Online Learning Algorithms for Stochastic Water-Filling

Yi Gai and Bhaskar Krishnamachari
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089, USA
Email: {ygai, bkrishna}@usc.edu

***Abstract*—Water-filling is the term for the classic solution to the problem of allocating constrained power to a set of parallel channels to maximize the total data-rate. It is used widely in practice, for example, for power allocation to sub-carriers in multi-user OFDM systems such as WiMax. The classic water-filling algorithm is deterministic and requires perfect knowledge of the channel gain to noise ratios. In this paper we consider how to do power allocation over stochastically time-varying (i.i.d.) channels with unknown gain to noise ratio distributions. We adopt an online learning framework based on stochastic multi-armed bandits. We consider two variations of the problem, one in which the goal is to find a power allocation to maximize $\sum_i \mathbb{E}[\log(1+SNR_i)]$, and another in which the goal is to find a power allocation to maximize $\sum_i \log(1+\mathbb{E}[SNR_i])$. For the first problem, we propose a *cognitive water-filling* algorithm that we call CWF1. We show that CWF1 obtains a regret (defined as the cumulative gap over time between the sum-rate obtained by a distribution-aware genie and this policy) that grows polynomially in the number of channels and logarithmically in time, implying that it asymptotically achieves the optimal time-averaged rate that can be obtained when the gain distributions are known. For the second problem, we present an algorithm called CWF2, which is, to our knowledge, the first algorithm in the literature on stochastic multi-armed bandits to exploit non-linear dependencies between the arms. We prove that the number of times CWF2 picks the incorrect power allocation is bounded by a function that is polynomial in the number of channels and logarithmic in time, implying that its frequency of incorrect allocation tends to zero.**

I. Introduction

A fundamental resource allocation problem that arises in many settings in communication networks is to allocate a constrained amount of power across many parallel channels in order to maximize the sum-rate. Assuming that the power-rate function for each channel is proportional to $\log(1+SNR)$ as per the Shannon's capacity theorem for AWGN channels, it is well known that the optimal power allocation can be determined by a water-filling strategy [1]. The classic water-filling solution is a deterministic algorithm, and requires perfect knowledge of all channel gain to noise ratios.

In practice, however, channel gain-to-noise ratios are stochastic quantities. To handle this randomness, we consider an alternative approach, based on online learning, specifically stochastic multi-armed bandits. We formulate the problem of stochastic water-filling as follows: time is discretized into slots; each channel's gain-to-noise ratio is modeled as an i.i.d. random variable with an unknown distribution. In our general formulation, the power-to-rate function for each channel is allowed to be any sub-additive function [1]. We seek a power allocation that maximizes the expected sum-rate (i.e., an optimization of the form $\mathbb{E}[\sum_i \log(1+SNR_i)]$). Even if the channel gain-to-noise ratios are random variables with known distributions, this turns out to be a hard combinatorial stochastic optimization problem. Our focus in this paper is thus on a more challenging case.

In the classical multi-armed bandit, there is a player playing K arms that yield stochastic rewards with unknown means at each time in i.i.d. fashion over time. The player seeks a policy to maximize its total expected reward over time. The performance metric of interest in such problems is regret, defined as the cumulative difference in expected reward between a model-aware genie and that obtained by the given learning policy. And it is of interest to show that the regret grows sub-linearly with time so that the time-averaged regret asymptotically goes to zero, implying that the time-averaged reward of the model-aware genie is obtained asymptotically by the learning policy.

We show that it is possible to map the problem of stochastic water-filling to an MAB formulation by treating each possible power allocation as an arm (we consider discrete power levels in this paper; if there are P possible power levels for each of N channels, there would be P^N total arms.) We present a novel combinatorial policy for this problem that we call CWF1, that yields regret growing polynomially in N and logarithmically over time. Despite the exponential growing set of arms, the CWF1 observes and maintains information for $P \cdot N$ variables, one corresponding to each power-level and channel, and exploits linear dependencies between the arms based on these variables.

Typically, the way the randomness in the channel gain to noise ratios is dealt with is that the mean channel gain to noise ratios are estimated first based on averaging a finite set of training observations and then the estimated gains are used in a deterministic water-filling procedure. Essentially this

This research was sponsored in part by the U.S. Army Research Laboratory under the Network Science Collaborative Technology Alliance, Agreement Number W911NF-09-2-0053.

