Stochastic Bandits with Pathwise Constraints

Author

Institute

Abstract. We consider the problem of stochastic bandits, with the goal of maximizing a reward while satisfying pathwise constraints. The motivation for this problem comes from cognitive radio networks, in which agents need to choose between different transmission profiles to maximize throughput under certain operational constraints such as limited average power. Stochastic bandits serve as a natural model for an unknown, stationary environment. We propose an algorithm, based on a steering approach, and analyze its regret with respect to the optimal stationary policy that knows the statistics of the different arms.

1 Introduction

In this paper we introduce a new approach to the problem of stochastic bandits with pathwise constraints, inspired by the field of cognitive radio networks.

Cognitive Radio (CR) [11] problems consist of multi-user communication networks, occupied by primary and secondary users. Primary users have precedence over secondary users in use of network resources. Thus, secondary users must identify and exploit available resources. Through their interaction with the network, secondary users, or Cognitive Agents (CA), characterize resources and choose transmission profiles.

In [5], the Multi-Armed Bandit (MAB) framework is proposed as a model for CR problems. MABs have been widely studied in the context of balancing exploration and exploitation in sequential decision problems [4], in which an agent repeatedly chooses a single arm from a set of arms whose characteristics are unknown, and receives a certain reward based on every choice. Over time, the agent characterizes the different arms' performance in order to make well-informed decisions (exploration) while maximizing some function of the reward (exploitation). The MAB setting fits the problem of CR quite naturally: secondary users may be viewed as playing a MAB whose arms are the available transmission profiles. The problem of identifying and choosing the best arm when playing a MAB has been addressed in a series of papers [6, 1, 3, 2], using the concept of index based selection. With each time step, a number is assigned to each of the bandit's arms, reflecting the profitability of choosing it. Choosing the arm with the maximal index yields logarithmic regret with respect to always choosing the optimal arm. A simple, optimal, algorithm, which uses an upper confidence bound (UCB) for calculating the aforementioned index, is proposed in [3]. We borrow ideas from this algorithm and incorporate them into our proposed solution.

An important aspect of the CR problem, is that chosen transmission profiles must meet operational constraints, such as maximal power consumption. We suggest applying the formalism of constrained MABs to incorporate constraints into this setting. A framework for enforcing constraints in the context of online learning is proposed in

[10], in the form of a stochastic game in which a penalty is incurred, in addition to the acquired reward. Unlike the average reward the agent seeks to maximize, the average penalty ought to converge to a certain set that reflects the constraints. Taking average values is a natural choice for the CR problem since the choices CAs make are valid for short periods of time and averages converge quickly enough to serve as reliable performance measures. Since the algorithm proposed in [10] is computationally inefficient, we propose a different solution with improved convergence rates by combining the framework of MABs with the concept of steering policies, introduced in [8, 7, 12, 9].

The remainder of this paper is structured as follows. Section 2 includes a detailed formulation of the problem and states its optimal solution. Section 3 introduces an algorithm for achieving the optimal solution and theoretical results concerning it. Section 4 displays results of simulations and Section 5 concludes our work.

2 Formulation and Optimal Solution

We model the CR problem as a MAB problem with a finite time horizon and a finite number of arms, in which every arm represents a transmission profile. For simplicity, we deal with the case of a scalar reward and a single scalar constraint. At every time step, t, the agent chooses an arm, $k \in \{1..K\}$, according to a mixed policy, $\pi = (p_1, \ldots, p_K)$, that assigns probabilities to the different arms. As a result, an instantaneous reward, r_t, is received and an instantaneous penalty, c_t, is incurred. These are the agent's only source of knowledge. The reward and penalty are drawn from distributions unknown to the agent. We assume stationary Gaussian distributions:

$$r_t \sim N\left(\mu^r(k), \sigma^r(k)\right)$$
$$c_t \sim N\left(\mu^c(k), \sigma^c(k)\right).$$

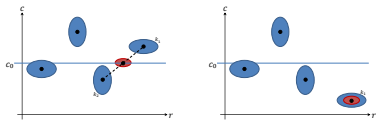
The agent seeks to maximize the acquired reward while minimizing the incurred penalty. This can be expressed as an optimization problem of the form

$$\max_{\pi} \{\hat{r}_T\} \quad s.t.\ \hat{c}_T \in \mathscr{C}. \tag{1}$$

The optimal solution to problem (1), $\pi^* = (p_1^* \ldots, p_K^*)$, depends on the characteristics of the different arms. We now introduce the concept of domination.

Definition 1. *An arm k is dominated by an arm j if $\mu^r(k) < \mu^r(j)$ and $\mu^c(k) \geq \mu^c(j)$.*

Clearly, an arm k which is dominated by one of the other arms cannot participate in the optimal solution, i.e. $p_k^* = 0$. Therefore, the optimal solution is obtained by applying a mixed policy over non-dominated arms. Specifically, if a single arm dominates all others, then the optimal policy involves this arm alone. We refer to this case as the degenerate case. Assuming a single, scalar penalty constraint, the condition $\hat{c}_t \in \mathscr{C}$ can be restated as $\hat{c}_t \leq C_0$, where C_0 denotes the maximal average penalty allowed. Fig. 1 illustrates the cases discussed.



