Adaptive Policies for Perimeter Surveillance Problems

James A. Grant[1], David S. Leslie[2], Kevin Glazebrook[3], Roberto Szechtman[4], and Adam Letchford[3]

[1]*STOR-i Centre for Doctoral Trainng, Lancaster University*
[2]*Department of Mathematics and Statistics, Lancaster University*
[3]*Department of Management Science, Lancaster University*
[4]*Department of Operations Research, Naval Postgraduate School*

October 5, 2018

Abstract

Maximising the detection of intrusions is a fundamental and often critical aim of perimeter surveillance. Commonly, this requires a decision-maker to optimally allocate multiple searchers to segments of the perimeter. We consider a scenario where the decision-maker may sequentially update the searchers' allocation, learning from the observed data to improve decisions over time. In this work we propose a formal model and solution methods for this sequential perimeter surveillance problem. Our model is a combinatorial multi-armed bandit (CMAB) with Poisson rewards and a novel *filtered* feedback mechanism - arising from the failure to detect certain intrusions. 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 intrusions along a border or perimeter. Tracking illegal fishing and smugglers, quantifying traffic flows and detecting adversarial intrusions are a few among many examples of potential civilian or military aims in this domain. These problems are often of high importance, and timeliness is critical for mission success. Thus approaches to the optimal design of surveillance 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.

In these surveillance tasks the notion of optimality can be equated to maximising the rate of detection of intrusions, or equivalently, detecting as many intrusions as possible over some fixed time horizon. We will consider a scenario where surveillance is performed by a fleet of searchers, coordinated by a central agent referred to as the *controller* who chooses which segments of the perimeter each searcher will survey. The precise definition of *searchers* will depend on the context but may include human patrollers, or sensors mounted on fixed or mobile platforms. The search may take place in the air (e.g., with unmanned aerial vehicles (UAVs)) or in the water (e.g., with unmanned underwater vehicles (UUVs) or static sonobuoys). In any case the problem of designing

arXiv:1810.02176v1 [cs.LG] 4 Oct 2018

an optimal surveillance strategy becomes truly challenging when the size of the fleet is insufficient to guarantee perfect detection of all intruders, which is typically 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 perimeter 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 intruders appear along the length of the perimeter and the probabilities with which searchers detect intruders that have appeared at particular points (under a particular allocation of searchers to perimeter segments). It is, of course, a strong assumption that such information is available, particularly at the beginning of a new surveillance project.

In this work we consider the more realistic setting where the rate at which intruders appear is unknown, and two cases in terms of the information available on the probability of successful detection. When this rate function 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 after appropriate experimentation. This sequential problem is our principal concern in this paper. Such a scenario is all the more plausible in the advent of technologies such as UAVs or UUVs whose allocations can be much more easily updated than fixed cameras, or slow moving human searchers.

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 perimeter 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. We argue that this is a reasonable simplification to make as over large distances, decisions about allocations will typically be made to the nearest hundred metres or kilometre. 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 intruders 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 an intruder is more likely to escape if it is not detected within some window of time. An optimal action may well be to assign each searcher to a single cell.

1.1 Related Literature

Search theory started 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.

More recently, Park et al. (2016) use approximate dynamic programming to find good policies for online detection on a closed perimeter. Their problem has a sequential aspect however their formulation differs substantially from ours as they focus on a particular scenario where static sensors flag up intrusions and the decisions to be made are around visiting sensor locations to gather further information. We focus on detecting as many intrusions as possible with mobile searchers. Papadaki et al. (2016) analyse a perimeter formed by a discrete linear network, defended by a single agent who can move one node per period, and a strategic attacker who wins if an attack on a node is not detected within a time window.

Our work, however, is the first to model the problem of allocating multiple searchers to detect multiple intruders with the capacity to sequentially adapt searcher allocation and learn the distribution of intrusions. The sequential problem we consider is structurally similar to a combinatorial multi-armed bandit (CMAB) problem (Chen et al., 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. consider a setting where this function may be non-linear. Numerous authors (Anantharam et al. (1987), Gai et al. (2012), Kveton et al. (2015b), Combes et al. (2015), and Luedtke et al. (2016)) 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 et al. (2015a) consider a setting where playing a subset of arms may randomly trigger additional rewards from other arms, and Chen et al. (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.

Reward maximisation in a CMAB problem requires addressing a similar exploration-exploitation

trade-off 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 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 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 et al. (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. 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 et al. (2015), 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 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 type approaches (Kaufmann et al., 2012; Agrawal and Goyal, 2012; Russo and Van Roy, 2016) 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 more general CMAB problem with non-linear rewards presents a significant analytical challenge, and therefore in this work we focus on the analysis of a UCB type approach.

1.2 Key Contributions

This work makes a number of contributions to the theory of optimal search, multi-armed bandits and broader online optimisation. Simultaneously we give a practically useful solution to a real problem in perimeter surveillance. We summarise the headline contributions below:

- Introduction of a formal model for sequential, multi-searcher perimeter surveillance 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 concentration inequality for filtered Poisson data;

- Regret analysis leading to an optimal order analytical 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 and the unpredictable performance of Thompson Sampling.

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 perimeter surveillance 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 for the sequential resource allocation problem, FP-CUCB, 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 Ω let the notation $[\Omega]$ represent the set $\{1, 2, ..., \Omega\}$.

