Persistent Surveillance of Stochastic Events with Unknown Statistics

Cenk Baykal

May 11, 2016

Abstract

We consider the use of a mobile agent to monitor stochastic, transient events that occur in discrete locations in the environment with the objective of maximizing the number of event observations in a balanced manner. We assume that the events of interest at each station follow a stochastic process with an initially unknown and station-specific rate parameter; Consequently, we are faced with a bandit problem -that is similar to the canonical Multi-Armed Bandit roblem- in which the inherent trade-off between exploration and exploitation must be balanced in an appropriate manner. We present a novel monitoring algorithm with provable guarantees that leverages variance estimates to generate policies capable of simultaneously taking into account the pertinent monitoring objectives and the balance between exploration and exploitation.

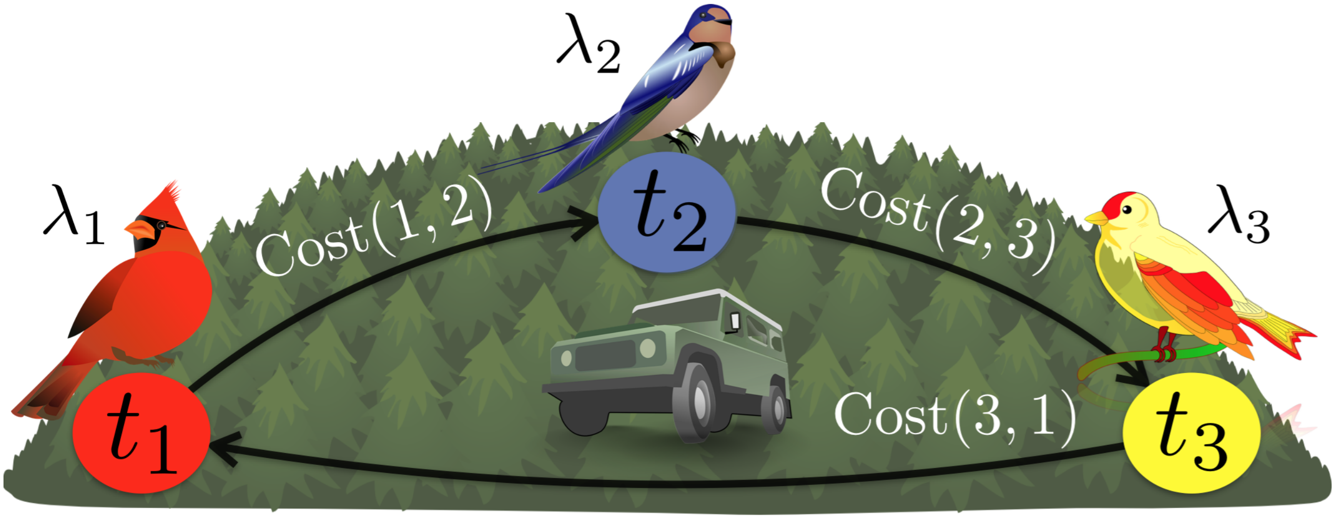

Figure 1: A persistent monitoring application in which a documentary maker would like to monitor three different species of birds is shown above. At each discrete station i, the sightings of birds (i.e. events) follow a stochastic process with rate λ_i that is initially unknown to the documentary maker and must be learned and approximated throughout the monitoring process. Given a cyclic path defining the sequence of stations to visit, the documentary maker would like to traverse this cyclic path repeatedly, stopping at each station for some defined amount of time in order to observe events. The overarching goal of the documentary maker is to generate and execute a policy $\pi := (t_1, t_2, t_3)$ that allows the documentary maker to collect the maximum number of bird sightings in a balanced manner across the different species so that no particular bird species receives too little or too much attention.

1 Introduction

We consider the problem of using a single mobile robot to monitor stochastic, transient events of interest occurring at discrete locations in the environment. We assume that events at each location follow a Poisson process with a rate parameter that is unknown prior to the monitoring process and is independent of other processes' rates located at the other stations in the environment. Since the events are stochastic and transient, their exact time of occurrence cannot be known apriori. Hence, the monitoring process requires the robot to travel from one station to the other and remain at each station for some amount of time in anticipation of events to occur.

We assume that we are given a cyclic patrolling route and seek to generate the optimal observation time to be spent at each location subject to a given optimality criteria [1, 2]. More specifically, our overarching monitoring objective is to maximize the expected number of observations made in a and simultaneously balance the monitoring effort across all stations during a cycle. An example of a monitoring task involving the monitoring of different bird species by a documentary maker is shown in Fig. 1.

The relaxed assumptions, pertaining to the knowledge of event statistics, we make in this paper are in contrast to previous problem definitions such as those in [1, 2, 3], where the statistics of events occurring at different locations – such as rate of occurrence – were assumed to be known. This implies that we are faced with the canonical exploration and exploitation trade-off, as the robot must simultaneously learn statistics about events in the environment and adjust its policy in order to optimize the pertinent monitoring objective. It is worth mentioning that the exploration and exploitation trade-off is also faced by the canonical multi-armed bandit problem [4, 5] and reinforcement learning [6].

In this paper, we introduce a novel persistent monitoring algorithm with provable guarantees that quantifies and employs the uncertainty of our rate approximations to generate policies in order to reason about and explicitly consider the inherent exploration and exploitation trade-off. The use of variance estimates enables the generation of appropriate monitoring policies that simultaneously considers the inherent exploration and exploitation trade-off and generates appropriate policies with respect to the aforementioned monitoring objectives. We present analysis proving probabilistic error bounds on the accuracy of rate approximations and the optimality of generated policies as a function of the number of the monitoring iterations. We present simulation results that compare the performance of our algorithm with that of an adaptive strategy and a state-of-the-art monitoring algorithm [2].

2 Related Work

In part due to the ubiquity of persistent monitoring tasks, the problem of persistent surveillance has been previously addressed with respect to a variety of applications and environments. For instance, in [7] the authors considered persistent surveillance of discrete locations -such as buildings, windows, doors- using a team of autonomous micro-aerial vehicles (MAVs). While UAVs are predominantly associated with persistent monitoring tasks, in [8] the authors considered the generation of monitoring policies for autonomous underwater robotic vehicles with the objective of facilitating efficient high-value data collection. Furthermore, in [2] and [9], the authors present different approaches to the min-max latency walk problem in the context of events occurring at discrete stations.

