2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

# Stochastic Bandits with Pathwise Constraints

Orly Avner and Shie Mannor

*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.**

## I. INTRODUCTION

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

### A. Cognitive Radio Networks

The term Cognitive Radio (CR), first introduced in [11], refers to a challenging field in the world of communications. Basic CR problems consist of multimedia, 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 face the challenge of identifying and exploiting available resources. Through their interaction with the network, secondary users, also called Cognitive Agents (CAs), characterize available resources and choose adequate transmission profiles. The complex dynamic and stochastic nature of these problems, combined with issues of partial observability, give rise to questions of sensing, estimation and action selection.

### B. The Multi-armed Bandit Framework

In [5], the Multi-Armed Bandit (MAB) framework is proposed as a model for CR problems. A simple instance of Markov Decision Processes (MDP), MABs have been widely studied in the context of balancing exploration and exploitation in sequential decision problems [4]. These problems comprise an agent repeatedly choosing a single arm from a set of arms whose characteristics are unknown, and receiving 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, attempting to make the best possible use of an unknown communication network, 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 [1]–[3], [6], all based on the concept of index based selection. With each time step, the algorithms proposed in these papers assign a number 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.

### C. Constrained Problems

An important aspect of the CR problem is that the transmission profiles chosen by the CA must meet certain operational constraints, such as, for example, maximal power consumption. Despite being an integral part of CR problems, this property has not been taken into account so far in the CR-MAB setting. We suggest applying the formalism of constrained MABs (or constrained MDPs, in general) in order to incorporate constraints into this setting. A framework that enables the incorporation of constraints into online learning problems is proposed in [10]. This framework introduces a stochastic game in which a penalty is incurred, in addition to the traditional notion of acquiring a 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.

### D. Steering Algorithms

The concept of steering policies in the context of average performance measures was introduced by [7], [8], [12]. Later, [9] suggested a class of policies based on steering for multi-criterion reinforcement problems. However, this approach requires knowledge of the environment and has problematic convergence rates. In order to solve the constrained MAB problem, we combine the framework of MABs with the concept of steering policies.

The remainder of this paper is structured as follows. Section II includes a detailed formulation of the problem and states its optimal solution. Section III introduces an algorithm

O. Avner and S. Mannor are with the Faculty of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel

This research was partially supported by the CORNET consortium (www.cornet.org.il).

for achieving the optimal solution and theoretical results concerning it. Section IV displays results of simulations and Section V concludes our work.

## II. FORMULATION AND OPTIMAL SOLUTION

We model the CR problem as a MAB problem: every transmission profile is represented by a single arm. We assume a finite time horizon and a finite number of arms. For simplicity, we deal with the case of a scalar reward and a single scalar constraint. We formulate our problem as follows. 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 of this choice, 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 that are unknown to the agent. We assume stationary Gaussian distributions:

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

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

$$\max_{\pi} \{W_T\} \quad (1)$$
$$s.t. \sum_{k=1}^{K} p_k = 1, \;\; p_k \leq 1 \quad k = 1, \ldots, K,$$

where $W_T = \mathbb{E}_\pi[\hat{r}_T - \lambda L(\hat{c}_T, \mathscr{C})]$; $\hat{r}_T \triangleq \frac{1}{T}\sum_{\tau=0}^{T-1} r_\tau$ and $\hat{c}_T \triangleq \frac{1}{T}\sum_{\tau=0}^{T-1} c_\tau$ are the average reward and average penalty accumulated up till time $T$, respectively, $L(\hat{c}_t, \mathscr{C})$ is a loss function with respect to a predetermined set of constraints $\mathscr{C}$, and $\lambda$ is a weight factor. We assume the loss function to be of the form:

$$L(\hat{c}_t, \mathscr{C}) = \begin{cases} 0 & \hat{c}_t \in \mathscr{C} \\ f(\hat{c}_t, \mathscr{C}) & \hat{c}_t \notin \mathscr{C}, \end{cases}$$

where $f$ is a function that is monotone in a chosen measure of distance between $\hat{c}_t$ and $\mathscr{C}$. For example, $f$ may be the set-to-point distance. Solving problem (1) proves to be very difficult even for a simple case in which $K = 2$ with linear or square loss. The problem is neither convex nor concave and cannot be solved analytically.