A domain of search (perimeter) comprises K cells while searches are conducted by U agents. We write
$$a_k = u, \quad k \in [K], \ u \in [U]$$
to denote the deployment of agent u to search 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 agents across the perimeter. We impose the requirement that $\mathbf{a} \in \mathcal{A}$, the *action set*, where
$$\mathcal{A} = \{\mathbf{a} : \ a_i = a_j = u \Rightarrow a_k = u, \ i \leq k \leq j \ \forall i < j, u \in [U]\}.$$
These conditions on $\mathcal{A}$ ensure that agents are assigned to disjoint connected sub-regions of the perimeter. This constraint eliminates unnecessary travel times for the agents and the possibility of multiple agents detecting the same event. The actions are uniquely defined by indicator variables
$$a_{iju} = 1 \Leftrightarrow \text{agent } u \text{ is assigned to the cells } \{i, i+1, ..., j\} \text{ only.}$$

Each action $\mathbf{a} \in \mathcal{A}$ gives rise to a certain detection probability $\gamma_k(\mathbf{a}) \in [0, 1]$ in each cell $k \in [K]$. The detection probabilities capture the effectiveness of each searcher in finding an intruder 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 agents. We assume that if a cell is searched there will be some non-zero probability of detecting intrusions that occur. That is to say $\gamma_k(\mathbf{a}) > 0$ for any $\mathbf{a} \in \mathcal{A} : a_k \neq 0$ for any $k \in [K]$.

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 agent 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 or the time it takes intruders to cross the perimeter.

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. Intrusion 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 $\mathbf{Y}_t = \{Y_{1t}, Y_{2t}, ..., Y_{Kt}\} \in \mathbb{N}^K$ 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

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

where T 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 $\{\pi_t : \mathbf{H}_t \rightarrow \mathcal{A}, t \in \mathbb{N}\}$ 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 $\boldsymbol{\lambda}, \boldsymbol{\gamma}$ and over partial horizons $\{1, 2, ..., n\} \subseteq \mathbb{N}$. We write

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

for the mean number of events observed up to time $t \in \mathbb{N}$ 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

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

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*

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

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.

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 $q_{\boldsymbol{\lambda},\boldsymbol{\gamma},iju}$ for all $1 \leq i \leq j \leq K$, and $u \in [U]$. We now have that

$$\text{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} = \max_{\{a_{iju}, 1 \leq i \leq j \leq K, u \in [U]\}} \sum_{i=1}^{K}\sum_{j=i}^{K}\sum_{u=1}^{U} q_{\boldsymbol{\lambda},\boldsymbol{\gamma},iju} a_{iju} \tag{2}$$

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

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

$$a_{iju} \in \{0,1\}, \quad 1 \leq i \leq j \leq K, \quad u \in [U].$$

The first constraint above guarantees that each agent u is assigned to at most one sub-region while the second constraint guarantees that each cell k is searched by at most one agent. 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 $\boldsymbol{\gamma}$ 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 (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 $\boldsymbol{\lambda}$, 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 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 surveillance 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 of 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. A second related point is that 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 less 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 $\boldsymbol{\lambda}$. 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 $q_{\boldsymbol{\lambda},\boldsymbol{\gamma},iju}$ 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 1. 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 $\bar{\lambda}_{k,t}$ is calculated for each cell k as the sum of an empirical mean for filtered data and an appropriate inflation term. An action which is optimal with respect to the K-vector of inflated rates $\bar{\boldsymbol{\lambda}}_t = (\bar{\lambda}_{1,t}, ..., \bar{\lambda}_{K,t})$ is then selected by solving the IP (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 1 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

$$\bar{\lambda}_{k,t} = \frac{\sum_{s=1}^{t-1} Y_{k,s}}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \frac{2\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}} + \sqrt{\frac{6\lambda_{max}\log(t)}{\sum_{s=1}^{t-1} \gamma_{k,s}}}, \quad k \in [K] \tag{3}$$

- 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,

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

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

where $\mathcal{A}_k = \{\mathbf{a} \in \mathcal{A} : a_k \neq 0\}$ 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 $\boldsymbol{\lambda},\boldsymbol{\gamma}$ but we drop this dependence in the notation for simplicity.

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 $\boldsymbol{\gamma}$ satisfies*

$$Reg_{\boldsymbol{\lambda},\boldsymbol{\gamma}}^{FP\text{-}CUCB}(n) \leq \sum_{k:\Delta_{min}^k>0} \frac{12K^2}{\gamma_{k,min}} \left[\frac{b(\Delta_{min}^k)}{\Delta_{min}^k} + \int_{\Delta_{min}^k}^{\Delta_{max}^k} \frac{b(x)}{x^2} dx \right] \log(n) + \left(\frac{\pi^2}{3} + 1 \right) K\Delta_{max}, \quad (4)$$

where $b(x) = \lambda_{max} + \sqrt{\lambda_{max}^2 + x^2/9K^2}$, and $\gamma_{k,min} = \min_{\mathbf{a}:a_k \neq 0} \gamma_k(\mathbf{a})$.

4.1.1 Proving Theorem 1

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 the observations 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 $\mathcal{G}_{k,B}$ 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 $\mathbf{g}_{k,B}^1, \mathbf{g}_{k,B}^2, ..., \mathbf{g}_{k,B}^{B_k}$ in increasing order of expected reward. We use the following notation for optimality gaps with respect to these ordered vectors

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

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 $k' = \arg\min_{j:g_{j,t}>0} N_{j,t-1}$, 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.

11

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:

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

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

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

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

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 $n > K$,*

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

The full proof of Proposition 1 is given in Appendix C, but in particular depends on the following Lemma describing the concentration of filtered Poisson data.

