left(\frac{\delta}{36\sqrt{2}}, \sqrt{\Lambda_n}, ||\cdot||_2\right). \tag{6.4.18}$$

As any ϵ-covering number is decreasing in ϵ, it follows by Lemma 6.4.5 that

$$\sup_{\delta>\bar{\delta}_n} \log N\left(\frac{\delta}{36\sqrt{2}}, \sqrt{\Lambda_{n,\delta}}, \frac{1}{n}\sum_{i=1}^n ||\gamma_i \times \cdot||_2\right) \leq c_3 n\bar{\delta}_n^2.$$

Thus we have satisfied constraint (6.4.10).

Remaining Mass Condition

Finally, Lemma 6.4.6 below is sufficient to validate condition (6.4.11) directly. Its proof is given in Section 6.6.4.

Lemma 6.4.6. *Under the QGCP model, with Λ_n as defined in* (6.4.16)*, and δ_n as defined in* (6.4.13) *there exist constants $c_4 > 0, c_5 \geq c_2 + 4$ such that*

$$\Pi(\lambda : \lambda \notin \Lambda_n) \leq c_4 e^{-c_5 n \delta_n^2},$$

for all n such that

$$n^{(2\alpha+d)/(4\alpha+d)} \log^{2\rho}(n) \geq \frac{q_1}{4c_5 ||g_0||_\infty^2} \log\Big((L_2 + L_3 + L_4) n^{(2\alpha+d)/(4\alpha d+d^2)} \log^{4\rho/d}(n)\Big).$$

Conlcuding the Proof

By Lemmas 6.4.4, 6.4.5, and 6.4.6 and the definitions of δ_n and $\bar{\delta}_n$ therein we have that the conditions of Theorem 6.4.3 are satisfied. Thus, for the QGCP model

$$E_{\lambda_0}\left[\Pi_n\left(\lambda : \Gamma_{n,2}^{1/2}(\lambda, \lambda_0) \geq \sqrt{2} J M \epsilon_n | \tilde{X}_{1:n}\right)\right] \leq \frac{1}{C^2 n \epsilon_n^2} + e^{-M^2 n \epsilon_n^2/4} \\ + 2e^{-(M^2/2 - c_3) n \epsilon_n^2} + \frac{2}{c_1} e^{-(c_2 M^2 J^2/4 - C - 1) n \epsilon_n^2}$$

holds with $\epsilon_n = \max(\delta_n, \bar{\delta}_n) = \bar{\delta}_n$ for any $C > 0, J \geq 1, M \geq 2$. Specific values of the remaining constants can be extracted from Lemmas 6.4.4, 6.4.5, 6.4.6, and 6.6.1.

Then, so long as M and J are sufficiently large, the second, third and fourth terms on the RHS of equation (6.4.12) decay much more quickly than the first and the bound is $\tilde{o}(n^{\frac{-d}{4\alpha+d}})$ as stated, for all n such that the conditions of Theorem 6.4.3 and Lemmas 6.4.5 and 6.4.6 are met. □

6.4.3 Proof of Theorem 6.4.2: Contraction of the SGCP model

Like the proof of Theorem 6.4.1, the proof of Theorem 6.4.2 relies on demonstrating the the SGCP model described in Section 6.3 meets the conditions of Theorem 6.4.3. In Kirichenko and

Van Zanten (2015) the conditions of Theorem 1 of Belitser et al. (2015) - the asymptotic and i.i.d. analogue of Theorem 6.4.3 - are verified for the SGCP model. However certain asymptotic arguments are used in said proof. In the following sections we handle each condition of Theorem 6.4.3 in turn under our setting. Throughout we have

$$\delta_n = n^{-\alpha/(2\alpha+d)}(\log(n))^{(1+d)/(2+d/\alpha)} \tag{6.4.19}$$

$$\bar{\delta}_n = n^{-\alpha/(2\alpha+d)}(\log(n))^{(1+d)/(2+d/\alpha)+d+1} \tag{6.4.20}$$

Prior Mass Condition

For the SGCP model, the prior mass condition (6.4.9) can be verified by the following lemma which we prove in Section 6.6.5.

Lemma 6.4.7. *If $\lambda_0 = ||\lambda_0||_\infty \sigma(g_0)$ where $g_0 \in \mathcal{C}^\alpha([0,1]^d)$ for some $\alpha > 0$ then under the SGCP model there exist constants c_1, c_2 for δ_n as defined in* (6.4.19) *such that the prior satisfies*

$$\Pi(\lambda : ||\lambda - \lambda_0||_\infty \leq \delta_n) \geq c_1 e^{-c_2 n \delta_n^2}$$

for all $n \geq 3$.

Then, by (6.4.15) and Lemma 6.4.7 we have that there exist constants $c_1, c_2 > 0$ such that

$$\Pi\bigg(\lambda : \Gamma_{n,\infty}(\lambda, \lambda_0) \leq \delta_n\bigg) \geq c_1 e^{-c_2 n \delta_n^2}$$

and we have shown condition (6.4.9) is satisfied under the SGCP model.

Definition of Sieves

We now define the sets Λ_n such that the remaining coniditions hold. Consider,

$$\Lambda_n = \bigcup_{\lambda \le \lambda_n} \lambda\sigma(\mathcal{G}_n) \tag{6.4.21}$$

where

$$\mathcal{G}_n = \left[\beta_n \sqrt{\frac{\zeta_n}{\chi_n}} \mathbb{H}_1^{\zeta_n} + \kappa_n \mathbb{B}_1\right] \cup \left[\bigcup_{a \le \chi_n} (\beta_n \mathbb{H}_1^a) + \kappa_n \mathbb{B}_1\right], \tag{6.4.22}$$

$\mathbb{B}_1$ is the unit ball in $\mathcal{C}([0,1]^d)$ with respect to the uniform norm, and $\mathbb{H}_1^l$ is the unit ball of the RKHS $\mathbb{H}^l$ of the GP g with covariance as given in (6.3.4). Though the structure of the sieves $\mathcal{G}_n$ is the same as in the proof of Theorem 6.4.1, the sequences are defined differently, as below

$$\zeta_n = L_8 n^{\frac{1}{2\alpha+d}} (\log(n))^{2\rho/d}, \quad \beta_n = L_9 n^{\frac{d}{2(2\alpha+d)}} (\log(n))^{d+1+2\rho},$$
$$\lambda_n = L_{10} n^{\frac{d}{\kappa(2\alpha+d)}} (\log(n))^{4\rho/\kappa}, \quad \kappa_n = \frac{1}{3}\bar{\delta}_n, \quad \chi_n = \frac{\kappa_n}{2\tau\sqrt{d}\beta_n},$$

for constants

$$L_8 > \max\left(A, 1, \left(\frac{2c_5}{D_1}\right)^{1/d}\right), \quad L_9 \ge \sqrt{8c_5}, \quad L_{10} > \left(\frac{c_5}{c_0}\right)^{1/\rho}$$

such that

$$L_8 L_9^3 L_{10}^{3/2} > \frac{2}{(6cL_1)^{3/2}\tau\sqrt{d}}, \quad L_9 L_{10}^{1/2} > \frac{1}{6cL_1\sqrt{||\mu||}},$$

where $L_1 = 1/(36\sqrt{2})$, $c = 2^{-5/2}$, and κ is a positive constant.

Entropy Condition

The following lemma will allow us to verify condition (6.4.10). We prove it in Section 6.6.6.

Lemma 6.4.8. *For Λ_n as defined in* (6.4.21)*, a constant $L_1 > 0$ and $\bar{\delta}_n$ as defined in* (6.4.20)*, there exists a constant $c_3 > 0$ such that*

$$\log N(L_1\bar{\delta}_n, \sqrt{\Lambda_n}, ||\cdot||_2) \leq c_3 n\bar{\delta}_n^2,$$

for all n such that

$$\log^{2d+2}(n) > K_1 L_8^d\Bigg(\log(\sqrt{2\tau L_8 L_9^3} L_{10}^{3/4} d^{1/4}) + \frac{\kappa(6d+6\alpha+2)+3d}{4\kappa(2\alpha+d)}\log(n) + \log\left(\log^{3\rho/2+3\rho/\kappa+\rho/d-d-1}(n)\right)\Bigg)^{1+d} \tag{6.4.23}$$

$$n^{\frac{d}{2\alpha+d}} > \max\left(2\log(12cL_1L_9L_{10}^{1/2}), 2\log(L_1L_{10}^{1/2})\right) + 1. \tag{6.4.24}$$

Then, as in the proof of Theorem 6.4.1, using (6.4.18) and Lemma 6.4.8 we have

$$\sup_{\delta>\bar{\delta}_n} \log N\left(\frac{\delta}{36\sqrt{2}}, \sqrt{\Lambda_{n,\delta}}, \frac{1}{n}\sum_{i=1}^{n}||\gamma_i \times \cdot||_2\right) \leq c_3 n\bar{\delta}_n^2,$$

verifying condition (6.4.10).

Remaining Mass Condition

Finally, Lemma 6.4.9 below is sufficient to validate condition (6.4.11) directly. Its proof is given in Section 6.6.7.

Lemma 6.4.9. *Under the SGCP model, with Λ_n as defined in* (6.4.21)*, and δ_n as defined in* (6.4.20) *there exist constants $c_4 > 0$, $c_5 \geq c_2 + 4$ such that*

$$\Pi(\lambda : \lambda \notin \Lambda_n) \leq c_4 e^{-c_5 n \delta_n^2}$$

for all n such that

$$\log^{2\rho}(n) > \frac{16 K_5 D_1 L_8^d}{L_9^2} \left(\frac{\log(\sqrt{L_{10}} L_8) + \log\left(n^{\frac{2\alpha\kappa+2\kappa+d}{2\kappa(2\alpha+d)}} \log^{\rho(2/\kappa+2/d-1)-d-1}(n)\right)}{\log(n)} \right)^{1+d}, \quad (6.4.25)$$

$$n^{\frac{d}{2\alpha+d}} > \frac{1}{c_5}\left(\log(L_8^{q_1-d+1}) + 1\right), \quad (6.4.26)$$

Concluding the Proof

Thus, the three conditions (6.4.9), (6.4.10), and (6.4.11) of Theorem 6.4.3 are satisfied by the SGCP model, and we have

$$E_{\lambda_0}\left[\Pi_n\left(\lambda : \Gamma_{n,2}^{1/2}(\lambda, \lambda_0) \geq \sqrt{2} J M \epsilon_n | \tilde{X}_{1:n}\right)\right] \leq \frac{1}{C^k (n\epsilon_n^2)^{k/2}} + e^{-M^2 n \epsilon_n^2/4} + 2e^{-(M^2/2 - c_3')n\epsilon_n^2} + \frac{2}{c_1'} e^{-(c_2' M^2 J^2/4 - C - 1)n\epsilon_n^2}$$

with $\epsilon_n = \max(\delta_n, \bar{\delta}_n) = \bar{\delta}_n$. Then, so long as M and J are sufficiently large, the second, third and fourth terms on the RHS of equation (6.4.12) decay much more quickly than the first and the bound is $\tilde{o}(n^{\frac{-d}{2\alpha+d}})$ as stated, for all n such that the conditions of Theorem 6.4.3 are met and that (6.4.23), (6.4.24), (6.4.25), and (6.4.26) hold. □

6.4.4 Proof of Theorem 6.4.3: Generic contraction in NHPPs

The proof of Theorem 6.4.3 depends on a general result for convergence of posterior parameter estimation given i.n.i.d. observations. Such a result is given in Ghosal and Van Der Vaart (2007), but without a finite time rate on the probability. We restate their result below as Theorem 6.4.10 but with a rate included.

Consider as in Ghosal and Van Der Vaart (2007), a model in which a parameter $\theta_0 \in \Theta$ gives rise to a i.n.i.d sequence of data. The data at time i are drawn independently from the data at other times from a distribution P_i^θ, which we assume admits a density p_i^θ with respect to a dominating measure.

We define the following subsets of the parameter space for $n \geq 1$ and $k > 1$

$$B_n(\theta_0, \epsilon; k) = \left\{\theta \in \Theta : \frac{1}{n}\sum_{i=1}^{n} K_i(\theta_0, \theta) \leq \epsilon^2, \ \frac{1}{n}\sum_{i=1}^{n} V_{k,0;i}(\theta_0, \theta) \leq C_k \epsilon^k\right\}$$

where $K_i(\theta_0, \theta) = \int p_i^{\theta_0} \log(p_i^{\theta_0}/p_i^\theta) d\mu$ is the Kullback-Leibler divergence and $V_{k,0;i}(\theta_0, \theta) = \int p_i^{\theta_0} |\log(p_i^{\theta_0}/p_i^\theta) - K(\theta_0, \theta)|^k d\mu$ is a variance discrepancy measure. Furthermore, let d_n be the averaged Hellinger distance, defined by

$$d_n^2(\theta, \theta') = \frac{1}{n}\sum_{i=1}^{n} \int (\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 d\mu_i.$$

Our modified version of Theorem 4 of Ghosal and Van Der Vaart (2007) is as below.

Theorem 6.4.10. *Suppose $Y_i \sim P_i^\theta$ independently for $i = 1, ..., n$ and let d_n be defined as the average Hellinger distance. Further, suppose that for a sequence $\epsilon_n \to 0$ such that $n\epsilon_n^2$ is bounded away from 0, some $k > 1$, all sufficiently large $j \in \mathbb{N}$, constants $c_1, c_2, c_3 > 0$, and sets $\Theta_n \subset \Theta$, the following conditions hold:*

$$\frac{\Pi_n(\theta \in \Theta_n : j\epsilon_n < d_n(\theta, \theta_0) \leq 2j\epsilon_n)}{\Pi_n(B_n^*(\theta_0, \epsilon_n; k))} \leq c_1 e^{\frac{c_2 n\epsilon_n^2 j^2}{4}} \tag{6.4.27}$$

$$\frac{\Pi_n(\Theta \setminus \Theta_n)}{\Pi_n(B_n^*(\theta_0, \epsilon_n; k))} = o(e^{-2n\epsilon_n^2}) \tag{6.4.28}$$

$$\sup_{\epsilon > \epsilon_n} \log N\biggl(\frac{\epsilon}{36}, \bigl\{\theta \in \Theta_n : d_n(\theta, \theta_0) < \epsilon\bigr\}, d_n\biggr) \leq c_3 n\epsilon_n^2. \tag{6.4.29}$$

Then for any $C > 0, J \geq 1,$ and $M \geq 2$,

$$\begin{aligned}\mathbb{E}_{\theta_0}\Pi_n(\theta : d_n(\theta, \theta_0) \geq JM\epsilon_n | Y^{(n)}) \leq \frac{1}{C^k(n\epsilon_n^2)^{k/2}} &+ e^{-M^2 n\epsilon_n^2/4} \\ &+ 2e^{-(M^2/2 - c_3)n\epsilon_n^2} + \frac{2}{c_1}e^{-(c_2 M^2 J^2/4 - C - 1)n\epsilon_n^2}\end{aligned}$$

for all n such that $e^{-n\epsilon_n^2 M^2/4} \leq 1/2$.

The proof of Theorem 6.4.10 is a modification of proof of Theorem 4 of Ghosal and Van Der Vaart (2007). We replace arguments that hold in the limit with finite-time versions and handle the introduction of the constants c_1, c_2, c_3, assumed to be 1 in Ghosal and Van Der Vaart (2007). We can set the constant K present in the original theorem to $1/2$ since we are dealing with the Hellinger distance.

Proof of Theorem 6.4.10: By Lemmas 9 and 10 of Ghosal and Van Der Vaart (2007) and given conditions (6.4.27), (6.4.28), and (6.4.29), we have for n such that $e^{-n\epsilon_n^2 M^2/4} \leq 1/2$ any $M \geq 2$, $J \geq 1$ and $C > 0$,

$$\begin{aligned}&\mathbb{E}_{\theta_0}\Pi_n(\theta : d_n(\theta, \theta_0) \geq JM\epsilon_n | Y^{(n)}) \\ &\leq \frac{1}{C^k(n\epsilon_n^2)^{k/2}} + e^{-M^2 n\epsilon_n^2/4} + \frac{e^{-(M^2/2 - c_3)n\epsilon_n^2}}{1 - e^{-M^2 n\epsilon_n^2/2}} + \sum_{j \geq J} \frac{1}{c_1} e^{-n\epsilon_n^2(c_2 M^2 j^2/4 - C - 1)} \\ &\leq \frac{1}{C^k(n\epsilon_n^2)^{k/2}} + e^{-M^2 n\epsilon_n^2/4} + 2e^{-(M^2/2 - c_3)n\epsilon_n^2} + \frac{1}{c_1}e^{-n\epsilon_n^2(c_2 M^2 J^2/4 - C - 1)} \sum_{j=0}^{\infty} \bigl(e^{-n\epsilon_n^2(c_2 M^2/4)}\bigr)^{j^2} \\ &\leq \frac{1}{C^k(n\epsilon_n^2)^{k/2}} + e^{-M^2 n\epsilon_n^2/4} + 2e^{-(M^2/2 - c_3)n\epsilon_n^2} + \frac{1}{c_1}e^{-n\epsilon_n^2(c_2 M^2 J^2/4 - C - 1)} \sum_{j=0}^{\infty} \bigl(e^{-n\epsilon_n^2(c_2 M^2/4)}\bigr)^{j^2} \\ &\leq \frac{1}{C^k(n\epsilon_n^2)^{k/2}} + e^{-M^2 n\epsilon_n^2/4} + 2e^{-(M^2/2 - c_3)n\epsilon_n^2} + \frac{e^{-n\epsilon_n^2(c_2 M^2 J^2/4 - C - 1)}}{c_1(1 - e^{-n\epsilon_n^2 c_2 M^2/4})}\end{aligned}$$

$$\leq \frac{1}{C^k (n\epsilon_n^2)^{k/2}} + e^{-M^2 n\epsilon_n^2/4} + 2e^{-(M^2/2-c_3)n\epsilon_n^2} + \frac{2}{c_1} e^{-(c_2 M^2 J^2/4-C-1)n\epsilon_n^2}. \ \square$$

To apply Theorem 6.4.10 we define averaged versions of the Hellinger distance, KL divergence and variance measure. Let $p_{\lambda\gamma_i}(N) = p(X^{(i)}|\lambda, \gamma_i)$ and we define the averaged Hellinger distance $h_n(p_\lambda, p_{\lambda'})$ by

$$h_n^2(p_\lambda, p_{\lambda'}) = \frac{1}{n}\sum_{i=1}^n 2\left(1 - E_{\lambda\gamma_i}\left(\sqrt{\frac{p_{\lambda\gamma_i}(N)}{p_{\lambda'\gamma_i}(N)}}\right)\right),$$

the averaged KL-divergence as

$$k_n(p_\lambda, p_{\lambda'}) = -\frac{1}{n}\sum_{i=1}^n E_{\lambda'\gamma_i}\left(\log\left(\frac{p_{\lambda\gamma_i}(N)}{p_{\lambda'\gamma_i}(N)}\right)\right),$$

and variance measure as

$$v_n(p_\lambda, p_{\lambda'}) = \frac{1}{n}\sum_{i=1}^n Var_{\lambda'\gamma_i}\left(\log\left(\frac{p_{\lambda\gamma_i}(N)}{p_{\lambda'\gamma_i}(N)}\right)\right).$$

Through component-wise application of the relations in Section A.1 of Belitser et al. (2015) we have deterministic expressions for these quantities as

$$\begin{aligned}
h_n^2(p_\lambda, p_{\lambda'}) &= \frac{1}{n}\sum_{i=1}^n 2\left(1 - \exp\left\{-\frac{1}{2}\int_{R_i}\left(\sqrt{\lambda(t)\gamma_i(t)} - \sqrt{\lambda'(t)\gamma_i(t)}\right)^2 dt\right\}\right),\\
k_n(p_\lambda, p_{\lambda'}) &= \frac{1}{n}\sum_{i=1}^n \left(\int_{R_i}(\lambda(t) - \lambda'(t))\gamma_i(t)dt + \int_{R_i}\lambda'(t)\gamma_i(t)\log\left(\frac{\lambda'(t)}{\lambda(t)}\right)dt\right),\\
v_n(p_\lambda, p_{\lambda'}) &= \frac{1}{n}\sum_{i=1}^n \int_{R_i}\lambda'(t)\gamma_i(t)\log^2\left(\frac{\lambda'(t)}{\lambda(t)}\right)dt.
\end{aligned}$$

Lemma 1 of Belitser et al. (2015) gives bounds on the non-averaged versions of these quantities, but as the bounds will hold for each component of the average, we can trivially extend these results to give the following inequalities:

$$\frac{1}{\sqrt{2}n}\sum_{i=1}^{n}\big(||\sqrt{\lambda\gamma_i}-\sqrt{\lambda'\gamma_i}||_2\wedge 1\big)\leq h_n(p_\lambda,p_{\lambda'})\leq\frac{\sqrt{2}}{n}\sum_{i=1}^{n}\big(||\sqrt{\lambda\gamma_i}-\sqrt{\lambda'\gamma_i}||_2\wedge 1\big) \quad (6.4.30)$$

$$k_n(p_\lambda,p_{\lambda'})\leq\frac{3}{n}\sum_{i=1}^{n}||\sqrt{\lambda\gamma_i}-\sqrt{\lambda'\gamma_i}||_2^2+v_n(p_\lambda,p_{\lambda'}) \quad (6.4.31)$$

$$\frac{1}{n}\sum_{i=1}^{n}||\sqrt{\lambda\gamma_i}-\sqrt{\lambda'\gamma_i}||_2^2\leq\frac{1}{4n}\sum_{i=1}^{n}\int_{R_i}\gamma_i(s)(\lambda(s)\vee\lambda'(s))\log^2\Big(\frac{\lambda(s)}{\lambda'(s)}\Big)ds \quad (6.4.32)$$

where for numbers x and y, the minimum is denoted $x\wedge y$ and the maximum is denoted $x\vee y$.

By assumption, λ_0 is bounded away from 0. It follows that any $\lambda\in\Lambda$ with $||\lambda_0-\lambda||_\infty\leq\lambda_{min}$ is also bounded away from 0, and that by the results (6.4.31) and (6.4.32) above $k_n(p_{\lambda_0},p_\lambda)$ and $v_n(p_{\lambda_0},p_\lambda)$ are both bounded by a constant times the averaged uniform norm $\frac{1}{n}\sum_{i=1}^{n}||\lambda_0\gamma_i-\lambda\gamma_i||_\infty$. Therefore for $n\geq n_1$ the ball

$$B_n^*(\epsilon_n)=\Big\{\lambda\in\Lambda:k_n(p_{\lambda_0},p_\lambda)\leq\epsilon_n^2,v_n(p_{\lambda_0},p_\lambda)\leq\epsilon_n^2\Big\}$$

is bounded by a multiple of the ball

$$\Big\{\lambda\in\Lambda:\frac{1}{n}\sum_{i=1}^{n}||\lambda_0\gamma_i-\lambda\gamma_i||_\infty\leq\epsilon_n\Big\}$$

for $\epsilon_n\leq\lambda_{min}$. It follows that for $n\geq n_1$, the condition (6.4.9) implies

$$\Pi_n(B_n^*(\delta_n))\geq c_1e^{-c_2n\delta_n^2}. \quad (6.4.33)$$

By (6.4.30) we have that

$$N\left(\frac{\epsilon}{36}, \Lambda_{n,\epsilon}, h_n\right) \leq N\left(\frac{\epsilon}{36\sqrt{2}}, \sqrt{\Lambda_{n,\epsilon}}, \frac{1}{n}\sum_{i=1}^{n} || \cdot ||_2\right)$$

where $\Lambda_{n,\epsilon} = \left\{\lambda \in \Lambda_n : h_n(p_\lambda, p_{\lambda_0}) \leq \epsilon\right\}$. Thus the condition (6.4.10) implies (6.4.29).

Combining these results we have

$$\frac{\Pi_n(\lambda \in \Lambda_n : j\delta_n < h_n(p_\lambda, p_{\lambda_0}) \leq 2j\delta_n)}{\Pi_n(B_n^*(\delta_n))} \leq \frac{1}{\Pi_n(B_n^*(\delta_n))} \leq c_1 e^{c_2 n\delta_n^2}$$

by (6.4.33) to satisfy condition (6.4.27), and

$$\frac{\Pi_n(\Lambda_n^c)}{\Pi_n(B_n^*(\delta_n))} \leq \frac{c_4 e^{-c_5 n\delta_n^2}}{c_1 e^{-c_2 n\delta_n^2}} = o(e^{-2n\delta_n^2})$$

by (6.4.11) and (6.4.33) for $c_5 - c_2 \geq 2$ to satisfy (6.4.28). Thus all the conditions of Theorem 6.4.10 are satisfied by the assumptions of Theorem 6.4.3 and the conclusion of Theorem 6.4.10 carries forward to Theorem 6.4.3 where we choose $k = 2$. □

6.5 Conclusion

We have derived finite time rates on the posterior contraction of the QGCP and SGCP models given i.n.i.d observations. This allows us to quantify the contraction of posterior estimates in the setting where events are not detected perfectly or the observation region is not sampled uniformly. As well as a new consistency result for the QGCP model, and the innovations of studying i.n.i.d data over i.i.d., the presentation of explicit rates on the posterior mass for the contraction of non-homogeneous Poisson process models is new. These results are of theoretical importance and practical interest in problems such as sequential decision making and experimental design.

We found that the SGCP model admitted a much tighter analysis than the QGCP model. For the simple setting of 1-dimensional 1-Hölder smooth rate functions, the SGCP model can be shown to have convergence of the near-optimal order $\tilde{o}(n^{-1/3})$. Our best result for the QGCP model only shows convergence of order $\tilde{o}(n^{-1/5})$. This discrepancy arises because of the different link functions used in the two models. In comparison to the bounded sigmoid function, the quadratic function induces a larger space of rate functions when the GP is transformed - meaning that wider sieves are required to give the desired results and the contraction guarantees are looser. Guarantees on the tightness of these bounds are currently unavailable, but this work provides some evidence to suggest that the SGCP model is superior to the QGCP in terms of rate of posterior contraction at least. This is an observation that would merit further empirical and analytical study in to the relationship between the models. We did not consider the LGCP model in this work as its exponential link function makes it very difficult to adapt the existing GP results of van der Vaart and van Zanten (2009) into meaningful results in the NHPP posterior contraction setting. In particular, the high probability bound $\{||g - g_0||_\infty \leq \eta_{\beta_n}\}$ on the GP model, does not imply a useful bound on $||e^g - e^{g_0}||_\infty$ - the distance to be bounded in the prior mass condition for the LGCP - that gives useful contraction results.

