53rd IEEE Conference on Decision and Control
December 15-17, 2014. Los Angeles, California, USA

Surveillance in an Abruptly Changing World via Multiarmed Bandits

Vaibhav Srivastava Paul Reverdy Naomi E. Leonard

Abstract— **We study a path planning problem in an environment that is abruptly changing due to the arrival of unknown spatial events. The objective of the path planning problem is to collect the data that is most evidential about the events. We formulate this problem as a multiarmed bandit (MAB) problem with Gaussian rewards and change points, and address the fundamental tradeoff between learning the true event (exploration), and collecting the data that is most evidential about the true event (exploitation). We extend the switching-window UCB algorithm for MAB problems with bounded rewards and change points to the context of correlated Gaussian rewards and develop the switching-window UCL (SW-UCL) algorithm. We extend the SW-UCL algorithm to an adaptive SW-UCL algorithm that utilizes statistical change detection to adapt the SW-UCL algorithm. We also develop a block SW-UCL algorithm that reduces the number of transitions among arms in the SW-UCL algorithm, and is more amenable to robotic applications.**

I. Introduction

Several robotic missions including persistent surveillance and environmental monitoring involve learning the environment while collecting mission specific data from the environment. For instance, in the context of environmental monitoring, we may want to collect as much data as possible about a particular type of algae in the ocean whose location depends on the environment. As we learn the environment (exploration), we would like to focus more and more on the region in the environment where the algae exists (exploitation). This is called the exploration-exploitation tradeoff, which is at the heart of in-situ robotic missions.

The multiarmed bandit (MAB) problems are canonical formulations of the exploration-exploitation tradeoff. In a stochastic MAB problem a set of options (arms) are given. A stochastic reward with an unknown mean is associated with each arm. A player can pick only one option at a time, and the objective of the player is to maximize the cumulative expected reward. In an MAB problem, the player needs to balance the tradeoff between learning the mean rewards at each arm (exploration), and picking the arm with maximum mean reward (exploitation). The *spatial* MAB problem, in which arms are spatially embedded, models animal and robotic foraging well [1]. In this paper, we study MAB problems in which the reward at each arm may abruptly change to another value, and we propose these MAB problems for use in modeling robotic missions like surveillance and environmental monitoring.

Surveillance and persistent monitoring problems have been studied extensively in the literature. In one of the common formulations of this problem, trajectories of the robots are planned such that (i) the information collected along the way-points is maximized, and (ii) the total distance traveled is minimized, or is within a predefined budget. Such informative path planning problems are studied in [2–11].

In the aforementioned works, the trajectories for the robots are designed to maximize the total gain in information or equivalently, the total reduction in uncertainty about the environment. However, in applications like tracking and detection of events, the objective is to collect observations that are most evidential about the event, e.g., collecting data that maximizes the likelihood of the event conditioned on the occurrence of the event. Since the occurrence of an event is not known apriori, exploration is needed to learn which event has occurred while exploitation is needed to collect the most evidential data about the event. This tradeoff can be formulated in an MAB framework as follows.

The environment can be partitioned into a set of regions according to the footprint of the sensors on the robot, and each region can be viewed as an arm. Conditioned on the occurrence of a particular event in the environment, a *feature map* of the environment can be constructed, e.g., if the event is “fire at a location”, then a possible feature map is a temperature profile across the environment, or if the event is the presence of fish at a location, then the feature map is a binary map telling if the fish are present or absent at an arm. The value of the feature map at each arm can be viewed as the reward from the arm. At each time the robot collects a noisy measurement of the reward (e.g., temperature or presence of fish) at the current arm, and the objective of the path planning algorithm is to maximize the cumulative expected reward. Note that such a path planning algorithm needs to maintain a good estimate of rewards at each arm while maximizing the number of measurements from the arm associated with the maximum reward. If the reward at each arm is identical, then this path planning problem reduces to the aforementioned informative path planning problems.

The MAB problem has been extensively studied; see [12] for a detailed review. A popular algorithm for stationary MABs with bounded rewards that achieves logarithmic (cumulative expected) regret is the UCB algorithm proposed in [13]. For the MABs with Gaussian rewards, the upper credible limit (UCL) algorithm is proposed in [14] and is shown to achieve a logarithmic regret for uninformative priors. The UCL algorithm is a variation of the Bayes-UCB algorithm proposed in [15]. We review the UCL algorithm and regret as a performance measure in Section II.

In this paper, we study the exploration-exploitation trade-

This research has been supported in part by ONR grants N00014-09-1-1074, N00014-14-1-0635, and ARO grant W911NG-11-1-0385.

V. Srivastava, P. Reverdy, and N. E. Leonard are with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA 08544 {vaibhavs, preverdy, naomi}@princeton.edu

978-1-4673-6090-6/14/$31.00 ©2014 IEEE

off in MAB problems with change points. In such problems, the mean rewards from arms are not stationary, and may abruptly change to an unknown value at some unknown time. It is assumed that the order of the number of possible changes within time T is known. We focus on MAB problems in which rewards at arms are modeled by a Gaussian process with an unknown mean and a known correlation structure.