Lemma 1 *For any set of independent Poisson random variables $Y_1, ..., Y_s$ with means $\gamma_1\mu, ...\gamma_s\mu$ respectively, and parameters $t \geq s$ and $\mu_{max} \geq \mu$ the following holds:*

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

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. We provide a proof of Lemma 1 in Appendix B.

For the under sufficiency counter we have the following proposition, also proved in Appendix C,

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

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

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

$$
\begin{aligned}
Reg_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(n) &\leq \mathbb{E}\bigg(\sum_{k=1}^{K}\Big(\Delta^{k,1} + \sum_{l=1}^{B_k}(N_{k,n}^{l,suf} + N_{k,n}^{l,und})\Delta^{k,l}\Big)\bigg) \\
&= \mathbb{E}\bigg(\sum_{k=1}^{K}\Big(\Delta^{k,1} + \sum_{l=1}^{B_k}N_{k,n}^{l,suf}\Delta^{k,l}\Big)\bigg) + \mathbb{E}\bigg(\sum_{k=1}^{K}\sum_{l=1}^{B_k}N_{k,n}^{l,und}\Delta^{k,l}\bigg) \\
&\leq K\Delta_{max} + \mathbb{E}\bigg(\sum_{k=1}^{K}\sum_{l=1}^{B_k}N_{k,n}^{l,suf}\Delta^{k,l}\bigg) + \sum_{k:\Delta_{min}^k>0}\bigg(h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \int_{\Delta^{k,B_k}}^{\Delta^{k,1}} h_{k,n}(x)dx\bigg) \\
&\leq \Big(\frac{\pi^2}{3}+1\Big)K\Delta_{max} + \sum_{k:\Delta_{min}^k>0}\bigg(h_{k,n}(\Delta_{min}^k)\Delta_{min}^k + \int_{\Delta_{min}^k}^{\Delta_{max}^k} h_{k,n}(x)dx\bigg) \\
&= \sum_{k:\Delta_{min}^k>0}\frac{12K^2}{\gamma_{k,min}}\bigg[\frac{b(\Delta_{min}^k)}{\Delta_{min}^k} + \int_{\Delta_{min}^k}^{\Delta_{max}^k}\frac{b(x)}{x^2}dx\bigg]\log(n) + \Big(\frac{\pi^2}{3}+1\Big)K\Delta_{max}. \quad \square
\end{aligned}
$$

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.

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 and Robbins, 1985) is one where

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

for every $\mathbf{g} : \Delta_{\mathbf{g}} > 0$ 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

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

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 configuration 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

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

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 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 $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ 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*

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

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

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

such that $\inf_{\boldsymbol{\theta} \in B(\boldsymbol{\lambda})} \sum_{\mathbf{g} \in S(\boldsymbol{\lambda})} d_{\mathbf{g}} \sum_{k=1}^{K} g_k kl(\lambda_k, \theta_k) \geq 1.$ $\tag{14}$

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 $c(\boldsymbol{\lambda})$ by following steps similar to those in the proof of Theorem 2 in Combes et al. (2015). However we omit this here because in the interests of succinctness as it is not an especially useful or elegant expression.

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 a UAV 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 decays 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

$$\gamma_k(\mathbf{a}) = \sum_{u=1}^{U} \phi_u(\mathbf{a})\omega_{ku}\mathbb{I}\{a_k = u\}, \quad k \in [K], \tag{15}$$

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 intrusions 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 it takes a some time for intruders to cross the perimeter.

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 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.

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 $\tau_{ku} = \omega_{k,u}\lambda_k$. $\mathcal{A}$ continues 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

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

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

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 2.

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*

$$Reg_{\boldsymbol{\lambda},\boldsymbol{\gamma}}^{FP\text{-}CUCB}(n) \leq \sum_{ku:\Delta_{min}^{ku}>0} \frac{12(KU)^2}{\phi_{ku,min}} \left[\frac{b'(\Delta_{min}^{ku})}{\Delta_{min}^{ku}} + \int_{\Delta_{min}^{ku}}^{\Delta_{max}^{ku}} \frac{b'(x)}{x^2}dx \right] \log(n) + \left(\frac{\pi^2}{3} + 1\right) KU\Delta_{max},$$

where $\tilde{b}(x) = \tau_{max} + \sqrt{\tau_{max}^2 + x^2/9(KU)^2}$, and $\phi_{ku,min} = \min_{\mathbf{a}:a_k=u} \phi_u(\mathbf{a})$.

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 $t = KU + 1, KU + 2, ...$

- Calculate indices

$$\bar{\tau}_{ku,t} = \frac{\sum_{s=1}^{t-1} Y_{ku,s}}{\sum_{s=1}^{t-1} \phi_{ku,s}} + \frac{2 \log(t)}{\sum_{s=1}^{t-1} \phi_{ku,s}} + \sqrt{\frac{6\tau_{max} \log(t)}{\sum_{s=1}^{t-1} \phi_{ku,s}}}, \ ku \in [K] \times [U]$$

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

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*

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

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 our 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.

We compare the FP-CUCB, Greedy and Thompson Sampling 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 (15) 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

---

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 $t = K + 1, K + 2, ...$

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

Algorithm 4 Thompson Sampling (TS)

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

- For each $k \in [K]$ sample $\tilde{\lambda}_{k,t}$ from a $\text{Gamma}(\alpha + \sum_{s=1}^{t-1} Y_{k,s}, \beta + \sum_{s=1}^{t-1} \gamma_k(\mathbf{a}_s))$ distribution

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

---

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 $\text{Uniform}(10, 20)$ distribution for $k \in [K]$. Baseline detection probabilities ω_{ku} are sampled from $\text{Beta}(u, 2)$ 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 $\text{Beta}(u + 2, 2)$ 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 $\text{Uniform}(90, 100)$ distribution for $k \in [K]$. Baseline detection probabilities ω_{ku} are sampled from a $\text{Beta}(30, 5)$ 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 $\text{Uniform}(0.4, 1)$ distribution for $k \in [K]$. Baseline detection probabilities ω_{ku} are sampled from a $\text{Beta}(1, 1)$ 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 our 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

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

We calculate this value for all algorithms, all 250 datasets and rounds $1 \le n \le 2000$. 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 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 it appears that with prior variance equal to 1 the performance can be almost as poor as in the greedy case but with larger prior variance TS can outperform the FP-CUCB algorithm.