We have focussed on particular choices of smoothness class, the link function used within the GCP construction and the width of the balls used in the contraction rate statements. There is of course potential to expand on these results by studying other choices. We believe however that the choices we have are consistent with the most common modelling choices in implementation of GCPs and useful for relating our results to the existing literature on posterior contraction of Bayesian nonparametric models.

The results we have provided in this chapter are, we believe, the best available with the current theory around contraction of nonparametric Bayesian inference. An open problem now is to utilise these to derive bounds on the performance of sequential event detection algorithms using GCP inference.

6.6 Further proofs

6.6.1 Proof of Lemma 6.4.4

Proving Lemma 6.4.4 relies on a bound on the uniform norm in the GP space. The following lemma gives a particular prior mass result which holds for all $n > 0$ and uses a general term $\eta_{\beta,n}$ which fits with the analysis of both the SGCP and QGCP models.

Lemma 6.6.1. *If $g_0 \in \mathcal{C}^\alpha([0,1]^d)$ for some $\alpha > 0$, then there exist constants $c_1, c_2 > 0$ such that*

$$\Pi\big(||g - g_0||_\infty \leq \eta_{\beta,n}\big) \geq c_1 e^{-c_2 n \eta_{\beta,n}^\beta}$$

for $\eta_{\beta,n} = n^{-\alpha/(\beta\alpha+d)}(\log(n))^{\rho_\beta}$ and $\rho_\beta = \frac{1+d}{\beta+d/\alpha}$ for all $n \geq 3$ where $\beta > 1$.

Furthermore, we rely on the following simple result which allows us to move between probabilistic bounds on the uniform norm of the GP and the squared GP.

Lemma 6.6.2. *Let w_1 and w_2 be functions defined on $[0,1]^d$ such that $||w_2||_\infty$ is finite, and c be a positive constant. Given the standard definition of the uniform norm, we have the following relation:*

$$\big\{||w_1 - w_2||_\infty \leq c\big\} \Rightarrow \big\{||w_1^2 - w_2^2||_\infty \leq 2c||w_2||_\infty + c^2\big\}.$$

We prove Lemma 6.6.1 in Section 6.6.2 and prove Lemma 6.6.2 below.

Proof of Lemma 6.6.2:

We have:

$$
\begin{aligned}
&||w_1 - w_2||_\infty \leq c \\
&\quad \Rightarrow w_1(x) \leq w_2(x) + c && \forall\ x \in S \\
&\quad \Rightarrow w_1^2(x) \leq w_2^2(x) + 2cw_2(x) + c^2 && \forall\ x \in S \\
&\Rightarrow ||w_1^2 - w_2^2|| \leq 2c||w_2||_\infty + c^2. && \square
\end{aligned}
$$

Proof of Lemma 6.4.4:

Recall the defintion $\delta_n = 2\eta_n||g_0||_\infty + \eta_n^2$, with $\eta_n = \eta_{4,n}$. By definition we have:

$$
\begin{aligned}
&\Pi\big(\lambda : ||\lambda - \lambda_0||_\infty \leq \delta_n\big) \\
&= \Pi\big(g : ||g^2 - g_0^2||_\infty \leq 2\eta_n||g_0||_\infty + \eta_n^2\big) \\
&\geq \Pi\big(g : ||g - g_0||_\infty \leq \eta_n\big) \\
&\geq c_1 e^{-c_2 n \eta_n^4} \geq c_1 e^{-c_2 n \delta_n^2},
\end{aligned}
$$

Here, the first inequality is due to Lemma 6.6.2. The second is by application of Lemma 6.6.1 and the third is by definition of δ_n. □

6.6.2 Proof of Lemma 6.6.1

We will utilise the following result from Section 5.1 of van der Vaart and van Zanten (2009), which holds for a constant H depending only on g_0 and μ, a constant K_2 depending only on

g_0, μ, α, d and D_1 and any $\epsilon > 0$

$$\Pi\big(||g - g_0||_\infty \leq 2\epsilon\big) \geq C_1 \exp\bigg\{ - K_2\Big(\frac{1}{\epsilon}\Big)^{d/\alpha}\bigg(\log\Big(\frac{1}{\epsilon}\Big)\bigg)^{1+d}\bigg\}\Big(\frac{H}{\epsilon}\Big)^{\frac{q_1+1}{\alpha}}.$$

Recall that C_1, D_1, and q_1 are constants from the assumption (6.3.5) on the length scale of the GP prior.

Substituting the particular form of $\epsilon = \epsilon_n = n^{-\alpha/(\beta\alpha+d)}(\log(n))^\rho$ and $\rho = \frac{1+d}{\beta+d/\alpha}$ from Lemma 6.6.1 into the above we have:

$$\begin{aligned}
&\Pi\big(||g - g_0||_\infty \leq \epsilon_n\big)\\
&\geq C_1 \exp\Bigg\{ - 2^{\frac{d}{\alpha}} K_2 n^{\frac{d}{\beta\alpha+d}}(\log(n))^{-\frac{d}{\alpha}\frac{1+d}{\beta+d/\alpha}}\Bigg(\log\bigg(2n^{\frac{\alpha}{\beta\alpha+d}}(\log(n))^{-\frac{1+d}{\beta+d/\alpha}}\bigg)\Bigg)^{1+d}\Bigg\}\\
&\quad\times\Bigg(2Hn^{\frac{\alpha}{\beta\alpha+d}}(\log(n))^{-\frac{1+d}{\beta+d/\alpha}}\Bigg)^{\frac{q_1+1}{\alpha}},
\end{aligned}$$

defining $Z(n) = \Big(Hn^{\frac{\alpha}{\beta\alpha+d}}(\log(n))^{-\frac{1+d}{\beta+d/\alpha}}\Big)^{\frac{q_1+1}{\alpha}}$ and expanding the logarithm,

$$\begin{aligned}
&= C_1 Z(n) \exp\Bigg\{ - 2^{\frac{d}{\alpha}} K_2 n^{\frac{d}{\beta\alpha+d}}(\log(n))^{-\frac{d}{\alpha}\frac{1+d}{\beta+d/\alpha}}\bigg(\frac{\alpha}{\beta\alpha+d}\log(2n) - (\log(n))^{\frac{1+d}{\beta+d/\alpha}}\bigg)^{1+d}\Bigg\}\\
&\geq C_1 Z(n) \exp\Bigg\{ - 2^{\frac{d}{\alpha}} K_2 n^{\frac{d}{\beta\alpha+d}}(\log(n))^{-\frac{d}{\alpha}\frac{1+d}{\beta+d/\alpha}}\bigg(\frac{\alpha}{\beta\alpha+d}\log(2n)\bigg)^{1+d}\Bigg\}
\end{aligned}$$

using $2\log(n) \geq \log(2n)$ for $n \geq 2$

$$\geq C_1 Z(n) \exp\Bigg\{ - 2^{1+d/\alpha} K_2 n \cdot n^{-\frac{\beta\alpha}{\beta\alpha+d}}(\log(n))^{(1+d)\frac{\beta+d/\alpha-d/\alpha}{\beta+d/\alpha}}\Bigg\}$$

$$= C_1 Z(n) \exp\Bigg\{ - 2^{1+d/\alpha} K_2 n \epsilon_n^\beta \Bigg\}$$

letting $K_3 = \min_{n \geq 3}(Z(n))$

$$\geq C_1 K_3 \exp\Bigg\{ - 2^{1+d/\alpha} K_2 n \epsilon_n^\beta \Bigg\} = c_1 e^{-c_2 n \epsilon_n^\beta},$$

where $c_1 = C_1 K_3$, and $c_2 = 2^{1+d/\alpha} K_2$. □

6.6.3 Proof of Lemma 6.4.5

As $\sqrt{\Lambda_n} = \mathcal{G}_n$, the covering numbers $N(L_1\bar{\delta}_n, \sqrt{\Lambda_n}, ||\cdot||_2)$ and $N(L_1\bar{\delta}_n, \mathcal{G}_n, ||\cdot||_2)$ are equivalent. It follows that

$$N(L_1\bar{\delta}_n, \sqrt{\Lambda_n}, ||\cdot||_2) \leq N(L_1\bar{\delta}_n, \mathcal{G}_n, ||\cdot||_\infty).$$

Defining $\mathcal{G}_n$ as in (6.4.17) allows us to use the following result, (5.4) of van der Vaart and van Zanten (2009):

$$\log N(L_1\bar{\delta}_n, \mathcal{G}_n, ||\cdot||_\infty) \leq m\zeta_n^d \Bigg(\log \frac{3^{3/2} d^{1/4} \beta_n^{3/2} \sqrt{2\tau\zeta_n}}{(L_1\bar{\delta}_n)^{3/2}} \Bigg)^{1+d} + 2\log\frac{6\beta_n\sqrt{||\mu||}}{L_1\bar{\delta}_n}$$

for $||\mu||$ the total mass of the spectral measure μ, τ^2 as the second moment of μ, positive constant m depending only on μ and d, and given

$$(3/L_1)^{3/2} d^{1/4} \beta_n^{3/2} \sqrt{2\tau\zeta_n} > 2\bar{\delta}_n^{3/2}, \qquad (3/L_1)\beta_n\sqrt{||\mu||} > \bar{\delta}_n.$$

By the definitions of β_n, and ζ_n we have that

$$m\zeta_n^d\biggl(\log\frac{3^{3/2}d^{1/4}\beta_n^{3/2}\sqrt{2\tau\zeta_n}}{(L_1\bar{\delta}_n)^{3/2}}\biggr)^{1+d} \leq n\bar{\delta}_n^2$$
$$2\log\frac{6\beta_n\sqrt{||\mu||}}{L_1\bar{\delta}_n} \leq n\bar{\delta}_n^2,$$

for the values of n specified in the statement of Lemma 6.4.5. It follows that the lemma is satisfied with $c_3 = 2$. □

6.6.4 Proof of Lemma 6.4.6

Firstly note that $\Pi(\lambda \notin \Lambda_n) = \Pi(g \notin \mathcal{G}_n)$. By a simplification of (5.3) of van der Vaart and van Zanten (2009) to account for our assumption that $q_2 = 0$, we have

$$\Pi(g \notin \mathcal{G}_n) \leq C_1\zeta_n^{q_1-d+1}e^{-D_1\zeta_n^d} + e^{-\beta_n^2/8}$$

$\bar{\delta}_n < \delta_0$ for small $\delta_0 > 0$, and β_n, ζ_n, and $\bar{\delta}_n$ satisfying

$$\beta_n^2 > 16K_5\zeta_n^d\biggl(\log\left(\frac{3\zeta_n}{\bar{\delta}_n}\right)\biggr)^{1+d}, \qquad \zeta_n > 1,$$

for a constant K_5 depending only on μ and g. The definitions of β_n, δ_n and ζ_n give us the following relations, for a constant $c_5 = c_2 + 4$

$$D_1\zeta_n^d \geq 2c_5n\delta_n^2,\ \ \beta_n^2 \geq 8c_5n\delta_n^2,\ \ \zeta_n^{q_1-d+1} \leq e^{c_5n\delta_n^2},$$

with the final of these holding for values of n as specified in the statement of Lemma 6.4.6. Using these we can obtain the necessary result as follows:

$$\begin{aligned}\Pi(\lambda : \lambda \notin \Lambda_n) = \Pi(g : g \notin \mathcal{G}_n) &\leq C_1 \zeta_n^{q_1-d+1} e^{-D_1 \zeta_n^d} + e^{-\beta_n^2/8} \\ &\leq C_1 e^{c_5 n\delta_n^2} e^{-2c_5 n\delta_n^2} + e^{-c_5 n\delta_n^2} \\ &= \big(C_1 + 1\big) e^{-2c_5 n\delta_n^2} \\ &\leq c_4 e^{-(c_2+4)n\delta_n^2}\end{aligned}$$

for $c_4 = C_1 + 1$. □

6.6.5 Proof of Lemma 6.4.7

Under the SGCP model we have

$$\begin{aligned}\Pi\Big((\lambda^*, g) : ||\lambda^*\sigma(g) - \lambda_0||_\infty \leq \delta_n\Big) \geq \Pi\Big(\lambda^* : |\lambda^* - 2||\lambda_0||_\infty| \leq \frac{\delta_n}{2}\Big) \\ \times \Pi\Big(g : ||\sigma(g) - \sigma(g_0)||_\infty \leq \frac{\delta_n}{4||\lambda_0||_\infty}\Big).\end{aligned}$$

By the assumption that λ^* has a positive continuous density, the first term on the RHS of the inequality can be bounded below by a constant times δ_n, which can itself be lower bounded by a constant for finite n. The second term can be bounded below by $\Pi_n(||g - g_0||_\infty \leq \delta_n/(16||\lambda_0|_\infty))$ since $1/4$ is the Lipschitz constant of the sigmoid transformation. Thus, by Lemma 6.6.1 (given in Section 6.6.1) we have:

$$\Pi\Big(\lambda : \Gamma_{n,\infty}(\lambda, \lambda_0) \leq \delta_n\Big) \geq c_0'\Pi\Big(g : ||g - g_0||_\infty \leq \frac{\delta_n}{16||\lambda_0||_\infty}\Big) \geq c_1' e^{-nc_2'\delta_n^2}$$

for positive constants c_1', c_2', showing condition (6.4.9) is satisfied under the SGCP model.

6.6.6 Proof of Lemma 6.4.8

Define $\psi_n = \bar{\delta}_n/(2L_1\sqrt{\lambda_n})$. We have

$$\begin{aligned}\log N(L_1\bar{\delta}_n, \sqrt{\Lambda_n}, ||\cdot||_2) &= \log N(2\psi_n\sqrt{\lambda_n}, \sqrt{\Lambda_n}, ||\cdot||_2)\\ &\leq \log N(\psi_n\sqrt{\lambda_n}, [0,\lambda_n], \sqrt{|\cdot|}) + \log N(\psi_n/c, \mathcal{G}_n, ||\cdot||_\infty)\\ &\leq \log\frac{1}{\psi_n} + \log N(\psi_n/c, \mathcal{G}_n, ||\cdot||_\infty) \qquad (6.6.34)\end{aligned}$$

for $c = 2^{-5/2}$, the Lipschitz constant of $\sqrt{\sigma}$.

Then, as in the proof of Lemma 6.4.5, by equation (5.4) of van der Vaart and van Zanten (2009), we have for $B_n > 0$,

$$\log N(B_n\bar{\delta}_n, \mathcal{G}_n, ||\cdot||_\infty) \leq m\zeta_n^d\bigg(\log\frac{3^{3/2}d^{1/4}\beta_n^{3/2}\sqrt{2\tau\zeta_n}}{(B_n\bar{\delta}_n)^{3/2}}\bigg)^{1+d} + 2\log\frac{6\beta_n\sqrt{||\mu||}}{B_n\bar{\delta}_n} \qquad (6.6.35)$$

subject to the conditions

$$(3/B_n)^{3/2}d^{1/4}\beta_n^{3/2}\sqrt{2\tau\zeta_n} > 2\bar{\delta}_n^{3/2}, \quad (3/B_n)\beta_n\sqrt{||\mu||} > \bar{\delta}_n, \quad \zeta_n > A$$

for a constant $A > 0$. These conditions hold by defintion for n as specified by (6.4.23), with $B_n = (2cL_1\sqrt{\lambda_n})^{-1}$. Then, combining (6.6.34) and (6.6.35) we have

$$\begin{aligned}\log N\big(L_1\bar{\delta}_n, \sqrt{\Lambda_n}, ||\cdot||_2\big) &\leq \log\frac{2L_1\sqrt{\lambda_n}}{\bar{\delta}_n} + m\zeta_n^d\bigg(\log\frac{(6cL_1)^{3/2}d^{1/4}\beta_n^{3/2}\sqrt{2\tau\lambda_n\zeta_n}}{\bar{\delta}_n^{3/2}}\bigg)^{1+d}\\ &\quad + 2\log\frac{12cL_1\beta_n\sqrt{\lambda_n||\mu||}}{\bar{\delta}_n}.\end{aligned}$$

For n as specified by (6.4.24), we have

$$\begin{aligned}\log\frac{2L_1\sqrt{\lambda_n}}{\bar{\delta}_n} &< n\bar{\delta}_n^2,\\ m\zeta_n^d\bigg(\log\frac{(6cL_1)^{3/2}d^{1/4}\beta_n^{3/2}\sqrt{2\tau\lambda_n\zeta_n}}{\bar{\delta}_n^{3/2}}\bigg)^{1+d} &< n\bar{\delta}_n^2,\\ 2\log\frac{12cL_1\beta_n\sqrt{\lambda_n||\mu||}}{\bar{\delta}_n} &< n\bar{\delta}_n^2.\end{aligned}$$

Thus for n satisfying (6.4.23) and (6.4.24) we have

$$\log N\bigg(L_1\bar{\delta}_n,\sqrt{\Lambda_n},||\cdot||_2\bigg)\leq 3n\bar{\delta}_n^2$$

proving Lemma 6.4.8 with $c_3=3$. □

6.6.7 Proof of Lemma 6.4.9

As in Kirichenko and Van Zanten (2015), we may decompose the probability of interest

$$\begin{aligned}\Pi\Big(\lambda:\lambda\notin\Lambda_n\Big) &= \Pi\Big((\lambda^*,g):\lambda^*\sigma(g)\notin\Lambda_n\Big)\\ &\leq\int_0^{\lambda_n}\Pi\Big((\lambda^*,g):\lambda^*\sigma(g)\notin\Lambda_n\Big)p_{\lambda^*}(\lambda)d\lambda+\int_{\lambda_n}^{\infty}p_{\lambda^*}(\lambda)d\lambda\\ &\leq\Pi\Big(g:g\notin\mathcal{G}_n\Big)+C_0e^{-c_0\lambda_n^{\rho}},\end{aligned}$$

by the assumption (6.3.6). As utilised in the proof of Lemma 6.4.6, equation (5.3) of van der Vaart and van Zanten (2009) states that

$$\Pi(g\notin\mathcal{G}_n)\leq C_1\zeta_n^{q_1-d+1}e^{-D_1\zeta_n^d}+e^{-\beta_n^2/8}$$

given conditions

$$\beta_n^2 > 16K_5\zeta_n^d\Big(\log\Big(\frac{\lambda_n^{1/2}\zeta_n}{\bar{\delta}_n}\Big)\Big)^{1+d}, \quad \zeta_n > 1$$

which are satisfied by our earlier definitions, for a constant K_5 depending only on μ and g. Then for n as specified by equations (6.4.25) and (6.4.26), we have the following results

$$c_0\lambda_n^\rho > c_5 n\delta_n^2, \quad D_1\zeta_n^d \geq 2c_5 n\delta_n^2, \quad \zeta_n^{q_1-d+1} \leq e^{c_5 n\delta_n^2}, \quad \beta_n^2 \geq 8c_5 n\delta_n^2.$$

The required result then follows. □

Chapter 7

Thompson Sampling for Lipschitz Bandits

An updated version of this chapter has been accepted for publication, to appear as Grant J.A., and Leslie, D.S. (2020). On Thompson Sampling for Smoother-than-Lipschitz Bandits. *In Proceedings of AISTATS 2020.*

7.1 Introduction

The posterior contraction results of Chapter 6 provide us with a deeper understanding of Gaussian Cox Processes (GCPs), but do not readily lead to a bound on the performance of GCP-Thompson Sampling (GCP-TS). As we described in Section 6.1.1, there is , generally speaking, a lack of understanding of TS based on non-parametric inference. The GCP-TS is a special case among many poorly understood algorithms. While the results of Chapter 6 may be useful for deriving a bound on the regret of GCP-TS, when coupled with appropriate regret analysis techniques, such sophisticated techniques are unfortunately not currently known.

In this chapter we will tackle the more general problem, of quantifying the performance of the TS approach based on a non-parametric prior over a class of smooth problems in its application to the continuum-armed bandit (CAB) problem which we introduced in Chapter 3.

We will give a bound on the Bayesian regret of TS for CAB problems where the reward function is a sample from a prior on the class of functions with M Lipschitz smooth derivatives, where $M \in \mathbb{N}$, and the feedback is of the form of a noisy realisation of the reward function at the location of the selected action in each round. While a prior with mass on the entirety of this class may be hard to define, there are a number of non-parametric models which may approximate it well or place all their mass on such a class. Some options are a Gaussian process (GP) with a covariance kernel which induces Lipschitz smooth realisations, certain Bayesian neural networks, or priors over smooth basis functions such as B-splines. The particular choice of prior will of course have an effect on the performance of TS, and if the true reward function is not supported by the chosen prior, the results in this chapter may no longer apply.

However, we assert that results pertaining to general priors and exact inference are valuable benchmarking tools, increase our understanding of the TS principle in general, and derivation of such results is timely as implementation of (approximate) TS based on the aforementioned non-parametric priors becomes increasingly viable, thanks to advanced sampling techniques.

Furthermore, these results are valuable because TS is a powerful method which may be more widely applicable than other approaches which require careful tuning to the concentration of parameter estimates, and thus any further understanding of its properties is helpful. CABs have applications in many of the same settings as simpler bandit models including clinical trials and dose design, website optimisation, parameter tuning, and optimal search. Indeed, in many cases the CAB provides a more realistic representation of the available action space and reward function than the necessarily discrete formulation under the MAB.

In our sequential event detection problem, if only the per-round reward (and not individual event locations) is observed by the decision-maker, the problem can be readily modelled by a CAB problem of the type described above. There are a number of reasons that this may in fact be a realistic assumption. Perhaps storage of event locations is too costly and they are discarded, or pinpointing event locations is somehow more challenging, or less reliable, than detecting the occurrence of an event and the information is not deemed suitable for inference.

7.1.1 Related Work

As we mentioned in Chapter 3, although TS was first proposed as a method long ago (Thompson, 1933), the majority of research on TS has been developed in the last decade. Numerous authors have studied the frequentist regret of TS in multi-armed bandit (MAB), combinatorial multi-armed bandit (CMAB), and contextual bandit problems, with varying assumptions on the feedback mechanism and reward noise distribution (May et al., 2012; Agrawal and Goyal, 2012; Kaufmann et al., 2012b; Korda et al., 2013; Komiyama et al., 2015; Wang and Chen, 2018).

Russo and Van Roy (2014) originated the study of the Bayesian regret of TS. They introduced the notion of the *ϵ-eluder dimension* of a function class (a new complexity measure useful for analysis, whose form we will fully specify in the main body of the chapter) and showed that by considering this along with the concentration properties of the least squares estimator, a bound on the Bayesian regret of TS for general action sets and reward function classes is available. They use this generic argument to derive bounds for bandit problems with (generalised) linear reward functions under sub-Gaussian noise. Quadratic functions and applications in model-based reinforcement learning are considered by Osband and Van Roy (2014). The eluder dimension-based technique may be generalized further. We extend the application of this theory to the broader setting of reward functions with Lipschitz derivatives and sub-exponential reward noise, through a new ϵ-eluder dimension bound, and the generalisation of Russo and van Roy's work. This extension is valuable as the application of TS to such a general notion of the CAB with smooth reward functions has not yet been studied.

The special case of TS for the CAB problem where the reward function is a sample from a

GP and the reward noise is sub-Gaussian, sometimes referred to as *GP optimisation* has received some 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). Hernández-Lobato et al. (2014) showed a method to approximately sample from a GP posterior, and later Bijl et al. (2016) show how Sequential Monte-Carlo methods can be used to implement approximate TS for the CAB with reward function drawn from a GP. Russo and Van Roy (2014) also derive a Bayesian regret bound for TS applied to GP optimisation. These results rely on the information-theoretic technqiues of Srinivas et al. (2012), meaning that they only hold for GPs with a covariance kernel for which the "maximum information gain" is known. This is an information theoretic property of a particular covariance kernel and is non-trivial to derive.

Basu and Ghosh (2017) study an ϵ-randomised variant of TS, however they are interested in the rate at which the selected action converges to the optimal action, rather than regret or Bayesian regret. They show that an exponential rate of convergence is achievable subject to certain conditions on the kernel function and its eigenfunctions. Kandasamy et al. (2018) provide methods for parallelising TS for Bayesian Optimisation in this setting, and carry forward versions of the guarantees of Russo and Van Roy (2014).

Further papers have considered the use of information theoretic ideas to bound the Bayesian regret of TS in multi-armed bandits (Russo and Van Roy, 2016; Dong and Van Roy, 2018). These bounds express regret in terms of the *information ratio* - a statistic which characterises the trade-off between exploration and exploitation performed by a particular algorithm. The techniques used to derive these bounds are quite different to the confidence set based analysis we use in this chapter, and as such we will not investigate them further here, though we note they could perhaps also be applied to the nonparametric bandits we consider in our work.

Several upper confidence bound approaches exist for CABs with a Lipschitz smooth reward function. In this setting, Kleinberg (2005) demonstrated that $\Omega(T^{2/3})$ regret is the best achievable. Order-optimal performance can be achieved by the zooming algorithm (Kleinberg, 2005), which was initially proposed for sub-Gaussian rewards. Recently the order optimal results have been extended to a version of the zooming algorithm adapted to heavy tailed rewards (Lu et al., 2019). Bubeck et al. (2011) propose the Hierarchical Online Optimisation (HOO) algorithm which can attain $O(\sqrt{T})$ regret, subject to further convexity assumptions on the reward function, which also reduce the order of the lower bound for the problem. Each of these algorithms achieves order optimal regret by an adaptive discretisation routine which imposes an appropriate amount of exploration on the sequence of selected actions.

