Toward Order Optimal Channel Access in Unknown Environments: An Online Learning Method

Shuai Ye and Shaowei Wang

***Abstract*—We investigate the opportunistic channel access problem in a centralized cognitive radio network, where the cognitive base station (CBS) periodically detects spectrum holes in the licensed network and coordinates the unlicensed secondary users to utilize the idle channels. Existing spectrum access mechanisms typically rest on the assumption that the spectrum environment is known in advance, i.e., licensed channel states exhibit either stochastic or adversarial variations across different time instances and locations. These approaches address the spectrum access problem from the perspective of online learning, which relies heavily on the prior knowledge of network parameters such as the time horizon and the user activities to tune hyperparameters. In this paper, we tackle the multiuser channel access task by formulating it as a combinatorial multi-armed bandit problem in an unknown environment, where we propose an online mirror descent-based channel access method for the CBS that does not require any prior knowledge of the environment. Our proposed method adaptively adjusts the probability of secondary users accessing the licensed channels based on the historical transmission feedback, achieving order-optimal performance in both stochastic and adversarial environments. Numerical results validate the theoretical analysis and also demonstrate that the proposed method outperforms others under various network settings.**

***Index Terms*—Cognitive radio, multi-armed bandit, online mirror descent, opportunistic channel access.**

I. Introduction

The rapid proliferation of wireless devices and mobile applications, such as smart home and healthcare systems, has led to a significant surge in demand for spectrum resources [1]. Meanwhile, global measurements of spectrum usage have revealed that a substantial portion of licensed spectrum remains underutilized due to the static spectrum management policies [2]. Cognitive radio (CR) addresses the spectrum scarcity problem by enabling unlicensed secondary users (SUs) to opportunistically access spectrum holes in licensed networks [3–5]. By leveraging these underutilized resources, CR enhances spectrum utilization and meets the diverse transmission requirements of wireless devices.

In a CR network, the unlicensed SUs are permitted to access the licensed channels only when those channels are not in use by the licensed primary users (PUs) [6–8]. A critical challenge in implementing such a spectrum sharing is acquiring accurate knowledge of PU activity, particularly the idle and busy states of the licensed channels. Spectrum sharing in CR networks includes two approaches: fixed spectrum assignment and dynamic spectrum access (DSA). Fixed spectrum assignment, which relies on a dependable spectrum usage database for channel availability information [9], requires significant signaling and raises privacy concerns due to the data exchange between the database and the SUs.

In the DSA case, each SU performs spectrum sensing to determine the availability of licensed channels. To avoid interference with PUs, SUs access the licensed channels only when their sensing results indicate that those channels are idle. DSA can be further categorized into centralized and decentralized approaches. For the decentralized approach, the SUs access licensed channels by executing their own DSA policies, which is suitable for ad-hoc networks [10–12]. This individual decision-making at each SU leads to increased computation overhead, power consumption, and additional implementation costs. Moreover, the SUs may act selfishly and fail to collaborate, leading to network failures such as collisions. The centralized DSA, where a cognitive base station (CBS) manages spectrum allocation and avoids collisions, is well-suited for smaller networks like smart homes, where an access point acts as the CBS and devices function as SUs [13, 14]. AS a result, the SUs need not to perform power-hungry spectrum sensing, with the primary cost being communication between the CBS and the SUs.

In both centralized and decentralized DSA systems, SUs perform spectrum sensing and access idle channels per time slot, adjusting their channel selection policies based on historical sensing results and transmission feedback, which falls into an online learning framework. Due to hardware constraints and power limitations, each SU can only choose a subset of channels to sense per slot. This multiuser DSA task is commonly formulated as a multi-armed bandit (MAB) problem [10, 11], where each SU is seen as a player facing a bandit machine with multiple arms. In each round, the player would pull an arm, representing a particular licensed channel here, and receive a reward related to the arm. Depending on the modeling of channel statistics, these works can be categorized as stochastic MAB [10–19] and non-stochastic (also referred to as adversarial) MAB [20–23]. For the stochastic MAB, the states of licensed channels are described by a parametric family of probability distributions. Typically, the state of each channel is modeled as a Bernoulli process with an unknown parameter [10–17]. In [18, 19], channel states evolve as a two-state Markov chain, with the stationary distribution equivalent to a Bernoulli distribution. The adversarial MAB, on the other

Manuscript received October 26, 2023; revised May 30, 2024 and July 13, 2024; accepted July 16, 2024. This work was partially supported by the National Natural Science Foundation of China under Grants 61931023. The associate editor coordinating the review of this article and approving it for publication was C. Li. (*Corresponding author: Shaowei Wang.*)

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: dz20230027@smail.nju.edu.cn; wangsw@nju.edu.cn).

hand, generally assumes malicious actors who can manipulate sensing results. A common strategy is the PU emulation attacks, where attackers forge PU signals on licensed channels, tricking SUs into believing they are occupied [24, 25].

Existing channel access strategies typically assume a fully stochastic or entirely adversarial environment for licensed channel states [26, 27]. These approaches aim to maximize the SU throughput by minimizing regret, the performance difference between the proposed strategy and the optimal one known in hindsight. While the stochastic MAB offers logarithmic regret bounds [28], the adversarial MAB guarantees square root bounds [29, 30]. However, a purely stochastic model is generally unrealistic even in the absence of adversarial behavior. Occasional disruptive events, such as burst movements of users and jitter effects of electromagnetic waves, can disrupt the stochastic nature of channels. Consequently, channel access strategies based on stochastic assumptions may face practical implementation challenges, and the validity of their theoretical performance becomes uncertain. Conversely, assuming a completely adversarial environment is overly pessimistic, as a substantial portion of channels can still exhibit stochastic behavior. In general, it is hard to decide which type of MAB model should be prioritized for a given scenario. Additionally, existing online learning methods for the DSA require prior network information, such as time horizon and PU activity, to tune hyperparameters and ensure theoretical performance guarantees, making them difficult for applications.

In this paper, we investigate the multiuser DSA problem in an unknown environment, formulating the optimization task as a combinatorial MAB and categorizing wireless environment features into three typical regimes: stochastic, periodically stochastic, and stochastic with adversarial corruptions. The latter two regimes account for time-varying activity patterns of the PUs or the PU emulation attacks. We propose an online mirror descent (OMD) based channel access algorithm with Tsallis entropy as the regularization function. OMD is a versatile online learning algorithm that computes the current action using a simple gradient update rule. The versatility of the OMD arises from performing updates in a dual space, defined by the choice of regularization, leading to improved bounds based on the geometry of the transformed space. Our algorithm achieves competitive performance guarantees in both adversarial and stochastic environments, with problem-dependent logarithmic regret bounds in the latter case. The main contributions of this work are summarized as follows.

- We make no assumptions about the characteristics of the licensed channel states and propose an OMD-based channel access algorithm for the CBS, which does not require any prior knowledge of the environment nature or network parameters.
- Theoretical analysis demonstrates that the proposed algorithm achieves order optimality in both stochastic and adversarial MAB settings. Specifically, for the stochastic case, the leading constant of regret matches the asymptotic lower bound within a multiplicative factor of 2 [28], while for the adversarial case, the constant matches the known lower bound within a multiplicative factor of less than 40 [31], representing the best leading constant in adversarial regret bounds known to date.
- We comprehensively compare the proposed algorithm with representative ones in the literature. The experimental setup takes practical features of wireless environments into consideration, encompassing existing stochastic and adversarial channel state models while providing enhanced adaptability to attacker strategies.

TABLE I.
LIST OF MAIN NOTATIONS

$a^\star$	best action in hindsight
a_t	action of CBS at slot t
$\mathcal{A}$	action space of CBS
A_i	i-th action in the action space
$k^\star$	best channel with the largest idle probability
K	number of channels
l_t	loss of actions at slot t
$\hat{l}_t$	estimated loss of actions at slot t
$\hat{L}_t$	estimated cumulative loss of actions till slot t
M	number of SUs
p_t	sampling distribution of actions at slot t
$q_{t,k}$	selection probability of channel k at slot t
$r_m(t)$	reward of SU m at slot t
$s_k(t)$	state of channel k at slot t
S	number of actions
T	total time slots
γ	power of Tsallis entropy
Δ_k	idle probability gap between channel k and $k^\star$
η_t	learning rate
Ω_t	set of attacked channels at slot t
μ_k	idle probability of channel k
Ψ_t	regularization function at slot t

The rest of the paper is organized as follows. Related work is discussed in Section II. In Section III, we present the communication model along with three environment regimes and the formulation of the problem. The proposed OMD-based channel access algorithm and its regret analysis are given in Section IV and Section V, respectively. Numerical results are provided in Section VI. Finally, we conclude the paper in Section VII. The main notations used in this paper are summarized in Table I.