For MAB problems with change points and bounded rewards, Garivier and Moulines [16] proposed the sliding-window UCB algorithm in which only the observations in a recent time-window are used to estimate the mean rewards and select the arms. The width of the time-window is chosen to achieve a provably efficient performance. In Section III, we extend the sliding-window UCB algorithm to the sliding-window UCL (SW-UCL) algorithm for correlated Gaussian MABs. For an uncorrelated and uninformative prior, the SW-UCL algorithm is practically identical to the frequentist algorithm in [16]; however, the Bayesian nature of the UCL algorithm and the assumption of Gaussian rewards allow us to encode the structure of the environment (feature map) through the covariance in the prior. The ability to encode the environment structure is a useful property for robotic applications.

One drawback of the sliding-window algorithms is that these algorithms do not utilize the collected data to detect changes in the mean, and accordingly, adapt the width of the time-window. In Section IV, we adopt the Page-Hinkley non-parametric change detection test to detect a change in the mean reward and adapt the width of the time-window accordingly.

The SW-UCL algorithm allows for transition among arms at each time. In the context of robotic missions, transition among arms corresponds to traveling between locations in the physical space. This is an undesirable feature if travel is costly. To remove this drawback, in Section V, we extend the SW-UCL algorithm to an algorithm that restricts the number of transitions among arms by incorporating a block allocation strategy in which the same arm is selected for a block of time instants. The block allocation strategy used in this paper is similar to the block allocation strategy used in [17, 14].

II. Preliminaries

In this section we review the MAB problem with Gaussian rewards and the UCL algorithm to solve this problem.

A. The Gaussian MAB Problem

Consider an N-armed bandit problem, i.e., a MAB problem with N arms. The reward associated with arm $i \in \{1,\dots,N\}$ is a Gaussian random variable with an unknown mean m_i, and a known variance σ_s^2. The mean of the Gaussian reward at arm i can be interpreted as the signal strength at the arm, while the variance can be interpreted as the sampling noise that is the same at each arm. Let the agent choose arm i_t at time $t \in \{1,\dots,T\}$ and receive a reward $r_t \sim \mathcal{N}(m_{i_t}, \sigma_s^2)$. The decision-maker's objective is to choose a sequence of arms $\{i_t\}_{t\in\{1,\dots,T\}}$ that maximizes the expected cumulative reward $\sum_{t=1}^{T} m_{i_t}$, where T is the horizon length of the sequential allocation process.

For an MAB problem, the expected *regret* at time t is defined by $R_t = m_{i^*} - m_{i_t}$, where $m_{i^*} = \max\{m_i \mid i \in \{1,\dots,N\}\}$. The objective of the decision-maker can be equivalently defined as minimizing the expected cumulative regret defined by $\sum_{t=1}^{T} R_t = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_i^T]$, where n_i^T is the cumulative number of times arm i has been chosen until time T and $\Delta_i = m_{i^*} - m_i$ is the expected regret due to picking arm i instead of arm i^*. It is known that the regret of any algorithm for an MAB problem is lower bounded by a logarithmic function of the horizon length T [12].

B. UCL Algorithm for Gaussian MAB Problem

The UCL algorithm proposed in [14] is a variation of the Bayes-UCB algorithm [15] and is described as follows. Let the prior distribution on the vector of mean reward at arms be a Gaussian random variable with mean $\boldsymbol{\mu}_0 \in \mathbb{R}^N$ and covariance $\Sigma_0 \in \mathbb{R}^{N\times N}$. Note that the environment structure can be encoded into the covariance Σ_0. Let $\{\boldsymbol{\phi}_t \in \mathbb{R}^N\}_{t\in\{1,\dots,T\}}$ be the indicator vector corresponding to the currently chosen arm i_t, where $(\boldsymbol{\phi}_t)_k = 1$ if $k = i_t$, and zero otherwise. Then the posterior belief about the mean rewards vector at time t is a Gaussian random variable with mean $\boldsymbol{\mu}_t \in \mathbb{R}^N$ and covariance $\Sigma_t \in \mathbb{R}^{N\times N}$ given by

$$\begin{aligned} \Lambda_t \mu_t &= r_t \boldsymbol{\phi}_t / \sigma_s^2 + \Lambda_{t-1} \boldsymbol{\mu}_{t-1} \\ \Lambda_t &= \boldsymbol{\phi}_t \boldsymbol{\phi}_t^T / \sigma_s^2 + \Lambda_{t-1},\ \Sigma_t = \Lambda_t^{-1}, \end{aligned} \tag{1}$$

where $\Lambda_t = \Sigma_t^{-1}$ is the *precision* matrix. Let μ_i^t and $(\sigma_i^t)^2$ be the posterior mean and variance of arm i at time t.

The UCL algorithm at each (discrete) time t first computes the $(1-1/Kt)$-upper credible limit Q_i^t associated with each arm $i \in \{1,\dots,N\}$ defined by

$$Q_i^t := \mu_i^t + \sigma_i^t \Phi^{-1}(1 - 1/Kt),$$

where $K = \sqrt{2\pi e}$ and $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for the standard normal random variable. The UCL algorithm then selects an arm $i_t := \arg\max\{Q_i^t \mid i \in \{1,\dots,N\}\}$. For an uninformative and uncorrelated prior, the UCL algorithm achieves a logarithmic cumulative expected regret, which is within a constant factor of the optimal.

