2010 International Symposium on Nonlinear Theory and its Applications
NOLTA2010, Krakow, Poland, September 5-8, 2010

Improvement of Tug-of-war Model for Two-armed Bandit Problem: Biologically Inspired Computing Method for Nonlocally-correlated Parallel Searches

Song-Ju Kim†, Masashi Aono†, and Masahiko Hara†

†Flucto-Order Functions Research Team, RIKEN-HYU Collaboration Research Center, Advanced Science Institute, RIKEN
Fusion Technology Center 5F, Hanyang University, 17 Haengdang-dong, Seongdong-gu, Seoul 133-791, Korea
2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan
Email: songju@riken.jp

Abstract—The "tug-of-war (TOW) model" proposed in our previous studies [1, 2, 3] is a unique method for parallel searches inspired by the photoavoidance behavior of the single-celled amoeba, the true slime mold *Physarum*. In the TOW model, many branches of the amoeba act as search agents to collect information on light stimulations while conserving the total sum of their resources (volume). We showed that the nonlocal correlation via resource conservation can be advantageous to manage the "exploration–exploitation dilemma" for solving the multi-armed bandit problem.

In this study, we investigate the effect of the information from the other branch on the TOW model's performance, for the purpose of improving the model. We improve the TOW model so that it can exhibit better performances regardless of the reward probabilities.

1. Introduction

We consider that there must be some crucial differences between biological organisms and digital computers with respect to their information processing. We expect that biological organisms are good at dealing with some kind of problems. In the amoeba's body (the true slime mold *Physarum* (Fig. 1A)), a constant amount of intracellular protoplasmic sol shuttles through tubular channels, while its extracellular gel layer (ectoplasm), like a sponge, rhythmically oscillates the contraction tension to squeeze and absorb the sol (Fig. 1B). While the amoeba oscillates its branches to collect environmental information, the volume of the sol flowing through its body remains constant, unless nutrients are provided. We are interested in how this physical conservation law affects the information processing of the amoeba [1, 2, 3, 4]. To elucidate this issue, we considered the "multi-armed bandit problem" because it is related to the difficulties of biological organisms faced while adapting to uncertain environments.

In this study, we focused on the two-armed bandit problem stated as follows. Consider a slot machine that has 2 arms. Both arms have individual reward probabilities P_A and P_B. At each trial, a player pulls one of the arms and obtains some reward, for example, a coin, with the corresponding probability[1]. The player wants to maximize the total reward sum obtained after a certain number of selections. However, it is supposed that the player does not know these probabilities. How can the player gain maximal rewards? The problem is to determine the optimal strategy for selecting the arm which yield maximum rewards by referring to past experiences. In the original form of the problem, the player was allowed to pull only one arm at each trial. However, to explore the advantages of parallel computing, we allowed the player to simultaneously pull both the arms. With this modification, the situation becomes more realistic, as it were a "two-bandit problem." The new form of the problem considers 2 slot machines A and B, each having only 1 arm. Machines A and B have reward probabilities P_A and P_B, respectively.

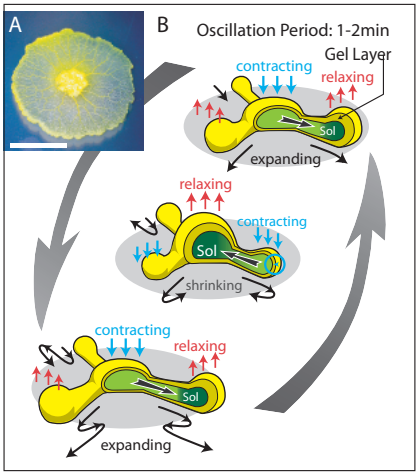

Figure 1: (A) An individual unicellular amoeba of the true slime mold *Physarum polycephalum* (scale bar = 7 mm). (B) Schematic illustration of the amoeba's body architecture.

The player has to "explore" many unknown machines to gather much information to determine the best machine. However, these explorations are risky because the player

[1] In this study, we assume that each pull results in a reward of fixed size with the given probability. We are dealing with the simplified variant of the general two-armed bandit problem.

may lose considerable rewards that could have been "exploited" from the already-known best machine. Thus, there is a trade-off between "exploration" and "exploitation." Living organisms generally encounter this "exploration–exploitation dilemma" because they have to survive in an unknown world. In order to survive, organisms need to adapt to the unknown situations by overcoming this dilemma. We speculate that organisms would have developed some efficient methods to overcome this dilemma.

In our previous studies [1, 2, 3], we proposed the "tug-of-war (TOW) model" which is a unique method for parallel searches inspired by the photoavoidance behavior of the true slime mold amoeba. The TOW model is a bio-inspired computing method capable of effectively solving problems without necessarily being a biological model for reproducing an amoeba's behavior. In our previous reports, we showed that the nonlocal correlation via resource conservation can be advantageous to manage the "exploration–exploitation dilemma" for solving the multi-armed bandit problem. We showed that the average accuracy rate of the TOW model is higher than those of well-known algorithms such as the modified ϵ-greedy algorithm and modified softmax algorithm. We also showed that the TOW model effectively adapts to a changing environment in which the reward probabilities dynamically switch.

In this study, we investigate the performances of the extended version of the TOW model for the two-armed bandit problem, and show that the optimized weight parameters depend on reward probabilities. This fact suggests that the performance of the TOW model can be further enhanced. We propose an improved TOW model which can exhibit better performances regardless of the reward probabilities.