[1] A function f is subadditive if $f(x+y) \leq f(x)+f(y)$; for any concave function g, if $g(0) \geq 0$ (such as $\log(1+x)$), g is subadditive.

approach tries to identify the power allocation that maximizes a pseudo-sum-rate, which is determined based on the power-rate equation applied to the mean channel gain-to-noise ratios (i.e., an optimization of the form $\sum_i \log(1 + \mathbb{E}[SNR_i])$. We also present a different stochastic water-filling algorithm that we call CWF2, which learns to do this in an online fashion. This algorithm observes and maintains information for N variables, one corresponding to each channel, and exploits non-linear dependencies between the arms based on these variables. To our knowledge, CWF2 is the first MAB algorithm to exploit non-linear dependencies between the arms. We show that the number of times CWF2 plays a non-optimal combination of powers is uniformly bounded by a function that is logarithmic in time. Under some restrictive conditions, CWF2 may also solve the first problem more efficiently.

II. Related Work

The classic water-filling strategy is described in [1]. There are a few other stochastic variations of water-filling that have been covered in the literature that are different in spirit from our formulation. When a fading distribution over the gains is known *a priori*, the power constraint is expressed over time, and *the instantaneous gains are also known*, then a deterministic joint frequency-time water-filling strategy can be used [2], [3]. In [4], a stochastic gradient approach based on Lagrange duality is proposed to solve this problem when the fading distribution is unknown but still instantaneous gains are available. By contrast, in our work we do not assume that the instantaneous gains are known, and focus on keeping the same power constraint at each time while considering unknown gain distributions.

Another work [5] considers water-filling over stochastic non-stationary fading channels, and proposes an adaptive learning algorithm that tracks the time-varying optimal power allocation by incorporating a forgetting factor. However, the focus of their algorithm is on minimizing the maximum mean squared error assuming imperfect channel estimates, and they prove only that their algorithm would converge in a stationary setting. Although their algorithm can be viewed as a learning mechanism, they do not treat stochastic water-filling from the perspective of multi-armed bandits, which is a novel contribution of our work. In our work, we focus on stationary setting with perfect channel estimates, but prove stronger results, showing that our learning algorithm not only converges to the optimal allocation, it does so with sub-linear regret.

There has been a long line of work on stochastic multi-armed bandits involving playing arms yielding stochastically time varying rewards with unknown distributions. Several authors [6]–[9] present learning policies that yield regret growing logarithmically over time (asymptotically, in the case of [6]–[8] and uniformly over time in the case of [9]). Our algorithms build on the UCB1 algorithm proposed in [9] but make significant modifications to handle the combinatorial nature of the arms in this problem. CWF1 has some commonalities with the LLR algorithm we recently developed for a completely different problem, that of stochastic combinatorial bipartite matching for channel allocation [10], but is modified to account for the non-linear power-rate function in this paper. Other recent work on stochastic MAB has considered decentralized settings [11]–[14], and non-i.i.d. reward processes [15]–[19]. With respect to this literature, the problem setting for stochastic water-filling is novel in that it involves a non-linear function of the action and unknown variables. In particular, as far as we are aware, our CWF2 policy is the first to exploit the non-linear dependencies between arms to provably improve the regret performance.

III. Problem Formulation

We define the stochastic version of the classic communication theory problem of power allocation for maximizing rate over parallel channels (water-filling) as follows.

We consider a system with N channels, where the channel gain-to-noise ratios are unknown random processes $X_i(n), 1 \leq i \leq N$. Time is slotted and indexed by n. We assume that $X_i(n)$ evolves as an i.i.d. random process over time (i.e., we consider block fading), with the only restriction that its distribution has a finite support. Without loss of generality, we normalize $X_i(n) \in [0, 1]$. We do not require that $X_i(n)$ be independent across i. This random process is assumed to have a mean $\theta_i = \mathbb{E}[X_i]$ that is unknown to the users. We denote the set of all these means by $\Theta = \{\theta_i\}$.

