Algorithms for Persistent Autonomy and Surveillance

by

Cenk Baykal

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author .....
Department of Electrical Engineering and Computer Science
May 18, 2017

Certified by .....
Daniela Rus
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Theses

Algorithms for Persistent Autonomy and Surveillance

by

Cenk Baykal

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

In this thesis, we consider the problem of monitoring stochastic, time-varying events occurring at discrete locations. Our problem formulation extends prior work in persistent surveillance by considering the objective of successfully completing monitoring tasks in unknown, dynamic environments where the rates of events are time-inhomogeneous and may be subject to abrupt changes. We propose novel monitoring algorithms that effectively strike a balance between exploration and exploitation as well as a balance between remembering and discarding information to handle temporal variations in unknown environments. We present analysis proving the favorable properties of the policies generated by our algorithms and present simulation results demonstrating their effectiveness in several monitoring scenarios inspired by real-world applications. Our theoretical and empirical results support the applicability of our algorithm to a wide range of monitoring applications, such as detection and tracking efforts at a large scale.

Thesis Supervisor: Daniela Rus
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I am grateful for the people who have encouraged and supported me throughout my research and my first years in graduate school. First and foremost, I would like to thank my advisor, Daniela Rus, for all her invaluable guidance and encouragement. Daniela has been an enthusiastic and remarkable advisor who has used every opportunity to teach me how to be a better researcher and help me grow as a person.

I would also like to thank my collaborators Guy Rosman, Sebastian Claici, Mark Donahue, and Kyle Kotowick for fruitful discussions and their invaluable contributions, without which this thesis would not have been possible. They have generously provided their time and expertise during inspiring discussions that culminated in many of the ideas in this thesis.

Most importantly, I would like to thank my friends and family for their unconditional support and understanding.

Contents

List of Figures

Chapter 1

Introduction

Persistent surveillance tasks often require the agent to monitor stochastic events of interest in unknown, dynamic environments over long periods of time in an autonomous manner. Equipped with limited a priori knowledge, uncertainty over time-varying event statistics necessitates the robot to travel from one landmark to another, identify the regions of importance, and adapt to the temporal variations in the environment. The overarching objective is to maximize the number of events observed in order to enable efficient data collection, which may be imperative for a successful surveillance mission. Applications include monitoring of wildlife, natural phenomena (e.g., floods, volcanic eruptions), urban events (see Fig. 1-1), and friendly and unfriendly activities.

In this thesis, we consider a novel persistent surveillance problem in which a mobile robot is tasked with monitoring transient events that occur in discrete, spatially-distributed landmarks according to station-specific Poisson processes with unknown, time-varying statistics. We assume that the monitoring task is conducted by a single robot equipped with a limited-range sensor that can only record measurements when the robot is stationary at a location, i.e., it cannot make measurements while traveling. Hence, the robot must travel to each location and wait for transient events to occur for an appropriately generated amount of time before traveling to another location. The persistent surveillance problem is to generate an optimal sequence of location-time pairs that optimizes the monitoring objectives at hand, e.g., maximizes the number of events observed in a balanced manner across all regions of interest.

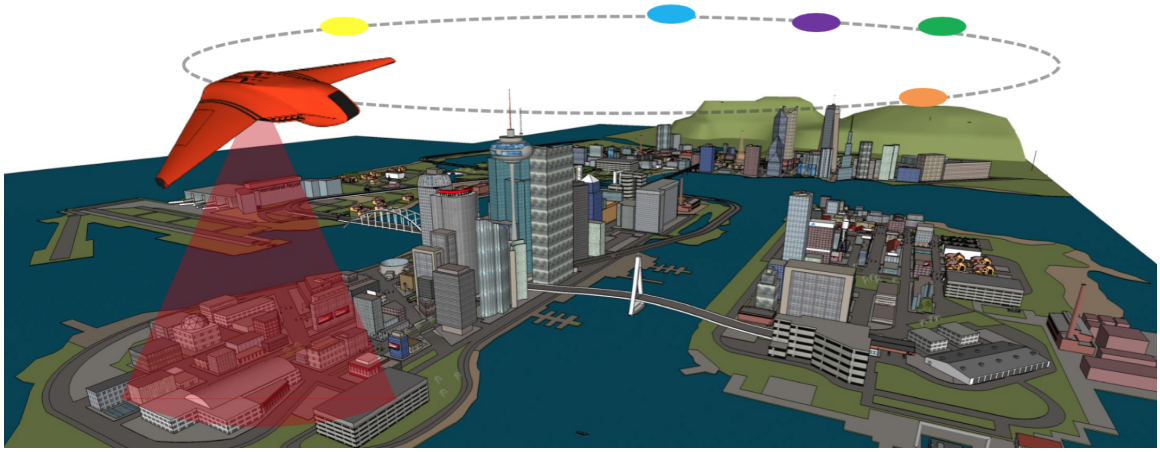


Figure 1-1: An example application of a monitoring procedure where a UAV is tasked with monitoring urban events. The colored nodes denote the discrete, spatially-distributed landmarks where pertinent events occur. The monitoring objective is to optimize the monitoring objectives, e.g., maximize the number of event sightings, in an efficient manner [38].

1.1 Contributions

This thesis contributes the following:

1. Persistent surveillance problem formulations that relax the assumptions of (i) known event statistics and (ii) static event statistics, commonly imposed by prior work.

2. A persistent surveillance problem formulation that bridges the monitoring objective of maximizing event observations with the objective of minimizing regret by introducing a new definition of weak regret for persistent surveillance.

3. A novel monitoring algorithm for generating appropriate policies to monitor transient events in unknown, dynamic environments where the total variation over time is bounded by a variation budget V_T that is known a priori.

4. An analysis proving that under the assumption that the total variation of the event rates is bounded by a variation budget $V_T = o(T)$, our algorithm generates long-run average optimal policies.

5. Simulation results that characterize our algorithm's effectiveness in minimizing

regret (i.e., maximizing the number of event observations) in several dynamic and random environments and comparing its performance to adaptive monitoring algorithms.

1.2 Outline of Thesis

The outline of this thesis is as follows. In Chapter 2, we present related work in persistent surveillance and discuss the limitations of current state-of-the-art monitoring approaches. In Chapter 3, we present the multi-objective persistent monitoring problem formulation that relaxes the assumption of known event statistics in otherwise static environments. We formulate and analyze a monitoring algorithm capable of balancing exploration and exploitation in the environment when the agent is initially equipped with limited a priori knowledge.

In Chapter 4, we further extend the canonical persistent surveillance problem formulation by relaxing the assumption of static event statistics and consider monitoring events in unknown and dynamic environments. We present and analyze a Multi-armed Bandits inspired approach that overcomes the limitations of prior approaches by generating policies in consideration of temporal variations and the exploration and exploitation trade-off. We conclude with a discussion of the presented approaches and their limitations, and provide suggestions for avenues of future research in Chapter 5,

Chapter 2

Related Work

Our work leverages and builds upon prior work in persistent surveillance, mobile sensor scheduling, and stochastic optimization.

2.1 Persistent Surveillance

The problem of persistent surveillance has been studied in a variety of real-world inspired monitoring applications such as underwater marine monitoring and detection of natural phenomena [6,11,21,22,25,26,30,32–34,36]. These approaches generally assume that the robot can obtain measurements while moving and generate paths that optimize an application-specific monitoring objective, such as mutual information. Further examples of monitoring objectives include facilitating high-value data collection for autonomous underwater vehicles [32], keeping a growing spatio-temporal field bounded using speed controllers [34], and generating the shortest watchman routes along which every point in a given space is visible [12].

Other related work in persistent surveillance includes variants and applications of the Orienteering Problem (OP) to generate informative paths that are constrained to a fixed length or time budget [18]. Yu et al. present an extension of OP to monitor spatially-correlated locations within a predetermined time [40]. In [15] and [39] the authors consider the OP problem in which the reward accumulated is characterized by a known function of the time spent at each point of interest. In contrast to

our work, approaches in OP predominantly consider known, static environments and budget-constrained policies that visit each location at most once.

Surveillance of discrete landmarks is of particular relevance to our work. Monitoring discrete locations such as buildings, windows, doors using a team of autonomous micro-aerial vehicles (MAVs) is considered in [26]. [1] presents different approaches to the min-max latency walk problem in a discrete setting. [38] extends this work to include multiple objectives, i.e. [38] considers the objective of minimizing the maximum latency and maximizing balance of events across stations using a single mobile robot. The authors show a reduction of the optimization problem to a quasi-convex problem and prove that a globally optimal solution can be computed in $O(\text{poly}(n))$ time where n is the number of discrete landmarks. Persistent surveillance in a discrete setting can be extended to the case of reasoning over different trajectories as shown in [30, 34, 35]. However, most prior work assumes that the rates of events are known prior to the surveillance mission, which is very often not the case in real world robotics applications. In this thesis, we relax the assumption of known rates and present an algorithm with provable guarantees to generate policies conducive to learning event rates and optimizing the monitoring objectives.

2.2 Mobile Sensor Scheduling and Coverage

The problem of persistent surveillance can be formulated as a mobile sensor scheduling problem and has been studied extensively in this context [16,19,20,27]. Persistent surveillance is closely related to sensor scheduling [19], sensor positioning [20], and coverage [16]. Previous approaches have considered persistent monitoring in the context of a mobile sensor [27].

Mobile sensor scheduling in environments with discrete landmarks are of particular relevance to our work. For instance, [26] considers monitoring discrete locations such as buildings, windows, and doors using a team of autonomous micro-aerial vehicles (MAVs). In [1], the authors present an approach to the min-max latency walk problem and [38] extends this work to the multi-objective mobile sensor scheduling problem for

surveillance of transient events occurring in discrete locations in the environment with known event rates and proposes an algorithm for generating the unique optimal policy maximizing the balance of observations while minimizing latency of observations at each station.

Yu et al. present an extension of OP to monitor spatially-correlated locations within a predetermined time [40]. In [15] and [39] the authors consider the OP problem in which the reward is a known function of the time spent at each point of interest. In contrast to our work, approaches in OP predominantly consider known environments and budget-constrained policies that visit each location at most once and optimize only a single objective.

[11] considers controlling multiple agents to minimize an uncertainty metric in the context of a 1D spatial domain. Decentralized approaches to controlling a network of robots for purposes of sensory coverage are investigated in [30], where a control law to drive a network of mobile robots to an optimal sensing configuration is presented. Persistent monitoring of dynamic environments has studied in [25,33,34]. For instance, [25] considers optimal sensing in a time-changing Gaussian Random Field and proposes a new randomized path planning algorithm to find the optimal infinite horizon trajectory. [10] presents a surveillance method based on Partially Observable Markov Decision Processes (POMDPs), however, POMDP-based approaches are often computationally intractable, especially when the action set includes continuous parameters, as in our case.

2.3 Multi-armed Bandits

As exemplified by the aforementioned works, a variety of monitoring algorithms have been presented and shown to perform well empirically. However, literature on methods with theoretical performance guarantees in unknown, time-varying environments has been sparse and limited. The added complexity stems from the inherent exploration and exploitation trade-off, which has been rigorously addressed and analyzed in Reinforcement Learning [23, 24, 28] and the Multi-armed Bandit (MAB) litera-

ture [3,4,9]. However, the traditional MAB problem considers minimizing regret with respect to the accumulated reward by appropriately pulling one of the $K \in \mathbb{N}_+$ levers at each discrete time step to obtain a stochastic reward that is generally assumed to bounded or subgaussian. Our work differs from the traditional MAB formulation in that we consider optimization in the face of travel costs, non-stationary processes, distributions with infinite support, and continuous state and parameter space. To the best of our knowledge, this thesis presents the first treatment of a MAB variant exhibiting all of the aforementioned complexities and a monitoring algorithm with provable regret guarantees with respect to the number of events observed.

MAB formulations that relax the assumptions of the traditional MAB problem are of particular pertinence to our work. Besbes et al. present a non-stationary MAB formulation where the variation of the rewards are bounded by a variation budget V_T and present policies that achieve a regret of order $(KV_T)^{1/3}T^{2/3}$, which is long-run average optimal if the variation V_T is sub-linear with respect to the time horizon T [5]. The authors mathematically show the difficulty of this problem by proving the lower bound of $\Omega((KV_T)^{1/3}T^{2/3})$, which implies that long-run average optimality is not achievable whenever V_T is linear in T.

Garivier et al. consider a non-stationary MAB setting where the distributions of the rewards change abruptly at unknown time instants, but the number of changes up to time T, Υ_T, is bounded and known in advance [17]. The authors present discounted and sliding window variants of the Upper Confidence Bound (UCB) algorithm [3, 4] that achieve a regret of $O(\sqrt{T\Upsilon_T \log T})$ and also prove the lower-bound of $\Omega(\sqrt{T\Upsilon_T})$ on the achievable regret in this setting, which is linear if the number of abrupt changes grows linearly with time. Prior work on the MAB formulation with switching costs tells a similar story regarding the difficulty of the aforementioned MAB extensions: Dekel et al. prove the lower-bound of $\tilde{\Omega}(T^{2/3})$ on the achievable regret in the presence of switching costs [14].

Recently, Srivastava et al. presented an approach with a provable upper bound on the number of visits to sub-optimal regions that bridges surveillance and MAB for monitoring phenomena in an unknown, abruptly changing environment [37]. However,

their approach considers a discrete state and parameter space (i.e., the generation of observation times is not considered), assumes Gaussian distributed random variables –which may be less suitable for monitoring instantaneous events (such as arrivals), assumes that the number of abrupt changes are bounded and known in advance, and does not explicitly take travel cost into consideration.

We build upon prior work and consider an unknown, dynamic environment where the robot is tasked with visiting each location more than once, observing stochastic, instantaneous events for an appropriately generated time, and adapting to the temporal variations in the environment over an unbounded amount of time. Unlike prior work in persistent surveillance which has focused on environments with a bounded number of abrupt changes, our problem formulation extends to continuously-varying as well as to abruptly-changing environments, as long as the total variation is bounded [5]. We introduce novel monitoring algorithms with provable guarantees with respect to the number of event sightings and present simulation results of real-world inspired monitoring scenarios that support our theoretical claims.

Chapter 3

Unknown, Static Environments

We begin by relaxing the assumption of known event statistics and consider the problem of persistent surveillance when the agent is equipped with limited a priori knowledge about the environment.

3.1 Problem Definition

Let there be $n \in \mathbb{N}_+$ spatially-distributed stations in the environment whose locations are known. At each station $i \in [n]$, stochastic events of interest occur according to a Poisson process with an unknown, station-specific rate parameter λ_i that is independent of other stations' rates. We assume that the robot executes a given cyclic path, taking $d_{i,j} > 0$ time to travel from station i to station j and let $D := \sum_{i=1}^{n-1} d_{i,i+1} + d_{n,1}$ denote the total travel time per cycle. The robot can only observe events at one station at any given time and cannot make observations while traveling.

We denote each complete traversal of the cyclic path as a *monitoring cycle*, indexed by $k \in \mathbb{N}_+$. We denote the observations times for all stations $\pi_k := (t_{1,k}, \ldots, t_{n,k})$ as the *monitoring policy* at cycle k. Our monitoring objective is to generate policies that maximize the number of events observed in a balanced manner across all stations within the allotted monitoring time $T_{\max}$ that is assumed to be unknown and unbounded. We introduce the function $f_{\text{obs}}(\Pi)$ that computes the total number of expected observations for a sequence of policies $\Pi := (\pi_k)_{k \in \mathbb{N}_+}$: $f_{\text{obs}}(\Pi) :=$

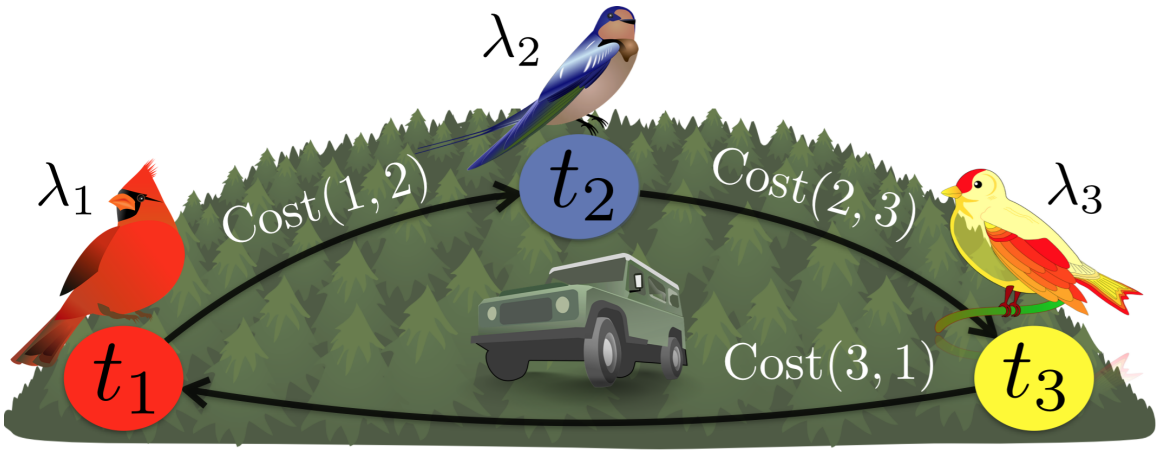


Figure 3-1: An example surveillance setting where the objective is to monitor three different species of birds that appear in discrete, species-specific locations. The overarching objective is to observe as many bird sightings as possible in a balanced way, so that an approximately uniform amount of data is collected across all bird species.

