Improving throughput using multi-armed bandit algorithm for wireless LANs

Kaori Kuroda$^{1a)}$, Hiroki Kato$^1$, Song-Ju Kim$^2$, Makoto Naruse$^3$, and Mikio Hasegawa$^1$

$^1$ Department of Electrical Engineering, Tokyo University of Science, 6-3-1 Niijuku, Katsushika-ku, Tokyo 125-8585, Japan
$^2$ Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan
$^3$ Strategic Planning Department, National Institute of Information and Communications Technology, 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan

$^{a)}$ kuroda@ee.kagu.tus.ac.jp

Received April 25, 2017; Revised August 17, 2017; Published January 1, 2018

Abstract: Recently, various mobile communication systems have been widely deployed, and mobile traffic is increasing. However, the bandwidth available for mobile communications is limited, hence the scarcity of radio resources in mobile communications is a serious problem. As an approach to solve this problem, cognitive wireless communication models have been proposed. These models search for vacant time slots in multi-channel wireless communication systems. Although previous studies have shown that frequency utilization efficiency can be improved by multi-armed bandit algorithms, channels are assumed to be independent. However, channels used in 2.4 GHz wireless LANs (such as IEEE802.11b or IEEE802.11g) are not independent because these channels overlap with adjacent channels. In this paper, we propose an extended multi-armed bandit algorithm that uses continuous-valued rewards, which is applicable to wireless communication systems with overlapping channels. We show the effectiveness of the proposed method by experimental demonstrations.

Key Words: multi-armed bandit algorithm, liquid tug-of-war model, cognitive radio model, wireless LAN

1. Introduction

Recently, various mobile communication devices have been widely deployed and the amount of the mobile traffic is significantly increasing. However, frequency bands for mobile communications are limited, and spectrum band scarcity is becoming a very serious problem. As an approach to solve this problem, cognitive radio systems have been studied [1, 2]. These systems improve the efficiency of radio resource usage by means of opportunistic spectrum access.

In a spectrum sharing-type cognitive radio system, cognitive users utilize vacant frequency bands
to improve radio resource usage. Vacant radio resources dynamically change as a result of changes in other users' usage of the resources. In order to detect vacant radio resources efficiently, an effective channel sensing scheme is necessary. In Refs. [3, 4], Lai et al. modeled a cognitive radio as a multi-armed bandit problem. The multi-armed bandit problem maximizes the total rewards provided from slot machines by optimizing the selection of slot machines that probabilistically provide rewards. In Refs. [3, 4], each channel is assumed to be probabilistically vacant in time division. Such a cognitive radio system can be defined as a multi-armed bandit problem. In previous studies [3, 4], channels are assumed to be independent. However, these channels are not independent because they overlap with adjacent channels (such as IEEE802.11b or IEEE802.11g), which results in interference. Therefore, it is preferable that rewards in a cognitive radio system take the form of continuous values, rather than binary values. From this point of view, we proposed a multi-armed bandit algorithm with continuous-valued rewards taking account of interference effects in wireless communications. To evaluate the proposed method, we implemented a wireless LAN access point based on a model described in Refs. [3, 4], concerning the fact that the wireless LAN is based on the CSMA/CA MAC protocol. We evaluated the effectiveness of our proposed method by experimental demonstrations.

2. Cognitive radio model as a multi-armed bandit problem

The multi-armed bandit problem [8] is a simple machine learning problem, based on a situation faced by a player attempting to earn the maximum reward from multiple slot machines. The purpose of the multi-armed bandit problem is to detect, through finite trials, which slot machine should be selected in order to maximize the reward amount. It is assumed that the player has no prior information about the reward rates of any of the machines. In the beginning, the player gathers information on each slot machine. Specifically, the player tries as many slots as possible and estimates which slot machine may have the highest expected value of reward. Next, the player plays the slot machine with the highest expected value of reward. As a result, the player can get many rewards. If the estimation time is long, the player can correctly estimate the expected value of reward of each slot machine. However, the player cannot get many rewards because the exploitation time is short. On the other hand, if the estimation time is short, the player may select a slot machine with a low reward and continue to invest in the slot machine. In the multi-armed bandit problem, it is important to resolve this trade-off. In order to resolve this trade-off, several multi-armed bandit algorithms have been proposed [5–7].