We analyse these behaviours further in Figures 2 and 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 2 and for tests (iii) and (iv) in Figure 3. We omit the greedy algorithm's performance from these figures as the values are so large. In Appendix E 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 it is clear 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 λ_{max} used in the algorithm is only 10% of the true problem value. We do see the performance of FP-CUCB start to suffer when λ_{max} is chosen to be 1, but it is still much better than TS.
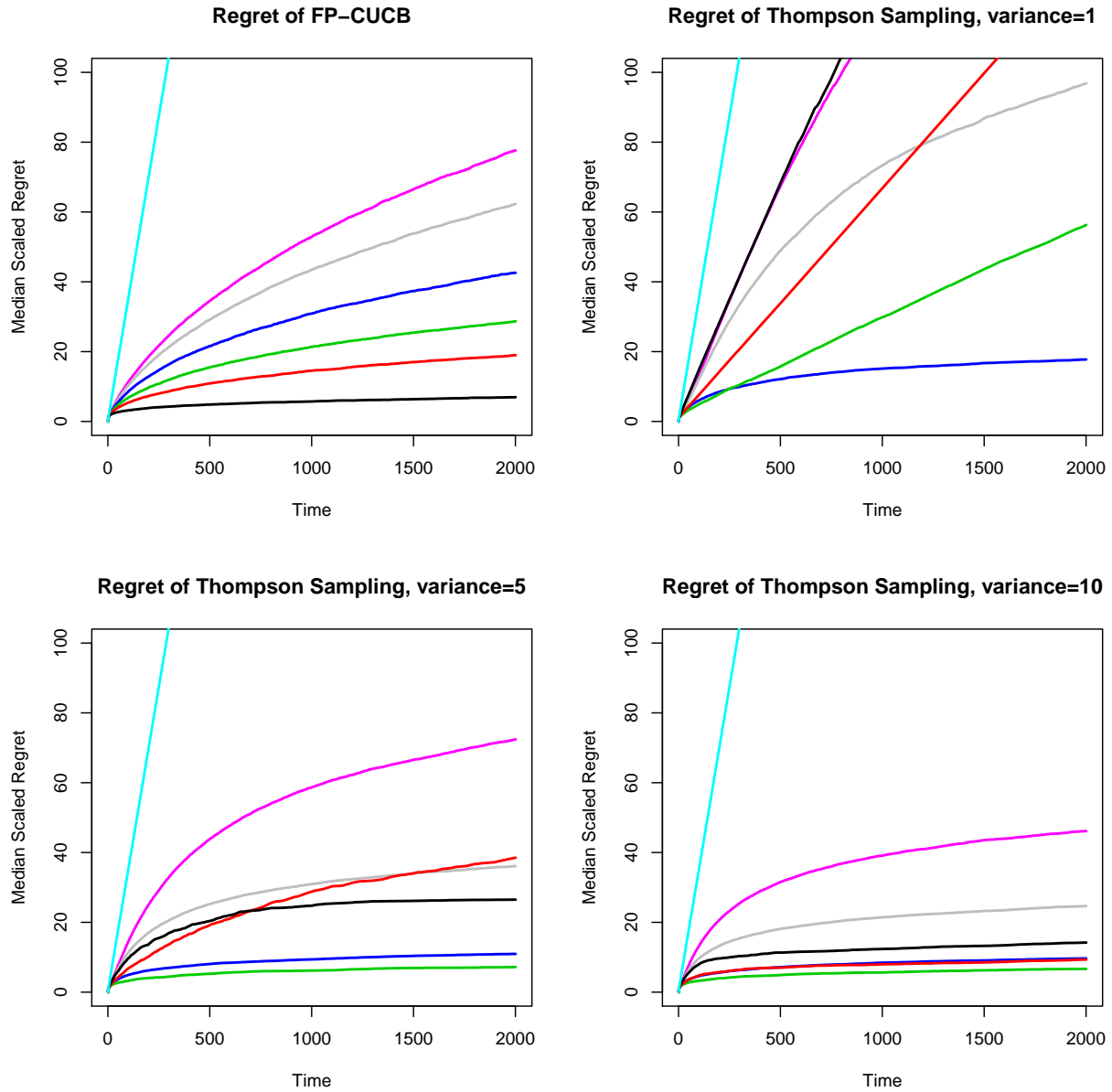
Figure 1: Cumulative Regret histories for Test (i). Upper left: FP-CUCB, upper right: TS with a prior variance of 1, lower left: TS with a prior variance of 5, lower right: TS with prior variance of 10. In each case the plotted lines are the median values of scaled regret calculated at each time point 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.
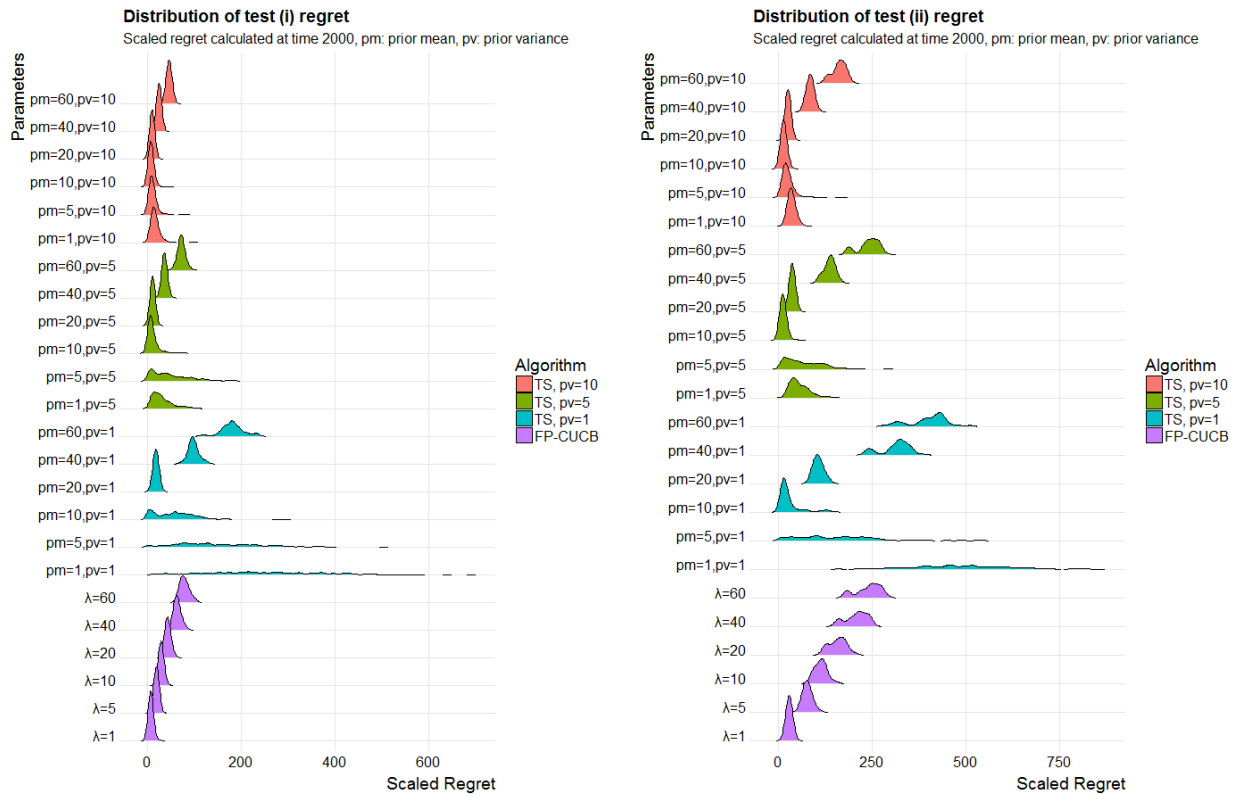
19