From the perspective of the persistent monitoring problem scenario that we address in this paper, our work can be seen as most similar to that of [2], where the authors also considered the monitoring of stochastic, transient events occurring in discrete locations in the environment. Nevertheless, the authors imposed the relatively strong assumption of having exact and full knowledge of the event statistics governing each stochastic process at each location prior to the monitoring process. In the context of this assumption, the authors presented a provably-optimal algorithm that generates the unique optimal policy that maximizes the balance of observations while minimizing the maximum time between two consecutive observations at each station [2].

Viewed from the perspective of sequential decision making in the context of uncertainty, there exists parallels between the monitoring problem that we consider in this paper and the canonical problem of prediction with expert advice where the best expert is unknown apriori. An even more profound relationship and similarity exists between our problem and the widely-studied Multi-Armed Bandit (MAB) problem, in which a gambler is faced with a row of K slot machines such that, pulling the lever of machine yields a stochastic reward according to a machine-specific probability distribution with a finite mean which is initially unknown [4, 5]. The overarching objective is to generate and pull the optimal lever at each discrete time step so that the regret with respect to the reward accumulated after a finite number of time is minimized.

There exist algorithms that provide regret guarantees even in the finite-horizon case for both the prediction for experts problem [10] and MAB [4, 5]. Unfortunately however, application or extension of these algorithms to the problem of persistent surveillance is rendered non-trivial due to key differences between the persistent monitoring problem that we consider and a widely-studied bandit problem such as MAB. Namely, our persistent surveillance problem exhibits a continuous state and parameter space, which is in contrast to MAB since the bandit attempts to choose the optimal lever to pull among a finite set of levers at discrete time steps, i.e. rounds. Furthermore, the monitoring problem we consider allows traveling the given cyclic path multiple times. This necessitates additional reasoning for iteration-dependent policies that consider the trade-off between the cost (i.e., wasted travel time that could otherwise be spent on observing) incurred by traveling from one station to the next and the total time that should be spent in traversing each monitoring cycle.

In contrast to all of the aforementioned prior work in the realm of persistent surveillance, we present algorithms with provable guarantees for the problem of monitoring of stochastic and transient events occurring in discrete stations in which the event statistics are unknown apriori. We employ Bayesian inference to efficiently learn, approximate, and reason about the event statistics at each station. Our algorithm explicitly quantifies and considers the uncertainties over our approximations to generate time-efficient, adaptive policies which simultaneously achieve near-optimal monitoring objective values and balance exploration and exploitation.

3 Problem Definition

Let there be $n \in \mathbb{N}_+$ stations, labeled by $i \in [n]$, whose locations are known. At each station there is a stream of events of interest, occurring at *unknown* rates. The events at each station $i \in [n]$ follow a Poisson process with an unknown rate parameter, denoted by λ_i, where the rate for each station is independent of other stations' rates. We assume that the stations are spatially distributed in the domain and hence the robot must spend a non-zero travel time $c_{i,j} \in \mathbb{R}_+$ as it travels from one station i to another station j.

We assume that we are given a cyclic path between the stations and our goal is to generate a policy stating the observation time that the robot should spend at each station to optimize an arbitrary monitoring objective. Over a monitoring period that is presumably bounded by resource constraints, a robot may traverse the given cyclic path multiple times and execute a variety of policies. We formally define a *monitoring iteration* as the complete execution of a monitoring policy and let $k \in \mathbb{N}_+$ denote each iteration. Under this terminology, a policy at iteration k, π_k, is defined as the sequence of observation times per station, i.e. $\pi_k := (t_{1,k}, t_{2,k}, \ldots, t_{n,k})$ where $t_{i,k} \in \mathbb{R}_+$ is the time to spend at each station $i \in [n]$.

To distinguish the randomness of events occurring in each iteration $k \in \mathbb{N}_+$, we let $N_i(\pi_k)$ be the random variable denoting the number of events observed at station i after having executed policy π_k. To summarize both the random number of events observed and the time spent observing at each station, we let $X_i^{(k)} := (N_i(\pi_k), t_{i,k})$ be the ordered pair of the number of events seen and the time spent observing at station i during iteration k and and let $X_i^{(1:k)} := \{X_i^1, X_i^2, \ldots, X^k\}$. After having executed k iterations, the realization of $X_i^{(1:k)}$ can be thought of as a summary of the monitoring iterations 1 through k for station $i \in [n]$.

As mentioned in Sec. 2, the majority of problems that face the exploration and exploitation trade-off, such as MAB, define the overarching optimization problem as the minimization of regret after a finite amount of time. MAB is concerned with the sole objective of maximizing the cumulative reward obtained, which is the sole objective function to be considered. However, in persistent monitoring, the overarching goal is to *simultaneously* maximize the number of events observed while maintaining a perfect balance of observations across all stations. The multi-objective problem we consider in persistent monitoring renders the definition of regret with respect to multiple objectives to be non-trivial.

The authors instead recast the problem of persistent monitoring in to an optimization in the context of a single monitoring cycle and present an alternative definition for the optimization problem. Defining the optimization problem in way that is local or greedy with respect to the respective cycle can be viewed as a heuristic for greedily generating high-quality policies which perform well in minimizing regret for both objectives when the sequence of policies is considered. In future work, the authors plan to conceive and present a monitoring objective function for which the problem of minimizing regret can easily be defined similar to the MAB definition.

Prior to presenting the optimization problem, we formalize our overarching objective functions. We let $f_{\text{obs}}(\pi_k)$ be the objective function regarding the expected number of observations made across all stations, i.e.

$$f_{\text{obs}}(\pi_k) := \sum_{i=1}^{n} \mathbb{E}[N_i(\pi_k)] \tag{1}$$

where $\mathbb{E}[N_i(\pi_k)] = \lambda_i t_{i,k}$ by definition of expectation. In order to reason about balanced attention, we formalize the notion of observation balance by letting the function $f_{\mathrm{bal}}(\pi_k)$ denote as in [2] the expected observations ratio for a given π_k which we seek to maximize,

$$f_{\mathrm{bal}}(\pi_k) := \min_i \frac{\mathbb{E}[N_i(\pi_k)]}{\sum_{j=1}^n \mathbb{E}[N_j(\pi_k)]}. \tag{2}$$

The theoretical and idealized definition of persistent surveillance is traditionally defined as an infinite-horizon problem in which the total monitoring time is unbounded. Intuitively, we expect the agent execute multiple monitoring iterations of varying time length depending on iteration-specific policies that consider past history and observations. In light of a possibly unbounded monitoring time, the two aforementioned objective functions defined above do not help establish an appropriate upper bound on the total time that should be spent per monitoring iteration. As a next step, we seek to establish an adaptive bound on the observation time for each station in a way that considers the trade-off between travel-cost and the need to execute multiple monitoring iterations so that each station can be visited more than once.