$\sum_{\pi_k} \sum_{i\in[n]} \mathbb{E}[N_i(\pi_k)]$, where $N_i(\pi_k)$ is the Poisson random variable, with realization $n_{i,k}$, denoting the number of events observed at station i under policy π_k and $\mathbb{E}[N_i(\pi_k)] := \lambda_i t_{i,k}$ by definition.

To reason about balanced attention, we let $f_{\text{bal}}(\Pi)$ denote as in [38] the expected observations ratio taken over the sequence of policies Π:

$$f_{\text{bal}}(\Pi) := \min_{i\in[n]} \frac{\sum_{\pi_k} \mathbb{E}[N_i(\pi_k)]}{\sum_{\pi_k} \sum_{j=1}^{n} \mathbb{E}[N_j(\pi_k)]}. \tag{3.1}$$

The *idealized* persistent surveillance problem is then:

Problem 1 (Idealized Persistent Surveillance Problem)**.** *Generate the optimal sequence of policies* $\Pi^* = \text{argmax}_{\Pi\in S} f_{obs}(\Pi)$ *where* S *is the set of all possible policies that can be executed within the allotted monitoring time* T_{max}*.*

Generating the optimal solution Π^* at the beginning of the monitoring process is challenging due to the lack of knowledge regarding both the upper bound $T_{\max}$ and the station-specific rates. Hence, instead of optimizing the entire sequence of policies at once, we take a greedy approach and opt to subdivide the problem into multiple, *per-cycle* optimization problems. For each cycle $k \in \mathbb{N}_+$, our goal is to adaptively generate the policy π_k^* that optimizes the monitoring objectives with respect to the

most up-to-date knowledge of event statistics. We let $\hat{f}_{\text{bal}}$ represent the per-cycle counterpart of f_{bal}

$$\hat{f}_{\text{bal}}(\pi_k) := \min_{i \in [n]} \frac{\mathbb{E}[N_i(\pi_k)]}{\sum_{j=1}^{n} \mathbb{E}[N_j(\pi_k)]}.$$

We note that the set of policies that optimize $\hat{f}_{\text{bal}}$ is uncountably infinite and policies of all possible lengths belong to this set [38]. To generate observation times that are conducive to exploration, we impose the hard constraint $t_{i,k} \geq t_{i,k}^{\text{low}}$ on each observation time, where $t_{i,k}^{\text{low}}$ is a lower bound that is a function of our uncertainty of the rate parameter λ_i (see Sec. 3.2). The optimization problem that we address in this thesis is then of the following form:

Problem 2 (Per-cycle Monitoring Optimization Problem)**.** *At each cycle $k \in \mathbb{N}_+$, generate a per-cycle optimal policy π_k^* satisfying*

$$\pi_k^* \in \operatorname*{argmax}_{\pi_k} \hat{f}_{bal}(\pi_k) \;\; s.t. \;\; \forall i \in [n] \;\; t_{i,k} \geq t_{i,k}^{low}. \tag{3.2}$$

3.2 Methods

In this section, we present our monitoring algorithm and detail the main subroutines employed by our method to generate dynamic, adaptive policies and interleave learning and approximating of event statistics with policy execution.

3.2.1 Algorithm for Monitoring Under Unknown Event Rates

The entirety of our persistent surveillance method appears as Alg. 1 and employs Alg. 2 as a subprocedure to generate adaptive, uncertainty-reducing policies for each monitoring cycle.

Algorithm 1: Core monitoring algorithm

$\alpha_i \leftarrow \alpha_{i,0}$;
$\beta_i \leftarrow \beta_{i,0}$;
$\hat{\lambda}_i \leftarrow \alpha_i/\beta_i$;
Loop
    $\pi^* \leftarrow$ Algorithm2(α_i, β_i);
    for $i \in [n]$ **do**
        Observe for t_i^* time to obtain n_i observations;
        $\alpha_i \leftarrow \alpha_i + n_i$,;
        $\beta_i \leftarrow \beta_i + t_i^*$;
        $\hat{\lambda}_i \leftarrow \alpha_i/\beta_i$;

Algorithm 2: Generates a per-cycle optimal policy π^*

for $i \in [n]$ **do**
    Compute t_i^{low} using (3.8);
$\pi_{\text{low}} \leftarrow (t_1^{\text{low}}, \ldots, t_n^{\text{low}})$;
$N_{\max} \leftarrow \max_{i\in[n]} t_i^{\text{low}} \alpha_i/\beta_i$;
for $i \in [n]$ **do**
    Compute t_i^* using $N_{\max}$ according to (3.10);
return $\pi^* = (t_1^*, \ldots, t_n^*)$;

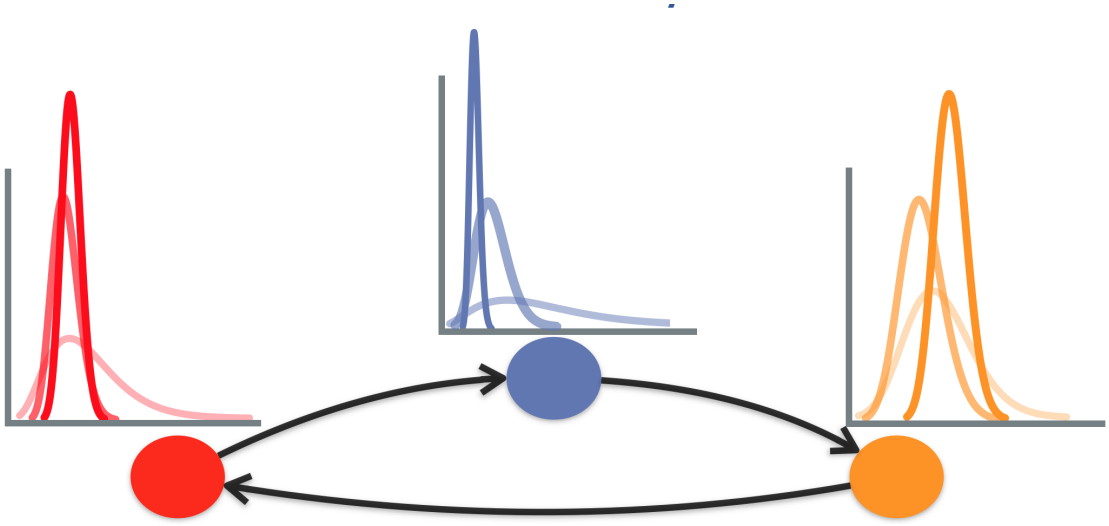


Figure 3-2: A depiction of the resulting distributions over event rates at each location over multiple cycles (faded colors) when the uncertainty constraint is incorporated into the optimization. The uncertainty constraint enables the uncertainty at each location to decay in a controlled and uniform way over multiple cycles.

3.2.2 Learning and Approximating Event Statistics

We use the Gamma distribution as the conjugate prior for each rate parameter because it provides a closed-form expression for updating the posterior distribution after observing events. We let Gamma(α_i, β_i) denote the Gamma distribution with hyper-parameters $\alpha_i, \beta_i \in \mathbb{R}_+$ that are initialized to user-specified values $\alpha_{i,0}, \beta_{i,0}$ for all stations i and are updated as new events are observed.

For any arbitrary number of events $n_{i,k} \in \mathbb{N}$ observed in $t_{i,k}$ time, the posterior distribution is given by Gamma$(\alpha_i + n_{i,k}, \beta_i + t_{i,k})$ for any arbitrary station $i \in [n]$ and cycle $k \in \mathbb{N}_+$. For notational convenience, we let $X_i^k := (n_{i,k}, t_{i,k})$ represent the summary of observations for cycle $k \in \mathbb{N}_+$ and define the aggregated set of observations up to any arbitrary cycle as $X_i^{1:k} := \{X_i^1, X_i^2, \ldots, X_i^k\}$ for all stations $i \in [n]$. After updating the posterior distribution using the hyper-parameters, i.e. $\alpha_i \leftarrow \alpha_i + n_{i,k}$, $\beta_i \leftarrow \beta_i + t_{i,k}$, we use the maximum probabiliy estimate of the rate parameter λ_i, denoted by $\hat{\lambda}_{i,k}$ for any arbitrary station i:

$$\hat{\lambda}_{i,k} := E[\lambda_i | X_i^{1:k}] = \frac{\alpha_{i,0} + \sum_{k=1}^{n} n_{i,k}}{\beta_{i,0} + \sum_{k=1}^{n} t_{i,k}} = \frac{\alpha_i}{\beta_i}. \tag{3.3}$$

3.2.3 Per-cycle Optimization and the Uncertainty Constraint

Inspired by confidence-based MAB approaches [2–4], our algorithm adaptively computes policies by reasoning about the uncertainty of our rate approximations. We introduce the *uncertainty-constraint*, an optimization constraint that enables the generating a station-specific observation time based on uncertainty of each station's parameter. The constraint helps bound the policy lengths adaptively over the course of the monitoring process so that approximation uncertainty decreases uniformly across all stations. We use the posterior variance of the rate parameter λ_i, $\text{Var}(\lambda_i | X_i^{1:k})$, as our uncertainty measure of each station i after executing k cycles. We note that in our Gamma-Poisson model, $\text{Var}(\lambda_i | X_i^{1:k}) := \frac{\alpha_i}{\beta_i^2}$ by definition of the Gamma distribution.

Uncertainty constraint For a given $\delta \in (0,1), \epsilon \in \left(0, 2(1+2e^{1/\pi})^{-1}\right)$ and arbitrary cycle $k \in \mathbb{N}_+$, π_k must satisfy the following

$$\forall i \in [n] \quad \mathbb{P}\left(\left(\mathrm{Var}(\lambda_i | X_i^{1:k}, \pi_k) \leq \delta \mathrm{Var}(\lambda_i | X_i^{1:k-1}) \middle| X_i^{1:k-1}\right)\right) > 1 - \epsilon. \tag{3.4}$$

We incorporate the uncertainty constraint as a hard constraint and recast the per-cycle optimization problem from Sec. 3.1 in terms of the optimization constraint.

Problem 3 (Recast Per-cycle Monitoring Optimization Problem)**.** *For each monitoring cycle $k \in \mathbb{N}_+$ generate a per-cycle optimal policy π_k^* that simultaneously satisfies the uncertainty constraint* (3.4) *and maximizes the balance of observations, i.e.,*

$$\begin{aligned} &\pi_k^* \in \underset{\pi_k}{\mathrm{argmax}}\, \hat{f}_{bal}(\pi_k) \\ &s.t. \;\; \forall i \in [n] \quad \mathbb{P}\left(\left(Var(\lambda_i | X_i^{1:k}, \pi_k) \leq \delta\, Var(\lambda_i | X_i^{1:k-1}) \middle| X_i^{1:k-1}\right)\right) > 1 - \epsilon. \end{aligned} \tag{3.5}$$

3.2.4 Controlling Approximation Uncertainty

We outline an efficient method for generating observation times that satisfy the uncertainty constraint and induce uncertainty reduction at each monitoring cycle. We begin by simplifying (3.4) to obtain

$$\mathbb{P}\left(\left(N_i(t_{i,k}) \leq \delta k(t_{i,k}) | X_i^{1:k-1}\right)\right) > 1 - \epsilon \tag{3.6}$$

where $N_i(t_{i,k}) \sim \mathrm{Pois}(\lambda_i t_{i,k})$ by definition of Poisson process and $k(t_{i,k}) := \delta\alpha_i(\beta_i + t_{i,k})^2/\beta_i^2 - \alpha_i$. Given that the distribution of the random variable $N_i(t_{i,k})$ is a function of the unknown parameter λ_i, we use interval estimation to reason about the cumulative probability distribution of $N_i(t_{i,k})$.

For each monitoring cycle $k \in \mathbb{N}_+$ we utilize previously obtained observations $X_i^{1:k-1}$ to construct the equal-tail credible interval for each parameter λ_i, $i \in [n]$ defined by the open set $(\lambda_i^{\mathrm{l}}, \lambda_i^{\mathrm{u}})$ such that

$$\forall \lambda_i \in \mathbb{R}_+ \quad \mathbb{P}\left((\lambda_i \in (\lambda_i^{\mathrm{l}}, \lambda_i^{\mathrm{u}}) | \, X_i^{1:k-1})\right) = 1 - \epsilon$$

where $\epsilon \in (0, 2(1+2e^{1/\pi})^{-1})$. By leveraging the relation between the Poisson and Gamma distributions, we compute the end-points of the equal-tailed credible interval:

$$\lambda_i^{\mathrm{l}} := \frac{Q^{-1}(\alpha_i, \frac{\epsilon}{2})}{\beta_i} \quad \lambda_i^{\mathrm{u}} := \frac{Q^{-1}(\beta_i, 1-\frac{\epsilon}{2})}{\beta_i}$$

where $Q^{-1}(a, s)$ is the Gamma quantile function and α_i and β_i are the posterior hyper-parameters after observations $X_i^{1:k-1}$. Given that we desire our algorithm to be *cycle-adaptive* (Sect. 3.1), we seek to generate the minimum feasible observation time satisfying the uncertainty constraint for each station $i \in [n]$, i.e.,

$$t_{i,k}^{\mathrm{low}} = \inf_{t_{i,k} \in \mathbb{R}_+} t_{i,k} \;\text{ s.t. }\; \mathbb{P}\left(\left(N_{i,k}(t_{i,k}) \leq \delta k(t_{i,k}) | X_i^{1:k-1}\right)\right) > 1-\epsilon. \tag{3.7}$$

For computational efficiency in the optimization above, we opt to use a tight and efficiently-computable lower bound for approximating the Poisson cumulative distribution function that improves upon the Chernoff-Hoeffding inequalities by a factor of at least two [31]. As demonstrated rigorously in Lemma 1, the expression for an approximately-minimal observation time satisfying constraint (3.4) is given by

$$t_{i,k}^{\mathrm{low}} := t \in \mathbb{R}_+ \mid D_{\mathrm{KL}}\big(\mathrm{Pois}(\lambda_i^{\mathrm{u}} t) \,||\, \mathrm{Pois}(k(t))\big) - W_\epsilon = 0 \tag{3.8}$$

where $D_{\mathrm{KL}}\big(\mathrm{Pois}(\lambda_1) \,||\, \mathrm{Pois}(\lambda_2)\big)$ is the Kullback-Leibler (KL) divergence between two Poisson distributions with mean λ_1 and λ_2 respectively and W_ϵ is defined using the Lambert W function [13]: $W_\epsilon = \frac{1}{2} W\big(\frac{(\epsilon-2)^2}{2\epsilon^2\pi}\big)$. An appropriate value for $t_{i,k}^{\mathrm{low}}$ can be obtained by invoking a root-finding algorithm such as Brent's method on the equation above [8].

The constant factor $\delta \in (0, 1)$ is the exploration parameter that influences the rate of uncertainty decay. Low values of δ lead to lengthy, and hence less cycle-adaptive policies, whereas high values lead to shorter, but also less efficient policies due to incurred travel time. We found that values generated by a logistic function with respect to problem-specific parameters as input worked well in practice for up to 50 stations: $\delta(n) := (1+\exp(-n/D))^{-1}$ where D is the total travel time per cycle.

