2020

Adaptive policies for perimeter surveillance problems

Grant, J.A.; Leslie, D.S.; Glazebrook, K.; Szechtman, R.; Letchford, A.N.

Stochastics and Statistics

Adaptive policies for perimeter surveillance problems

James A. Grant [a,*], David S. Leslie [b], Kevin Glazebrook [c], Roberto Szechtman [d], Adam N. Letchford [c]

[a] *STOR-i Centre for Doctoral Training, Lancaster University, UK*
[b] *Department of Mathematics and Statistics, Lancaster University, UK*
[c] *Department of Management Science, Lancaster University, UK*
[d] *Department of Operations Research, Naval Postgraduate School, CA, USA*

A B S T R A C T

We consider the problem of sequentially choosing observation regions along a line, with an aim of maximising the detection of events of interest. Such a problem may arise when monitoring the movements of endangered or migratory species, detecting crossings of a border, policing activities at sea, and in many other settings. In each case, the key operational challenge is to learn an allocation of surveillance resources which maximises successful detection of events of interest. We present a combinatorial multiarmed bandit model with Poisson rewards and a novel filtered feedback mechanism - arising from the failure to detect certain intrusions - where reward distributions are dependent on the actions selected. Our solution method is an upper confidence bound approach and we derive upper and lower bounds on its expected performance. We prove that the gap between these bounds is of constant order, and demonstrate empirically that our approach is more reliable in simulated problems than competing algorithms.

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.

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 according to a Poisson process and the likelihood of an event being detected depends on the allocation of resource chosen by the controller.

The problem of designing an optimal deployment of searchers becomes truly challenging when the number of available searchers is insufficient to guarantee perfect detection of all events, which is often the case in tight fiscal environments. In such a setting the controller faces a classic *resource allocation problem*, where the action set is the set of possible allocations of searchers to segments of the line and the controller aims to find an action which maximises the rate of detection. To compute this rate of detection the controller must know the rate at which events occur along the length of the line and the probabilities with which searchers detect events that have appeared at particular points (under a particular allocation of searchers to parts of the line). It is, of course, a strong assumption that such information is available, particularly at the beginning of a new project.

* Corresponding author.

E-mail addresses: j.grant@lancaster.ac.uk (J.A. Grant), d.leslie@lancaster.ac.uk (D.S. Leslie), k.glazebrook@lancaster.ac.uk (K. Glazebrook), rszechtm@nps.edu (R. Szechtman), a.n.letchford@lancaster.ac.uk (A.N. Letchford).

In this work we consider the more realistic setting where the rate at which events occur is unknown. When the rate at which events occur is unknown the controller has two broad options:

(a) to select an allocation which performs best in expectation according to some prior information (if it exists) and stick to that,
(b) (if possible) to take an adaptive strategy, which alters the allocation of searchers as observations are collected.

In this second scenario a *sequential resource allocation problem* is faced - where the controller wishes to quickly and confidently converge on an optimal allocation while also ensuring appropriate experimentation. This sequential problem is our principal concern in this paper.

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 *multi-armed bandit problems* when designing and analysing solutions to the problem.

Multi-armed bandit 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. 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.

1.1. Related literature

We select a Poisson process as the data generating model for our problem. The Poisson process is famously widely used as a model for spatial and spatiotemporal event data in many settings, such as ecology (Heikkinen & Arjas, 1999; Serra, Saez, Juan, Varga, & Mateu, 2014), and arrival process modelling (Benes, 1957; Weinberg, Brown, & Stroud, 2007). There is a large literature on inference for Poisson processes, which has lead to a variety of sophisticated techniques, such as those involving Gaussian processes (Adams, Murray, & MacKay, 2009; John & Hensman, 2018) or kernel-based smoothing (Diggle, 1985). However the theoretical properties of the more complex methods are typically only understood asymptotically (Gugushvili, van der Meulen, Schauer, & Spreij, 2018; Helmers, Mangku, & Zitikis, 2005; Kirichenko & Van Zanten, 2015) and therefore in the interest of developing tight guarantees on the performance of sequential decision making algorithms, we favour a simple piecewise-constant model for the Poisson process rate in this paper.

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, Kress, Lin, and Cfir (2008), who study the perimeter protection problem when the parameters of the arrival process are fully known, for mobile and fixed searchers. Carlsson, Carlsson, and Devulapalli (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 sequential problem we consider is structurally similar to a combinatorial multi-armed bandit (CMAB) problem (Chen, Wang, & Yuan, 2013). To permit discussion of a CMAB we first describe the simpler multi-armed bandit (MAB) problem (first attributed to Thompson, 1933), which is a special case. The (stochastic) MAB problem models a scenario where an agent is faced with a series of potential actions (or arms), each associated with some underlying probability distribution. In each of a series of rounds, the agent selects a single action and receives a *reward* drawn from the underlying distribution associated with the selected action. The agent's aim is to maximise her cumulative expected reward over some number of rounds, or equivalently minimise her cumulative *regret* - defined as the difference in expected reward between optimal actions and actions actually selected. To succeed in this the agent must manage an exploration-exploitation trade-off as she learns which actions have high expected reward.

The CMAB problem models a richer framework where the agent may select multiple actions in each round and her 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. Numerous authors (Anantharam, Varaiya, & Walrand, 1987, Gai, Krishnamachari, & Jain, 2012, Kveton, Wen, Ashkan, & Szepesvari, 2015b, Combes, Shahi, Proutiere, & Lelarge, 2015, and Luedtke, Kaufmann, & Chambaz, 2019) consider a special case (known as a *multiple play bandit*) where the reward is simply a sum of the random observations and the number of actions which may be selected in one round is limited. A number of other works have since extended the framework of Chen et al. (2013) to model other novel features. Chen et al. (2016a) and Kveton, Szepesvari, Wen, and Ashkan (2015a) consider a setting where playing a subset of arms may randomly trigger additional rewards from other arms, and Chen, Wang, Yuan, and Wang (2016b) considers a broader set of non-linear reward functions. However the CMAB model and UCB approach 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.