II. RELATED WORK

In the stochastic DSA, the state of each channel in each slot is independently and identically distributed (i.i.d.). One way to learn the unknown channel statistics is to randomly sense the licensed channels and estimate the idle probability of each channel by averaging historical sensing results [10, 15]. Random sensing ensures comprehensive exploration but requires prior knowledge of the idle probability gap among channels for optimal sensing allocation. To get rid of the prior knowledge, an upper confidence bound (UCB) algorithm is introduced to DSA [11, 32]. The UCB uses sample mean and sample variance observed so far to give an overestimate of the unknown idle probability, which achieves asymptotically optimal performance. Another approach calculates the UCB

index of each channel considering both the idle probability and the transmission rate of licensed channels [27]. A Bayesian method called Thompson sampling (TS) represents its uncertainty about the true idle probability of each channel by a prior distribution and at each slot selects the channel with the highest posterior probability of being optimal [19, 33].

Adversarial DSA considers malicious attacks and frequent changes in the activity patterns of PUs, where the states of licensed channels are not i.i.d. In [34], a random algorithm called exponential weights for exploration and exploitation (EXP3) is proposed with a pre-defined timeframe to adjust its learning rate, achieving a square-root regret with respect to time. It utilizes the cumulative reward of each channel as its weight exponent and calculates the selection probability of each channel based on the normalization of the weights. In [21], a new control parameter is introduced into EXP3, which uses transmission rate as feedback and maintains order optimality in the adversarial regime while simultaneously achieving a logarithmic squared regret in stochastic regime. One drawback of the EXP3 algorithm is that the distribution of its regret has a large variance when the selection probability of a licensed channel becomes small. In [35], a reduced-variance version of the EXP3 is discussed by introducing an additional exploration parameter when calculating the selection probability of each channel.

Existing DSA researches rarely jointly consider the stochastic and the adversarial cases. Designing algorithms that achieve optimal regret rates in both of the MAB models has recently gained significant attention [36]. One strategy to accomplish this goal involves starting to play under the assumption that the characteristics of the environment is i.i.d., while continually monitoring whether the assumption holds [37]. If a deviation from the stochastic assumption is detected, the algorithm undergoes an irreversible switch to an adversarial mode of operation. This approach requires prior knowledge of the time horizon, and its regret in the adversarial case suffers from a multiplicative logarithmic factor. Another approach involves altering the EXP3 algorithm to achieve improved regret in the stochastic case without compromising adversarial guarantees [38], which introduces an extra logarithmic factor in the adversarial case. However, the leading constant in the stochastic regret is quite large.

The multi-SU scenario, where multiple SUs are in close proximity, necessitates collision avoidance mechanisms to prevent invalid transmissions. Collisions occur when multiple SUs attempt to access the same channel simultaneously, resulting in unsuccessful transmissions for all involved [10, 11]. In stochastic DSA, licensed channels can be sorted based on the estimated idle probabilities obtained through learning algorithms. Decentralized coordination methods, such as pre-agreement, can be employed to enable SUs to access the estimated best channels in a round-robin fashion [16, 17]. Another distributed coordination method allows an SU to persistently sense the same channel if it achieves a successful transmission on that channel [12]. This method, inspired by the game Musical Chairs, ensures that the SUs choose different channels to access within a limited number of slots. On the other hand, centralized coordination methods require that the CBS arranges for the SUs to sense different channels, ensuring collision-free access at each slot [14, 32, 33]. In adversarial DSA, where channel states are not i.i.d. and the optimal channel can change dynamically, index-based methods like pre-agreement and Musical Chairs become ineffective. Consequently, centralized coordination through a CBS is typically employed to manage the challenges posed by the adversarial environment and ensure efficient spectrum access [20].

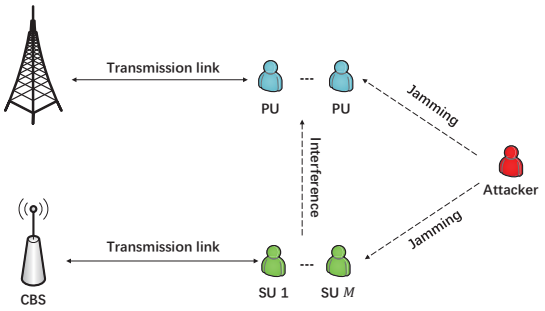


Fig. 1. A cognitive radio network with one CBS and M SUs.

III. System Model And Problem Formulation

A. System Model

We consider a centralized cognitive radio network comprising one CBS and $\mathcal{M} = \{1, 2, \ldots, M\}$ SUs, as shown in Fig. 1. This centralized CR network coexists with a primary network, which provides $\mathcal{K} = \{1, 2, \ldots, K\}$ non-overlapping orthogonal frequency channels with equal bandwidth. To avoid congestion, we assume that $M \leq K$. While these channels are licensed to the PUs, they are also shared with the unlicensed SUs through the coordination of the CBS. Due to the limited geographical coverage of the SUs, we assume that the state of each channel is consistent for all SUs.

The communication system operates in a time-synchronized manner with discrete units of time called slots. In each slot $t \in [1, T]$, where T represents the time horizon, the CBS selects M out of K channels from the channel pool $\mathcal{K}$ and assigns each channel to an SU. Subsequently, the SUs sense their assigned channels and transmit data over them only if the sensing results indicate an idle state. If the sensing result indicates a busy state, i.e., a PU signal is present, the SU remains in sleep mode. At the end of slot t, the SUs report their sensing results back to the CBS. The states of the licensed channels are not known to the CBS scheduler a priori.

B. Environment Regime

We denote the state of channel $k \in \mathcal{K}$ in slot t as $s_k(t) \in \{0, 1\}$, where $s_k(t) = 1$ represents that the channel is idle and $s_k(t) = 0$ represents that the channel is busy. The unknown features of the environment are categorized into three typical regimes: stochastic, periodically stochastic, and stochastic with adversarial corruptions.

1) Stochastic Regime: The state $s_k(t)$ of channel k is sampled from a Bernoulli distribution at each slot t. Let μ_k denote the idle probability of channel k, i.e., $\mathbb{E}[s_k(t) = 1] = \mu_k$. The best channel is defined as the one with the largest

idle probability, denoted as $k^\star = \text{argmax}_{k\in\mathcal{K}} \mu_k$. The idle probability gap between a suboptimal channel k and the best channel $k^\star$ is denoted by $\Delta_k = \mu_{k^\star} - \mu_k$.

2) Periodically Stochastic Regime: The idle probabilities of the licensed channels vary with time due to changes in PU activity, e.g., off time and peak time. The idle probability gaps between channels are kept fixed and the best channel remains the same for all time slots.

3) Stochastic Regime with Adversarial Corruptions: The idle probability settings are consistent with those in the periodically stochastic regime. PU emulation attacks can take place in every slot and target different channels, leading to a time-varying best channel in the perspective of the SUs. In this paper, we consider an oblivious attacker who does not react to the actions of the CBS. For adaptive attackers who adjust their attacking strategy based on the actions of the CBS, it has been proven that no policy can achieve a sub-linear regret, i.e., no learning algorithm can be effective [35].

C. Problem Formulation

We formulate the DSA problem as a combinatorial MAB with semi-bandit feedback. The action space of the CBS is denoted as $\mathcal{A} = \{a \subset \mathcal{K} : |a| = M\}$, where $|a|$ represents the cardinality of set a. The action space $\mathcal{A}$ consists all possible combinations of M channels from the channel pool $\mathcal{K}$, resulting in a total of $S = \binom{K}{M}$ actions. Note that the order of SUs is not considered. No priorities exist among the SUs. In each slot t, the CBS selects an action $a_t = \{a_{t,1}, a_{t,2}, \ldots, a_{t,M}\} \in \mathcal{A}$, where $a_{t,m}$ represents the channel assigned to SU $m \in \mathcal{M}$. Meantime, an attacker chooses a subset of channels $\Omega_t \subset \mathcal{K}$ to attack. SU m receives a busy sensing result if $a_{t,m} \in \Omega_t$. We define the reward of SU m at slot t as

$$r_m(t) = s_{a_{t,m}}(t)\mathbb{1}(a_{t,m} \notin \Omega_t), \tag{1}$$

where $\mathbb{1}$ represents the indication function. Thus, only when the sensed channel is idle and not attacked by the adversary will SU m transmit successfully. The goal is to maximize the cumulative reward of all SUs over time. To achieve this, the CBS employs an online learning algorithm ν for channel selection. The expected cumulative reward that this algorithm ν achieves over T slots is denoted as

$$G_\nu(T) = \mathbb{E}[\sum_{t=1}^{T}\sum_{m\in\mathcal{M}} r_m(t)], \tag{2}$$

where the expectation is taken with respect to the internal randomness of the algorithm ν and the environment.