(a) (b)

Fig. 1: (a) non-degenerate scenario; (b) a scenario with a single dominating arm (k_1). The ellipses represent distribution variances and the optimal solution is drawn in red.

Proposition 2. *In the 2-dimensional, non-degenerate case, a stationary solution of optimization problem* (1) *is*

$$(c^*, r^*) = \left(C_0, \max_{k_1 \in S_1, k_2 \in S_2} \left\{ \alpha \mu^r(k_1) + (1-\alpha)\mu^r(k_2) \right\} \right),$$

where S_1 and S_2 are sets defined by

$$S_1 = \{k \in \{1 \dots K\} : \mu^c(k) > C_0\}, \quad S_2 = \{k \in \{1 \dots K\} : \mu^c(k) \leq C_0\}.$$

The parameter $\alpha = \alpha(k_1, k_2)$ for each pair of arms $k_1 \in S_1$, $k_2 \in S_2$ is deterministic and is calculated based on knowledge of distribution parameters:

$$\alpha \mu^c(k_1) + (1-\alpha)\mu^c(k_2) = C_0, \;\; \alpha = \frac{C_0 - \mu^c(k_2)}{\mu^c(k_1) - \mu^c(k_2)}.$$

The proof of Proposition 2 is trivial. Before defining our objective, we define constraint satisfaction:

Definition 3. *A policy π is Probabilistically Constraint Satisfying (PCS) if there exist $f(t)$ and $g(t)$ such that*

$$\mathbb{P}\left\{ [\hat{c}_t - C_0]^+ < f(t) \right\} \geq 1 - g(t),$$

where $f(t) \to 0, g(t) \to 0$ as $t \to \infty$.

Next, we define a performance measure for the reward. Our definition is based on the classical notion of regret that compares the expected reward obtained by a applying a certain policy to the reward that could have been obtained by applying an optimal stationary policy with hindsight. In order to reflect the specific nature of the constrained problem, we use an adapted definition of the regret and restrict ourselves only to policies that are PCS.

Definition 4. *The expected regret for applying a certain PCS policy π when playing a constrained MAB with K arms is defined by*

$$R_t \triangleq \mu^r(p^*)t - \sum_{\tau=1}^{t} r_\tau,$$

where $\mu(p^*)$ is the expected reward of an optimal stationary PCS policy.

Our objective is to propose a policy that is PCS and minimizes this regret. We now turn to our algorithm and analyze its performance compared to the optimal solution.

3 Proposed Algorithm

In this section we suggest a steering-based approach, that attempts to reach a certain goal by adapting to changing conditions. Our policy steers the average penalty, $\hat{c}_t$, into the set $\mathscr{C}$, ensuring constraint satisfaction. It also attempts to maximize the average reward, $\hat{r}_t$. Satisfying the constraint is achieved by predicting the average penalty after the next step, $\hat{c}^p_{t+1}$, based on arm characteristics and the average penalty incurred so far. Prediction is made using an augmented form of the penalty, incorporating a version of the UCB algorithm introduced in [3]. Once the subset of constraint-satisfying arms has been determined, an arm is selected based on an augmented form of the reward. We assume Gaussian reward and penalty distributions, and implement the UCB1-NORMAL algorithm [3]. The proposed algorithm is designed for the 2-dimensional case, in which the reward and penalty are both scalar. The extension to more constraints is natural.

Algorithm 1 A steering policy incorporating UCB

loop
  if one of the arms has been sampled less than $\lceil 8\log t \rceil$ times **then**
    Sample it, or if more than one such arm - sample arm which has been sampled least.
  else
    Calculate augmented penalty and reward:
    $\bar{\mu}^c_t(k) \leftarrow \widetilde{\mu}^c_t(k) - 4\sqrt{\frac{q^c_t(k) - n_t(k)\left(\widetilde{\mu}^c_t(k)\right)^2}{n_t(k)-1}\frac{\ln(t-1)}{n_t(k)}}$
    $\bar{\mu}^r_t(k) \leftarrow \widetilde{\mu}^r_t(k) + 4\sqrt{\frac{q^r_t(k) - n_t(k)\left(\widetilde{\mu}^r_t(k)\right)^2}{n_t(k)-1}\frac{\ln(t-1)}{n_t(k)}}$
    Calculate projected average penalty for next step: $\hat{c}^p_{t+1}(k) \leftarrow {}^{\left(\bar{\mu}^c_t(k) + t\hat{c}_t\right)}/_{t+1}$
    Feasible arms are those for which $\hat{c}^p_{t+1}(k) \le C_0$; choose an arm using Algorithm 2
  end if
  Receive reward r_t and penalty c_t, calculate average reward $\hat{r}_t$ and penalty $\hat{c}_t$
  Update empirical means, $\widetilde{\mu}^r_t(k), \widetilde{\mu}^c_t(k)$ and sums of square rewards $q^r_t(k), q^c_t(k)$
end loop
Note: $n_t(k)$ is the number of times arm k was played up till time t