Figure 1 shows the cognitive radio channel model proposed by Lai et al. [3, 4]. This model has \( N \) channels, and each channel is separated by a time slot. In this model, we consider a primary network consisting of \( N \) non-overlapping channels. The users in the primary network operate in a synchronous time-slotted fashion. We assume that channel \( i \) is vacant with probability \( P_i \) at each time slot. When the cognitive users communicate, a cognitive user senses the channel, and utilizes the vacant channel. In this model, a channel corresponds to the slot machine in the multi-armed bandit problem. If the time slot at a selected channel is vacant, the player gets a reward.

---

![Fig. 1. Cognitive radio channel model.](image-url)
3. Multi-armed bandit algorithm

To solve multi-armed bandit problems, previous studies have proposed the ε-greedy algorithm [5], softmax algorithm [6] and UCB1-tuned algorithm [7] as multi-armed bandit algorithms. Although the UCB1-tuned algorithm is known as the best algorithm among parameter-free algorithms, the liquid type tug-of-war (LTOW) model [9, 10], which has approximately the same performance as the UCB1-tuned algorithm, has been proposed recently. The LTOW model adapts to changes in the environment as the reward probability dynamically changes. Therefore, it is suitable for use with cognitive radio systems.

3.1 Liquid-type tug-of-war model

The liquid-type tug-of-war model is a multi-armed bandit algorithm inspired by the behavior of the amoeboid organism [9, 10]. Let us consider incompressible fluid in a branched cylinder as shown in Fig. 2. For each branch \( i \) at time \( t \), let \( X_i(t) \) correspond to the displacement of machine \( i \) from an initial position. When a machine \( i \) is played at time \( t \), +1 is added to \( X_i(t) \) if the player receives a reward. Otherwise, \(-\omega\) is added to \( X_i(t) \). The LTOW model learning rule is defined by

\[
Q_i(t) = N_i(t) - (1 + \omega)L_i(t),
\]

where \( N_i(t) \) is the number of machines being played at time \( t \), \( L_i(t) \) is the number of loss in machine \( i \) at time \( t \), and \( \omega \) is a parameter. The displacement of branch \( i \) from its initial position is defined by

\[
X_i(t) = Q_i(t) - \frac{1}{n-1} \sum_{k=1, k \neq i}^{n} Q_k(t) + \zeta(t),
\]

where \( n \) is the number of slot machines and \( \zeta(t) \) is a fluctuation. In this study, we used

\[
\zeta(t) = \alpha \cos\left(\frac{2\pi t}{n} + \frac{2(i-1)\pi}{n}\right),
\]

where \( \alpha \) is a parameter. The player selects the machine having the highest value of \( X_i(t) \).

![Fig. 2. An LTOW model with three channels.](image)

3.2 Extended liquid-type tug-of-war model for continuous-valued reward

In the conventional channel model of cognitive radio [3, 4], it is assumed that channels are independent. Then, the reward can be defined binary (0 or 1). However, channels in 2.4 GHz Wi-Fi (such as IEEE802.11b or IEEE802.11g) are not independent because they overlap with adjacent channels (Fig. 3). Then, adjacent channel interference occurs when channels with adjacent frequencies are used simultaneously. Therefore, we considered effects from adjacent channels and defined a continuous-valued reward for wireless LANs (denoted as \( R_i(t) \)) as follows:

\[
R_i(t) = (\beta + \sum_{k=-l}^{l} a_k n_{i+k}(t))^{-1},
\]

where \( \beta \) is a parameter that decides the maximum value of the reward, \( l \) is the number of adjacent channels considering the effect of interference, \( a_k \) is a weight parameter, and \( n_i(t) \) represents whether...
channel $i$ is vacant or not. If channel $i$ is vacant, $n_i(t) = 0$. Otherwise, $n_i(t) = 1$. Thus, the reward is reduced when adjacent channels are used. The maximum reward becomes $\frac{1}{A}$ when the selected channel and adjacent channels are vacant.