At each decision period n (also referred to interchangeably as a time slot), an N-dimensional action vector $\mathbf{a}(n)$, representing a power allocation on these N channels, is selected under a policy $\pi(n)$. We assume that the power levels are discrete, and we can put any constraint on the selections of power allocations such that they are from a finite set $\mathcal{F}$ (i.e., the maximum total power constraint, or an upper bound on the maximum allowed power per subcarrier). We assume $a_i(n) \geq 0$ for all $1 \leq i \leq N$. When a particular power allocation $\mathbf{a}(n)$ is selected, the channel gain-to-noise ratios corresponding to nonzero components of $\mathbf{a}(n)$ are revealed, i.e., the value of $X_i(n)$ is observed for all i such that $a_i(n) \neq 0$. We denote by $\mathcal{A}_{\mathbf{a}(n)} = \{i : a_i(n) \neq 0, 1 \leq i \leq N\}$ the index set of all $a_i(n) \neq 0$ for an allocation $\mathbf{a}$.

We adopt a general formulation for water-filling, where the sum rate [2] obtained at time n by allocating a set of powers $\mathbf{a}(n)$ is defined as:

$$R_{\mathbf{a}(n)}(n) = \sum_{i \in \mathcal{A}_{\mathbf{a}(n)}} f_i(a_i(n), X_i(n)). \tag{1}$$

where for all i, $f_i(a_i(n), X_i(n))$ is a nonlinear continuous increasing sub-additive function in $X_i(n)$, and $f_i(a_i(n), 0) = 0$ for any $a_i(n)$. We assume f_i is defined on $\mathbb{R}^+ \times \mathbb{R}^+$.

Our formulation is general enough to include as a special case of the rate function obtained from Shannon's capacity theorem for AWGN, which is widely used in communication

[2] We refer to rate and reward interchangeably in this paper.

networks:

$$R_{\mathbf{a}(n)}(n) = \sum_{i=1}^{N} \log(1 + a_i(n)X_i(n))$$

In the typical formulation there is a total power constraint and individual power constraints, the corresponding constraint is

$$\mathcal{F} = \{\mathbf{a} : \sum_{i=1}^{N} a_i \leq P_{\text{total}} \wedge 0 \leq a_i \leq P_i, \forall i\}.$$

where P_{total} is the total power constraint and P_i is the maximum allowed power per channel.

Our goal is to maximize the expected sum-rate when the distributions of all X_i are unknown, as shown in (2). We refer to this objective as $\mathbf{O_1}$.

$$\max_{\mathbf{a}\in\mathcal{F}} \mathbb{E}[\sum_{i\in\mathcal{A}_\mathbf{a}} f_i(a_i, X_i))] \tag{2}$$

Note that even when X_i have known distributions, this is a hard combinatorial non-linear stochastic optimization problem. In our setting, with unknown distributions, we can formulate this as a multi-armed bandit problem, where each power allocation $\mathbf{a}(n) \in \mathcal{F}$ is an arm and the reward function is in a combinatorial non-linear form. The optimal arms are the ones with the largest expected reward, denoted as $\mathcal{O}^* = \{\mathbf{a}^*\}$. For the rest of the paper, we use $*$ as the index indicating that a parameter is for an optimal arm. If more than one optimal arm exists, $*$ refers to any one of them.

We note that for the combinatorial multi-armed bandit problem with linear rewards where the reward function is defined by $R_{\mathbf{a}(n)}(n) = \sum_{i\in\mathcal{A}_{\mathbf{a}(n)}} a_i(n)X_i(n)$, $\mathbf{a}^*$ is a solution to a deterministic optimization problem because $\max_{\mathbf{a}\in\mathcal{F}} \mathbb{E}[\sum_{i\in\mathcal{A}_\mathbf{a}} a_i X_i] = \max_{\mathbf{a}\in\mathcal{F}} \sum_{i\in\mathcal{A}_\mathbf{a}} a_i \mathbb{E}[X_i]$. Different from the combinatorial multi-armed bandit problem with linear rewards, $\mathbf{a}^*$ here is a solution to a stochastic optimization problem, i.e.,

$$\mathbf{a}^* \in \mathcal{O}^* = \{\tilde{\mathbf{a}} : \tilde{\mathbf{a}} = \arg\max_{\mathbf{a}\in\mathcal{F}} \mathbb{E}[\sum_{i\in\mathcal{A}_\mathbf{a}} f_i(a_i, X_i))]\}. \tag{3}$$

We evaluate policies for $\mathbf{O_1}$ with respect to *regret*, which is defined as the difference between the expected reward that could be obtained by a genie that can pick an optimal arm at each time, and that obtained by the given policy. Note that minimizing the regret is equivalent to maximizing the expected rewards. Regret can be expressed as:

$$\mathfrak{R}^\pi(n) = nR^* - \mathbb{E}[\sum_{t=1}^{n} R_{\pi(t)}(t)], \tag{4}$$

where $R^* = \max_{\mathbf{a}\in\mathcal{F}} \mathbb{E}[\sum_{i\in\mathcal{A}_\mathbf{a}} f_i(a_i, X_i))]$, the expected reward of an optimal arm.