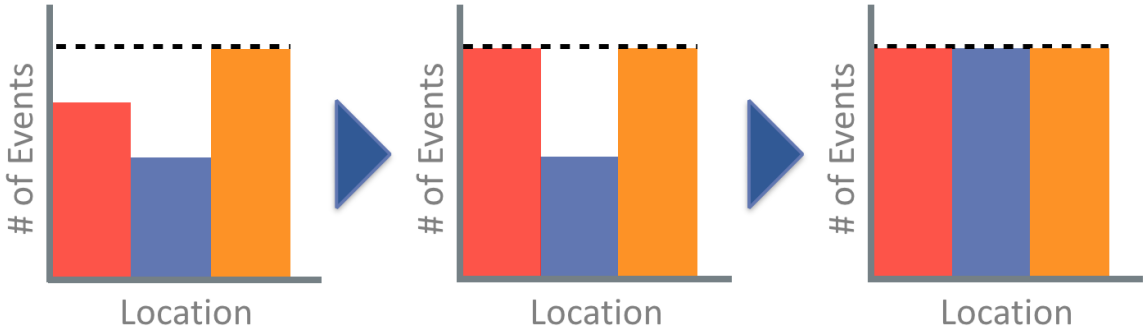


Figure 3-3: The water-filling algorithm described by Sec. 3.2.5 where the colors denote three different discrete locations. The observation times for stations below the bottle-neck expected number of events (dashed black line) is increased until the (approximate) expected number of observations at each station is equivalent.

3.2.5 Generating Balanced Policies that Consider Approximation Uncertainty

We build upon the method introduced in the previous section to generate a policy π_k^* that simultaneously satisfies the uncertainty constraint and balances attention given to all stations in approximately the minimum time possible. The key insight is that the value of $t_{i,k}^{\text{low}}$ given by (3.8) acts as a lower bound on the observation time for each station $i \in [n]$ for satisfying the uncertainty constraint (see Lemma 2). We also leverage the following fact from [38] regarding the optimality of the balance objective for a policy π_k:

$$\mathbb{E}[N_1(\pi_k)] = \cdots = \mathbb{E}[N_n(\pi_k)] \Leftrightarrow \pi_k \in \operatorname*{argmax}_{\pi} \hat{f}_{\text{bal}}(\pi). \tag{3.9}$$

We use a combination of this result and the fact that any observation time satisfying $t_{i,k} \geq t_{i,k}^{\text{low}}$ also satisfies the uncertainty constraint to arrive at an expression for the optimal observation time for each station. In constructing the optimal policy $\pi_k^* = (t_{1,k}^*, \ldots, t_{n,k}^*)$, we first identify the "bottleneck" value, $N_{\max}$, which is computed using the lower bounds for each $t_{i,k}$, i.e., $N_{\max} := \max_{i \in [n]} \hat{\lambda}_{i,k} t_{i,k}^{\text{low}}$. Given (3.9), we use the bottleneck value $N_{\max}$ to set the value of each observation time $t_{i,k}^*$ appropriately so that each $t_{i,k}^* \geq t_{i,k}^{\text{low}}$ and the policy defined by $\pi_k^* := (t_{1,k}^*, \ldots, t_{n,k}^*)$ maximizes

the balance objective function. Namely, the optimal observation times for all stations which constitute the per-cycle optimal policy $\pi_k^* = (t_{1,k}^*, \ldots, t_{n,k}^*)$ are computed individually:

$$\forall k \in \mathbb{N}_+ \quad \forall i \in [n] \quad t_{i,k}^* := \frac{N_{\max}}{\hat{\lambda}_{i,k}} = N_{\max} \frac{\beta_i}{\alpha_i}. \tag{3.10}$$

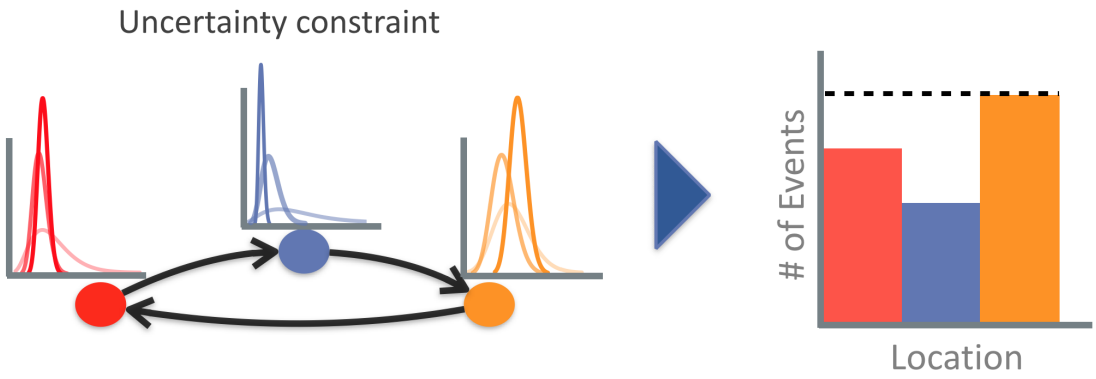


Figure 3-4: A unified visualization of the components of our algorithm. The lower bounds for the observation times are generated according to the uncertainty constraint in order to ensure controlled decay of uncertainty (Left). The lower bounds are then utilized by the water-filling algorithm detailed in Sec. 3.2.5 to generate policies that are balanced. The conjunction of these two components culminates in monitoring policies that are conducive to both exploration and exploitation.

3.3 Analysis

The outline of results in this section is as follows: we begin by proving the uncertainty-reducing property and per-cycle optimality of policies generated by Alg. 2 with respect to the rate approximations. We present a probabilistic bound on posterior variance and error of our rate approximations with respect to the ground-truth rates by leveraging the properties of each policy. We use the previous results to establish a probabilistic bound on the per-cycle optimality of any arbitrary policy generated by Alg. 2 with respect to the ground-truth optimal solution of Problem 3.

We impose the following assumption on user-specified input.

Assumption 1. *The parameters ϵ and δ are confined to the intervals $(0, 2(1 + 2e^{1/\pi})^{-1})$ and $(0, 1)$ respectively, i.e., $\epsilon \in (0, 2(1 + 2e^{1/\pi})^{-1})$, $\delta \in (0, 1)$.*

A policy π_k is said to be *approximately-optimal* at cycle $k \in \mathbb{N}_+$ if π_k is an optimal solution to Problem 3 with respect to the rate approximations $\hat{\lambda}_{1,k}, \ldots, \hat{\lambda}_{n,k}$, i.e., if it is optimal under the approximation of expectation: $\mathbb{E}[N_i(\pi_k)] \approx \hat{\lambda}_{i,k} t_{i,k}$ $\forall i \in [n]$. In contrast, a policy π_k is *ground-truth optimal* if it is an optimal solution to Problem 3 with respect to the ground-truth rates $\lambda_1, \ldots, \lambda_n$. For sake of notational brevity, we introduce the function $g : \mathbb{R} \to \mathbb{R}$ denoting

$$g(x) := 1 - \frac{e^{-x}}{\max\left\{2, 2\sqrt{\pi x}\right\}},$$

and note the bound established by [31] for a Poisson random variable Y with mean m and $k \in \mathbb{R}_+$ such that $k \geq m$

$$\mathbb{P}\left(\left(Y \leq k\right)\right) > g\big(D_{\mathrm{KL}}(\mathrm{Pois}(m) \,||\, \mathrm{Pois}(k))\big). \tag{3.11}$$

We begin by proving that each policy generated by Alg. 2 is optimal with respect to the per-cycle optimization problem (Problem 3).

Lemma 1 (Satisfaction of the uncertainty constraint)**.** The observation time $t_{i,k}^{\mathrm{low}}$ given by (3.8) satisfies the uncertainty constraint (3.4) for any arbitrary station $i \in [n]$ and iteration $k \in \mathbb{N}_+$.

Proof. We consider the left-hand side of (3.6) from Sect. 3.2 and marginalize over the unknown parameter $\lambda_i \in \mathbb{R}_+$:

$$\mathbb{P}\left((N_i(t_{i,k}) \leq k(t_{i,k}) | X_i^{1:k-1})\right) = \int_0^\infty \mathbb{P}\left((N_i(t_{i,k}) \leq k(t_{i,k}) | X_i^{1:k-1}, \lambda)\right) \mathbb{P}\left((\lambda | X_i^{1:k-1})\right) \, d\lambda$$

where the probability is with respect to the random variable $N_i(t_{i,k}) \sim \mathrm{Pois}(\lambda t_{i,k})$ $\forall \lambda \in \mathbb{R}_+$ by definition of a Poisson process with parameter λ. Using the equal-tails

credible interval constructed in Alg. 2, i.e. the interval $(\lambda_i^{\mathrm{l}}, \lambda_i^u)$ satisfying

$$\forall i \in [n] \;\; \forall \lambda_i \in \mathbb{R}_+ \; \mathbb{P}\left((\lambda_i^{\mathrm{l}} > \lambda_i | X_i^{1:k-1})\right) = \mathbb{P}\left((\lambda_i^{\mathrm{u}} < \lambda_i | X_i^{1:k-1})\right) = \frac{\epsilon}{2},$$

we establish the inequalities:

$$\begin{aligned}
\mathbb{P}\left((N_i(t_{i,k}) \leq k(t_{i,k}) | X_i^{1:k-1})\right) &> \int_0^{\lambda_i^u} \mathbb{P}\left((N_i(t_{i,k}) \leq k(t_{i,k}) | X_i^{1:k-1}, \lambda)\right) \mathbb{P}\left((\lambda | X_i^{1:k-1})\right) \, d\lambda \\
&\geq \mathbb{P}\left((N_i(t_{i,k}) \leq k(t_{i,k}) | X_i^{1:k-1}, \lambda_i^u)\right) \int_0^{\lambda_i^u} \mathbb{P}\left((\lambda | X_i^{1:k-1})\right) \, d\lambda \\
&= (1 - \frac{\epsilon}{2}) \, \mathbb{P}\left((N_i(t_{i,k}) \leq k(t_{i,k}) | X_i^{1:k-1}, \lambda_i^u)\right). \qquad (3.12)
\end{aligned}$$

where we utilized the fact that $\mathbb{P}\left(\left(N_i(t_{i,k}) \leq k(t_{i,k}) | X_i^{1:k-1}, \lambda_i^{\mathrm{u}}\right)\right)$ is monotonically decreasing with respect to λ. By construction, $t_{i,k}^{\mathrm{low}}$ satisfies
$D_{\mathrm{KL}}(\mathrm{Pois}(\lambda_i^{\mathrm{u}} t_{i,k}^{\mathrm{low}}) \,||\, \mathrm{Pois}(k(t_{i,k}^{\mathrm{low}}))) = W_\epsilon$ which yields $1 - g(W_\epsilon) = 1 - \frac{\epsilon}{2-\epsilon}$ by definition and thus by (3.11) we have:

$$\mathbb{P}\left((N_i(t_{i,k}^{\mathrm{low}}) \leq k(t_{i,k}^{\mathrm{low}}) | X_i^{1:k-1}, \lambda_i^{\mathrm{u}})\right) > 1 - g(W_\epsilon) = 1 - \frac{\epsilon}{2-\epsilon}.$$

Combining this inequality with the expression of (3.12) establishes the result. □

Lemma 2 (Monotonicity of solutions satisfying (3.4))**.** For any arbitrary station $i \in [n]$ and monitoring cycle $k \in \mathbb{N}_+$, the observation time $t_{i,k}$ satisfying $t_{i,k} \geq t_{i,k}^{\mathrm{low}}$, where $t_{i,k}^{\mathrm{low}}$ is given by (3.8), satisfies the uncertainty constraint.

Theorem 1 (Per-cycle approximate-optimality of solutions)**.** For any arbitrary cycle $k \in \mathbb{N}_+$, the policy $\pi_k^* := (t_{1,k}^*, \ldots, t_{n,k}^*)$ generated by Alg. 2 is an approximately-optimal solution with respect to Problem 3.

Proof. By definition of (3.10), we have for any arbitrary cycle $k \in \mathbb{N}_+$ and station $i \in [n]$, $t_{i,k}^* = N_{\max}/\hat{\lambda}_{i,k} \geq t_{i,k}^{\mathrm{low}}$ by definition of $N_{\max} := \max_{i \in [n]} \hat{\lambda}_{i,k} t_{i,k}^{\mathrm{low}}$. Applying Lemma 2 and observing that

$$\hat{\lambda}_{1,k} t_{1,k}^* = N_{\max}, \; \hat{\lambda}_{2,k} t_{2,k}^* = N_{\max}, \; \ldots, \hat{\lambda}_{n,k} t_{n,k}^* = N_{\max}$$

implies that the uncertainty constraint is satisfied for all stations $i \in [n]$ and that $\pi_k^* \in \operatorname{argmax}_{\pi_k} \hat{f}_{\text{bal}}(\pi_k)$, which establishes the optimality of π_k with respect to Problem 3. □

Using the fact that each policy satisfies the uncertainty constraint, we establish probabilistic bounds on uncertainty, i.e. posterior variance, and rate approximations.

Lemma 3 (Bound on posterior variance)**.** After executing an arbitrary number of cycles $k \in \mathbb{N}_+$, the posterior variance $\text{Var}(\lambda_i|X_i^{1:k})$ is bounded above by $\delta^k \text{Var}(\lambda_i)$ with probability at least $(1-\epsilon)^k$, i.e.,

$$\forall i \in [n] \;\; \forall k \in \mathbb{N}_+ \;\; \mathbb{P}\left(\left(\text{Var}(\lambda_i|X_i^{1:k}) \leq \delta^k \text{Var}(\lambda_i)|X_i^{1:k}\right)\right) > (1-\epsilon)^k$$

for all stations $i \in [n]$ where $\text{Var}(\lambda_i) := \alpha_{i,0}/\beta_{i,0}^2$ is the prior variance.

Proof. Iterative application of the inequality $\text{Var}(\lambda_i|X_i^{1:k}) \leq \delta \text{Var}(\lambda_i|X_i^{1:k-1})$ each with probability $1-\epsilon$ by the uncertainty constraint (3.4) yields the result. □

Corollary 1 (Bound on variance of the posterior mean)**.** After executing an arbitrary number of cycles $k \in \mathbb{N}_+$, the variance of our approximation $\text{Var}\big(\hat{\lambda}_{i,k}|X_i^{1:k-1}\big)$ is bounded above by $\delta^{k-1}\text{Var}(\lambda_i)$ with probability greater than $(1-\epsilon)^{k-1}$, i.e.,

$$\forall i \in [n] \;\; \mathbb{P}\left(\left(\text{Var}(\hat{\lambda}_{i,k}|X_i^{1:k-1}) \leq \delta^{k-1}\text{Var}(\lambda_i)|X_i^{1:k-1}\right)\right) > (1-\epsilon)^{k-1}.$$

Proof. Application of the law of total conditional variance and invoking Lemma 3 yields the result. □

Theorem 2 (ξ-bound on approximation error)**.** For all $\xi \in \mathbb{R}_+$ and cycles $k \in \mathbb{N}_+$, the inequality $|\hat{\lambda}_{i,k} - \lambda_i| < \xi$ holds with probability at least $(1-\epsilon)^{k-1}(1 - \frac{\delta^{k-1}\text{Var}(\lambda_i)}{\xi^2})$, i.e.,

$$\forall i \in [n] \;\; \mathbb{P}\left(\left(|\hat{\lambda}_{i,k} - \lambda_i| < \xi|X_i^{1:k-1}\right)\right) > (1-\epsilon)^{k-1}\big(1 - \frac{\delta^{k-1}\text{Var}(\lambda_i)}{\xi^2}\big).$$

Proof. Applying Corollary 1 and using Chebyshev's inequality gives the result. □

Theorem 3 (Δ-bound on optimality with respect to Problem 3)**.** For any $\xi_i \in \mathbb{R}_+$, $i \in [n], k \in \mathbb{N}_+$, given that $|\hat{\lambda}_{i,k} - \lambda_i| \in (0, \xi_i)$ with probability as given in Theorem 2, let $\sigma_{\min} := \sum_{i=1}^{n} (\lambda_i - \xi_i)^{-1}$ and $\sigma_{\max} := \sum_{i=1}^{n} (\lambda_i + \xi_i)^{-1}$. Then, the objective value of the policy π_k^* at iteration k is within a factor of Δ of the ground-truth optimal solution, where $\Delta := \frac{\sigma_{\min}}{\sigma_{\max}}$ with probability greater than $(1-\epsilon)^{n(k-1)}\big(1 - \frac{\delta^{k-1}\mathrm{Var}(\lambda_i)}{\xi^2}\big)^n$.