Following Srinivas et al. (2010, 2012) a number of works have considered Bayesian upper confidence bound algorithms (not to be confused with the Bayes-UCB algorithm mentioned in Chapter 3) for GP optimisation. Under this the regime $O(\sqrt{T})$ regret is possible. This is because the reward function is assumed to be a sample from a GP and thus is restricted to be smoother than in the Lipschitz bandit setting. Srinivas et al. (2010) initially proposed the GP-UCB algorithm (an extension of the UCB idea to CABs) and demonstrated $O(\sqrt{T})$ regret. Several extensions of the algorithm are proposed and found to have similar theoretical guarantees. Contal et al. (2013); Bogunovic et al. (2016); Wang et al. (2016) propose methods where UCB decision making steps are combined with pure exploration steps or where pure exploration is performed on a subset of actions determined by a UCB function. Contal et al. (2014) propose a variant of the method where the exploration bonus incorporates the expected information gain. In Shekhar and Javidi (2018) the GP-UCB ideas are combined with ideas from tree-search algorithms for Lipschitz bandit problems to give an approach which avoids any non-convex optimisation. Grünewälder et al. (2010) and Scarlett et al. (2017) derive algorithm-independent lower bounds for GP optimisation in the noise-free and noisy settings respectively, and Krause and Ong (2011) extend the GP-UCB algorithm to a contextual bandit setting.

7.1.2 Key Contributions

The main contribution of this chapter 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 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 non-parametric inference that considers such a general framework. We derive an upper bound of order $O(T^{1-\frac{1}{2}\frac{2M^2+3M+2}{2M^2+7M+6}})$. This suggests that TS may not perform as well as UCB methods with adaptive discretisation for problems with M small, but that it seems to perform at the optimal order 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.

7.1.3 Chapter Outline

The remainder of the chapter is structured as follows. Section 7.2 introduces the general bandit model relevant to this chapter. In Section 7.3 we present a general bound on the Bayesian regret under sub-exponential reward noise. Then in Section 7.4 we specialise this bound to particular reward function classes, including the case of a reward function having Lipschitz smooth derivatives. Finally we conclude with a discussion in Section 7.5.

7.2 Model

Throughout the chapter we will use the following general representation of a bandit problem. There exists a set of actions $\mathcal{A} \in \mathbb{R}^d$, which can be selected by a decision-maker. Each action $a \in \mathcal{A}$

has an expected reward given by a reward function $f_\theta : \mathcal{A} \to \mathbb{R}$ parameterised by a potentially infinite dimensional parameter $\theta \in \Theta$, with prior p_θ, where Θ is a parameter space. The implication of this parameterisation is that $f_\theta \in \mathcal{F}$, where $\mathcal{F}$ is a class of functions parameterised by $\theta \in \Theta$, that we will assume is known. Furthermore, we assume that $\forall f \in \mathcal{F},\ \ \forall a \in \mathcal{A},\ \ f(a) \in [0, C]$, i.e. that all functions in $\mathcal{F}$ are bounded on $\mathcal{A}$. 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 $R_t = f_\theta(a_t) + \eta_t$, which is a noisy pertubation of the reward function at a_t. Let $\mathcal{H}_t = \sigma(a_1, R_1, \dots, a_t, R_t)$ 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 $(\mathcal{H}_{t-1}, \theta, a_t)$, meaning

$$\mathbb{E}\big(e^{\lambda \eta_t} | \mathcal{H}_{t-1}, \theta, a_t\big) \leq e^{\frac{\lambda^2 \sigma^2}{2}}, \quad \forall\ |\lambda| \leq \frac{1}{b}. \tag{7.2.1}$$

We are interested in the performance of TS as a policy to select actions a_t for $t \in [T]$. Let $p_{\theta,t}$ denote the posterior distribution on θ conditioned on $\mathcal{H}_t$ and let $\tilde{\theta}_t$ be a sample from $p_{\theta,t}$. Set $p_{\theta,0} = p_\theta$. The TS approach, π^{TS}, is the one which chooses an action $a_t \in \operatorname{argmax}_{a \in \mathcal{A}} f_{\tilde{\theta}_{t-1}}(a)$ in round t, breaking ties arbitrarily if the maximiser is non-unique.

We concern ourselves with the Bayesian regret of π^{TS} in T rounds, given as

$$BReg(T, \pi^{TS}) = \sum_{t=1}^{T} \mathbb{E}_{p_\theta}\bigg(\max_{a \in \mathcal{A}} f_\theta(a) - f_\theta(a_t)\bigg), \tag{7.2.2}$$

where $\mathbb{E}_{p_\theta}$ denotes expectation with respect to the prior p_θ. In particular, we are interested in bounding the Bayesian regret as a function of T for particular $\mathcal{A}$ and $\mathcal{F}$, and the order with respect to T that such bounds possess. The choice to study Bayesian regret is a natural one in the Bayesian framework. Guarantees on the frequentist regret are also available for TS in other settings. However, since these guarantees are generally constructed via markedly different analytical techniques, we will not consider frequentist performance measures in this chapter. In the following section we proceed to give a bound on the Bayesian regret for this very general representation of a bandit problem.

7.3 Bounds on the Bayesian Regret

We first give a bound on the Bayesian regret for general function classes, $\mathcal{F}$, and action sets, $\mathcal{A}$, and later specialise these expressions for particular important choices of $\mathcal{F}$ and $\mathcal{A}$. This general result is similar to the general bound given in 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.

The difficulty of a bandit problem is often related to the complexity of the function class, and the size of the action set. This is natural, since in more complex function classes, it will be more challenging to learn the true function. Thus bounds on the Bayesian regret include measures of the compexity of $\mathcal{F}$. Specifically, Russo and Van Roy (2014) show that the performance of TS can be linked to two notions of the complexity of $\mathcal{F}$, the ϵ-eluder dimension, and ball-width function, which we introduce below.

Firstly, to define the ϵ-eluder dimension, we first introduce the notion of ϵ-dependence. An action $a \in \mathcal{A}$ is said to be *ϵ-dependent* on actions $\{a_1, \dots, a_n\} \in \mathcal{A}$ 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 $\{a_1, \dots, a_n\}$ if a is not ϵ-dependent on $\{a_1, \dots, a_n\}$. The *ϵ-eluder dimension* $dim_E(\mathcal{F}, \epsilon)$, which we will often refer to simply as the eluder dimension, 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 smoothness 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 greater the smoothness of the functions in a function class, the smaller the eluder dimension of that function class is.

Second, we have the following function, which we will refer to as the *ball-width function*. The ball-width function, β_n^*, defines the size of high-probability confidence sets in the function class $\mathcal{F}$, in terms of n, a number of reward observations. In particular it depends on $N(\alpha, \mathcal{F}, ||\cdot||_\infty)$, the α-covering number of the function class $\mathcal{F}$ with respect to the uniform norm, $||\cdot||_\infty$, the sub-exponential parameters of the reward noise distribution, σ^2 and b, further parameters $\alpha, \delta > 0$ which will be chosen to optimise the regret bound and $\lambda : |\lambda| \leq b^{-1}$ which keeps the same interpretation as the free parameter in Equation (7.2.1). The ball-width function is specified as follows:

$$\begin{aligned}\beta_n^*(\mathcal{F}, \delta, \alpha, \lambda) := \frac{\log(N(\alpha, \mathcal{F}, ||\cdot||_\infty)/\delta)}{\lambda(1-2\lambda\sigma^2)} + \frac{2\alpha n(4C+\alpha)(1-\lambda\sigma^2)}{1-2\lambda\sigma^2} \\ + \frac{2\alpha\sum_{i\leq\lfloor n_0\rfloor}\sqrt{2\sigma^2\log(4i^2/\delta)} + 2\alpha\sum_{i\geq\lceil n_0\rceil}^{n} 2b\log(4i^2/\delta)}{1-2\lambda\sigma^2},\end{aligned} \tag{7.3.3}$$

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

The ball-width function presented here is the analogue of the simpler equation (8) given by Russo and Van Roy (2014) in the case of sub-Gaussian noise. The properties of sub-exponential distributions mean that our expression is more complex, but the interpretation of both functions is the same. The functions $\{\beta_n^*\}_{n=1}^\infty$ define the widths of certain high-probability confidence sets for the true reward function, based on n actions and realisations. In particular, they depend on the α-covering number of the function class. This is natural, since in larger function classes, a greater coverage is required to include the true reward function with high probability.

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 7.3.1. *For all problem horizons $T \in \mathbb{N}$, parameters $\alpha > 0, \delta \leq 1/(2T)$, and $|\lambda| \leq b^{-1}$, and nonincreasing functions $\kappa : \mathbb{N} \to \mathbb{R}_+$, we have that the Bayesian regret of π^{TS} applied to the general bandit problem with action set $\mathcal{A}$ where the reward function $f_\theta \in \mathcal{F}$ is drawn from p_θ and reward noise is (σ^2, b)-sub-exponential is bounded as follows,*

$$BReg(T, \pi_\theta^{TS}) \leq T\kappa(T) + (dim_E(\mathcal{F}, \kappa(T)) + 1)C + 4\sqrt{dim_E(\mathcal{F}, \kappa(T))\beta_T^*(\mathcal{F}, \alpha, \delta, \lambda)T}. \quad (7.3.4)$$

The bound (7.3.4) 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 $\mathcal{F}$. Then 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 7.4.

In Russo and Van Roy (2014) a similar bound to (7.3.4) is constructed, but a material difference is that $\kappa(T)$ is fixed to T^{-1}, which constrains the quality of the results which can be obtained for specific function classes. We show that the analysis may be extended to allow for more general choices of $\kappa(T)$ as a nonincreasing function of T, allowing for greater flexibility in deriving function class specific results.

In the remainder of this section we provide a proof of Theorem 7.3.1. The ideas of the proof are similar to those employed in Chapter 5. Central to the proof is the observation that, when concerned with Bayesian regret, TS can be shown to achieve the best performance of any upper confidence bound sequence. That is to say, that given a sequence of high probability confidence sets for the reward function, the Bayesian regret of TS may be decomposed in terms of the sums of the widths of these sets, which should be decreasing functions of the number of rounds. This means that if the confidence sets are chosen appropriately, this sum may be written as being sublinear in the problem horizon T.

In Chapter 5, this property was exploited by selecting a bespoke set of confidence intervals for the empirical mean of Poisson data. In this chapter we require a more general sequence of confidence sets, around function estimators, as opposed to univariate parameters. The proof of Theorem 7.3.1 makes use of general properties of the least squares estimator, which can be defined abstractly for a general estimation problem on $\mathcal{F}$, even if it does not admit a convenient analytical form. What is key, is that the width of certain high probability confidence sets around the least squares estimator *can* be defined in closed-form. These widths, which are specified in terms of the eluder dimension and ball-width function, are then used to bound the regret.

7.3.1 Proof of Theorem 7.3.1: General regret bound

We begin the proof with the following martingale concentration result, an extension of Lemma 3 of Russo and Van Roy (2014) (which holds for sub-Gaussian noise). The result below says that with high probability, for any function $f : \mathcal{A} \to \mathbb{R}$, its squared error $L_{2,t}(f) = \sum_{i=1}^{t-1}(f(A_i) - R_i)^2$ is lower bounded. In particular, we say that the squared error of f will not fall below the sum of the squared error of the true reward generating function, f_θ, and a measure of the distance between f and f_θ, by more than a fixed constant.

Lemma 7.3.2. *For any action sequence $A_1, A_2, \dots \in \mathcal{A}$, inducing (σ^2, b)-sub-exponential reward observations $R_1, R_2, \dots$ and any function $f : \mathcal{A} \to \mathbb{R}$, we have*

$$\mathbb{P}\bigg(L_{2,n+1}(f) \geq L_{2,n+1}(f_\theta) + (1 - 2\lambda\sigma^2)\sum_{i=1}^{n}(f(A_i) - f_\theta(A_i))^2 - \frac{\log(1/\delta)}{\lambda}, \;\; \forall n \in \mathbb{N}\bigg) \geq 1 - \delta, \tag{7.3.5}$$

for all λ with $|\lambda| \leq b^{-1}$.

A proof of Lemma 7.3.2 is provided in Section 7.6, it is based on the sub-exponential property of the reward noise. Lemma 7.3.2 allows us to construct high-probability confidence sets for the true reward function, f_θ. These sets are defined with respect to the least squares estimate of f_θ. That is a function $\hat{f}_t^{LS} \in \operatorname{argmin}_{f \in \mathcal{F}} L_{2,t}(f)$, with minimal squared error, in reference to the observed rewards. The following lemma gives the definition and high-confidence property of said confidence sets.

Lemma 7.3.3. *For all $\delta > 0$, $\alpha > 0$, $|\lambda| \leq b^{-1}$, and $\{A_1, \dots A_n\} \in \mathcal{A}^n$ we have*

$$\mathbb{P}\bigg(f_\theta \in \bigcap_{n=1}^{\infty} \mathcal{F}_n \bigg) \geq 1 - 2\delta.$$

for all $n \in \mathbb{N}$ where

$$\mathcal{F}_n = \bigg\{ f \in \mathcal{F} : \sum_{i=1}^{n} (\hat{f}_n^{LS}(A_i) - f(A_i))^2 \leq \beta^*(\mathcal{F}, \delta, \alpha, \lambda) \bigg\}.$$

The proof of Lemma 7.3.3 is also reserved for Section 7.6. The confidence sets $\{\mathcal{F}_n\}_{n=1}^{\infty}$ defined in Lemma 7.3.3, allow us to bound the Bayesian regret of TS. Specifically, we can decompose the Bayesian regret in terms of a notion of the width of these confidence intervals. By Lemma 4 of Russo and Van Roy (2014) we have for all problem horizons $T \in \mathbb{N}$, that if $\inf_{f \in \mathcal{F}_t} f(a) \leq f_\theta(a) \leq \sup_{f \in \mathcal{F}_t} f(a)$ for all $t \in \mathbb{N}$ and $a \in \mathcal{A}$ with probability at least $1 - 1/T$ then

$$BReg(T, \pi_\theta^{TS}) \leq C + \mathbb{E}\bigg(\sum_{t=1}^{T} \sup_{f \in \mathcal{F}_t} f(a) - \inf_{f \in \mathcal{F}_t} f(a) \bigg). \tag{7.3.6}$$

The proof of Theorem 7.3.1 can then be completed by bounding the widths of the confidence sets, $w_{\mathcal{F}_t}(a) = \sup_{f \in \mathcal{F}_t} f(a) - \inf_{f \in \mathcal{F}_t} f(a)$. The following Lemma provides such a result by bounding the sum of the widths in terms of the $\kappa(T)$-eluder dimension, $\dim_E(\mathcal{F}, \kappa(T))$. It is a generalisation of Lemma 5 of Russo and Van Roy (2014) which fixes $\kappa(t) = t^{-1}$ and we provide its proof in Section 7.6.

Lemma 7.3.4. *If* $(\beta_t \geq 0 \ \ |t \in \mathbb{N})$ *is a non-decreasing sequence and* $\mathcal{F}_t$ *is*

$$\mathcal{F}_t := \left\{ f \in \mathcal{F} : \sum_{i=1}^{t} (\hat{f}_t^{LS}(A_i) - f(A_i))^2 \leq \beta_t \right\}$$

then for all $T \in \mathbb{N}$*, and non-increasing functions* $\kappa : \mathbb{N} \to \mathbb{R}_+$

$$\sum_{t=1}^{T} w_{\mathcal{F}_t}(A_t) \leq T\kappa(T) + dim_E(\mathcal{F}, \kappa(T))C + 4\sqrt{dim_E(\mathcal{F}, \kappa(T))\beta_T T}. \tag{7.3.7}$$

□

7.4 Bounds for Specific Function Classes

Equipped with the general bound of Theorem 7.3.1, providing regret bounds for specific function classes and action sets is a matter of bounding the eluder dimension $dim_E(\mathcal{F}, \kappa(T))$ and ball width function $\beta_t^*(\mathcal{F}, \delta, \alpha, \lambda)$. In this section we will do so for finite, (generalised) linear, and Lipschitz function classes.

7.4.1 Finite and (Generalised) Linear Function Classes

In the setting of sub-Gaussian reward noise, Russo and Van Roy (2014) provide bounds for $dim_E(\mathcal{F}, T^{-1})$ and the sub-Gaussian version of the ball-width function for three simple function settings: finitely many actions, linear function classes, and generalised linear function classes. We can produce analogous results for these settings under sub-exponential reward noise.

Eluder Dimension

For finite function classes, we may bound the eluder dimension as $dim_E(\mathcal{F}, \epsilon) \leq |\mathcal{A}|$ for all $\epsilon > 0$. For linear reward functions $f_\theta(a) = \theta^T\phi(a)$ where $\theta \in \Theta \subset \mathbb{R}^d$ such that $\mathcal{F} = \{f_\rho, \rho \in \Theta\}$. If there exist constants S and γ, such that $||\rho||_2 \leq S$ and $||\phi(a)||_2 \leq \gamma$ for all $a \in \mathcal{A}$ then the eluder dimension may be bounded as $dim_E(\mathcal{F}, \epsilon) \leq 3d\frac{e}{e-1}\log(3 + 3(\frac{2S}{\epsilon})^2) + 1$. Finally, consider generalised linear reward functions $f_\theta(a) = g(\theta^T\phi(a))$ where again $\theta \in \Theta \subset \mathbb{R}^d$ and $\mathcal{F} = \{f_\rho, \rho \in \Theta\}$, and where $g(\cdot)$ is a differentiable and strictly increasing function. If there exist constants $\underline{h}, \overline{h}, S$ and γ such that for all $\rho \in \Theta$ and $a \in \mathcal{A}$, $0 \leq \underline{h} \leq g'(\rho^T\phi(a)) \leq \overline{h}$, $||\rho||_2 \leq S$, and $||\phi(a)||_2 \leq \gamma$ then the eluder dimension can be bounded as $\dim_E(\mathcal{F}, \epsilon) \leq 3dr^2\frac{e}{e-1}\log(3r^2 + 3r^2(\frac{2S\overline{h}}{\epsilon})^2) + 1$, where $r = \sup_{\tilde{\theta},a} g'(< \phi(a), \tilde{\theta} >)/\inf_{\tilde{\theta},a} g'(< \phi(a), \tilde{\theta} >)$ bounds the ratio between the maximal and minimal slope of g.

Ball Width Function

For finite function classes, and $\alpha = 0$ we have $\beta_n^*(\mathcal{F}, \delta, 0, \lambda) = \frac{\log(|\mathcal{F}|/\delta)}{\lambda(1-2\lambda\sigma^2)}$. For both the class of linear and of generalised linear reward functions we have $\log N(\alpha, \mathcal{F}, || \cdot ||_\infty) = O(d\log(1/\alpha))$ from Russo and Van Roy (2014). It follows that in both cases $\beta_T^*(\mathcal{F}, \delta, 1/T^2, \lambda) = O(d\log(T/\delta))$.

Regret Bounds

As a result, for finite function classes we have,

$$BReg(T, \pi_\theta^{TS}) \leq 1 + (|\mathcal{A}| + 1)C + 4\sqrt{\frac{|\mathcal{A}|\log(2|\mathcal{F}|T)}{\lambda(1 - 2\lambda\sigma^2)}T}. \tag{7.4.8}$$

For linear and generalised linear function classes we have, for $\delta \leq 1/2T$,

$$BReg(T, \pi_\theta^{TS}) = O\Big(d\log(T) + \sqrt{d^2\log(T)\log(T/\delta)T}\ \Big). \tag{7.4.9}$$

The orders, with respect to T, of these bounds match those of Russo and Van Roy's bounds for the sub-Gaussian case, and are optimal up to the small contribution of the logarithmic factors. This relationship between the regret under sub-Gaussian and sub-exponential noise is consistent with the findings of Chapters 4 and 5, and other works on bandits with heavier than sub-Gaussian tailed rewards, in that the heavier tails affect only the coefficients of the regret bound, not its order with respect to T.

7.4.2 Reward Functions with Lipschitz Derivatives

We now present the main contribution of this chapter, a specification of the general Bayesian regret bound to the classes of functions with Lipschitz derivatives. We define the class of C-bounded functions with M L-Lipschitz smooth derivatives on $[0,1]$ as

$$\mathcal{F}_{C,M,L} = \left\{ f : [0,1] \to [0,C] : |f^{(m)}(a) - f^{(m)}(a')| \leq L|a-a'|, \forall a, a' \in [0,1], m \leq M \right\}, \tag{7.4.10}$$

for some $C, L > 0$, and $M \in \mathbb{N}$. Note that when $k = 0$ this is simply the class of bounded Lipschitz functions. Our main result, below, is a bound on the Bayesian regret of TS applied where f_θ is drawn from a prior on $\mathcal{F}_{C,M,L}$. Its proof is given in the following sub-section, Section 7.4.3. As in the case of (generalized) linear functions, it relies on bounding the terms of (7.3.4) which are specific to the function class $\mathcal{F}_{C,M,L}$. Note, that while we have focussed on the case of $\mathcal{A} = [0,1]$ we do not believe that this is the limit in terms of the application of this theory. We believe the techniques used to prove the theorem can be extended at least to $\mathcal{A} = [0,1]^d$ for $d \in \mathbb{N}$, and possibly to general compact $\mathcal{A} \in \mathbb{R}^d$.

Theorem 7.4.1. *For $M \in \mathbb{N}$ and the bandit problem with sub-exponential reward noise, and reward function drawn from a prior on $\mathcal{F}_{C,M,L}([0,1])$ the Bayesian regret of the TS algorithm which uses* *this prior is bounded as*

$$BReg(T, \pi^{TS}) = O(T^{1-\frac{1}{2}\frac{2M^2+3M+2}{2M^2+7M+6}}). \tag{7.4.11}$$

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. Generally, as the number of Lipschitz derivatives $M \to \infty$ the order of the Bayesian regret approaches $O(T^{1/2})$.

Interestingly, the bound (7.4.11) suggests that the performance of the proposed TS approach could be suboptimal when M is small. Recall that for $M = 0$ (i.e. where f_θ is Lipschitz) the lower bound of Kleinberg (2005) is $\Omega(T^{2/3})$ and upper confidence bound approaches can achieve $O(T^{2/3})$ regret. It is not clear from the analysis (neither that leading to our upper bound nor that leading to Kleinberg's lower bound) why there is a discrepancy. 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 $t \in \{1, \dots, T\}$, and in doing so force a certain level of exploration. It could be that the TS algorithm proposed as Algorithm 11 which has access to the entire action set $\mathcal{A}$ somehow carries a greater risk of conducting insufficient exploration.

Another possibility is that the true performance of the TS approach analysed here does match in fact the lower bound, and analysis of Russo and Van Roy (2014) which we have adapted to this setting is too loose in this framework. We notice, for instance, that even if the eluder dimension were to be $O(1)$, the best bound available for $M = 0$ via the general bound of Theorem 7.3.1 would then be $O(T^{3/4})$ because the general bound induces a $\Omega(\sqrt{T})$ result and the ball width function (which is square rooted in the regret bound) is $O(\sqrt{T})$ for $M = 0$ because of the covering number of the function class - a well-established theoretical result.

In the case where $M \to \infty$, the performance of TS would seem to be order optimal. While lower bounds on regret for the setting of reward functions with infinitely many smooth derivatives have not, to the best of our knowledge, explicitly been considered, there are some related results. In settings such as (generalised) linear bandits, the reward function class is contained entirely within $\mathcal{F}_{C,\infty,L}$ (trivially so since higher order derivatives are uniformly zero). The optimal order performance for such problems is $\Omega(\sqrt{T})$. Similarly, in GP-optimisation $O(\sqrt{T})$ regret is achievable, and certain choices of covariance kernel imply that the reward function will always have infinitely many smooth derivatives and lie in a subset of $\mathcal{F}_{C,\infty,L}$. So while further analysis is required to derive a lower bound for the setting where the reward function may be any function in $\mathcal{F}_{C,\infty,L}$, there is a reasonable suggestion that the $O(\sqrt{T})$ that we show the regret of TS converges to is optimal, at least for certain special cases.

7.4.3 Proof of Theorem 7.4.1: Regret under Lipschitz reward functions

As in the case of (generalised) linear functions, the proof of Theorem 7.4.1 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. We prove this result in the following sub-section, Section 7.4.4. 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.

Theorem 7.4.2. *Let $M \in \mathbb{N}$, $C, L, \epsilon > 0$ and $\mathcal{F}_{C,M,L}$ be the class of functions with M L-Lipschitz derivatives as defined in* (7.4.10). *We have the following bound on the ϵ-eluder dimension of $\mathcal{F}_{C,M,L}$,*

$$\dim_E(\mathcal{F}_{C,M,L}, \epsilon) = o((\epsilon/L)^{-1/(M+1)}). \tag{7.4.12}$$

We are interested in the $\kappa(T)$-eluder dimension, and since $\kappa(T)$ is a nondecreasing function of T, $dim_E(\mathcal{F}, \kappa(T))$ will be an increasing function. However, the presence of the $(-1/(M+1))^{th}$ order power means it makes only a minimal contribution to the overall order of regret for M large.

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,

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

Recall the definition of the ball-width function

$$\begin{aligned}\beta_T^*(\mathcal{F},\delta,\alpha,\lambda) := &\frac{\log(N(\alpha,\mathcal{F},||\cdot||_\infty)/\delta)}{\lambda(1-2\lambda\sigma^2)} + \frac{2\alpha T(4C+\alpha)(1-\lambda\sigma^2)}{1-2\lambda\sigma^2}\\ &+ \frac{2\alpha\sum_{i\leq\lfloor N\rfloor}\sqrt{2\sigma^2\log(4i^2/\delta)} + 2\alpha\sum_{i\geq\lceil N\rceil}^{n} 2b\log(4i^2/\delta)}{1-2\lambda\sigma^2},\end{aligned}$$

for (σ^2, b) sub-exponential rewards. 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,

$$\beta_T^*(\mathcal{F}_{C,M,L}, \delta, T^{-(M+1)/(M+2)}, \lambda) = O(T^{1/(M+2)}), M \in \mathbb{N} \tag{7.4.13}$$

as the best result available result.