Our convergence results hold during almost all stages, except for an initial exploration period and stages in which forced exploration is dictated by the UCB approach: $\mathscr{T} = \left\{ t > K\lceil 8\log t\rceil \cap \mathscr{T}^C_{\text{jump}} \right\}$, where $\mathscr{T}_{\text{jump}} = \left\{ t : \lceil 8\log t\rceil < \lceil 8\log(t+i)\rceil,\ 1 \le i \le K \right\}$.

We now state and prove the main results of this paper.

Theorem 5. *For the problem of a K-armed bandit with normally distributed penalties and rewards, the steering policy described in Algorithm 1 is PCS for all $t \in \mathscr{T}$, with*

$$f(t) = \frac{1}{t}\left[|C_0| + |\mu| + \delta + \sqrt{2\sigma^2}\right]$$
$$g(t) = 3t^{-3/2} + \frac{1}{2}\frac{1}{1 - e^{-\delta^2/\left(2\sigma^2\right)}} t^{-3\delta^2/\left(2\sigma^2\right)},$$

where μ, σ are the distribution parameters of one of the arms and $\delta > 0$ is a parameter.

Algorithm 2 A procedure for optimal arm selection

1: **Input:** Set of feasible arms - S_t; vectors $\widetilde{\mu}_t^c, \bar{\mu}_t^c, \bar{\mu}_t^r$; C_0
2: **if** $S_t = \{\emptyset\}$ **then**
3: Play arm which minimizes $\bar{\mu}_t^c(k)$.
4: **else**
5: Establish set of non-dominated arms, N_t, according to Definition 1.
6: Find best match for each arm in N_t for which $\bar{\mu}_t^c(k) \leq C_0$:
Best matches are arms for which $\bar{\mu}_t^c(j) > C_0$ that minimize the slope of the line connecting arms k and j in the reward-penalty plane:

$$j = \arg\min_i \left\{ {}^{(\widetilde{\mu}_t^c(i) - \widetilde{\mu}_t^c(k))} \big/ {}_{(\bar{\mu}_t^r(i) - \bar{\mu}_t^r(k))} \right\}$$

7: Find intersections with constraint, $r_t^*(k)$, for all pairs. For single arms, take $\bar{\mu}_t^r(j)$.
8: Choose pair with maximal $r_t^*(k)$; play feasible arm with highest reward.
9: **end if**

Our proof is based on the fact that for all $t \in \mathscr{T}$, the proposed algorithm chooses the next arm to be played from a set of feasible arms, whose projected augmented penalty meets the constraint. Explicitly, the condition $\forall t \in \mathscr{T}$ is

$$\frac{\bar{\mu}_t^c(k) + t\hat{c}_t}{t+1} \leq C_0, . \tag{2}$$

We derive an upper bound on the convergence rate of the average penalty, $\hat{c}_t$, to the optimal penalty, $c^* = C_0$. We begin with a result that appears as a conjecture in [3].

Lemma 6. *Let X be a χ^2 random variable with $K \geq 2$ degrees of freedom. Then*

$$\mathbb{P}\{X \geq 4K\} \leq e^{-\frac{K+1}{2}}.$$

Proof. We derive a Chernoff bound for X. For any $\alpha > 0$ and $t > 0$,

$$\mathbb{P}\{X \geq \alpha K\} = \mathbb{P}\left\{e^{tX} \geq e^{\alpha K t}\right\} \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{\alpha K t}} = \frac{(1-2t)^{-K/2}}{e^{\alpha K t}},$$

where we use the fact that $\mathbb{E}\left[e^{tX}\right] = (1-2t)^{-K/2}$. The expression is minimized when $t = \frac{\alpha - 1}{2\alpha}$; substituting this and then substituting $\alpha = 4$ we have

$$\mathbb{P}\{X \geq 4K\} \leq 4^{K/2} e^{-3K/2} \leq e^{\frac{1}{2}K\ln 4 - K + \frac{1}{2}} e^{-\frac{K+1}{2}}.$$

Since the first factor is smaller than one for every $K \geq 2$ our proof is complete.

We now proceed to prove Theorem 5.

Proof. Our proof consists of three stages. First, we state a feasibility condition in terms of penalty. Next, we establish a bound on the convergence rate of the average penalty to the optimal penalty, by separately treating the confidence bound and the empirical mean.

Finally, we calculate an exact expression for a parametric bound on the convergence rate. The degenerate case is treated separately at the end of this section.

