Performance Analysis of Multi-armed Bandit Algorithm with Negative Autocorrelation

Tomohiro Kato, Hiroshi Kurita and Mikio Hasegawa

Graduate School of Engineering, Tokyo University of Science
6–3–1 Niijuku, Katsushika, Tokyo 125-8585, Japan
Phone/FAX:+81-03-5876-1717
E-mail: \{kato,kurita\}@haselab.ee.kagu.tus.ac.jp,hasegawa@ee.kagu.tus.ac.jp

Abstract

We analyze the effectiveness of a multi-armed bandit algorithm, which utilizes ideal spatiotemporal chaotic dynamics generated by an FIR filter. In the previous research on additive chaotic noise to heuristic searches for combinatorial optimization problems, it has been shown that the chaotic sequences with negative autocorrelation improve the performance of asynchronously updated algorithms. The effectiveness of chaos can be understood in terms of the conventional theory of the chaotic CDMA, which showed that the cross-correlation between the sequences with negative autocorrelation becomes lowest. The spatiotemporal chaotic searching dynamics with such lowest cross-correlation has been shown to be effective in improving asynchronously updated algorithms. In this paper, as an asynchronously updated multi-armed bandit algorithm, we apply the FIR filter to the softmax algorithm, and analyze the spatiotemporal dynamics of this method. Our numerical simulation results show that the cross-correlation of this method can be minimized and its performance can be improved by using negative autocorrelation.

1. Introduction

The effectiveness of chaotic searching dynamics for combination optimization problems has been shown by many researchers [1]-[8]. The chaotic dynamics is more effective than stochastic searching methods like the simulated annealing and deterministic methods such as the Taboo Search. There are three chaotic optimization approaches: mutually connected neural network with additive chaos noise [3], mutually connected chaos neural network [1], and the heuristics method driven by chaotic neural networks [6]. The effective chaotic search has also been analyzed in previous research [4, 5]. In [5], it was shown, by the method of surrogate data, that a particular autocorrelation of additive chaotic noise has an important role in performance improvement of the optimization neural networks that should be asynchronously updated. From the analysis, it has been clarified that effective autocorrelation takes a negative value when lag is 1, and converges to zero with damped oscillation. The chaos with such negative autocorrelation has been used also in the research of asynchronous chaotic CDMA. The chaotic series with negative autocorrelation, $C(\tau) \approx C \times r^\tau, r = -2 + \sqrt{3}$, is effective for reducing the mutual interference [9]-[11]. It has been mathematically shown that the cross-correlation among the time series can be minimized by using those with such negative autocorrelation [10].

The negative autocorrelation used in the asynchronous chaotic CDMA has been applied also to the optimization neural networks, and its effectiveness in improving the performance has been shown [11]. By using negative autocorrelation, the searching dynamics of asynchronous algorithms can be ideally diversified and a high-performance search for an optimal solution can be realized. In order to generate such an ideal dynamics, in the research on chaotic CDMA, several methods of generating the negative autocorrelation time series have been proposed [9, 12, 13]. As one of those methods, the Lebesgue spectrum filter (LSF) has been employed to adjust the autocorrelation of the time series to negative autocorrelation [11]. The LSF has also been applied to several optimization methods, such as the neural network and the 2-opt heuristic method for QAP and TSP [14], and its effectiveness has been clearly demonstrated. However, the conventional research works only concern such benchmark optimization problems.

In this paper, we apply the LSF to an algorithm for the multi-armed bandit problem, which has practical applications such as in stochastic selection to maximize a reward. We analyze the performance of the proposed approach by evaluating the improvement of the performance of the softmax algorithm [15] for 10- and 20-armed bandit problems. We also analyze the best autocorrelation for the multi-armed bandit problems by tuning the parameter of the LSF.

2. Conventional Multi-armed Bandit Algorithms

The multi-armed bandit problem is a problem to maximize a total reward by selecting the best stochastic system. One of the examples is the maximization of the reward by selecting the best slot machine. When there are two or more slot
machines whose reward probabilities are unknown, a player attempts to maximize the total sum of rewards obtained after playing the machines. That is, the player wants to find, as quickly and accurately as possible, the optimal machine such that its reward probability becomes maximum. There are various algorithms to solve the multi-armed bandit problem, such as the ε-greedy algorithm and the softmax algorithm [15, 16]. In this study, we select the softmax algorithm and improve its performance by adopting the LSF.

In the softmax algorithm, the estimated reward probability \( g_i(t) \) for all the machines is used for making the decisions. \( g_i(t) \) is calculated by the following equation,

\[
g_i(t) = \frac{P_i(t)}{U_i(t)}
\]

where \( P_i(t) \) is the total number of rewards for machine \( i \) at time \( t \) and \( U_i(t) \) is the total number of playing machines \( i \) at time \( t \). The softmax algorithm selects an arm using the probability \( \epsilon_i(t) \), which is calculated using the probability \( g_i(t) \).

\[
\epsilon_i(t) = \frac{\exp[\beta(t)g_i(t)]}{\sum_{k \in I} \exp[\beta(t)g_k(t)]}
\]

Here, \( \beta(t) = \tau \times t \), where \( \tau \) is a parameter that determines the decay rate. \( \beta(t) = 0 \) corresponds to uniform probability for all selections. With increasing time \( t \), the selected arm gradually converges to the maximum \( g_i(t) \), corresponding to the greedy action. Some studies have revealed that the softmax algorithm is the best algorithm for the multi-armed bandit problem [16].