The proof of Theorem 7.4.1 is then completed by utilising the general bound of Theorem 7.3.1,

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

Choosing $\kappa(T) = T^{-\frac{1}{2}\frac{2M^2+3M+2}{2M^2+7M+6}}$, we have by Theorem 7.4.2 and (7.4.13) that

$$BReg(T, \pi^{TS}) \leq O(T^{1-\frac{1}{2}\frac{2M^2+3M+2}{2M^2+7M+6}}). \quad \square$$

7.4.4 Proof of Theorem 7.4.2: Eluder dimension of Lipschitz function class

To bound the eluder dimesnion, 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 elduer dimension. Functions $g \in \mathcal{G}_{C,M,L}$ also possess M Lipschitz derivatives. Specifically, we have the following result:

Proposition 7.4.3. *All functions $g \in \mathcal{G}_{C,M,L}$ are $[-C, C]$-bounded and possess M $2L$-Lipschitz smooth derivatives.*

Proof of Proposition 7.4.3: We have that any function $g \in \mathcal{G}_{C,M,L}$ is bounded since, $f(a) \in [0, C]$ for all $a \in [0, 1]$. The Lipschitz-smoothness of the m^{th} derivatives can be shown as follows. For any function $g = f - f'$ where $f, f' \in \mathcal{F}_{C,M,L}$, $m = 0, \ldots, M$, and pair of actions $a, a' \in [0, 1]$,

$$\begin{aligned} |g^{(m)}(a) - g^{(m)}(a')| &= |f^{(m)}(a) - f'^{(m)}(a) - f^{(m)}(a') + f'^{(m)}(a')| \\ &\leq |f^{(m)}(a) - f^{(m)}(a')| + |f'^{(m)}(a') - f'^{(m)}(a)| \\ &\leq 2L||a - a'||, \end{aligned}$$

where the first inequality holds by the triangle inequality, and the second by the L-Lipschitz smoothness of the M^{th} derivatives of functions in $\mathcal{F}_{C,M,L}$. □

We may define the eluder dimension in terms of $\mathcal{G}_{C,M,L}$. Doing so will make the definition more compact and also be useful for the proof of Theorem 7.4.2. Let $a_{1:k} \in [0, 1]^k$ denote a sequence of actions $(a_1, \dots, a_k)$ and define

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

We then define the ϵ-eluder dimension as follows:

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

We may now proceed with the proof of the eluder dimension bound.

Proof of Theorem 7.4.2: For any $k \in \mathbb{N}$ and sequence $a_{1:k} \in [0,1]^k$, it follows from the definition of $w_k(a_{1:k}, \epsilon')$ that the event $\{w_k(a_{1:k}, \epsilon') > \epsilon'\}$ implies that there exists $g \in \mathcal{G}_{C,M,L}$ such that $g(a_k) > \epsilon'$ and $\sum_{i=1}^{k-1}(g(a_i))^2 \leq (\epsilon')^2$. Conversely if for all $g \in \mathcal{G}_{C,M,L}$ the event $\{g(a_k) > \epsilon'\}$ is known to imply $\sum_{i=1}^{k-1}(g(a_i))^2 > (\epsilon')^2$, then the event $\{g(a_k) > \epsilon'\}$ also implies that $w_k(a_{1:k}, \epsilon') \leq \epsilon'$. This second idea will be central to proving Theorem 7.4.2.

We will show that for functions $g \in \mathcal{G}_{C,M,L}$ if $g(a_k) > \epsilon'$ then $g^2(b) > (\epsilon')^2/9$ for all b in a certain region around a_k. This is a consequence of functions in $\mathcal{G}_{C,M,L}$ having M smooth derivatives. If g takes value greater than ϵ' at a given point, then it must take relatively large values within a certain neighbourhood of that given point. The size of this neighbourhood is a function of the level of smoothness of g. As M increases, the size of this region where $g^2(b) > (\epsilon')^2/9$ increases. It follows that as M increases, the eluder dimension decreases, because if $g(a_k) > \epsilon'$, the previous actions $a_{1:k-1}$ must be increasingly far from a_k for $\sum_{i=1}^{k-1}(g(a_i))^2 \leq (\epsilon')^2$ to be satisfied.

To be precise about this behaviour and derive the required bound on the eluder dimension, we will first lower bound the size of the neighbourhood in which g must take large absolute values if $g(a) > \epsilon'$ for some $a \in [0,1]$. For a function $g : [0,1] \to [-C, C]$ define the region where it takes absolute value greater than $\epsilon/3$ as

$$B(g) := |\{b : g(b)^2 > \epsilon^2/9\}|. \tag{7.4.14}$$

Then for an action $a \in [0,1]$ define the minimum size of the set such that g^2 must exceed $\epsilon^2/9$ if $g(a) > \epsilon$ and $g \in \mathcal{G}_{C,M,L}$ as

$$B^*_{C,M,L}(a) := \min_{g \in \mathcal{G}_{C,M,L} : g(a) > \epsilon} B(g), \tag{7.4.15}$$

and the set of functions attaining this minimum as

$$\mathcal{G}^*_{C,M,L}(a) = \operatorname*{argmin}_{g \in \mathcal{G}_{C,M,L} : g(a) > \epsilon} B(g). \tag{7.4.16}$$

Bounds on $B^*_{C,M,L}(a)$, derived by identifying and considering the form of functions in $\mathcal{G}^*_{C,M,L}(a)$ will allow us to bound the eluder dimension.

We will first provide lower bounds on $B^*_{C,M,L}$ for the special cases of $M = 0$ and $M = 1$, and then show a general result for $M \geq 2$. In the case of $M = 0$ the lower bound follows from the Lipschitz property of all functions $g \in \mathcal{G}_{C,M,L}$. We give the lower bound on $B^*_{C,0,L}(a)$ for all $a \in [0,1]$ in the following lemma.

Lemma 7.4.4. *For $a \in [0,1]$, and $C, L > 0$ we have $B^*_{C,0,L}(a) \geq \frac{\epsilon}{3L}$.*

Proof of Lemma 7.4.4: We have that $|g(b) - g(b')| \leq 2L||b - b'||$ for all $g \in \mathcal{G}_{C,M,L}$ and $b, b' \in [0,1]$. Thus if $g(a) > \epsilon$ for some $a \in [0,1]$ we have that $(g(b))^2 > \epsilon^2/9$ for all $b \in [0,1] : (\min(0, \epsilon - 2L|a-b|))^2 \geq \epsilon^2/9$, equivalently $b \in [0,1] : |a - b| > \frac{\epsilon}{3L}$. The conclusion that $B_{C,0,L} \geq \frac{\epsilon}{3L}$ then follows immediately. □

The following lemma gives a similar result for the case of $M = 1$. In this case the proof relies on the observation that g', the gradient of a function $g \in \mathcal{G}^*_{C,M,L}(a)$, should satisfy $g'(a) = 0$, i.e.

a should be a maximiser of g. The bound on the size of $B^*_{C,1,L}(a)$ then follows from the Lipschitz property of g'. The result holds only for a sufficiently from the edges of $[0,1]$, since $g'(a)$ need not take value 0 to minimise $|\{b : g^2(b) > (\epsilon')^2/9\}|$ if a is close to an edge. Fortunately, however, the impact of these special edge cases is negligible when it comes to bounding the eluder dimension.

Lemma 7.4.5. *For $a \in [0,1]$ such that $a > \sqrt{\frac{2\epsilon}{3L}}$ and $1-a > \sqrt{\frac{2\epsilon}{3L}}$, and $C, L > 0$ we have $B^*_{C,1,L}(a) \geq 2\sqrt{\frac{2\epsilon}{3L}}$.*

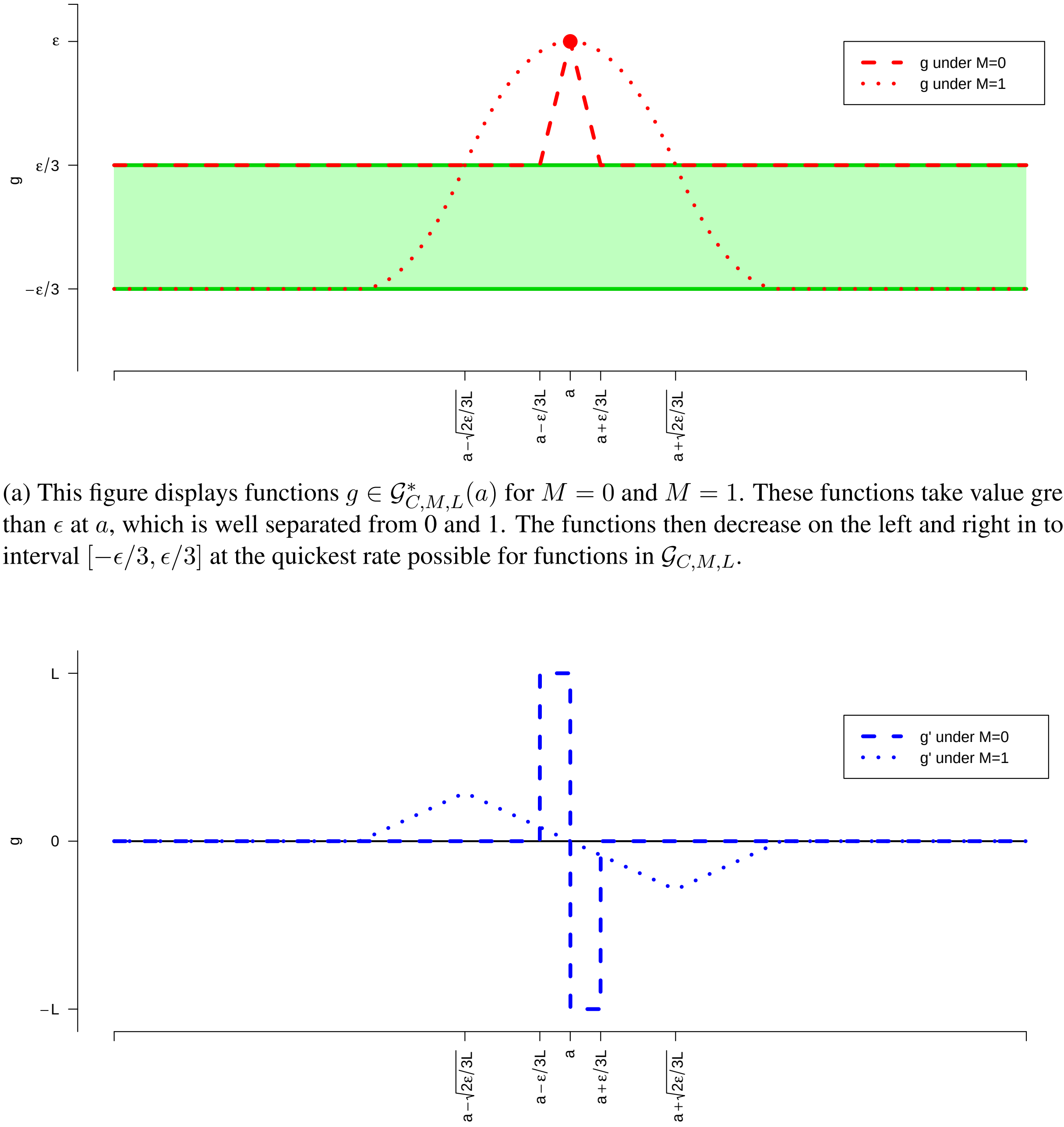
Proof of Lemma 7.4.5: We have that $|g'(b)-g'(b')| \leq 2L||b-b'||$ for all $g \in \mathcal{G}_{C,1,L}$ and $b, b' \in [0,1]$. Thus, for g with $g'(a) = 0$, we have $|g'(b)| \leq 2L||a-b||$ for all $b \in [0,1]$. For any $b' < b \in [0,1]$ we have $g(b) - g(b') = \int_{b'}^{b} g'(x)dx$. It follows that for $b < a \in [0,1]$

$$\begin{aligned} g(b) = g(a) - g(a) + g(b) &= g(a) - \int_b^a g'(x)dx \\ &\geq g(a) - \int_b^a 2L(a-x)dx \\ &= g(a) - La^2 + 2Lab - Lb^2 \\ &> \epsilon' - L(a-b)^2. \end{aligned}$$

A similar argument follows for $b > a \in [0,1]$ and thus $g(b) > \epsilon' - L||a-b||^2$ for all $b \in [0,1]$ given $g(a) > \epsilon'$ and $g'(a) = 0$. It follows that under these conditions we have $g^2(b) > \epsilon^2/9$ for all $b \in [0,1] : (\min(0, \epsilon - L|a-b|^2))^2 \geq \epsilon^2/9$, equivalently $b \in [0,1] : |a-b| \leq \sqrt{\frac{2\epsilon}{3L}}$.

If $g'(a) \neq 0$ then $\exists\, c \in [0,1]$ with $g(c) > g(a) > \epsilon'$ and $g'(c) = 0$. Then by the logic used for the case with $g'(a) = 0$ it follows that $g^2(b) > \epsilon^2/9$ for all $b \in [0,1] : ||b-c|| \leq \sqrt{\frac{1}{L}(g(c) - \epsilon/3)}$. Since $g(c) > \epsilon'$ it follows that if $g(a) > \epsilon'$ then the region such that $g^2(b) > \epsilon^2/9$ is larger if $g'(a) \neq 0$ than if $g(a) = 0$. Thus we have $g'(a) = 0$ for all $g \in \mathcal{G}^*_{C,1,L}(a)$ and $B_{C,1,L}(a) \geq \sqrt{\frac{2\epsilon}{3L}}$ for all $a \in [0,1]$ such that $a > \sqrt{\frac{2\epsilon}{3L}}$ and $1-a > \sqrt{\frac{2\epsilon}{3L}}$. □

Figure 7.4.1a provides an illustration of the bounds on $B^*_{C,M,L}(a)$. When $M = 0$, functions



(a) 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 $\mathcal{G}_{C,M,L}$.

(b) This figure displays the first derivatives of functions $g \in \mathcal{G}^*_{C,M,L}(a)$ for $M = 0$ and $M = 1$.

Figure 7.4.1: Functions $g \in \mathcal{G}_{C,M,L}(a)$ for $M = 0, 1$ and their first derivatives.

in $\mathcal{G}^*_{C,0,L}(a)$ may decrease in value more quickly than when $M = 1$ and the functions in $\mathcal{G}^*_{C,1,L}$ are smoother. As a result the size of the region where g takes value greater than $\epsilon/3$ is larger - i.e. $B^*_{C,M,L}(a)$ is larger. This intuition carries forward as M continues to increase, since the smoothness assumptions implied by the definition of $\mathcal{G}_{C,M,L}$ become increasingly strong.

In Figure 7.4.1b we illustrate the derivatives of the functions in $\mathcal{G}^*_{C,0,L}(a)$ and $\mathcal{G}^*_{C,1,L}(a)$. In the $M = 0$ case, since the first derivative need not be Lipschitz smooth, we have discontinuities at $a \pm \epsilon/3L$. In the $M = 1$ case, $B^*_{C,M,L}$ is much larger because the first derivative is Lipschitz and constrained to change gradually. As M increases, the first derivative of a function $g \in \mathcal{G}^*_{C,M,L}(a)$ will become increasingly smooth because discontinuities in the higher-order derivatives will also not be permitted by the definition of $\mathcal{G}_{C,M,L}$.

Bounding $B^*_{C,M,L}$ for larger values of M is more involved. To do so we will first define a particular function $h_{a,M} \in \mathcal{G}_{C,M,L}$ for each $M \geq 2$ and $a \in [0,1]$ and bound $B(h_{a,M})$. We will then show that this function is in $\mathcal{G}^*_{C,M,L}$, and thus that $B^*_{C,M,L}(a) = B(h_{a,M})$. The form of $h_{a,M}$ will vary depending on whether M is even or odd. We will first specify $h_{a,M}$ for M even.

For $M \geq 2$ even, let $h_{a,M}$ be maximised at a with $h_{a,M}(a) > \epsilon'$, and let $x_{1,M} = x_{1,a,M} = \max_{x<a, h_{a,M}(x)=\epsilon/3} x$ be the point closest to a on the left where $h_{a,M}$ takes value $\epsilon/3$. Define $\Delta_M = a - x_{1,M}$, and then further points $y_{1,M} = x_{1,M} - \Delta_M$, $x_{2,M} = a + \Delta_M$, and $y_{2,M} = a + 2\Delta_M$. We then specify $h_{a,M}$ as a function with M^{th} derivative given as

$$\frac{1}{2L} h^{(M)}_{a,M}(z) = \begin{cases} x_{1,M} - z, & z \in (y_{1,M}, a), \\ z - x_{2,M}, & z \in [a, y_{2,M}), \end{cases} \tag{7.4.17}$$

and whose lower order derivatives satisfy the following properties:

$$h^{(m)}_{a,M}(x_1) = h^{(m)}_{a,M}(x_2) = 0, 2 \leq m \leq M, m \text{ even}, \tag{7.4.18}$$

$$h^{(m)}_{a,M}(y_1) = h^{(m)}_{a,M}(a) = h^{(m)}_{a,M}(y_2) = 0, m \leq M, m \text{ odd}. \tag{7.4.19}$$

Since $h_{a,M}^{(M)}$ is Lipschitz this defines the function that can have $h_{a,M}^{(M)}(x) = 0$ where $h_{a,M}$ crosses $\epsilon/3$ and change most rapidly elsewhere. To bound $B(h_M)$ we first require expressions for the lower order derivatives of h_M. Having the restricted behaviour on $\{y_{1,M}, x_{1,M}, a, x_{2,M}, y_{2,M}\}$ means that these functions can be identified from $h_{a,M}^{(M)}$ alone. The following lemma specifies the form of these lower order derivatives. We focus on the left of a, as a symmetry argument will give an analogous result for the right.

Lemma 7.4.6. *For the function $h_{a,M}$ with M^{th} derivative given by* (7.4.17) *where M is even, the lower order derivatives satisfy*

$$\frac{1}{2L} h_{a,M}^{(M-m)}(z) = \begin{cases} j_{m+1}(x_{1,M}) - j_{m+1}(z), & m \in \{0, 2, 4, \ldots, M\} \\ j_{m+1}(a) - j_{m+1}(z), & m \in \{1, 3, \ldots, M-1\} \end{cases} \qquad z \in (y_{1,M}, a) \quad (7.4.20)$$

where

$$j_k(z) = \sum_{i=1}^{k} \frac{z^i}{i!} (-1)^{k-i} J_{k-i}, \quad k \in \{1, \ldots, M+1\},$$

$$J_k = j_k(a\mathbb{I}\{k \textit{ even}\} + x_1\mathbb{I}\{k \textit{ odd}\}),$$

and $j_0(z) = 1$ for all $z \in (y_1, a)$.

We prove this lemma in Section 7.6.5 using an induction argument and the assumed zeros of the m^{th} derivatives. Since $h_{a,M}$ is unimodal, and symmetric about a, we have $B(h_{a,M}) > x_{2,M} - x_{1,M} = 2(a - x_{1,M}) = 2\Delta_M$. In the following lemma, we determine the order of $B(h_{a,M})$ by bounding Δ_M for each even $M \geq 2$.

Lemma 7.4.7. *For the function $h_{a,M}$ with M^{th} derivative given by* (7.4.17) *where M is even, there* *exist finite constants* $K_{1,M}, K_{2,M} > 0$ *such that*

$$K_{1,M}(\epsilon/L)^{1/(M+1)} \leq B(h_{a,M}) \leq K_{2,M}(\epsilon/L)^{1/(M+1)}.$$

Proof of Lemma 7.4.7: Firstly observe that since $h_{a,M}(x_{1,M}) = \epsilon/3$ we have by definition that

$$h_{a,M}(a) - h_{a,M}(x_{1,M}) = \int_{x_{1,M}}^{a} h'_{a,M}(z)dz > \frac{2\epsilon}{3}.$$

Using the definition of $h'_{a,M}$ in (7.4.20), we expand the centre term of the above display as follows,

$$\begin{aligned}
&\int_{x_{1,M}}^{a} h'_{a,M}(z)dz = \int_{x_{1,M}}^{a} h_{a,M}^{(M-(M-1))}(z)dz \\
&= 2L\int_{x_{1,M}}^{a} j_M(a) - j_M(z)dz \\
&= 2L\int_{x_{1,M}}^{a} j_M(a) - \sum_{i=1}^{M} \frac{z^i}{i!}(-1)^{M-i} j_{M-i}\big(x_{1,M}\mathbb{I}\{M-i \text{ odd}\} + a\mathbb{I}\{M-i \text{ even}\}\big)dz \\
&= 2L\Bigg[j_M(a)z - \sum_{i=1}^{M} \frac{z^{i+1}}{(i+1)!}(-1)^{M-i} j_{M-i}\big(x_{1,M}\mathbb{I}\{M-i \text{ odd}\} + a\mathbb{I}\{M-i \text{ even}\}\big)\Bigg]_{x_{1,M}}^{a} \\
&= 2L\sum_{i=1}^{M}\left(\frac{a^{i+1}}{i!} - \frac{a^{i+1}}{(i+1)!} - \frac{x_{1,M}a^i}{i!} + \frac{x_{1,M}^{i+1}}{(i+1)!}\right)(-1)^{M-i} j_{M-i}\big(x_{1,M}\mathbb{I}\{M-i \text{ odd}\} + a\mathbb{I}\{M-i \text{ even}\}\big) \\
&= 2L\sum_{i\in\{2,4,\ldots,M\}}\left(\frac{a^{i+1}}{i!} - \frac{a^{i+1}}{(i+1)!} - \frac{x_{1,M}a^i}{i!} + \frac{x_{1,M}^{i+1}}{(i+1)!}\right) j_{M-i}(a) \\
&\quad - 2L\sum_{i\in\{1,3,\ldots,M-1\}}\left(\frac{a^{i+1}}{i!} - \frac{a^{i+1}}{(i+1)!} - \frac{x_{1,M}a^i}{i!} + \frac{x_{1,M}^{i+1}}{(i+1)!}\right) j_{M-i}(x_{1,M})
\end{aligned}$$

From the definition of the recurrence relation j, we have that for k even $j_k(a)$ may be written, for some $\kappa_{l,k}$, $l = 1, \ldots k$ as $j_k(a) = \sum_{l=1}^{k} \kappa_{l,k} a^l x_{1,M}^{k-l}$, i.e. for k even $j_k(a)$ is $O(a^k)$ and $O(x_{1,M}^{k-1})$. Similarly for k odd $j_k(x_{1,M})$ may be written, for some $\tau_{l,k}$, $l = 1, \ldots, k$ as $j_k(x_{1,M}) =$

$\sum_{l=1}^{k} \tau_{l,k} x_{1,M}^{l} a^{k-l}$, i.e. for k odd $j_k(x_{1,M})$ is $O(x_{1,M}^k)$ and $O(a^{k-1})$.

It follows from this and the above display, that we may write

$$\int_{x_{1,M}}^{a} h'_{a,M}(z)dz = 2L \sum_{i \in \{2,4,\dots,M\}} \left(\frac{a^{i+1}}{i!} - \frac{a^{i+1}}{(i+1)!} - \frac{x_{1,M}a^i}{i!} + \frac{x_{1,M}^{i+1}}{(i+1)!} \right) \sum_{l=1}^{M-i} \kappa_{l,M-i} a^l x_{1,M}^{M-i-l}$$
$$- 2L \sum_{i \in \{1,3,\dots,M-1\}} \left(\frac{a^{i+1}}{i!} - \frac{a^{i+1}}{(i+1)!} - \frac{x_{1,M}a^i}{i!} + \frac{x_{1,M}^{i+1}}{(i+1)!} \right) \sum_{l=1}^{M-i} \tau_{l,M-i} x_{1,M}^{l} a^{M-i-l},$$

and that there exist constants $H_{M,L,i}$, $i = 0, \dots, M+1$ such that

$$h_{a,M}(a) - h_{a,M}(x_1) = \sum_{i=0}^{M+1} H_{M,L,i} a^{M+1-i} x_{1,M}^{i} = O((a - x_{1,M})^{M+1}).$$

Since $h_{a,M}(a) - h_{a,M}(x_{1,M}) = 2\epsilon/(3L)$ we have that $x_{1,M} = a - o((\epsilon/L)^{1/(M+1)})$. By a symmetry argument about a we will also have that $x_{2,M} = a + o((\epsilon/L)^{1/M+1})$. Furthermore, by symmetry of g' about $x_{1,M}$ and $x_{2,M}$ we have that $h_{a,M}$ need not fall below $-\epsilon/3$, as $y_{1,M}$ and $y_{2,M}$ may be global minimisers of $h_{a,M}$ Thus for $h_{a,M}$ as described above, and $M \geq 2$ even, we have

$$B(h_{a,M}) = 2\Delta_M = o((\epsilon/L)^{1/(M+1)})$$

for all a sufficiently far from the edges of $[0, 1]$. □

Lemmas 7.4.6 and 7.4.7 pertain only to the case where M is even. We must now consider the complementary case of M odd. The function $h_{a,M}$ is different, but the argument used to bound $B(h_{a,M})$ is very similar.

For $M \geq 3$ odd let $h_{a,M}$ be a function in $\mathcal{G}^0_{C,M,L}(a)$ with M^{th} derivative specified as

$$\frac{1}{2L} h^{(M)}_{a,M}(z) = \begin{cases} z - y_{1,M}, & z \in (y_{1,M}, x_{1,M}), \\ a - z, & z \in [x_{1,M}, x_{2,M}), \\ z - y_{2,M}, & z \in [x_{2,M}, y_{2,M}), \end{cases} \tag{7.4.21}$$

and whose lower order derivatives satisfy conditions (7.4.18) and (7.4.19). This is chosen similarly to in the case of M even as the fastest varying function which meets the constraints on the derivatives on $\{y_{1,M}, x_{1,M}, a, x_{2,M}, y_{2,M}\}$. Again, we derive expressions for the lower order derivatives of $h_{a,M}$ and focus on the left of a, since similar expressions follow for the right hand side by symmetry.