We compare the proposed algorithm with the best action in hindsight. We assume the existence of an ideal policy that possesses full prior knowledge of the environment and always selects the best action for the SUs in each slot. The maximum accumulated reward of this ideal policy is given by

$$G_{\max}(T) = \max_{a\in\mathcal{A}} \mathbb{E}[\sum_{t=1}^{T}\sum_{k\in a} s_k(t)\mathbb{1}(k \notin \Omega_t)], \tag{3}$$

where the expectation is taken with respect to the randomness of environment. The regret of algorithm ν is then defined as the performance gap between it and the ideal policy,

$$R(T) = G_{\max}(T) - G_\nu(T). \tag{4}$$

In the stochastic regime, the regret $R(T)$ grows at least logarithmically with time and the leading constant of $R(T)$ is determined by the idle probability gap between channels. For the special case of a single SU $M = 1$, an algorithm ν is considered asymptotically optimal if its regret scales as [28]

$$\lim_{T\to\infty} \frac{R(T)}{\log(T)} = \sum_{k\neq k^\star} \frac{\Delta_k}{\text{KL}(\mu_{k^\star}||\mu_k)}, \tag{5}$$

where $\text{KL}(x||y) = x\log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y})$ is the Kullback-Leibler divergence between two Bernoulli distributions with biases x and y. In the adversarial regime, the known minimax lower bound for regret is expressed as [31]

$$R(T) \geq \frac{1}{20}\sqrt{KT}. \tag{6}$$

To the best of our knowledge, the regret of the EXP3 algorithm is $\sqrt{KT\log(K)}$, which is nearly optimal since this regret cannot be improved significantly in the worst case. We expect to design an algorithm ν that achieves a $\log(T)$ regret rate in the stochastic regime while simultaneously achieving a $\sqrt{T}$ regret rate in the adversarial regime. Moreover, the leading constant of the regret is within a constant multiple of the known optimal bound for the single-SU case.

IV. Proposed Algorithm

The proposed channel access algorithm leverages OMD with Tsallis entropy as the regularization function to calculate the selection probability of each action. We first introduce the general framework of the OMD and then delve into the finer points of implementing the algorithm, includes determining the power of Tsallis entropy and achieving unbiased estimation of channel rewards.

Let A_i denote the i-th action in $\mathcal{A}$, $1 \leq i \leq S$. Within each slot t, the proposed OMD algorithm maintains a distribution $p_t = [p_{t,1}, p_{t,2}, \ldots, p_{t,S}]$ over all possible actions, where $p_{t,i}$ represents the probability associated with selecting action A_i. The CBS samples an action a_t for the SUs according to the distribution, i.e., $a_t \sim p_t$. As each SU reports back the sensing result $s_{a_{t,m}}(t)$, the distribution p_t is updated. The OMD is an online learning algorithm rooted in the concept of loss $l_t = [l_{t,1}, l_{t,2}, \ldots, l_{t,S}]$ for each action, where $l_{t,i}$ signifies the loss encountered at action A_i in slot t. For a given reward model, OMD can be effectively applied through a simple manipulation, where $l_{t,i} = M - \sum_{k\in A_i} s_k(t)\mathbb{1}(k \notin \Omega_t)$. The overarching framework of OMD for updating p_t is then defined as follows

$$\begin{aligned} p_1 &= \underset{p\in\Theta^{S-1}}{\text{argmin}}\, \Psi_1(p), \\ p_{t+1} &= \underset{p\in\Theta^{S-1}}{\text{argmin}} \langle p, l_t\rangle + B_{\Psi_{t+1}}(p, p_t), \end{aligned} \tag{7}$$

where Θ^{S-1} is the S-dimensional simplicity, Ψ_t is the regularization function at slot t and $B_\Psi(x,y) = \Psi(x) - \Psi(y) - \langle x-y, \triangledown\Psi(y)\rangle$ is the Bregman divergence between x and y. On the one hand, random sampling gives each action a chance to be selected, facilitating the exploration of actions. On the other hand, actions with lower losses, as will be discussed

later, are granted higher selection probabilities, promoting the exploitation of more favorable actions.

The regularization function Ψ_t is constructed based on the Tsallis entropy $H_\gamma(x) = \frac{1}{1-\gamma}(1 - \sum_i x_i^\gamma)$ with power γ. We modify the scaling and introduce linear terms to the Tsallis entropy, resulting in the following time-varying regularizer

$$\Psi_t(p) = -\frac{1}{\eta_t}\sum_{i=1}^{S}\frac{p_i^\gamma - \gamma p_i}{\gamma(1-\gamma)}, \tag{8}$$

where η_t represents the learning rate of the OMD algorithm. As previously observed, EXP3 can be viewed as an OMD algorithm utilizing the negative Shannon entropy $\sum_i p_i \log(p_i)$ as the regularization function [39, 40]. When $\gamma \to 1$, our proposed algorithm is essentially equivalent to the negative Shannon entropy, differing only in linear and constant terms. Thus, the EXP3 algorithm can be regarded as a specialized instance of our proposed OMD algorithm. In this paper, we set $\gamma = 1/2$, which works well in both stochastic and adversarial cases. For Tsallis entropy with power $\gamma \neq 1/2$, the OMD algorithm requires prior information about the performance gap between channels to change the scaling of regularization function Ψ to meet theoretical guarantee [36]. Thus, $\gamma = 1/2$ is the only value we care about and it serves as the focus of the theoretical analysis in Section V.

In the considered bandit setting, the CBS does not have access to the entire loss vector l_t. Instead, only the element of action a_t is disclosed to the CBS at the end of each slot t. To update p_t, an unbiased estimator $\hat{l}_t$ is employed, satisfying the condition $\mathbb{E}_{a_t \sim p_t}[\hat{l}_t] = l_t$. A prevalent technique for creating an unbiased loss estimator is through importance-weighted (IW) sampling, formulated as follows

$$\hat{l}_{t,i} = \sum_{k \in A_i} \frac{\mathbb{1}(k \in a_t)(1 - s_k(t)\mathbb{1}(k \notin \Omega_t))}{q_{t,k}}, \tag{9}$$

where $q_{t,k} = \sum_{i:k\in A_i} p_{t,i}$ is the selection probability of channel k. However, an inherent limitation of the IW estimator lies in its susceptibility to increased estimation variance when the selection probability $q_{t,k}$ becomes small. To address this concern, we adopt a reduced variance (RV) version of the IW estimator, expressed as

$$\hat{l}_{t,i} = \sum_{k \in A_i} \frac{\mathbb{1}(k \in a_t)(1 - s_k(t)\mathbb{1}(k \notin \Omega_t) - b_k)}{q_{t,k}} + b_k, \tag{10}$$

where $b_k = \mathbb{1}(q_{t,k} \geq \eta_t^2)/2$. Notably, the RV estimator can be seen as the IW estimator plus $b_k(1 - \frac{\mathbb{1}(k\in a_t)}{q_{t,k}})$. The second term acts as a control variate, which preserves unbiasedness since the expectation of it is zero. Besides, the second term is negatively correlated with the IW estimator, thus reducing the second moment and mitigating the issue of variance amplification. In the classic OMD, only the loss of the selected action a_t is updated at each slot. In our proposed algorithm, the losses for actions that share one or more common channels with the selected action are also updated. It can be regarded that we first construct an unbiased loss estimator $\hat{l}_t(k) = \frac{\mathbb{1}(k\in a_t)(1-s_k(t)\mathbb{1}(k\notin\Omega_t)-b_k)}{q_{t,k}} + b_k$ for each channel k, and then update the loss of each action based on the channel it contains.