Intuitively, we would like the regret $\mathfrak{R}^\pi(n)$ to be as small as possible. If it is sub-linear with respect to time n, the time-averaged regret will tend to zero and the maximum possible time-averaged reward can be achieved. Note that the number of arms $|\mathcal{F}|$ can be exponential in the number of unknown random variables N.

We also note that for the stochastic version of the water-filling problems, a typical way in practice to deal with the unknown randomness is to estimate the mean channel gain to noise ratios first and then find the optimized allocation based on the mean values. This approach tries to identify the power allocation that maximizes the power-rate equation applied to the mean channel gain-to-noise ratios. We refer to maximizing this as the sum-pseudo-rate over averaged channels. We denote this objective by $\mathbf{O_2}$, as shown in (5).

$$\max_{\mathbf{a}\in\mathcal{F}} \sum_{i\in\mathcal{A}_\mathbf{a}} f_i(a_i, \mathbb{E}[X_i]) \tag{5}$$

We would also like to develop an online learning policy for $\mathbf{O_2}$. Note that the optimal arm $\mathbf{a}^*$ of $\mathbf{O_2}$ is a solution to a deterministic optimization problem. So, we evaluate the policies for $\mathbf{O_2}$ with respect to the expected total number of times that a non-optimal power allocation is selected. We denote by $T_\mathbf{a}(n)$ the number of times that a power allocation is picked up to time n. We denote $r_\mathbf{a} = \sum_{i\in\mathcal{A}_\mathbf{a}} f_i(a_i, \mathbb{E}[X_i])$. Let $T^\pi_{non}(n)$ denote the total number of times that a policy π select a power allocation $r^\mathbf{a} < r^{\mathbf{a}^*}$. Denote by $\mathbb{1}^\pi_t(\mathbf{a})$ the indicator function which is equal to 1 if $\mathbf{a}$ is selected under policy π at time t, and 0 else. Then

$$\begin{aligned}\mathbb{E}[T^\pi_{non}(n)] &= n - \mathbb{E}[\sum_{t=1}^{n} \mathbb{1}^\pi_t(\mathbf{a}^*) = 1] \\ &= \sum_{r_\mathbf{a} < r_{\mathbf{a}^*}} \mathbb{E}[T_\mathbf{a}(n)].\end{aligned} \tag{6}$$

IV. Online Learning for Maximizing the Sum-Rate

We first present in this section an online learning policy for stochastic water-filling under object $\mathbf{O_1}$.

A. Policy Design

A straightforward, naive way to solve this problem is to use the UCB1 policy proposed [9]. For UCB1, each power allocation is treated as an arm, and the arm that maximizes $\hat{Y}_k + \sqrt{\frac{2\ln n}{m_k}}$ will be selected at each time slot, where $\hat{Y}_k$ is the mean observed reward on arm k, and m_k is the number of times that arm k has been played. This approach essentially ignores the underlying dependencies across the different arms, and requires storage that is linear in the number of arms and yields regret growing linearly with the number of arms. Since there can be an exponential number of arms, the UCB1 algorithm performs poorly on this problem.

We note that for combinatorial optimization problems with linear reward functions, an online learning algorithm LLR has been proposed in [6] as an efficient solution. LLR stores the mean of observed values for every underlying unknown random variable, as well as the number of times each has been observed. So the storage of LLR is linear in the number of unknown random variables, and the analysis in [6] shows LLR achieves a regret that grows logarithmically in time, and polynomially in the number of unknown parameters.

However, the challenge with stochastic water-filling with objective $\mathbf{O_1}$, where the expectation is outside the non-linear reward function, directly storing the mean observations of X_i will not work.

To deal with this challenge, we propose to store the information for each a_i, X_i combination, i.e., $\forall 1 \leq i \leq N, \forall a_i$, we define a new set of random variables $Y_{i,a_i} = f_i(a_i, X_i)$. So now the number of random variables Y_{i,a_i} is $\sum_{i=1}^{N} |\mathcal{B}_i|$, where $\mathcal{B}_i = \{a_i : a_i \neq 0\}$. Note that $\sum_{i=1}^{N} |\mathcal{B}_i| \leq PN$.