Lemma 7.4.8. *For the function $h_{a,M}$ with M^{th} derivative given by* (7.4.21) *where M is odd, the lower order derivatives satisfy*

$$\frac{1}{2L} h^{(M-m)}_{a,M}(z) = \begin{cases} j_{m+1}(z) - J_{m+1}, & z \in (y_{1,M}, x_{1,M}), \\ L_{m+1} - l_{m+1}(z), & z \in [x_{1,M}, a), \end{cases} \tag{7.4.22}$$

where

$$\begin{aligned} j_k(z) &= \sum_{i=1}^{k} \frac{z^i}{i!} (-1)^{k-i} J_{k-i}, \quad z \in (y_{1,M}, x_{1,M}), \\ J_k &= j_k(y_{1,M}\mathbb{I}\{k \textit{ odd}\} + x_{1,M}\mathbb{I}\{k \textit{ even}\}), \\ l_k(z) &= \sum_{i=1}^{k} \frac{z^i}{i!} (-1)^{k-i} L_{k-i}, \quad z \in [x_{1,M}, a) \\ L_k &= l_k(a\mathbb{I}\{k \textit{ odd}\} + x_1\mathbb{I}\{k \textit{ even}\}), \end{aligned}$$

for $k \in \{1, \dots M+1\}$ *and where* $j_0(z) = l_0(z) = 1$ *for all* $z \in (y_{1,M}, a)$.

We prove Lemma 7.4.8 in Section 7.6.6. As in the case of M even, we can use this definition to bound the size of $B(h_{a,M})$, given in the lemma below. The proof is very similar to that of Lemma 7.4.7, and therefore we reserve it for Section 7.6.7.

Lemma 7.4.9. *For the function* $h_{a,M}$ *with* M^{th} *derivative given by* (7.4.21) *where* M *is odd, there exist finite constants* $K_{3,M}, K_{4,M} > 0$ *such that*

$$K_{3,M}(\epsilon/L)^{1/(M+1)} \leq B(h_{a,M}) \leq K_{4,M}(\epsilon/L)^{1/(M+1)}$$

The combined insight from Lemmas 7.4.7 and 7.4.9 is that for any $M \geq 2$ and $a \in [2\Delta_M, 1 - 2\Delta_M]$ there exists a function $h_{a,M} \in \mathcal{G}_{C,M,L}$ with $B(h_{a,M}) = o((\epsilon/L)^{1/(M+1)})$. We will demonstrate that this $o((\epsilon/L)^{1/(M+1)})$ result is optimal, in the sense that $B^*_{C,M,L}(a) = o((\epsilon/L)^{1/(M+1)})$ also.

Firstly, notice that $g'(a) = 0$ necessarily for all $g \in \mathcal{G}^*_{C,M,L}(a)$. If for some $g \in \mathcal{G}_{C,M,L}$ with $g(a) > \epsilon'$, $g'(a) \neq 0$ then either there exists $c \in [0,1]$ such that $g(c) > g(a)$ and $g'(c) = 0$ or else $g(b) > g(a)$ for all b in either $[0,a)$ or $(a,1]$. If the first event happens, by the same theory that says Δ_M is increasing in $g(a)$, there will be a region of width greater than $2\Delta_M$ centred c where $g(b) > \epsilon/3$. If the second event happens, $B(g)$ is plainly greater than $2\Delta_M$ since $a > 2\Delta_M$ and $1 - a > 2\Delta_M$. We therefore deduce that $g'(a) = 0$ for all $g \in \mathcal{G}^*_{C,M,L}(a)$ since $B(h_{a,M}) < B(g)$ for any g with $g(a) > \epsilon'$ and $g'(a) \neq 0$.

Next we observe that $B(h_{a,M})$ is the optimal value of $B(g)$ among functions $g \in \mathcal{G}_{C,M,L}$ with $g(a) > \epsilon'$ and derivatives constrained as in (7.4.18) and (7.4.19). For any such $g \in \mathcal{G}_{C,M,L}$ it is true that $B(g) = x_{2,g} - x_{1,g}$ where $x_{1,g} = \max_{x<a:g(x)=\epsilon/3} x$ and similarly $x_{2,g} = \min_{x>a:g(x)=\epsilon/3} x$. For $h_{a,M}$, we know that $x_{1,h_{a,M}} = a - \Delta_M$ and $x_{2,h_{a,M}} = a + \Delta_M$, thus that $x_{2,h_{a,M}} - x_{1,h_{a,M}} = 2\Delta_M$. The value of Δ_M is determined by $h'_{a,M}$, which we have previously pointed out changes at the fastest rate possible for a function with derivatives constrained according to (7.4.18) and (7.4.19). Thus for any other function g with derivatives constrained according to (7.4.18) and (7.4.19), $x_{2,g} - x_{1,g} \geq 2\Delta_M$ and $B(g) \geq B(h_{a,M})$.

On the other hand, functions whose derivatives are not constrained according to (7.4.18) and (7.4.19) may have $x_{2,g} - x_{1,g} < 2\Delta_M$. However, such functions will take value less than $-\epsilon/3$ at some points in $[0, 1]$. That is to say $B(g) \neq x_{2,g} - x_{1,g}$ for such functions, since $y_{1,g}$ and $y_{2,g}$ cannot not be global minimisers. We will show that $B(g) > B(h_{a,M})$ for functions $g \in \mathcal{G}_{C,M,L}$ with $g(a) > \epsilon$ and $x_{2,g} - x_{1,g} > 2\Delta_M$.

As before, we will consider the left hand side of a and allow the behaviour on the right hand to be explained by a symmetry argument. If, for a function $g \in \mathcal{G}_{C,M,L}$ with $g(a) > \epsilon'$ and $g'(a) = 0$ (otherwise it would not be optimal anyway) we have $x_{1,g} > x_{1,M}$ - i.e. the point on the left where g takes value $\epsilon/3$ is nearer to a than under $h_{a,M}$ - then we have that $\int_{x_{1,g}}^{a} g'(z)dz > \int_{x_{1,g}}^{a} h'_M(z)dz$. Since $g'(a) = h'_{a,M}(a) = 0$, this implies that $g'(y_{1,g}) = 0$ is not possible. There instead exists a point $y_{1,min} < y_{1,g}$ with $g(y_{1,min}) < -\epsilon/3$ and $g'(y_{1,min}) = 0$. The contribution to $B(g)$ from the left side of a is then at least $a - x_{1,g} + 2(y_{1,g} - y_{1,min})$. $y_{1,g} - y_{1,min} = x_{1,g} - x_{1,M}$ by the smoothness properties of functions in $\mathcal{G}_{C,M,L}$ and thus the contribution to $B(g)$ from the left of a will be greater than that of $B(h_{a,M})$. A similar result follows on the right of a, and we thus have that $B(g) > B(h_{a,M})$ for functions with $x_{2,g} - x_{1,g} < 2\Delta_M$. If $x_{2,g} - x_{1,g} > 2\Delta_M$ then the function g is obviously not optimal.

By showing that $h_{a,M}$ is optimal amongst functions with similarly constrained derivatives, and that $B(h_{a,M}) \leq B(g)$ for functions g without these constraints, we have therefore demonstrated that $B^*_{C,M,L}(a) = o((\epsilon/L)^{1/(M+1)})$ for $a \in [2\Delta_M, 1 - 2\Delta_M]$.

We complete the proof of Theorem 7.4.2 by noticing that if $k = 9/B^*_{C,M,L} + 2$ then for any sequence $a_{1:k} \in [0, 1]$ there must exist an index $j \in \{1, \dots, k\}$ such that $a_j \in [2\Delta_M, 1 - 2\Delta_M]$ and there exist distinct at least 9 distinct points a_{l_i}, $l_i \in \{1, \dots, j-1\}$, $i = 1, \dots, 9$ with $|a_j - a_{l_i}| \leq$

$B^*_{C,M,L}/2$. Then if $g(a_j) > \epsilon'$ and $g \in \mathcal{G}_{C,M,L}$ it follows that $(g(a_{l_i}))^2 > (\epsilon')^2/9$ for $i \in \{1, \dots 9\}$ and $\sum_{i=1}^{j-1}(g(a_i))^2 > (\epsilon')^2$.

Therefore if $k \geq 9/B^*_{C,M,L} + 2$ there exists no sequence $a_{1:k} \in [0,1]^k$ such that $w_\tau(a_{1:\tau}, \epsilon') > \epsilon'$ for every $\tau \leq k$, and thus $dim_E(\mathcal{F}_{C,M,L}, \epsilon) \leq k = o((\epsilon/L)^{1/(M+1)})$. □

7.5 Conclusion

The work in this chapter 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 subexponential. We achieved these results by utilising two general notions, first stated in Russo and Van Roy (2014): that the Bayesian regret of Thompson Sampling can be bounded in terms of any valid upper confidence bound sequence, and that the least squares estimator possesses a general theory of its convergence which can be applied for many function classes.

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. Furthermore, by considering sub-exponential reward noise, as opposed to the common sub-Gaussian assumption, these results are applicable to settings where the reward distribution may have somewhat heavier tails - such as applications in finance, or our own in Poisson process event detection.

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.

7.6 Further Proofs

7.6.1 Proof of Lemma 7.3.2

Consider random variables $(Z_i|i \in \mathbb{N})$ adapted to the filtration $(\mathcal{H}_n : n = 0, 1, ...)$. Assume that $\mathbb{E}(e^{\lambda Z_i})$ is finite for $\lambda \geq 0$, and define the conditional mean $\mu_i = \mathbb{E}(Z_i|\mathcal{H}_{i-1})$ and conditional cumulant generating function of the centred random variable $[Z_i - \mu_i]$ as $\psi_i(\lambda) = \log \mathbb{E}(\exp(\lambda[Z_i - \mu_i])|\mathcal{H}_{i-1})$. Let

$$M_n(\lambda) = \exp\left\{ \sum_{i=1}^{n} \lambda[Z_i - \mu_i] - \psi_i(\lambda) \right\}. \tag{7.6.23}$$

We then have by Lemmas 6 and 7 of Russo and Van Roy (2014) that $(M_n(\lambda)|n \in \mathbb{N})$ is a martingale, $\mathbb{E}(M_n(\lambda)) = 1$ for all n, and that for all $x \geq 0$, and $\lambda \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \lambda Z_i \leq x + \sum_{i=1}^{n} [\lambda \mu_i + \psi_i(\lambda)], \ \ \forall n \in \mathbb{N} \right) \geq 1 - e^{-x}. \tag{7.6.24}$$

We may use this result to build a confidence ball for the generic bandit problem with sub-exponential noise. These confidence balls will be expressed in terms of least squares function estimators. Define

$$\begin{aligned} Z_t &= (f_\theta(A_t) - R_t)^2 - (f(A_t) - R(t))^2 \\ &= -(f(A_t) - f_\theta(A_t))^2 + 2(f(A_t) - f_\theta(A_t))\epsilon_i. \end{aligned}$$

The conditional mean and conditional cumulant generating function of the centred version of Z_i

are as follows:

$$\mu_i = \mathbb{E}(Z_i|\mathcal{H}_{i-1}) = -(f(A_i) - f_\theta(A_i))^2, \tag{7.6.25}$$

$$\psi_i(\lambda) = \log \mathbb{E}(\exp(\lambda[Z_i - \mu_i]|\mathcal{H}_{i-1}) = \log \mathbb{E}(\exp(2\lambda(f(A_i) - f_\theta(A_i))\epsilon_i)|\mathcal{H}_{i-1}). \tag{7.6.26}$$

Therefore, by the sub-exponentiality assumption we have that

$$\psi_i(\lambda) \leq \frac{4\lambda^2(f(A_i) - f_\theta(A_i))^2\sigma^2}{2}, \quad \text{for } |\lambda| \leq b^{-1}.$$

Thus by (7.6.24), (7.6.26), and the observation that $\sum_{i=1}^n Z_i = L_{2,n+1}(f_\theta) - L_{2,n+1}(f)$ by definition,

$$\mathbb{P}\bigg(L_{2,n+1}(f_\theta) - L_{2,n+1}(f) \leq \frac{x}{\lambda} + (2\lambda\sigma^2 - 1)\sum_{i=1}^n (f(A_i) - f_\theta(A_i))^2, \ \forall n \in \mathbb{N}\bigg) \geq 1 - e^{-x},$$

for all λ with $|\lambda| \leq b^{-1}$. Substituting $\lambda =$ and $x = \log(1/\delta)$, we have

$$\mathbb{P}\bigg(L_{2,n+1}(f) \geq L_{2,n+1}(f_\theta) + (1 - 2\lambda\sigma^2)\sum_{i=1}^n (f(A_i) - f_\theta(A_i))^2 - \frac{\log(1/\delta)}{\lambda}, \ \forall n \in \mathbb{N}\bigg) \geq 1 - \delta, \tag{7.6.27}$$

for all λ with $|\lambda| \leq b^{-1}$, completing the proof. □

7.6.2 Proof of Proposition 7.3.3

Let $\mathcal{F}^\alpha$ be an α-covering of $\mathcal{F}$ in the sense that for any $f \in \mathcal{F}$ there is an $f^\alpha \in \mathcal{F}^\alpha$ such that $||f^\alpha - f||_\infty \leq \alpha$. Then by Lemma 7.3.2 and a union bound over $\mathcal{F}^\alpha$ we have with probability at least $1-\delta$,

$$L_{2,n+1}(f^\alpha)-L_{2,n+1}(f_\theta) \geq (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f^\alpha(A_i)-f_\theta(A_i))^2-\frac{1}{\lambda}\log\left(\frac{|\mathcal{F}^\alpha|}{\delta}\right), \ \ \forall n \in \mathbb{N}, \ \ \forall f^\alpha \in \mathcal{F}^\alpha.$$

Then, by some simple addition and subtraction, we have for all $f \in \mathcal{F}$, with probability at least $1-\delta$,

$$\begin{aligned}L_{2,n+1}(f) - L_{2,n+1}(f_\theta) \geq{} & (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f(A_i)-f_\theta(A_i))^2 - \frac{1}{\lambda}\log\left(\frac{|\mathcal{F}^\alpha|}{\delta}\right)\\ & + L_{2,n+1}(f) - L_{2,n+1}(f^\alpha)\\ & + (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f^\alpha(A_i)-f_\theta(A_i))^2 - (f(A_i)-f_\theta(A_i))^2, \ \ \forall n \in \mathbb{N}, \ \ \forall f^\alpha \in \mathcal{F}^\alpha.\end{aligned}$$

We may get the tightest version of this bound by introducing a minimum over the α-covering $\mathcal{F}^\alpha$, giving the result that for all $f \in \mathcal{F}$, with probability at least $1-\delta$,

$$\begin{aligned}L_{2,n+1}(f) - L_{2,n+1}(f_\theta) \geq{} & (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f(A_i)-f_\theta(A_i))^2 - \frac{1}{\lambda}\log\left(\frac{|\mathcal{F}^\alpha|}{\delta}\right)\\ & + \min_{f^\alpha\in\mathcal{F}^\alpha}\Bigg\{ L_{2,n+1}(f) - L_{2,n+1}(f^\alpha)\\ & \quad + (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f^\alpha(A_i)-f_\theta(A_i))^2 - (f(A_i)-f_\theta(A_i))^2\Bigg\}, \ \ \forall n \in \mathbb{N}.\end{aligned}$$

We refer to the term $\min_{f^\alpha\in\mathcal{F}^\alpha}\big\{L_{2,n+1}(f) - L_{2,n+1}(f^\alpha) + (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f^\alpha(A_i)-f_\theta(A_i))^2 - (f(A_i)-f_\theta(A_i))^2\big\}$ as the *discretisation error*. The following result gives a high-probability bound on its absolute value.

Lemma 7.6.1. *If f^α satisfies $||f - f^\alpha||_\infty \leq \alpha$, and $|\lambda| \leq b^{-1}$, then with probability at least $1-\delta$,*

$$\left| L_{2,n+1}(f) - L_{2,n+1}(f^\alpha) + (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f^\alpha(A_i) - f_\theta(A_i))^2 - (f(A_i) - f_\theta(A_i))^2 \right|$$
$$\leq 2\alpha n(4C+\alpha)(1-\lambda\sigma^2) + 2\alpha \sum_{i\leq\lfloor n_0\rfloor} \sqrt{2\sigma^2\log(4i^2/\delta)} + 2\alpha \sum_{i\geq\lceil n_0\rceil}^{n} 2b\log(4i^2/\delta),$$

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

By the definition of the least squares estimator, $L_{2,n+1}(\hat{f}_n^{LS}) \leq L_{2,n+1}(f_\theta)$. Therefore, with probability at least $1-2\delta$

$$(1-2\lambda\sigma^2)\sum_{i=1}^{n}(\hat{f}_n^{LS}(A_i) - f_\theta(A_i))^2 \leq \frac{1}{\lambda}\log\left(\frac{|\mathcal{F}^\alpha|}{\delta}\right) + 2\alpha n(4C+\alpha)(1-\lambda\sigma^2)$$
$$+ 2\alpha \sum_{i\leq\lfloor n_0\rfloor} \sqrt{2\sigma^2\log(4i^2/\delta)} + 2\alpha \sum_{i\geq\lceil n_0\rceil}^{n} 2b\log(4i^2/\delta)$$

for n_0 as defined in Lemma 7.6.1. Thus taking the infimum over the size of α-covers we have, with probability at least $1-2\delta$,

$$\sum_{i=1}^{n}(\hat{f}_n^{LS}(A_i) - f_\theta(A_i))^2 \leq \frac{\log(N(\alpha,\mathcal{F},||\cdot||_\infty)/\delta)}{\lambda(1-2\lambda\sigma^2)} + \frac{2\alpha n(4C+\alpha)(1-\lambda\sigma^2)}{1-2\lambda\sigma^2}$$
$$+ \frac{2\alpha\sum_{i\leq\lfloor n_0\rfloor}\sqrt{2\sigma^2\log(4i^2/\delta)} + 2\alpha\sum_{i\geq\lceil n_0\rceil}^{n} 2b\log(4i^2/\delta)}{1-2\lambda\sigma^2}$$

as required. □

7.6.3 Proof of Lemma 7.6.1

As in the proof of Lemma 8 of Russo and Van Roy (2014) we have

$$\begin{aligned}|(f^{\alpha}(a)-f_{\theta}(a))^2-(f(a)-f_{\theta}(a))^2| &\leq 4C\alpha+\alpha^2\\ |(R_i-f(a))^2-(R_i-f^{\alpha}(a))^2| &\leq 2\alpha|R_i|+2C\alpha+\alpha^2\end{aligned}$$

for all $a\in\mathcal{A}$ and $\alpha\in[0,C]$. Then summing over time, we have that

$$\begin{aligned}&\left|L_{2,n+1}(f)-L_{2,n+1}(f^{\alpha})+(1-2\lambda\sigma^2)\sum_{i=1}^{n}(f^{\alpha}(A_i)-f_{\theta}(A_i))^2-(f(A_i)-f_{\theta}(A_i))^2\right|\\ &\leq\sum_{i=1}^{n}(1-2\lambda\sigma^2)(4C\alpha+\alpha^2)+2\alpha|R_i|+2C\alpha+\alpha^2\\ &\leq\sum_{i=1}^{n}(1-2\lambda\sigma^2)(4C\alpha+\alpha^2)+2\alpha(C+|\epsilon_i|)+2C\alpha+\alpha^2\\ &=\sum_{i=1}^{n}2(4C\alpha+\alpha^2)(1-\lambda\sigma^2)+2\alpha|\epsilon_i|.\end{aligned}$$

Since ϵ_i is (σ^2,b)-sub-exponential we have the following exponential bound

$$\mathbb{P}(|\epsilon_i|\geq x)\leq\begin{cases}2\exp(-x^2/2\sigma^2) & \text{if } 0\leq x\leq\sigma^2/b\\ 2\exp(-x/2b) & \text{if } x>\sigma^2/b.\end{cases}$$

Then, by the independence of reward noises, and union bound:

$$\begin{aligned}\mathbb{P}\Bigg(\exists i\in\mathbb{N}:|\epsilon_i|\geq\sqrt{2\sigma^2\log(4i^2/\delta)}\mathbb{I}\{i:\sqrt{2\sigma^2\log(4i^2/\delta)}\leq\sigma^2/b\}&\\ +2b\log(4i^2/\delta)\mathbb{I}\{i:2b\log(4i^2/\delta)>\sigma^2/b\}\Bigg)&\end{aligned}$$

$$\leq \frac{\delta}{2}\sum_{i=1}^{\infty}\frac{1}{i^2} \leq \delta.$$

Thus, with probability at least $1-\delta$,

$$\begin{aligned}
&\left| L_{2,n+1}(f) - L_{2,n+1}(f^{\alpha}) + (1-2\lambda\sigma^2)\sum_{i=1}^{n}(f^{\alpha}(A_i) - f_{\theta}(A_i))^2 - (f(A_i) - f_{\theta}(A_i))^2 \right| \\
&\leq \sum_{i=1}^{n} 2(4C\alpha + \alpha^2)(1-\lambda\sigma^2) \\
&\qquad + 2\alpha\left(\sqrt{2\sigma^2\log\left(\frac{4i^2}{\delta}\right)}\mathbb{I}\left\{\log\left(\frac{4i^2}{\delta}\right) \leq \frac{\sigma^2}{2b^2}\right\} + 2b\log\left(\frac{4i^2}{\delta}\right)\mathbb{I}\left\{\log\left(\frac{4i^2}{\delta}\right) > \frac{\sigma^2}{2b^2}\right\}\right) \\
&= 2\alpha n(4C+\alpha)(1-\lambda\sigma^2) \\
&\quad + 2\alpha\sum_{i=1}^{n}\left(\sqrt{2\sigma^2\log\left(\frac{4i^2}{\delta}\right)}\mathbb{I}\left\{i \leq \sqrt{\frac{\delta}{4}\exp\frac{\sigma^2}{2b^2}}\right\} + 2b\log\left(\frac{4i^2}{\delta}\right)\mathbb{I}\left\{i > \sqrt{\frac{\delta}{4}\exp\frac{\sigma^2}{2b^2}}\right\}\right)
\end{aligned}$$

and the required result follows. □

7.6.4 Proof of Lemma 7.3.4

The proof of Lemma 7.3.4 depends the following proposition of Russo and Van Roy (2014).

Proposition 7.6.2 (Proposition 8 of Russo and Van Roy (2014)). *If* $(\beta_t \geq 0 | t \in \mathbb{N})$ *is a nondecreasing sequence and* $\mathcal{F}_t := \{f \in \mathcal{F} : ||f - \hat{f}_t^{LS}||_{2,E_t} \leq \sqrt{\beta_t}\}$ *then*

$$\sum_{t=1}^{T}\mathbb{I}\{w_{\mathcal{F}_t}(A_t) > \epsilon\} \leq \left(\frac{4\beta_T}{\epsilon}+1\right)\dim_E(\mathcal{F},\epsilon) \tag{7.6.28}$$

for all $T \in \mathbb{N}$ *and* $\epsilon > 0$.