Rather than imposing an arbitrary bound on the monitoring time per cycle, we let the bound be a function of the uncertainty over the rates at each station. In what follows, we introduce a class policy optimization problems subject to the *uncertainty constraint*, a hard constraint that adaptively balances exploration and exploitation by controlling the decay of uncertainty over time.

The premise of the uncertainty constraint is to induce a rapid decrease of approximation uncertainty, which enables more accurate evaluations of prospective policies in the subsequent cycle, leading to the generation of high-quality policies within a short amount of time. More specifically, we note that the previously introduced objective functions, 1 and 2, are both functions of the (random) rates of the events at each station. Consequently, generating optimal policies requires approximate evaluation of these objective functions using the best approximations of the rates available at each iteration k, which implies that in expectation, accurate evaluations are dependent on rate approximations with low uncertainty. Transitively, we expect to generate higher-quality policies when the uncertainty pertaining our rate approximations is low in contrast to when it is high.

Let $\upsilon_i : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be a function that quantifies the uncertainty in our estimate of the rate of each station i after a certain number of iterations. Now at the beginning of each iteration $k \in \mathbb{N}_+$, having the posterior knowledge of the events that transpired in the previous $k-1$ monitoring iterations, we would like to generate a policy π_k such that our uncertainty in our approximations decreases by some factor after executing π_k, with high probability. More formally, for a given $\delta \in (0,1), \epsilon \in (0, \frac{1}{2})$, each policy π_k must satisfy the following uncertainty constraint

$$\mathbb{P}\left(\upsilon_i(k|\pi_k) \leq \delta \upsilon_i(k-1) \middle| X_i^{(1:k-1)}\right) > 1 - \epsilon \qquad \forall i \in [n] \tag{3}$$

.

In light of our over-arching goal of maximizing the number of observations made in a balanced manner across all stations, we present an optimization problem with respect to each monitoring cycle that is defined by the objective functions 1 and 2.

Per-cycle Monitoring Optimization Problem In each iteration $k \in \mathbb{N}_+$, among all policies that satisfy the uncertainty constraint (3) generate the monitoring policy π_k^* that maximizes the balance of observations, i.e.,

$$\begin{aligned} \pi_k^* \in & \operatorname*{argmax}_{\pi_k} f_{\mathrm{bal}}(\pi_k) \qquad (4) \\ & \text{s.t. } \mathbb{P}\left(\upsilon_i(k|\pi_k) \leq \delta \upsilon_i(k-1) \middle| X_i^{(1:k-1)}\right) > 1 - \epsilon \qquad \forall i \in [n]. \end{aligned}$$

The fact that the uncertainty constraint is a hard constraint implies that it is possible to trivially combine this constraint with a monitoring objective in order to generate policies that optimize a monitoring-related objective function while simultaneously balancing exploitation vs. exploration by controlling uncertainty decay.

4 Methods

In this section, we introduce a novel monitoring algorithm that generates optimal policies with respect to each optimization problem defined in Sec. 3. We describe the sub-procedure for learning and approximating event statistics using Bayesian inference, which enables the incorporation of apriori knowledge and the generation of rate approximations for each station. We outline and provide pseudo-code for generating dynamic, adaptive policies that appropriately interleave learning and approximating event statistics (exploration) with the generating and executing policies (exploitation).

4.1 Learning and Approximating Event Statistics

In this subsection, we outline the use of Bayesian inference in order to leverage prior knowledge about environments' statistics and update posterior beliefs after spending time observing events at a station. Prior to the monitoring process, we may have prior beliefs about what the rate λ_i could be for each station i. To model and incorporate any beforehand knowledge regarding the rate parameter, we use a Gamma distribution defined by the shape hyper-parameter α_i and the scale hyper-parameter β_i as the conjugate prior for the parameter λ_i. The hyper-parameters α_i and β_i will be initialized to values representing the prior beliefs, which we denote as the hyper-parameters $\alpha_{i,0}$ and $\beta_{i,0}$ and will then be updated during the monitoring process to represent our posterior beliefs given observations.

We can obtain the posterior distribution for the rate of any arbitrary station by updating the hyper-parameters α_i and β_i. Given the current values of α_i and β_i at iteration k, consider observing $n_{i,k}$ observations in $t_{i,k}$ time. Then, our posterior distribution in light of the observations $X^{(1:k)}$ is defined as:

$$\begin{aligned}\mathbb{P}\left(\lambda_i \mid X^{(1:k)}\right) &= \frac{\mathbb{P}\left(X^{(1:k)} \mid \lambda_i\right)\mathbb{P}\left(\lambda_i\right)}{\mathbb{P}\left(X^{(1:k)}\right)} \propto \mathbb{P}\left(X^{(1:k)} \mid \lambda_i\right)\mathbb{P}\left(\lambda_i\right) \\ &\propto \mathrm{Gamma}(\alpha_i + n_{i,k}, \beta_i + t_{i,k}). \qquad (5)\end{aligned}$$

where (5) follows by conjugacy.

Hence, we note that for any arbitrary number of $n_{i,k}$ events observed during $t_{i,k}$ time, the posterior update procedure simply entails updating the hyper-parameters based on their previous values and the values of $n_{i,k}$ and $t_{i,k}$, i.e. $\alpha_i \leftarrow \alpha_i + n_{i,k}$ and $\beta_i \leftarrow \beta_i + t_{i,k}$ at each iteration. More generally, after k monitoring iterations our posterior distribution is given by $\mathrm{Gamma}(\alpha_{i,0} + \eta_{i,k}, \beta_i + \tau_{i,k})$ where $\eta_{i,k} := \sum_{j=1}^{k} n_{i,j}$ and $\tau_{i,k} := \sum_{j=1}^{k} t_{i,j}$ denote the sum of observations and observation time in each station $i \in [n]$ up to iteration k respectively.