Then the reward function can be expressed as

$$R_{\mathbf{a}} = \sum_{i \in \mathcal{A}_{\mathbf{a}}} Y_{i,a_i}, \tag{7}$$

Note that (7) is in a combinatorial linear form.

For this redefined MAB problem with $\sum_{i=1}^{N} |\mathcal{B}_i|$ unknown random variables and linear reward function (7), we propose the following online learning policy CWF1 for stochastic water-filling as shown in Algorithm 1.

Algorithm 1 Online Learning for Stochastic Water-Filling: CWF1

1: // INITIALIZATION
2: If $\max_{\mathbf{a}} |\mathcal{A}_{\mathbf{a}}|$ is known, let $L = \max_{\mathbf{a}} |\mathcal{A}_{\mathbf{a}}|$; else, $L = N$;
3: **for** $n = 1$ to N **do**
4: Play any arm $\mathbf{a}$ such that $n \in \mathcal{A}_{\mathbf{a}}$;
5: $\forall i \in \mathcal{A}_{\mathbf{a}}, \forall a_i \in \mathcal{B}_i, \overline{Y}_{i,a_i} := \frac{\overline{Y}_{i,a_i} m_i + f_i(a_i, X_i)}{m_i + 1}$;
6: $\forall i \in \mathcal{A}_{\mathbf{a}}, m_i := m_i + 1$;
7: **end for**
8: // MAIN LOOP
9: **while** 1 **do**
10: $n := n + 1$;
11: Play an arm $\mathbf{a}$ which solves the maximization problem

$$\sum_{i \in \mathcal{A}_{\mathbf{a}}} (\overline{Y}_{i,a_i} + \sqrt{\frac{(L+1)\ln n}{m_i}}); \tag{8}$$

12: $\forall i \in \mathcal{A}_{\mathbf{a}}, \forall a_i \in \mathcal{B}_i, \overline{Y}_{i,a_i} := \frac{\overline{Y}_{i,a_i} m_i + f_i(a_i, X_i)}{m_i + 1}$;
13: $\forall i \in \mathcal{A}_{\mathbf{a}}, m_i := m_i + 1$;
14: **end while**

To have a tighter bound of regret, different from the LLR algorithm, instead of storing the number of times that each unknown random variables Y_{i,a_i} has been observed, we use a 1 by N vector, denoted as $(m_i)_{1\times N}$, to store the number of times that X_i has been observed up to the current time slot.

We use a 1 by $\sum_{i=1}^{N} |\mathcal{B}_i|$ vector, denoted as $(\overline{Y}_{i,a_i})_{1\times \sum_{i=1}^{N} |\mathcal{B}_i|}$ to store the information based on the observed values. $(\overline{Y}_{i,a_i})_{1\times \sum_{i=1}^{N} |\mathcal{B}_i|}$ is updated in as shown in line 12. Each time an arm $\mathbf{a}(n)$ is played, $\forall i \in \mathcal{A}_{\mathbf{a}(n)}$, the observed value of X_i is obtained. For every observed value of X_i, $|\mathcal{B}_i|$ values are updated: $\forall a_i \in \mathcal{B}_i$, the average value $\overline{Y}_{i,a_i}$ of all the values of Y_{i,a_i} up to the current time slot is updated. CWF1 policy requires storage linear in $\sum_{i=1}^{N} |\mathcal{B}_i|$.

B. Analysis of regret

***Theorem** 1:* The expected regret under the CWF1 policy is at most

$$\left[\frac{4L^2(L+1)N\ln n}{(\Delta_{\min})^2} + N + \frac{\pi^2}{3}LN\right]\Delta_{\max}. \tag{9}$$

where $\Delta_{\min} = \min_{\mathbf{a}\neq\mathbf{a}^*} R^* - \mathbb{E}[R_{\mathbf{a}}]$, $\Delta_{\max} = \max_{\mathbf{a}\neq\mathbf{a}^*} R^* - \mathbb{E}[R_{\mathbf{a}}]$. Note that $L \leq N$.

The proof of Theorem 1 is omitted.

***Remark** 1:* For CWF1 policy, although there are $\sum_{i=1}^{N} |\mathcal{B}_i|$ random variables, the upper bound of regret remains $O(N^4 \log n)$, which is the same as LLR, as shown by Theorem 2 in [6]. Directly applying LLR algorithm to solve the redefined MAB problem in (7) will result in a regret that grows as $O(P^4 N^4 \log n)$.