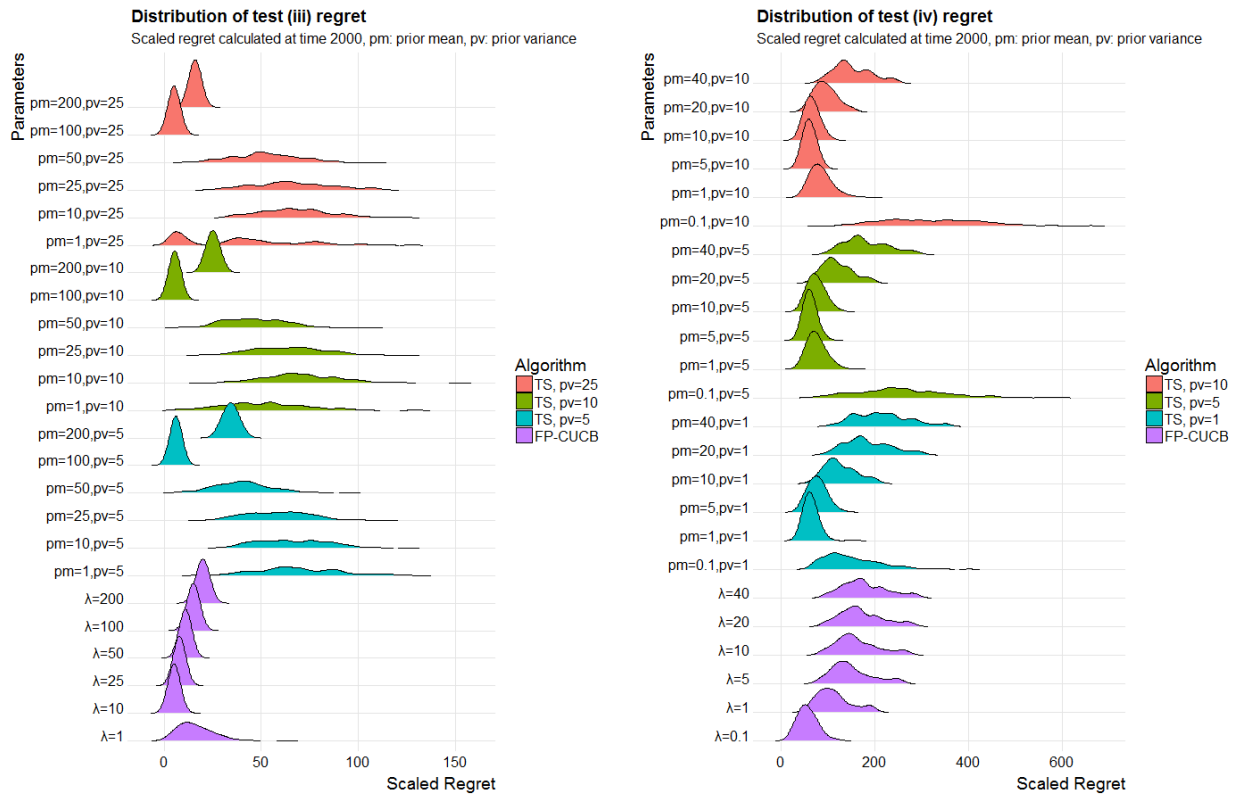Figure 2: Scaled regret distributions in tests (i) and (ii).