After updating, we can employ the posterior distribution to generate a refined approximation, i.e. a point estimate, of the intensity parameter for station i. Since our approximations will be iteratively changing, let $(\hat{\lambda}_i)_{k \in \mathbb{N}_+}$ denote the sequence of approximations for λ_i with respect to iteration k. For each iteration k we can leverage the fact that our updated posterior distribution is $\mathrm{Gamma}(\alpha_{i,0} + \eta_{i,k}, \beta_i + \tau_{i,k})$ and set our approximation to be the posterior mean, i.e.,

$$\hat{\lambda}_{i,k} := E[\lambda_i | X^{(1:k)}] = \frac{\alpha_{i,0} + \sum_{i=1}^{n} n_{i,k}}{\beta_{i,0} + \sum_{i=1}^{n} t_{i,k}} = \frac{\alpha_{i,0} + \eta_{i,k}}{\beta_{i,0} + \tau_{i,k}} = \frac{\alpha_i}{\beta_i}$$

which follows by definition of the Gamma distribution.

4.2 Controlling Approximation Uncertainty

We formalize the definitions of the uncertainty function and the uncertainty constraint (3) introduced in Sec. 3 and present a method to generate policies subject to the uncertainty constraint. The premise of the uncertainty approach is to enable efficient generation of high-quality policies by enforcing a controlled and rapid expected decay of the uncertainty of our approximations with high probability.

Recall from Sec. 3, at iteration k, $\upsilon_i : \mathbb{N}_+ \to \mathbb{R}_{\geq 0}$ is a function that quantifies our uncertainty in our rate estimate for the rate at station i. By the definition of the posterior update procedure, we have that after each iteration k, $\alpha_i = \alpha_{i,0} + \eta_{i,k}$ and $\beta_i = \beta_{i,0} + \tau_{i,k}$ where $\alpha_{i,0}$ and $\beta_{i,0}$ were defined to be the initial hyper-parameters of the prior. The uncertainty function evaluated after k iterations is then simply defined as the variance of the posterior distribution, i.e.,

$$\upsilon_i(k) := Var(\lambda_i | X^{(1:k)}) = \frac{\alpha_i}{\beta_i^2} = \frac{\alpha_{i,0} + \sum_{i=1}^{n} n_{i,k}}{(\beta_{i,0} + \sum_{i=1}^{n} t_{i,k})^2}. \qquad (6)$$

Under this setting, for a given policy π_k at iteration $k \in \mathbb{N}_+$ the uncertainty constraint as defined in Sec. 3 is equivalent to:

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

for all stations $i \in [n]$. We further simplify the uncertainty constraint by employing the definition of posterior variance and obtain

$$\mathbb{P}\big(\frac{\alpha_i + N_{i,k}(\pi_k)}{(\beta_i + t_{i,k})^2} \leq \delta \frac{\alpha_i}{\beta_i^2} | X_i^{(1:k-1)}\big) = \mathbb{P}\left(N_{i,k}(t_{i,k}) \leq \delta K(t_{i,k}) | X_i^{(1:k-1)}\right) > 1 - \epsilon \tag{7}$$

where $N_{i,k}(t_{i,k}) \sim \text{Poisson}(\lambda_i t_{i,k})$ and $K(t_{i,k}) := \delta \frac{\alpha_i}{\beta_i^2}(\beta_i + t_{i,k})^2 - \alpha_i$.

Generating an appropriate $t_{i,k}$ that satisfies the inequality given by (7) above requires that we reason about the possible values that the random variable $N_{i,k}(\pi_k)$ can assume. Hence, in order to make a more informed decision in generating the observation time $t_{i,k}$, we leverage the notion of a credible interval in our policy generation process. Namely, given a fixed $\epsilon \in (0, \frac{1}{2})$ and past observations $X_i^{(1:k-1)}$ at station i, we construct a credible interval for λ_i denoted by the open set $C_i(X_i^{(1:k-1)}) := (\lambda_i^{\mathrm{l}}, \lambda_i^{\mathrm{u}})$ such that:

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

for the rate parameter λ_i of each station i. We generate the end-points of the interval $C_i(X_i^{(1:k-1)})$ by employing the function $Q^{-1}(a, s)$ which denotes the inverse of the regularized Gamma function $Q(a, z)$, i.e. $Q^{-1}(a, s) = z : Q(a, z) = s$. Putting it all together, given $X_i^{(1:k-1)}$ we define the end-points of the credible interval as follows:

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

In addition, we have the property that $C_i(X_i^{(1:k-1)})$ is a symmetric (equal-tails) credible interval, which further implies

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

Now, putting it all together, given observations $X_i^{(1:k-1)}$ after having executed $k - 1$ iterations and the end points of the confidence interval λ_i^{l} and λ_i^{u}, generating an optimal observation time $t_{i,k}$ for each station i entails efficiently generating an observation time $t_{i,k}^*$ (7). As demonstrated in the analysis section (Sec. 5), an appropriate choice of $t_{i,k}^*$ is given by

$$t_{i,k}^* := t \in \mathbb{R}_+ \;\mid\; H(\lambda_i^{\mathrm{u}} t, K(t)) - \frac{1}{2} W\big(\frac{(\epsilon - 2)^2}{2\epsilon^2 \pi}\big) = 0, \tag{8}$$

where $H(m, k)$ is the Kullback-Leibler (KL) divergence between two Poisson distributed random variables with means m and k and W is the Lambert W function. An appropriate value for $t_{i,k}^*$ can be efficiently obtained by invoking a root-finding algorithm such as Brent's method on equation above.

4.3 Generating Balanced Policies that Consider Approximation Uncertainty

We extend the method of defined in the previous section so that the generated policy π_k^* simultaneously satisfies the uncertainty constraint and balances attention given to all stations in the minimum time possible. The key insight behind our approach is that if we first compute a $\pi_{\text{low}} := (t_1^{\text{low}}, \ldots, t_n^{\text{low}})$ where each t_i^{low} is defined by the expression given by Eq. (9) acts as a lower bound on each of the observation times. In other words, any observation time t_i for a particular station i that is higher than t_i^{low} given by π_{low} is ensured to satisfy uncertainty constraint by monotonicity as described in Sec. 5. Now, we can initially set $\pi_k := \pi_{\text{low}}$ to ensure that π_k and any policy with higher observation times satisfies the uncertainty constraint.