Stage 1 - feasibility condition: The next arm to be played must fulfill the condition

$$\frac{\bar{\mu}_t^c(k)+t\hat{c}_t}{t+1} \leq C_0 \iff \hat{c}_t - C_0 \leq \frac{C_0-\bar{\mu}_t^c(k)}{t}. \tag{3}$$

Using the definition of $\bar{\mu}_t^c(k)$ which appears in Algorithm 1,

$$\hat{c}_t - C_0 \leq \frac{1}{t}\left(C_0 - \widetilde{\mu}_t^c(k) + 4\sqrt{\frac{q_t^c(k)-n_t(k)\left(\widetilde{\mu}_t^c(k)\right)^2}{n_t(k)-1}\frac{\ln(t)}{n_t(k)}}\right).$$

Stage 2 - confidence bound convergence: As shown in [13], given $n_t(k)$, the random variable

$$X_t = \frac{1}{\left(\sigma^c(k)\right)^2}\left(q_t^c(k) - n_t(k)\left(\widetilde{\mu}_t^c(k)\right)^2\right)$$

is χ^2-distributed with $n_t(k)-1$ degrees of freedom. Thus, using Lemma 6, we have that

$$\begin{aligned}\mathbb{P}\{X_t \geq 4(n_t(k)-1)\} &= \sum_{n=\lceil 8\log t\rceil}^{\infty} \mathbb{P}[X_t \geq 4(n_t(k)-1)\,|\,n_t(k)=n]\,\mathbb{P}\{n_t(k)=n\} \\ &\leq \sum_{n=\lceil 8\log t\rceil}^{\infty} e^{-n/2} \leq 3t^{-3/2},\end{aligned}$$

where we use the fact that $n_t(k) \geq \lceil 8\log t\rceil \geq 3\ln t$ by definition of the UCB1-NORMAL algorithm. Thus, for every feasible arm, with probability greater than $1-3t^{-3/2}$,

$$\hat{c}_t - C_0 \leq \frac{1}{t}\left(C_0 - \widetilde{\mu}_t^c(k) + 4\sqrt{4\left(\sigma_k^c\right)^2\frac{\ln t}{n_t(k)}}\right).$$

Using the lower bound on $n_t(k)$ once again, with probability greater than $1-3t^{-3/2}$,

$$\hat{c}_t - C_0 \leq \frac{1}{t}\left(C_0 - \widetilde{\mu}_t^c(k) + \sqrt{2\left(\sigma^c(k)\right)^2}\right). \tag{4}$$

Stage 3 - empirical mean convergence: Using the fact that, given $n_t(k)$, $\widetilde{\mu}_t^c(k) \sim N\left(\mu^c(k), \sigma^c(k)/\sqrt{n_t(k)}\right)$, we have for any $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}\{\widetilde{\mu}_t^c(k) \geq \mu^c(k)+\varepsilon\} &= \sum_{n=\lceil 8\log t\rceil}^{\infty} \mathbb{P}[\widetilde{\mu}_t^c(k) \geq \mu^c(k)+\varepsilon\,|\,n_t(k)=n]\,\mathbb{P}\{n_t(k)=n\} \\ &\leq \frac{1}{2}\sum_{n=\lceil 8\log t\rceil}^{\infty} e^{-\frac{n\varepsilon^2}{2(\sigma^c(k))^2}} \\ &\leq \frac{1}{2}\frac{1}{1-e^{-\varepsilon^2/\left(2(\sigma^c(k))^2\right)}} t^{-3\varepsilon^2/\left(2(\sigma^c(k))^2\right)}.\end{aligned}$$

Thus, for any $\delta > 0$, we have that

$$C_0 - \widetilde{\mu}_t^c(k) \leq |C_0| + |\mu^c(k)| + \delta,$$

with probability which is greater than

$$1 - \frac{1}{2}\frac{1}{1 - e^{-\delta^2/\left(2(\sigma^c(k))^2\right)}} t^{-3\delta^2/\left(2(\sigma^c(k))^2\right)}.$$

Incorporating this into (4) and using the union bound yields

$$\hat{c}_t - C_0 \leq \frac{1}{t}\left(|C_0| + |\mu^c(k)| + \delta + \sqrt{2\left(\sigma^c(k)\right)^2}\right), \tag{5}$$

with a probability of at least

$$1 - 3t^{-3/2} - \frac{1}{2}\frac{1}{1 - e^{-\delta^2/\left(2(\sigma^c(k))^2\right)}} t^{-3\delta^2/\left(2(\sigma^c(k))^2\right)}.$$

Finally, we maximize over k in order to reflect the worst possible choice in terms of penalty. Such an event may occur, since the choice between feasible arms is made according to the reward. For the arm which maximizes the right hand side of (5) we denote $\mu^c(k) \triangleq \mu$ and $\sigma^c(k) \triangleq \sigma$. Thus, the convergence bound for the penalty of Algorithm 1 for any $\delta > 0$ is

$$\hat{c}_t - C_0 \leq \frac{1}{t}\left[|C_0| + |\mu| + \delta + \sqrt{2\sigma^2}\right],$$

with probability

$$1 - 3t^{-3/2} - \frac{1}{2}\frac{1}{1 - e^{-\delta^2/\left(2\sigma^2\right)}} t^{-3\varepsilon^2/\left(2\sigma^2\right)}.$$