III. Gaussian MAB Problem with Change Points: A Sliding Window Approach

We now consider the MAB problem with non-stationary Gaussian rewards. In particular, we focus on the scenario in which rewards may change abruptly at some unknown time. We assume that the number of times the rewards may change until time T is upper bounded by a known constant ζ_T. We present the SW-UCL algorithm for the Gaussian MAB problem with change points. The SW-UCL algorithm is an adaptation of the sliding-window UCB algorithm proposed in [16] to the context of correlated Gaussian rewards. The following analysis combines ideas from [16] and [14] to derive performance metrics similar to [16].

A. The SW-UCL Algorithm

For Gaussian MAB problems with change points, the SW-UCL algorithm is similar to the UCL algorithm, except that it uses only recent observations; in particular, observations collected within a fixed width time-window before the current time. For a given width of time-window t_w, the SW-UCL algorithm at time t:

(i) sets the posterior distribution of mean rewards at time $(t-t_w)^+$ to $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$, where $(\cdot)^+ = \max\{0, \cdot\}$;

(ii) performs the estimation in (1) using observations at times $\{(t-t_w)^+ + 1, \ldots, t\}$;

(iii) selects the arm i with the maximum value of

$$Q_i^{t,t_w} := \mu_i^{t,t_w} + \sigma_i^{t,t_w} \Phi^{-1}(1 - 1/K\tilde{t}_w),$$

where $\tilde{t}_w = \min\{t_w, t\}$, and μ_i^{t,t_w} and $(\sigma_i^{t,t_w})^2$ are, respectively, the posterior mean and variance of the mean reward at arm i, obtained at step (ii).

B. Regret Analysis of the SW-UCL Algorithm

We now analyze the regret associated with the SW-UCL algorithm. Let n_i^T be the number of times arm i is selected until time T when it is suboptimal. Let m_i^t be the mean reward from arm i at time t, i_t be the arm selected at time t, and i_t^* be the best arm at time t. Define $\Delta_{i,\min} = \min\{m_{i_t^*} - m_i^t \mid t \in \{1, \ldots, T\}, \text{ and } i_t^* \neq i\}$.

*Theorem 1 (**Regret of the SW-UCL algorithm**):* For the Gaussian MAB problem with change points and the SW-UCL algorithm with an uncorrelated and uninformative prior, the following statements hold:

(i) the expected number of times arm i is selected when it is suboptimal satisfies

$$\mathbb{E}[n_i^T] \leq \left\lceil \frac{T}{t_w} \right\rceil \left\lceil \frac{4\sigma_s^2 \beta^2}{\Delta_{i,\min}^2} (1 + 2\log t_w - \log 2 - \log\log t_w) \right\rceil + \zeta_T t_w + \frac{2}{K} \log t_w + \frac{2T}{K t_w},$$

where $\beta = 1.02$;

(ii) for a number of abrupt changes $\zeta_T = O(T^\nu), \nu \in [0,1)$ and $t_w = \lceil \sqrt{T \log T / \zeta_T} \rceil$,

$$\mathbb{E}[n_i^T] \leq O(T^{\frac{1+\nu}{2}} \sqrt{\log T});$$

(iii) for a number of abrupt changes $\zeta_T \leq \lambda T$, for some $\lambda \in [0,1)$, and $t_w = \lceil \sqrt{-\log \lambda / \lambda} \rceil$,

$$\mathbb{E}[n_i^T] \leq O(T \sqrt{-\lambda \log \lambda}).$$

Proof: Let n_i^{t,t_w} be the number of times a suboptimal arm i has been selected within time interval $\{(t-t_w)^+ + 1, \ldots, t\}$. For an uncorrelated and uninformative prior, Q_i^{t,t_w} reduces to

$$Q_i^{t,t_w} = \bar{m}_i^{t,t_w} + \sigma_s \Phi^{-1}(1 - 1/K\tilde{t}_w) / \sqrt{n_i^{t,t_w}},$$

where $\bar{m}_i^{t,t_w}$ is the empirical mean of the rewards collected from arm i within time-interval $\{(t-t_w)^+ + 1, \ldots, t\}$. Note that

$$\begin{aligned} n_i^T &= \sum_{t=1}^{T} \mathbf{1}(i_t = i \neq i_t^*) \\ &= \sum_{t=1}^{T} \Big(\mathbf{1}(i_t = i \neq i_t^*, n_i^{t,t_w} \leq \eta) + \mathbf{1}(i_t = i \neq i_t^*, n_i^{t,t_w} > \eta) \Big) \\ &\leq \left\lceil \frac{T}{t_w} \right\rceil \eta + \zeta_T t_w + \sum_{t \in \mathcal{T}} \mathbf{1}(i_t = i \neq i_t^*, n_i^{t,t_w} > \eta), \end{aligned} \qquad (2)$$

where the second term corresponds to windows with change-points, and $\mathcal{T} = \{t \in \{1, \ldots, T\} \mid m_i^{(t-t_w)^+ + 1} = \ldots = m_i^t, \ \forall i\}$, i.e., the set of consecutive t_w times at which the mean is constant.