In addition to satisfying the uncertainty constraint on the observation times, we must also satisfy the balance constraint, i.e. maximize objective function 2. The idea is to generate a new policy π_k^* by increasing the observation times of π_k minimally so that π_k^* satisfies the balance and uncertainty constraints within the shortest amount of observation time. Note that π_k^* achieves the optimal balance value if and only if:

$$\mathbb{E}[N_1(\pi_k^*)] = \hat{\lambda}_{1,k} t_{1,k} = \mathbb{E}[N_2(\pi_k^*)] = \hat{\lambda}_{2,k} t_{2,k}^* = \cdots = \mathbb{E}[N_n(\pi_k^*)] = \hat{\lambda}_{n,k} t_{n,k}^*.$$

For the initial lower bound policy $\pi_k = \pi_{\text{low}} = (t_1^{\text{low}}, \ldots, t_n^{\text{low}})$ from the expression in the previous section, it may very well be the case that the above equality does not hold. However, π_k can be modified by first looking at the maximum number of expected events that needs to be matched, i.e. $N_{\max} := \max_{i \in [n]} \hat{\lambda}_i t_i^{\text{low}}$.

Now, using $N_{\max}$ we can minimally increase the observation times in π_k to generate the true optimal policy $\pi_k^* = (t_{1,k}^*, \ldots, t_{n,k}^*)$. The generation procedure for each observation time is as follows:

$$t_{i,k}^* := \frac{N_{\max}}{\hat{\lambda}_{i,k}} = N_{\max} \frac{\beta_i}{\alpha_i}.$$

5 Analysis

In this section, we present analysis proving the fact that the each iterations-specific policy generated by our algorithm described in Sec. 4 is an optimal solution with respect to the optimization problem defined in Sec. 3 with respect to the generated rate approximations $\hat{\lambda}_{i,k}, i \in [n]$ at each iteration $k \in \mathbb{N}_+$. Subsequently, we establish guarantees on the posterior variance and absolute error of our rate approximations as a function of optimization iterations. We conclude by employing the aforementioned properties to establish a bound on the quality of our generated solutions with respect to those generated by an *Oracle Algorithm* that is assumed to have perfect knowledge of the ground-truth rates.

We begin by showing that at every monitoring iteration $k \in \mathbb{N}_+$, the policy defined by the sequence of observation times, each generated according to the expression in 9 presented in Sec. 4, is optimal with respect to the rate approximations.

Lemma 1 (Satisfaction of the Uncertainty Constraint). *For a given $\epsilon \in (0, \frac{1}{2}), \delta \in (0, 1)$ at iteration $k \in \mathbb{N}$, a value of $t_{i,k}^*$ expressed by*

$$t_{i,k}^* := t \in \mathbb{R}_+ \mid H(\lambda_u(x_i)t, K(t)) - \frac{1}{2} W\big(\frac{(\epsilon - 2)^2}{2\epsilon^2 \pi}\big) = 0, \tag{9}$$

satisfies the uncertainty constraint.

Lemma 2 (Optimality of Generated Observation Times). *For all $\epsilon \in (0, \frac{1}{2}), \delta \in (0, 1)$ at iteration $k \in \mathbb{N}$, an optimal observation time for each station with respect to Problem 2 is given by*

$$t_{i,k}^* := \frac{N_{max}}{\hat{\lambda}_{i,k}} = N_{max} \frac{\beta_i}{\alpha_i}$$

where $N_{max} := \max_{i \in [n]} \hat{\lambda}_{i,k} t_{i,k}^{low}$ and $t_{i,k}^{low}$ is given by the expression in (9).

In light of the appropriateness of our choices for the observation time, we can establish further guarantees that pertain to the posterior variance and the error of our approximations.

Lemma 3 (Bound on Posterior Variance). *For any $\epsilon \in (0, \frac{1}{2}), \delta \in (0, 1)$, after $k \in \mathbb{N}_+$ iterations, the posterior variance $Var(\lambda_i | X^{(1:k)})$ is bounded above by $\delta^k Var(\lambda_i)$ with probability at least $(1 - \epsilon)^k$, i.e.,*

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

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

Corollary 4 (Bound on Approximation Variance). *For any $\epsilon \in (0, \frac{1}{2}), \delta \in (0, 1)$, after $k \in \mathbb{N}_+$ iterations, the variance of our approximation $Var\big(\hat{\lambda}_{i,k} | X^{(1:k-1)}\big)$ is bounded above by $\delta^{k-1} Var(\lambda_i)$ with probability greater than $(1 - \epsilon)^{k-1}$, i.e.,*

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

for all stations $i \in [n]$.

Theorem 5 (ξ-Bound on the Approximation Error)**.** *For any $\epsilon \in (0, \frac{1}{2}), \delta \in (0, 1)$, after $k \in \mathbb{N}_+$ iterations, for any $\xi \in \mathbb{R}_+$, our approximation $\hat{\lambda}_{i,k}$ lies within a ball of radius ξ centered at λ_i with probability at least $(1-\epsilon)^{k-1}(1 - \frac{\delta^{k-1} Var(\lambda_i)}{\xi^2})$, i.e.,*

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

for all $i \in [n]$.

Theorem 6 (Δ-Bound on Policy Optimality)**.** *For any $\xi_i \in \mathbb{R}_+$, $i \in [n]$, given that $0 < |\hat{\lambda}_{i,k} - \lambda_i| < \xi_i$ with probability as given in Theorem 5, 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 our 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 $(1-\epsilon)^{n(k-1)}\left(1 - \frac{\delta^{k-1} Var(\lambda_i)}{\xi^2}\right)^n$.*

6 Results

In this section, we present simulation results that portray the performance of our algorithm in a simulated monitoring scenario and contrast the quality of policies generated by our algorithm to that of a current state-of-the-art algorithm for persistent surveillance [2] and a dynamic algorithm that represents a naive method of generating adaptive policies from one monitoring iteration to the other. The simulation framework and the aforementioned monitoring algorithms were implemented in Python. The experiments were conducted on a MacBook Pro with one 3.1 GHz Intel Core i7 (4 cores total) processor and 16 GB of RAM. In what follows, we present the experimental environments and the respective results.

We obtained results from 10,000 trials (per algorithm) of a simulated persistent monitoring scenario involving the monitoring of events in 3 discrete stations for a monitoring period of 10 hours (600 minutes). The settings for the environment and the ground-truth rates were randomly generated by generating random variables from the following distributions for each of the three stations:

1. Prior Hyper-parameter $\alpha_{i,0} \sim \text{Uniform}(1, 20)$
2. Prior Hyper-parameter $\beta_{i,0} \sim \text{Uniform}(0.5, 1)$
3. Rate parameter $\lambda_i \sim Uniform(0.25\frac{\alpha_{i,0}}{\beta_{i,0}}), 4\frac{\alpha_{i,0}}{\beta_{i,0}})$ events per
4. Cost of travel to an adjacent station $\sim$ Uniform$(2, 5)$ minutes of travel time.