Figure 3: Scaled regret distributions in tests (iii) and (iv).

6 Discussion

In this paper we have modelled and derived solution methods for a family of sequential decision problems motivated by multi-searcher perimeter surveillance scenarios with incomplete information. Namely, we have considered the problem of adaptively assigning multiple searchers to cells along a perimeter in order to detect the maximum number of intrusion events. 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 comparing our FP-CUCB algorithm to Thompson Sampling (TS) we see that its principal benefit lies in its reliability. 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 our algorithm still performs well even when a key assumption does not hold is likely due to the fact that Bernstein'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 our UCB algorithm we required a symmetric tail bound for which an inflation term giving the type of concentration in Lemma 1 could be identified. Chernoff bounds (see e.g. Boucheron et al. (2013)) are tighter but lack these properties. 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.

In general we note the performance of TS is much more variable than that of FP-CUCB, as shown through the lower and upper quantiles of scaled regret. This 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 concerning disadvantage of TS. The reduced variability and theoretical guarantees of our FP-CUCB method make it a more reliable option for real surveillance operations.

A   Proof of NP-hardness of the IP (2)

Theorem 3 *Integer Linear Programs of the following type are $\mathcal{NP}$-hard in the strong sense:*

$$\max_{a_{iju}, 1 \leq i \leq j \leq K, u \in [U]} \sum_{i=1}^{K} \sum_{j=i}^{K} \sum_{u=1}^{U} q_{iju} a_{iju}$$

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

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

$$a_{iju} \in \{0, 1\}, \ \ 1 \leq i \leq j \leq K, u \in [U].$$

Proof of Theorem 3:

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 $U = nt$ and $K = 3n$. For $k = 1, ..., 3n$, set $q_{iju} = w_k$ if $j - i = w_k$ and the 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. $\square$

B   Lemma 1 Proof: Concentration of filtered Poisson estimator

First consider the result from Chapter 2 of Boucheron et al. (2013) that *Bernstein's Inequality* Bernstein (1946) holds for Poisson random variables. That is to say that if Z follows a Poisson distribution with parameter λ then

$$\mathbb{P}\left(Z \geq \lambda + \epsilon \right) \leq \exp\left(\frac{-\epsilon^2}{2\lambda + \frac{2}{3}\epsilon} \right) \quad \text{for } \epsilon > 0, \tag{16}$$

$$\mathbb{P}\left(Z \leq \lambda - \epsilon \right) \leq \exp\left(\frac{-\epsilon^2}{2\lambda + \frac{2}{3}\epsilon} \right) \quad \text{for } 0 < \epsilon < \lambda.$$

We prooced to prove Lemma 1 by first bounding the probability in (10)

$$\mathbb{P}\left(\left|\frac{\sum_{j=1}^s Y_j}{\sum_{j=1}^s \gamma_j} - \mu\right| \geq \frac{2\log(t)}{\sum_{j=1}^s \gamma_j} + \sqrt{\frac{6\mu_{max}\log(t)}{\sum_{j=1}^s \gamma_j}}\right)$$

$$\leq \mathbb{P}\left(\frac{\sum_{j=1}^s Y_j}{\sum_{j=1}^s \gamma_j} \geq \mu + \frac{2\log(t)}{\sum_{j=1}^s \gamma_j} + \sqrt{\frac{6\mu_{max}\log(t)}{\sum_{j=1}^s \gamma_j}}\right) + \mathbb{P}\left(\frac{\sum_{j=1}^s Y_j}{\sum_{j=1}^s \gamma_j} \leq \mu - \frac{2\log(t)}{\sum_{j=1}^s \gamma_j} + \sqrt{\frac{6\mu_{max}\log(t)}{\sum_{j=1}^s \gamma_j}}\right)$$

$$= \mathbb{P}\left(\sum_{j=1}^s Y_j \geq \mu\sum_{j=1}^s \gamma_j + \left(2\log(t) + \sqrt{6\mu_{max}\log(t)\sum_{j=1}^s \gamma_j}\right)\right)$$

$$+ \mathbb{P}\left(\sum_{j=1}^s Y_j \leq \mu\sum_{j=1}^s \gamma_j - \left(2\log(t) + \sqrt{6\mu_{max}\log(t)\sum_{j=1}^s \gamma_j}\right)\right). \tag{17}$$