Therefore, in the terms of Definition 3, we have

$$f(t) = \frac{1}{t}\left[|C_0| + |\mu| + \delta + \sqrt{2\sigma^2}\right], \quad g(t) = 3t^{-3/2} + \frac{1}{2}\frac{1}{1 - e^{-\delta^2/\left(2\sigma^2\right)}} t^{-3\varepsilon^2/\left(2\sigma^2\right)}.$$

Theorem 7. *The expected reward regret for running Algorithm 1 on K machines with normally distributed rewards and penalties, defined in Definition 4, is bounded for all $t \in \mathscr{T}$:*

$$R_t \leq 8\ln t \sum_{k=1}^{K} \Delta^r(k) + \sum_{j=1}^{\widetilde{K}} \Delta^r(p,j)\left[1 + \frac{5\pi^2}{3} + \left(256\left(\left(\frac{\sigma^r(k_1)}{\Delta_{k_1}}\right)^2 + \left(\frac{\sigma^r(k_2)}{\Delta_{k_2}}\right)^2\right)\right)\ln t\right],$$

where $\widetilde{K}$ is the number of pairs of non-dominated arms, $\Delta^r(k) \triangleq \mu^r(p^*) - \mu^r(k)$, $\Delta^r(p,j) \triangleq \mu^r(p^*) - \mu^r(p,k)$, $\mu^r(p^*)$ is the expected reward of the optimal combination of arms, $\mu^r(p,k)$ is the expected reward of the k'th pair and k_1 and k_2 are the arms which make up the k'th pair.

For our proof we assume the non-degenerate scenario, in which the optimal reward is obtained by choosing a combination of exactly two arms. We treat the degenerate scenario immediately afterwards. Before proceeding, we restate Conjecture 1 from [3].

Conjecture 1. *Let X be a Student's t-distributed random variable with s degrees of freedom. Then, for all $0 \leq a \leq \sqrt{2(s+1)}$,*

$$\mathbb{P}\{X \geq a\} \leq e^{-a^2/4}.$$

We now prove Theorem 7.

Proof. We define a MAB whose arms represent pairs of arms of the original bandit. Next we bound the expected regret in the reward sense. Since we have already proved convergence of the average penalty to the optimal penalty, once we converge to the optimal pair of arms, the correct balance between them (see Proposition 2) is guaranteed.

Stage 1: Definition of pairs-MAB We define a new MAB, whose arms represent pairs of arms of the original bandit. In general, such a bandit has $\frac{1}{2}K(K-1)$ arms, but the efficiency of the pairing process is greatly improved by Algorithm 2. Every arm is assigned an index reflecting its empirical mean reward with an upper confidence bound. The penalty of all arms (i.e., pairs) is $c(p,k) = c^* = C_0$. Denoting the reward confidence bound of a single arm by $b^r(k)$, the reward index of each pair of arms is

$$\bar{\mu}^r(p,k) = \alpha\left(\widetilde{\mu}^r(k_1) + b^r(k_1)\right) + (1-\alpha)\left(\widetilde{\mu}^r(k_2) + b^r(k_2)\right) \triangleq \widetilde{\mu}^r(p,k) + b^r(p,k),$$

where $b^r(p,k) = \alpha b^r(k_1) + (1-\alpha) b^r(k_2)$ is the confidence bound of the k'th pair and $\widetilde{\mu}^r(p,k) = \alpha\widetilde{\mu}^r(k_1) + (1-\alpha)\widetilde{\mu}^r(k_2)$ its empirical mean reward.

Stage 2: Bounding the expected regret The expected reward regret is

$$\begin{aligned} R_t &\triangleq \mu^r(p^*)t - \sum_{\tau=1}^{T} r_\tau \\ &= \sum_{k:\mu^r(p,k)<\mu^r(p^*)} \left(\mu^r(p^*) - \mu^r(p,k)\right)\mathbb{E}[n_t(p,k)] + \sum_{k=1}^{K}\left(\mu^r(p^*) - \mu^r(k)\right)\mathbb{E}[n_t(k)], \end{aligned}$$

"$*$" indicating the optimal pair. We bound the number of times every suboptimal pair of arms is sampled, $n_t(p,k)$, and the number of times every single arm is sampled, $n_t(k)$.