Reward maximisation in a CMAB problem requires addressing a similar trade-off between exploration and exploitation to that faced in the MAB problem. For MAB-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, Glazebrook, & Weber, 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 2. More recently, so called upper confidence bound (UCB) algorithms,

first proposed by Lai and Robbins (1985) and Burnetas and Katehakis (1996), and popularised by Auer, Cesa-Bianchi, and Fischer (2002), have attracted much attention as approaches that enjoy efficient implementation and strong theoretical guarantees. These heuristic methods balance exploration and exploitation by selecting actions based on optimistic estimates of the associated expected rewards and can be applied to both MAB and CMAB problems.

Auer et al. (2002) originally proposed a UCB approach for MAB problems 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, Honda, and Katehakis (2018), Bubeck and Cesa-Bianchi (2012), Bubeck, Cesa-Bianchi, and Lugosi (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. (2019) have studied multiple-play bandits 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 (Agrawal & Goyal, 2012; Kaufmann, Korda, & Munos, 2012; Russo & Van Roy, 2016; Wang & Chen, 2018) and approaches which utilise the KL divergence of the reward distributions (Cappé, Garivier, Maillard, Munos, & Stoltz, 2013; Kaufmann, 2017). 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, as the existing theory around KL-based UCB indices relies on independence of the reward generation and action selection mechanisms. Therefore in this work we focus on the theoretical analysis of a more traditional UCB type approach. A TS alternative is presented and evaluated numerically in Section 5.

1.2. Key contributions

This work 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.

1.3. Paper outline

The remainder of the paper is structured as follows. Section 2 introduces a model of the sequential problem. In Section 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 we introduce a solution method, the *Filtered Poisson Combinatorial Upper Confidence Bound* algorithm, for the sequential resource allocation problem, and derive a performance guarantee in the form of an upper bound 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 5 and 6 with numerical experiments and a discussion respectively.

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 $\{1, 2, \ldots, W\}$.

The observation domain (line) 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* $\mathbf{a} := (a_1, a_2, \ldots, a_K) \in \{0, 1, \ldots, U\}^K$ describes a deployment of the searchers across the line. We impose the requirement that $\mathbf{a} \in \mathcal{A}$, the *action set*, where

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

These conditions on $\mathcal{A}$ ensure that 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

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

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 $i \leq j$,

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

Hence, the detection probability in a cell depends only on the subregion 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 $\mathbf{a}_t \in \mathcal{A}$ 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

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

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 $\mathbf{Y}_t = \{Y_{1t}, Y_{2t}, \ldots, Y_{Kt}\} \in \mathbb{N}^K$ is observed. Events in the underlying X-process are observed or not independently of each other. We write

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

It follows from standard theory that

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

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

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

where T denotes vector transposition and λ is the K-vector whose kth component is λ_k.
4. We write

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

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 $\{\pi_t : \mathbf{H}_t \to \mathcal{A}, t \in \mathbb{N}\}$ 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 λ, γ and over partial horizons $\{1, 2, \ldots, n\} \subseteq \mathbb{N}$. We write

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

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

$$opt_{\lambda,\gamma} := \max_{\mathbf{a} \in \mathcal{A}} r_{\lambda,\gamma}(\mathbf{a}),$$

then it is plain that, for any choice of A

$$n \cdot opt_{\lambda,\gamma} \geq \mathbb{E}_A \left(\sum_{t=1}^n r_{\lambda,\gamma}(\mathbf{a}_t) \right),$$

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

$$Reg_{\lambda,\gamma}^A(n) := n \cdot opt_{\lambda,\gamma} - \mathbb{E}_A \left(\sum_{t=1}^n r_{\lambda,\gamma}(\mathbf{a}_t) \right), \tag{1}$$

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 λ. 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

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

where the outer expectation is taken over some prior distribution ρ for the unknown λ. 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 λ. 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 λ. 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 λ. 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 λ at that point.

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

$$opt_{\lambda,\gamma} := \max_{\mathbf{a} \in \mathcal{A}} r_{\lambda,\gamma}(\mathbf{a})$$

for any pre-specified $\lambda \in (\mathbb{R}_+)^K$. 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 λ_t of the unknown λ available at time t. While such a proposal would fail to adequately address the challenge of learning about λ, we will in the succeeding sections develop effective algorithms which choose allocations determined by solutions of full information problems for carefully chosen λ-values.

A challenge to the solution of the full information problem is the non-linearity in $\mathbf{a}$ of the objective $r_{\lambda,\gamma}(\mathbf{a})$ inherited from the non-linearity of the detection mechanism $\gamma(\mathbf{a})$. 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

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

for the mean number of events detected when agent u is allocated to the subregion $\{i, i+1, \ldots, j\}$ 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 $q_{\lambda,\gamma,iju}$ for all $1 \le i \le j \le K$, and $u \in [U]$. We now have that

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

$$\text{such that} \quad \sum_{i=1}^{K} \sum_{j=i}^{K} a_{iju} \le 1, \quad u \in [U]$$

$$\sum_{i=1}^{k} \sum_{j=k}^{K} \sum_{u=1}^{U} a_{iju} \le 1, \quad k \in [K]$$

$$a_{iju} \in \{0, 1\}, \quad 1 \le i \le j \le K, \quad u \in [U]. \quad (2)$$

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 (2) as the optimal allocation strategy and the optimal value function as the best achievable performance for an agent with perfect knowledge of γ and λ.

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 λ) by means of branch and bound. While it can be shown that the IP (2) belongs to a class of problems which is NP-hard (see Appendix A) we find that the solution of this IP is very efficient in practice. We believe that this is because the solution of the LP-relaxation of (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. Sequential problem

In the sequential problem, the controller's objective is to minimise regret (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 λ, 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 2 is closely related to the *Combinatorial Multi Armed Bandit problem* (CMAB) model of Chen et al. (2013). 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 her cumulative reward over a given horizon. Chen et al. study a CMAB problem where the decision-maker receives *semibandit feedback* on her actions, meaning she observes the overall reward but also all realisations of the random variables associated with the selected arms. Realisations of the random variables are identically distributed for a given arm and independent both across time and arms.

In our adaptive searching 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 real 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 exploitative ones.

4.1. Case (I): Known detection probabilities

In our first version of the problem, case (I), the only unknowns are the underlying rate parameters λ. We assume that detection probability vectors $\gamma(\mathbf{a})$ 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 $q_{\lambda,\gamma,iju}$ quantities with the optimistic estimate of λ in place of known λ.