We therefore take a different approach, and treat the penalty constraint as a hard constraint. Thus, the optimization problem becomes (we omit the constraints on $p_k$ for clarity)

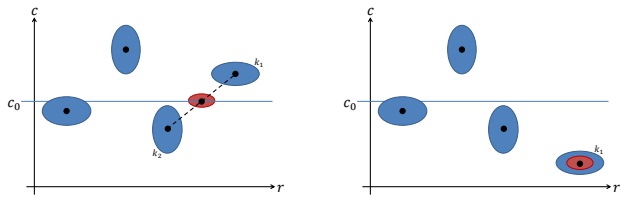
$$\max_{\pi} \{\hat{r}_T\} \quad (2)$$
$$s.t. \; \hat{c}_t \in \mathscr{C}.$$

The optimal solution to problem (2), $\pi^* = (p_1^* \ldots, p_K^*)$, depends on the characteristics of the different arms. We 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.

As mentioned above, we assume a single, scalar penalty constraint. Thus, 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. We focus our interest on cases in which there is no single dominating arm and the penalty constraint can be met (i.e. $\exists k : \mu^c(k) \leq C_0$). Fig. 1 illustrates the cases discussed.




(a) (b)

Fig. 1: Examples of different scenarios with $K = 4$; (a) is the scenario on which we focus in our derivation; (b) is a scenario in which there is a single dominating arm ($k_1$). The ellipses around the centers represent the distribution variances, and the optimal solution is drawn in red.

We continue with our analysis of the non-degenerate, 2-dimensional (scalar reward, scalar constraint) case.

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

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

where $S_1$ and $S_2$ are sets defined by

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

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:

$$\alpha \mu^c(k_1) + (1-\alpha)\mu^c(k_2) = C_0.$$

*Proof:* Being a set of linear combinations of points in a 2-dimensional space, the set of all possible reward-penalty combinations is a convex polytope. Thus, any point on its perimeter, including the optimal solution, can be represented as a convex combination of at most two extreme points. Since the optimal constraint for the non-degenerate case is $c^* = C_0$, the two arms which make up the optimal solution must be situated on opposite sides of the constraint. ■

We are now ready to formally define our objective. First, we define constraint satisfaction:

*Definition 2:* A policy $\pi$ 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 $f(t) \to 0$ and $g(t) \to 0$ 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. Generally, the expected regret for applying a certain policy $\pi$ when playing a MAB with $K$ arms is defined by

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

where $\mu(k^*)$ is the expected reward of the optimal arm.

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 3:* The expected regret 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, as defined in Proposition 1.

Our objective is to propose a policy that is PCS and minimizes the expected reward regret, compared to an optimal stationary PCS policy.

We now turn to our proposed algorithm and analyze its performance compared to the optimal solution, $(r^*, c^*)$.

## III. PROPOSED ALGORITHM

In this section we suggest an approach that is based on steering. The motivation for using this approach is the temporal character of our problem—every choice the agent makes is valid for a very short interval. Thus, single choices have a small effect on the average, and convergence to asymptotic values is rather fast.

Steering policies attempt to reach a certain goal by adapting their actions to changing conditions. In our case, the policy steers the average penalty incurred, $\hat{c}_t$, into the set $\mathscr{C}$, thus ensuring that the constraint is satisfied. While doing so, it attempts to maximize the average reward obtained, $\hat{r}_t$. Satisfying the constraint is achieved by predicting the average penalty after the next step, $\hat{c}^p_{t+1}$, based on arm characteristics (either known or learned) and on the average penalty incurred so far. This 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, a single arm is selected based on an augmented form of the reward. Since we assume Gaussian reward and penalty distributions, we implement the UCB1-NORMAL algorithm [3]. We note that 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; if there is more than one such arm - sample the arm which has been sampled the least.
  **else**
    Calculate augmented penalty and reward:
    $\bar{\mu}^c_t(k) \leftarrow \widetilde{\mu}^c_t(k) - 4\sqrt{\frac{q^c_t(k) - n_t(k)(\widetilde{\mu}^c_t(k))^2}{n_t(k)-1}\frac{\ln(t-1)}{n_t(k)}}$
    $\bar{\mu}^r_t(k) \leftarrow \widetilde{\mu}^r_t(k) + 4\sqrt{\frac{q^r_t(k) - n_t(k)(\widetilde{\mu}^r_t(k))^2}{n_t(k)-1}\frac{\ln(t-1)}{n_t(k)}}$
    Calculate projected average penalty for next step:
    $\hat{c}^{pr}_{t+1}(k) \leftarrow \frac{1}{t+1}\left(\bar{\mu}^c_t(k) + t\hat{c}_t\right)$
    Feasible set consists of arms $k$ for which $\hat{c}^{pr}_{t+1}(k) \leq C_0$; choose a feasible arm based on 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 $t$