We now analyze the last term in equation (2). For the set $\mathcal{T}$, the MAB problem with switching points is the same as the standard MAB problem, and it follows similar to the case of UCL algorithm [14] that

$$\begin{aligned} \mathbb{P}[i_t = i \neq i_t^*] &\leq \mathbb{P}[Q_i^{t,t_w} \geq Q_{i_t^*}^{t,t_w}] \\ &\leq 2\mathbb{P}\big[z \geq \Phi^{-1}(1 - 1/K\tilde{t}_w)\big] + \\ &\qquad \mathbb{P}\big[m_{i_t}^* \leq m_i + 2\sigma_s \Phi^{-1}(1 - 1/K\tilde{t}_w) / \sqrt{n_i^{t,t_w}}\big]. \end{aligned} \qquad (3)$$

Again, similar to the case of the UCL algorithm [14], using an upper bound on Φ^{-1}, it can be shown that the argument of the last term in equation (3) is true only if

$$n_i^{t,t_w} \leq \frac{4\sigma_s^2 \beta^2}{\Delta_{i,\min}^2} (1 + 2\log t_w - \log 2 - \log\log t_w).$$

Therefore, picking $\eta = \lceil \frac{4\sigma_s^2 \beta^2}{\Delta_{i,\min}^2} (1 + 2\log t_w - \log 2 - \log\log t_w) \rceil$, yields

$$\mathbb{E}[n_i^T] \leq \left\lceil \frac{T}{t_w} \right\rceil \left\lceil \frac{4\sigma_s^2 \beta^2}{\Delta_{i,\min}^2} (1 + 2\log t_w - \log 2 - \log\log t_w) \right\rceil + \zeta_T t_w + \frac{2}{K} \log t_w + \frac{2T}{K t_w}. \qquad (4)$$

This establishes the first statement.

The second and third statements can be verified by substituting the expressions for ζ_T and t_w in (4). ■

Theorem 1 implies that if the number of change points is a sublinear function of the time-horizon T, then the fraction of times the SW-UCL algorithm selects a suboptimal arm approaches zero as T approaches infinity. However, this fraction may not approach zero if the number of change points is a linear function of horizon length.

The convergence of the fraction of suboptimal arm selections may be very slow if the number of arms is high. In particular, for small time horizons, the width t_w of the time-window may be smaller than the number of arms, and the algorithm may not select each arm even once in each time-window. In such cases, the width of the time-window should be chosen as the maximum of N and the designed width.

IV. Gaussian MAB Problem with Change Points: An Adaptive Sliding Window Approach

A drawback of the SW-UCL algorithm studied in the previous section is that it uses a fixed width of the observation widow. This leads to all the observations within the given time-window being used to estimate the true mean of the reward from an arm, even if there is an abrupt change within the time-window. However, for a given arm, we can utilize the rewards collected within the time-window to estimate a change point, and accordingly, can use only the observations made after the change point to obtain a better estimate of the mean rewards.

In this section we use the Page-Hinkley change detection algorithm to adapt the width of the time-window in the aforementioned way, and develop the adaptive SW-UCL algorithm. In the following, we first review the Page-Hinkley change point detection algorithm and then utilize it to adaptively select the width of the sliding window.

A. The Page-Hinkley algorithm for change detection

Consider a sequence of observations $\{y_\tau\}_{\tau\in\mathbb{N}}$. Suppose the set of observations $\{y_1,\dots,y_v\}$, with $v\in\mathbb{N}$ unknown, are i.i.d. with a Gaussian distribution and unknown mean $\hat{m}_0$. Similarly, the set of observations $\{y_{v+1},\dots\}$ are i.i.d. with a Gaussian distribution with unknown mean $\hat{m}_1\neq\hat{m}_0$. Let the variance of each observation be the same. The Page-Hinkley algorithm [18] detects the change point as follows:

(i) at each time t, it maintains a running estimate of the mean $\bar{y}_t=\frac{1}{t}\sum_{\tau=1}^{t}y_\tau$;

(ii) it integrates the observations adjusted by the estimated mean: $\Lambda_t=\Lambda_{t-1}+y_t-\bar{y}_t$, with $\Lambda_0=0$;

(iii) it computes the maximum and minimum evidence

$$\Lambda_{\max}=\max\{\Lambda_\tau\mid\tau\in\{1,\dots,t\}\},\ \text{and}$$
$$\Lambda_{\min}=\min\{\Lambda_\tau\mid\tau\in\{1,\dots,t\}\};$$

(iv) it compares the current evidence with the maximum evidence to declare a decrease in mean, i.e., if $\Lambda_{\max}-\Lambda_t$ is greater than a threshold η, then a decrease in mean is declared, and the change point is picked as $t_{\text{chng}}=\operatorname{argmax}\{\Lambda_\tau\mid\tau\in\{1,\dots,t\}\}$;

(v) it compares the current evidence with the minimum evidence to declare an increase in mean, i.e., if $\Lambda_t-\Lambda_{\min}$ is greater than a threshold η, then an increase in mean is declared, and the change point is picked as $t_{\text{chng}}=\operatorname{argmin}\{\Lambda_\tau\mid\tau\in\{1,\dots,t\}\}$;

B. The adaptive SW-UCL algorithm

The adaptive SW-UCL algorithm works similar to the SW-UCL algorithm and at each time t:

(i) estimates the change point at each arm t_i^{chng}, initialized to 1;

(ii) selects the width of the time-window at arm i at time t as $\hat{t}_w^i=\min\{t_w,t-t_i^{\text{chng}}+1\}$;

(iii) performs the estimation in (1) using observations from arm i in the time-window $\{(t-\hat{t}_w^i)^++1,\dots,t\}$;

(iv) selects arm i with the maximum value of

$$Q_i^{t,\hat{t}_w^i}:=\mu_i^{t,\hat{t}_w^i}+\sigma_i^{t,\hat{t}_w^i}\Phi^{-1}(1-1/K\min\{\hat{t}_w^i,t\}),$$

where $\mu_i^{t,\hat{t}_w^i}$ and $(\sigma_i^{t,\hat{t}_w^i})^2$ are, respectively, the posterior mean and variance of the mean reward at arm i, obtained at step (iii).