***Remark** 2:* Algorithm 1 will even work for rate functions that do not satisfy subadditivity.

***Remark** 3:* We can develop similar policies and results when X_i are Markovian rewards as in [19] and [20].

V. Online Learning for Sum-Pseudo-Rate

We now show our novel online learning algorithm CWF2 for stochastic water-filling with object $\mathbf{O_2}$. Unlike CWF1, CWF2 exploits non-linear dependencies between the choices of power allocations and requires lower storage. Under condition where the power allocation that maximize $\mathbf{O_2}$ also maximize $\mathbf{O_1}$, we will see through simulations that CWF2 has better regret performance.

A. Policy Design

Our proposed policy CWF2 for stochastic water filling with objective $\mathbf{O_2}$ is shown in Algorithm 2.

We use two 1 by N vectors to store the information after we play an arm at each time slot. One is $(\overline{X}_i)_{1\times N}$ in which $\overline{X}_i$ is the average (sample mean) of all the observed values of X_i up to the current time slot (obtained through potentially different sets of arms over time). The other one is $(m_i)_{1\times N}$ in which m_i is the number of times that X_i has been observed up to the current time slot. So CWF2 policy requires storage linear in N.

B. Analysis of regret

For the analysis of the upper bound for $\mathbb{E}[T^{\pi}_{non}(n)]$ of CWF2 policy, we use the inequalities as stated in the Chernoff-Hoeffding bound as follows:

***Lemma** 1 (Chernoff-Hoeffding bound [21]):*
$X_1, \ldots, X_n$ are random variables with range $[0, 1]$, and

Algorithm 2 Online Learning for Stochastic Water-Filling: CWF2

// INITIALIZATION
If $\max_{\mathbf{a}} |\mathcal{A}_{\mathbf{a}}|$ is known, let $L = \max_{\mathbf{a}} |\mathcal{A}_{\mathbf{a}}|$; else, $L = N$;
for $n = 1$ to N **do**
    Play any arm $\mathbf{a}$ such that $n \in \mathcal{A}_{\mathbf{a}}$;
    $\forall i \in \mathcal{A}_{\mathbf{a}}$, $\overline{X}_i := \frac{\overline{X}_i m_i + X_i}{m_i+1}$, $m_i := m_i + 1$;
end for
// MAIN LOOP
while 1 **do**
    $n := n + 1$;
    Play an arm $\mathbf{a}$ which solves the maximization problem

$$\max_{\mathbf{a}\in\mathcal{F}} \sum_{i\in\mathcal{A}_{\mathbf{a}}} \left(f_i(a_i, \overline{X}_i) + f_i(a_i, \sqrt{\frac{(L+1)\ln n}{m_i}}) \right); \quad (10)$$

    $\forall i \in \mathcal{A}_{\mathbf{a}(n)}$, $\overline{X}_i := \frac{\overline{X}_i m_i + X_i}{m_i+1}$, $m_i := m_i + 1$;
end while

$E[X_t|X_1, \ldots, X_{t-1}] = \mu$, $\forall 1 \le t \le n$. Denote $S_n = \sum X_i$. Then for all $a \ge 0$

$$\begin{aligned} \mathbb{P}\{S_n \ge n\mu + a\} &\le e^{-2a^2/n} \\ \mathbb{P}\{S_n \le n\mu - a\} &\le e^{-2a^2/n} \end{aligned} \quad (11)$$

***Theorem** 2:* Under the CWF2 policy, the expected total number of times that non-optimal power allocations are selected is at most

$$\mathbb{E}[T^{\pi}_{non}(n)] \le \frac{N(L+1)\ln n}{B^2_{\min}} + N + \frac{\pi^2}{3}LN, \quad (12)$$

where $B_{\min}$ is a constant defined by $\delta_{\min}$ and L; $\delta_{\min} = \min_{\mathbf{a}: r_{\mathbf{a}} < r^*} (r^* - r_{\mathbf{a}})$.