Algorithm 1 Newton Method Approximation of x

Require: x, $\hat{L}_t$ and η_t
1: **repeat**
2: $\quad p_{t,i} \leftarrow 4(\eta_t(\hat{L}_{t,i} - x))^{-2}$
3: $\quad x \leftarrow x - (\sum_i p_{t,i} - 1)/(\eta_t \sum_i p_{t,i}^{\frac{3}{2}})$
4: **until** Convergence

Algorithm 2 OMD Based Channel Selection for CBS

1: **Input** $\mathcal{K}$, $\hat{L}_1 = \mathbf{0}$ and $\{\eta_t, t \in [1, T]\}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad$ set $p_{t,i}$ for each action $A_i \in \mathcal{A}$ as in Algorithm 1
4: $\quad$ sample $a_t \sim p_t$
5: $\quad$ receive results $s_{a_{t,1}}(t)$, $s_{a_{t,2}}(t)$, …, $s_{a_{t,M}}(t)$
6: $\quad$ **for** $k \in \mathcal{K}$ **do**
7: $\quad\quad$ compute $q_{t,k} = \sum_{i:k\in A_i} p_{t,i}$
8: $\quad$ **end for**
9: $\quad$ use RV (10) to construct $\hat{l}_{t,i}, 1 \leq i \leq S$
10: $\quad$ update loss $\hat{L}_{t+1,i} = \hat{L}_{t,i} + \hat{l}_{t,i}, 1 \leq i \leq S$
11: **end for**

The determination of the selection probability $p_{t,i}$ for each action A_i involves solving an optimization problem. Below, we present an implicit form of the solution, detailing the method. The cumulative estimated loss of actions is denoted as $\hat{L}_t = [\hat{L}_{t,1}, \hat{L}_{t,2}, \ldots, \hat{L}_{t,S}]$ with $\hat{L}_{t,i} = \sum_{n=1}^{t-1} \hat{l}_{n,i}$ representing the cumulative loss for action A_i until time slot t. It has been established that the solution for the optimization problem (7) with Tsallis entropy power $\gamma = 1/2$ and learning rate η_t takes the following structure

$$p_{t,i} = 4(\eta_t(\hat{L}_{t,i} - x))^{-2}. \tag{11}$$

Here x denotes a normalization factor, implicitly defined through the constraint

$$\sum_{i=1}^{S} 4(\eta_t(\hat{L}_{t,i} - x))^{-2} = 1. \tag{12}$$

Efficiently approximating the normalization factor x is attainable through Newton's Method, achieving a desired level of precision in just a few iterations. Algorithm 1 outlines the computational steps in more detail, with x from the preceding iteration utilized as a warm start.

The operational procedure of the OMD is outlined below. During each time slot t, the CBS computes the selection probability of each action in $\mathcal{A}$ using the method laid out in Algorithm 1. Subsequently, the CBS samples an access action $a_t \sim p_t$ for the SUs and the SUs relay their respective sensing results $[s_{a_{t,1}}(t), s_{a_{t,2}}(t), \ldots, s_{a_{t,M}}(t)]$ back to the CBS. The CBS calculates the selection probability for each licensed channel. The estimated loss for each action is calculated based on the RV estimator (10). Further details of the proposed algorithm are provided in Algorithm 2.

V. Regret Analysis

We present the main results of our proposed OMD algorithm with RV estimator in different regimes. We will collectively

refer to the stochastic regime and the periodically stochastic regime as stochastic regime. We follow the standard OMD analysis in [35] and introduce the potential function $\Phi_t(-L) = \max_{p\in\Theta^{S-1}}\{\langle p,-L\rangle - \Psi_t(p)\}$ to decompose the regret into *stability* and *penalty* terms

$$\begin{aligned} R(T) &= \mathbb{E}[\sum_{t=1}^{T}(\sum_{k\in a_t} l_t(k) - \sum_{k\in a^\star} l_t(k))] \\ &= \underbrace{\mathbb{E}[\sum_{t=1}^{T}(\sum_{k\in a_t} l_t(k) + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1}))]}_{\textit{stability}} \quad (13) \\ &+ \underbrace{\mathbb{E}[\sum_{t=1}^{T}(-\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \sum_{k\in a^\star} l_t(k))]}_{\textit{penalty}}, \quad (14) \end{aligned}$$

where $l_t(k) = 1 - s_k(t)\mathbb{1}(k \notin \Omega_t)$. Note that we use loss model in the regret analysis and $a^\star$ is the best action in hindsight that yields the lowest loss over T slots. In each regime, the analysis bounds the *stability* and *penalty* terms separately.

A. Adversarial Regime

Theorem 1: The regret of OMD using RV estimator with $\gamma = 1/2$ and learning rate $\eta_t = 4\sqrt{1/t}$ in any adversarial regime satisfies

$$R(T) \le (MK^{\frac{1}{2}} + S^{\frac{1}{2}})\sqrt{T} + 16M + (6MK + 8SM^3)\log(T).$$

The bound is order optimal since it achieves $\sqrt{T}$ regret rate. Besides, the leading constant matches the known lower bound within a multiplicative factor of less than 40 for $M = 1$.

Proof: Let $A_{i,k}$ denote the $k-$th channel in action A_i and $a_{t,k^\star}$ denote the channel with the lowest loss in the selected action a_t. For positive learning rate $0 < \eta_t \le 1$ and power $\gamma = \frac{1}{2}$, the instantaneous *stability* of OMD with Tsallis entropy at any slot t, denoted as s_t, satisfies

$$\begin{aligned} s_t &\le \mathbb{E}[\sum_{i=1}^{S}\frac{\eta_t}{2}p_{t,i}^{\frac{3}{2}}(\sum_{k\in A_i}\hat{l}_t(k) - \sum_{k\in a_t} l_t(k))^2] \\ &+ \mathbb{E}[\sum_{i=1}^{S}\frac{\eta_t^2}{2}p_{t,i}^2|\sum_{k\in A_i}\hat{l}_t(k) - \sum_{k\in a_t} l_t(k)|^3] \\ &\le \mathbb{E}[\sum_{i=1}^{S}\frac{\eta_t}{2}p_{t,i}^{\frac{3}{2}}M\sum_{k=1}^{M}(\hat{l}_t(A_{i,k}) - l_t(a_{t,k}))^2] + \frac{SM^3\eta_t^2}{2} \\ &\le \mathbb{E}[\sum_{k=1}^{K}\frac{M\eta_t}{2}(\hat{l}_t(k) - l_t(a_{t,k^\star}))^2\sum_{i:k\in A_i}p_{t,i}^{\frac{3}{2}}] + \frac{SM^3\eta_t^2}{2} \\ &\le M\mathbb{E}[\sum_{k=1}^{K}\frac{\eta_t}{2}q_{t,k}^{\frac{3}{2}}(\hat{l}_t(k) - l_t(a_{t,k^\star}))^2] + \frac{SM^3\eta_t^2}{2} \\ &\le M\sum_{k=1}^{K}\frac{\eta_t}{8}\mathbb{E}[q_{t,k}]^{\frac{1}{2}}(1 - \mathbb{E}[q_{t,k}]) + \frac{M\eta_t^2(3K + 4SM^2)}{8} \\ &\le \frac{M\sqrt{K}\eta_t + M\eta_t^2(3K + 4SM^2)}{8}. \quad (15) \end{aligned}$$

The first and penultimate inequalities follow the results in [36]. In the second inequality, Cauchy-Schwarz inequality is used to bound the first expectation. The second part of expectation is derived using $|\sum_{k\in A_i}\hat{l}_t(k) - \sum_{k\in a_t} l_t(k)| \le \sum_{k=1}^{M}|\hat{l}_t(A_{i,k}) - l_t(a_{t,k})|$ and Lemma 11 in [36]. In the third and fourth equation, we transform the regret analysis for each action into the regret analysis for each channel and use the inequality $\sum_{i:k\in A_i} p_{t,i}^{\frac{3}{2}} \le (\sum_{i:k\in A_i} p_{t,i})^{\frac{3}{2}} = q_{t,k}^{\frac{3}{2}}$. The last inequation is derived using $\sum_{k=1}^{K}\mathbb{E}[q_{t,k}]^{\frac{1}{2}}(1 - \mathbb{E}[q_{t,k}]) \le \sum_{k=1}^{K}\sqrt{\mathbb{E}[q_{t,k}]} \le \sqrt{K}$. For $T < 16$ where $\eta_t \ge 1$, the instantaneous *stability* at each slot is smaller than M. Thus, for $T \ge 16$, we have

$$\begin{aligned} \textit{stability} &\le 15M + \sum_{t=16}^{T}(\frac{M\sqrt{K}\eta_t}{8} + \frac{3MK\eta_t^2}{8} + \frac{SM^3\eta_t^2}{2}) \\ &= 15M + \sum_{t=16}^{T}(\frac{M\sqrt{K}}{2\sqrt{t}} + \frac{6MK}{t} + \frac{8SM^3}{t}) \\ &\le 15M + M\sqrt{KT} + (6MK + 8SM^3)\log(T). \quad (16) \end{aligned}$$