In order to state our convergence results, we define the set of stages during which our bounds hold. These are almost all stages, except for an initial exploration period and stages in which forced exploration is needed, according to the UCB approach : $\mathscr{T} = \left\{t > K\lceil 8\log t\rceil \cap \mathscr{T}^C_{\text{jump}}\right\}$, where $\mathscr{T}_{\text{jump}} = \{t : \lceil 8\log t\rceil < \lceil 8\log(t+i)\rceil,\ 1 \leq i \leq K\}$.

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

*Theorem 2:* 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

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

where $\mu, \sigma$ are the distribution parameters of one of the arms and $\delta > 0$ is a parameter (see proof for details).

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

$$\frac{\bar{\mu}^c_t(k) + t\hat{c}_t}{t+1} \leq C_0,\ \ k \in \{1, \ldots, K\}, \tag{3}$$

where we use the assumptions introduced above regarding the scalar reward and penalty and $\bar{\mu}^c_t(k)$ is defined in Algorithm 1. Using (3), we derive an upper bound on the convergence rate of the average penalty, $\hat{c}_t$, to the optimal penalty, $c^* = C_0$.

**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, $N_t$.
An arm $k$ is dominated if there exists an arm $j$ such that
$$\bar{\mu}_t^c(j) \leq \bar{\mu}_t^c(k)$$
$$\bar{\mu}_t^r(j) > \bar{\mu}_t^r(k)$$
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 $\bar{\mu}_t^c(j) > C_0$ and which minimize the slope of the line connecting arms $k$ and $j$ in the reward-penalty plane:
$$j = \arg\min_i \left\{ \frac{\widetilde{\mu}_t^c(i) - \widetilde{\mu}_t^c(k)}{\bar{\mu}_t^r(i) - \bar{\mu}_t^r(k)} \right\}$$
Note: only positive slopes are considered, in order to avoid pairing arms with arms they dominate.
When a match does not exist, $k$ is its own best match.
7: Find intersection with constraint, $r_t^*(p,k)$, for all pairs.
For single arms, replace the intersection by $\bar{\mu}_t^r(k)$.
8: Choose pair for which $r_t^*(p,k)$ is maximal; play the feasible arm.
If both arms are feasible play the one with higher reward.
9: **end if**

Before proceeding with the proof, we state and prove an intermediate result that appears as a conjecture in [3].

*Lemma 3:* Let $X$ be a $\chi^2$ random variable with $K$ 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 the moment generating function for a central Chi-square distribution is $\mathbb{E}\left[e^{tX}\right] = (1-2t)^{-K/2}$. The expression is minimized when $t = \frac{\alpha-1}{2\alpha}$; substituting this and then substituting $\alpha = 4$ we have
$$\mathbb{P}\{X \geq \alpha K\} \leq \alpha^{K/2} e^{-(\alpha-1)K/2} \leq 4^{K/2} e^{-3K/2}.$$
Next, we rewrite $\frac{-3K}{2} = -\frac{K+1}{2} + \frac{-2K+1}{2}$:
$$\begin{aligned}\mathbb{P}\{X \geq 4K\} &\leq 4^{K/2} e^{\frac{-2K+1}{2}} e^{-\frac{K+1}{2}} \\ &= e^{\frac{1}{2}K\ln 4 - K + \frac{1}{2}} e^{-\frac{K+1}{2}}.\end{aligned}$$
In order to reach the desired bound, we need to ensure that the first factor is smaller than one, i.e. its exponent is smaller than or equal to zero. This condition is met for every $K \geq 2$, and therefore $\mathbb{P}\{X \geq 4K\} \leq e^{-\frac{K+1}{2}} \quad \forall K \geq 2$. ■

We now proceed to prove Theorem 2.

