Channel Selection for Cognitive Radio Terminals

Ling-Hung Kung; SUID: 04906103

1 Introduction

Due to the excessive need of wireless spectrum and the inefficiency in utilizing it, the technology of cognitive radio (CR) addresses the issue of allowing unlicensed users to make use of the frequency bands where licensed users is currently not active. From the hierarchical structure, CR users can only grab resources under the premise of not interfering with the normal operation of the primary system (PS), and this extra constraint complicates the original time-varying wireless communication. By assuming Bernoulli distribution for each channel and independence across channels, this dynamic spectrum access scheme can be treated as a multi-armed bandit (MAB) problem for a single CR user, where each channel is considered as a slot machine with some expected reward, and this user is trying to get as much available bandwidth as possible. The key component of MAB problem is the tradeoff between exploitation and exploration, where the CR terminal tries to pick the channel that has highest estimated reward from past history, and look for new channels that might give even higher rewards at the same time.

There are different versions of MAB formulation. In the case of stationary distribution, Gittins index is shown to be the optimal strategy for discounted MAB in [6], and [8] apply it to CR. By allowing channel distributions to change over time, Whittle's index is proved to be asymptotically optimal under some constraints in [12], and it is shown in [9] that opportunistic spectrum access is indexable and hence able to apply this strategy. However, the above approaches both assume infinite horizon and maximize discounted reward, whereas in the wireless environment, we only care about the reward obtained in a finite observation period, which leads us to the finite-time MAB introduced in [3] and others. So far there is no optimal strategy to our knowledge, and we would refer to different finite-time algorithms with tuned parameters. In this paper we basically follow the algorithms in [1] and [11], and proceeds as follows: In section 2 we describe the network model in detail, and in section 3 we examine some common finite-time MAB algorithms. Numerical simulations are provided in section 4 to compare algorithms in different probability distributions, and followed by the conclusion as well as possible extensions in section 5.

2 Network Model

Consider a set of channels $\mathcal{M} = \{1, \cdots, M\}$ in a PS, and a CR terminal tries to use these channels when they are free, or not occupied by the PS. The channels are temporally divided into discrete time slots, and the CR terminal synchronizes to the PS such that the beginning and end of each time slot is known. The probability that channel i is free is $p_i, i \in \mathcal{M}$. In general we model the channels using a stochastic process, but here we assume that p_i is stationary to simplify the problem. The terminal operates as follows: for each time slot t, the terminal chooses some channel $i^{(t)}$, senses to determine whether it is free (with probability $p_{i^{(t)}}$), and conducts its own transmission if it is; if the channel turns out to be occupied, then the terminal needs to wait till the next time slot, and choose some channel (maybe the same one) again. Normally the terminal has no prior information about $\mathbf{p} = \{p_1, \cdots, p_M\}$, and will learn some empirical distribution in the process of transmission. Let the reward of choosing channel i at time t be $x_i^{(t)}$, then the goal of CR terminal is to maximize the accumulated reward up to observation period T, i.e. $\sum_{i=1}^{M} \sum_{t=1}^{T} x_i^{(t)}$, or to minimize the regret of adopting this strategy, calculated by $Tp^* - \sum_i \sum_t x_i^{(t)}$, where $p^* = \max_{i \in \mathcal{M}} p_i$ is the optimal expected reward per time slot. From now on we simply assign 1 to $x_i^{(t)}$ if channel i is selected at time t and not occupied, and 0 otherwise.

3 Learning Algorithms

3.1 Static environment

Most algorithms for finite-time MAB assumes stationary probability, as in [3], [11], and the references therein. Here we introduce some basic prototypes to compare their performance under our network model.

3.1.1 Upper confidence bound

This algorithm is derived from the index-based policy developed in [3], where the index is the sum of two terms: one is the current average reward, and the second term corresponds to the confidence interval that both the true and average rewards fall in with high probability. The upper confidence bound (UCB) algorithm first initializes by selecting each channel once. After that, for each time t, UCB chooses channel $i^{(t)}$ such that

$$i^{(t)} = \arg\max_{i \in \mathcal{M}} \left(\bar{x}_i^{(t)} + \sqrt{\frac{\xi \log t}{n_i^{(t)}}} \right)$$