For the *penalty* term, it is proved in [36] that for any $\gamma \in [0,1]$,

$$\begin{aligned} &\mathbb{E}[\sum_{t=1}^{T}(-\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \sum_{k\in a^\star} l_t(k))] \\ &\le \mathbb{E}[\frac{\Psi(\theta) - \Psi(p_1)}{\eta_T}] + \langle\theta^\star, L_T\rangle, \quad (17) \end{aligned}$$

where θ is a vector belong to Θ^{S-1} and $\theta^\star$ is equal to θ except for $\theta^\star_{a^\star} = 0$. With a slight abuse of notation, we also use $a\star$ to represent the index of it within the set $\mathcal{A}$. Following the trick of [40], we set $\theta_{a^\star} = 1 - T^{-1}$ and $\theta_i^\star = \theta_i = \frac{T^{-1}}{S-1}$ for $i \ne a^\star$. The loss of each action A_i is bounded in $[0, MT]$, which implies $\langle\theta^\star, L_T\rangle \le M$. Since the loss of each action is zero initially, the explicit form of p_1 is $p_{1,i} = \frac{1}{S}$ for each action A_i. Function Ψ is equal to $\frac{\Psi_t}{\eta_t}$. Substituting in equation (17), we have

$$\begin{aligned} \textit{penalty} &\le \frac{4(S^{\frac{1}{2}} - (S-1)^{\frac{1}{2}}T^{-\frac{1}{2}} - \sqrt{1 - T^{-1}})}{\eta_T} + M \\ &\le \frac{4(S^{\frac{1}{2}} - 1)(1 - T^{-\frac{1}{2}})}{\eta_T} + M \\ &\le \sqrt{ST} + M, \quad (18) \end{aligned}$$

where in the second equation we use Taylor's expansion $u^{\frac{1}{2}} \le (u-1)^{\frac{1}{2}} + \frac{1}{2}(u-1)^{-\frac{1}{2}}$ for any $u > 1$. Combining the *stability* and *penalty* terms, we get the result.

B. Stochastic Regime

Theorem 2: The regret of OMD using RV estimator with $\gamma = 1/2$ and learning rate $\eta_t = 4\sqrt{1/t}$ in any stochastic regime satisfies

$$R(T) \le \sum_{k\ne k^\star}(\frac{\alpha^2\log(T)}{4\Delta_k} + \frac{\alpha^2\beta}{\Delta_k(4 - 2\beta)}) + 2C,$$

where $\alpha = M + \sqrt{\frac{S}{KM}}$, $\beta = 2M - \sqrt{\frac{S}{KM}}$ and C is equal to $\frac{M}{2\Delta_{\min}} + \frac{3}{4}\sqrt{S} + 16M + (6MK + 8SM^3)\log(T)$ for short. $\Delta_{\min} = \min_{k\ne k^\star}\Delta_k$ is the minimal idle probability gap between channels. The bound is order optimal as it achieves $\log(T)$ regret growth rate. Since the well known divergence-dependent lower bound (5) is larger than $\sum_{k\ne k^\star}\frac{1}{2\Delta_k}$, we show that for the special case $M = 1$,

$$\lim_{T\to\infty}\frac{R(T)}{\log(T)} = \sum_{k\ne k^\star}\frac{1}{\Delta_k} + 28K, \quad (19)$$

which matches the asymptotic lower bound within a multiplicative factor of 2.

Proof: In the stochastic case, we continue bounding the *stability* term up from the penultimate inequality from (15). For $k \neq k^\star$, we use $\sqrt{\mathbb{E}[q_{t,k}]}(1-\mathbb{E}[q_{t,k}]) \leq \sqrt{\mathbb{E}[q_{t,k}]}$. For best channel $k^\star$, we use $\sqrt{\mathbb{E}[q_{t,k^\star}]}(1-\mathbb{E}[q_{t,k^\star}]) \leq (1-\mathbb{E}[q_{t,k^\star}]) = \sum_{k\neq k^\star}\mathbb{E}[q_{t,k}]$. Further, we define $T_0 = \lceil (\frac{1}{2\Delta_{\min}})^2 \rceil$ and bound the last expression as $\sum_{k\neq k^\star}\mathbb{E}[q_{t,k}] \leq 1$ for $t < T_0$. The *stability* term can be rebounded as

$$\begin{aligned} \textit{stability} &\leq 15M + (6MK + 8SM^3)\log(T) + M\sqrt{T_0} \\ &\quad + M\sum_{k\neq k^\star}\Big(\sum_{t=T_0+1}^{T}\frac{\mathbb{E}[q_{t,k}]}{2\sqrt{t}} + \sum_{t=16}^{T}\frac{\sqrt{\mathbb{E}[q_{t,k}]}}{2\sqrt{t}}\Big). \end{aligned} \tag{20}$$

For the power $\gamma = 1/2$ and any unbiased estimators, the *penalty* term of OMD satisfies

$$\begin{aligned} \textit{penalty} &\leq \sum_{t=1}^{T}\Big(\sum_{i:A_i\neq a^\star}\frac{\sqrt{\mathbb{E}[p_{t,i}]}-\frac{1}{2}\mathbb{E}[p_{t,i}]}{2\sqrt{t}}\Big) + \frac{3}{4}\sqrt{S} \\ &\leq \sqrt{\frac{S}{KM}}\sum_{t=1}^{T}\Big(\sum_{k\neq k^\star}\frac{\sqrt{\mathbb{E}[q_{t,k}]}-\frac{1}{2}\mathbb{E}[q_{t,k}]}{2\sqrt{t}}\Big) + \frac{3}{4}\sqrt{S}, \end{aligned} \tag{21}$$

where the first equation follows the results in [36]. In the second equation, we use the inequality $\sum_{i:k\in A_i}\sqrt{p_{t,i}} \leq \sqrt{\binom{K-1}{M-1}q_{t,k}}$ to transform the regret analysis for each action into the regret analysis for each channel. The inequality is a generalization of mean value theorem $u_1+u_2+\ldots+u_n \leq \sqrt{n(u_1^2+u_2^2+\cdots+u_n^2)}$. Since each action contains M channels, which means $p_{t,i}$ is repeated calculation M times, the leading constant reduces to $\sqrt{\frac{S}{KM}}$. Combining the *stability* and *penalty* terms gives the bound

$$\begin{aligned} R(T) \leq \sum_{k\neq k^\star}\Big(&\sum_{t=1}^{T}\frac{(M+\sqrt{\frac{S}{KM}})\sqrt{\mathbb{E}[q_{t,k}]}}{2\sqrt{t}} \\ &+ \sum_{T_0+1}^{T}\frac{(2M-\sqrt{\frac{S}{KM}})\mathbb{E}[q_{t,k}]}{4\sqrt{t}}\Big) + C. \end{aligned} \tag{22}$$

In the stochastic regime, the regret of any online learning algorithm satisfies

$$R(T) = \sum_{t=1}^{T}\sum_{k\neq k^\star}\Delta_k\mathbb{E}[q_{t,k}]. \tag{23}$$

Then we can obtain

$$\begin{aligned} R(T) &\leq 2R(T) - \sum_{t=1}^{T}\sum_{k\neq k^\star}\Delta_k\mathbb{E}[q_{t,k}] \\ &\leq \sum_{k\neq k^\star}\sum_{t=1}^{T_0}\Big(\frac{\alpha\sqrt{\mathbb{E}[q_{t,k}]}}{\sqrt{t}} - \Delta_k\mathbb{E}[q_{t,k}]\Big) + 2C \\ &\quad + \sum_{k\neq k^\star}\sum_{t=T_0+1}^{T}\Big(\frac{\alpha\sqrt{\mathbb{E}[q_{t,k}]}}{\sqrt{t}} + \frac{\beta\mathbb{E}[q_{t,k}]}{2\sqrt{t}} - \Delta_k\mathbb{E}[q_{t,k}]\Big). \end{aligned} \tag{24}$$

By using the simple optimization that $\max_{x>0} 2\lambda\sqrt{x}-\omega x = \frac{\lambda^2}{\omega}$, we obtain

$$\max_{\mathbb{E}[q_{t,k}]\geq 0}\frac{\alpha\sqrt{\mathbb{E}[q_{t,k}]}}{\sqrt{t}} - \Delta_k\mathbb{E}[q_{t,k}] = \frac{\alpha^2}{4\Delta_k t},$$

and

$$\begin{aligned} &\max_{\mathbb{E}[q_{t,k}]\geq 0}\frac{\alpha\sqrt{\mathbb{E}[q_{t,k}]}+\frac{1}{2}\beta\mathbb{E}[q_{t,k}]}{\sqrt{t}} - \Delta_k\mathbb{E}[q_{t,k}] \\ &= \frac{\alpha^2}{4t(\Delta_k-\frac{\beta}{4\sqrt{t}})} = \frac{\alpha^2}{4\Delta_k t} + \frac{\alpha^2\beta}{4t(4\Delta_k^2 t^{\frac{1}{2}}-\beta\Delta_k)}. \end{aligned} \tag{25}$$