Proof. Note that for any arbitrary total observation time $T \in \mathbb{R}_+$, a policy $\pi_k = (t_{1,k}^*, \ldots, t_{n,k}^*)$ satisfying

$$\forall i \in [n] \quad t_{i,k}^* := \frac{T}{\lambda_i \sum_{l=1}^{n} \frac{1}{\lambda_l}}. \tag{3.13}$$

optimizes the balance objective function $\hat{f}_{\text{bal}}$ [38]. Using the fact that $|\hat{\lambda}_{i,k} - \lambda_i| < \xi_i$ with probability given by Theorem 2, we arrive at the following inequality for $\hat{f}_{\text{bal}}(\pi_k^*)$

$$\hat{f}_{\text{bal}}(\pi_k^*) > \frac{\frac{T}{\sum_{l=1}^{n}(\lambda_i+\xi_l)^{-1}}}{\frac{nT}{\sum_{l=1}^{n}(\lambda_l-\xi_l)^{-1}}} = \frac{\sum_{l=1}^{n}(\lambda_l - \xi_l)^{-1}}{n\sum_{l=1}^{n}(\lambda_l + \xi_l)^{-1}}$$

with probability at least $(1-\epsilon)^{n(k-1)}\big(1 - \frac{\delta^{k-1}\mathrm{Var}(\lambda_i)}{\xi^2}\big)^n$. □

3.4 Results

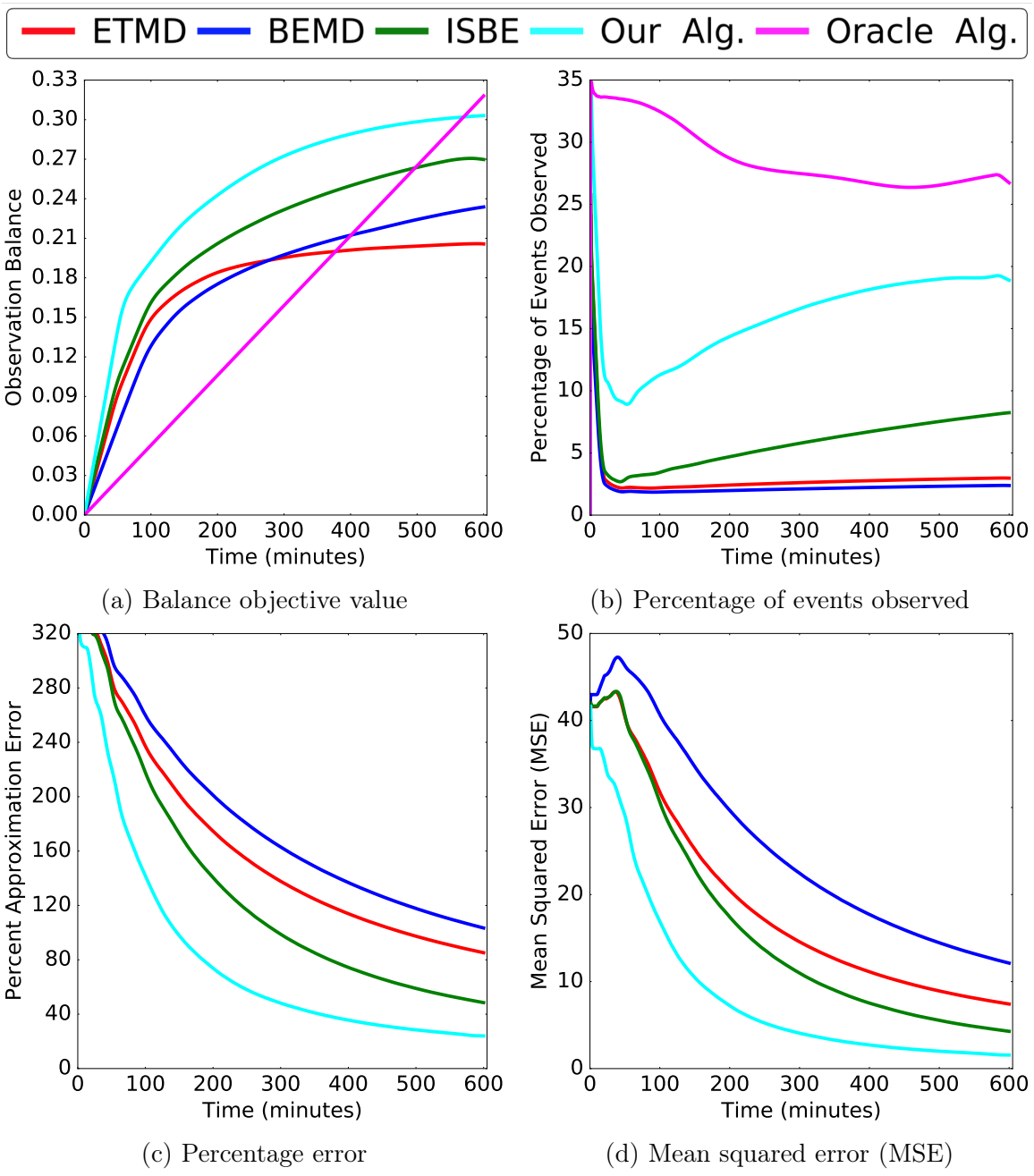
We evaluate the performance of Alg. 1 in two simulated scenarios modeled after real-world inspired monitoring tasks: (i) a synthetic simulation in which events at each station precisely follow a station-specific Poisson process and (ii) a scenario simulated in Armed Assault (ARMA) [7], a military simulation game, involving detections of suspicious agents. We note the statistics do not match our assumed Poisson model, and yet our algorithm performs well compared to other approaches. We compare Alg. 1 to the following monitoring algorithms:

1. Equal Time, Min. Delay (ETMD): computes the total cycle time to minimize latency T_{obs} [38] and partitions T_{obs} evenly across all stations.

2. Bal. Events, Min. Delay (BEMD): the algorithm introduced by [38] which

generates policies that minimize latency and maximize observation balance.

3. Incremental Search, Bal. Events (ISBE): generates policies to maximize balance that increase in length by a fixed amount $\Delta_{\text{obs}} \in \mathbb{R}_+$ after each cycle.



(a) Balance objective value

(b) Percentage of events observed

(c) Percentage error

(d) Mean squared error (MSE)

Figure 3-5: Results of the synthetic simulation averaged over 10,000 trials that characterize and compare our algorithm to the four monitoring algorithms in randomized environments containing three discrete stations.

4. Oracle Algorithm (Oracle Alg.): an omniscient algorithm assuming perfect knowledge of ground-truth rates and monitoring time $T_{\max}$ where each observation time is generated according to (3.13).

3.4.1 Synthetic Scenario

We consider the monitoring scenario involving the surveillance of events in three discrete stations over a monitoring period of 10 hours. We characterize the average performance of each monitoring algorithm with respect to 10,000 randomly generated problem instances with the following statistics:

1. Prior hyper-parameters: $\alpha_{i,0} \sim \text{Uniform}(1, 20)$ and $\beta_{i,0} \sim \text{Uniform}(0.75, 1.50)$.

2. Rate parameter of each station: $\mu_{\lambda_i} = 2.23$ and $\sigma_{\lambda_i} = 1.02$ events per minute.

3. Initial percentage error of the rate estimate $\lambda_{i,0}$, denoted by ρ_i: $\mu_{\rho_i} = 358.29\%$ and $\sigma_{\rho_i} = 221.32\%$.

4. Travel cost from station i to another j: $\mu_{d_{i,j}} = 9.97$ and $\sigma_{d_{i,j}} = 2.90$ minutes.

where μ and σ refer to standard deviation and variance of each parameter respectively and the transient events at each station $i \in [n]$ are simulated precisely according to $\text{Pois}(\lambda_i)$.

The performance of each algorithm with respect to the the monitoring objectives defined in Sect. 3.1 is shown in Figs. 3-5a and 3-5b respectively. The figures show that our algorithm is able to generate efficient policies that enable the robot to observe significantly more events that achieve a higher balance in comparison to those computed by other algorithms (with exception of Oracle Alg.) at all times of the monitoring process. Figs. 3-5c and 3-5d depict the efficiency of each monitoring algorithm in rapidly learning the events' statistics and generating accurate approximations. The error plots show that our algorithm achieves lower measures of error at any given time in comparison to those of other algorithms and supports our method's practical efficiency in generating exploratory policies conducive to rapidly obtaining accurate

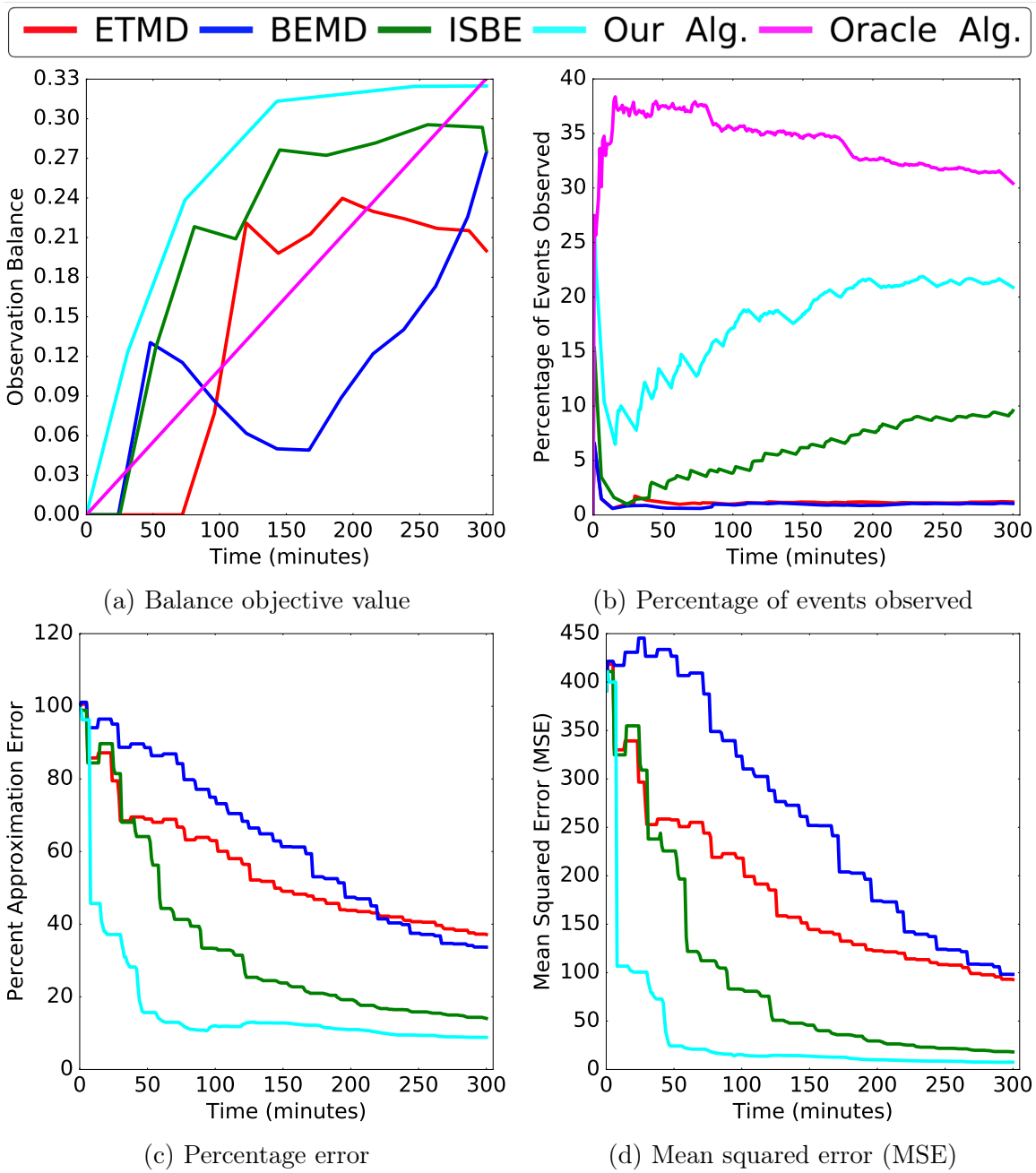
approximations of event statistics. Figs. 3-5a-3-5d show our algorithm's dexterity in balancing the inherent trade-off between exploration vs. exploitation.

Figure 3-6: Viewpoints from two stations in the ARMA simulation of the yellow backpack scenario. Agents wearing yellow backpacks whose detections are of interest appear in both figures.

3.4.2 Yellow Backpack Scenario

In this subsection, we consider the evaluation of our monitoring algorithm in a real-world inspired scenario, labeled the *yellow backpack scenario*, that entails monitoring of suspicious events that do not adhere to the assumed Poisson model (Sect. 3.1). Using the military strategy game ARMA, we simulate human agents that wander

around randomly in a simulated town. A subset of the agents wear yellow backpacks (see Fig. 3-6). Under this setting, our objective is to optimally monitor the yellow backpack-wearing agents using three predesignated viewpoints, i.e. stations. We considered a monitoring duration of 5 hours under the following simulation configuration:



(a) Balance objective value

(b) Percentage of events observed

(c) Percentage error

(d) Mean squared error (MSE)

Figure 3-7: The performance of each monitoring algorithm evaluated in the ARMA-simulated yellow backpack scenario.

1. Environment dimensions: 250 meters x 250 meters ($62,500$ meters^2).

2. Number of agents with a yellow backpack: 10 out of 140 ($\approx 7.1\%$ of agents).

3. Travel cost (minutes): $d_{1,2} = 3$, $d_{2,3} = 2$, $d_{3,1} = 12$.

We used the Faster Region-based Convolutional Neural Network (Faster R-CNN, [29]) for recognizing yellow backpack-wearing agents in real-time at a frequency of 1 Hertz. We ran the simulation for a sufficiently long time in order to obtain estimates for the respective ground-truth rates of 23.3, 20.3, and 18.5 yellow backpack recognitions per minute, which were used to generate Figs. 3-7c and 3-7d. The results of the yellow backpack scenario, shown in Figs. 3-7a-3-7d, tell the same story as did the results of the synthetic simulation. We note that at all instances of the monitoring process, our approach that leverages uncertainty estimates outperforms others in generating balanced policies conducive to efficiently observing more events and obtaining accurate rate approximations.

Chapter 4

Unknown, Dynamic Environments

In this chapter we relax the assumption of static event statistics and consider the problem of persistent monitoring in unknown, dynamic environments.

4.1 Problem Definition

Let there be $n \in \mathbb{N}_+$ discrete stations in the environment where transient events of interest occur according to inhomogeneous Poisson processes. The temporal variations at each station $i \in [n]$ are governed by an integrable rate function $\lambda_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ that is station-specific and independent of those of other stations. We assume that the rate functions can exhibit an unbounded number of abrupt changes, however, we require that the total variation of each function λ_i within the time horizon $T \in \mathbb{R}_+$ be bounded by a variation budget $V_T \in \mathbb{R}_+$ [5], i.e.,

$$\sup_{P \in \mathcal{P}} \sum_{j=1}^{n_p - 1} \max_{i \in [n]} \left| \lambda_i(p_{j+1}) - \lambda_i(p_j) \right| \leq V_T, \tag{4.1}$$

where $\mathcal{P} = \left\{ P = \{p_1, \ldots, p_{n_P}\} \mid P \text{ is a partition of } [0, T] \right\}$. We note that since our problem is intimately linked with the MAB problem, we address the case of a known surveillance duration T as is common in MAB literature.

We assume that there exists a travel cost, $c : [n] \times [n] \to \mathbb{R}_{\geq 0}$, associated with going from one station to another. Due to sensor constraints that mandate the robot to be

stationary to make accurate measurements, the robot cannot make observations while traveling. Our overarching monitoring objective is to generate an optimal sequence of station-time pairs that dictates the appropriate station visit order and respective observation windows in order to maximize the number of sighted events.

More formally, a policy $\pi = \big((s_1, t_1), \ldots, (s_m, t_m)\big)$ is a sequence of $m \in \mathbb{N}_+$ ordered pairs where each ordered pair, (s, t), denotes an observation window of $t \in \mathbb{R}_{\geq 0}$ time at station $s \in [n]$. For any non-negative reals a, b such that $a \leq b$, let $N_i(a, b]$ denote the random number of events that occur in the time interval $(a, b]$ at station $i \in [n]$. It follows then that $\mathbb{E}\big[N_i(a, b]\big] = \int_a^b \lambda_i(\tau)\ d\tau$ by definition of an inhomogeneous Poisson process at each station i. The expected number of events obtained for any policy $\pi = \big((s_1, t_1), \ldots, (s_m, t_m)\big)$ constrained by the total surveillance time T can then be computed as follows:

$$\mathbb{E}[N(\pi, T)] := \sum_{j=1}^{m} \int_{o_j(\pi)}^{o_j(\pi)+t_j} \lambda_{s_j}(\tau)\ d\tau, \tag{4.2}$$

where $o_j(\pi)$ denotes the start of the j^{th} observation window, i.e., $o_1(\pi) = 0$ and for any integral value $j > 1$,

$$o_j(\pi) := \sum_{k=1}^{j-1} t_k + c(s_k, s_{k+1}).$$

Our notion of weak regret is defined relative to the maximum number of *expected* events, N_T^* at a single best station after an allotted monitoring time of $T \in \mathbb{R}_+$:

$$N_T^* = \max_{i \in [n]} \mathbb{E}\big[N_i(0, T]\big] = \max_{i \in [n]} \int_0^{\mathrm{T}} \lambda_i(\tau)\ d\tau.$$

We seek to generate policies that minimize the *expected regret* with respect to the quantity N_T^*. We let $R(\pi, T)$ denote the regret accrued by policy $\pi = \big((s_1, t_1), \ldots, (s_m, t_m)\big)$ after time T

$$R(\pi, T) := N_T^* - N(\pi, T) \tag{4.3}$$

and define our optimization problem with respect to the expectation of $R(\pi, T)$.