We examine $b^r_{t,s}(p,k)$, the reward confidence bound for the k'th pair at time t, after this pair has been sampled s times. We denote by $b^r_{t,s}(p^*)$ the same term for the optimal pair, and follow the proof of Theorem 1 of [3]. Defining the event of pair k being chosen as $\{I_t = p(k)\}$ and using the notation $\tau_0 = 1 + \lceil 8\log t\rceil$, we have for some $l \geq \lceil 8\log t\rceil$

$$\begin{aligned} n_t(p,k) &= \lceil 8\log t\rceil + \sum_{\tau=\tau_0}^{t} \mathbf{1}\{I_\tau = p(k)\} \\ &\leq l + \sum_{\tau=\tau_0}^{t} \mathbf{1}\{I_\tau = p(k), n_{\tau-1}(p,k) \geq l\} \\ &\leq l + \sum_{\tau=\tau_0}^{t} \mathbf{1}\left\{\widetilde{\mu}^r_{n_{\tau-1}}(p,k^*) + b^r_{\tau-1,n_{\tau-1}}(p,k^*) \leq \widetilde{\mu}^r_{n_{\tau-1}}(p,k) + b^r_{\tau-1,n_{\tau-1}}(p,k)\right\} \\ &\qquad \mathbf{1}\{n_{\tau-1}(p,k) \geq l\}. \end{aligned}$$

Comparing the worst case of the optimal arm with the best case of the sub-optimal arm,

$$n_t(p,k) \leq l + \sum_{\tau=\tau_0}^{t} \mathbf{1}\Big\{ \min_{0<s<\tau} \left[\widetilde{\mu}_s^r(p,k^*) + b_{\tau-1,s}^r(p,k^*)\right] \leq \max_{l<s_k<\tau} \left[\widetilde{\mu}_{s_k}^r(p,k) + b_{\tau-1,s_k}^r(p,k)\right] \Big\}$$
$$\leq l + \sum_{\tau=1}^{\infty} \sum_{s=1}^{\tau-1} \sum_{s_k=l}^{\tau-1} \mathbf{1}\left\{\widetilde{\mu}_s^r(p,k^*) + b_{\tau,s}^r(p,k^*) \leq \widetilde{\mu}_{s_k}^r(p,k) + b_{\tau,s_k}^r(p,k)\right\}.$$

Denoting

$$S \triangleq \left\{\widetilde{\mu}_s^r(p^*) + b_{\tau,s}^r(p^*) \leq \widetilde{\mu}_{s_k}^r(p,k) + b_{\tau,s_k}^r(p,k)\right\},$$
$$A \triangleq \left\{\widetilde{\mu}_s^r(p^*) \leq \mu^r(p^*) - b_{\tau,s}^r(p^*)\right\},$$
$$B \triangleq \left\{\widetilde{\mu}_{s_k}^r(p,k) \geq \mu^r(p,k) + b_{\tau,s_k}^r(p,k)\right\},$$
$$C \triangleq \left\{\mu^r(p^*) \leq \mu^r(p,k) + 2b_{\tau,s_k}^r(p,k)\right\},$$

we have that $S \subseteq A \cup B \cup C$.

Breaking up the optimal pair into two arms, we note that $A \subseteq A_1 \cup A_2$, where

$$A_1 \triangleq \left\{\widetilde{\mu}_{s_1}^r(k_1^*) \leq \mu^r(k_1^*) - b_{\tau,s_1}^r(k_1^*)\right\}$$
$$A_2 \triangleq \left\{\widetilde{\mu}_{s_2}^r(k_2^*) \leq \mu^r(k_2^*) - b_{\tau,s_2}^r(k_2^*)\right\},$$

and k_1^* and k_2^* are the arms which make up the optimal pair, p^*. We bound the probabilities of events A_1 and A_2 by following the proof of Theorem 4 in [3]. For any single arm k, the random variable $\left(\widetilde{\mu}_{s_k}^r(k) - \mu^r(k)\right) / \sqrt{\left(q_{s_k}^r - s_k\left(\widetilde{\mu}_{s_k}^r(k)\right)^2\right) / \left(s_k(s_k-1)\right)}$ has a Student's t-distribution with $s_k - 1$ degrees of freedom [13]. Combining this with Conjecture 1 using $s = s_k - 1$ and $a = 4\ln\tau$, we have for arm k_1^*, for example

$$\mathbb{P}\left\{\widetilde{\mu}_{s_1}^r(k_1^*) \leq \mu^r(k_1^*) - b_{\tau,s_1}^r(k_1^*)\right\} = \mathbb{P}\left\{\frac{\widetilde{\mu}_{s_1}^r(k_1^*) - \mu^r(k_1^*)}{\sqrt{\left(q_{s_1}^r - s_1\left(\widetilde{\mu}_{s_1}^r(k_1^*)\right)^2\right) / \left(s_1(s_1-1)\right)}} \leq 4\sqrt{\ln\tau}\right\}$$
$$\leq \tau^{-4}. \tag{6}$$