The inference in step (iii) can be performed by setting the sampling variance to infinity for all observations from arm i not collected in the time-window $\{(t-\hat{t}_w^i)^++1,\dots,t\}$. The estimate of the change point t_i^{chng} at arm i is determined as follows.

(i) an empirical mean of the reward from arm i is calculated using the time-window determined using the current estimate of change point, i.e., $\hat{m}_i^t:=\sum_{\tau=t-\hat{t}_w^i+1}^{t}r_t\mathbf{1}(i_t=i)/\sum_{\tau=t-\hat{t}_w^i+1}^{t}\mathbf{1}(i_t=i)$;

(ii) mean adjusted reward $y_t=r_t-\hat{m}_{i_t}^t$ is calculated;

(iii) the Page-Hinkley statistics are calculated

$$\Lambda_t^i=\sum_{\tau=(t-\hat{t}_w^i)^++1}^{t}y_\tau\mathbf{1}(i_\tau=i)$$
$$\Lambda_{\max}^i=\max\{\Lambda_\tau^i\mid\tau\in\{(t-\hat{t}_w^i)^++1,\dots,t\}\}$$
$$\Lambda_{\min}^i=\min\{\Lambda_\tau^i\mid\tau\in\{(t-\hat{t}_w^i)^++1,\dots,t\}\}$$

(iv) if $(\Lambda_{\max}^i-\Lambda_t^i)>\eta$, then t_i^{chng} is updated to $\operatorname{argmax}\{\Lambda_\tau^i\mid\tau\in\{(t-\hat{t}_w^i)^++1,\dots,t\}\}$;

(v) if $(\Lambda_\tau^i-\Lambda_{\min}^i)>\eta$, then t_i^{chng} is updated to $\operatorname{argmin}\{\Lambda_\tau^i\mid\tau\in\{(t-\hat{t}_w^i)^++1,\dots,t\}\}$.

The performance of the Page-Hinkley test is characterized only in the asymptotic limit, and this makes the analysis of the adaptive SW-UCL algorithm hard. In this paper, we will only numerically characterize the performance of the adaptive SW-UCL algorithm.

V. Gaussian MAB Problem with Change Points: A Block Allocation Strategy

We now study a block allocation strategy for the Gaussian MAB problem with change points. The purpose of the block allocation strategy is to restrict the number of transitions among arms while maintaining a performance similar to the SW-UCL algorithm.

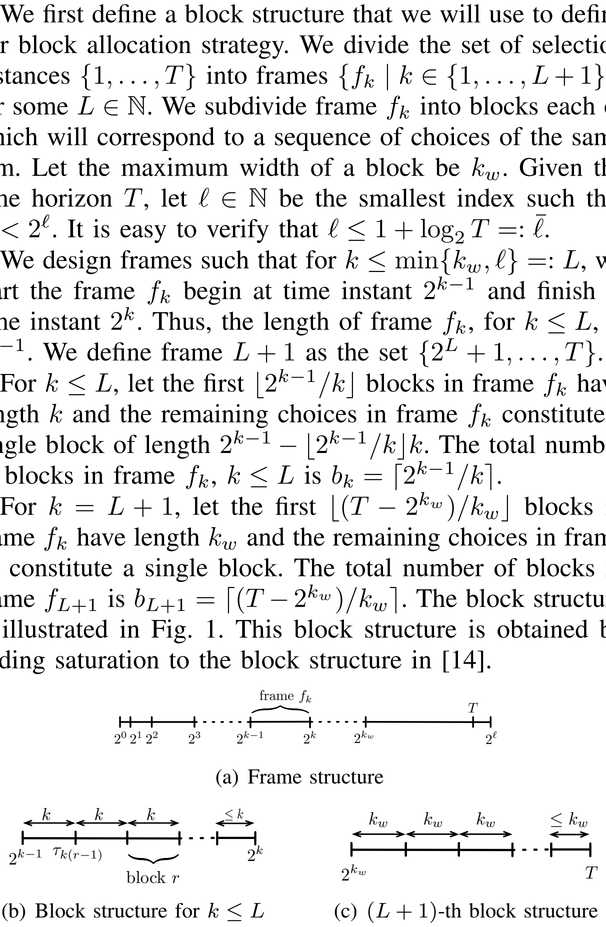
A. The block SW-UCL algorithm

We first define a block structure that we will use to define our block allocation strategy. We divide the set of selection instances $\{1,\dots,T\}$ into frames $\{f_k\mid k\in\{1,\dots,L+1\}\}$, for some $L\in\mathbb{N}$. We subdivide frame f_k into blocks each of which will correspond to a sequence of choices of the same arm. Let the maximum width of a block be k_w. Given the time horizon T, let $\ell\in\mathbb{N}$ be the smallest index such that $T<2^\ell$. It is easy to verify that $\ell\leq 1+\log_2 T=:\bar{\ell}$.