Plugging the above results into (24), we obtain

$$\begin{aligned} R(T) &\leq \sum_{k\neq k^\star}\Big(\sum_{t=1}^{T}\frac{\alpha^2}{4\Delta_k t} + \sum_{t=T_0+1}^{T}\frac{\alpha^2\beta}{16\Delta_k^2 t^{\frac{3}{2}}-4\beta\Delta_k t}\Big) + 2C \\ &\leq \sum_{k\neq k^\star}\frac{\alpha^2\log(T)}{4\Delta_k} + \sum_{k\neq k^\star}\frac{\alpha^2\beta}{\Delta_k(8\Delta_k\sqrt{T_0}-2\beta)} + 2C \\ &\leq \sum_{k\neq k^\star}\frac{\alpha^2\log(T)}{4\Delta_k} + \sum_{k\neq k^\star}\frac{\alpha^2\beta}{\Delta_k(\frac{4\Delta_k}{\Delta_{\min}}-2\beta)} + 2C \\ &\leq \sum_{k\neq k^\star}\frac{\alpha^2\log(T)}{4\Delta_k} + \sum_{k\neq k^\star}\frac{\alpha^2\beta}{\Delta_k(4-2\beta)} + 2C, \end{aligned} \tag{26}$$

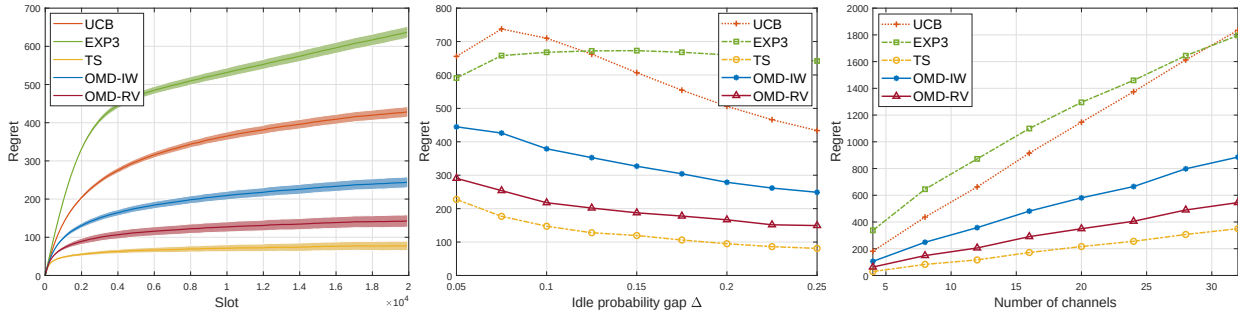
where the second inequality is derived using inequality $\sum_{T_0+1}^{T}\frac{1}{bt^{\frac{3}{2}}-ct} \leq \frac{2}{b\sqrt{T_0}-c}$ and the third inequation is achieved by the definition of T_0, which satisfies $\frac{1}{2\Delta_{\min}} \leq \sqrt{T_0} \leq \frac{1}{2\Delta_{\min}}+1$.

VI. Numerical Results

We verify our theoretical analysis through a series of numerical experiments in the three environmental regimes. In the stochastic regime, we set the idle probability to $(1+\Delta)/2$ for the single best channel and $(1-\Delta)/2$ for all suboptimal channels. In the periodically stochastic regime, we designate the time horizon as $T = 15810$, partitioned into five segments $[1000, 1600, 2560, 4096, 6554]$ with the segment lengths increasing exponentially by a factor of 1.6 [20, 36]. During odd segments, the idle probability of the best channel is set to 1, and that of the suboptimal channels is set to $1-\Delta$. Conversely, in even segments, the idle probabilities are set to Δ and 0, respectively. In the last regime, the idle probability of each channel aligns with that in the periodically stochastic regime. The attacker launches an attack every τ slots, referred to as the attack interval, and at the slot chooses ξ channels to attack, referred to as the attack strength. For the multi-SU scenario, we select M channels out of pool $\mathcal{K}$ as the best channels, maintaining identical idle probabilities as those in the single-SU case.

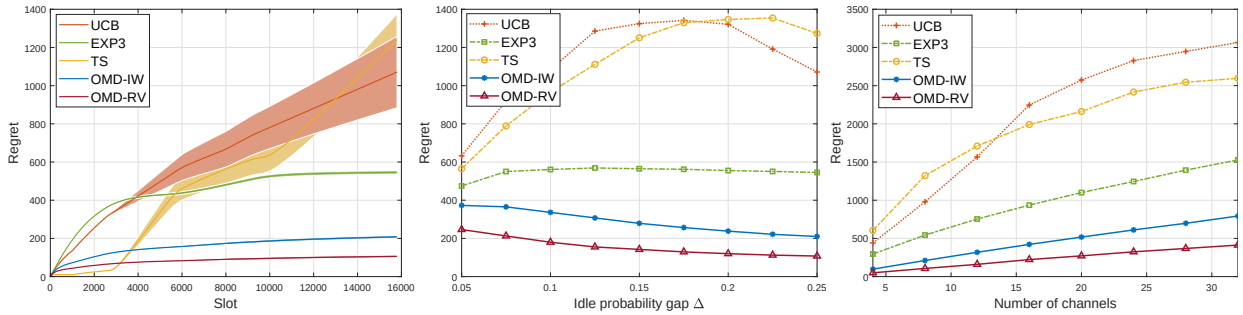
We present the numerical results of our proposed OMD algorithm with the RV estimator (OMD-RV) and the OMD algorithm with the IW estimator (OMD-IW). In the single-SU scenario, we perform a comparative analysis of the proposed algorithm against three foundational MAB algorithms, namely UCB [26] and TS [41] for the stochastic MAB, and the reduced variance variant of EXP3 [35] for the adversarial MAB. In the multi-SU scenario, our proposed algorithm is compared with the latest combinatorial versions of the UCB [32], TS [33], and EXP3 [20] algorithms, referred to as CUCB, CTS and CEXP3, respectively. In both scenarios, the learning rate parameter of the EXP3 algorithm is configured using a known time horizon T. Each experimental iteration involves 1000 Monte Carlo simulations, and the results are then averaged.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3434960



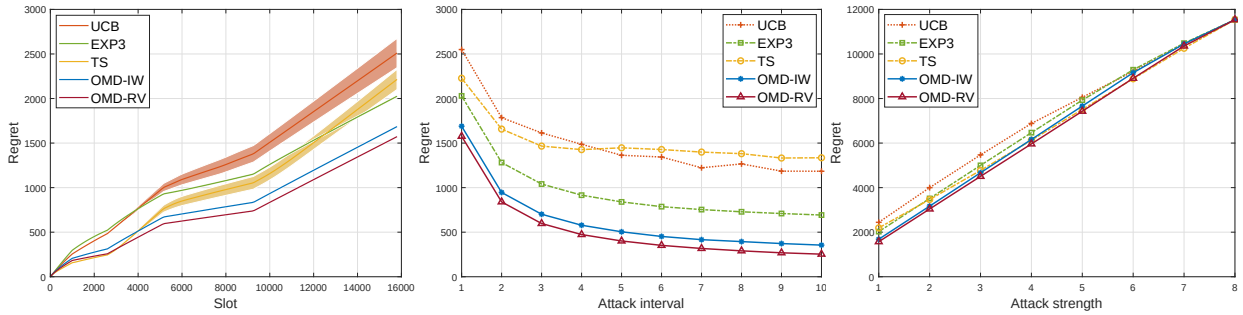
(a) Time horizon (b) Idle probability gap (c) Channel number

Fig. 2. Regrets in the stochastic regime for the single-SU case with three different system parameters: time horizon T, idle probability gap Δ and number of channels K.



(a) Time horizon (b) Idle probability gap (c) Channel number

Fig. 3. Regrets in the periodically stochastic regime for the single-SU case with three different system parameters: time horizon T, idle probability gap Δ and number of channels K.