*Proof:* Our proof consists of three stages. First, we state and develop a condition which all feasible arms (in terms of penalty) must fulfill. Next, we use this condition in order to establish a bound on the convergence rate of the average penalty to the optimal penalty. We do so by separately characterizing the convergence rates of the confidence bound and the empirical mean. Finally, we use the characteristics of the problem and our algorithm to calculate an exact expression for a parametric bound on the convergence rate. Throughout our proof we assume the non-degenerate case, as described in Section II; the trivial case in which a single dominating arm exists is treated at the end of the proof.

**Stage 1 - feasibility condition:** As mentioned above, the next arm to be played must fulfill the condition
$$\frac{\bar{\mu}_t^c(k) + t\hat{c}_t}{t+1} \leq C_0 \iff \hat{c}_t - C_0 \leq \frac{C_0 - \bar{\mu}_t^c(k)}{t}. \quad (4)$$
Using the definition of $\bar{\mu}_t^c(k)$ which appears in Algorithm 1,
$$\begin{aligned}&\hat{c}_t - C_0 \\ &\leq \frac{1}{t}\left(C_0 - \widetilde{\mu}_t^c(k) + 4\sqrt{\frac{q_t^c(k) - n_t(k)\left(\widetilde{\mu}_t^c(k)\right)^2}{n_t(k)-1}\frac{\ln(t-1)}{n_t(k)}}\right) \\ &\leq \frac{1}{t}\left(C_0 - \widetilde{\mu}_t^c(k) + 4\sqrt{\frac{q_t^c(k) - n_t(k)\left(\widetilde{\mu}_t^c(k)\right)^2}{n_t(k)-1}\frac{\ln(t)}{n_t(k)}}\right).\end{aligned}$$

**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 $\chi^2$-distributed with $n_t(k) - 1$ degrees of freedom. Thus, using Lemma 3, we have that
$$\begin{aligned}&\mathbb{P}\{X_t \geq 4(n_t(k)-1)\} \\ &= \sum_{n=1}^{\infty} \mathbb{P}[X_t \geq 4(n_t(k)-1) | n_t(k) = n]\,\mathbb{P}\{n_t(k) = n\} \\ &= \sum_{n=\lceil 8\log t\rceil}^{\infty} \mathbb{P}[X_t \geq 4(n_t(k)-1) | n_t(k) = n]\,\mathbb{P}\{n_t(k) = n\} \\ &\leq \sum_{n=\lceil 8\log t\rceil}^{\infty} \mathbb{P}[X_t \geq 4(n_t(k)-1) | n_t(k) = n] \\ &\leq \sum_{n=\lceil 8\log t\rceil}^{\infty} e^{-n/2} \leq 3t^{-3/2},\end{aligned}$$
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 arm in the feasible set, 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, we have that with probability greater than $1 - 3t^{-3/2}$,
$$\hat{c}_t - C_0 \leq \frac{1}{t}\left(C_0 - \widetilde{\mu}_t^c(k) + \sqrt{2\left(\sigma^c(k)\right)^2}\right). \quad (5)$$

**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 $\varepsilon > 0$

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

where $Q(\cdot)$ is the tail probability of the standard normal distribution, and the exponential bound for it appears in, e.g., [14]. 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 (5) and using the union bound yields

$$\hat{c}_t - C_0 \leq \frac{1}{t}\left(|C_0| + |\mu^c(k)| + \delta + \sqrt{2(\sigma^c(k))^2}\right), \quad (6)$$

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 (in terms of penalty) is made according to the reward. For the arm which maximizes the right hand side of (6) we denote $\mu^c(k) \triangleq \mu$ and $\sigma^c(k) \triangleq \sigma$. Therefore, the convergence bound for the average 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

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

Therefore, in the terms of Definition 2, we have

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

■

*Theorem 4:* The expected reward regret for running Algorithm 1 on $K$ arms with normally distributed rewards and penalties, defined in Definition 3, is bounded for all $t \in \mathscr{T}$:

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

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.

We note that the bound sums only over pairs of non-dominated arms. Thus, depending on the geometry of the problem, the confinement of Algorithm 1 to the non-dominated arms considerably decreases its expected regret.

For the proof of Theorem 4 we assume the non-degenerate scenario, an example of which is shown in Fig. 1a. In this case, as explained above, the optimal reward is obtained by choosing a certain combination of exactly two arms. We treat the degenerate scenario, in which a single arm dominates the others, immediately after the proof.