where $n_i^{(t)}$ is the number of times channel i has been chosen so far, $\bar{x}_i^{(t)} = \sum_{\tau=1}^{t} x_i^{(\tau)} / n_i^{(t)}$ is the current average reward, and ξ is some parameter chosen to be 2 in [3]. By letting $\xi = 0.5$, not only it performs better in our simulation, but we also effectively reduce the upper bound of expected regret from a factor of 4. An improved algorithm, UCB-V, that considers the effect of the empirical variance, is proposed in [2] and chooses channel $i^{(t)}$ such that

$$i^{(t)} = \arg\max_{i \in \mathcal{M}} \left(\bar{x}_i^{(t)} + \sqrt{\frac{\left(\bar{x}_i^{(t)} - (\bar{x}_i^{(t)})^2\right) \xi \log t}{n_i^{(t)}}} + \frac{c \log t}{n_i^{(t)}} \right)$$

where we are free to adjust ξ and c.

3.1.2 ϵ-Greedy and its variants

The ϵ-greedy strategy consists of choosing a random channel with probability ϵ, and select the channel with highest current average reward otherwise. Here the choice of $\epsilon \in (0, 1)$ is not specified. However, this simple form of ϵ-greedy strategy is sub-optimal for stationary probability distribution because the constant ϵ will prevent the terminal from choosing the optimal channel asymptotically. A natural variant, GreedyT, is to decrease ϵ gradually by choosing $\epsilon_t = \min\{1, \frac{\epsilon_0}{t}\}$. We can also use the decreasing factor $\log(t)/t$ instead of $1/t$ to get another strategy GreedyLogT. Some discussion on the regret bounds of the greedy-family algorithms are given in [3] and [4].

3.1.3 SoftMax and its variants

Recall that $\bar{x}_i^{(t)}$ is the current average reward of channel i at time t, then the SoftMax strategy chooses channel i at time $t+1$ with probability $\exp(\bar{x}_i^{(t)}/\tau)/Z^{(t)}$, where $Z^{(t)}$ is the normalization factor. $\tau \in \mathbb{R}^+$ is called the temperature and is free to user's choice. Similar to the case in ϵ-greedy, we can gradually increase the probability that the channel with highest average reward being chosen by setting $\tau_t = \tau_0/t$ or $\tau_t = \tau_0 \log(t)/t$, which we call them SoftMaxT and SoftMaxLogT.

3.2 Stochastically changing environment

So far the algorithms above all use average reward as an index to compute which channel to choose. However, in the time-varying wireless channel, it is not reasonable to assign equal weights to all observations no matter when we acquire them. One intuition is to forget old data and introduce "backward-discounted" reward by calculating the weighted average reward

$$\hat{x}_i^{(t)} = \frac{1}{\hat{n}_i^{(t)}} \sum_{\tau=1}^{t} \gamma_i^{t-\tau} x_i^{(\tau)}, \quad \hat{n}_i^{(t)} = \sum_{\tau=1}^{t} \gamma_i^{t-\tau} \mathbf{1}\{i^{(t)} = i\}$$

where $0 < \gamma_i < 1$ is the discount factor for channel i that depends on how fast channel i changes, and the weighting function decreases as $t - \tau$ increases, as in [7]. Now we can replace the average reward used in UCB, ϵ-greedy, and SoftMax with this weighted average reward. Notice that this new reward may not be applied directly to the variants of ϵ-greedy and SoftMax since exploration is comparatively important in the dynamic environment. Another possibility is to use "sliding window" with width depending on how fast channel changes, which is proposed in [5] along with some related regret bounds.

4 Numerical Simulation

In our simulation, we assume that one channel is either occupied by the PS (hence has a low free probability) or not, and we test the stationary algorithms against the following three distributions:

	CH1	CH2	CH3	CH4	CH5	CH6	CH7	CH8	CH9	CH10
Distribution 1	.9	.8	.8	.7	.7	.3	.3	.2	.2	.1
Distribution 2	.9	.3	.3	.3	.2	.2	.2	.1	.1	.1
Distribution 3	.9	.8	.8	.8	.8	.8	.8	.8	.8	.8