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 1. The algorithm consists of an initialisation phase of length K where allocations are

---

Algorithm 1 FP-CUCB (case (I)).

Inputs: Upper bound $\lambda_{max} \ge \lambda_k$, $k \in [K]$.

Initialisation Phase: For $t \in [K]$

- Select an arbitrary allocation $\mathbf{a} \in \mathcal{A}$ such that $a_t \neq 0$

Iterative Phase: For $t = K + 1, K + 2, \ldots$

- Calculate indices

$$\bar{\lambda}_{k,t} = \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{6 \max(1, \sqrt{\lambda_{max}}) \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}$$
$$+ \sqrt{\frac{6 \lambda_{max} \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}}, \quad k \in [K]$$

- Select an allocation $\mathbf{a}_{\bar{\lambda}_t}^*$ such that $r_{\bar{\lambda}_t,\gamma}(\mathbf{a}_{\bar{\lambda}_t}^*) = \max_{\mathbf{a} \in \mathcal{A}} r_{\bar{\lambda}_t,\gamma}(\mathbf{a})$.

selected such that every cell is searched in some capacity at least once. Then in every subsequent round $t > K$, a UCB index

$$\bar{\lambda}_{k,t} = \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{6 \max(1, \sqrt{\lambda_{max}}) \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \sqrt{\frac{6\lambda_{max} \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}}, \quad (3)$$

is calculated for each cell k. This particular UCB index is chosen because it can be shown to bound λ_k with high probability. Specifically, using de la Peña's inequality (de la Peña, 1999), it can be shown that $P(\bar{\lambda}_{k,t} \geq \lambda_k)$ approaches 1 as $t \to \infty$ at an appropriate rate. A full derivation of this term is given in the proof of the following theorem.

An action which is optimal with respect to the K-vector of inflated rates $\bar{\lambda}_t = (\bar{\lambda}_{1,t}, \ldots, \bar{\lambda}_{K,t})$ is then selected by solving the IP (2) with $\bar{\lambda}_t$ in place of λ. 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.

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,

$$\Delta_{max}^k = \text{opt}_{\lambda, \gamma} - \min_{\mathbf{a} \in \mathcal{A}_k} r_{\lambda, \gamma}(\mathbf{a}),$$

$$\Delta_{min}^k = \text{opt}_{\lambda, \gamma} - \max_{\mathbf{a} \in \mathcal{A}_k} r_{\lambda, \gamma}(\mathbf{a}),$$

where $\mathcal{A}_k = \{\mathbf{a} \in \mathcal{A} : a_k \neq 0\}$ for $k \in [K]$, and $\Delta_{max} = \max_{k \in [K]} \Delta_{max}^k$, and $\Delta_{min} = \min_{k \in [K]} \Delta_{min}^k$. 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_{max}^k and Δ_{min}^k 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 λ, γ but we drop this dependence in the notation for simplicity.

4.1.1. Upper bound on regret

Now, in Theorem 1 we provide an analytical bound on the expected regret of the FP-CUCB algorithm in n rounds.

Theorem 1. *The regret of the FP-CUCB algorithm with λ_{max} applied to the sequential surveillance problem with known γ satisfies*

$$Reg_{\lambda, \gamma}^{\text{FP-CUCB}}(n) \leq \sum_{k: \Delta_{min}^k > 0} \frac{12K^2}{\gamma_{k,min}} \left[\frac{b(\Delta_{min}^k)}{\Delta_{min}^k} + \int_{\Delta_{min}^k}^{\Delta_{max}^k} \frac{b(x)}{x^2} dx \right] \log(n)$$

$$+ \left(\frac{\pi^2}{3} + 1 \right) K \Delta_{max}, \quad (4)$$

where

$$b(x) = \lambda_{max} + \frac{x \max(1, \sqrt{\lambda_{max}})}{K}$$

$$+ \sqrt{\lambda_{max}^2 + \frac{2x\lambda_{max} \max(1, \sqrt{\lambda_{max}})}{K}},$$

and $\gamma_{k,min} = \min_{\mathbf{a}: a_k \neq 0} \gamma_k(\mathbf{a})$.

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, $\gamma(\mathbf{a})$, 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

$\mathcal{G} = \{\gamma(\mathbf{a}), \forall \mathbf{a} \in \mathcal{A}\}$ of possible detection probability vectors most parsimoniously describes the set of possible actions in this problem.

For an element $\mathbf{g} = (g_1, \ldots, g_K) \in \mathcal{G}$ we then have expected reward $\mathbf{g}^T \cdot \lambda$ and optimality gap $\Delta_{\mathbf{g}} = \text{opt}_{\lambda, \gamma} - \mathbf{g}^T \cdot \lambda$. 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 $\mathbf{g}_{k,B}^1, \mathbf{g}_{k,B}^2, \ldots, \mathbf{g}_{k,B}^{B_k}$ in increasing order of expected reward. We use the following notation for optimality gaps with respect to these ordered vectors

$$\Delta^{k,j} = \text{opt}_{\lambda, \gamma} - (\mathbf{g}_{k,B}^j)^T \cdot \lambda \quad j \in [B_k], \quad k \in [K] \quad (5)$$

and thus the gaps defined previously can be expressed as $\Delta_{max}^k = \Delta^{k,1}$ and $\Delta_{min}^k = \Delta^{k,B_k}$. 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 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 $k' = \arg \min_{j:g_{j,t}>0} N_{j,t-1}$ (i.e. k' indexes the cell involved in the current action which has the lowest counter), where if k' is non-unique, we choose a single value randomly from the minimising set. If $\mathbf{g}_t^T \cdot \lambda \neq \text{opt}_{\lambda, \gamma}$ 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 sub-optimal plays in t rounds (since one of the first K actions may be optimal), and $D_{k,t} \geq \gamma_{k,min} N_{k,t}$ for all k and t (since cell k is always searched with detection probability at least $\gamma_{k,min}$). While tracking the sub-optimal plays in this way is more complex than maintaining a single counter of the number of sub-optimal actions, it permits a convenient decomposition of regret that allows us to prove Theorem 1.

Proof of Theorem 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). The proofs of these propositions is reserved for Appendix C.