In our proof we show that the optimal pair of arms is chosen with high probability and that the correct balance is achieved; the correct balance is the one which achieves the optimal penalty. In order to derive a bound on the algorithm's regret, we follow the lines of the proofs of Theorem 1 and 4 of [3]. We consider an equivalent problem, which consists of choosing the optimal pair of arms among all available pairs. Since we have already proved the convergence of the average penalty incurred by our algorithm to the optimal penalty, $c^* = C_0$, once we converge to the optimal pair of arms, convergence to the correct balance is ensured. The correct balance, as mentioned in Section II, is

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

Before proceeding with the proof, we restate Conjecture 1 from the proof of Theorem 4 in [3], which we will use further along the proof.

*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 4.

*Proof:* We begin our proof by defining and characterizing another MAB, whose arms represent pairs of arms of the original bandit. Then, using this definition, we bound the expected regret in the reward sense.

**Stage 1: Definition of pairs-MAB** We define a new MAB, whose every arm represents a pair 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 can be greatly improved

by considering only arms which are not dominated (see Definition 1) and by pairing only arms which are situated on opposite sides of the penalty constraint.

As in the case of the arms of the original bandit problem, every arm is represented by an index which reflects its empirical mean reward together with an upper confidence bound. The penalty of all arms (i.e., pairs) converges to the same value: $c(p,k) = c^* = C_0$. Denoting the reward confidence bound of a single arm by $b^r(k)$, we set the reward index of each pair of arms to be

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

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 $\widetilde{\mu}^r(p,k) = \alpha\widetilde{\mu}^r(k_1) + (1-\alpha)\widetilde{\mu}^r(k_2)$ is its empirical mean reward. We note that the parameter $\alpha(k_1,k_2)$ is determined for every pair of arms based on their parameters, and is used only for the sake of the analysis.

**Stage 2: Bounding the expected regret** Based on Definition 3, the expected reward regret is

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

where the "$*$" notation indicates the optimal pair. In order to bound the regret, we must 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, $n_t(k)$.

Let us examine $b^r_{t,s}(p,k)$, which is 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 $l \geq \lceil 8\log t\rceil$

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

We further develop our bound by comparing the worst case of the optimal arm with the best case of the sub-optimal arm:

$$\begin{aligned}n_t(p,k) &\leq l + \sum_{\tau=\tau_0}^{t}\mathbf{1}\Big\{\min_{0<s<\tau}\left[\widetilde{\mu}^r_s(p,k^*) + b^r_{\tau-1,s}(p,k^*)\right] \\ &\qquad \leq \max_{l<s_k<\tau}\left[\widetilde{\mu}^r_{s_k}(p,k) + b^r_{\tau-1,s_k}(p,k)\right]\Big\} \\ &\leq l + \sum_{\tau=1}^{\infty}\sum_{s=1}^{\tau-1}\sum_{s_k=l}^{\tau-1}\mathbf{1}\Big\{\widetilde{\mu}^r_s(p,k^*) + b^r_{\tau,s}(p,k^*) \\ &\qquad \leq \widetilde{\mu}^r_{s_k}(p,k) + b^r_{\tau,s_k}(p,k)\Big\}.\end{aligned}$$

Denoting

$$\begin{aligned}S &\triangleq \left\{\widetilde{\mu}^r_s(p^*) + b^r_{\tau,s}(p^*) \leq \widetilde{\mu}^r_{s_k}(p,k) + b^r_{\tau,s_k}(p,k)\right\}, \\ A &\triangleq \left\{\widetilde{\mu}^r_s(p^*) \leq \mu^r(p^*) - b^r_{\tau,s}(p^*)\right\}, \\ B &\triangleq \left\{\widetilde{\mu}^r_{s_k}(p,k) \geq \mu^r(p,k) + b^r_{\tau,s_k}(p,k)\right\}, \\ C &\triangleq \left\{\mu^r(p^*) \leq \mu^r(p,k) + 2b^r_{\tau,s_k}(p,k)\right\},\end{aligned}$$

we have that $S \subseteq A \cup B \cup C$. This follows from the definitions of these events: $A$ is the event in which the optimal pair underperforms, $B$ is the event in which a sub-optimal arm overperforms and $C$ is the event in which the expected rewards of the optimal and suboptimal pairs are too close to be distinguishable. Thus, event $S$ requires that *at least* one of the events $A, B, C$ occur.