Problem 4 (Persistent Surveillance Problem)**.** *Compute the optimal monitoring policy,* $\pi^* = \big((s_1^*, t_1^*), \ldots, (s_m^*, t_m^*)\big)$*, that minimizes the expected regret with respect to the allotted monitoring time* T

$$\pi^* = \underset{\pi}{\operatorname{argmin}}\ \mathbb{E}[R(\pi, T)]. \tag{4.4}$$

In what follows, we seek to minimize a long-term variant of Eqn. 4.4. We define the long-run average optimal policy as

Definition 1 (Long-run Average Optimal Policy)**.** A policy $\pi = \big((s_1, t_1), \ldots, (s_m, t_m)\big)$ is called a long-run average optimal policy if and only if

$$\limsup_{T \to \infty} \frac{\mathbb{E}[R(\pi, T)]}{T} = \frac{\mathbb{E}[N_T^*] - \mathbb{E}[N(\pi, T)]}{T} \leq 0. \tag{4.5}$$

4.2 Methods

In this section, we describe the intuition behind our approach, and present a monitoring algorithm (Alg. 3). Our approach trades off exploration and exploitation by leveraging information gained within bounded time steps. Specifically, we partition our alloted time into equal length intervals called epochs. Within each epoch we reason about the currently known best station and attempt to cleverly remove stations that are suboptimal with high probability. By removing suboptimal stations, future passes through the list of remaining stations are expected to yield better long-term rewards and require less time to be spent on traveling relative to observing stations.

The algorithm begins by computing the length of each epoch, τ, as a function of the total time T, variation budget V_T, and maximum travel time between each station $T_{\text{travel}} = \max_{i,j \in [n]: i \neq j} c(i, j)$. The variables N_i and T_i, denoting the total number of observations and the total time spent at station i respectively, are reset at the beginning of each epoch. Discarding out-dated information in this way enables us to balance remembering and forgetting by computing the average rate, $\hat{\lambda}_i = N_i / T_i$,

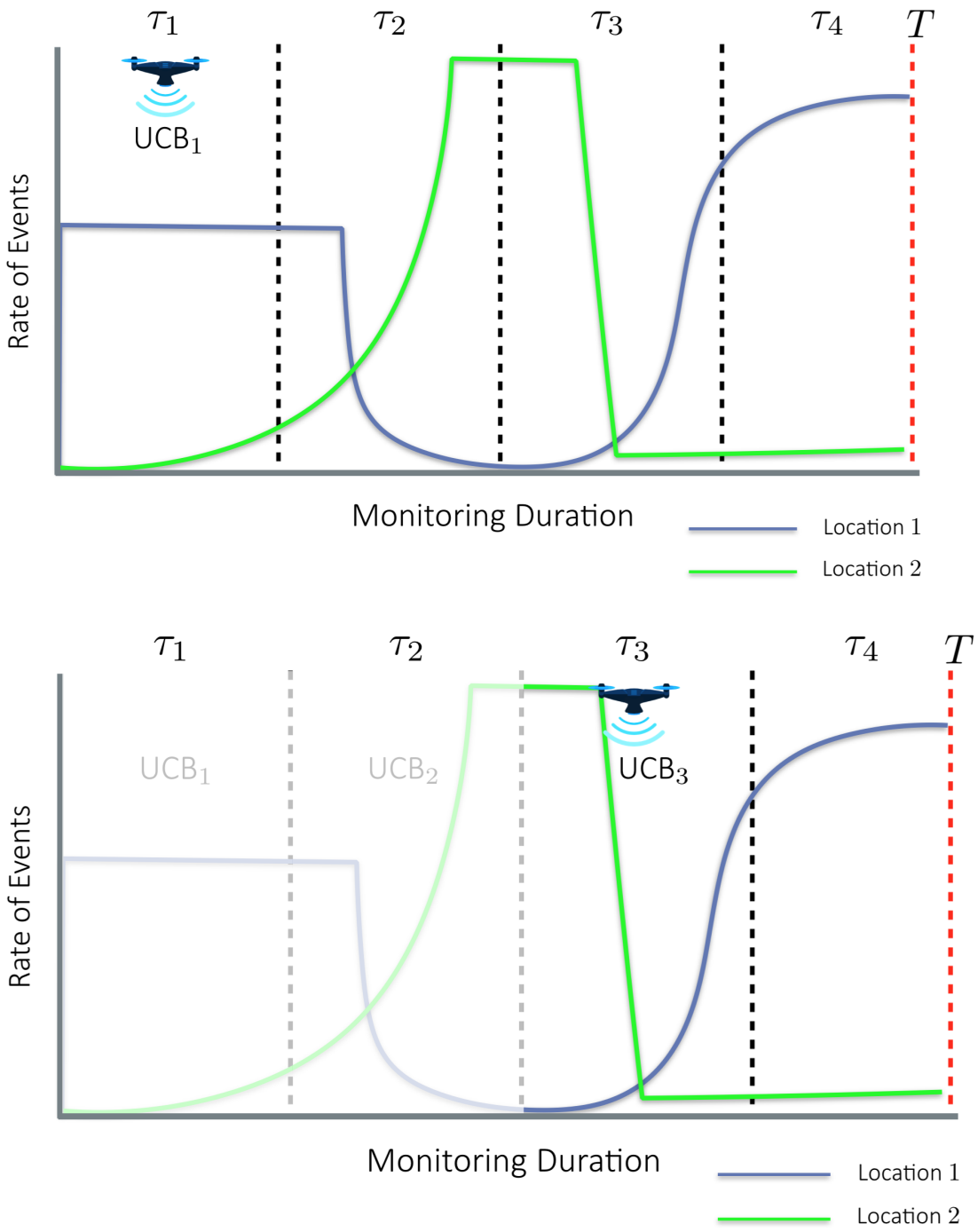


Figure 4-1: An example instance of our algorithm, depicting the partitioning of the monitoring period T into epochs of length τ. Within each epoch, the robot executes our variant of the Upper Confidence Bounds (UCB) algorithm to balance exploration and exploitation. Information obtained from prior epochs is purposefully discarded at the start of each epoch in order to adapt to the temporal variations in the environment.

for each station using only information obtained within that epoch. For each epoch, our method employs an algorithm based on the Improved UCB Algorithm [4] and seeks to balance the inherent exploration/exploitation trade-off.

4.3 Analysis

In this section, we present a regret bound analysis proving that the policy π generated by Alg. 3 is long-run average optimal with respect to our definition of weak regret. To establish our result, we proceed by bounding the total regret in each epoch of length τ, and then sum the regret over all $\lceil \frac{T}{\tau} \rceil$ epochs to obtain an upper bound on the entire monitoring horizon of length T.

4.3.1 Preliminaries

Assumption 2 (Bounded Rates)**.** *For a given time horizon $T \in \mathbb{R}_+$, the rate parameters $\forall i \in [n]$ $\lambda_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ are bounded above by a known constant λ_{max}.*

Define a *stage* indexed by $m \in \mathbb{N}$ as the completion of the inner while loop of Alg. 3 (i.e., execution of lines 13-30) and denote the partition of the time horizon T into $k = \lceil \frac{T}{\tau} \rceil$ epochs as $\tau_1, \ldots, \tau_k$ of length τ each (with the possible exception of τ_k). For a given stage m, we define $\tilde{\Delta}(m)$, $T_{\text{obs}}(m)$, $S(m)$, and $\xi(m)$ as the values of each variable at stage m (see Alg. 3). Let $w_{i,0}, \ldots, w_{i,m}$ be the $m + 1 \in \mathbb{N}$ observation windows at station $i \in [n]$, where each observation window $w_{i,j}$ is defined by the time interval $(a_{i,j}, b_{i,j})$. Note that $\sum_{j=0}^{m}(b_{i,j} - a_{i,j}) = T_{\text{obs}}(m)$.

For an arbitrary epoch τ_j, we let $\hat{\lambda}_i(m)$ denote the *sample mean* of the rate parameter and let $\bar{\lambda}_i(m)$ denote the ground-truth mean rate after observing at station i for m stages. We define $\bar{\lambda}_i$ to denote the average rate of a station over the specific epoch τ_j (that is clear from the context) and let $\bar{\lambda}^* = \max_{i \in [n]} \bar{\lambda}_i$ denote the epoch-specific optimal rate of the best station $*$. Finally, we let $\Delta_i = \bar{\lambda}^* - \bar{\lambda}_i$ denote the difference in the rates of a station i in comparison to that of the optimal station over epoch τ_j.

Algorithm 3: Dynamic Upper Confidence Bound Monitoring Algorithm

Input: Time horizon T, variation budget V_T, number of stations n, and travel costs $c : [n] \times [n] \to \mathbb{R}_{\geq 0}$.

Effect: Monitors locations of interest for T time.

$T_{\text{travel}} \leftarrow \max_{i,j \in [n]: i \neq j} c(i,j)$;

// Compute the length of each epoch

$\tau \leftarrow (n\lambda_{\max} T / V_T)^{\frac{2}{3}}$;

while $t_{current} \leq T$ **do**

  // Initialize parameters, discarding all previously obtained information from previous epochs

  $T_i \leftarrow 0 \ \ \forall i \in [n]; \quad N_i \leftarrow 0 \ \ \forall i \in [n]$;

  // Initialize the set of all station indices

  $S \leftarrow \{1, \ldots, n\}$;

  $\tilde{\Delta} \leftarrow \lambda_{\max}$;

  // Determine the end point of the current epoch

  $T_{\text{end}} \leftarrow t_{\text{current}} + \tau$;

  while $t_{current} \leq T_{end}$ **do**

    // Compute the goal observation time

    $T_{\text{obs}} \leftarrow \frac{8\lambda_{\max} \log\left(\tau \tilde{\Delta}^2\right)}{3\tilde{\Delta}^2}$;

    if $|S| > 1$ **then**

      for $i^* \in S$ *such that* $T_{i^*} < T_{obs}$ **do**

        $t_{i^*} \leftarrow \min\left\{T_{\text{end}} - t_{\text{current}}, T_{\text{obs}} - T_{i^*}\right\}$;

        Observe at station i^* for t_{i^*} time;

        $T_{i^*} \leftarrow T_{i^*} + t_{i^*}$;

    else

      // Only one station remains in S

      $t_{i^*} \leftarrow T_{\text{end}} - t_{\text{current}}$;

      Observe at the sole station $i^* \in S$ until T_{end};

      $T_{i^*} \leftarrow T_{i^*} + t_{i^*}$;

    // Identify and remove suboptimal stations

    $\xi \leftarrow \sqrt{\frac{8\lambda_{\max} \log\left(\tau \tilde{\Delta}^2\right)}{3T_{\text{obs}}}}$;

    $\hat{\lambda}^* \leftarrow \max_{i \in S} \hat{\lambda}_i - \xi$;

    $B \leftarrow \{i \in S \mid \hat{\lambda}_i + \xi < \hat{\lambda}^*\}$;

    $S \leftarrow S \setminus B$;

    $\tilde{\Delta} \leftarrow \frac{\tilde{\Delta}}{2}$;

Fact 1 (Chernoff Bounds)**.** Let $X \sim \text{Poisson}(\lambda)$, then the following holds for $\xi \in \mathbb{R}_{\geq 0}$:

$$\mathbb{P}\left(X > \lambda + \xi\right) \leq \exp\left(-\frac{\xi^2}{2\lambda}\psi(\xi/\lambda)\right),$$

and for $\xi \in [0, \lambda_{\max}]$

$$\mathbb{P}\left(X < \lambda - \xi\right) \leq \exp\left(-\frac{\xi^2}{2\lambda}\psi(-\xi/\lambda)\right) \leq \exp\left(-\frac{\xi^2}{2\lambda}\right),$$

where

$$\psi(x) = \frac{(1+x)\log(1+x) - x}{x^2/2} \geq (1 + x/3)^{-1}.$$

Lemma 4 (Concentration Inequalities)**.** For any station $i \in [n]$ and arbitrary sequence of observation windows $w_{i,0}, \ldots, w_{i,m}$ such that $T_{\text{obs}}(m) = \sum_{j=0}^{m}(b_{i,j} - a_{i,j})$ and $\xi \in \mathbb{R}_{\geq 0}$:

$$\mathbb{P}\left(\hat{\lambda}_i(m) > \bar{\lambda}_i(m) + \xi\right) \leq \exp\left(-\frac{3T_{\text{obs}}(m)\xi^2}{8\lambda_{\max}}\right)$$

and for $\xi \in [0, \lambda_{\max}]$

$$\mathbb{P}\left(\hat{\lambda}_i(m) < \bar{\lambda}_i(m) - \xi\right) \leq \exp\left(-\frac{3T_{\text{obs}}(m)\xi^2}{8\lambda_{\max}}\right).$$

Proof. Let N be the random variable denoting the number of events observed during the observation windows summing up to a total of T_{obs} time. Then, it follows by definition of a Poisson process that $N \sim \text{Poisson}(\bar{\lambda}_i T_{\text{obs}})$ and thus by the aforementioned Chernoff bounds, we have for $\xi \in \mathbb{R}_{\geq 0}$:

$$\begin{aligned}\mathbb{P}\left(\hat{\lambda}_i \leq \bar{\lambda}_i + \xi\right) &= \mathbb{P}\left(N > \mathbb{E}[N] + \xi T_{\text{obs}}\right) \\ &\leq \exp\left(\frac{T_{\text{obs}}\xi^2}{2\bar{\lambda}_i}\psi(\xi/\bar{\lambda}_i)\right)\end{aligned}$$

and for $\xi \in [0, \bar{\lambda}_i]$:

$$\mathbb{P}\left(\hat{\lambda}_i \leq \bar{\lambda}_i - \xi\right) \leq \exp\left(\frac{T_{\text{obs}}\xi^2}{2\bar{\lambda}_i}\psi(-\xi/\bar{\lambda}_i)\right).$$

Now, note that $\xi \leq \lambda_{\max}$ by definition, and consider the expression $\frac{T_{\text{obs}}\xi^2}{2\bar{\lambda}_i}\psi(\xi/\bar{\lambda}_i)$:

$$\begin{aligned}\frac{T_{\text{obs}}\xi^2}{2\bar{\lambda}_i}\psi(\xi/\bar{\lambda}_i) &\geq \frac{T_{\text{obs}}\xi^2}{2\bar{\lambda}_i}\psi(\xi/\bar{\lambda}_i) \\ &\geq \frac{T_{\text{obs}}\xi^2}{2\bar{\lambda}_i}\left(\frac{3\bar{\lambda}_i}{3\bar{\lambda}_i+\xi}\right) \\ &\geq \frac{3T_{\text{obs}}\xi^2}{8\lambda_{\max}}\end{aligned}$$

where we used the inequality $\psi(x) \geq (1 + x/3)^{-1}$ mentioned in Fact 1. The result above implies that

$$\exp\left(-\frac{T_{\text{obs}}\xi^2}{2\lambda}\psi(\xi/\bar{\lambda}_i)\right) \leq \exp\left(-\frac{3T_{\text{obs}}\xi^2}{8\lambda_{\max}}\right). \tag{4.6}$$

Application of (4.6) yields the desired result:

$$\forall \xi \in \mathbb{R}_{\geq 0} \quad \mathbb{P}\left(\hat{\lambda}_i > \bar{\lambda}_i + \xi\right) \leq \exp\left(-\frac{3T_{\text{obs}}\xi^2}{8\lambda_{\max}}\right)$$

and for $\xi \in [0, \lambda_{\max}]$:

$$\mathbb{P}\left(\hat{\lambda}_i > \bar{\lambda}_i + \xi\right) \leq \exp\left(-\frac{T_{\text{obs}}\xi^2}{2\bar{\lambda}_i}\right) \leq \exp\left(-\frac{3T_{\text{obs}}\xi^2}{8\lambda_{\max}}\right)$$

□

4.3.2 Regret over an epoch

We decompose the total expected regret over an arbitrary epoch τ_j of length τ, $\mathbb{E}[R(\pi, \tau)]$, and consider the regret incurred by observing and traveling separately, i.e., $\mathbb{E}[R(\pi, \tau)] = \mathbb{E}[R_{\text{obs}}(\pi, \tau)] + \mathbb{E}[R_{\text{travel}}(\pi, \tau)]$.