3. Ideal Search Realized Using Lebesgue Spectrum Filter

In asynchronous chaotic CDMA, the LSF is used to minimize cross-correlation by negative autocorrelation for each sequence. The LSF is denoted as

\[
f'(t) = \sum_{\tau=0}^{M} r^\tau f(t - \tau)
\]

where \( f(t) \) is an input at time \( t \), \( f'(t) \) is the output of LSF at time \( t \), and \( r \) is a parameter for adjusting the autocorrelation. Equation (3) can be transformed to the following form suitable for numerical computation.

\[
f'(t) = rf'(t - 1) + f(t)
\]

It is possible to adjust autocorrelation to \( C(\tau) \approx C \times r^\tau, r = -2 + \sqrt{3} \), using the LSF. When the input has white autocorrelation, the cross-correlation among the output sequences can be minimized by setting \( r = -2 + \sqrt{3} \).

4. Application of LSF to Bandit Problem

We apply the LSF to the softmax algorithm, and realize ideal selection dynamics, which realizes wide-area search by the lowest cross-correlation among the selections for each arm. The selections are decided using \( g_i'(t) \), which is the output of the LSF applied to the estimated reward probability \( g_i(t) \).

\[
g_i'(t) = rg_i'(t - 1) + g_i(t)
\]

Using the LSF, asynchronous cross-correlation among \( g_i(t) \) can be minimized by using negative \( r \). Therefore, in order to apply this merit, it is necessary to update asynchronously, though the original Softmax algorithm is updated synchronously. At each step of updating one \( g_i(t) \), the probability \( \epsilon_i(t) \) for selecting the corresponding \( r \)th arm is calculated and the play is applied with the probability \( \epsilon_i(t) \). It is noted that \( t \) is not the number of plays for the proposed asynchronously updated algorithm.

5. Results

In this paper, the performance characteristics of the algorithms are evaluated using two problems, the 10- and 20-armed bandit problems shown in Table 1. Each result is the average value over 1000 runs.

<table>
<thead>
<tr>
<th>Arms</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

The parameter of the softmax algorithm is \( \tau = 0.04 \). We analyze the effectiveness of applying the LSF to the softmax algorithm in Section 5.1, and that of the negative autocorrelation in Section 5.2.

5.1 Evaluating effectiveness of the LSF applied to softmax algorithm

First, the effectiveness of applying the LSF to the softmax algorithm is evaluated. The relationship between the number of plays and the probability of selecting the optimal arm is shown in Figs. 1 and 2 for the two problems, respectively. The results of using the method with the LSF given for in the case of using the best \( r \). From these results, it is clear that the LSF improves the performance of the softmax algorithm. Improvement of the performance is greater in the case with 20 arms. It is clarified that the performance can be improved by applying the LSF, especially in difficult high-dimensional problems.
5.2 Evaluating effectiveness of negative autocorrelation for bandit problems

It has been shown in previous research that negative autocorrelation minimizes cross-correlation and improves the performance of the chaotic CDMA and combinatorial optimization algorithms. In this study, the effectiveness of such a low cross-correlation dynamics is analyzed on the Softmax algorithm. The relationship between the parameter $r$ and the rate of selection of the optimal arm is shown in Fig. 3. The number of plays for each is 10000, and $r$ is changed from $-1$ to 1 at 0.1 intervals. Figure 3 clarifies that the performance is higher when parameter $r$ becomes a negative value. It can be seen that the best $r$ is around $-0.4$, which is close to the theoretical value $r = -2 + \sqrt{3}$ for minimizing the asynchronous cross-correlation. Figures 4 and 5 show the relationship between the actual autocorrelation of $g_i(t)$ and rates of selection of the optimal arm and cross-correlation. From these results, negative autocorrelation actually minimizes cross-correlation and the performance of the softmax algorithm is improved.
6. Conclusion

In this research, we showed that the performance of the softmax algorithm, which is one of the algorithms for the multi-armed bandit problems, can be improved by applying the LSF. The negative autocorrelation generated by the LSF minimizes the cross-correlation among the selections of the arms and realizes ideally wide selections of the arms. In the previous research, the effectiveness of the LSF was shown in asynchronous chaotic CDMA and combinatorial optimization problems. Our results showed that the LSF can also improve the performance of the bandit algorithms, which have various and wide application fields.

Acknowledgment

This research is partially supported by the Aihara Innovative Mathematical Modelling Project of the Japan Society for the Promotion of Science (JSPS) through the “Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program),” initiated by the Council for Science and Technology Policy (CSTP).

References