Proof: See [22]. ∎

***Remark** 4:* CWF2 can be used to solve the stochastic water-filling with objective $\mathbf{O_1}$ as well if $\exists \mathbf{a}^* \in \mathcal{O}^*$, such that $\forall \mathbf{a} \notin \mathcal{O}^*$,

$$\sum_{i\in\mathcal{A}_{\mathbf{a}^*}} f_i(a_i, \theta_i)) > \sum_{j\in\mathcal{A}_{\mathbf{a}}} f_j(a_j, \theta_j). \quad (13)$$

Then the regret of CWF2 is at most

$$\mathfrak{R}^{CWF2}(n) \le \left[\frac{N(L+1)\ln n}{B^2_{\min}} + N + \frac{\pi^2}{3}LN \right] \Delta_{\max}, \quad (14)$$

REFERENCES

[1] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[2] A. J. Goldsmith and P. P. Varaiya, "Capacity of Fading MIMO Channels with Channel Estimation Error," *IEEE International Conference on Communications (ICC)*, June, 2004.

[3] A. J. Goldsmith, *Wireless Communications.* New York: Cambridge University Press, 2005.

[4] X. Wang, D. Wang, H. Zhuang, and S. D. Morgera, "Energy-Efficient Resource Allocation in Wireless Sensor Networks over Fading TDMA," vol. 28, no. 7, pp. 1063-1072, 2010.

[5] I. Zaidi and V. Krishnamurthy, "Stochastic Adaptive Multilevel Waterfilling in MIMO-OFDM WLANs," *the 39th Asilomar Conference on Signals, Systems and Computers*, October, 2005.

[6] Y. Gai, B. Krishnamachari and R. Jain, "Combinatorial Network Optimization with Unknown Variables: Multi-Armed Bandits with Linear Rewards and Individual Observations", to appear in *IEEE/ACM Transactions on Networking*.

[7] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part I: IID Rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968-976, 1987.

[8] R. Agrawal, "Sample Mean Based Index Policies with O(log n) Regret for the Multi-Armed Bandit Problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054-1078, 1995.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, no. 2, pp. 235-256, 2002.

[10] Y. Gai, B. Krishnamachari, and R. Jain, "Learning Multiuser Channel Allocations in Cognitive Radio Networks: A Combinatorial Multi-armed Bandit Formulation", *IEEE International Dynamic Spectrum Access Networks (DySPAN) Symposium*, Singapore, April, 2010.

[11] A. Anandkumar, N. Michael, and A.K. Tang, "Opportunistic Spectrum Access with Multiple Users: Learning under Competition," *IEEE International Conference on Computer Communications (INFOCOM)*, March, 2010.

[12] A. Anandkumar, N. Michael, A. Tang, and A. Swami, "Distributed Learning and Allocation of Cognitive Users with Logarithmic Regret," *IEEE JSAC on Advances in Cognitive Radio Networking and Communications*, vol. 29, no. 4, pp. 781-745, 2011.

[13] K. Liu and Q. Zhao, "Distributed Learning in Cognitive Radio Networks: Multi-Armed Bandit with Distributed Multiple Players", *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March, 2010.

[14] Y. Gai and B. Krishnamachari, "Decentralized Online Learning Algorithms for Opportunistic Spectrum Access," *IEEE Global Communications Conference (GLOBECOM)*, December, 2011.

[15] C. Tekin and M. Liu, "Online Algorithms for the Multi-Armed Bandit Problem with Markovian Rewards," *the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, September, 2010.

[16] C. Tekin and M. Liu, "Online Learning in Opportunistic Spectrum Access: a Restless Bandit Approach," *IEEE International Conference on Computer Communications (INFOCOM)*, April, 2011.

[17] H. Liu, K. Liu, and Q. Zhao,"Learning and Sharing in a Changing World: Non-Bayesian Restless Bandit with Multiple Players," *Information Theory and Applications Workshop (ITA)*, January, 2011.

[18] W. Dai, Y. Gai, B. Krishnamachari and Q. Zhao, "The Non-Bayesian Restless Multi-Armed Bandit: a Case of Near-Logarithmic Regret," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May, 2011.

[19] Y. Gai, B. Krishnamachari and M. Liu, "On the Combinatorial Multi-Armed Bandit Problem with Markovian Rewards," *IEEE Global Communications Conference (GLOBECOM)*, December, 2011.

[20] Y. Gai, B. Krishnamachari and M. Liu, "Online learning for combinatorial network optimization with restless markovian rewards," arXiv:1109.1606.

[21] D. Pollard, *Convergence of Stochastic Processes*. Berlin: Springer, 1984.

[22] Y. Gai and B. Krishnamachari, "Online Learning Algorithms for Stochastic Water-Filling," arXiv:1109.2088.