However, to check the vacancy of the adjacent channels, the same number of wireless LAN interfaces as adjacent channels are required in a terminal. Then, by estimating the vacancy of the adjacent channels, we defined the reward given when only one wireless LAN interface is used as

$$ R_i(t) = (\beta + a_0n_i(t) + \sum_{k=-l}^{l} a_k\hat{n}_{i+k}(t))^{-1}, $$

where $\hat{n}_i$ is an estimated vacancy probability in adjacent channels, which is defined by

$$ \hat{n}_r(t) = \frac{\sum_{s=1}^{N_r(t)} \sum_{i=1}^{N_r(t)} n_r(s)\delta(t-t^r_i)}{N_r(t)}, $$

where $N_r(t)$ is the number of times that channel $r$ is selected by time $t$, $t^r_i$ is the time at which channel $r$ is selected ($i = 1, 2, 3, \ldots, N_r(t)$), and $\delta(\cdot)$ is the Delta function ($\delta(x) = 1$ if $x = 0$, and $\delta(x) = 0$ otherwise). In this study, we set the parameter $l$ to 2, because we have determined that the interference of adjacent channels that are 2 channels away from the selected channel is large through experiments using a real wireless LAN system. The parameter $\beta$ is set to 1, in order to set the maximum reward to 1. The parameters $a_0$, $a_{\pm 1}$ and $a_{\pm 2}$ are set according to the following rules: $a_0 > \sum_{k=-l}^{l} a_k$, $a_{\pm 1} > a_{\pm (i+1)}$, and $a_i = a_{-i}$. When the selected channel is occupied and all the adjacent channels are vacant, the reward $R^A$ is $(\beta + \sum_{k=-l}^{l} a_k)^{-1}$. In contrast, when the selected channel is vacant and all the adjacent channels are occupied, the reward $R^B$ is $(\beta + a_0)^{-1}$. The reward when the selected channel is vacant should be larger than the reward when the selected channel is occupied ($R^A > R^B$). Therefore, we set the rule, $a_0 > \sum_{k=-l}^{l} a_k$. In addition, the effect of interference becomes small when the channel is away from the selected channel. Then, we set the rule, $a_{\pm i} > a_{\pm (i+1)}$. To satisfy these rules, the parameters are set to $a_0 = 2.0$, $a_{\pm 1} = 0.6$, and $a_{\pm 2} = 0.3$.

The original LTOW model can be used only for multi-armed bandit problems with binary rewards. To apply the model to multi-armed bandit problems with continuous-valued rewards, we reformulate the LTOW model based on Refs. [9, 10].

For the sake of simplicity, we assume that the number of slot machines is two. We consider that the sum of the average rewards $\gamma = \bar{R}_A(t) + \bar{R}_B(t)$ is given. When we estimate the probabilities based on the selection of machine $A$ a total of $N_A$ times, the average rewards at time $t$ are given as follows,

$$ \bar{R}_A(t) = \frac{\sum_{i=1}^{N_A(t)} R_A(i)}{N_A(t)}, \quad \bar{R}_B(t) = \gamma - \frac{\sum_{i=1}^{N_A(t)} R_A(i)}{N_A(t)}. $$

Similarly, when we estimate the probabilities based on selection of the machine $B$ a total of $N_B$ times, the reward probabilities at time $t$ are given as follows,

$$ \bar{R}_A(t) = \gamma - \frac{\sum_{i=1}^{N_B(t)} R_B(i)}{N_B(t)}, \quad \bar{R}_B(t) = \frac{\sum_{i=1}^{N_B(t)} R_B(i)}{N_B(t)}. $$
From the average rewards (Eqs. (7) and (8)), the expected values of rewards from machine A and machine B are given as follows,

\[
E_A(t) = \frac{\sum_{i=1}^{N_A(t)} R_A(i)}{N_A(t)} N_A(t) + (\gamma - \sum_{i=1}^{N_B(t)} R_B(i) N_B(t)) N_B(t)
\]

\[
= \sum_{i=1}^{N_A(t)} R_A(i) - \sum_{i=1}^{N_B(t)} R_B(i) + N_B(t) \gamma,
\]

(9)

\[
E_B(t) = (\gamma - \sum_{i=1}^{N_B(t)} R_A(i)) N_A(t) + \sum_{i=1}^{N_B(t)} R_B(i) N_B(t)
\]

\[
= \sum_{i=1}^{N_B(t)} R_B(i) - \sum_{i=1}^{N_A(t)} R_A(i) + N_A(t) \gamma.
\]