Lemma 5 (Regret over an Epoch)**.** The per-epoch expected observation regret of our algorithm, $\mathbb{E}[R_{\text{obs}}(\pi, \tau)]$, with respect to an arbitrary epoch j of length τ and with

variation budget V_j is at most

$$(\Delta + V_j)\tau + \sum_{i \in \mathcal{B}} \left(\frac{1024}{\Delta_i} + \frac{2560\lambda_{\max} \log\left(\tau \Delta_i^2 / 64\right)}{3\Delta_i} \right),$$

where $\Delta = \max\left\{4V_j, \sqrt{\frac{8\exp(1-3/(5\lambda_{\max}))}{\tau}}\right\}$ and $\mathcal{B} = \{i \in [n] \mid \Delta_i > \Delta\}$.

Proof. Our proof employs results established in, and follows a similar structure as the proof given by Auer [4] and [5]. Let V_j denote the total variation in the rates during epoch j, i.e.,

$$V_j = \sup_{P \in \mathcal{P}_j} \sum_{k=1}^{n_p - 1} \max_{i \in [n]} \left| \lambda_i(p_{k+1}) - \lambda_i(p_k) \right|, \tag{4.7}$$

where $\mathcal{P}_j$ is a partition of epoch τ_j. Summing over all epochs $j = 1, \ldots, \lceil \frac{T}{\tau} \rceil$, note that $\sum_{j=1}^{\lceil T/\tau \rceil} V_j \leq V_T$.

Let $m_i = \min\{m : \tilde{\Delta}(m) < \Delta_i/8\}$ denote the first stage index in which our guess $\tilde{\Delta}(m)$ is close to the actual difference in the rates for stations $i \in S$. The following inequalities follow by definition

$$\tilde{\Delta}(m_i) = \frac{\lambda_{\max}}{2^{m_i}} < \frac{\Delta_i}{8} \leq 2\tilde{\Delta}(m_i) = \frac{2\lambda_{\max}}{2^{m_i}} \tag{4.8}$$

and

$$\xi(m_i) = \sqrt{\frac{8\lambda_{\max} \log\left(\tau \tilde{\Delta}^2(m_i)\right)}{3T_{\text{obs}}(m_i)}} = \tilde{\Delta}(m_i) < \Delta_i/8.$$

We will consider bounding the regret incurred by monitoring clearly suboptimal locations, i.e., $\mathcal{B} = \{i \in S \mid \Delta_i > \Delta\}$, instead of monitoring the optimal station $*$, where $\Delta = \max\{4V_j, \sqrt{\frac{8\exp(1-3/(5\lambda_{\max}))}{\tau}}\}$.

Case 1 (At stage m_i, there exists a sub-optimal station $i \in S(m_i)$ and the optimal station $*$ is in $S(m_i)$)**.** We will proceed by finding an upper bound for the probability that the sub-optimal station i is *not* removed during the duration of the epoch τ_j.

Consider the following inequalities for some stage m

$$\hat{\lambda}_i(m) \leq \bar{\lambda}_i(m) + \xi(m) \tag{4.9}$$

$$\hat{\lambda}^*(m) \geq \bar{\lambda}^*(m) - \xi(m). \tag{4.10}$$

If conditions (4.9) and (4.10) hold at stage $m = m_i$ under the assumption that $* \in S(m_i)$, then it follows that i will be removed from $S(m_i)$ at stage m_i

$$\begin{aligned}
\hat{\lambda}_i(m_i) + \xi(m_i) &\leq \bar{\lambda}_i(m_i) + 2\xi(m_i) && \text{by (4.9)} \\
&\leq \bar{\lambda}_i + V_j + 2\xi(m_i) && \text{by (4.7)} \\
&< \bar{\lambda}_i + \Delta_i - V_j - 2\xi(m_i) \\
&\leq \bar{\lambda}^* - V_j - 2\xi(m_i) \\
&\leq \bar{\lambda}^*(m_i) - 2\xi(m_i) && \text{by (4.7)} \\
&\leq \hat{\lambda}^*(m_i) - \xi(m_i) && \text{by (4.10)}
\end{aligned}$$

where we used the fact that $\Delta_i > 2V_j + 4\xi(m_i)$.

Using Lemma 4, the probability that either (4.9) or (4.10) does *not* hold is as follows.

$$\mathbb{P}\left(\hat{\lambda}_i(m_i) > \bar{\lambda}_i(m_i) + \xi(m_i)\right) \leq \frac{1}{\tau\tilde{\Delta}^2(m_i)},$$

and similarly for condition (4.10)

$$\mathbb{P}\left(\hat{\lambda}^*(m_i) < \bar{\lambda}^*(m_i) - \xi(m_i)\right) \leq \frac{1}{\tau\tilde{\Delta}^2(m_i)}$$

By the union bound, the probability that the sub-optimal station is not eliminated in stage m_i (or before) is bounded above by $\frac{2}{\tau\tilde{\Delta}_i^2(m_i)}$. Taking the sum of conditional expectations over each station $i \in \mathcal{B}$ and bounding the regret over the epoch by $\tau\Delta_i$

yields the accumulated regret for this case:

$$\mathbb{E}[R_1] \leq \sum_{i \in \mathcal{B}} \frac{2\tau\Delta_i}{\tau\tilde{\Delta}^2(m_i)} \leq \sum_{i \in \mathcal{B}} \frac{32}{\tilde{\Delta}(m_i)} \leq \sum_{i \in \mathcal{B}} \frac{512}{\Delta_i}.$$

Case 2 (For each sub-optimal station $i \in \mathcal{B}$, either i is eliminated at stage m_i (or before) or the optimal station $*$ is not in $S(m_i)$)**.**

Case 2.1 (Station i is eliminated at stage m_i or before)**.** Since each sub-optimal arm $i \in \mathcal{B}$ is eliminated in round m_i at the latest, the total observation time spent at i follows by the definition of $T_{\text{obs}}(m_i)$:

$$\begin{aligned} T_{\text{obs}}(m_i) &= \frac{8\lambda_{\max} \log\left(\tau\tilde{\Delta}^2(m_i)\right)}{3\tilde{\Delta}^2(m_i)} \\ &\leq \frac{2048\lambda_{\max} \log\left(\tau\Delta_i^2/64\right)}{3\Delta_i^2}. \end{aligned}$$

Multiplying the observation time by $\Delta_i + V_j$ for each $i \in \mathcal{B}$ to account for variation up to stage m_i (i.e., since $\bar{\lambda}_i(m) > \bar{\lambda}_i$ may hold, but $\bar{\lambda}_i(m) \leq \bar{\lambda}_i + V_j$ is assured for all m), we establish the regret bound for this case:

$$\begin{aligned} \mathbb{E}[R_2] &\leq \sum_{i \in \mathcal{B}} (\Delta_i + V_j) T_{\text{obs}}(m_i) \\ &< \sum_{i \in \mathcal{B}} \frac{5\Delta_i T_{\text{obs}}(m_i)}{4} \\ &\leq \sum_{i \in \mathcal{B}} \frac{2560\lambda_{\max} \log\left(\tau\Delta_i^2/64\right)}{3\Delta_i}. \end{aligned}$$

Case 2.2 (There is no optimal station $*$ in $S(m_i)$)**.** Consider the event that an arbitrary sub-optimal station $i \in \mathcal{B}$ removes the optimal station $*$ at stage m_*. This implies that the following removal condition holds at m_*

$$\hat{\lambda}_i(m_*) - \xi(m_*) > \hat{\lambda}^*(m_*) + \xi(m_*). \tag{4.11}$$

If we assume that the inequalities (4.9) and (4.10) hold at stage m_*, then (4.11) leads

to the contradiction $\bar{\lambda}_i + 2V_j > \bar{\lambda}^*$ since $\Delta_i > 4V_j$:

$$\begin{aligned}
\bar{\lambda}_i + V_j &\geq \bar{\lambda}_i(m_*) \\
&\geq \hat{\lambda}_i(m_*) - \xi(m_*) && \text{by (4.9)} \\
&> \hat{\lambda}^*(m_*) + \xi(m_*) && \text{by (4.11)} \\
&\geq \bar{\lambda}^*(m_*) && \text{by (4.10)} \\
&\geq \bar{\lambda}^* - V_j.
\end{aligned}$$

Thus, it follows that if (4.9) and (4.10) hold at stage m_*, the optimal station $*$ will not be removed at this stage. Thus, using previously established results, we have that the probability that an arbitrary sub-optimal station $i \in \mathcal{B}$ removes $*$ at stage m_* is at most $\frac{2}{\tau\tilde{\Delta}^2(m_*)}$.

If $*$ is indeed removed by a sub-optimal station $i \in \mathcal{B}$, then by definition of the considered case, we have that $* \in \mathcal{B}$ for all stages $m_k < m_*$ and thus, all stations k such that $m_k < m_*$ were eliminated at or before round m_k. This also implies that $*$ can only be eliminated in round m_* by some $i \in \mathcal{B}$ satisfying $m_i \geq m_*$. The regret incurred after having removed the optimal station $*$ at stage m_* is bounded above by $\tau \max_{k\in\mathcal{B}:m_k\geq m_*} \Delta_k \leq 16\tau\tilde{\Delta}(m_*)$. Taking the sum of the conditional expectations over all $i \in \mathcal{B}$ and considering the probability of removing $*$ yields the expected regret for this case, R_3:

$$\begin{aligned}
\mathbb{E}[R_3] &\leq \sum_{i\in\mathcal{B}} \sum_{m_*=0}^{m_i} \frac{32\tau\tilde{\Delta}(m_*)}{\tau\tilde{\Delta}^2(m_*)} \\
&= \sum_{i\in\mathcal{B}} \sum_{m_*=0}^{m_i} \frac{32}{\tilde{\Delta}(m_*)} = \sum_{i\in\mathcal{B}} \sum_{m_*=0}^{m_i} 32\frac{2^{m_*}}{\lambda_{\max}} \\
&= \sum_{i\in\mathcal{B}} 32\frac{(2^{m_i+1}-1)}{\lambda_{\max}} \\
&< \sum_{i\in\mathcal{B}} 32\frac{2^{m_i+1}}{\lambda_{\max}} = \sum_{i\in\mathcal{B}} \frac{32}{\tilde{\Delta}(m_{i+1})} \\
&\leq \sum_{i\in\mathcal{B}} \frac{1024}{\Delta_i}.
\end{aligned}$$

Finally, summing over the expected regrets $\mathbb{E}[R_1], \mathbb{E}[R_2]$, and $\mathbb{E}[R_3]$ and bounding the regret of sub-optimal stations not in $\mathcal{B}$ by $\Delta_i \tau \leq (\Delta + V_j)\tau$ yields the result. $\square$

Lemma 6 (Distribution-independent Epoch Regret)**.** The per-epoch expected regret of our algorithm, $\mathbb{E}[R_{\text{obs}}(\pi, \tau)]$, with respect to an arbitrary epoch j of length τ and with total variation budget V_j is at most

$$\mathbb{E}[R_{\text{obs}}(\pi, \tau)] = \mathcal{O}\left(V_j \tau + n\sqrt{\tau}\lambda_{\max}\right).$$

Proof. Consider the right-hand side of the bound on $\mathbb{E}[R_{\text{obs}}(\pi, \tau)]$ from Lemma 5, which is a function of the variable Δ_i, $R_{\Delta_i}(\pi, \tau)$, i.e.,

$$R_{\Delta_i}(\pi, \tau) = \sum_{i \in \mathcal{B}} \left(\frac{1024}{\Delta_i} + \frac{2560 \lambda_{\max} \log\left(\tau \Delta_i^2 / 64\right)}{3\Delta_i} \right).$$

Differentiating $R_{\Delta_i}(\pi, \tau)$ with respect to Δ_i, setting the resulting expression to 0, and solving for Δ_i^* to yields

$$\Delta_i^* = \frac{8 \exp\left(1 - \frac{3}{5\lambda_{\max}}\right)}{\sqrt{\tau}}.$$

Observe that the $\Delta\tau$ term on the left-hand side of the regret bound is bounded above by the expression in the sum for all values of $\Delta \geq \sqrt{\frac{e}{T}}$ and reasonable $\lambda_{\max}$ (i.e., $\lambda_{\max} \geq 1$). Thus, the expression in the sum is the dominant term in the bound. Evaluating the regret $R_{\Delta_i}(\pi, \tau)$ by plugging in Δ_i^* establishes the upper bound on $R_{\Delta_i}(\pi, \tau)$:

$$\begin{aligned} R_{\Delta_i}(\pi, \tau) &\leq 214 n \sqrt{\tau} \lambda_{\max} \exp(3/(5\lambda_{\max}) - 1) \\ &= \mathcal{O}(n\sqrt{\tau}\lambda_{\max}) \end{aligned}$$

which is independent of Δ_i. The bound above in conjunction with the $V_j \tau$ term establishes the asymptotic distribution-independent bound. $\square$

Lemma 7 (Bound on Travel Time Per Epoch)**.** In an epoch length of duration τ, the total regret incurred by traveling from one station to the other is bounded above

by

$$\mathbb{E}[R_{\text{travel}}(\pi, \tau)] = \mathcal{O}\left(\log(\tau) n \lambda_{\max} T_{\text{travel}}\right).$$

Proof. By definition of our algorithm, no more than $\mathcal{O}(\log(\tau))$ stages can be executed within an epoch of length τ. Moreover, since the cardinality of S is at most n at each stage, the regret incurred per stage is bounded above by $n\lambda_{\max} T_{\text{travel}}$, which yields the result. □

4.3.3 Total Regret

Theorem 4 (Long-run Average Optimality)**.** The total expected regret of Alg. 3, $\mathbb{E}[R(\pi, T)]$, over the entire monitoring duration T and total variation budget V_T is bounded by

$$\mathbb{E}[R(\pi, T)] = \mathcal{O}\left(V_T^{1/3} (T n \lambda_{\max})^{2/3}\right),$$

for a choice of epoch length $\tau = \mathcal{O}((n\lambda_{\max} T / V_T)^{2/3})$ as long as $n\lambda_{\max} T_{\text{travel}}$ is negligible relative to T, i.e., $n\lambda_{\max} T_{\text{travel}} = \mathcal{O}(1)$.

Proof. Invoking Lemmas 6 and 7 and summing over $\lceil \frac{T}{\tau} \rceil$ epochs yields

$$\begin{aligned}
&\mathbb{E}[R(\pi, T)] \\
&\quad = \sum_{j=1}^{\lceil T/\tau \rceil} \mathcal{O}\left(V_j \tau + n\lambda_{\max}(\sqrt{\tau} + \log(\tau) T_{\text{travel}})\right) \\
&\quad = \mathcal{O}(V_T \tau) + \mathcal{O}\left((\frac{T}{\tau} + 1)(n\lambda_{\max}\sqrt{\tau})\right) \\
&\quad = \mathcal{O}\left(V_T \tau + \frac{T n \lambda_{\max}}{\sqrt{\tau}}\right).
\end{aligned}$$

Setting $\tau = (n\lambda_{\max}T/V_T)^{2/3}$ we have

$$\begin{aligned}\mathbb{E}[R(\pi, T)] &= \mathcal{O}\left(V_T^{1/3}(Tn\lambda_{\max})^{2/3}\right) \\ &= o(T),\end{aligned}$$

which establishes the long-run average optimality of our algorithm. □

4.4 Results

We evaluate the performance of our algorithm in simulated environments subject to temporal variations and compare its effectiveness in maximizing the number of observations within the allotted monitoring time. In particular, we compare our algorithm (Alg. 3) to the following baseline and adaptive procedures that are inspired by state-of-the-art methods:

1. Random Choice & Time: picks a station i^* uniformly at random and observes for a random time $t_i^* \sim \text{Exp}(\lambda_{\text{exp}}(t_{\text{current}}, i^*))$.

2. ϵ-greedy: explores with probability $\epsilon(t_{\text{current}})$ (see Random Choice & Time) and exploits otherwise, i.e. $i^* = \text{argmax}_{i\in[n]}\, \bar{\lambda}_i$ and $t_i^* \sim \text{Exp}(\lambda_{\text{exp}}(t_{\text{current}}, i^*))$.