Now, define $w_t = w_{\mathcal{F}_t}(A_t)$ and reorder the sequence $(w_1, \ldots, w_T) \rightarrow (w_{i_1}, \ldots, w_{i_T})$ in descending order such that $w_{i_1} \geq w_{i_2} \geq \cdots \geq w_{i_T}$. We have

$$\begin{aligned}\sum_{t=1}^{T} w_{\mathcal{F}_t}(A_t) &= \sum_{t=1}^{T} w_{i_t} \\ &= \sum_{t=1}^{T} w_{i_t}\mathbb{I}\{w_{i_t} \leq \kappa(T)\} + \sum_{t=1}^{T} w_{i_t}\mathbb{I}\{w_{i_t} > \kappa(T)\} \\ &\leq T\kappa(T) + \sum_{t=1}^{T} w_{i_t}\mathbb{I}\{w_{i_t} > \kappa(T)\}.\end{aligned}$$

As a consequence of $(w_{i_1}, \ldots, w_{i_T})$ being arranged in descending order we have for $t \in [T]$ that $w_{i_t} > \epsilon \Rightarrow \sum_{k=1}^{t} \mathbb{I}\{w_{\mathcal{F}_k}(A_k) > \epsilon\} \geq t$. By Proposition 7.6.2, $w_{i_t} > \epsilon$ is only possible if $t \leq \left(\frac{4\beta_T}{\epsilon}+1\right) \dim_E(\mathcal{F}, \epsilon)$. Furthermore, $\epsilon \geq \kappa(T) \Rightarrow \dim_E(\mathcal{F}, \epsilon) \leq \dim_E(\mathcal{F}, \kappa(T))$ since $\dim_E(\mathcal{F}, \epsilon')$ is non-increasing in ϵ'. Therefore if $w_{i_t} > \epsilon \geq \kappa(T)$ we have that $t < \left(\frac{4\beta_T}{\epsilon} + 1\right) \dim_E(\mathcal{F}, \epsilon)$, i.e. $\epsilon^2 \leq \sqrt{\frac{4\beta_T \dim_E(\mathcal{F},\kappa(T))}{t-\dim_E(\mathcal{F},\kappa(T))}}$. Thus, if $w_{i_t} > \kappa(T) \Rightarrow w_{i,t} \leq \min(C, \sqrt{\frac{4\beta_T \dim_E(\mathcal{F},\kappa(T))}{t-\dim_E(\mathcal{F},\kappa(T))}})$, and finally

$$\begin{aligned}\sum_{t=1}^{T} w_{i_t}\mathbb{I}\{w_{i_t} > \kappa(T)\} &\leq \dim_E(\mathcal{F}, \kappa(T))C + \sum_{t=\dim_E(\mathcal{F},\kappa(T))+1}^{T} \sqrt{\frac{4\beta_T \dim_E(\mathcal{F}, \kappa(T))}{t - \dim_E(\mathcal{F}, \kappa(T))}} \\ &\leq \dim_E(\mathcal{F}, \kappa(T))C + 2\sqrt{\beta_T \dim_E(\mathcal{F}, \kappa(T))} \int_{t=0}^{T} \frac{1}{\sqrt{t}}dt \\ &\leq \dim_E(\mathcal{F}, \kappa(T))C + 4\sqrt{\beta_T \dim_E(\mathcal{F}, \kappa(T))T}. \quad \square\end{aligned}$$

7.6.5 Proof of Lemma 7.4.6:

We prove this Lemma via an induction argument over m. Firstly, for $m = 1$, we have $\frac{1}{2L}h^{(M-m)}(z) = \frac{1}{2L}h^{(M-1)}(z) = \int x_1 - z dz = x_1 z - z^2/2 + D$. Since $M-1$ is odd and $h \in \mathcal{G}^0_{C,M,L}(a)$ we have that $h^{(M-1)}(a) = 0$ and the integration constant, D, must be $a^2/2 - x_1 a$, i.e. we have

$$\frac{1}{2L}h^{(M-1)}(z) = x_1 z - z^2/2 + a^2/2 - a x_1 = j_2(a) - j_2(z).$$

Second, for some m' with $2 \leq m' < M$ let us assume that

$$\frac{1}{2L}h^{(M-m')}(z) = J_{m'+1} - j_{m'+1}(z) \quad z \in (y_1, a).$$

Finally we consider we consider $h^{(M-m'-1)}$. We have,

$$\begin{aligned}
&\frac{1}{2L}h^{(M-m'-1)}(z) \\
&= \int J_{m'+1} - j_{m'+1}(z)dz \\
&= \int \sum_{i=1}^{m'+1} \frac{\big(x_1\mathbb{I}\{m'+1 \text{ odd}\} + a\mathbb{I}\{m'+1 \text{ even}\}\big)^i - z^i}{i!}(-1)^{m'+1-i}J_{m'+1-i}dz \\
&= \sum_{i=1}^{m'+1} \frac{z\big(x_1\mathbb{I}\{m'+1 \text{ odd}\} + a\mathbb{I}\{m'+1 \text{ even}\}\big)^i}{i!}(-1)^{m'+1-i}J_{m'+1-i} \\
&\qquad - \sum_{i=1}^{m'+1} \frac{z^{i+1}}{(i+1)!}(-1)^{m'+1-i}J_{m'+1-i} + D \\
&= \sum_{i=1}^{m'+1} \frac{z\big(x_1\mathbb{I}\{m'+1 \text{ odd}\} + a\mathbb{I}\{m'+1 \text{ even}\}\big)^i}{i!}(-1)^{m'+1-i}J_{m'+1-i} \\
&\qquad - \sum_{i=1}^{m'+1} \frac{z^{i+1}}{(i+1)!}(-1)^{m'+1-i}J_{m'+1-i} + \sum_{i=1}^{m'+1} \frac{\big(x_1\mathbb{I}\{m' \text{ odd}\} + a\mathbb{I}\{m' \text{ even}\}\big)^{i+1}}{(i+1)!}(-1)^{m'+1-i}J_{m'+1-i} \\
&\qquad - \sum_{i=1}^{m'+1} \frac{(x_1\mathbb{I}\{m' \text{ odd}\} + a\mathbb{I}\{m' \text{ even}\})\big(x_1\mathbb{I}\{m'+1 \text{ odd}\} + a\mathbb{I}\{m'+1 \text{ even}\}\big)^i}{i!}(-1)^{m'+1-i}J_{m'+1-i} \\
&= zJ_{m'+2-1} - \sum_{s=2}^{m'+2} \frac{z^s}{s!}(-1)^{m'+2-s}J_{m'+2-s} - (x_1\mathbb{I}\{m'+2 \text{ odd}\} + a\mathbb{I}\{m'+2 \text{ even}\})J_{m'+2-1} \\
&\qquad + \sum_{s=2}^{m'+2} \frac{\big(x_1\mathbb{I}\{m'+2 \text{ odd}\} + a\mathbb{I}\{m'+2 \text{ even}\}\big)^s}{s!}(-1)^{m'+2-s}J_{m'+2-s} \\
&= \sum_{s=1}^{m'+2} \frac{\big(x_1\mathbb{I}\{m'+2 \text{ odd}\} + a\mathbb{I}\{m'+2 \text{ even}\}\big)^s - z^s}{s!}(-1)^{m'+2-s}J_{m'+2-s}
\end{aligned}$$

$$= J_{m'+2} - j_{m'+2}(z)$$

The first equality uses the assumed form of $h^{(M-m')}$, the fourth evaluates the integration constant D based on the knowledge that if $m'+1$ is odd, we will have $h^{(M-m'-1)}(a) = 0$ and if $m'+1$ is even, we will have $h^{(M-m'-1)}(x_1) = 0$, and the fifth uses a change of variable $s = i+1$. □

7.6.6 Proof of Lemma 7.4.8:

As in the case of M even, we prove this lemma via an induction argument over m. Firstly, for $m = 1$ we have for $z \in (y_1, x_1)$, $\frac{1}{2L}h^{(M-1)}(z) = \int z - ydz = z^2/2 - yz + D$. Since $M-1$ is even and $h \in \mathcal{G}^0_{C,M,L}(a)$ we have that $h^{(M-1)}(x_1) = 0$ and the integration constant, D, must be $yx_1 - x_1^2/2 = -J_2$. For $z \in [x_1, a)$, $\frac{1}{2L}h^{(M-1)}(z) = \int a - zdz = az - z^2/2 + D$, and $D = x_1^2/2 - ax_1 = L_2$. Thus,

$$\frac{1}{2L}h^{(M-1)}(z) = \begin{cases} j_2(z) - J_2, & z \in (y_1, x_1) \\ L_2 - l_2(z), & z \in [x_1, a). \end{cases}$$

Secondly, for some m', $2 \leq m' < M$ we assume that

$$\frac{1}{2L}h^{(M-m')}(z) = \begin{cases} j_{m'+1}(z) - J_{m'+1}, & z \in (y_1, x_1) \\ L_{m'+1} - l_{m'+1}(z), & z \in [x_1, a). \end{cases}$$

We now consider $h^{(M-m'-1)}$. For $z \in (y_1, x_1)$ we have,

$$\frac{1}{2L}h^{(M-m'-1)}(z)$$
$$= \int j_{m'+1}(z) - J_{m'+1}dz$$

$$
\begin{aligned}
&= \int \sum_{i=1}^{m'+1} \frac{z^i - \big(y_1 \mathbb{I}\{m'+1 \text{ odd}\} + x_1 \mathbb{I}\{m'+1 \text{ even}\}\big)^i}{i!} (-1)^{m'+1-i} J_{m'+1-i} dz \\
&= \sum_{i=1}^{m'+1} \left(\frac{z^{i+1}}{(i+1)!} - \frac{z\big(y_1 \mathbb{I}\{m'+1 \text{ odd}\} + x_1 \mathbb{I}\{m'+1 \text{ even}\}\big)^i}{i!} \right) (-1)^{m'+1-i} J_{m'+1-i} + D \\
&= \sum_{s=2}^{m'+2} \frac{z^s}{s!} (-1)^{m'+2-s} J_{m'+2-s} - z J_{m'+2-1} + (y_1 \mathbb{I}\{m'+2 \text{ odd}\} + x_1 \mathbb{I}\{m'+2 \text{ even}\}) J_{m'+2-1} \\
&\qquad - \sum_{s=2}^{m'+2} \frac{\big(y_1 \mathbb{I}\{m'+2 \text{ odd}\} + x_1 \mathbb{I}\{m'+2 \text{ even}\}\big)^s}{s!} (-1)^{m'+2-s} J_{m'+2-s} \\
&= j_{m'+2}(z) - J_{m'+2}
\end{aligned}
$$

This follows the same steps as the proof for M even, but with the opposite sign and slightly different definition of j. The proof for $z \in [x_1, a)$ follows the same steps as the above and the proof for M even. The required result follows by induction. □

7.6.7 Proof of Lemma 7.4.9

By the definition of $x_{1,M}$ we have $h_{a,M}(a) - h_{a,M}(x_{1,M}) = \int_{x_{1,M}}^{a} h'_{a,M}(z)dz > 2\epsilon/3$. We rewrite the LHS of this relation as follows,

$$
\begin{aligned}
\int_{x_{1,M}}^{a} h'_{a,M}(z)dz &= 2L \int_{x_{1,M}}^{a} L_M - l_M(z)dz \\
&= 2L \left[L_M z - \sum_{i=1}^{M} \frac{z^{i+1}}{(i+1)!} (-1)^{M-i} L_{m-i} \right]_{z=x_{1,M}}^{a} \\
&= 2L \sum_{i=1}^{M} \left(\frac{a^{i+1}}{i!} - \frac{a^{i+1}}{(i+1)!} - \frac{x_{1,M} a^i}{i!} + \frac{x_{1,M}^{i+1}}{(i+1)!} \right) (-1)^{M-i} L_{M-i}.
\end{aligned}
$$

This is the same expression derived for $h_{a,M}(a) - h_{a,M}(x_{1,M})$ as in the M even case, and thus the same conclusion follows. □

Chapter 8

Conclusions

In this closing chapter, we review the contributions of this thesis and discuss some open problems and opportunities for further work that have been uncovered during the process of this research.

8.1 Contributions

In Chapter 1 we presented the sequential event detection problem. This is a problem with applications across multiple disciplines and industries, arising when point process data need to be observed, resource needs to be intelligently allocated to do so, and there is the opportunity to receive feedback on the quality of previous resource allocations and improve them. We outlined that a successful strategy to solve such a problem must combine three effective components - an inference scheme; an optimisation approach; and a policy to balance exploration and exploitation.

In this thesis we have tackled the challenge of designing, implementing, and analysing such strategies. We developed useful models which capture the challenges of these problems in a variety of settings. We have proposed some simple but powerful algorithms which are tailored to the problem and readily implementable. We have conducted detailed theoretical and empirical analysis of these algorithms, showing their efficacy and contributing to the broader understanding of problems in sequential decision making, applied probability, and Bayesian non-parametrics. We summarise these contributions in further detail below.

Models

We are aware of no formal models for *sequential* event detection problems existing prior to this work. We have proposed two widely applicable models of the problem, as multi-armed bandit problems which have been useful tools for our algorithm design and analysis, and which we hope will provide value for further research in these areas and beyond.

Specifically, in Chapter 4 we proposed a combinatorial bandit model of the problem with a discrete action space. This model captures the problem of deploying multiple sensors to disjoint subintervals and allows us to model filtering of events with a wide range of detection probability functions. Filtering refers to the phenomenon where the events may be detected or not probabilistically - for instance because the observation quality depends on the allocation of resource. The formulation also permits efficient solution via integer programming.

In Chapter 5 we proposed a continuum-armed bandit model of the problem. This version is similar in that it captures the problem of deploying multiple sensors to disjoint subintervals but it is more flexible in that it allows these subintervals to start and end at any point along an observable region, so long as they are nonoverlapping. The model allows for a flexible cost of searching to be included, letting the user capture the trade-off between increasing the size of the observable region and increasing the cost of their surveillance.

Algorithms

For the two models described above, we have combined effective inference, optimisation, and bandit methods to produce provably useful solution algorithms.

For the combinatorial model with filtering, we proposed an upper confidence bound approach which is adapted to the concentration of the inference on the rate function under filtering. The approach is simple to implement and can be deployed quickly in practice thanks to a bespoke integer programming formulation which enjoys fast solvability. We also empirically demonstrate the reliability of the approach. Compared to Thompson Sampling and greedy approaches for the problem, the upper confidence bound approach behaves in a more consistent manner. This is an important quality for decision-makers who in practice may only encounter a single learning problem and therefore must focus on the variability of an algorithm's performance, not only its expected regret.

For the CAB model, we proposed a Thompson Sampling approach, which used a progressive discretisation of the action space to handle the challenges of there being infinitely many arms to choose between and make inference on the reward distribution of. We showed that the algorithm outperformed UCB and greedy competitors. In this setting, rewards were downweighted by an additive cost per unit searching. Under this reward function UCB algorithms performed poorly, as they were overly optimistic and took many rounds to assign indices below the cost level to any part of the discretised space. This was not an issue in the setting of Chapter 4 because the filtering model which downweighted the reward of playing actions covering larger spaces was multiplicative, rather than additive. This meant UCB prioritised making observations in regions with *relatively* high indices (with respect to those of other regions) rather than in regions with whose indices have absolute value above the cost level.

Theoretical Understanding

Throughout the thesis we have been able to validate our choices of inference scheme, optimisation approach, and sequential decision making policy by deriving theoretical results guaranteeing their efficacy.

In Chapter 4 we adapted de la Peña's inequality to give a new martingale result for the frequentist mean of filtered Poisson observations. This allowed us to bound the variation of the UCB indices used in our approach to the CMAB version of the sequential event detection problem and derive an $O(\log(T))$ bound on regret. This showed that the FP-CUCB approach is order optimal, as it matched the $\Omega(\log(T))$ lower bound we derived on the regret of any uniformly good algorithm. We also demonstrated that the full-information optimisation problem encountered at each stage of the CMAB problem is part of a class of NP-hard problems.

In Chapter 5 we presented a bound on the Bayesian regret of our progressive discretisation-based Thompson Sampling algorithm for the CAB variant of the problem. We showed that the regret is $\tilde{O}(T^{2/3})$. A lower bound has not been identified for this particular problem, although the Lipschitz bandit of Kleinberg (2005) is similar and has a $\Omega(T^{2/3})$ lower bound, suggesting that the proposed Thompson Sampling approach is a good algorithm. We also demonstrated the feasibility of the algorithm by proving that the complexity of the optimisation step in each iteration is of the order the number of bins, which we fix to $O(T^{1/3})$.

In the aforementioned algorithms of Chapters 4 and 5, the inference is based on the assumption of a piecewise constant rate function with independent levels. A natural extension is to study algorithms based on inference schemes which capture more complex, spatially smooth rate functions. As a key component of any regret analysis is exploiting tight bounds on the variability of the decision-making indices, we studied the posterior contraction properties of Gaussian Cox Processes in Chapter 6.

We derived finite-time bounds on the posterior contraction of the Sigmoidal and Quadratic

Gaussian Cox Processes given independent, non-identically distributed data. Existing results had focussed only on the asymptotic contraction rate of the sigmoidal version under independent full realisations of the underlying Poisson process. Our results therefore extend beyond this work in a number of aspects that are relevant to design and analysis of sequential decision making algorithms with Cox process inference. Finite-time concentration properties enable finite-time analysis of algorithms, handling non-identically distributed data covers the case where not all parts of the observable region are observed in each round and studying multiple models adds to the general understanding of Gaussian Cox Processes and in our case has suggested that the sigmoidal version is preferable.

Finally, in Chapter 7 we studied the performance of TS applied to CABs in a setting more general than sequential event detection. We considered a CAB problem with reward function drawn from a class of functions with $M \in \mathbb{N}$ bounded, Lipschitz derivatives and sub-exponentially distributed observation noise. The main results of the chapter are generalisations of Bayesian regret analysis of parametric bandit problems in Russo and Van Roy (2014). We showed that these techniques, which bound regret in terms of the eluder dimension and complexity of the function class from which the reward distribution is drawn, can be extended to non-parametric function classes, and derive sublinear bounds on the Bayesian regret of TS.

8.2 Further Work

In addition to delivering solution approaches and answering questions around how to optimally make sequential decisions in event detection problems, we have identified numerous opportunities for further study through the research of this thesis. There are opportunities to develop research in each of the three areas in which we have made contributions, and we will divide our discussion of potential further work accordingly.

Models

The problem formulation we have used (and modified only slightly across the chapters) throughout the thesis has been intentionally straightforward. Both since this maintains a generalisability over many applications and since it has allowed us to focus on particular issues pertaining to inference and the analysis of sequential decision making problems.

Going forward, the ideas developed in this thesis could certainly be applied to more complex variants of the sequential event detection problem, and certain variations on our simple problem present themselves as natural alternatives motivated by real applications. We shall discuss a selection of these and some intuition as to how to handle them in the remainder of this subsection.

One may consider higher dimensional observable regions, e.g. detecting events with satellite or drone technology may frequently necessitate looking at 2D regions rather than simply lines or borders. In the CAB setting of Chapter 7, one may consider d dimensional action sets $[0, 1]^d$ instead of just the unit interval $[0, 1]$ we studied.

In principle these higher dimensional problems can be tackled with very similar strategies - the change in the action space and dimension of the observable region will not change the necessity to balance exploration and exploitation and Thompson Sampling or upper confidence bound principles will still be effective. The implementation and associated theoretical analyses may however become more complex.

In the settings of Chapter 4 and 5, as the cell means are modelled as independent, changing the dimension of the observable region should not alter the theoretical analysis. A non-trivial modelling question would centre around what restrictions to place on the shape of viable contiguous subregions (combinations of cells/bins) and how to represent this usefully in an integer programming formulation of the full information problem.

In the setting of Chapter 7, as we note within the chapter, we anticipate that it should be feasible to extend the bound of the elduer-dimension of the Lipschitz smooth function classes to higher dimensional inputs. The same analytical techniques which construct functions with derivative behaviour at the limits of what is permitted within the function class can be extended to functions in multiple dimensions, and should give an expression for the eluder dimension that is polynomial in the dimension of the function support. These bounds could then be propagated through the existing analyses to give yet more general regret guarantees for Thompson Sampling on smoother-than-Lipschitz bandits.

Similarly, one may wish to relax the assumption the subregions assigned to sensors must be disjoint. For instance if detection probabilities are low, but the rate function only takes values in one small area, an action which permits deploying multiple sensors to said area may be far better than those permitted by our action sets. Again, this is unlikely to alter the optimal principles for sequential decision making, and variants of our existing algorithms could likely be deployed successfully. The challenge would come in formulating efficient optimisation approaches to the more complex problem where overlapping is permitted, and dealing with domain specific issues around multiple sensors detecting the same event etc.

Extensions which will have more of an impact on the sequential decision making aspect of our algorithms are those which alter the assumptions around how events are generated and the stationarity of the reward function.

If event locations are supposed to be non-independent, we may favour a different model. Self-excitation processes such as the Hawkes process (Hawkes, 1971) capture the phenomenon where the occurrence of an event increases the probability of further events nearby. Determinantal point processes (see e.g. Lavancier et al. (2015)), on the other hand, are one class of model which capture the opposite phenomenon, where events may repel each other. These models could, for instance, be chosen instead of the NHPP to capture settings where event locations are non-independent. Thompson Sampling approaches could be readily proposed based on Bayesian Inference in these models, but theoretical analysis and/or the design of UCB policies may be more challenging due to the complex nature of the likelihoods.

If we maintain the NHPP model, but the rate function of the changes during the problem horizon, then observed data should not be handled in the same way we have. Designing an effective strategy will rely on some assumptions as to how the rate function changes. If we believe it is likely to change gradually, a sliding-window or discounting approach which either discard data after a certain number of rounds or downweight older data when making inference would be a reasonable choice. Such approaches have been deployed in simpler bandit problems (Garivier and Moulines, 2011; Kocsis and Szepesvári, 2006). If sudden changes are more likely, an approach which incorporates a changepoint detection algorithm may be more appropriate.

If we cannot make even these assumptions on the generation of events, an *adversarial* formulation of the problem may be necessary - i.e. we may wish to consider a non-stochastic bandit model of sequential event detection and design randomised algorithms whose worst case performance is sublinear for *any* sequence of rewards. If we wish to go further and assume that events are actively placed in patterns that are hard to learn or generally to minimise the number of events detected, a fully game-theoretic formulation (Fudenberg and Levine, 1998) of the problem may be appropriate.

Algorithms

The core algorithms we proposed and analyse extend the (frequentist) UCB and Thompson Sampling principles to Poisson process bandit problems. As we reviewed in Chapter 3, there are a number of other algorithmic approaches to simpler bandit problems - i.e. those with lighter tailed noise or simpler feedback - such as the KL-UCB algorithms which form an index by numerical maximisation of a function of KL-divergence and Bayes-UCB algorithms which use quantiles of the posterior distribution as decision making indices. For continuum-armed bandits there is the GP-UCB algorithm which maximises a upper confidence bound on the reward function to select actions, and a variety of other methods from Bayesian optimisation which can be extended to bandit problems.

There is scope to extend and implement these methods for Poisson process bandit problems, and to tackle the (potentially more complex) analysis required to derive bounds on their regret. In simpler bandit problems, methods such as KL-UCB and Bayes-UCB can be shown to outperform UCB1 and Thompson Sampling, and as such there is a possibility that such methods could perform better than those introduced in this thesis. While extension of these principles to the finite-action space problem may be quite straightforward - from an algorithm design point of view if not with regards to the analysis - a particular challenge would be incorporating the spatial information present in the continuum-armed version of the problem in to such a problem. We know that under Gaussian Cox Process models, the posterior is intractable, and as such designing an algorithm which forms upper confidence bounds or can draw quantiles of the posterior may be challenging.

Theoretical Understanding

All of the proposed modelling and algorithmic developments above bring with them the opportunity to derive the kinds of concentration and regret analyses we have presented in this thesis for their specific modification. Such contributions will carry with them varying levels of intellectual challenge and novelty. We will not discuss these here. Rather we will focus on certain open questions and opportunities for further study that remain around the methods and problems considered in this work.

- **Regret of Thompson Sampling with Gaussian Cox Process inference:** The bounds in Chapter 6 are derived with analysis of a Thompson Sampling approach in mind. An open problem remains to determine whether these bounds can actually be used to bound the Bayesian regret of such an algorithm. Knowing that posterior mass is concentrated near the true function means that the distance between the sampled rate function $\tilde{\lambda}$ which TS makes decisions based on and the true rate function λ_0 is bounded with high probability. The structure of the problem means that the instantaneous regret will then also be bounded with high probability. The challenge in using this high probability bound is one of non-identifiability - we do not have any guarantee of the contraction at specific points, only the overall contraction over the space. We intend to continue investigation in to what results can be derived using the GCP bounds in the future.

- **Lower bounds for problems with point process feedback:** Under the combinatorial model of Chapter 4, we were able to derive a logarithmic order lower bound on regret. In this model of the problem, while we may have observed event locations, they provided no additional information beyond that given by the count of observed events in each cell, since we assumed the rate function to be piecewise constant with independent levels. In the more complex continuum armed setting where the rate function is not piecewise constant however, the feedback of event locations affects how we infer the rate function. Intuitively, it follows that it may therefore affect the regret lower bounds for the continuum armed bandit version of the problem. For this reason in Chapter 5, we do not claim to have knowledge of the true lower bound for our problem and point only to those for related problems as being suggestive as to the ballpark of the true lower bound. Deriving the correct lower bound for the problem with observed event locations as feedback remains an open problem and interesting research opportunity. Solving this open problem may also carry insights which can be carried forward to producing better algorithms for our problem and for lower bounding regret in other learning problems with complex feedback.

- **Contraction of Gaussian (Cox) Processes with non-independent samples:** The results of Chapter 6 may be useful for studying the contraction of the GCP posterior under the reward model of Chapter 5 - where all events in a selection region are observed, subject to cost. To consider the reward model of Chapter 4 in the CAB setting would require martingale versions of the posterior contraction results, since the filtering introduces dependencies between the rewards and actions, and thus across the reward sequence. Ghosal and Van Der Vaart (2001) study a version of non-independent data where successive observations are realisations of a Markov Chain. Work to derive martingale versions of the GCP contraction results would commence with looking at the extension of this fundamental work. As done with the non-identically distributed theory in Chapter 6, the ideas would then need to be carried through the Gaussian process and Poisson process theory, deriving analogues of the results of van der Vaart and van Zanten (2009) and Belitser et al. (2015) for non-independent data. The impact of such results (particularly the Gaussian process contraction) could extend to other problems with non-independent sequential observations such as Bayesian optimisation and active learning.

- **Performance of Thompson Sampling based on Variational Inference:** As we described in Chapter 2, the most efficient implementations of GCP inference (Donner and Opper, 2018; John and Hensman, 2018) are based on variational inference. As a result, efficient implementations of Thompson Sampling will in practice rely on approximate inference and therefore sampled indices will necessarily be drawn from approximations of the true posterior. Indeed, even Markov Chain Monte Carlo inference is only exact in the limit, so any implementations of Thompson Sampling where closed-form updates of the posterior are not available will inevitably include such an approximation.

  Existing analyses of Thompson Sampling approaches typically assume exact inference, ignoring such approximations, which is of course possible when closed form posteriors are available. As the use of variational methods increase, and the study of Thompson Sampling moves to ever more complex problems, an understanding of the effect of such approximations on sequential decision making will become important, not only in the Poisson process bandit. A few papers have begun to explore the use of variational inference in Thompson Sampling. However these either only present empirical evaulation as in Urteaga and Wiggins (2018) or consider very specific models where bespoke bounds on the quality of the variational approximation are available (Qi et al., 2018).

- **Variance of Thompson Sampling:** In the empirical analysis of Chapter 4, we find evidence that the variance of the reward accumulated by the Thompson Sampling approach changes depending on the prior used. Indeed there is an order of magnitude difference between the empirical variances of the different parametrisations of Thompson Sampling in some experiments. This raises the question of how the variance of the reward obtained by Thompson Sampling is linked to the prior parameters and to the problem instance - can we derive theoretical guarantees on variance akin to those on expected (Bayesian) regret? We are not the only ones to raise this question - Lattimore and Szepesvári (2018) have the following to say on the subject: "*We should be wary [...] that injecting noise into our algorithms might come at a cost in terms of variance. What is gained or lost by the randomization in Thompson sampling is still not clear, but we leave [...] a suggestion to the reader to think about some of the costs and benefits*".

  This is not, merely, an academically interesting question. Consider the case of a decision-maker wishing to make an informed choice of which bandit algorithm to apply to some sequential decision-making problem which is of importance to them. While decisions are made repeatedly within a learning process, the decision of which algorithm to employ is only made once for a given problem. For this reason, the decision-maker should think carefully about the variance of their options (potential algorithms) as well as the expected return (or regret), just as they should when evaluating any other decision or investment.

  There are existing attempts to consider *risk* within stochastic bandits. Sani et al. (2012) consider an alternative framework to expected regret minimisation, where they aim to reduce risk by minimising a mean-variance version of regret. Galichet et al. (2013) consider a variant of the bandit problem centred around identifying arms with maximal lower quantiles -

reducing the risk of large losses. These works consider the variance and/or the distribution of the regret to an extent, but both propose and analyse new algorithms which perform well for these performance metrics. What is crucially missing from the literature (to the best of our knowledge) is an understanding of the variance of the reward accumulated by existing, popular algorithms.

We note that the variance of common algorithms has been addressed to some extent in adversarial bandits - for instance in Section 11.5 of Lattimore and Szepesvári (2018), Bubeck et al. (2018), Bubeck and Sellke (2019), and Pogodin and Lattimore (2019). However as with other aspects of bandit theory, there is little transfer across between the worst-case analysis in the unstructured environment of adversarial bandits and stochastic bandits.