The parameters adopted for our algorithms are UCB with $\xi = 0.5$, UCB-V with $\xi = 0.2$ and $c = 0.3$, Greedy with $\epsilon = 0.1$, GreedyT with $\epsilon_0 = 25$, GreedyLogT with $\epsilon_0 = 4$, SoftMax with $\tau = 0.05$, SoftMaxT with $\tau_0 = 8$, and SoftMaxLogT with $\tau_0 = 2.5$. After 10000 iterations, the results of average regret, variance of regret, and the percentage of time choosing the optimal channel (CH1) of different algorithms are shown in Figure 1. These comparisons show that for ϵ-Greedy and SoftMax, gradually decreasing the percentage of exploration helps the algorithm to converge to choosing the optimal channel. Notice that in distribution 2, though the SoftMax family have small average regret and high percentage of optimal choice, they exhibit extreme large variance in regret, which is not a sign for good algorithm. Besides, the regret bounds derived for these algorithms may be too loose for smaller T. For instance, the modified bound for UCB in distribution 3 gives 1243.4, which is not only much larger than our empirical result, but also larger than the total reward. Based on the three indices that we tested, UCB-V gives the best performance, but this superiority may depend on the parameters that we choose. For the ϵ-Greedy family, if we choose

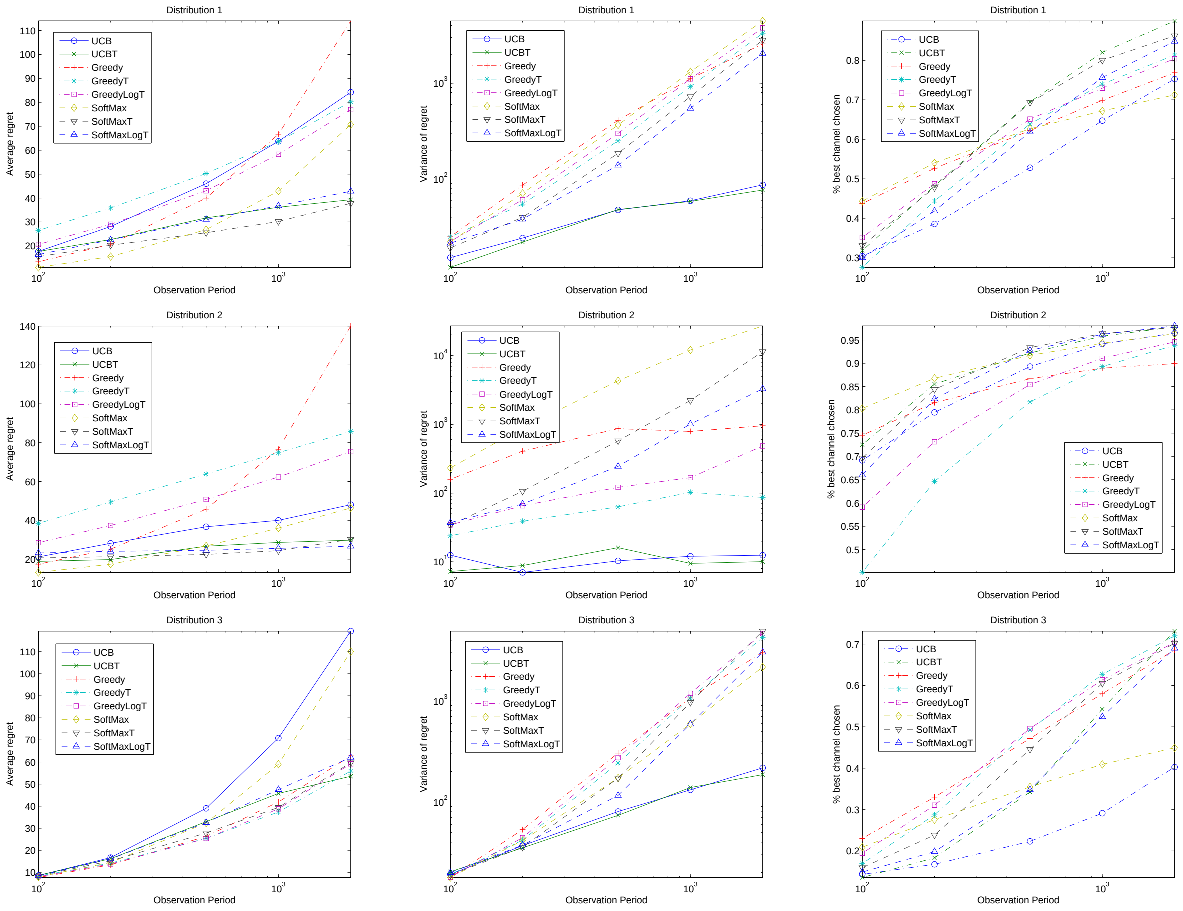

Figure 1: Regrets and percentage of optimal action for algorithms under different distributions