We design frames such that for $k\leq\min\{k_w,\ell\}=:L$, we start the frame f_k begin at time instant 2^{k-1} and finish at time instant 2^k. Thus, the length of frame f_k, for $k\leq L$, is 2^{k-1}. We define frame $L+1$ as the set $\{2^L+1,\dots,T\}$.

For $k\leq L$, let the first $\lfloor 2^{k-1}/k\rfloor$ blocks in frame f_k have length k and the remaining choices in frame f_k constitute a single block of length $2^{k-1}-\lfloor 2^{k-1}/k\rfloor k$. The total number of blocks in frame f_k, $k\leq L$ is $b_k=\lceil 2^{k-1}/k\rceil$.

For $k=L+1$, let the first $\lfloor(T-2^{k_w})/k_w\rfloor$ blocks in frame f_k have length k_w and the remaining choices in frame f_k constitute a single block. The total number of blocks in frame f_{L+1} is $b_{L+1}=\lceil(T-2^{k_w})/k_w\rceil$. The block structure is illustrated in Fig. 1. This block structure is obtained by adding saturation to the block structure in [14].



(a) Frame structure

(b) Block structure for $k\leq L$

(c) $(L+1)$-th block structure

Fig. 1. Block allocation strategy

Each block is characterized by the tuple (k, r), for some $k \in \{1, \ldots, L+1\}$, and $r \in \{1, \ldots, b_k\}$, where k identifies the frame and r identifies the block within the frame. We denote the time at the beginning of block (k, r) by $\tau_{kr} \in \mathbb{N}$.

The block SW-UCL algorithm at time τ_{kr}:

(i) sets the posterior distribution of mean rewards at time $(\tau_{kr} - 2^{k_w})^+$ to $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$;

(ii) performs the estimation in (1) using the observations collected in the time-window $\{(\tau_{kr} - 2^{k_w})^+ + 1, \ldots, \tau_{kr}\}$;

(iii) selects the arm i with the maximum value of

$$Q_i^{\tau_{kr}, k_w} := \mu_i^{kr, k_w} + \sigma_i^{kr, k_w} \Phi^{-1}(1 - 1/K\tilde{\tau}_{kr}^{w}),$$

for the duration of the block, where $\tilde{\tau}_{kr}^{w} = \min\{2^{kw}, \tau_{kr}\}$, μ_i^{kr,k_w} and $(\sigma_i^{kr,k_w})^2$ are, respectively, the posterior mean and variance of the mean reward at arm i, obtained at step (ii).

B. Regret analysis of the block SW-UCL algorithm

We now analyze the performance of the block SW-UCL algorithm and show that it achieves a performance similar to the performance of the SW-UCL algorithm.

*Theorem 2 (**Regret of the block SW-UCL algorithm**):* For the Gaussian MAB problem with change points and the block SW-UCL algorithm with an uncorrelated and uninformative prior, the following statements hold:

(i) the expected number of times arm i is selected when it is suboptimal satisfies

$$\mathbb{E}[n_i^T] \leq \left\lceil \frac{T}{2^{k_w}} \right\rceil \left\lceil \frac{4\beta^2 \sigma_s^2}{\Delta_{i,\min}^2}(1 + 2k_w \log 2) + \min\{k_w, \bar{\ell}\} \right\rceil + \zeta_T(2^{k_w} + k_w) + \frac{8}{K} + \frac{2\log 2}{K} \min\{k_w, \bar{\ell}\} + \frac{2k_w}{K 2^{k_w}} \max\left\{0, \left\lceil \frac{T - 2^{k_w}}{k_w} \right\rceil \right\};$$

(ii) for a number of abrupt changes $\zeta_T = O(T^\nu), \nu \in [0, 1)$ and $k_w = \lceil \frac{1}{2} \log_2 \frac{T \log T}{\zeta_T} \rceil$,

$$\mathbb{E}[n_i^T] \leq O(T^{\frac{1+\nu}{2}} \sqrt{\log T});$$

(iii) for a number of abrupt changes $\zeta_T \leq \lambda T$, for some $\lambda \in [0, 1)$, and $k_w = \lceil \frac{1}{2} \log_2 \frac{-\log \lambda}{\lambda} \rceil$,

$$\mathbb{E}[n_i^T] \leq O(T\sqrt{-\lambda \log \lambda}).$$

Proof: The proof proceeds similarly to Theorem 1. In the following, we only highlight the key differences in the proofs. Let $n_i^{kr}(k_w)$ be the number of times a suboptimal arm i has been selected within time interval $\{(\tau_{kr} - 2^{k_w})^+ + 1, \ldots, \tau_{kr}\}$. Let (k_t, r_t) be the largest tuple such that $\tau_{k_t r_t} \leq t$. We note that

$$\begin{aligned} n_i^T &= \sum_{t=1}^{T} \mathbf{1}(i_t = i \neq i_t^*) \\ &= \sum_{t=1}^{T} \Big(\mathbf{1}(i_t = i \neq i_t^*, n_i^{k_t r_t} \leq \eta) + \mathbf{1}(i_t = i \neq i_t^*, n_i^{k_t r_t} > \eta) \Big) \\ &\leq \left\lceil \frac{T}{2^{k_w}} \right\rceil (\eta + \min\{\ell, k_w\}) + \zeta_T(2^{k_w} + k_w) + \\ &\qquad + \sum_{t \in \mathcal{T}} \mathbf{1}(i_t = i \neq i_t^*, n_i^{k_t r_t}(k_w) > \eta), \end{aligned}$$