Appendix A

Verifying conditions on sieves

Throughout the analysis of Chapter 6 the sequences used in defining the sieves are subject to numerous conditions and assumptions, in order that we may demonstrate the conditions of Theorem 6.4.3 are met for the GCP models. By choosing $L_{2:10}$ as specified in the main body, these conditions are met by definition for values of n as specified. There are numerous such conditions to verify, and doing so can be non-trivial. In this section we show the link between the conditions and constraints on $L_{2:10}, n$ and demonstrate fully that the necessary results hold.

QGCP model

Recall, the definitions of the following sequences:

$$
\begin{aligned}
\delta_n &= 2||g_0||_\infty n^{-\alpha/(4\alpha+d)} \log^{\rho}(n) + n^{-2\alpha/(4\alpha+d)} \log^{2\rho}(n),\\
\bar{\delta}_n &= 2||g_0||_\infty n^{-\alpha/(4\alpha+d)} \log^{\rho+d+1}(n) + n^{-2\alpha/(4\alpha+d)} \log^{2\rho+2d+2}(n)\\
\zeta_n &= L_2 n^{(2\alpha+d)/(4\alpha d+d^2)} \log^{2\rho/d}(n) + L_3 n^{(\alpha d+d^2)/(4\alpha d+d^2)} \log^{3\rho/d}(n) + L_4 n^{d/(4\alpha+d)} \log^{4\rho/d}(n)\\
\beta_n &= L_5 n^{(2\alpha+d)/(8\alpha+2d)} \log^{2\rho+(d+1)/2}(n) + L_6 n^{(\alpha+d)/(8\alpha+2d)} \log^{3\rho+(d+1)/2}(n)\\
&\qquad + L_7 n^{d/(8\alpha+2d)} \log^{4\rho+(d+1)/2}(n)
\end{aligned}
$$

with $L_2, ..., L_7$ satisfying

$$\begin{aligned}
L_2 + L_3 + L_4 &> \max(A, e) \\
L_2 L_5^3 &> \left(\frac{8\max(1, ||g_0||_\infty)}{(3/L_1)^{3/2} d^{1/4}\sqrt{2\tau}}\right)^2 \\
L_5 + L_6 + L_7 &> \frac{4L_1 \max(1, ||g_0||_\infty)}{3\sqrt{||\mu||}} \\
L_2 &\geq (8c_5||g_0||_\infty^2)/D_1 \\
L_3 &\geq (8c_5||g_0||_\infty)/D_1 \\
L_4 &\geq 2c_5/D_1 \\
L_5 &\geq \max\left(\sqrt{\frac{16K_5 L_2^d \mathcal{K}_1^{1+d}}{\log^{2\rho}(3)}}, \sqrt{32||g_0||_\infty^2 c_5}\right) \\
L_6 &\geq \max\left(\sqrt{\frac{16K_5 L_3^d \mathcal{K}_1^{1+d}}{\log^{3\rho}(3)}}, \sqrt{32||g_0||_\infty c_5}\right) \\
L_7 &\geq \max\left(\sqrt{\frac{16K_5 L_4^d \mathcal{K}_1^{1+d}}{\log^{4\rho}(3)}}, \sqrt{8c_5}\right)
\end{aligned}$$

for $n \geq \max(3, n_3, n_4, n_5)$. Here n_3 is the smallest integer n such that

$$\begin{aligned}
4||g_0||_\infty^2 \log^{2d+2-2\rho}(n) \geq \frac{m\sum_{i=2}^4 L_i^d}{2^{1+d}}\Bigg(&\log\left(\frac{27\tau\sqrt{d}(\sum_{i=5}^7 L_i)^3 \sum_{i=2}^4 L_i}{4||g_0||_\infty^3}\right) \\
&+ \left(4 + \frac{12 + d + d^2}{8\alpha d + 2d^2}\right)\log(n)\Bigg)^{1+d}
\end{aligned}$$

n_4 is the smallest integer n such that

$$\begin{aligned}
2\log\left(\frac{6\sqrt{||\mu||}(L_5 + L_6 + L_7)}{2L_1||g_0||_\infty + L_1}\right) \leq\ & 4||g_0||_\infty^2 n^{(4\alpha+2d)/(8\alpha+2d)} \log^{2\rho+2d+2}(n) \\
&- \log\left(n^{(6\alpha+d)/(4\alpha+d)} \log^{6\rho}(n)\right)
\end{aligned}$$

and n_5 is the smallest integer n such that

$$n^{(2\alpha+d)/(4\alpha+d)}\log^{2\rho}(n) \geq \frac{q_1}{4c_5||g_0||_\infty^2}\log\Big((L_2+L_3+L_4)n^{(2\alpha+d)/(4\alpha d+d^2)}\log^{4\rho/d}(n)\Big).$$

In the remainder of this subsection, we show that the following conditions, which are all restatements of required results in our main analysis, hold for the sequences described above.

$$\zeta_n > \max(A,1) \tag{A.0.1}$$

$$(3/L_1)^{3/2}d^{1/4}\beta_n^{3/2}\sqrt{2\tau\zeta_n} > 2\bar{\delta}_n^{3/2} \tag{A.0.2}$$

$$(3/L_1)\beta_n\sqrt{||\mu||} > \bar{\delta}_n \tag{A.0.3}$$

$$m\zeta_n^d\Bigg(\log\left(\frac{(3/L_1)^{3/2}d^{1/4}\beta_n^{3/2}\sqrt{2\tau\zeta_n}}{\bar{\delta}_n^{3/2}}\right)\Bigg)^{1+d} \leq n\bar{\delta}_n^2 \tag{A.0.4}$$

$$2\log\left(\frac{6\beta_n\sqrt{||\mu||}}{L_1\bar{\delta}_n}\right) \leq n\bar{\delta}_n^2 \tag{A.0.5}$$

$$\beta_n^2 > 16K_5\zeta_n^d\Bigg(\log\left(\frac{3\zeta_n}{\bar{\delta}_n}\right)\Bigg)^{1+d} \tag{A.0.6}$$

$$D_1\zeta_n^d\Bigg(\log^{q_2}(\zeta_n)\Bigg) \geq 2c_5n\delta_n^2 \tag{A.0.7}$$

$$\beta_n^2 \geq 8c_5n\delta_n^2 \tag{A.0.8}$$

$$\zeta_n^{q_1-d+1} \leq \exp(c_5n\delta_n^2), \tag{A.0.9}$$

Verifying (A.0.1)

For $n = 3$, $\log(n) > 1$ thus $\zeta_n > L_2 + L_3 + L_4$ for all $\alpha \in [0,1]$, and $d \geq 1$. It follows that (A.0.1) is satisfied for $n = 3$ given $L_2 + L_3 + L_4 > \max(A, e)$. To show it holds for all $n > 3$ we simply note that ζ_n is an increasing function.

Verifying (A.0.2)

First consider,

$$\begin{aligned}
&\bar{\delta}_n = 2||g_0||_\infty n^{-\alpha/(4\alpha+d)} \log^{\rho+d+1}(n) + n^{-2\alpha/(4\alpha+d)} \log^{2\rho+2d+2}(n) \\
&\quad \leq 4\max(1, ||g_0||_\infty) n^{-\alpha/(4\alpha+d)} \log^{2\rho+2d+2}(n) \\
\Rightarrow \quad & 2\bar{\delta}_n^{3/2} \leq (4\max(1, ||g_0||_\infty))^{3/2} n^{-3\alpha/(8\alpha+2d)} \log^{3\rho+3d+3}(n))^{3/2} \\
&\quad = (4\max(1, ||g_0||_\infty))^{3/2} n^{-3\alpha/(8\alpha+2d)} \log^{3\rho+3(d+1)/4}(n) \log^{9(d+1)/4}(n)
\end{aligned}$$

Let $z_1 = (3/L_1)^{3/2} d^{1/4}\sqrt{2\tau}$,

$$\begin{aligned}
z_1\sqrt{\beta_n^3\zeta_n} &\geq z_1 \log^{4\rho+3(d+1)/4}(n)\big(L_5 n^{\frac{2\alpha+d}{8\alpha+2d}} + L_6 n^{\frac{\alpha+d}{8\alpha+2d}} + L_7 n^{\frac{d}{8\alpha+2d}}\big)^{3/2} \\
&\quad \times \big(L_2 n^{\frac{2\alpha+d}{4\alpha d+d^2}} + L_3 n^{\frac{\alpha+d}{4\alpha d+d^2}} + L_4 n^{\frac{d}{4\alpha d+d^2}}\big)^{1/2} \\
&\geq z_1 \log^{3\rho+3(d+1)/4}(n)\sqrt{L_5^3 L_2 n^{\frac{6\alpha+3d}{8\alpha+2d}}}.
\end{aligned}$$

Thus values of L_2, L_5 such that

$$z_1\sqrt{L_2 L_5^3 n^{\frac{6\alpha+3d}{16\alpha+4d}+\frac{3\alpha}{8\alpha+2d}}} > 8\max(1, ||g_0||_\infty)\log^{9(d+1)/4}(n)$$

are sufficient to verify (A.0.2). For $n \geq 3$, $d > 1$ and $\alpha \in [0,1]$ $n^{\frac{12\alpha+3d}{16\alpha+4d}} > \log^{9(d+1)/4}(n)$ so

$$L_2 L_5^3 > \left(\frac{8\max(1, ||g_0||_\infty)}{(3/L_1)^{3/2} d^{1/4}\sqrt{2\tau}}\right)^2$$

is a sufficient condition to verify (A.0.2).

Verifying (A.0.3)

First consider,

$$L_1\bar{\delta}_n \leq 4L_1 \max(1, ||g_0||_\infty) n^{-\alpha/(4\alpha+d)} \log^{2\rho+2d+2}(n),$$

and

$$3\sqrt{||\mu||}\beta_n \geq 3\sqrt{||\mu||}(L_5 + L_6 + L_7) n^{(2\alpha+d)/(8\alpha+2d)} \log^{2\rho+(d+1)/2}(n).$$

Plainly $n^{(2\alpha+d)/(8\alpha+2d)} \log^{2\rho+(d+1)/2}(n) > n^{-\alpha/(4\alpha+d)} \log^{2\rho+2d+2}(n)$ for $n \geq 3$, so

$$L_5 + L_6 + L_7 > \frac{4L_1 \max(1, ||g_0||_\infty)}{3\sqrt{||\mu||}}$$

is a sufficient condition to verify (A.0.3).

Verifying (A.0.4)

Consider,

$$\begin{aligned}
& m\zeta_n^d \Bigg(\log\left(\frac{3^{3/2} d^{1/4} \beta_n^{3/2} \sqrt{2\tau\zeta_n}}{(L_1\bar{\delta}_n)^{3/2}} \right) \Bigg)^{1+d} \\
& \leq m(L_2^d + L_3^d + L_4^d) n^{\frac{2\alpha+d}{4\alpha+d}} \log^{4\rho}(n) \left[\frac{3}{2} \log\left(\frac{3\beta_n}{L_1\bar{\delta}_n} \right) + \frac{1}{2}\log(2\tau d^{1/2}\zeta_n) \right]^{1+d}, \\
& \leq m(L_2^d + L_3^d + L_4^d) n^{\frac{2\alpha+d}{4\alpha+d}} \log^{4\rho}(n) \Bigg[\frac{3}{2} \log\left(\frac{3(L_5 + L_6 + L_7)}{2||g_0||_\infty} n^{\frac{6\alpha+d}{8\alpha+2d}} \log^{\rho}(n) \right) \\
& \qquad\qquad + \frac{1}{2}\log(2\tau\sqrt{d}(L_2 + L_3 + L_4) n^{\frac{2\alpha+d}{4\alpha d+d^2}} \log^{4\rho/d}(n)) \Bigg]^{1+d} \\
& \leq \frac{m(L_2^d + L_3^d + L_4^d)}{2^{1+d}} \Bigg[3\log\left(\frac{3(L_5 + L_6 + L_7)}{2||g_0||_\infty} \right) + \frac{18\alpha + 3d}{8\alpha + 2d}\log(n) + 3\rho\log(\log(n)) \\
& \quad + \log(2\tau\sqrt{d}(L_2 + L_3 + L_4)) + \frac{2\alpha + d}{4\alpha d + d^2}\log(n) + \frac{4\rho}{d}\log(\log(n)) \Bigg]^{1+d} n^{\frac{2\alpha+d}{4\alpha+d}} \log^{4\rho}(n)
\end{aligned}$$

$$\begin{aligned}
&= \frac{m(L_2^d + L_3^d + L_4^d)}{2^{1+d}} \Bigg[\log\Big(\frac{27\tau\sqrt{d}(L_5 + L_6 + L_7)^3}{4||g_0||_\infty^3}(L_2 + L_3 + L_4)\Big) + \left(\frac{18\alpha d + 4\alpha + 2d + 3d^2}{8\alpha d + 2d^2}\right)\log(n) \\
&\qquad\qquad + \left(3\rho + \frac{4\rho}{d}\right)\log(\log(n))\Bigg]^{1+d} n^{\frac{2\alpha+d}{4\alpha+d}}\log^{4\rho}(n) \\
&\le \frac{m\sum_{i=2}^4 L_i^d}{2^{1+d}} \Bigg[\log\Big(\frac{27\tau\sqrt{d}(\sum_{i=5}^7 L_i)^3}{4||g_0||_\infty^3}\sum_{i=2}^4 L_i\Big) + \Big(\frac{18\alpha d + 4\alpha + 2d + 3d^2}{8\alpha d + 2d^2} + \frac{3\rho d + 4\rho}{d}\Big)\log(n)\Bigg]^{1+d} \\
&\quad \times n^{\frac{2\alpha+d}{4\alpha+d}}\log^{4\rho}(n) \\
&\le \frac{m\sum_{i=2}^4 L_i^d}{2^{1+d}} \Bigg[\log\Big(\frac{27\tau\sqrt{d}(\sum_{i=5}^7 L_i)^3}{4||g_0||_\infty^3}\sum_{i=2}^4 L_i\Big) + \frac{32\alpha d + 12\alpha + 2d + 3d^2 + 6\alpha d^2}{8\alpha d + 2d^2}\log(n)\Bigg]^{1+d} n^{\frac{2\alpha+d}{4\alpha+d}}\log^{4\rho}(n) \\
&\le \frac{m\sum_{i=2}^4 L_i^d}{2^{1+d}} \Bigg[\log\Big(\frac{27\tau\sqrt{d}(\sum_{i=5}^7 L_i)^3}{4||g_0||_\infty^3}\sum_{i=2}^4 L_i\Big) + \Big(4 + \frac{12 + d + d^2}{8\alpha d + 2d^2}\Big)\log(n)\Bigg]^{1+d} n^{\frac{2\alpha+d}{4\alpha+d}}\log^{4\rho}(n)
\end{aligned}$$

and

$$n\bar{\delta}_n^2 \ge 4||g_0||_\infty^2 n^{\frac{2\alpha+d}{4\alpha+d}}\log^{2\rho+2(d+1)}(n)$$

Condition (A.0.4) is then satisfied for all n such that

$$\begin{aligned}
4||g_0||_\infty^2 \log^{2d+2-2\rho}(n) \ge \frac{m\sum_{i=2}^4 L_i^d}{2^{1+d}} \Bigg(&\log\Big(\frac{27\tau\sqrt{d}(\sum_{i=5}^7 L_i)^3\sum_{i=2}^4 L_i}{4||g_0||_\infty^3}\Big) \\
&+ \Big(4 + \frac{12 + d + d^2}{8\alpha d + 2d^2}\Big)\log(n)\Bigg)^{1+d}
\end{aligned}$$

Verifying (A.0.5)

Consider,

$$\begin{aligned}
2\log\left(\frac{6\beta_n\sqrt{||\mu||}}{L_1\bar{\delta}_n}\right) &\le 2\log\left(\frac{6\sqrt{||\mu||}(L_5 + L_6 + L_7)n^{(2\alpha+d)/(8\alpha+2d)}\log^{4\rho+(d+1)/2}(n)}{(2||g_0||_\infty + 1)L_1 n^{-2\alpha/(4\alpha+d)}\log^{\rho+(d+1)/2}(n)}\right) \\
&= 2\log\left(\frac{6\sqrt{||\mu||}(L_5 + L_6 + L_7)}{2L_1||g_0||_\infty + L_1}n^{(6\alpha+d)/(8\alpha+2d)}\log^{3\rho}(n)\right)
\end{aligned}$$

$$= 2\log\left(\frac{6\sqrt{||\mu||}(L_5+L_6+L_7)}{2L_1||g_0||_\infty + L_1}\right) + \log\left(n^{(6\alpha+d)/(4\alpha+d)}\log^{6\rho}(n)\right)$$

and

$$\begin{aligned} n\bar{\delta}_n^2 &= 4||g_0||_\infty^2 n^{(2\alpha+d)/(4\alpha+d)}\log^{2\rho+2d+2}(n) + 4||g_0||_\infty n^{(\alpha+d)/(4\alpha+d)}\log^{3\rho+3d+3}(n) + n^{d/(4\alpha+d)}\log^{4\rho+4d+4}(n) \\ &\geq 4||g_0||_\infty^2 n^{(4\alpha+2d)/(8\alpha+2d)}\log^{2\rho+2d+2}(n) \end{aligned}$$

Therefore, condition (A.0.5) holds for all n such that

$$2\log\left(\frac{6\sqrt{||\mu||}(L_5+L_6+L_7)}{2L_1||g_0||_\infty + L_1}\right) \leq 4||g_0||_\infty^2 n^{(4\alpha+2d)/(8\alpha+2d)}\log^{2\rho+2d+2}(n) - \log\left(n^{(6\alpha+d)/(4\alpha+d)}\log^{6\rho}(n)\right)$$

Verifying (A.0.6)

Consider

$$\beta_n^2 = L_5^2 n^{\frac{2\alpha+d}{4\alpha+d}}\log^{4\rho+d+1}(n) + L_6^2 n^{\frac{\alpha+d}{4\alpha+d}}\log^{6\rho+d+1}(n) + L_7^2 n^{\frac{d}{4\alpha+d}}\log^{8\rho+d+1}(n)$$

and

$$\begin{aligned} &\zeta_n^d\Bigg(\log\left(\frac{3\zeta_n}{\bar{\delta}_n}\right)\Bigg)^{1+d} \\ &= (L_2^d n^{\frac{2\alpha+d}{4\alpha+d}}\log^{2\rho}(n) + L_3^d n^{\frac{\alpha+d}{4\alpha+d}}\log^{3\rho}(n) + L_4^d n^{\frac{d}{4\alpha+d}}\log^{4\rho}(n)) \\ &\quad\times\left(\log\left(\frac{3L_2 n^{(2\alpha+d)/(4\alpha d+d^2)}\log^{2\rho/d}(n) + 3L_3 n^{(\alpha+d)/(4\alpha d+d^2)}\log^{3\rho/d}(n) + 3L_4 n^{d/(4\alpha d+d^2)}\log^{4\rho/d}(n)}{2||g_0||_\infty n^{-\alpha/(4\alpha+d)}\log^{\rho+d+1}(n) + n^{-2\alpha/(4\alpha+d)}\log^{2\rho+2d+2}(n)}\right)\right)^{1+d} \\ &\leq (L_2^d n^{\frac{2\alpha+d}{4\alpha+d}}\log^{2\rho}(n) + L_3^d n^{\frac{\alpha+d}{4\alpha+d}}\log^{3\rho}(n) + L_4^d n^{\frac{d}{4\alpha+d}}\log^{4\rho}(n)) \\ &\quad\times\left(\log\left(\frac{3(L_2+L_3+L_4)n^{(2\alpha+d)/(4\alpha d+d^2)}\log^{4\rho/d}(n)}{\min(1, 2||g_0||_\infty)n^{-2\alpha/(4\alpha+d)}\log^{\rho+d+1)}(n)}\right)\right)^{1+d} \\ &\leq (L_2^d n^{\frac{2\alpha+d}{4\alpha+d}}\log^{2\rho}(n) + L_3^d n^{\frac{\alpha+d}{4\alpha+d}}\log^{3\rho}(n) + L_4^d n^{\frac{d}{4\alpha+d}}\log^{4\rho}(n)) \end{aligned}$$

$$\times \left[\log\left(\frac{3(L_2+L_3+L_4)}{\min(1, 2||g_0||_\infty)}\right) + \log\left(n^{\frac{2\alpha+d}{4\alpha d+d^2}+\frac{2\alpha}{4\alpha+d}} \log^{4\rho-\rho/d-d-1}(n)\right)\right]^{1+d}.$$

Define

$$\begin{aligned}\mathcal{K}(n) &= \left[\log\left(\frac{3(L_2+L_3+L_4)}{\min(1, 2||g_0||_\infty)}\right) + \frac{2\alpha d+2\alpha+d}{4\alpha d+d^2}\log(n) + (4\rho-\rho/d-d-1)\log(\log(n))\right]^{1+d} \\ &\leq \log^{1+d}(n)\Bigg(\log\left(\frac{3(L_2+L_3+L_4)}{\min(1, 2||g_0||_\infty)}\right) + \frac{2\alpha d+2\alpha+d}{4\alpha d+d^2} + (4\rho-\rho/d-d-1)\Bigg)^{1+d},\end{aligned}$$

for $n \geq 3$. Let $\mathcal{K}_1 = \left(\log\left(\frac{3(L_2+L_3+L_4)}{\min(1,2||g_0||_\infty)}\right) + \frac{2\alpha d+2\alpha+d}{4\alpha d+d^2} + (4\rho-\rho/d-d-1)\right)$. Grouping terms of the same order we require the following for all sufficiently large n

$$\begin{aligned} L_5^2\log^{2\rho}(n) &\geq 16K_5L_2^d\mathcal{K}_1^{1+d}, \\ L_6^2\log^{3\rho}(n) &\geq 16K_5L_3^d\mathcal{K}_1^{1+d}, \\ L_7^2\log^{4\rho}(n) &\geq 16K_5L_4^3\mathcal{K}_1^{1+d}, \end{aligned}$$

to satisfy (A.0.6). Thus, the following are sufficient conditions to satisfy (A.0.6) for all $n \geq 3$

$$L_5 \geq \sqrt{\frac{16K_5L_2^d\mathcal{K}_1^{1+d}}{\log^{2\rho}(3)}}, \quad L_6 \geq \sqrt{\frac{16K_5L_3^d\mathcal{K}_1^{1+d}}{\log^{3\rho}(3)}}, \quad L_7 \geq \sqrt{\frac{16K_5L_4^d\mathcal{K}_1^{1+d}}{\log^{4\rho}(3)}}.$$

Verifying (A.0.7)

Consider,

$$2c_5n\delta_n^2 = 2c_5\Bigg(4||g_0||_\infty^2 n^{\frac{2\alpha+d}{4\alpha+d}}\log^{2\rho}(n) + 4||g_0||_\infty n^{\frac{\alpha+d}{4\alpha+d}}\log^{3\rho}(n) + n^{\frac{d}{4\alpha+d}}\log^{4\rho}(n)\Bigg),$$

and

$$D_1\zeta_n^d \log^{q_2}(\zeta_n) \geq D_1\bigg(L_2 n^{\frac{2\alpha+d}{4\alpha+d}} \log^{2\rho}(n) + L_3 n^{\frac{\alpha+d}{4\alpha+d}} \log^{3\rho}(n) + L_4 n^{\frac{d}{4\alpha+d}} \log^{4\rho}(n)\bigg)$$

for $\zeta_n > e$ - i.e. such that $\log(\zeta_n) \geq 1$. Then condition (A.0.1) and $L_2 \geq (8c_5||g_0||_\infty^2)/D_1$, $L_3 \geq (8c_5||g_0||_\infty)/D_1$ and $L_4 \geq 2c_5/D_1$ are sufficient conditions to verify (A.0.7).

Verifying (A.0.8)

Consider,

$$8c_5 n\delta_n^2 = 8c_5\bigg(4||g_0||_\infty^2 n^{\frac{2\alpha+d}{4\alpha+d}} \log^{2\rho}(n) + 4||g_0||_\infty n^{\frac{\alpha+d}{4\alpha+d}} \log^{3\rho}(n) + n^{\frac{d}{4\alpha+d}} \log^{4\rho}(n)\bigg),$$

and

$$\beta_n^2 \geq L_5^2 n^{\frac{2\alpha+d}{4\alpha+d}} \log^{2\rho}(n) + L_6^2 n^{\frac{\alpha+d}{4\alpha+d}} \log^{3\rho}(n) + L_7^2 n^{\frac{d}{4\alpha+d}} \log^{4\rho}(n).$$

Then (A.0.7) is satisfied with $L_5^2 > 32||g_0||_\infty^2 c_5$, $L_6^2 > 32||g_0||_\infty c_5$, and $L_7^2 > 8c_5$.