By breaking up the optimal pair into the arms of which it consists, event $A$ can be rewritten as $A \subseteq A_1 \cup A_2$, where

$$\begin{aligned}A_1 &\triangleq \left\{\widetilde{\mu}^r_{s_1}(k_1^*) \leq \mu^r(k_1^*) - b^r_{\tau,s_1}(k_1^*)\right\} \\ A_2 &\triangleq \left\{\widetilde{\mu}^r_{s_2}(k_2^*) \leq \mu^r(k_2^*) - b^r_{\tau,s_2}(k_2^*)\right\},\end{aligned}$$

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 $\left(\widetilde{\mu}^r_{s_k}(k) - \mu^r(k)\right) / \sqrt{\left(q^r_{s_k} - s_k\left(\widetilde{\mu}^r_{s_k}(k)\right)^2\right)/\left(s_k(s_k-1)\right)}$ 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

$$\begin{aligned}&\mathbb{P}\left\{\widetilde{\mu}^r_{s_1}(k_1^*) \leq \mu^r(k_1^*) - b^r_{\tau,s_1}(k_1^*)\right\} \\ &= \mathbb{P}\left\{\frac{\widetilde{\mu}^r_{s_1}(k_1^*) - \mu^r(k_1^*)}{\sqrt{\left(q^r_{s_1} - s_1\left(\widetilde{\mu}^r_{s_1}(k_1^*)\right)^2\right)/\left(s_1(s_1-1)\right)}} \leq 4\sqrt{\ln\tau}\right\} \\ &\leq \tau^{-4}.\end{aligned} \tag{7}$$

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

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

We rewrite event $B$ similarly: $B \subseteq B_1 \cup B_2$, where

$$\begin{aligned}B_1 &\triangleq \left\{\widetilde{\mu}^r_{s_1}(k_1) \geq \mu^r(k_1) + b^r_{\tau,s_1}(k_1)\right\} \\ B_2 &\triangleq \left\{\widetilde{\mu}^r_{s_2}(k_2) \geq \mu^r(k_2) + b^r_{\tau,s_2}(k_2)\right\},\end{aligned}$$

and $k_1$ and $k_2$ are the arms which make up the $k$'th pair. Using an argument analogous to (7),

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

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

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

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 $\Delta_{k_1} \triangleq \mu^r(k_1^*) - \mu^r(k_1)$, using the explicit expression for $b^r_{\tau,s}(k_1)$ and reorganizing the equation 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 3 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 $\mathbb{P}\{C_1\}$:

$$\begin{aligned}\mathbb{P}\{C_1\} &= \sum_{s=1}^{\infty}\mathbb{P}[C_1|s_1=s]\mathbb{P}\{s_1=s\} \\ &= \sum_{s=m_1\ln\tau}^{\infty}\mathbb{P}[C_1|s_1=s]\mathbb{P}\{s_1=s\} \\ &\leq \sum_{s=m_1\ln\tau}^{\infty} e^{-s/2} \\ &\leq 3\tau^{-m_1/2}.\end{aligned}$$

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

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

Finally, the bound for the expected number of times a suboptimal pair of arms is sampled is

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

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

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

At this point we note that this expression bounds only the mean number of times pairs of non-dominated arms are sampled. *All* arms are sampled at least $\lceil 8\ln t\rceil$ times, and therefore the bound for the expected reward-regret is

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

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:* The proofs of Theorem 2 and Theorem 4 deal with the non-degenerate case, in which there is no single dominating arm. When such an arm exists, the optimal solution is to sample it alone. Thus, the optimal penalty is the expected penalty of this arm, $\mu^c(k^*)$, and the optimal reward is its expected reward, $\mu^r(k^*)$. Algorithm 1 treats the selection of the next arm to be played in a pairwise manner; a single dominating arm is paired with itself. Thus, the problem of convergence to the optimal reward is analyzed in the same manner as in the non-degenerate case, and the expected reward regret is bounded as stated in Theorem 4. 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 2) 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 4.