To ensure consistency and compare the algorithms in a fair manner, we incorporated the same learning and approximation procedure detailed in Sec. 4 for the other two algorithms. This integration enabled us to measure the performance of an algorithm by that operated under the assumption of known rates prior to the monitoring procedure [2].

The label and description of each algorithm along with its corresponding color in the figures is as follows:

1. Bal. Events, Min. Delay (Red): the algorithm introduced by [2] which, as mentioned in Sec. 2, assumes that the event statistics are available apriori.

2. Incremental Search, Bal. Events (Dark Blue): is an algorithm that acknowledges the presence of the exploration/exploitation trade-off and attempts to generate adaptive and lengthier policies. The algorithm initially begins with a random upper bound on the total cycle time. After each monitoring iteration, the algorithm increases the upper bound monotonically by a small random amount (with an expected increase of 5 minutes) of by generating observation times that balance expected observations subject to an arbitrary upper bound on the total cycle time.

3. Bal. Events, Min σ^2 (Cyan): our algorithm introduced in this paper that employs variance estimates to simultaneously generate policies and balance the exploration/exploitation trade-off in a near-optimal way.

We show plots of relative approximation error as a function of time, the total number of events observed on average after 10 hours of monitoring, the balance of observations with respect to all of the observations made in the 10 hour monitoring period, and the total computation time spent for generating policies during the execution of a trial. As expected, the results show that our algorithm shown in cyan in all figures is able to relatively outperform the other two evaluated algorithms with respect to every metric. Namely, from the figures we can see that our algorithm is able to efficiently generate balanced policies leading to policies capable of achieving near-optimal monitoring objective values while simultaneously inducing a rapid decline of approximation uncertainty.

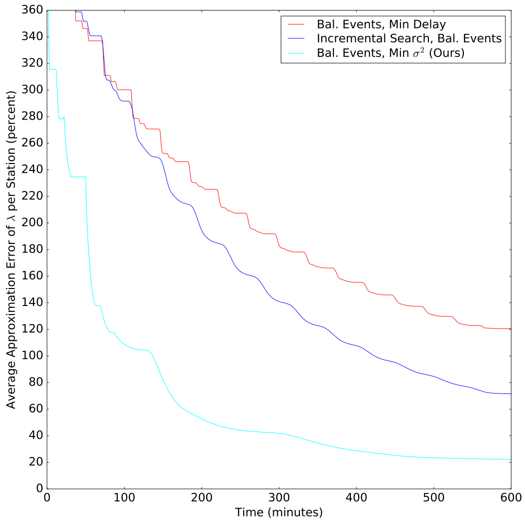

Figure 2: We present results conveying the quality of the statistics approximations as a function of monitoring time. The rapid rate of approximation error for the performance of our algorithm (cyan) supports the conjecture that our algorithm is able to generate adatpvie policies conducive to an accelerated rate of error decrease in contrast to the other algorithms' performance.

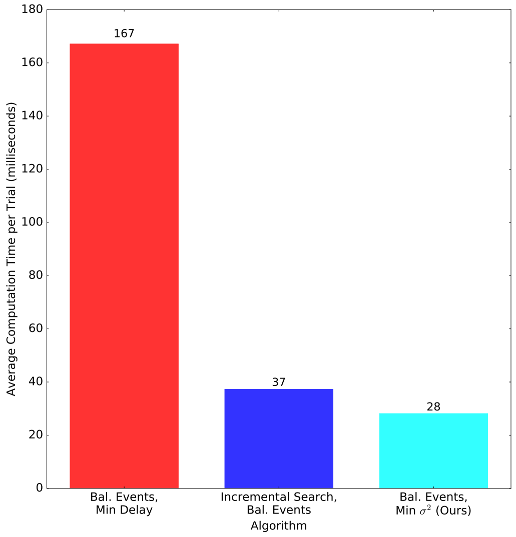

Figure 3: Simulation results comparing the computational efficiency of each algorithm measured by the total computation time spent for the generation of policies per trial.

7 Conclusion

In this paper we introduced novel algorithms and objective criteria for the task of persistent monitoring of events with statistics that are unknown a priori. Our algorithms bridged previous literature and tools pertaining persistent surveillance and machine learning in order to introduce algorithms that were able to simultaneously explore and exploit the environment with respect to a given monitoring objective. Namely, our algorithms considered maximizing the number of observations across all stations in a balanced manner while simultaneously ensuring the controlled decay of uncertainty in our rate approximations . We presented analysis showing the favorable properties of our algorithm with regard to uncertainty and policy optimality. We performed computational experiments with a diverse environment in terms of event statistics and compared our monitoring approach to the state-of-the-art. In future work we intend to relax the assumptions imposed on the events further and extend our work to dynamic and large-scale environments.

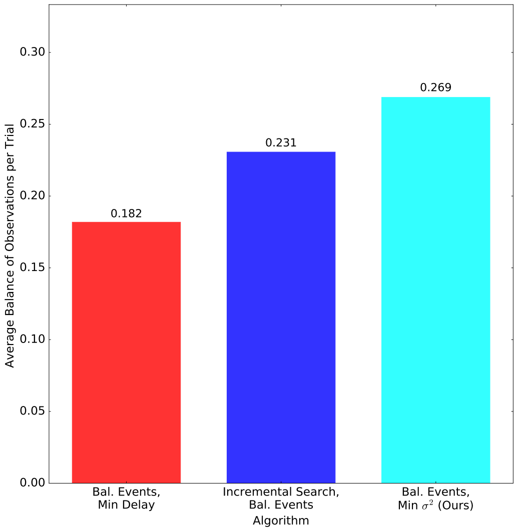

Figure 4: The performance of our monitoring algorithms with respect to the objective function pertaining to the balance of observations is shown above. We can see that when the balance of observations is considered with respect to the entire 10 hour monitoring window, the policies generated by our algorithm achieve a significantly higher objective value than do the those generated by the other two algorithms.