Since $Y_1, ..., Y_s$ are all Poisson distributions, their sum $\sum_{j=1}^s Y_j$ follows a Poisson distribution with parameter $\mu\sum_{j=1}^s \gamma_j$ and we can apply (16) to bound the first component of (17),

$$\mathbb{P}\left(\sum_{j=1}^s Y_j \geq \mu\sum_{j=1}^s \gamma_j + \left(2\log(t) + \sqrt{6\mu_{max}\log(t)\sum_{j=1}^s \gamma_j}\right)\right)$$

$$\leq \exp\left(\frac{-\left(2\log(t) + \sqrt{6\mu_{max}\log(t)\sum_{j=1}^s \gamma_j}\right)^2}{2\mu\sum_{j=1}^s \gamma_j + \frac{2}{3}\left(2\log(t) + \sqrt{6\mu_{max}\log(t)\sum_{j=1}^s \gamma_j}\right)}\right)$$

$$= \exp\left(-3\log(t)\frac{4\log(t) + 4\sqrt{6\mu_{max}\log(t)\sum_{j=1}^s \gamma_j} + 6\mu_{max}\sum_{j=1}^s \gamma_j}{4\log(t) + 2\sqrt{6\mu_{max}\log(t)\sum_{j=1}^s \gamma_j} + 6\mu\sum_{j=1}^s \gamma_j}\right) \leq t^{-3}.$$

The second probability in (17) can also be bounded by t^{-3} by applying Bernstein's Inequality, thus completing the proof. $\square$

C Theorem 1 Proof: Expected regret of FP-CUCB

To complete the proof of Theorem 1 provided in the main text, we separately prove Propositions 1 and 2.

Proof of Proposition 1:

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

$$\mathcal{N}_t = \left\{\left|\frac{\sum_{j=1}^{t-1} Y_{k,j}}{D_{k,t-1}} - \lambda_k\right| < \frac{2\log(t)}{D_{k,t-1}} + \sqrt{\frac{6\lambda_{max}\log(t)}{D_{k,t-1}}} \ \forall k \in [K]\right\}.$$

Define random variables $\Lambda_{k,t} = \frac{2\log(t)}{D_{k,t-1}} + \sqrt{\frac{6\lambda_{max}\log(t)}{D_{k,t-1}}}$ for $k \in [K]$ and $\Lambda_t = \max_{k:g_{k,t}>0}(\Lambda_{k,t})$. Define $\Lambda^{k,l} = \frac{2\log(t)}{\gamma_{k,min}h_{k,n}(\Delta^{k,l})} + \sqrt{\frac{6\lambda_{max}\log(t)}{\gamma_{k,min}h_{k,n}(\Delta^{k,l})}}$ for $l \in [B_k]$, $k \in [K]$, which are not random variables. By

these definitions and the definition of UCB indices $\bar{\lambda}_{k,t}$ we have the following properties.

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

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

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

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

$$\mathbf{g}_t^T \cdot \boldsymbol{\lambda} + 2K\Lambda^{k,l} > \mathbf{g}_t^T \cdot \boldsymbol{\lambda} + 2K\Lambda_t \geq \mathbf{g}_t^T \cdot \bar{\boldsymbol{\lambda}}_t \geq (\mathbf{g}_{\boldsymbol{\lambda}}^*)^T \cdot \bar{\boldsymbol{\lambda}}_t \geq (\mathbf{g}_{\boldsymbol{\lambda}}^*)^T \cdot \boldsymbol{\lambda} = \mathrm{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} \qquad (18)$$

where $\mathbf{g}_{\boldsymbol{\lambda}}^*$ is an action that is optimal with respect to rate vector $\boldsymbol{\lambda}$. However, by definition $2K\Lambda^{k,l} \geq \Delta^{k,l}$ and therefore (18) is a contradiction of the definition of $\Delta^{k,l} = \mathrm{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}} - \mathbf{g}_{k,B}^l \cdot \boldsymbol{\lambda}$. 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 1 since

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

$$\leq \sum_{t=K+1}^{n} 2Kt^{-2} \leq \frac{\pi^2}{3} \cdot K. \quad \square$$

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 $\{j \in [K] | \Delta_{min}^j > 0\}$

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

$$= \sum_{t=K+1}^{n} \sum_{l=1}^{B_k} \sum_{j=1}^{l} \mathbb{I}\left\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \left(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j})\right)\right\} \Delta^{k,l}$$

$$\leq \sum_{t=K+1}^{n} \sum_{l=1}^{B_k} \sum_{j=1}^{l} \mathbb{I}\left\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \left(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j})\right)\right\} \Delta^{k,j}$$

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

$$\leq \sum_{t=K+1}^{n} \sum_{l=1}^{B_k} \sum_{j=1}^{B_k} \mathbb{I}\left\{\mathbf{g}_t = \mathbf{g}_{k,B}^l, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \left(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j})\right)\right\}\Delta^{k,j}$$

$$= \sum_{t=K+1}^{n} \sum_{j=1}^{B_k} \mathbb{I}\left\{\mathbf{g}_t \in \mathcal{G}_{k,B}, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \left(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j})\right)\right\}\Delta^{k,j}$$

$$= \sum_{j=1}^{B_k} \sum_{t=K+1}^{n} \mathbb{I}\left\{\mathbf{g}_t \in \mathcal{G}_{k,B}, N_{k,t} > N_{k,t-1}, N_{k,t-1} \in \left(h_{k,n}(\Delta^{k,j-1}), h_{k,n}(\Delta^{k,j})\right)\right\}\Delta^{k,j}$$

$$\leq \sum_{j=1}^{B_k} \left(h_{k,n}(\Delta^{k,j}) - h_{k,n}(\Delta^{k,j-1})\right)\Delta^{k,j}$$

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

$$= h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \sum_{j=1}^{B_k-1} h_{k,n}(\Delta^{k,j}) \cdot (\Delta^{k,j} - \Delta^{k,j+1})$$

$$\leq h_{k,n}(\Delta^{k,B_k})\Delta^{k,B_k} + \int_{\Delta^{k,B_k}}^{\Delta^{k,1}} h_{k,n}(x)dx.$$