where the second term corresponds to windows with change-points, and $\mathcal{T} = \{t \in \{1, \ldots, T\} \mid m_i^{(t - 2^{k_w})^+ + 1} = \ldots = m_i^{t + k_w}, \ \forall i\}$. It follows that

$$\begin{aligned} n_i^T \leq{} & \left\lceil \frac{T}{2^{k_w}} \right\rceil (\eta + \min\{\ell, k_w\}) + \zeta_T(2^{k_w} + k_w) \\ & + \sum_{k=1}^{L+1} \sum_{r=1}^{b_k} \tilde{k} \mathbf{1}(i_{\tau_{kr}} = i \neq i_{\tau_{kr}}^*, \tau_{kr} \in \mathcal{T}, n_i^{kr}(k_w) > \eta), \end{aligned} \tag{5}$$

where $\tilde{k} = \min\{k, k_w\}$.

Following the same reasoning as in the proof of Theorem 1, we can show that the expect number of selections of suboptimal instances of arm i is upper bounded by

$$\mathbb{E}[n_i^T] \leq \left\lceil \frac{T}{2^{k_w}} \right\rceil \left\lceil \frac{4\beta^2 \sigma_s^2}{\Delta_{i,\min}^2}(1 + 2k_w \log 2) + \min\{k_w, \bar{\ell}\} \right\rceil + \zeta_T(2^{k_w} + k_w) + \frac{2}{K} \sum_{k=1}^{L} \sum_{r=1}^{b_k} \frac{k}{2^{k-1} + (r-1)k} + \frac{2k_w b_{L+1}}{K 2^{k_w}}.$$

It can be shown that

$$\sum_{k=1}^{L} \sum_{r=1}^{b_k} \frac{k}{2^{k-1} + (r-1)k} \leq 4 + (\log 2) \min\{k_w, \bar{\ell}\}.$$

Therefore, it follows that

$$\begin{aligned} \mathbb{E}[n_i^T] \leq{} & \left\lceil \frac{T}{2^{k_w}} \right\rceil \left\lceil \frac{4\beta^2 \sigma_s^2}{\Delta_{i,\min}^2}(1 + 2k_w \log 2) + \min\{k_w, \bar{\ell}\} \right\rceil \\ & + \zeta_T(2^{k_w} + k_w) + \frac{8}{K} + \frac{2\log 2}{K} \min\{k_w, \bar{\ell}\} \\ & + \frac{2k_w}{K 2^{k_w}} \max\left\{0, \left\lceil \frac{T - 2^{k_w}}{k_w} \right\rceil \right\}. \end{aligned} \tag{6}$$

This establishes the first statement.

The second and the third statement follow by substituting expressions for ζ_T and k_w in (6). ■

The convergence of the fraction of suboptimal arm selections for the block SW-UCL algorithm may be very slow if the number of arms is high. In particular, for small time horizons, the number of blocks within each time-window may be smaller than the number of arms, and the algorithm may not select each arm even once in each time-window. In such cases, the width of the maximum block width k_w should be chosen as the maximum of $\lceil \{k > 1 \mid 2^k = Nk\} \rceil$ and the designed width.

VI. Numerical Illustrations

Consider a 5×5 square grid environment. Let the centers of the cells in the grid coincide with points $(k, l) \in \{1, \ldots, 5\}^2$ in the Euclidean space. Let each cell in the grid denote a region in the environment that we identify as arms. Let arms $\{1, \ldots, 25\}$ be numbered in the lexicographically increasing order of the centers of the cells.

Suppose if an event occurs at arm i, then we define the mean reward at arm i as $m_i = 10$, and define the reward at another arm j as $m_j = m_i \exp(-0.3 d_{ij})$, where d_{ij} is the Euclidean distance between arms i and j. We pick the sampling variance $\sigma_s^2 = 1$, the mean in the prior $\mu_{0i} = 0$, and the covariance in the prior $(\Sigma_0)_{ij} = \sigma_0^2 \rho_{ij}$, for each $i, j \in \{1, \dots, 5\}$, where $\sigma_0^2 = 10$, and $\rho_{ij} = \exp(-0.3 d_{ij})$.

For a given horizon length T, we pick the number of abrupt changes as $\lfloor \sqrt{T} \rfloor$, and pick instances of abrupt changes uniformly in the set $\{1, \dots, T\}$. At each instance of abrupt change, we pick an arm uniformly in $\{1, \dots, 25\}$, and place an event at that arm. The reward profile at each arm is accordingly modified.

For the above set of parameters and a set of horizon lengths, we simulated the SW-UCL algorithm, the adaptive SW-UCL algorithm, and the block SW-UCL algorithm. For each value of the horizon length, 20 trials of each algorithm were performed. The time-window for the SW-UCL and the block SW-UCL algorithms was picked according to Theorems 1 and 2. The time window for the adaptive SW-UCL algorithm was chosen as $\lceil T/\zeta_T \rceil$.