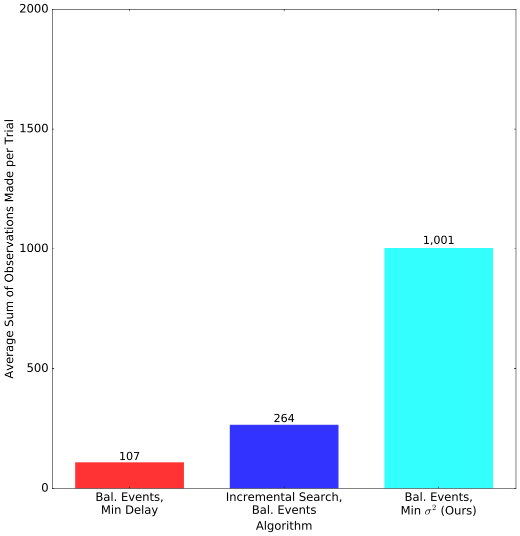

Figure 5: We present the total number of observations (across all 3 stations) that our monitoring algorithm made during the monitoring time of 600 minutes. We note that our algorithm (cyan) enables the agent to observe significantly more events than does the state of the art (red).[2]

References

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

[2] 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.

[3] Daniel E. Soltero, Mac Schwager, and Daniela Rus. Generating informative paths for persistent sensing in unknown environments. In *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots Systems (IROS)*, pages 2172–2179, 2012.

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

[5] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[6] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, pages 237–285, 1996.

[7] Nathan Michael, Ethan Stump, and Kartik Mohta. Persistent surveillance with a team of mavs. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.

[8] 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.

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

[10] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

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

8 Appendix

8.1 Technical Proofs

8.1.1 Proof of Lemma 1

Lemma 1 (Satisfaction of the Uncertainty Constraint)**.** *For a given $\epsilon \in (0, \frac{1}{2}), \delta \in (0,1)$ at iteration $k \in \mathbb{N}$, a value of $t^*_{i,k}$ expressed by*

$$t^*_{i,k} := t \in \mathbb{R}_+ \mid H(\lambda_u(x_i)t, K(t)) - \frac{1}{2} W\big(\frac{(\epsilon-2)^2}{2\epsilon^2\pi}\big) = 0, \tag{9}$$

satisfies the uncertainty constraint.

Proof. We first show that the proposed value of $t^*_{i,}$ satisfies the uncertainty condition (3) for all stations $i \in [n]$. Recall from Sec. 4 that the uncertainty constraint is equivalent to the following:

$$\mathbb{P}\left(N_{i,k}(t^*_{i,k}) \le \delta K(t^*_{i,k}) | X_i^{(1:k-1)}\right) > 1 - \epsilon \tag{10}$$

with $N_i(t^*_{i,k}) \sim \text{Poisson}(\lambda_i t^*_{i,k})$ and $K(t^*_{i,k}) := \delta \frac{\alpha_i}{\beta_i^2}(\beta_i + t^*_{i,k})^2 - \alpha_i$. We further simplify the left-hand side of (10) as follows:

$$\begin{aligned}
\mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}\right) &= \int_0^\infty \mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda\right) \mathbb{P}\left(\lambda | X_i^{(1:k-1)}\right) d\lambda \\
&> \int_0^{\lambda_\text{u}} \mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda\right) \mathbb{P}\left(\lambda | X_i^{(1:k-1)}\right) d\lambda \\
&\ge \mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_\text{u}\right) \int_0^{\lambda_\text{u}} \mathbb{P}\left(\lambda | X_i^{(1:k-1)}\right) d\lambda \\
&= (1 - \frac{\epsilon}{2}) \mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_\text{u}\right)
\end{aligned} \tag{11}$$

where we utilized the generated credible interval for λ_i and the fact that

$$\mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_\text{u}\right) = \inf_{\lambda \in (0, \lambda_\text{u})} \mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda\right)$$

to establish the inequalities.

We can further simplify the expression in (11) by establishing a lower bound for the cumulative distribution function of a Poisson random variable with mean $m(t^*_{i,k}) = \lambda_\text{u}(x_i) t^*_{i,k}$, given the value $K(t^*_{i,k})$. Using the inequality established by [11], we have that the following holds for $k \ge m$:

$$\mathbb{P}\left(N_i(t^*_{i,k}) \le k\right) > 1 - \frac{e^{-H(m,k)}}{\max\left\{2, \sqrt{4\pi H(m,k)}\right\}} \tag{12}$$

where $m := \mathbb{E}[N_i(t^*_{i,k}) | \lambda_\text{u}] = \lambda_\text{u} t^*_{i,k}$, $k = K^*_{i,k}$, and $H(m,k)$ is the Kullback-Leibler (KL) divergence between two Poisson distributed random variables with means m and k defined as

$$H(m,k) := m - k + k \ln\left(\frac{k}{m}\right).$$

We note that by definition of $t^*_{i,k}$, $H(m(t^*_{i,k}), K(t^*_{i,k})) = H^* = W\big(\frac{(\epsilon-2)^2}{2\epsilon^2\pi}\big)$, we have:

$$1 - \frac{e^{-H^*}}{\max\left\{2, \sqrt{4\pi H^*}\right\}} = 1 - \frac{\epsilon}{2-\epsilon},$$

and thus

$$\mathbb{P}\left(N_i(t^*_{i,k}) \le K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_\text{u}\right) > 1 - \frac{\epsilon}{2-\epsilon}.$$

Continuing from (11) in light of this inequality yields

$$\begin{aligned}\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k})|X_i^{(1:k-1)}\right) &> (1-\frac{\epsilon}{2})\,\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k})|X_i^{(1:k-1)}, \lambda_{\text{u}}\right) \\ &> (1-\frac{\epsilon}{2})(1-\frac{\epsilon}{2-\epsilon}) = 1-\epsilon.\end{aligned}$$

Putting it all together, we have for our choice of $t^*_{i,k}$ given by (9) that for any $\epsilon \in (0, \frac{1}{2})$

$$\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k})|X_i^{(1:k-1)}\right) > 1-\epsilon.$$

□

8.1.2 Proof of Lemma 2

Lemma 2 (Optimality of Generated Observation Times)**.** *For all $\epsilon \in (0, \frac{1}{2}), \delta \in (0,1)$ at iteration $k \in \mathbb{N}$, an optimal observation time for each station with respect to Problem 2 is given by*

$$t^*_{i,k} := \frac{N_{max}}{\hat{\lambda}_{i,k}} = N_{max}\frac{\beta_i}{\alpha_i}$$

where $N_{max} := \max_{i\in[n]} \hat{\lambda}_{i,k} t^{low}_{i,k}$ and $t^{low}_{i,k}$ is given by the expression in (9).