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:

$$N_{k,n}^{l,suf} = \sum_{t=K+1}^n \mathbb{I}\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{k,t-1} > h_{k,n}(\Delta^{k,l})\}, \quad (6)$$

$$N_{k,n}^{l,und} = \sum_{t=K+1}^n \mathbb{I}\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{k,t-1} \leq h_{k,n}(\Delta^{k,l})\}, \quad (7)$$

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 $\mathbf{g}_{k,B}^l$ if $N_{k,t-1} > h_{k,n}(\Delta^{k,l})$, and thus $N_{k,n}^{l,und}, N_{k,n}^{l,suf}$ count the suboptimal plays leading to incrementing $N_{k,n}^l$ up to and after the sufficient level, respectively.

From the definitions (6) and (7) we have $N_{k,n} = 1 + \sum_{l=1}^{B_k}(N_{k,n}^{l,suf} + N_{k,n}^{und})$. The expected regret at time horizon n can also be bounded above using this notation as

$$Reg_{\lambda,\gamma}(n) \leq \mathbb{E}\left[\sum_{k=1}^{K}\left(\Delta^{k,1} + \sum_{l=1}^{B_k}(N_{k,n}^{l,suf} + N_{k,n}^{l,und}) \cdot \Delta^{k,l}\right)\right] \quad (8)$$

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 (8).

Firstly, for the beyond sufficiency counter we have

Proposition 1. *For any time horizon $n > K$,*

$$\mathbb{E}\left(\sum_{k=1}^{K}\sum_{l=1}^{B_k} N_{k,n}^{l,suf}\right) \leq \frac{\pi^2}{3} \cdot K. \quad (9)$$

The full proof of Proposition 1 is given in Appendix C, 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 $Y_1, \ldots, Y_t$ 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 $\gamma_1, \ldots, \gamma_s$ 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 1. *Let $Y_1, \ldots, Y_s$ be any sequence of Poisson random variables with means $\gamma_1\lambda, \ldots \gamma_s\lambda$ respectively, such that the sequence $\{Z_j\}_{j=1}^{s}$ is a martingale where $Z_j = \sum_{i=1}^{j}(Y_i - \mathbb{E}(Y_i|Y_{i-1}, \ldots, Y_1))$. Then, given parameters $t \geq s$ and $\lambda_{max} \geq \lambda$ the following holds:*

$$\mathbb{P}\left(\left|\frac{\sum_{j=1}^{s} Y_j}{\sum_{j=1}^{s} \gamma_j} - \lambda\right| \geq \frac{6\max(1, \sqrt{\lambda_{max}})\log(t)}{\sum_{j=1}^{s}\gamma_j} + \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{j=1}^{s}\gamma_j}}\right)$$
$$\leq 2t^{-3}. \quad (10)$$

The proof of this Lemma is given in Appendix B. The consequence of this Lemma is that the UCB indices (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 Appendix C,

Proposition 2. *For any time horizon $n > K$ and $k : \Delta_{min}^k > 0$,*

$$\sum_{l=1}^{B_k} N_{k,n}^{l,und}\Delta^{k,l} \leq h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \int_{\Delta^{k,B_k}}^{\Delta^{k,1}} h_{k,n}(x)dx. \quad (11)$$

Combining the decomposition (8), with the bounds (9) and (11) we have

$$Reg_{\lambda,\gamma}(n) \leq \mathbb{E}\left(\sum_{k=1}^{K}\left(\Delta^{k,1} + \sum_{l=1}^{B_k}(N_{k,n}^{l,suf} + N_{k,n}^{l,und})\Delta^{k,l}\right)\right)$$

$$= \mathbb{E}\left(\sum_{k=1}^{K}\left(\Delta^{k,1} + \sum_{l=1}^{B_k} N_{k,n}^{l,suf}\Delta^{k,l}\right)\right)$$

$$+ \mathbb{E}\left(\sum_{k=1}^{K}\sum_{l=1}^{B_k} N_{k,n}^{l,und}\Delta^{k,l}\right)$$

$$\leq K\Delta_{max} + \mathbb{E}\left(\sum_{k=1}^{K}\sum_{l=1}^{B_k} N_{k,n}^{l,suf}\Delta^{k,l}\right)$$

$$+ \sum_{k:\Delta_{min}^k > 0}\left(h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \int_{\Delta^{k,B_k}}^{\Delta^{k,1}} h_{k,n}(x)dx\right)$$

$$\leq \left(\frac{\pi^2}{3} + 1\right)K\Delta_{max}$$

$$+ \sum_{k:\Delta_{min}^k > 0}\left(h_{k,n}(\Delta_{min}^k)\Delta_{min}^k + \int_{\Delta_{min}^k}^{\Delta_{max}^k} h_{k,n}(x)dx\right)$$

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

$$+ \left(\frac{\pi^2}{3} + 1\right)K\Delta_{max}. \quad \square$$

In the remainder of this section we show that the bound obtained in Theorem 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_{min}^k and Δ_{max}^k, $\forall k \in [K]$.

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

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

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

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

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

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

as the set of mean vectors such that all actions in $J(\lambda)$ are suboptimal but this cannot be discerned by playing only actions in $J(\lambda)$. The larger the set $B(\lambda)$, the more challenging the problem is. If $\mathcal{I}(\lambda) = [K]$, then the problem is trivial as 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 2. *For any $\lambda \in \mathbb{R}_+^K$ such that $B(\lambda) \neq \emptyset$, and for any uniformly good policy π for the sequential surveillance problem with known γ, we have*

$$\liminf_{n \to \infty} \frac{Reg_{\lambda,\gamma}^{\pi}(n)}{\log(n)} \geq c(\lambda) \tag{12}$$

where $c(\lambda)$ is the optimal value of the following optimisation problem over non-negative coefficients $\mathbf{d} = \{d_{\mathbf{g}}, \mathbf{g} \in S(\lambda)\}$,

$$\inf_{\mathbf{d} \geq \mathbf{0}} \sum_{\mathbf{g} \in S(\lambda)} d_{\mathbf{g}} \Delta_{\mathbf{g}} \tag{13}$$

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

We note that the lower bound is based on the KL-divergence of the cell means, and this suggests that, as in simpler MAB problems, an approach incorporating the KL-divergence in the UCB indices could be asymptotically optimal. However, existing theory on the convergence of such approaches (Combes et al., 2015; Garivier & Cappé, 2011) pertains only to the case of independent reward generation action selection mechanisms. Therefore, it is not clear how to approach the optimal design and analysis of such an approach.