(a) Time horizon (b) Attack interval (c) Attack strength

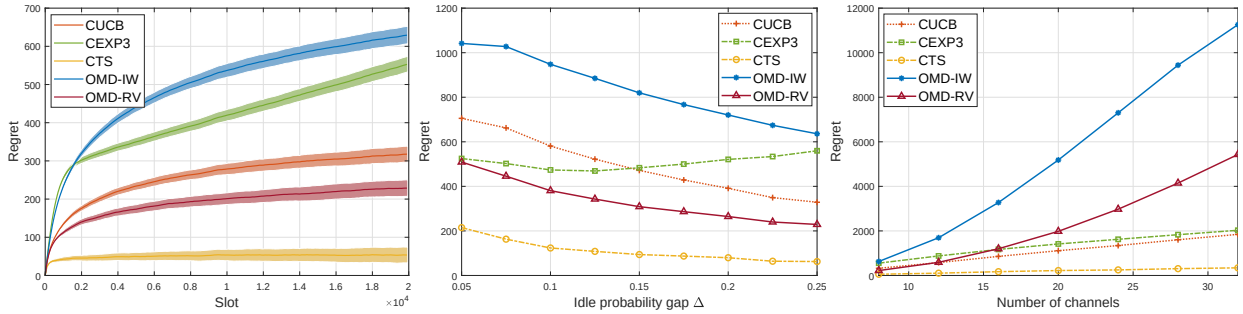
Fig. 4. Regrets in the stochastic regime with adversarial corruptions for the single-SU case with three different system parameters: time horizon T, attack interval τ and attack strength ξ.

A. Single-SU Case

Fig. 2 depicts the comparative regrets within the stochastic regime. In Fig. 2a, regrets along with their 99% confidence intervals are plotted as a function of time slot for $K = 8$, $\Delta = 0.25$ and $T = 20000$. The regrets of all algorithms exhibit sublinear growth over time. While TS showcases the lowest regret, the proposed OMD algorithm, coupled with the RV estimator, closely trails TS and notably outperforms all other contenders by a substantial margin.

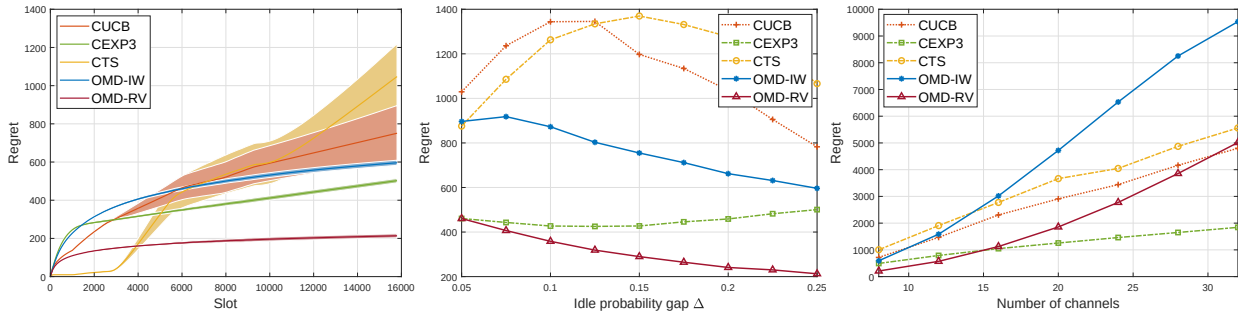
Fig. 2b shows the regrets of different algorithms as a function of idle probability gap Δ for $K = 8$ and $T = 20000$. As Δ increases, distinguishing the best channel from the channel pool becomes easier, resulting in diminishing regrets across all algorithms. However, larger Δ also implies higher regret when exploring suboptimal channels. For the UCB and EXP3 algorithms, the increase in regret due to the second effect initially outweighs the decrease from the first effect. Therefore, the regrets of them initially increase with Δ and then decrease.

Fig. 2c shows the regrets of different algorithms as a function of channel number K for $\Delta = 0.25$ and $T = 20000$. The increase of channels raises the exploration cost for each algorithm, as more time slots are required to learn the statistics



(a) Time horizon (b) Idle probability gap (c) Channel number

Fig. 5. Regrets in the stochastic regime for the multi-SU case with three different system parameters: time horizon T, idle probability gap Δ and number of channels K.



(a) Time horizon (b) Idle probability gap (c) Channel number

Fig. 6. Regrets in the periodically stochastic regime for the multi-SU case with three different system parameters: time horizon T, idle probability gap Δ and number of channels K.

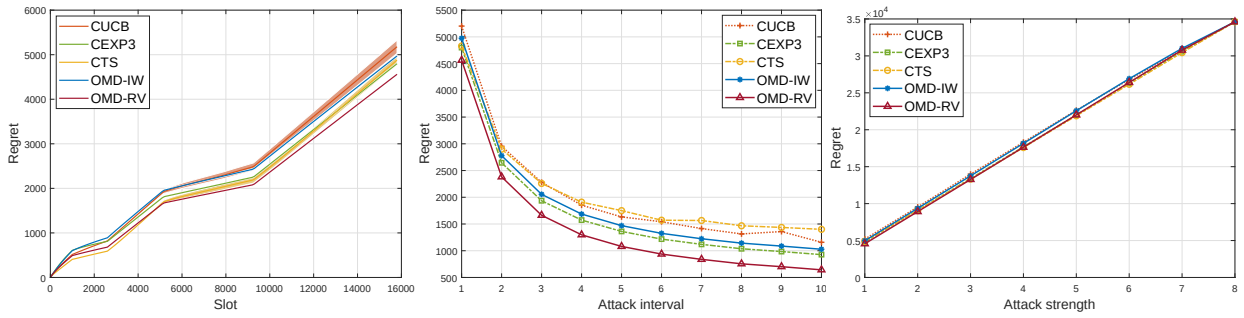
of the additional suboptimal channels. Thus, the regrets of all algorithms amplify with K. Notably, while the performance gap between the proposed OMD-RV algorithm and TS grows gradually with K, the gap between OMD-RV and its competitors experiences a pronounced increase as K expands.

Fig. 3 illustrates algorithmic regrets in the periodically stochastic regime. Fig. 3a shows the regrets along with their 99% confidence intervals as a function of slot for $K = 8$ and $\Delta = 0.25$. The UCB and the TS exhibit nearly linear regrets due to their lack of theoretical performance guarantees when faced with changing idle probabilities. Our proposed OMD-RV algorithm effectively responds to the stochastic nature of licensed channels while capitalizing on swift adaptations in channel selection under non-stochastic channel statistics, resulting in the lowest regret.

Fig. 3b shows the regrets as a function of idle probability gap Δ for $K = 8$. Similar to the stochastic regime, the proposed OMD-RV algorithm exhibits declining regret with increased Δ. The UCB experiences initial regret increase due to suboptimal channel selection outweighing best channel distinction, later improving as Δ increases. Fig. 3c shows the regrets as a function of channel number K for $\Delta = 0.25$. The regrets escalate across algorithms with rising K due to additional slots required to account for more suboptimal channels. The proposed OMD-RV algorithm consistently surpasses the others, with the performance gap growing notably as K increases.

Fig. 4 depicts algorithmic regrets within the stochastic regime with adversarial corruptions. To the best of our knowledge, a fully adversarial regime is very hard to simulate. We assume an oblivious attacker who randomly chooses licensed channels to attack. To better showcase the influence of PU emulation attacks, the reward of each algorithm is compared with the ideal action with the knowledge of attacks, i.e., $G_{\max}(T)$ is the same as that in the periodically stochastic regime. Fig. 4a shows the regrets along with their 99% confidence intervals as a function of time slot for $K = 8$ and $\Delta = 0.25$. Due to the emulation attacks, the regrets of all algorithms are higher than that in the periodically stochastic regime. The proposed algorithm enjoys strong theoretical guarantee like EXP3 in the adversarial regime while leveraging the stochastic nature of the problem, which shows the lowest regret.

Fig. 4b shows the regrets as a function of attack interval τ for $\xi = 1$. A higher τ indicates less frequent attacker launches, pushing the environment closer to the periodically stochastic regime. Consequently, the regrets of all algorithms decreases with τ. Fig. 4c shows the regrets as a function of attack strength ξ for $\tau = 1$. The number of idle sensing results decreases with the number of attacked channels. Thus, the regrets of all algorithms grow with ξ. When ξ approaches K, which means no idle sensing results can be achieved by the SUs, there is little difference between algorithms. In various attack scenarios, the proposed OMD-RV algorithm consistently achieves the lowest regret.



(a) Time horizon (b) Attack interval (c) Attack strength

Fig. 7. Regrets in the stochastic regime with adversarial corruptions for the multi-SU case with three different system parameters: time horizon T, attack interval τ and attack strength ξ.