Proof. We argue by contradiction, suppose that there exists some $t^*_{i,k}$ that happens to not be the optimal solution to problem 4. This implies that $t^*_{i,k}$ either (i) violates the uncertainty constraint (3) or (ii) induces an unbalanced observation scheme.

We immediately see that case (i) leads to a contradiction since $t^*_{i,k}$ is defined to be bounded below by the solution to given by the expression in Eq. (9), $t^{\text{low}}_{i,k}$, hence by monotonicity of the uncertainty condition, any value greater than or equal to also satisfies the inequality given by (3). Similarly, we note that (ii) also leads to a contradiction and thus cannot occur since by definition of each $t^*_{i,k}$, 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 $\pi^*_k = (t^*_{1,k}, \ldots, t^*_{n,k})$ maximizes balance (i.e., objective function 2)

$$\mathbb{E}[N_1(\pi^*_k)] = \mathbb{E}[N_2(\pi^*_k)] = \cdots = \mathbb{E}[N_n(\pi^*_k)] \Longleftrightarrow \pi^*_k \in \operatorname*{argmax}_{\pi_k} f_{\text{bal}}(\pi_k)$$

hence, we have that (ii) leads to a contradiction. Since we have exhausted all the cases of sub-optimality, it must be the case that for all stations $i \in [n]$ and all iterations $k \in \mathbb{N}$, the value of $t^*_{i,k}$ is optimal, implying that the policy $\pi^*_k = (t^*_{1,k}, \ldots, t^*_{n,k})$ with respect to the per-cycle optimization problem.

□

8.1.3 Proof of Lemma 3

Lemma 3 (Bound on Posterior Variance)**.** *For any $\epsilon \in (0, \frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, the posterior variance $Var(\lambda_i|X^{(1:k)})$ is bounded above by $\delta^k Var(\lambda_i)$ with probability at least $(1-\epsilon)^k$, i.e.,*

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

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

Proof. From Lemma 1 we have that each $t^*_{i,k}$ is ensured to satisfy the uncertainty condition (3) $\forall i \in [n]$

$$\mathbb{P}\left(Var(\lambda_i|X^{(1:k)}) \leq \delta Var(\lambda_i|X_i^{(1:k-1)})|X_i^{(1:k-1)}\right) > 1-\epsilon \tag{13}$$

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:

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

□

8.1.4 Proof of Corollary 4

Corollary 4 (Bound on Approximation Variance)**.** *For any $\epsilon \in (0, \frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, the variance of our approximation $Var\big(\hat{\lambda}_{i,k}|X^{(1:k-1)}\big)$ is bounded above by $\delta^{k-1} Var(\lambda_i)$ with probability greater than $(1-\epsilon)^{k-1}$, i.e.,*

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

for all stations $i \in [n]$.

Proof. Employing the law of total conditional variance, we have for each $i \in [n]$

$$\begin{aligned}
Var(\lambda_i|X_i^{(1:k-1)}) &= \mathbb{E}[Var(\lambda_i|X^{(1:k)})] + Var(\mathbb{E}[\lambda_i|X^{(1:k)}]|X^{(1:k-1)}) \\
&= \mathbb{E}[Var(\lambda_i|X^{(1:k)})] + Var(\hat{\lambda}_{i,k}|X^{(1:k-1)}) \\
&\geq Var(\hat{\lambda}_{i,k}|X^{(1:k-1)})
\end{aligned}$$

Invoking Lemma 3, we have that $Var(\lambda_i|X_i^{(1:k-1)}) \leq \delta^{k-1} Var(\lambda_i)$ with probability greater than $(1-\epsilon)^{k-1}$. Combining this inequality with the above application of law of total conditional variance yields the result. □

8.1.5 Proof of Theorem 5

Theorem 5 (ξ-Bound on the Approximation Error)**.** *For any $\epsilon \in (0, \frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, for any $\xi \in \mathbb{R}_+$, our approximation $\hat{\lambda}_{i,k}$ lies within a ball of radius ξ centered at λ_i with probability at least $(1-\epsilon)^{k-1}(1 - \frac{\delta^{k-1} Var(\lambda_i)}{\xi^2})$, i.e.,*

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

for all $i \in [n]$.

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

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

In light of Corollary 4, we have that

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

employing this inequality and Chebyshev's inequality yields:

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

□

8.1.6 Proof of Theorem 6

Theorem 6 (Δ-Bound on Policy Optimality)**.** *For any $\xi_i \in \mathbb{R}_+$, $i \in [n]$, given that $0 < |\hat{\lambda}_{i,k} - \lambda_i| < \xi_i$ with probability as given in Theorem 5, 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 our 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 $(1-\epsilon)^{n(k-1)}\big(1 - \frac{\delta^{k-1} Var(\lambda_i)}{\xi^2}\big)^n$.*

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

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

Now recall that the objective function pertaining to balance (2) is given by:

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

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 π_k^*

$$\begin{aligned}
f_{\text{bal}}(\pi_k^*) &= \min_i \frac{\mathbb{E}[N_i(\pi_k^*)]}{\sum_{j=1}^n \mathbb{E}[N_j(\pi_k^*)]} \\
&= \frac{\min_i \hat{\lambda}_{i,k} t_{i,k}^*}{\sum_{j=1}^n \hat{\lambda}_{j,k} t_{j,k}^*} \\
&= \frac{\min_i \frac{T}{\sum_{l=1}^n (\hat{\lambda}_{l,k})^{-1}}}{\sum_{j=1}^n \frac{T}{\sum_{l=1}^n (\hat{\lambda}_{l,k})^{-1}}} \\
&> \frac{\frac{T}{\sum_{l=1}^n (\lambda_i + \xi_l)^{-1}}}{\frac{nT}{\sum_{l=1}^n (\lambda_l - \xi_l)^{-1}}} \\
&= \frac{\sum_{l=1}^n (\lambda_l - \xi_l)^{-1}}{n \sum_{l=1}^n (\lambda_l + \xi_l)^{-1}} \\
&= \frac{1}{n}\Big(\frac{\sigma_{\min}}{\sigma_{\max}}\Big)
\end{aligned}$$

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