Thus, the probability of event A is bounded by applying the union bound:

$$\mathbb{P}\{A\} \leq \mathbb{P}\{A_1\} + \mathbb{P}\{A_2\} \leq 2\tau^{-4}.$$

Similarly for event B, we have

$$\mathbb{P}\{B\} \leq \mathbb{P}\{B_1\} + \mathbb{P}\{B_2\} \leq 2\tau^{-4}.$$

Finally, we address event C, which can also be rewritten as $C \subseteq C_1 \cup C_2$, where

$$C_1 \triangleq \left\{\mu^r(k_1^*) \leq \mu^r(k_1) + 2b_{\tau,s_1}^r(k_1)\right\}$$
$$C_2 \triangleq \left\{\mu^r(k_2^*) \leq \mu^r(k_2) + 2b_{\tau,s_2}^r(k_2)\right\}.$$

We examine C_1, for example.

$$\mathbb{P}[C_1|s_1=s]=\mathbb{P}\left[\left(\mu^r(k_1^*)-\mu^r(k_1)\right)^2<4\left(b^r_{\tau,s}(k_1)\right)^2\middle|s_1=s\right].$$

Denoting $\Delta_{k_1}\triangleq\mu^r(k_1^*)-\mu^r(k_1)$ and using the explicit expression for $b^r_{\tau,s}(k_1)$ yields

$$\mathbb{P}[C_1|s_1=s]=\mathbb{P}\left[\frac{q^r_s(k_1)-s\left(\widetilde{\mu}^r_s(k_1)\right)^2}{\left(\sigma^r(k_1)\right)^2}>(s-1)\frac{\Delta^2_{k_1}}{\left(\sigma^r(k_1)\right)^2}\frac{s}{64\ln t}\middle|s_1=s\right],$$

which by using Lemma 6 is bounded for $s\geq 256\left(\sigma^r(k_1)/\Delta_{k_1}\right)^2\ln\tau$:

$$\mathbb{P}[C_1|s_1=s]\leq\mathbb{P}\left[\frac{q^r_s(k_1)-s\left(\widetilde{\mu}^r_s(k_1)\right)^2}{\left(\sigma^r(k_1)\right)^2}>4(s-1)\middle|s_1=s\right]\leq e^{-s/2}.$$

Denoting $m_1\triangleq 256\left(\sigma^r(k_1)/\Delta_{k_1}\right)^2$, we calculate $\mathbb{P}\{C_1\}$:

$$\mathbb{P}\{C_1\}=\sum_{s=m_1\ln\tau}^{\infty}\mathbb{P}[C_1|s_1=s]\mathbb{P}\{s_1=s\}\leq\sum_{s=m_1\ln\tau}^{\infty}e^{-s/2}\leq 3\tau^{-m_1/2}.$$

The bound for $\mathbb{P}\{C_2\}$ is similar, and thus we have that

$$\mathbb{P}\{C\}\leq\mathbb{P}\{C_1\}+\mathbb{P}\{C_2\}\leq 3\tau^{-m_1/2}+3\tau^{-m_2/2}.$$

Using the bounds for events A,B,C we have that

$$\mathbb{P}\{S\}\leq\mathbb{P}\{A\}+\mathbb{P}\{B\}+\mathbb{P}\{C\}\leq 4\tau^{-4}+3\tau^{-m_1/2}+3\tau^{-m_2/2}.$$

Thus, the expected number of times a suboptimal pair of arms is sampled is bounded

$$\mathbb{E}[n_t(p,k)]\leq\lceil m_p\ln t\rceil+\sum_{\tau=1}^{\infty}\sum_{s=1}^{\tau}\sum_{s_i=l}^{\tau}\left(4\tau^{-4}+6\tau^{-m_p/2}\right),$$

where $m_p=\max\{8,\min\{m_1,m_2\}\}$. Since $m_p\geq 8$, we have

$$\mathbb{E}[n_t(p,k)]\leq\left(8+256\left(\left(\frac{\sigma^r(k_1)}{\Delta_{k_1}}\right)^2+\left(\frac{\sigma^r(k_2)}{\Delta_{k_2}}\right)^2\right)\right)\ln t+\frac{5\pi^2}{3}+1.$$

This expression bounds the mean number of times pairs of non-dominated arms are sampled. *All* arms are sampled at least $\lceil 8\ln t\rceil$ times, making the bound for the expected reward-regret

$$R_t\leq 8\ln t\sum_{k=1}^{K}\Delta^r(k)+\sum_{j=1}^{\widetilde{K}}\Delta^r(p,j)\left[1+\frac{5\pi^2}{3}+\left(256\left(\left(\frac{\sigma^r(k_1)}{\Delta_{k_1}}\right)^2+\left(\frac{\sigma^r(k_2)}{\Delta_{k_2}}\right)^2\right)\right)\ln t\right],$$

where $\widetilde{K}$ is the number of pairs of non-dominated arms. $\Delta^r(k)\triangleq\mu^r(p^*)-\mu^r(k)$, and $\Delta^r(p,j)\triangleq\mu^r(p^*)-\mu^r(p,k)$.