B. Multi-SU Case

In Fig. 5a, regrets along with their 99% confidence intervals in the multi-SU scenario are presented as a function of slot within the stochastic regime. The regrets of all algorithms increase with time and the proposed OMD-RV algorithm achieves the lowest regret except for the CTS algorithm. Due to the simplicity of identifying the best channel brought by large Δ, the regrets of almost all algorithms decrease with Δ in Fig. 5b. As channel number K increases, the action space of the proposed algorithm grows in factorial form and the regret of it exceeds the other algorithms when K exceeds 16, as shown in Fig. 5c.

Fig. 6 illustrates algorithmic regrets within the periodically stochastic regime. The CTS algorithm cannot adapt to the change in channel idle probabilities and the regret of it increases significantly with time in Fig. 6a. Fig. 6b shows the regrets as a function of idle probability gap Δ. The trends are similar to those in the single-SU case. Our proposed algorithm can keep tracking the best channel even with small Δ. While the CUCB and CTS algorithms suffer from the cost of selecting suboptimal channels and the regrets of them grow with Δ initially. Fig. 6c shows the regrets as a function of channel number K. The OMD-RV and CEXP3 algorithms maintain their order optimum in the adversarial regime and outperform the other competitors.

Fig. 7 depicts algorithmic regrets within the stochastic regime with adversarial corruptions. Similar to that in the single-SU case, the regrets of all algorithms show almost linear growth with time in Fig. 7a. In Fig. 7b, as the attack interval increases, the proposed algorithm deals with the malicious attacks better than the latest CEXP3 algorithm, achieving a lower regret. As the number of attacked channels increases in Fig. 7c, the SUs are cheated by the attackers and get more incorrect sensing results. Due to the worst case regret guarantee, the proposed algorithm still brings more channel access opportunities for the SUs.

VII. CONCLUSION

In this paper, we investigated the centralized DSA problem in an unknown environment. We formulate it as a combinatorial MAB problem and propose an online mirror descent algorithm for channel selection, which utilizes Tsallis entropy as the regularization function by generalizing the EXP3 algorithm for adversarial bandits. Our proposed algorithm maintains a $\sqrt{T}$ regret rate guarantee in the adversarial regime while simultaneously achieving a $\log(T)$ regret growth in the stochastic regime. Numerical results validate our theoretical analysis, moreover, the proposed algorithm outperforms others in almost all regimes with different network parameters such as idle probability gap, channel number, attack interval and attack strength.

REFERENCES

[1] Y. Chen *et al.*, "Age of information for short-packet relay communications in cognitive-radio-based Internet of Things with outdated channel state information," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 3, pp. 722–737, Jun. 2023.
[2] A. Ahmad *et al.*, "A survey on radio resource allocation in cognitive radio sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 888–917, 2nd Quart. 2015.
[3] F. Li *et al.*, "Advances and emerging challenges in cognitive Internet-of-Things," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5489–5496, Aug. 2020.
[4] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.
[5] M. Zhou, T. Wang, and S. Wang, "Spectrum sensing across multiple service providers: A discounted Thompson sampling method," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2402–2406, Dec. 2019.
[6] X. Zhu *et al.*, "Dynamic channel selection and transmission scheduling for cognitive radio networks," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 24 429–24 443, Dec. 2022.
[7] S. Ye, T. Wang, and S. Wang, "Dynamic spectrum access in non-stationary environments: A Thompson sampling based method," in *Proc. IEEE GLOBECOM'22*, Rio de Janeiro, Brazil, Dec. 2022.
[8] S. Ye and S. Wang, "An improved Thompson sampling method for dynamic spectrum access in non-stationary environments," in *Proc. IEEE APSCON'23*, Bengaluru, India, Jan. 2023.
[9] C. Xin and M. Song, "Analysis of the on-demand spectrum access architecture for CBRS cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 970–978, Feb. 2020.
[10] M. Hanawal and S. Darak, "Multiplayer bandits: A trekking approach," *IEEE Trans. Autom. Control*, vol. 67, no. 5, pp. 2237–2252, May 2022.
[11] A. Anandkumar *et al.*, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Apr. 2011.
[12] L. Besson and E. Kaufmann, "Multi-player bandits revisited," in *Proc. Algorithmic Learn. Theory*, Lanzarote, Spain, Apr. 2018.
[13] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.
[14] J. Oksanen and V. Koivunen, "An order optimal policy for exploiting idle spectrum in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1214–1227, Mar. 2015.

[15] A. Magesh and V. Veeravalli, "Decentralized heterogeneous multi-player multi-armed bandits with non-zero rewards on collisions," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2622–2634, Apr. 2022.

[16] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, Mar. 2013.

[17] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6184–6193, Dec. 2014.

[18] N. Modi, P. Mary, and C. Moy, "Qos driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 1, pp. 49–66, Mar. 2017.

[19] S. Ye, T. Wang, and S. Wang, "Thompson sampling based dynamic spectrum access in non-stationary environments," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 3, pp. 593–603, Jun. 2023.

[20] A. Alipour-Fanid *et al.*, "Multiuser scheduling in centralized cognitive radio networks: A multi-armed bandit approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1074–1091, Jun. 2022.

[21] P. Zhou and T. Jiang, "Toward optimal adaptive wireless communications in unknown environments," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3655–3667, May 2016.

[22] W. Wang *et al.*, "Decentralized learning for channel allocation in IoT networks over unlicensed bandwidth as a contextual multi-player multi-armed bandit game," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3162–3178, May 2022.

[23] S. Feng and S. Haykin, "Coordinated cognitive risk control for bridging vehicular radar and communication systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4135–4150, May 2022.

[24] S. Han *et al.*, "Channel-correlation-enabled transmission optimization for MISO wiretap channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 858–870, Feb. 2021.

[25] X. Sheng and S. Wang, "Online primary user emulation attacks in cognitive radio networks using Thompson sampling," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8264–8273, Dec. 2021.

[26] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, May 2002.

[27] A. Javanmardi, M. Qureshi, and C. Tekin, "Decentralized dynamic rate and channel selection over a shared spectrum," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3787–3801, Jun. 2021.

[28] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math*, vol. 6, no. 1, pp. 4–22, 1985.

[29] J. Zimmert and T. Lattimore, "Connections between mirror descent, Thompson sampling, and the information ratio," in *Proc. Adv. Neural Inf. Process. Sys.*, Vancouver, Canada, Dec. 2019.

[30] J. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proc. Int. Conf. Comput. Learn. Theory*, Montreal, Canada, Jun. 2009.

[31] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends® Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[32] B. Kveton *et al.*, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proc. Int. Conf. Artif. Intell. Statist.*, San Diego, CA, USA, May 2015.

[33] S. Wang and W. Chen, "Thompson sampling for combinatorial semi-bandits," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018.

[34] P. Auer *et al.*, "The non-stochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[35] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[36] J. Zimmert and Y. Seldin, "Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits," *J. Mach. Learn. Res.*, vol. 22, no. 28, pp. 1–49, 2021.

[37] P. Auer and C. Chiang, "An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits," in *Proc. Int. Conf. Comput. Learn. Theory*, New York, USA, Jun. 2016.

[38] C. Wei and H. Luo, "More adaptive algorithms for adversarial bandits," in *Proc. Int. Conf. Comput. Learn. Theory*, Stockholm, Sweden, Jul. 2018.

[39] J. Abernethy, C. Lee, and A. Tewari, "Fighting bandits with a new kind of smoothness," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, Canada, Dec. 2015.

[40] A. Agarwal *et al.*, "Corralling a band of bandit algorithms," in *Proc. Int. Conf. Comput. Learn. Theory*, Amsterdam, Netherlands, Jul. 2017.

[41] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. Annu. Conf. Learn. Theory*, Edinburgh, Scotland, Jun. 2012.

Shuai Ye received the BS degree from Nanjing University, Nanjing, China, in 2020, where he is currently pursuing the PhD degree at the School of Electronic Science and Engineering. His current research interests include dynamic spectrum access and online learning.

Shaowei Wang received the Ph.D. degree from Wuhan University, Wuhan, China, in 2006. He joined the School of Electronic Science and Engineering, Nanjing University, Nanjing, China, as a faculty member, in 2006, where he is currently a full professor. From 2012 to 2013, he was a Visiting Scholar/a Professor with Stanford University, Stanford, CA, USA, and The University of British Columbia, Vancouver, BC, Canada. His research interests include communications and networking, operations research, and machine learning.