4.1.3. Gap-free bound on regret

The logarithmic order bounds of Theorems 1 and 2 are useful as they establish the order-optimality of the UCB algorithm. We note that the coefficients of the two bounds are not the same, and the upper bound may be very large in problem instances where the Δ_{min}^{k} terms are very small.

The main purpose, however, of the bounds in Theorems 1 and 2 is analytical, not numerical as they can be challenging to compute in practice. The computation of the Δ_{min}^{k} and Δ_{max}^{k} terms used in the upper bound requires evaluating the expected reward of every possible action, which quickly becomes computationally challenging for even modest values of K and U. The computation of the lower bound again requires computation of the expected reward of every possible action, to calculate the $\Delta_{\mathbf{g}}$ terms, and also a minimisation over $|S(\lambda)|$ variables, subject to a non-linear constraint. This optimisation problem lacks an convenient analytical solution and must be resolved numerically.

Moreover, in absence of knowledge of the true reward generating parameters these bounds 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 $\lambda \in [0, \lambda_{max}]^K$ and does not depend on the optimality gaps.

Theorem 3. *The regret of the FP-CUCB algorithm with λ_{max} applied to the sequential surveillance problem with known γ satisfies*

$Reg_{\lambda,\gamma}^{FP-CUCB}(n)$

$$\leq \frac{5K\lambda_{max}}{2} + \frac{12K \max(1, \sqrt{\lambda_{max}})}{\gamma_{min}} \log(n)(1 + \log(n))$$

$$+ \sqrt{\frac{92K^2 \lambda_{max} n \log(n)}{\gamma_{min}}}. \tag{15}$$

Proof of Theorem 3: We first consider the following decomposition of regret,

$$Reg_{\lambda,\gamma}^{FP-CUCB}(n) = \mathbb{E}\left(\sum_{t=1}^{n} r_{\lambda,\gamma}(\mathbf{a}^*) - r_{\lambda,\gamma}(\mathbf{a}_t)\right)$$

$$= \mathbb{E}\left(\sum_{t=1}^{n} r_{\lambda,\gamma}(\mathbf{a}^*) - r_{\lambda,\gamma}(\mathbf{a}_t) + r_{\bar{\lambda}_t,\gamma}(a_t) - r_{\bar{\lambda}_t,\gamma}(a_t)\right)$$

$$= \mathbb{E}\left(\sum_{t=1}^{n} \sum_{k=1}^{K} \lambda_k g_k^* - \bar{\lambda}_{kt} g_{kt} + \bar{\lambda}_{kt} g_{kt} - \lambda_k g_{kt}\right)$$

$$\leq \mathbb{E}\left(\sum_{t=1}^{n} \sum_{k=1}^{K} (\lambda_k - \bar{\lambda}_{kt}) g_k^* + (\bar{\lambda}_{kt} - \lambda_k) g_{kt}\right)$$

$$= \sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left((\lambda_k - \bar{\lambda}_{kt}) g_k^*\right) + \sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left((\bar{\lambda}_{kt} - \lambda_k) g_{kt}\right) \tag{16}$$

The terms of the first sum in (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:

$$\sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{E}((\lambda_k - \bar{\lambda}_{kt})) g_k^*$$

$$\leq \sum_{t=1}^{n} \sum_{k=1}^{K} g_k^* \mathbb{P}(\lambda_k > \bar{\lambda}_{kt}) \mathbb{E}(\lambda_k - \bar{\lambda}_{kt} | \lambda_k > \bar{\lambda}_{kt})$$

$$\leq \sum_{t=1}^{n} \lambda_{max} \sum_{k=1}^{K} \mathbb{P}(\lambda_k > \bar{\lambda}_{kt})$$

$$= K\lambda_{max} \sum_{t=1}^{n} \mathbb{P}\left(\sum_{k=1}^{K} \lambda_k > \sum_{k=1}^{K} \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{6 \max(1, \sqrt{\lambda_{max}}) \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}\right.$$

$$\left. + \sqrt{\frac{6\lambda_{max} \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}}\right)$$

$$\leq K\lambda_{max} \sum_{t=1}^{n} t^{-3} \leq \frac{5K\lambda_{max}}{4}$$

where the penultimate inequality is due to Lemma 1.

Now consider the second sum in (16)

$$\sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left((\bar{\lambda}_{kt} - \lambda_k) g_{kt}\right)$$

$$= \sum_{t=1}^{n} \mathbb{E}\left(\sum_{k=1}^{K} \gamma_{k,t} \left(\frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{6 \max(1, \sqrt{\lambda_{max}}) \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}\right.\right.$$

$$\left.\left. + \sqrt{\frac{6\lambda_{max} \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}} - \lambda_k\right)\right)$$

$$\leq \sum_{t=1}^{n} K\lambda_{max} \mathbb{P}\left(\sum_{k=1}^{K} \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} - \lambda_k > \sum_{k=1}^{K} \frac{6 \max(1, \sqrt{\lambda_{max}}) \log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}\right.$$

$$+ \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Bigg)$$

$$+ \sum_{t=1}^{n}\mathbb{E}\Bigg(\sum_{k=1}^{K}2\gamma_{k,t}\Bigg(\frac{6\max(1,\sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}} + \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Bigg)\Bigg)$$

$$\leq \frac{5K\lambda_{max}}{4} + \mathbb{E}\Bigg(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{12\gamma_{k,t}\max(1,\sqrt{\lambda_{max}})\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}\Bigg)$$

$$+ \mathbb{E}\Bigg(\sum_{t=1}^{n}\sum_{k=1}^{K}\gamma_{k,t}\sqrt{\frac{24\lambda_{max}\log(t)}{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Bigg)$$

$$\leq \frac{5K\lambda_{max}}{4} + 12\max(1,\sqrt{\lambda_{max}})\log(n)\mathbb{E}\Bigg(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{\gamma_{k,t}}{\sum_{s=1}^{t-1}\gamma_{k,s}}\Bigg)$$

$$+ \sqrt{24\lambda_{max}\log(n)}\mathbb{E}\Bigg(\sum_{t=1}^{n}\sum_{k=1}^{K}\frac{\gamma_{k,t}}{\sqrt{\sum_{s=1}^{t-1}\gamma_{k,s}}}\Bigg)$$