The threshold for the Page-Hinkley algorithm for each arm was chosen equal to 10. At each time the estimate of the change-point was chosen as the maximum of the estimates of change-points at each arm. This resulted in a uniform time-window across all arms, and consequently, simplified the inference procedure.

A comparison of the performance of the three algorithms in our simulations is shown in Fig. 2. The fraction of times a suboptimal arm is selected decreases with the horizon length for each algorithm; however, its rate of decrease is much smaller for the block-SWUCL algorithm as compared to the SW-UCL and the adaptive SW-UCL algorithms.

The fraction of times a transition occurs among arms also decreases with the horizon length, and it is substantially smaller for the block SW-UCL algorithm as compared to the SW-UCL and the adaptive SW-UCL algorithms. The number of transitions for the adaptive SW-UCL algorithm is smaller as compared to the SW-UCL algorithm.

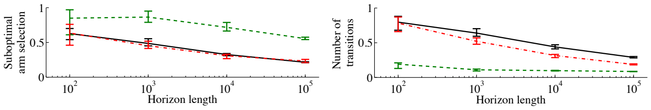


Fig. 2. Number of suboptimal arm selections and number of transitions among arms normalized with respect to the horizon length. The solid black, dashed-dotted red, and dashed green lines represent the performance of the SW-UCL, the adaptive SW-UCL, and the block SW-UCL algorithms, respectively. The error bars show the minimum and the maximum value of corresponding quantities.

VII. Conclusions

In this paper, we formulated the path planning problem in an environment with abrupt changes as an MAB problem with Gaussian rewards and change points. We extended the switching-window UCB algorithm for MAB problems with bounded rewards to the context of correlated Gaussian rewards, and developed the SW-UCL algorithm. The SW-UCL algorithm is a Bayesian algorithm that allows us to encode the environment structure through the covariance matrix in the prior. We utilized the Page-Hinkley test for change point detection to adapt the SW-UCL algorithm, and developed the adaptive SW-UCL algorithm. Finally, we incorporated a block allocation strategy in the SW-UCL algorithm to develop the block SW-UCL algorithm. The block allocation strategy restricts the number of transitions among arms in the SW-UCL algorithm, and hence, is more amenable to robotic applications.

There are several possible directions of future research. First, the policies considered in this paper involve a single robot. It is of interest to study policies for multiple robots. One way to extend the results in this paper to multiple robots is to construct an equitable partition of the environment such that the total number of abrupt changes in each partition is the same, and then use the single robot policy in each partition. It is of interest to analyze the performance of such policies, and also to consider alternative settings that require coordination strategies. Second, we assumed that in between the change points, the environment remains stationary. An interesting direction is to consider environments that evolve according to a known dynamic.

References

[1] V. Srivastava, P. Reverdy, and N. E. Leonard. Optimal foraging and multi-armed bandits. In *Allerton Conf. on Communications, Control and Computing*, pages 494–499, Monticello, IL, USA, October 2013.

[2] J. L. Ny, M. Dahleh, and E. Feron. Multi-agent task assignment in the bandit framework. In *IEEE Conference on Decision and Control*, pages 5281–5286, December 2006.

[3] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. E. Davis. Collective motion, sensor networks, and ocean sampling. *Proceedings of the IEEE*, 95(1):48–74, 2007.

[4] A. Singh, A. Krause, C. Guestrin, and W. J. Kaiser. Efficient informative sensing using multiple robots. *Journal of Artificial Intelligence Research*, 34(2):707–755, 2009.

[5] N. E. Leonard, D. A. Paley, R. E. Davis, D. M. Fratantoni, F. Lekien, and F. Zhang. Coordinated control of an underwater glider fleet in an adaptive ocean sampling field experiment in Monterey Bay. *Journal of Field Robotics*, 27(6):718–740, 2010.

[6] R. Graham and J. Cortés. Adaptive information collection by robotic sensor networks for spatial estimation. *IEEE Transactions on Automatic Control*, 57(6):1404–1419, 2012.

[7] S. L. Smith, M. Schwager, and D. Rus. Persistent robotic tasks: Monitoring and sweeping in changing environments. *IEEE Transactions on Robotics*, 28(2):410–426, 2012.

[8] D. E. Soltero, M. Schwager, and D. Rus. Generating informative paths for persistent sensing in unknown environments. In *IEEE/RSJ Int Conf on Intelligent Robots and Systems*, pages 2172–2179, Vilamoura, Algarve, Portugal, October 2012.

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

[10] G. A. Hollinger, S. Choudhary, P. Qarabaqi, C. Murphy, U. Mitra, G. S. Sukhatme, M. Stojanovic, H. Singh, and F. Hover. Underwater data collection using robotic sensor networks. *IEEE Journal on Selected Areas in Communications*, 30(5):899–911, 2012.

[11] N. Sydney and D. A. Paley. Multivehicle coverage control for a nonstationary spatiotemporal field. *Automatica*, 50(5):1381–1390, 2014.

[12] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

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

[14] P. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision making in generalized Gaussian multiarmed bandits. *Proceedings of the IEEE*, 102(4):544–571, 2014.

[15] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Int. Conf. on Artificial Intelligence and Statistics*, pages 592–600, April 2012.

[16] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[17] R. Agrawal, M. V. Hedge, and D. Teneketzis. Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.

[18] D. V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 1970.