Remark: In the degenerate case, in which there is a single dominating arm, the optimal solution is to sample it alone. Thus, $(c^*, r^*) = (\mu^c(k^*), \mu^r(k^*))$. Algorithm 1 always treats the selection of the next arm to be played in a pairwise manner; a single dominating arm is paired with itself, and the expected reward regret is bounded as stated in Theorem 7. The penalty aspect, however, is a bit different. The structure of Algorithm 1 allows exploration, based on confidence bounds, as long as the penalty constraint is met (in our case, as long as $\hat{c}_t \leq C_0$). Therefore, the average penalty incurred converges to the constraint C_0 linearly (as shown in Theorem 5) and then continues to converge towards the optimal penalty, $\mu^c(k^*)$, at a logarithmic rate which is the convergence rate of the procedure for optimal arm selection, bounded in Theorem 7.

4 Simulations

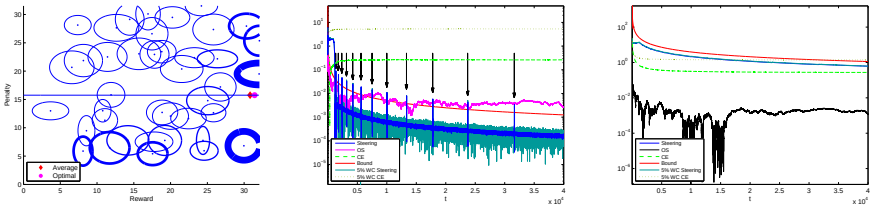
We demonstrate our results using computer simulations of a CR problem. In our scenario, the CA repeatedly has to choose one of 40 possible transmission profiles that have unknown Gaussian reward and penalty distributions. The CA interacts with the system, implementing Algorithm 1. For reference, we implement two algorithms: an ideal one that applies an optimal stationary (OS) policy, based on full knowledge of arm characteristics, and another that applies a certainty equivalence (CE) approach, updating its estimate of the solution based on the empirical means of the reward and penalty.

The results, averaged over 100 repetitions, are presented in Fig. 2. Fig. 2a displays the problem layout in the reward-penalty plane. The ellipses represent arm distributions, with their mean values and variances. The thickness of ellipse contours represents the number of times an arm was sampled, and the optimal solution and the average performance of our algorithm are annotated. Fig. 2b displays the convergence of the average penalty to the optimal penalty, together with the bound derived in Theorem 5 and with the reference policies described above. The times during which exploration overrides the penalty constraint, defined by $t \notin \mathcal{T}$, are annotated by arrows.We also display the convergence of the average of the worst 5% of the runs, where the advantage of the steering policy is clear. Finally, we present the convergence of the average reward to the optimal value. We compare our steering algorithm to the optimal mixed policy, to the certainty equivalence policy and to the theoretical bound derived in Theorem 7. As expected, we pay for the steering policy's strict adherence to the constraint in terms of reward convergence. However, reward convergence is identical in the average and worst case scenarios, unlike that of the certainty equivalence approach.

5 Conclusions and Future Work

We introduced a formulation of the CR problem using stochastic MABs with pathwise constraints. In order to solve this problem, we proposed a steering policy which results in convergence of the average reward and penalty to their optimal values.

Future directions include examining the proposed formulation from a multiple agent point of view, in order to understand issues of cooperation and competition in this setting. We also plan to examine the issue of bandits with correlated arms, in which the distributions of sub-groups of arms are not independent. These may provide a realistic



(a) Problem layout (b) Penalty convergence (c) Reward convergence

Fig. 2: Simulation results

model for closely related transmission profiles. Finally, we hope to be able to apply our framework to real-world data. An extension of our work to the case of bounded reward and penalty distributions, using the UCB1 algorithm [3], is straightforward.

References

[1] R. Agrawal. Sample mean based index policies with O (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

[2] J.Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165. Springer, 2007.

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

[4] D.A. Berry and B. Fristedt. *Bandit problems: sequential allocation of experiments*. Chapman and Hall London, 1985.

[5] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *Signals, Circuits and Systems (SCS), 2009 3rd International Conference on*, pages 1–6. IEEE, 2010.

[6] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[7] D.J. Ma and A.M. Makowski. A class of steering policies under a recurrence condition. In *Decision and Control, 1988., Proceedings of the 27th IEEE Conference on*, pages 1192–1197. IEEE, 1988.

[8] A.M. Makowski and A. Shwartz. Implementation issues for Markov decision processes. *TR 1986-63*, 1986.

[9] S. Mannor and N. Shimkin. A geometric approach to multi-criterion reinforcement learning. *The Journal of Machine Learning Research*, 5:325–360, 2004.

[10] S. Mannor, J.N. Tsitsiklis, and J.Y. Yu. Online learning with sample path constraints. *The Journal of Machine Learning Research*, 10:569–590, 2009.

[11] J. Mitola and G.Q. Maguire. Cognitive radio: making software radios more personal. *Personal Communications, IEEE*, 6(4):13 –18, August 1999.

[12] K.W. Ross. Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Operations Research*, 37(3):pp. 474–477, 1989.

[13] S. S. Wilks. *Mathematical statistics*. Wiley, 1962.