Consider the expectation in the final term, we have,

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

Similarly for the other expectation, we have

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

Pulling this all together we have the following gap-free bound on regret:

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

as stated in Theorem 3. □

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

$$\gamma_k(\mathbf{a}) = \sum_{u=1}^{U}\phi_u(\mathbf{a})\omega_{ku}\mathbb{I}\{a_k = u\}, \quad k \in [K], \tag{17}$$

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 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 λ, the K-vector of rate parameters, and $\omega = (\omega_{1,1}, \ldots, \omega_{1,U}, \omega_{2,1}, \ldots, \omega_{K,U})$, the KU-vector of baseline detection probabilities are unknown as opposed to solely λ in case (I). Due to nonidentifiability we cannot make direct inference on λ or ω. 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 $\tau = (\omega_{1,1}\lambda_1, \ldots, \omega_{1,U}\lambda_1, \omega_{2,1}\lambda_2, \ldots, \omega_{K,U}\lambda_K)$ when referring to the unknown parameters.

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 $\phi(\mathbf{a}) = (\phi_1(\mathbf{a}), \ldots, \phi_U(\mathbf{a}))$. 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

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

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

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

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

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 2.

Algorithm 2 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 $t = KU + 1, KU + 2, \ldots$

- Calculate indices

$$\bar{\tau}_{ku,t} = \frac{\sum_{s=1}^{t-1}Y_{ku,s}}{\sum_{s=1}^{t-1}\phi_{ku,s}} + \frac{6\max(1,\sqrt{\tau_{max}})\log(t)}{\sum_{s=1}^{t-1}\phi_{ku,s}}$$
$$+ \sqrt{\frac{6\tau_{max}\log(t)}{\sum_{s=1}^{t-1}\phi_{ku,s}}}, \quad ku \in [K] \times [U]$$

- Select an allocation $\mathbf{a}_{\lambda_t^*}^*$ such that $\tilde{r}_{\bar{\tau}_t,\phi}(\mathbf{a}_{\lambda_t^*}^*) = \max_{\mathbf{a} \in \mathcal{A}}\tilde{r}_{\bar{\tau}_t,\phi}(\mathbf{a})$.

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 1 and 2 can be derived. Specifically we have a regret upper bound for FP-CUCB in Corollary 1 and a lower bound for regret of any uniformly good algorithm in Corollary 2.

Corollary 1. *The regret of the FP-CUCB algorithm in case (b) defined by τ_{max} applied to the sequential search problem as defined previously*

satisfies

$$Reg_{\lambda,\gamma}^{\text{FP-CUCB}}(n)$$

$$\leq \sum_{ku:\Delta_{min}^{ku}>0} \frac{12(KU)^2}{\phi_{ku,min}} \left[\frac{b'(\Delta_{min}^{ku})}{\Delta_{min}^{ku}} + \int_{\Delta_{min}^{ku}}^{\Delta_{max}^{ku}} \frac{b'(x)}{x^2} dx \right] \log(n)$$

$$+ \left(\frac{\pi^2}{3} + 1 \right) KU \Delta_{max},$$

where

$$\tilde{b}(x) = \tau_{max} + \frac{x \max(1, \sqrt{\tau_{max}})}{KU}$$

$$+ \sqrt{\tau_{max}^2 + \frac{2x\tau_{max}\max(1,\sqrt{\tau_{max}})}{KU}},$$

and $\phi_{ku,min} = \min_{\mathbf{a}:a_k=u} \phi_u(\mathbf{a})$.

Corollary 2. 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

$$\lim_{n\to\infty} \inf \frac{Reg_{\lambda,\gamma}^{\pi}(n)}{\log(n)} \geq \tilde{c}(\boldsymbol{\tau})$$

where $\tilde{c}(\boldsymbol{\tau})$ is the solution of an optimisation problem analogous to (13).

Precise specification of $\tilde{c}(\boldsymbol{\tau})$ requires redefining notation from Section 4.1.2 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 1 and 2 can be easily extended to prove Corollaries 1 and 2.

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 3 and 4 respectively.

Algorithm 3 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 $t = K + 1, K + 2, \ldots$

- 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 $\mathbf{a}_{\hat{\lambda}_t}^*$ such that $r_{\hat{\lambda}_t,\gamma}(\mathbf{a}_{\hat{\lambda}_t}^*) = \max_{\mathbf{a}\in\mathcal{A}} r_{\hat{\lambda}_t,\gamma}(\mathbf{a})$.

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 (17) but we know both the ϕ functions and ω values.

Algorithm 4 Thompson Sampling (TS).

Inputs: Gamma prior parameters α, β
Iterative Phase: For $t = 1, 2, \ldots$

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

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 $n = 2000$. 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 $k = 1, \ldots, 10$, $[20 - k, 30 - k]$ for $k = 11, \ldots, 20$, $[k - 20, k - 10]$ for $k = 21, \ldots, 30$, $[40 - k, 50 - k]$ for $k = 31, \ldots, 40$, and $[k - 40, k - 30]$ for $k = 41, \ldots, 50$. 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 prior parameters 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