## IV. SIMULATIONS

We demonstrate our results using simulations of a CR problem. In our scenario, the CA repeatedly has to choose one of 5 channels, applying one of two possible coding techniques and transmitting at one of four possible power levels. These parameters constitute 40 possible transmission profiles. Using a hypothetical model, we map every profile to Gaussian reward and penalty distributions. We base this model on reasonable assumptions, such as the positive impact of higher power levels on both reward and penalty.

We allow the CA to interact with the system for $T = 40,000$ cycles, monitoring its average reward (throughput) and penalty (power) together with the number of times it sampled every arm. For reference, we implement two algorithms: an ideal one that applies an optimal stationary (OS) policy, based on full knowledge of the arm characteristics, and another that applies a certainty equivalence (CE) approach, updating its estimate of the optimal solution based on the empirical means of the reward and penalty.

The results of our simulations, averaged over 100 repetitions, are presented in Fig. 2. Fig. 2a displays the problem layout in the reward-penalty plane. The ellipses represent the distributions of the arms, with their mean values and

variances. The thickness of ellipse contours represents the number of times an arm was sampled. The optimal solution and the average performance of our algorithm are also annotated. Clearly, dominated arms are sampled a minimal number of times, while other arms are sampled proportionally to their chance of participating in the optimal solution.

Fig. 2b displays the convergence of the average penalty to the optimal penalty, together with the bound derived in Theorem 2 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. Note that the graph is plotted using a logarithmic scale. We also display the convergence of the average of the worst 5% of the runs, where the advantage of the steering policy is clear.

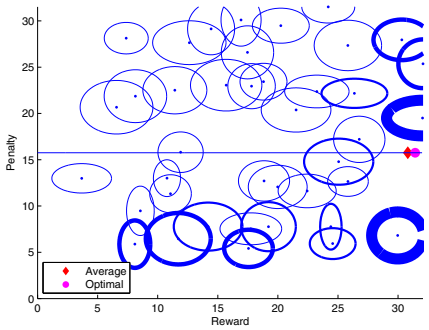
Finally, we present the convergence of the average reward to the optimal value. We compare our steering algorithm to the optimal mixed policy, to the certainty equivalence policy and to the theoretical bound derived in Theorem 4. As expected, we pay for the steering policy's strict adherence to the constraint in terms of reward convergence. However, reward convergence is identical in the average and worst case scenarios, unlike that of the certainty equivalence approach.

## V. CONCLUSIONS AND FUTURE WORK

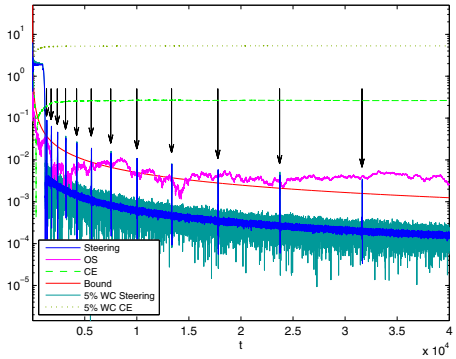
We introduced a formulation of the CR problem using stochastic MABs with pathwise constraints. In order to solve this problem, we proposed a steering policy that results in convergence of the average reward and penalty to their optimal values.

An extension of our work to the case of bounded reward and penalty distributions, using the UCB1 algorithm [3], is straightforward using Hoeffding's inequality.

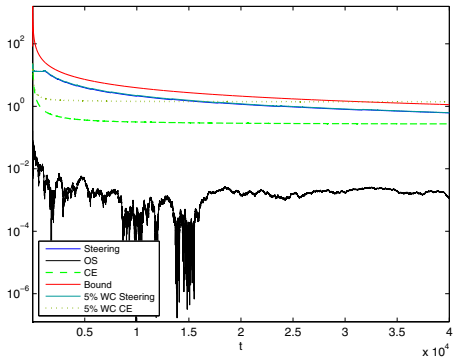
Future directions include examining the proposed formulation from a multiple agent point of view, in order to understand issues of cooperation and competition in this setting. We also plan to examine the issue of bandits with correlated arms, in which the distributions of sub-groups of arms are not independent. These may provide a realistic model for closely related transmission profiles. Finally, we hope to be able to apply our framework to real-world data.




(a) Problem layout




(b) Penalty convergence to optimal value




(c) Reward regret convergence

Fig. 2: Simulation results

## 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.

[14] J. M. Wozencraft and I. M. Jacobs. *Principles of communication engineering*, volume 28. Wiley New York, 1965.