3. Discounted ϵ-greedy: same procedure as ϵ-greedy except discounted sample means, $\tilde{\lambda}_i$, are used instead.

4. Discounted Cyclic Policy: generates cyclic policies using an extended version of the algorithm introduced by [38] where discounted sample means, $\tilde{\lambda}_i$, are used to ensure adaptiveness.

where $\lambda_{\text{exp}}(t, i) = \bar{\lambda}_i/(n\epsilon(t))$ and $\epsilon : \mathbb{R}_{\geq 0} \to [0, 1]$ denotes the exploration function defined as $\epsilon(t) = 1/\log t$. Methods 3 and 4 employ discounted sample means which is computed as follows at time t_{current}

$$\forall i \in [n] \quad \tilde{\lambda}_i = \frac{\sum_{t=0}^{\lceil t_{\text{current}}\rceil} \gamma^{(\lceil t_{\text{current}}\rceil - t)} N_i(t)}{\sum_{t=0}^{\lceil t_{\text{current}}\rceil} \gamma^{(\lceil t_{\text{current}}\rceil - t)} T_i(t)}, \tag{4.12}$$

where $\gamma = 0.99$ and $N_i(t)$ and $T_i(t)$ denote the sum of events observed and the total observation time spent at station i up to time t respectively.

4.4.1 Sinusoidal Variations

We consider the simulated scenario involving the surveillance of two spatially-distributed stations where events occur according to unknown event statistics that are subject sinusoidal temporal variations. The rate functions of the two stations are given as a function of the variation budget V_T where V_T depends sub-linearly on the allotted surveillance time, $V_T = \sqrt{T}$ (see Fig. 4-2)(a):

$$\lambda_1(t) = \frac{1}{2} + \frac{1}{2}\sin\left(\frac{\pi V_T t}{T}\right)$$
$$\lambda_2(t) = \frac{1}{2} + \frac{1}{2}\sin\left(\frac{\pi V_T t}{T} + \pi\right).$$

The cost of travel from one station to the other is assumed to be 3 minutes of travel during which the robot is unable to record any observations.

Figures 4-3(a)-(b) show the average performance of each monitoring algorithm given a time horizon $T = 20,000$ minutes, over 100 trials. Our algorithm (shown in cyan) achieves sub-linear regret over time, reaffirming the theoretical property of our algorithm established in Sec. 3.3. In comparison to the other adaptive monitoring algorithms, our algorithm is the only procedure that achieves $\mathbb{E}[R(\pi, T)]/T \approx 0$ (see Def. 1). Furthermore, Fig. 4-3(c) shows that our algorithm observes the highest percentage of event sighting with respect to the cumulative number of all events that occurred across all stations.

4.4.2 Discrete Random Walk

In the previous subsection, we considered environments with temporal variations with a relatively small variation budget $V_T = \sqrt{T}$ where the changes in the environment were continuous and sinusoidal. In this subsection, we consider a significantly more erratic and challenging scenario in which we increase the budget to $V_T = T^{2/3}$ and

allow discontinuous, abrupt temporal variations in the event rates. In particular, we consider monitoring 3 stations where the rate functions follow a bounded discrete random walk that is generated as a function of the variation budget $V_T = T^{2/3}$.

Namely, for each station $i \in [n]$, we construct a station-specific random sequence defined by $X_0 \sim \text{Uniform}(0, 1)$ and X_t for $t \in \mathbb{N}_+$, $t \leq T$:

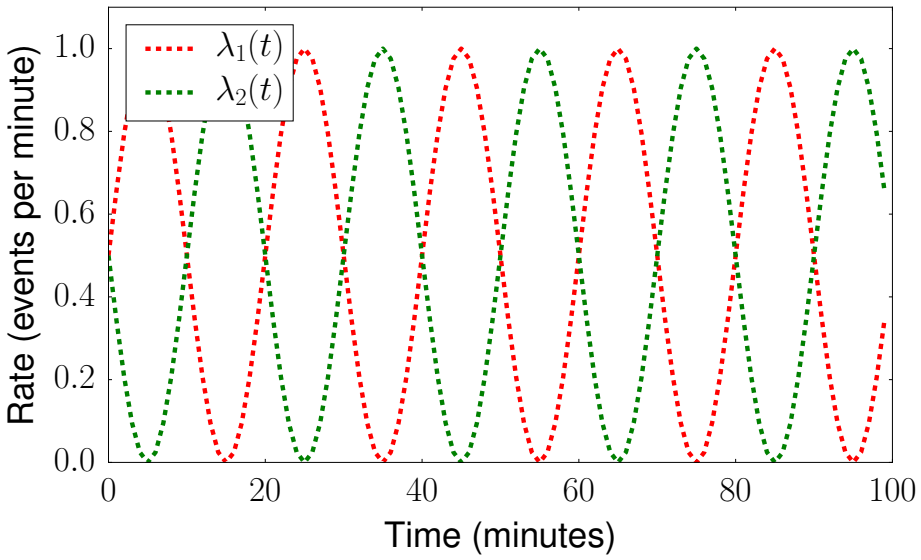
$$X_t = \begin{cases} X_{t-1} + U_t & \text{if } X_{t-1} + U_t > 0 \\ X_{t-1} + |U_t| & \text{otherwise} \end{cases} \tag{4.13}$$

where $U_t \sim \text{Uniform}(-V_T/T, V_T/T)$. Then, the rate function associated with station $i \in [n]$ is defined as

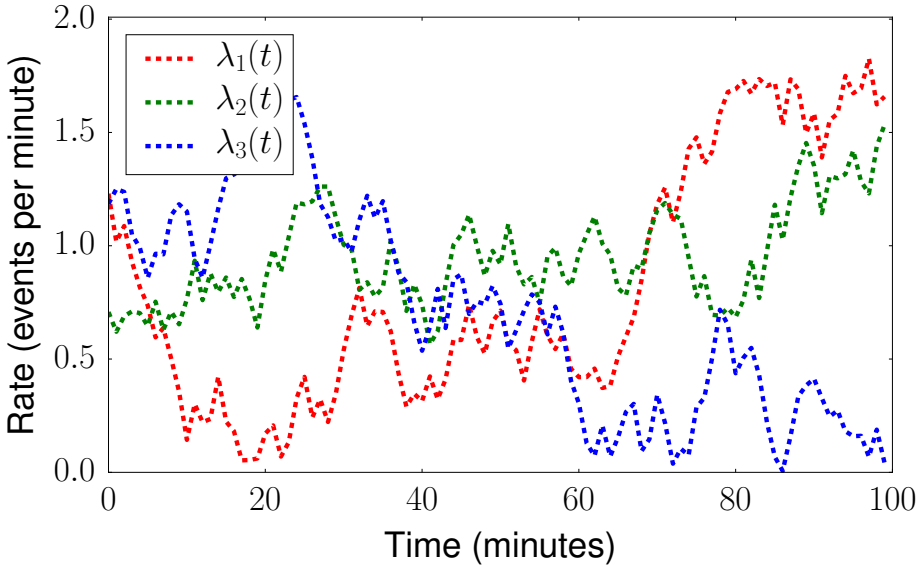
$$\lambda_i(t) = X_{\lceil t \rceil}. \tag{4.14}$$

Figure 4-2(b) depicts an example generated by this construction performed with a curtailed time horizon of $T = 100$ minutes. The travel time between one station i to the other j, $i \neq j$, is uniformly drawn, i.e., $c(i, j) \sim \text{Uniform}(1, 5)$ minutes.

Figures 4-4(a)-(c) show the performance of each monitoring algorithm for a time horizon $T = 20,000$ minutes, averaged over 100 trials. The figures tell the same story as did those from the previous subsection: our algorithm (cyan) is the only method to achieve sub-linear regret over time (Figs. 4-4)(a)-(b), which reaffirms the long-run average optimality of the policies generated by our algorithm, depicted in Fig. 4-4(b). In addition to minimizing the regret metric formalized in Sec. 3.1, the policies generated by our method achieve the highest percentage of observed events taken with respect to all of the transpired events at 3 stations, as shown in Fig. 4-4(c).

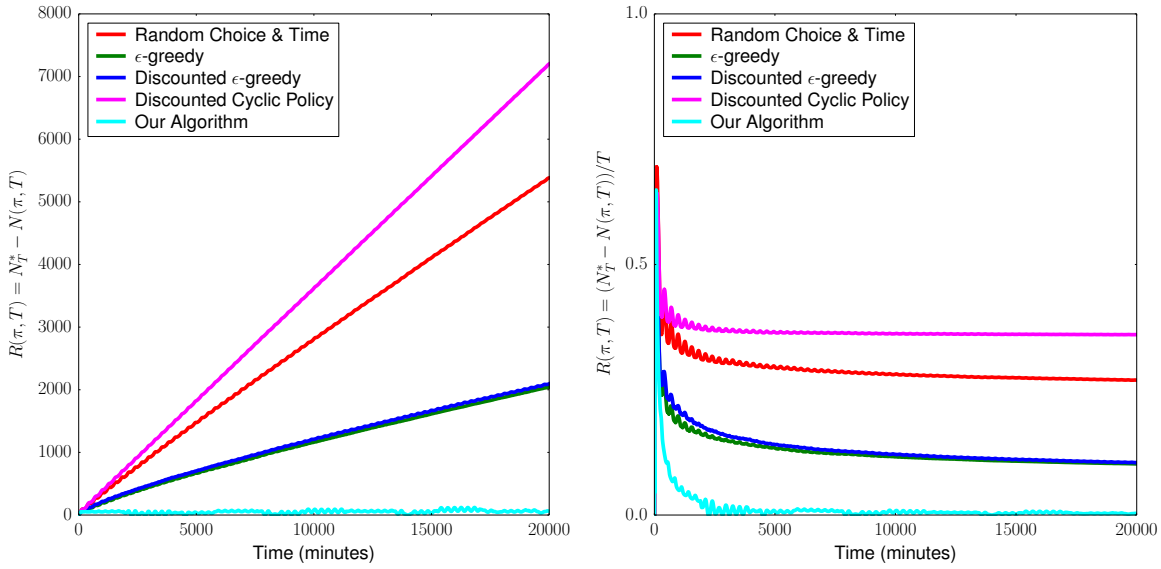


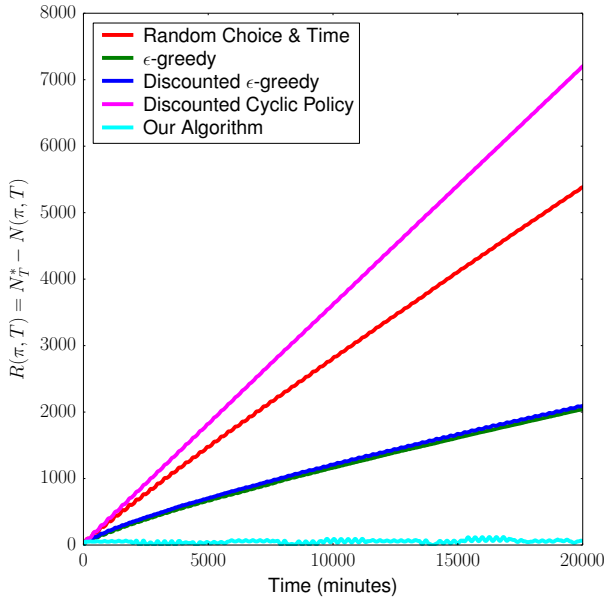
(a)



(b)

Figure 4-2: The two scenarios explored in our experiments. a) The sinusoidal rates of each Poisson process as a function of time with $V_T = \sqrt{T}$ and $T = 100$ minutes. b) The rates of each Poisson process as a function of time generated by a discrete random walk as described in Sec. 4.4.2. The figure depicts the rates of three stations over a time horizon $T = 100$ minutes and variation budget $V_T = T^{2/3}$.





(c)

Figure 4-3: a) Plot of total regret $R(\pi, T) = N_T^* - N(\pi, T)$ over time. The figure depicts sub-linear growth of regret over time for our algorithm (cyan), as expected from our theoretical results (Sec. 3.3). b) Growth of total regret over time expressed as the quotient $R(\pi, T)/T$. Our algorithm achieves sub-linear regret over time and that $R(\pi, T)/T \to 0$. c) Percentage of events observed with respect to the sum of events that occurred across all stations in the environment subject to sinusoidal variation over time. Our algorithm approximately attains optimal number of expected events in this setting consisting of 2 stations.

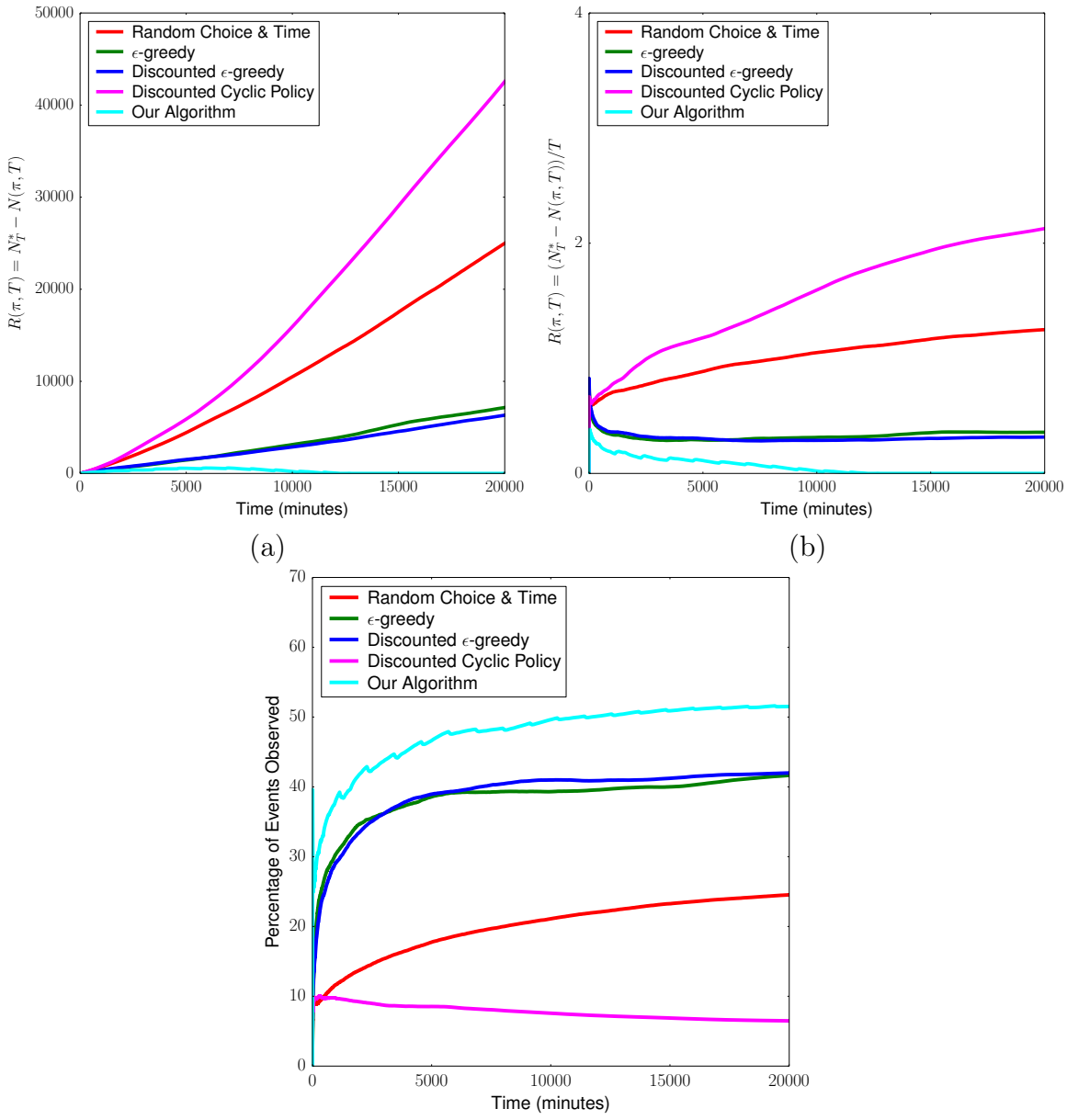


Figure 4-4: Clockwise a) Total regret as a function of time, i.e. $R(\pi, T) = N_T^* - N(\pi, T)$, in the simulated scenario involving discontinuous, abrupt changes. Our algorithm (shown in cyan) achieves the lowest regret at all times of the allotted monitoring time $T = 20,000$ minutes. b) Growth of the total regret over time, $R(\pi, T)/T$, in an abruptly changing environment. c) Percentage of events observed with respect to all of the events that transpired in the abruptly changing environment (Sec. 4.4.2) at all stations during the time horizon T.