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

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 Fig. 1 we illustrate how regret evolves over time by plotting the median scaled regret across the 250 runs of each algorithm in
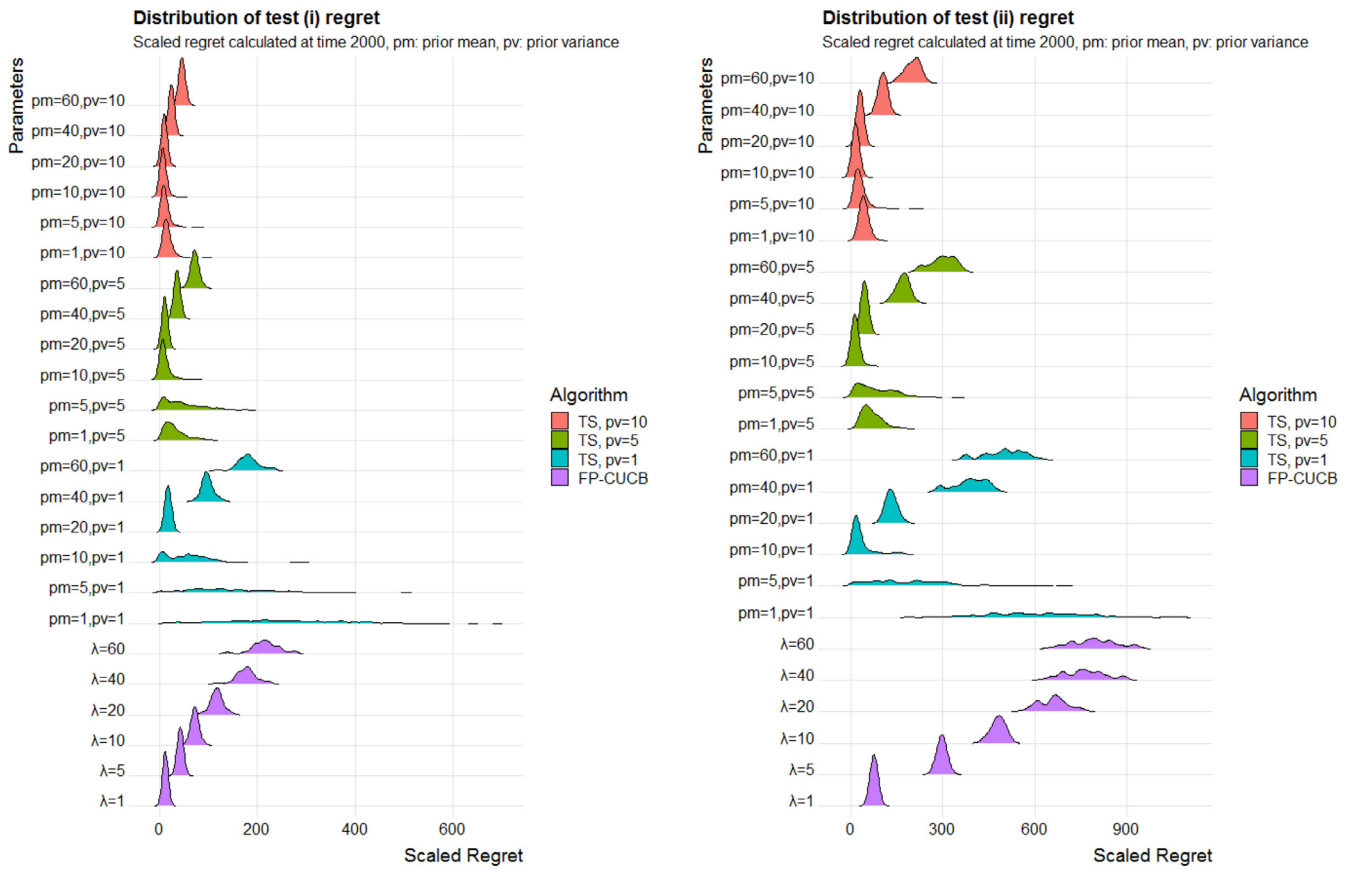
Fig. 2. Scaled regret distributions in tests (i) and (ii). In both tests we have a true largest rate of 20.
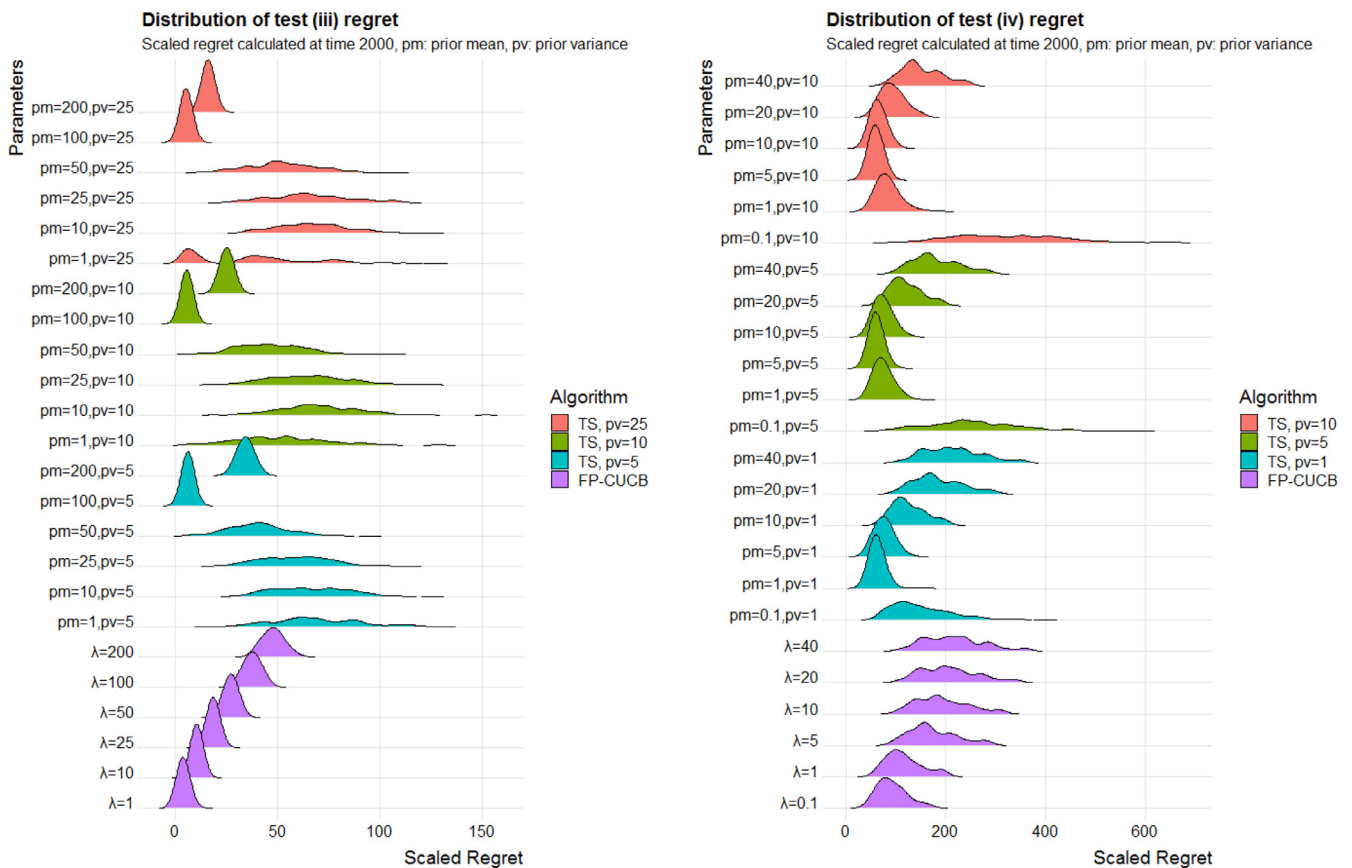