larger ϵ, then in general we have larger mean and smaller variance, whereas the same thing holds for larger τ in SoftMax family. We demonstrate the relative variations of performance indices by choosing different parameters, which is shown in Figure 2. For GreedyT, the optimal ϵ_0's are 2.5, 100, and 25 for these indices individually, and the actual choice of ϵ_0, if we decide to use GreedyT, depend on how the system evaluates these indices.

5 Conclusion and Future work

In this paper we transform channel selection in CR into an equivalent MAB problem, examine several approaches and algorithms that deal with it, and run simulations to compare their performance under stationary environment. There are several topics that we can keep working on. Besides the non-stationarity of channels mentioned in Section 3.2, we can study different channel models, such as the Gilbert-Elliot model used in [10], which treats one channel as a Markov chain with two states, busy and idle, yet preserve the independence across channels. More generally, channels can be modeled as a partially observable Markov decision process (POMDP) by introducing the correlation across channels, which is examined in [13]. One other dimension is to introduce imperfect sensing to make the scenario more realistic, as discussed in [10]. Finally, we can extend our topic into multi-agent system, where all terminals perform distributed learning without changing any information explicitly, allowing the CR network to be effectively established in a simple manner.

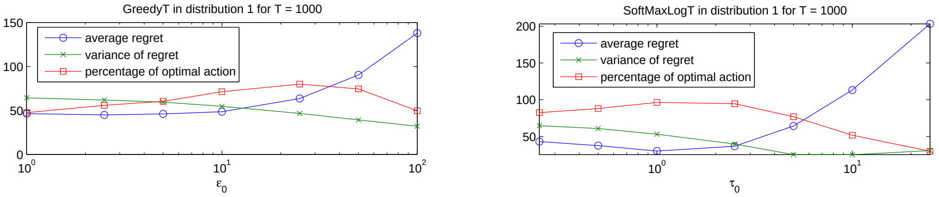

Figure 2: Relative variation of performance indices versus parameter choice

References

[1] A. Alaya-Feki, E. Moulines, and A. LeCornec, "Dynamic spectrum access with non-stationary multi-armed bandit,", in *IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, pp. 416-420, 2008.

[2] J-Y. Audibert, R. Munos, and A. Szepesvari, "Tuning bandit algorithms in stochastic environments," in *Algorithmic Learning Theory*, pp. 150-165, 2007.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," in *Machine Learning*, vol. 47, pp. 235-256, 2002.

[4] N. Cesa-Bianchi and P. Fischer, "Finite-time regret bounds for the multiarmed bandit problem," in *Proceedings of the 15th International Conference on Machine Learning*, pp. 100-108, 1998.

[5] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," 2008. Availble from http://arxiv.org/PS_cache/arxiv/pdf/0805/0805.3415v1.pdf

[6] J. Gittins and D. Jones, "A dynamic allocation indices for the sequential design of experiments," in *Progress in Statistics, European Meeting of Statisticians*, vol. 1, pp. 241-266, 1974.

[7] L. Kocsis and C. Szepesvari, "Discounted UCB," in *2nd PASCAL Challenges Workshop*, 2006.

[8] L. Lai, H. El Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: exploration, exploitation and competition," in *IEEE/ACM Trans. on Networking*, Oct. 2007. Submitted.

[9] K. Liu and Q. Zhao, "A restless bandit formulation of opportunistic access: indexablity and index policy," in *5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*, pp. 1-5, 2008.

[10] O. Mehanna, A. Sultan, and H. El Gamal, "Blind cognitive mac protocols," 2008. Available from http://arxiv.org/PS_cache/arxiv/pdf/0810/0810.1430v1.pdf

[11] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Proceedings of the 16th European Conference on Machine Learning*, pp. 437 - 448, 2005.

[12] P. Whittle,"Restless bandits: Activity allocation in a changing world", in Journal of Applied Probability, Vol. 25, 1988.

[13] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework," in *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589-600, 2007.