Chapter 5

Conclusion

5.1 Conclusion

In this thesis, we presented two main results that remedy the limitations of current persistent monitoring algorithms.

First, we presented a novel algorithm for monitoring transient events in unknown, but otherwise static environments over a long period of time. We presented a novel monitoring algorithm capable of optimizing multiple monitoring objectives while simultaneously balancing the inherent exploration and exploitation trade-off. Our favorable theoretical and empirical results show promise of our algorithm's application to a wide variety of monitoring scenarios where the agent is equipped with limited a priori knowledge.

Second, we extended the aforementioned problem formulation and considered the problem of persistent surveillance in unknown and dynamic environments, where the event statistics are subject to temporal variations. We proposed an algorithm that builds upon and extends the state-of-the-art in persistent surveillance by introducing a method of constructing policies that are provably long-run average optimal even in scenarios subject to discontinuous, abrupt temporal variations. We presented rigorous regret-analysis of our algorithm and proved its long-run average optimality as long as the travel costs and the number of stations are not exceedingly high. Our method hinged on novel connections between persistent surveillance and the Multi-armed

Bandit (MAB) problem variants, which may be of independent theoretical interest.

Our favorable simulation results in both continuously and abruptly changing environments reaffirm our theoretical results and show the potential applications of our algorithms to a wide range of monitoring applications. We envision that our algorithm may be employed to facilitate persistent surveillance missions, such as detection and tracking efforts at a large scale.

5.2 Limitations and Future Work

There are many practically and theoretically avenues of interest for future research. From a practical standpoint, it would to be interesting to investigate the performance of our algorithms' against real-world surveillance data. A natural extension of this setting would be to integrate our algorithms with hardware to conduct surveillance missions using an UAV, as an example.

Many interesting algorithmic and theoretical questions also remain:

1. Can the algorithm of Chapter 4 be extended so that the monitoring duration T does not have to be given as input?

2. Can the asymptotic sufficient conditions imposed on travel time and the number of stations be made more rigorous to obtain necessary and sufficient conditions?

3. Can instance-dependent lower bounds on the performance of monitoring algorithms in unknown, dynamic environments be established?

Appendix A

Technical Supplement to the Theoretical Results in Chapter 3

A.1 Proof of Lemma 2

Proof. Differentiation of g yields $\forall x \in \mathbb{R}_+$

$$g'(x) \geq \frac{e^{-x}}{\max\left\{2, 2\sqrt{\pi x}\right\}} > 0$$

which establishes that g is monotonically increasing in its argument. Now, consider the KL divergence function $D_{\mathrm{KL}}(\mathrm{Pois}(m) \,||\, \mathrm{Pois}(k))$ which is a function of m and k. From Sec. 3.2 we know that m and k are in turn functions of $t_{i,k}$, namely, $m(t_{i,k}) := \lambda_i^u t_{i,k}$ and $k(t_{i,k}) := k(t_{i,k}) = \delta \frac{\alpha_i}{\beta_i^2}(\beta_i + t_{i,k})^2 - \alpha_i$. Taking the total derivative of D_{KL} with respect to $t_{i,k}$ yields:

$$\begin{aligned}
\frac{dD_{\mathrm{KL}}}{dt_{i,k}} &= \frac{\partial D_{\mathrm{KL}}}{\partial m}\frac{dm}{dt_{i,k}} + \frac{\partial D_{\mathrm{KL}}}{\partial k}\frac{dk}{dt_{i,k}} \\
&= \big(1 - \frac{k}{m}\big)\lambda_i^u + \ln\frac{k}{m}\frac{2\alpha_i\delta(\beta_i + t_{i,k})}{\beta_i^2} \\
&\geq \big(1 - \frac{k}{m}\big)\lambda_i^u - \big(1 - \frac{k}{m}\big)\frac{2\alpha_i\delta(\beta_i + t_{i,k})}{\beta_i^2} \\
&= \big(1 - \frac{k}{m}\big)\big(\lambda_i^u - \frac{2\alpha_i\delta(\beta_i + t_{i,k})}{\beta_i^2}\big).
\end{aligned}$$

By definition of each $t^*_{i,k}$ in π^*_k, we have that for all stations $i \in [n]$ and iterations $k \in \mathbb{N}_+$,

$$t^*_{i,k} = \frac{N_{\max}}{\hat{\lambda}_{i,k}} \geq t^{\text{low}}_{i,k}.$$

Invoking Lemma 2, we conclude that $t^*_{i,k}$ satisfies the uncertainty constraint for all stations, which establishes that condition (i) holds for the constructed policy π^*_k.

Now, to establish that (ii) holds for π^*_k, note that for any arbitrary policy π_k, we have that

$$\mathbb{E}[N_1(\pi_k)] = \cdots = \mathbb{E}[N_n(\pi_k)] \Leftrightarrow \pi_k \in \operatorname*{argmax}_{\pi} \hat{f}_{\text{bal}}(\pi).$$

In other words, π_k optimizes the objective function for balance if and only if the expected number of observations under π_k is equal for each station [38]. Now, note that for the constructed $\pi^*_k = (t^*_{1,k}, \ldots, t^*_{n,k})$, we have:

$$\hat{\lambda}_{1,k} t^*_{1,k} = N_{\max}, \hat{\lambda}_{2,k} t^*_{2,k} = N_{\max}, \ldots, \hat{\lambda}_{n,k} t^*_{n,k} = N_{\max}$$

which implies that

$$\mathbb{E}[N_1(\pi^*_k)] = N_{\max} = \cdots = \mathbb{E}[N_n(\pi^*_k)]$$

and thus we conclude that condition (ii) holds for π^*_k, i.e., $\pi^*_k \in \operatorname{argmax}_{\pi_k} \hat{f}_{\text{bal}}(\pi_k)$. $\square$

A.2 Proof of Lemma 3

Proof. From Lemma 2 we have that each $t^*_{i,k}$ is ensured to satisfy the uncertainty condition (3.4) $\forall i \in [n]$

$$\mathbb{P}\left(\left(\text{Var}(\lambda_i | X_i^{1:k}) \leq \delta \text{Var}(\lambda_i | X_i^{1:k-1}) | X_i^{1:k-1}\right)\right) > 1 - \epsilon \tag{A.1}$$

for each iteration k regardless of the events that transpire in the other iterations. Hence, the probability of satisfying this condition for k consecutive iterations is

greater than $(1-\epsilon)^k$. This implies that, with probability at least $(1-\epsilon)^k$, we have that the following chain of of inequalities holds:

$$\begin{aligned} Var\big(\lambda_i|X_i^1\big) &\leq \delta Var\big(\lambda_i\big), \\ Var\big(\lambda_i|X_i^{1:2}\big) &\leq \delta Var\big(\lambda_i|X_i^1\big) = \delta^2 Var\big(\lambda_i\big), \\ &\vdots \\ Var\big(\lambda_i|X_i^{1:k}\big) &\leq \delta Var\big(\lambda_i|X_i^{1:k-1}\big) = \delta^k Var\big(\lambda_i\big). \end{aligned}$$

□

A.3 Proof of Theorem 2

Proof. Note that by Chebyshev's inequality states the following:

$$\mathbb{P}\left(\big(|\hat{\lambda}_{i,k} - \lambda_i| < \xi|X_i^{1:k-1}\big)\right) > 1 - \frac{\mathrm{Var}(\hat{\lambda}_{i,k}|X_i^{1:k-1})}{\xi^2}.$$

In light of Corollary 1, we have that

$$\mathbb{P}\left(\big(\mathrm{Var}(\hat{\lambda}_{i,k}|X_i^{1:k-1}) \leq \delta^{k-1}\mathrm{Var}(\lambda_i)|X_i^{1:k-1}\big)\right) > (1-\epsilon)^{k-1}$$

employing this inequality and Chebyshev's inequality yields:

$$\begin{aligned} \mathbb{P}\left(\big(|\hat{\lambda}_{i,k} - \lambda_i| < \xi\big|X_i^{1:k-1}\big)\right) &> (1-\epsilon)^{k-1}\big(1 - \frac{\mathrm{Var}(\hat{\lambda}_{i,k}|X_i^{1:k-1})}{\xi^2}\big) \\ &> (1-\epsilon)^{k-1}\big(1 - \frac{\delta^{k-1}\mathrm{Var}(\lambda_i)}{\xi^2}\big) \end{aligned}$$

□

A.4 Proof of Theorem 3

Proof. Let $T = \sum_{i=1}^n t_{i,k}^*$ be the total observation time allocated by the generated policy. Then, by the optimality of policy $\pi_k^* = (t_{1,k}^*, \ldots, t_{n,k}^*)$ with respect to the rate

approximations, we have the following equalities

$$\hat{\lambda}_{1,k}t^*_{1,k} = N_{\max}, \hat{\lambda}_{2,k}t^*_{2,k} = N_{\max}, \ldots, \hat{\lambda}_{n,k}t^*_{n,k} = N_{\max}.$$

which implies that

$$\forall i \in [n] \quad t^*_{i,k} := \frac{T}{\lambda_i \sum_{l=1}^{n} \frac{1}{\lambda_l}}.$$

Now recall that the objective function pertaining to balance (3.1) is given by:

$$\hat{f}_{\text{bal}}(\pi_k) := \min_i \frac{\mathbb{E}[N_i(\pi_k)]}{\sum_{j=1}^{n} \mathbb{E}[N_j(\pi_k)]}.$$

and the optimal (maximal) value of this function is $\frac{1}{n}$. Now, using the fact that $|\hat{\lambda}_{i,k} - \lambda_i| < \xi_i$, we have the following inequalities for $\hat{f}_{\text{bal}}(\pi^*_k)$

$$\begin{aligned}
\hat{f}_{\text{bal}}(\pi^*_k) &= \frac{\min_i \lambda_i t^*_{i,k}}{\sum_{j=1}^{n} \lambda_j t^*_{j,k}} = \frac{\min_i \frac{T}{\sum_{l=1}^{n}(\lambda_l)^{-1}}}{\sum_{j=1}^{n} \frac{T}{\sum_{l=1}^{n}(\lambda_l)^{-1}}} \\
&> \frac{\frac{T}{\sum_{l=1}^{n}(\lambda_i+\xi_l)^{-1}}}{\frac{nT}{\sum_{l=1}^{n}(\lambda_l-\xi_l)^{-1}}} = \frac{\sum_{l=1}^{n}(\lambda_l-\xi_l)^{-1}}{n\sum_{l=1}^{n}(\lambda_l+\xi_l)^{-1}} \\
&= \frac{1}{n}\big(\frac{\sigma_{\min}}{\sigma_{\max}}\big)
\end{aligned}$$

with probability at least $(1-\epsilon)^{n(k-1)}\big(1 - \frac{\delta^{k-1}\text{Var}(\lambda_i)}{\xi^2}\big)^n$. □

Bibliography

[1] Soroush Alamdari, Elaheh Fata, and Stephen L Smith. Persistent monitoring in discrete environments: Minimizing the maximum weighted latency between observations. *IJRR*, 33(1):138–154, 2014.

[2] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3(Nov):397–422, 2002.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[4] Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

[5] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

[6] Jonathan Binney, Andreas Krause, and Gaurav S Sukhatme. Informative path planning for an autonomous underwater vehicle. In *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4791–4796, 2010.

[7] Bohemia Interactive. ARMA 3. `http://arma3.com/`.

[8] Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

[9] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[10] Jesus Capitan, Luis Merino, and Anibal Ollero. Decentralized cooperation of multiple uas for multi-target surveillance under uncertainties. In *ICUAS*, pages 1196–1202, 2014.

[11] Christos Cassandras, Xuchao Lin, and Xuchu Ding. An optimal control approach to the multi-agent persistent monitoring problem. *Automatic Control, IEEE Transactions on*, 58(4):947–961, 2013.

[12] Wei-pang Chin and Simeon Ntafos. Optimum watchman routes. *Information Processing Letters*, 28(1):39–44, 1988.

[13] Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambert W function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

[14] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 459–467. ACM, 2014.

[15] Güneş Erdoğan and Gilbert Laporte. The orienteering problem with variable profits. *Networks*, 61(2):104–116, 2013.

[16] Yoav Gabriely and Elon Rimon. Competitive on-line coverage of grid environments by a mobile robot. *Computational Geometry*, 24(3):197–224, 2003.

[17] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

[18] Aldy Gunawan, Hoong Chuin Lau, and Pieter Vansteenwegen. Orienteering problem: A survey of recent variants, solution approaches and applications. *European Journal of Operational Research*, 2016.

[19] Ying He and Edwin KP Chong. Sensor scheduling for target tracking in sensor networks. In *ICDC*, volume 1, pages 743–748. IEEE, 2004.

[20] Alfred O Hero III, Christopher M Kreucher, and Doron Blatt. Information theoretic approaches to sensor management. In *Foundations and applications of sensor management*, pages 33–57. Springer, 2008.

[21] Geoffrey A Hollinger, Sunav Choudhary, Parastoo Qarabaqi, Christopher Murphy, Urbashi Mitra, Gaurav S Sukhatme, Milica Stojanovic, Hanumant Singh, and Franz Hover. Underwater data collection using robotic sensor networks. *IEEE Journal on Selected Areas in Communications*, 30(5):899–911, 2012.

[22] Geoffrey A Hollinger and Gaurav S Sukhatme. Sampling-based motion planning for robotic information gathering. In *Robotics: Science and Systems*, pages 72–983. Citeseer, 2013.

[23] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *JMLR*, 11(Apr):1563–1600, 2010.

[24] Kenji Kawaguchi. Bounded optimal exploration in MDP. In *AAAI*, pages 1758–1764. AAAI Press, 2016.

[25] Xiaodong Lan and Mac Schwager. Planning periodic persistent monitoring trajectories for sensing robots in gaussian random fields. In *ICRA*, pages 2415–2420. IEEE, 2013.

[26] Nathan Michael, Ethan Stump, and Kartik Mohta. Persistent surveillance with a team of mavs. In *IROS*, 2011.

[27] Jerome Le Ny, Munther A Dahleh, Eric Feron, and Emilio Frazzoli. Continuous path planning for a data harvesting mobile server. In *ICDC*, pages 1489–1494. IEEE, 2008.

[28] P Ortner and R Auer. Logarithmic online regret bounds for undiscounted reinforcement learning. *NIPS*, 19:49, 2007.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. in Neural Inf. Proc. Sys.*, pages 91–99, 2015.

[30] Mac Schwager, Daniela Rus, and Jean-Jacques E. Slotine. Decentralized, adaptive coverage control for networked robots. *IJRR*, 28(3):357–375, 2009.

[31] Michael Short. Improved inequalities for the Poisson and binomial distribution and upper tail quantile functions. *ISRN Probability and Statistics*, 2013.

[32] Ryan N Smith, Mac Schwager, Stephen L Smith, Burton H Jones, Daniela Rus, and Gaurav S Sukhatme. Persistent ocean monitoring with underwater gliders: Adapting sampling resolution. *Journal of Field Robotics*, 28(5):714–741, 2011.

[33] Stephen L Smith, Mac Schwager, and Daniela Rus. Persistent robotic tasks: Monitoring and sweeping in changing environments. *Robotics, IEEE Transactions on*, 28(2):410–426, 2012.

[34] Daniel E. Soltero, Mac Schwager, and Daniela Rus. Generating informative paths for persistent sensing in unknown environments. In *IROS*, pages 2172–2179, 2012.

[35] Daniel E Soltero, Mac Schwager, and Daniela Rus. Decentralized path planning for coverage tasks using gradient descent adaptive control. *IJRR*, page 0278364913497241, 2013.

[36] Vaibhav Srivastava, Fabio Pasqualetti, and Francesco Bullo. Stochastic surveillance strategies for spatial quickest detection. *The International Journal of Robotics Research*, 32(12):1438–1458, 2013.

[37] Vaibhav Srivastava, Paul Reverdy, and Naomi E Leonard. Surveillance in an abruptly changing world via multiarmed bandits. In *53rd IEEE Conference on Decision and Control*, pages 692–697. IEEE, 2014.

[38] J. Yu, S. Karaman, and D. Rus. Persistent monitoring of events with stochastic arrivals at multiple stations. *IEEE Transactions on Robotics*, 31(3):521–535, 2015.

[39] Jingjin Yu, Javed Aslam, Sertac Karaman, and Daniela Rus. Anytime planning of optimal schedules for a mobile sensing robot. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 5279–5286. IEEE, 2015.

[40] Jingjin Yu, Mac Schwager, and Daniela Rus. Correlated orienteering problem and its application to informative path planning for persistent monitoring tasks. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 342–349. IEEE, 2014.