(10)

In the LTOW model, the difference between expected rewards is given by

\[
\frac{E_A(t) - E_B(t)}{2} = (\sum_{i=1}^{N_A(t)} R_A(i) - \sum_{i=1}^{N_B(t)} R_B(i)) - \frac{\gamma}{2} (N_A(t) - N_B(t)).
\]

(11)

When we transform the expected reward \(E_j(t)\) into

\[
Q'_j(t) = E_j(t)/2,
\]

(12)

where \(j \in \{A, B\}\), we can obtain the difference

\[
Q'_A(t) - Q'_B(t) = (\sum_{i=1}^{N_A(t)} R_A(i) - \sum_{i=1}^{N_B(t)} R_B(i)) - \frac{\gamma}{2} (N_A(t) - N_B(t)).
\]

(13)

From Eq. (13), the learning rule of the extended LTOW model for continuous-valued rewards is defined by

\[
Q'_j(t) = \sum_{i=1}^{N_j(t)} R_j(i) - \frac{\gamma}{2} N_j(t).
\]

(14)

When the number of machines is greater than 2, \(\gamma\) is given by \(\bar{R}_1(t) + \bar{R}_2(t)\), where \(\bar{R}_1(t)\) and \(\bar{R}_2(t)\) are the first- and second- highest average reward values, respectively [10]. Using the learning rule of the extended LTOW model (Eq. (14)), branch displacements are calculated in the same manner as in the original LTOW model:

\[
X_i(t) = Q'_j(t) - \frac{1}{n-1} \sum_{k=1, k \neq i}^{n} Q'_k(t) + \zeta(t).
\]

(15)

4. Results

To check the validity of the proposed method, we experimentally implemented a system composed multiple terminals and an access point (Fig. 4). We have developed a new access point which can use all the channels in the 2.4 GHz band by using 14 wireless LAN interfaces. In the access point, we configured a network bridge that constantly keeps connecting from a static IP address to change the channels quickly. The terminals run the extended LTOW model, which has five interfaces and one interface, and the original LTOW model to decide which channel is to be selected.

We evaluated the performance of the proposed methods that utilize continuous-valued rewards (using one interface and five interfaces), and the original LTOW model with binary rewards. We measured the TCP throughput of an 800 MByte file during morning (9-10am), daytime (1-2pm), evening (5-6pm) and nighttime (9-10pm) hours (Fig. 5). In all time periods, the download time of the proposed method was shorter, or becomes more effective, than that of the original LTOW model.
The method using one interface provided approximately the same performance as the method using five interfaces. The results also indicate that the proposed method is effective in the morning and evening hours when there are many users.

To check whether the proposed method can select an optimal channel that changes over time, we compared the download time of the proposed method with that of a conventional wireless LAN system.
In this experiment, we downloaded a 200 MByte file every hour for two days, using 14 terminals that assign fixed channels and one terminal with one wireless LAN interface running the proposed LTOW model. Figure 6 shows the download time results. The download time of the proposed method is shorter than that of the conventional wireless LAN system with the fixed channels, even though the proposed method's time includes delays caused by changing channels. Although the download time of ch.14, which does not receive much interference from adjacent channels is the longest, it is assumed that the traffic in ch.14 was heavy when we conducted this experiment. From the result, we conclude that the proposed method selects channels adaptively.

5. Conclusion
In this paper, we proposed and experimentally demonstrated a multi-armed bandit algorithm with continuous-valued rewards taking account of the interference effects in wireless communications in order to realize enhanced efficiency.

In the experiments, the proposed LTOW model with continuous-valued rewards shows better performances than the original LTOW model which does not consider the effect of interference from adjacent channels. By estimating the vacancy of adjacent channels, the proposed method exhibits higher performances even when only one wireless LAN interface is used in the terminal. Moreover, in the experiment that measured performance over longer time periods, the performance provided by the proposed method was better than that of a conventional wireless LAN system with fixed channels, even though the proposed method’s performance is affected by delays caused by changing channels. From these results, the effectiveness of the proposed method is clarified.

Acknowledgments
This work was supported in part by the Core-to-Core Program, A. Advanced Research Networks and Grants-in-Aid for Scientific Research (A) from Japan Society for the Promotion of Science.

References