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

D   Theorem 2 Proof: Lower bound on regret

To prove Theorem 2, 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 $\mathbf{g} \in \mathcal{G}$ then has an associated set of $\boldsymbol{\lambda}$ parameter vectors under which it is an optimal control $\Lambda_{\mathbf{g}} = \{\boldsymbol{\lambda} \in \mathbb{R}_+^K : \mathbf{g}^T \cdot \boldsymbol{\lambda} = \mathrm{opt}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}\}$, 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 $\mathbf{g} \in \mathcal{G}$:

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

With these quantities and those defined in Section 4.1.2 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. $\square$

E   Numerical Results

Algorithm	Parameters	0.025 Quantile	Median	0.975 Quantile
FP-CUCB	$\lambda_{max} = 1$	4.10	6.94	14.08
	$\lambda_{max} = 5$	12.72	18.97	24.46
	$\lambda_{max} = 10$	22.35	28.67	36.30
	$\lambda_{max} = 20$	35.72	42.59	54.98
	$\lambda_{max} = 40$	53.26	62.31	79.86
	$\lambda_{max} = 60$	64.06	77.63	100.34
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 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
	$\lambda_{max} = 1$	18.13	27.63	39.34
	$\lambda_{max} = 5$	57.39	76.02	104.67
FP-CUCB	$\lambda_{max} = 10$	83.99	113.86	150.29
	$\lambda_{max} = 20$	116.59	159.85	200.11
	$\lambda_{max} = 40$	152.79	212.42	248.46
	$\lambda_{max} = 60$	176.58	245.46	285.71
	Mean=1, Variance=1	248.37	481.61	783.17
	Mean=5, Variance=1	6.34	150.56	451.04
	Mean=10, Variance=1	6.38	16.93	129.87
	Mean=20, Variance=1	84.59	104.91	134.17
	Mean=40, Variance=1	234.03	321.99	365.10
	Mean=60, Variance=1	297.34	412.77	481.44
	Mean=1, Variance=5	22.56	48.46	121.32
	Mean=5, Variance=5	7.75	54.94	175.94
Thompmson Sampling	Mean=10, Variance=5	5.26	11.12	32.86
	Mean=20, Variance=5	29.47	36.23	45.42
	Mean=40, Variance=5	106.17	137.60	162.94
	Mean=60, Variance=5	183.13	245.48	279.97
	Mean=1, Variance=10	20.34	33.24	55.34
	Mean=5, Variance=10	9.44	21.12	79.06
	Mean=10, Variance=10	8.24	12.40	25.07
	Mean=20, Variance=10	19.19	24.56	31.66
	Mean=40, Variance=10	69.19	84.84	100.94
	Mean=60, Variance=10	126.94	162.99	188.46
Greedy		243.30	576.35	963.96

Table 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
	$\lambda_{max} = 1$	2.82	14.72	37.18
	$\lambda_{max} = 10$	3.97	5.14	6.73
FP-CUCB	$\lambda_{max} = 25$	6.38	7.86	9.44
	$\lambda_{max} = 50$	9.42	10.94	12.51
	$\lambda_{max} = 100$	12.60	15.15	17.45
	$\lambda_{max} = 200$	16.59	20.23	22.85
	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
Thompson Sampling	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 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	$\lambda_{max} = 0.1$	22.71	53.35	101.82
	$\lambda_{max} = 1$	59.48	106.59	193.57
	$\lambda_{max} = 5$	84.44	140.40	250.38
	$\lambda_{max} = 10$	91.32	151.69	263.00
	$\lambda_{max} = 20$	96.01	162.47	270.39
	$\lambda_{max} = 40$	98.81	172.28	284.63
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: Quantiles of scaled regret at horizon $n = 2000$ for algorithms applied to Test (iv) data

References

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

Anantharam, V., Varaiya, P., and 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., and Fischer, P. (2002). Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256.

Bernstein, S. (1946). The theory of probabilities.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University Press.

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

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

Burnetas, A. N. and 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., and 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., and Devulapalli, R. (2016). Shadow Prices in Territory Division. *Networks and Spatial Economics*, 16(3):893–931.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016a). Combinatorial multi-armed bandit with general reward functions. In *Advances in Neural Information Processing Systems*, pages 1651–1659.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159.

Chen, W., Wang, Y., Yuan, Y., and 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., and Lelarge, M. (2015). Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124.

Cowan, W., Honda, J., and Katehakis, M. N. (2015). Normal bandits of unknown means and variances: Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. *arXiv preprint arXiv:1504.05823*.

Gai, Y., Krishnamachari, B., and 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.

Garey, M. R. and Johnson, D. S. (1979). Computers and intractability: a guide to np-completeness.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons.

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

Kaufmann, E. (2016). On Bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*.

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

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

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. (2015a). Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015b). Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543.

Lai, T. L. and 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 *Advances in Neural Information Processing Systems*, pages 1583–1592.

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

Papadaki, K., Alpern, S., Lidbetter, T., and Morton, A. (2016). Patrolling a border. *Operations Research*, 64(6):1256–1269.

Park, M., Kalyanam, K., Darbha, S., Khargonekar, P. P., Pachter, M., and Chandler, P. R. (2016). Performance guarantee of an approximate dynamic programming policy for robotic surveillance. *IEEE Transactions on Automation Science and Engineering*, 13(2):564–578.

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

Stone, L. D. (1976). *Theory of Optimal Search*, volume 118. Elsevier.

Szechtman, R., Kress, M., Lin, K., and 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.

Washburn, A. R. (2002). *Search and Detection*. INFORMS.