Fig. 3. Scaled regret distributions in tests (iii) and (iv). In test (iii) the true largest rate is 100, and in test (iv) the true largest rate is 1.

6. Discussion

In this paper 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. The problem is real, and has important applications in ecology, security, defence and other areas. 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.

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 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, Cesa-Bianchi, Freund, & Schapire, 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 & 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.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2019.11.004

References

Adams, R. P., Murray, I., & MacKay, D. J. (2009). Tractable nonparametric bayesian inference in poisson processes with Gaussian process intensities. In *Proceedings of the 26th annual international conference on machine learning* (pp. 9–16). ACM.

Agrawal, S., & Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the conference on learning theory* (pp. 39–41).

Anantharam, V., Varaiya, P., & Walrand, J. (1987). 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.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning, 47*(2–3), 235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science* (pp. 322–331). IEEE.

Benes, V. (1957). On queues with poisson arrivals. *The Annals of Mathematical Statistics, 28*(3), 670–677.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems. *Foundations and Trends in Machine Learning, 5*(1), 1–122.

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

Burnetas, A. N., & Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics, 17*(2), 122–142.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., & Stoltz, G. (2013). Kullback-–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics, 41*(3), 1516–1541.

Carlsson, J. G., Carlsson, E., & Devulapalli, R. (2016). Shadow Prices in Territory Division. *Networks and Spatial Economics, 16*(3), 893–931.

Cesa-Bianchi, N., & Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences, 78*(5), 1404–1422.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., & Lu, P. (2016a). Combinatorial multi-armed bandit with general reward functions. In *Proceedings of the advances in neural information processing systems* (pp. 1651–1659).

Chen, W., Wang, Y., & Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th international conference on machine learning* (pp. 151–159).

Chen, W., Wang, Y., Yuan, Y., & Wang, Q. (2016b). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research, 17*(50), 1–33.

Combes, R., Shahi, M. S. T. M., Proutiere, A., & Lelarge, M. (2015). Combinatorial bandits revisited. In *Proceedings of the advances in neural information processing systems* (pp. 2116–2124).

Cowan, W., Honda, J., & Katehakis, M. N. (2018). Normal bandits of unknown means and variances. *J. Mach. Learn. Res., 18*(158), 1–28.

Diggle, P. (1985). A kernel method for smoothing point process data. *Applied statistics, 34*(2), 138–147.

Gai, Y., Krishnamachari, B., & Jain, R. (2012). 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.

Garivier, A., & Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the Colt* (pp. 359–376).

Gittins, J., Glazebrook, K., & Weber, R. (2011). *Multi-armed bandit allocation indices.* John Wiley & Sons.

Graves, T. L., & Lai, T. L. (1997). Asymptotically efficient adaptive choice of control laws incontrolled Markov chains. *SIAM Journal on Control and Optimization, 35*(3), 715–743.

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

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

Helmers, R., Mangku, I. W., & Zitikis, R. (2005). Statistical properties of a kernel–type estimator of the intensity function of a cyclic poisson process. *Journal of Multivariate Analysis, 92*(1), 1–23.

John, S. T., & Hensman, J. (2018). Large-scale Cox process inference using variational Fourier features. In *International Conference on Machine Learning* (pp. 2367–2375).

Kaufmann, E. (2017). On bayesian index policies for sequential resource allocation. *Annals of Statistics, 46*(2), 842–865.

Kaufmann, E., Korda, N., & Munos, R. (2012). Thompson sampling: an asymptotically optimal finite-time analysis. In *Proceedings of the international conference on algorithmic learning theory* (pp. 199–213). Springer.

Kirichenko, A., & Van Zanten, H. (2015). Optimality of poisson processes intensity learning with gaussian processes. *The Journal of Machine Learning Research, 16*(1), 2909–2919.

Koopman, B. (1946). *Search and screening, operations evaluation group report 56.* Alexandria, Virginia: Center for Naval Analysis.

Kveton, B., Szepesvari, C., Wen, Z., & Ashkan, A. (2015a). Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the international conference on machine learning* (pp. 767–776).

Kveton, B., Wen, Z., Ashkan, A., & Szepesvari, C. (2015b). Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the artificial intelligence and statistics* (pp. 535–543).

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

Lattimore, T. (2017). A scale free algorithm for stochastic bandits with bounded kurtosis. In *Proceedings of the advances in neural information processing systems* (pp. 1583–1592).

Luedtke, A., Kaufmann, E., & Chambaz, A. (2019). Asymptotically optimal algorithms for multiple play bandits with partial feedback. To Appear in Machine Learning.

de la Peña, V. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability, 27*(1), 537–564.

Russo, D., & Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research, 17*(1), 2442–2471.

Serra, L., Saez, M., Juan, P., Varga, D., & Mateu, J. (2014). A spatio-temporal poisson hurdle point process to model wildfires. *Stochastic environmental research and risk assessment, 28*(7), 1671–1684.

Stone, L. D. (1976). *Theory of optimal search*: 118. Elsevier.

Szechtman, R., Kress, M., Lin, K., & Cfir, D. (2008). Models of sensor operations for border surveillance. *Naval Research Logistics (NRL), 55*(1), 27–41.

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

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

Washburn, A. R. (2002). Search and detection. *Topics in Operations Research Series* (4th Edition). Linthicum, MD: Institute for Operations Research and the Management Sciences.

Weinberg, J., Brown, L. D., & Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association, 102*(480), 1185–1198.