Verifying (A.0.9)

Consider

$$\begin{aligned}\exp(c_5 n\delta_n^2) &= \exp\left(c_5\bigg(4||g_0||_\infty^2 n^{\frac{2\alpha+d}{4\alpha+d}} \log^{2\rho}(n) + 4||g_0||_\infty n^{\frac{\alpha+d}{4\alpha+d}} \log^{3\rho}(n) + n^{\frac{d}{4\alpha+d}} \log^{4\rho}(n)\bigg)\right), \\ &\geq \exp\left(4c_5||g_0||_\infty^2 n^{(2\alpha+d)/(4\alpha+d)} \log^{2\rho}(n)\right)\end{aligned}$$

and

$$\begin{aligned}\zeta_n^{q_1-d+1} &\leq \Big(L_2 n^{(2\alpha+d)/(4\alpha d+d^2)}\log^{2\rho/d}(n) + L_3 n^{(\alpha+d)/(4\alpha d+d^2)}\log^{3\rho/d}(n) + L_4 n^{d/(4\alpha d+d^2)}\log^{4\rho/d}(n)\Big)^{q_1},\\ &\leq \Big((L_2+L_3+L_4)n^{(2\alpha+d)/(4\alpha d+d^2)}\log^{4\rho/d}(n)\Big)^{q_1}\end{aligned}$$

The condition is then satisfied for all n such that

$$n^{(2\alpha+d)/(4\alpha+d)}\log^{2\rho}(n) \geq \frac{q_1}{4c_5||g_0||_\infty^2}\log\Big((L_2+L_3+L_4)n^{(2\alpha+d)/(4\alpha d+d^2)}\log^{4\rho/d}(n)\Big).$$

SGCP model

Recall the definitions of the following sequences:

$$\begin{aligned}\delta_n &= n^{-\alpha/(2\alpha+d}\log^{\rho}(n)\\ \bar{\delta}_n &= n^{-\alpha/(2\alpha+d}\log^{\rho+d+1}(n)\\ \zeta_n &= L_8 n^{\frac{1}{2\alpha+d}}(\log(n))^{2\rho/d},\\ \beta_n &= L_9 n^{\frac{d}{2(2\alpha+d)}}(\log(n))^{d+1+2\rho},\\ \lambda_n &= L_{10} n^{\frac{d}{\kappa(2\alpha+d)}}(\log(n))^{4\rho/\kappa}\end{aligned}$$

with L_8, L_9, L_{10} satisfying

$$\begin{aligned}L_8 &> \max\left(A, 1, \left(\frac{2c_5}{D_1}\right)^{1/d}\right)\\ L_9 &\geq \sqrt{8c_5}\\ L_{10} &> \left(\frac{c_5}{c_0}\right)^{1/\rho}\end{aligned}$$

$$L_8 L_9^3 L_{10}^{3/2} > \frac{2}{()6cL_1)^{3/2}\tau\sqrt{d}}$$
$$L_9 L_{10}^{1/2} > \frac{1}{6cL_1\sqrt{||\mu||}}$$

for $n \geq \max(3, n_6, n_7, n_8)$. Here n_6 is the smallest integer such that

$$n^{\frac{d}{2\alpha+d}} > \max\left(2\log(12cL_1L_9L_{10}^{1/2}) + 1, \log(2L_1L_{10}^{1/2}) + 1, \frac{1}{c_5}\big(\log(L_8^{q_1-d+1}) + 1\big)\right),$$

n_7 is the smallest integer such that

$$\log^{2d+2}(n) > mL_8^d\Bigg(\log((6cL_1)^{3/2}\sqrt{2\tau L_8 L_9^3}L_{10}^{3/4}d^{1/4}) + \frac{\kappa(6d+6\alpha+2)+3d}{4\kappa(2\alpha+d)}\log(n) + \log\Big(\log^{3\rho/2+3\rho/\kappa+\rho/d-d-1}(n)\Big)\Bigg)^{1+d},$$

and n_8 is the smallest integer such that

$$\log^{2\rho}(n) > \frac{16K_5D_1L_8^d}{L_9^2}\left(\frac{\log(\sqrt{L_{10}}L_8) + \log\big(n^{\frac{2\alpha\kappa+2\kappa+d}{2\kappa(2\alpha+d)}}\log^{\rho(2/\kappa+2/d-1)-d-1}(n)\big)}{\log(n)}\right)^{1+d}.$$

In the remainder of this subsection, we show that the following conditions, which are all restatements of required results in our main analysis, hold for the sequences described above.

$$(6cL_1)^{3/2}d^{1/4}\beta_n^{3/2}\lambda_n^{3/4}\sqrt{2\tau\zeta_n} > 2\bar{\delta}_n^{3/2} \tag{A.0.10}$$

$$6cL_1\beta_n\sqrt{\lambda_n||\mu||} > \bar{\delta}_n \tag{A.0.11}$$

$$\zeta_n > \max(A, 1) \tag{A.0.12}$$

$$m\zeta_n^d\Bigg(\log\Big(\frac{(6cL_1)^{3/2}\lambda_n^{3/4}\beta_n^{3/2}d^{1/4}\sqrt{2\tau\zeta_n}}{\bar{\delta}_n^{3/2}}\Big)\Bigg)^{1+d} < K_3n\bar{\delta}_n^2 \tag{A.0.13}$$

$$2\log\Big(\frac{12cL_1\beta_n\sqrt{\lambda_n}||\mu||}{\bar{\delta}_n}\Big) < K_4 n\bar{\delta}_n^2 \tag{A.0.14}$$

$$\log\Big(\frac{2L_1\lambda_n^{1/2}}{\bar{\delta}_n}\Big) < K_5 n\bar{\delta}_n^2 \tag{A.0.15}$$

$$\beta_n^2 > 16K_5\zeta_n^d\Bigg(\log\Big(\frac{\lambda_n^{1/2}\zeta_n}{\bar{\delta}_n}\Big)\Bigg)^{1+d} \tag{A.0.16}$$

$$c_0\lambda_n^\rho > c_5 n\delta_n^2 \tag{A.0.17}$$

$$D_1\zeta_n^d \geq 2c_5 n\delta_n^2 \tag{A.0.18}$$

$$\zeta_n^{q_1-d+1} \leq e^{c_5 n\delta_n^2} \tag{A.0.19}$$

$$\beta_n^2 \geq 8c_5 n\delta_n^2 \tag{A.0.20}$$

In turn we demonstrate that each of the conditions (A.0.10) through (A.0.20) hold.

Verifying (A.0.10)

Consider

$$\begin{aligned}
&(6cL_1)^{3/2}d^{1/4}\beta_n^{3/2}\lambda_n^{3/4}\sqrt{2\tau\zeta_n} \\
&= (6cL_1)^{3/2}d^{1/4}L_9^{3/2}n^{\frac{3d}{4(2\alpha+d)}}\log^{\frac{3d+3}{4}+3\rho}(n)L_{10}^{3/4}n^{\frac{3d}{4\kappa(2\alpha+d)}}\log^{3\rho/\kappa}(n)\sqrt{2\tau L_8 n^{\frac{1}{2\alpha+d}}(\log(n))^{2\rho/d}} \\
&= (6cL_1)^{3/2}\sqrt{2\tau L_8 L_9^3}L_{10}^{3/4}d^{1/4}n^{\frac{3d}{4(2\alpha+d)}+\frac{3d}{4\kappa(2\alpha+d)}+\frac{1}{2(2\alpha+d)}}\log^{\frac{3d+3}{4}+3\rho+3\rho/\kappa+\rho/d}(n)
\end{aligned}$$

and

$$2\bar{\delta}_n^{3/2} = 2n^{\frac{-3\alpha}{2(2\alpha+d)}}\log^{3\rho/2+3(d+1)/2}(n)$$

So (A.0.10) can be rewritten:

$$(6cL_1)^{3/2}\sqrt{2\tau L_8 L_9^3}L_{10}^{3/4}d^{1/4}n^{\frac{3d}{4(2\alpha+d)}+\frac{3d}{4\kappa(2\alpha+d)}+\frac{3\alpha+1}{2(2\alpha+d)}}\log^{3\rho/2+3\rho/\kappa+\rho/d-\frac{3d-3}{4}}(n) > 2,$$

which holds for L_8, L_9, L_{10} such that $L_8 L_9^3 L_{10}^{3/2} > \frac{2}{(6cL_1)^{3/2}\tau\sqrt{d}}$.

Verifying (A.0.11)

We may rewrite (A.0.11) as

$$6cL_1\sqrt{||\mu||}L_9 L_{10}^{1/2}n^{\frac{1}{2}}\log^{\rho+2\rho/\kappa-d-1}(n) > 1$$

which holds for all L_9, L_{10} such that $L_9 L_{10}^{1/2} > 1/(6cL_1\sqrt{||\mu||})$.

Verifying (A.0.12)

If $n \geq 3$ then ζ_n holds for all $L_8 \geq \max(A, 1)$.

Verifying (A.0.13)

Consider

$$m\zeta_n^d\Bigg(\log\Big(\frac{(6cL_1)^{3/2}\lambda_n^{3/4}\beta_n^{3/2}d^{1/4}\sqrt{2\tau\zeta_n}}{\bar{\delta}_n^{3/2}}\Big)\Bigg)^{1+d}$$
$$= mL_8^d n^{\frac{d}{2\alpha+d}}\log^{2\rho}(n)\log\left((6cL_1)^{3/2}\sqrt{2\tau L_8 L_9^3}L_{10}^{3/4}d^{1/4}n^{\frac{3d}{4(2\alpha+d)}+\frac{3d}{4\kappa(2\alpha+d)}+\frac{3\alpha+1}{2(2\alpha+d)}}\log^{3\rho/2+3\rho/\kappa+\rho/d-d-1}(n)\right)^{1+d}$$

and

$$n\bar{\delta}_n^2 = n^{\frac{d}{2\alpha+d}}\log^{2\rho+2d+2}(n)$$

Thus (A.0.13) holds for all n such that

$$\begin{aligned}\log^{2d+2}(n) > mL_8^d\bigg(& \log((6cL_1)^{3/2}\sqrt{2\tau L_8 L_9^3}L_{10}^{3/4}d^{1/4}) + \frac{\kappa(6d+6\alpha+2)+3d}{4\kappa(2\alpha+d)}\log(n) \\ & + \log\Big(\log^{3\rho/2+3\rho/\kappa+\rho/d-d-1}(n)\Big)\bigg)^{1+d}\end{aligned}$$

Verifying (A.0.14)

We may rewrite (A.0.14) as

$$2\log\left(12cL_1L_9L_{10}^{1/2}n^{\frac{1}{2}+\frac{d}{2\kappa(2\alpha+d)}}\log^{2\rho/\kappa+\rho-d-1}(n)\right) < n^{\frac{d}{2\alpha+d}}\log^{2\rho+d+1}(n)$$

which holds for all n such that

$$n^{\frac{d}{2\alpha+d}} > 2\log(12cL_1L_9L_{10}^{1/2}) + 1.$$

Verifying (A.0.15)

We may rewrite (A.0.15) as

$$\log\left(2L_1L_{10}^{1/2}n^{\frac{d}{2\kappa(2\alpha+d)}+\frac{\alpha}{2\alpha+d}}\log^{2\rho/\kappa-\rho-d-1}(n)\right) < n^{\frac{d}{2\alpha+d}}\log^{2\rho+d+1}(n)$$

which holds for all n such that

$$n^{\frac{d}{2\alpha+d}} > \log(2L_1L_{10}^{1/2}) + 1.$$

Verifying (A.0.16)

Consider

$$\beta_n^2 = L_9^2 n^{\frac{d}{2\alpha+d}} \log^{d+1+4\rho}(n)$$

and

$$\begin{aligned}
&16K_5\zeta_n^d\Bigg(\log\Big(\frac{\lambda_n^{1/2}\zeta_n}{\bar{\delta}_n}\Big)\Bigg)^{1+d} \\
&= 16K_5D_1L_8^d n^{\frac{d}{2\alpha+d}} \log^{2\rho}(n)\Bigg(\log\Big(\frac{\sqrt{L_{10}}L_8 n^{\frac{2\kappa+d}{2\kappa(2\alpha+d)}} \log^{2\rho/\kappa 2+2\rho/d}(n)}{n^{\frac{-\alpha}{2\alpha+d}} \log^{\rho+d+1}(n)}\Big)\Bigg)^{1+d} \\
&= 16K_5D_1L_8^d n^{\frac{d}{2\alpha+d}} \log^{2\rho}(n)\Bigg(\log(\sqrt{L_{10}}L_8) + \log\Big(n^{\frac{2\alpha\kappa+2\kappa+d}{2\kappa(2\alpha+d)}} \log^{\rho(2/\kappa+2/d-1)-d-1}(n)\Bigg)^{1+d}
\end{aligned}$$

So (A.0.16) can then be rewritten as

$$L_9^2 \log^{d+1+2\rho}(n) > 16K_5D_1L_8^d\Bigg(\log(\sqrt{L_{10}}L_8) + \log\Big(n^{\frac{2\alpha\kappa+2\kappa+d}{2\kappa(2\alpha+d)}} \log^{\rho(2/\kappa+2/d-1)-d-1}(n)\Big)\Bigg)^{1+d}$$

which holds for all n such that

$$\log^{2\rho}(n) > \frac{16K_5D_1L_8^d}{L_9^2}\Bigg(\frac{\log(\sqrt{L_{10}}L_8) + \log\big(n^{\frac{2\alpha\kappa+2\kappa+d}{2\kappa(2\alpha+d)}} \log^{\rho(2/\kappa+2/d-1)-d-1}(n)\big)}{\log(n)}\Bigg)^{1+d}$$

Verifying (A.0.17)

We may rewrite (A.0.17) as

$$c_0L_{10}^{\rho} n^{\frac{d\rho}{\kappa(2\alpha+d)}} \log^{4\rho^2/\kappa}(n) > c_5 n^{\frac{d}{2\alpha+d}} \log^{2\rho}(n)$$

If $\rho/\kappa > 1$ this holds for all $L_{10} > (c_5/c_0)^{1/\rho}$.

Verifying (A.0.18)

We may rewrite (A.0.18) as

$$D_1 L_8^d n^{\frac{d}{2\alpha+d}} \log^{2\rho}(n) > 2c_5 n^{\frac{d}{2\alpha+d}} \log^{\frac{2+2d}{2+d/\alpha}}(n)$$

which is satisfied for all $L_8 > (2c_5/D_1)^{1/d}$.

Verifying (A.0.19)

Consider

$$\zeta_n^{q_1-d+1} = L_8^{q_1-d+1} n^{\frac{q_1-d+1}{2\alpha+d}} \log^{\frac{2\rho(q_1-d+1)}{d}}(n)$$

and

$$\exp(c_5 n\delta_n^2) = \exp\big(c_5 n^{\frac{d}{2\alpha+d}} \log^{2\rho}(n)\big).$$

Then (A.0.19) holds for all n such that

$$n^{\frac{d}{2\alpha+d}} > \frac{1}{c_5}\big(\log(L_8^{q_1-d+1}) + 1\big).$$

Verifying (A.0.20)

We may rewrite (A.0.20) as

$$L_9^2 n^{\frac{d}{2\alpha+d}} \log^{d+1+4\rho}(n) \geq 8c_5 n^{\frac{d}{2\alpha+d}} \log^{2\rho}(n).$$

which is satisfied for all $L_9 \geq \sqrt{8c_5}$.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Ryan Prescott Adams, Iain Murray, and David JC MacKay. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM, 2009.

Virginia Aglietti, Edwin V Bonilla, Theodoros Damoulas, and Sally Cripps. Structured variational inference in continuous cox process models. *arXiv preprint arXiv:1906.03161*, 2019.

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

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

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

FJ Anscombe. Sequential medical trials. *Journal of the American Statistical Association*, 58(302): 365–383, 1963.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.

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

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *COLT*, pages 454–468, 2007.

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

Eduard Belitser, Paulo Serra, and Harry van Zanten. Rate-optimal bayesian intensity smoothing for inhomogeneous poisson processes. *Journal of statistical planning and inference*, 166:24–35, 2015.

Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3/4):221–229, 1956.

Donald A Berry. Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association*, 73(362):339–345, 1978.

Donald A Berry and Stephen G Eick. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in medicine*, 14(3):231–246, 1995.

Dimitris Bertsimas and José Niño-Mora. Conservation laws, extended polymatroids and multi-armed bandit problems; a polyhedral approach to indexable systems. *Mathematics of Operations Research*, 21(2):257–306, 1996.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

Hildo Bijl, Thomas B Schön, Jan-Willem van Wingerden, and Michel Verhaegen. A sequential monte carlo approach to thompson sampling for bayesian optimization. *arXiv preprint arXiv:1604.00169*, 2016.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Ilija Bogunovic, Jonathan Scarlett, Andreas Krause, and Volkan Cevher. Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. In *Advances in Neural Information Processing Systems*, pages 1507–1515, 2016.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems. *In Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *NeurIPS*, pages 638–646, 2013.

Sébastien Bubeck and Mark Sellke. First-order regret analysis of thompson sampling. *arXiv preprint arXiv:1902.00681*, 2019.

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems*, pages 201–208, 2009.

Sébastien Bubeck, Remi Munos, Gilles Stoltz, and Csaba Szepesvári. $\mathcal{X}$-armed bandits. *J. Mach. Learn. Res.*, 12:1655–1695, 2011.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pages 111–127, 2018.

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

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

John Gunnar Carlsson, Erik Carlsson, and Raghuveer Devulapalli. Shadow Prices in Territory Division. *Networks and Spatial Economics*, 16(3):893–931, 2016.

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

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

Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In *Advances in Neural Information Processing Systems*, pages 1651–1659, 2016a.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50): 1–33, 2016b.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, volume 70, pages 844–853, 2017.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124, 2015.

Emile Contal, David Buffoni, Alexandre Robicquet, and Nicolas Vayatis. Parallel gaussian process optimization with upper confidence bound and pure exploration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–240. Springer, 2013.

Emile Contal, Vianney Perchet, and Nicolas Vayatis. Gaussian process optimization with mutual information. In *International Conference on Machine Learning*, pages 253–261, 2014.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *The Journal of Machine Learning Research*, 18(1):5638–5665, 2017.

David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*, 2008.

Victor H. de la Peña. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564, 1999.

Peter Diggle. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147, 1985.

Peter J Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

Shi Dong and Benjamin Van Roy. An information-theoretic analysis for thompson sampling with many actions. In *Advances in Neural Information Processing Systems*, pages 4157–4165, 2018.

Christian Donner and Manfred Opper. Efficient bayesian inference of sigmoidal gaussian cox processes. *The Journal of Machine Learning Research*, 19(1):2710–2743, 2018.

James Edwards, Paul Fearnhead, and Kevin Glazebrook. On the identification and mitigation of weaknesses in the knowledge gradient policy for multi-armed bandits. *Probability in the Engineering and Informational Sciences*, 31(2):239–263, 2017.

Peter I Frazier, Warren B Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

Drew Fudenberg and David K Levine. *The Theory of Learning in Games*, volume 2. MIT press, 1998.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5):1466–1478, 2012.

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

Michael R Garey and David S Johnson. Computers and intractability: a guide to np-completeness, 1979.

Aurélien Garivier and Olivier Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *COLT*, pages 359–376, 2011.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

Subhashis Ghosal and Aad Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

Subhashis Ghosal and Aad W Van Der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5): 1233–1263, 2001.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

John C Gittins and David M Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Progress in Statistics*, 9:241–266, 1974.

John C Gittins and David M Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

James A Grant, David S Leslie, Kevin Glazebrook, Roberto Szechtman, and Adam Letchford. Adaptive policies for perimeter surveillance problems, 2018. arXiv:1810.02176.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

Jean-Bastien Grill, Michal Valko, and Remi Munos. Black-box optimization of noisy functions with unknown smoothness. In *NeurIPS*, pages 667–675, 2015.

Steffen Grünewälder, Jean-Yves Audibert, Manfred Opper, and John Shawe-Taylor. Regret bounds for gaussian process bandit problems. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 273–280, 2010.

Shota Gugushvili, Frank van der Meulen, Moritz Schauer, and Peter Spreij. Fast and scalable nonparametric bayesian inference for poisson point processes. *arXiv preprint arXiv:1804.03616*, 2018.

Tom Gunter, Chris Lloyd, Michael A. Osborne, and Stephen J. Roberts. Efficient Bayesian Nonparametric Modelling of Structured Point Processes. In *30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014. URL `http://arxiv.org/abs/1407.6949`.

Eli Gutin and Vivek Farias. Optimistic gittins indices. In *Advances in Neural Information Processing Systems*, pages 3153–3161, 2016.

John R Hauser, Glen L Urban, Guilherme Liberali, and Michael Braun. Website morphing. *Marketing Science*, 28(2):202–223, 2009.

Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.

Juha Heikkinen and Elja Arjas. Modeling a poisson forest in variable elevations: a nonparametric bayesian approach. *Biometrics*, 55(3):738–745, 1999.

Roelof Helmers, I Wayan Mangku, and Ričardas Zitikis. Statistical properties of a kernel-type estimator of the intensity function of a cyclic poisson process. *Journal of Multivariate Analysis*, 92(1):1–23, 2005.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. In *AISTATS*. JMLR, 2015.

José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.

Yu-Chi Ho, Qian-Chuan Zhao, and Qing-Shan Jia. *Ordinal optimization: Soft optimization for hard problems*. Springer Science & Business Media, 2008.

Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors. In *Artificial Intelligence and Statistics*, pages 375–383, 2014.

Alihan Huyuk and Cem Tekin. Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms, 2019.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

ST John and James Hensman. Large-scale cox process inference using variational fourier features. In *International Conference on Machine Learning*, pages 2367–2375, 2018.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Jeff Schneider, and Barnabás Póczos. Parallelised bayesian optimisation via thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 133–142, 2018.

Edward PC Kao and Sheng-Lin Chang. Modeling Time-Dependent Arrivals to Service Systems: A Case in using a Piecewise-polynomial Rate Function in a Nonhomogeneous Poisson Process. *Management Science*, 34(11):1367–1379, 1988.

Alan F Karr. Inference for stationary random fields given poisson samples. *Advances in applied probability*, 18(2):406–422, 1986.

Emilie Kaufmann. On Bayesian index policies for sequential resource allocation. *Annals of Statistics*, 46(2):842–865, 2016.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *AISTATS*, pages 592–600, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012b.

Jaya Kawale, Hung H Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient thompson sampling for online matrix-factorization recommendation. In *Advances in neural information processing systems*, pages 1297–1305, 2015.

Seong-Hee Kim and Barry L Nelson. Recent advances in ranking and selection. In *2007 Winter Simulation Conference*, pages 162–172. IEEE, 2007.

John Frank Charles Kingman. Poisson processes. *Encyclopedia of biostatistics*, 6, 2005.

Alisa Kirichenko and Harry Van Zanten. Optimality of poisson processes intensity learning with gaussian processes. *The Journal of Machine Learning Research*, 16(1):2909–2919, 2015.

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

Robert D. Kleinberg, Alexsandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proc. 40th Annu. ACM Symp. on Theory of Computing*, pages 681–690, 2008.

Levente Kocsis and Csaba Szepesvári. Discounted ucb. In *2nd PASCAL Challenges Workshop*, volume 2, 2006.

Andrey N Kolmogorov and Vladimir Mikhaılovich Tikhomirov. -entropy and -capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364, 1961.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays, 2015.

Bernard O Koopman. Search and screening, operations evaluation group report 56. *Center for Naval Analysis, Alexandria, Virginia*, 1946.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *NeurIPS*, pages 1448–1456, 2013.

Andreas Krause and Cheng S Ong. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2011.

Michael E Kuhl and James R Wilson. Least squares estimation of nonhomogeneous poisson processes. *Journal of Statistical Computation and Simulation*, 67(1):699–712, 2000.

Michael E Kuhl, James R Wilson, and Mary A Johnson. Estimating and Simulating Poisson Processes having Trends or Multiple Periodicities. *IIE transactions*, 29(3):201–211, 1997.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 420–429. AUAI Press, 2014.

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 1987.

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

Tor Lattimore. Regret analysis of the finite-horizon gittins index strategy for multi-armed bandits. In *Conference on Learning Theory*, pages 1214–1245, 2016.

Tor Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. In *Advances in Neural Information Processing Systems*, pages 1583–1592, 2017.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77 (4):853–877, 2015.

Peter AW Lewis and Gerald S Shedler. Statistical Analysis of Non-stationary Series of Events in a Data Base System. *IBM Journal of Research and Development*, 20(5):465–482, 1976.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548. ACM, 2016.

Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts. Variational inference for gaussian process modulated poisson processes. In *International Conference on Machine Learning*, pages 1814–1822, 2015.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163, 2019.

Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, pages 485–492, 2010.

Alexander Luedtke, Emilie Kaufmann, and Antoine Chambaz. Asymptotically optimal algorithms for multiple play bandits with partial feedback. *arXiv preprint arXiv:1606.09388*, 2016.

Benedict C May, N Korda, Anthony Lee, and David S Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *J. Mach. Learn. Res.*, 13:2069–2106, 2012.

Jesper Moller and Rasmus Plenge Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, 2003.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

Iain Murray, Zoubin Ghahramani, and David MacKay. Mcmc for doubly-intractable distributions. *Uncertainty in Artificial Intelligence*, 22:359–366, 2006.

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

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Roman Pogodin and Tor Lattimore. Adaptivity, variance and separation for adversarial bandits. *arXiv preprint arXiv:1903.07890*, 2019.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Yi Qi, Qingyun Wu, Hongning Wang, Jie Tang, and Maosong Sun. Bandit learning with implicit feedback. In *Advances in Neural Information Processing Systems*, pages 7276–7286, 2018.

Stephen L Rathbun and Noel Cressie. A space-time survival point process for a longleaf pine forest in southern georgia. *Journal of the American Statistical Association*, 89(428):1164–1174, 1994.

John Riordan. Moment recurrence relations for binomial, poisson and hypergeometric frequency distributions. *The Annals of Mathematical Statistics*, 8(2):103–111, 1937.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Found. Trends Mach. Learn.*, 11:1–96, 2018.

Ilya O Ryzhov, Warren B Powell, and Peter I Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

Thomas L Saaty. *Elements of queueing theory: with applications*, volume 34203. McGraw-Hill New York, 1961.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742, 2017.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016.

Shubhanshu Shekhar and Tara Javidi. Gaussian process bandits with adaptive discretization. *Electronic Journal of Statistics*, 12(2):3829–3874, 2018.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

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

Lawrence D Stone. *Theory of Optimal Search*, volume 118. Elsevier, 1976.

Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Roberto Szechtman, Moshe Kress, Kyle Lin, and Dolev Cfir. Models of Sensor Operations for Border Surveillance. *Naval Research Logistics (NRL)*, 55(1):27–41, 2008.

Yee W Teh and Vinayak Rao. Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2011.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

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

John N Tsitsiklis. A short proof of the gittins index theorem. *The Annals of Applied Probability*, 4 (1):194–199, 1994.

Iñigo Urteaga and Chris Wiggins. Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 698–706, 2018.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27, 2013.

Aad W van der Vaart and J Harry van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Aad W van der Vaart and J Harry van Zanten. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675, 2009.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

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

Zi Wang, Bolei Zhou, and Stefanie Jegelka. Optimization as estimation with gaussian processes in bandit settings. In *Artificial Intelligence and Statistics*, pages 1022–1031, 2016.

Alan R Washburn. *Search and Detection*. INFORMS, 2002.

Richard Weber. On the gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2 (4):1024–1033, 1992.

Peter Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):143–149, 1980.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA, 2006.

S Faye Williamson, Peter Jacko, Sofía S Villar, and Thomas Jaki. A bayesian adaptive design for clinical trials in rare diseases. *Computational statistics & data analysis*, 113:136–153, 2017.