2. Models

2.1. Tug-of-war Model

On the basis of the photoavoidance behavior of an amoeba, we proposed the tug-of-war (TOW) model in our previous studies [1, 2, 3]. Consider that the shape of an amoeba is like a slug, as shown in Fig. 2. Variables x_A and x_B correspond the volume increments in branch A and B, respectively. If x_A (x_B) is greater than 0, we consider that the amoeba selects A (B). Subsequently, light stimuli are applied to the branch A (B) with the probability $1 - P_A$ ($1 - P_B$) as a "punishment," i.e., an effect opposite to a "reward." In this model, there can be 4 types of selections: A, B, A and B, and no selection at each time step.

The volume increments x_A and x_B are determined by the following difference equations:

$$x_A(t+1) = x_A(t) + v_A(t), \quad (1)$$
$$x_B(t+1) = x_B(t) + v_B(t), \quad (2)$$

$$v_A(t) = v_A(t-1) + a_A(t), \quad (3)$$
$$v_B(t) = v_B(t-1) + a_B(t). \quad (4)$$

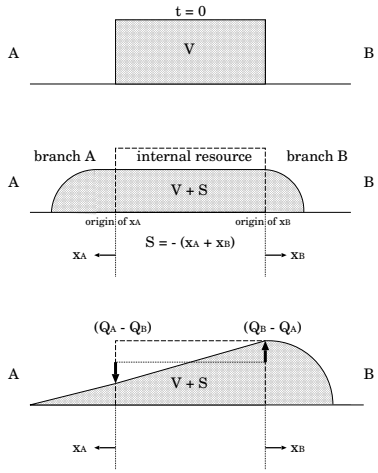

Figure 2: TOW model.

Here, $v_A(t)$ and $v_B(t)$ denote velocities of the corresponding volume increment, and $a_A(t)$ and $a_B(t)$ denote accelerations of the corresponding volume increment.

The internal resource deviation from the constant amount of the total resource V, $S(t)$, is determined by the following equation:

$$S(t+1) = S(t) - \Big(v_A(t) + v_B(t)\Big). \quad (5)$$

If the initial conditions ($x_A(0)$, $x_B(0)$, $v_A(0)$, $v_B(0)$ and $S(0)$) are set to zero, the value $x_A(t) + x_B(t) + S(t)$ will always be zero. This implies that $S(t)$=$-(x_A(t)+x_B(t))$, ensuring the conservation of the total resource V.

In order to incorporate the learning mechanism into this model, we introduced local biases of internal resource Q_A and Q_B for the resource on branches A and B, respectively (see the bottom figure in Fig. 2). For every time t, the number of selections and number of stimulations are accumulated in $Q_X(t)$ such that

$$Q_X(t) = \mu\left(N_X - 2\,L_X\right), \quad (6)$$

where N_X is the number of selection X (A or B) until time t, and L_X is the number of light stimulations on X (A or B) side until time t. Here, μ is the learning parameter.

By referring to the information on the number of selections and number of light stimulations, we assumed that a local bias of the internal resource is formed on each branch A or B. Thus, the local resource deviations $S_A(t)$ and $S_B(t)$ are given by

$$S_A(t) = S(t) + Q_A(t-1) - Q_B(t-1), \quad (7)$$
$$S_B(t) = S(t) + Q_B(t-1) - Q_A(t-1). \quad (8)$$

This implies that the communication between branch A and branch B is realized via resource conservation.

In the model, accelerations are essential variables (driving force). The acceleration $a_X(t)$ (X= A or B) is determined from Table 1; it depends on the local resource deviation $S_X(t)$ and light ON-OFF condition. The intrinsic

Table 1: Rule for determining acceleration a_X (X= A or B)

	$S_X < 0$	$S_X = 0$	$S_X > 0$
OFF	0	+1	+1
ON	–1	–1	0

dynamics of the model including the rule given in Table 1 are deterministic. However, variables $a_A(t)$ and $a_B(t)$ are determined stochastically, as the external light stimuli are applied in a probabilistic manner.

If the amoeba selects A, without the light stimuli (OFF), the acceleration $a_A(t) = +1$ will be added to $v_A(t)$, except in the case of $S_A(t) < 0$. This implies that if the local resource is abundant (S_A is zero or a positive value), the no light stimuli (OFF) induce an increase in v_A. If the amoeba selects A in the presence of the light stimuli (ON), the acceleration $a_A(t) = -1$ will be added to $v_A(t)$, except in the case of $S_A(t) > 0$. This implies that if the local resource is scarce (S_A is zero or a negative value), the light stimuli (ON) induce a decrease in v_A. In this way, the photoavoidance behavior of the amoeba is implemented in this model.

3. Results

3.1. Optimization of Weight Parameters

In order to investigate the effect of L_X in Eq. (6) on the TOW model's performance, we adopt the following form, instead of Eq. (6):

$$Q_X(t) \quad = \quad \mu \left(N_X - (1 + w)\, L_X \right). \tag{9}$$

Here, w is the weight parameter. In the original TOW model, the weight parameter w is always 1.

The above form, Eq. (9), is equivalent to the following form:

$$Q'_X(t) \quad = \quad \mu \left(N_X - \ L_X + w\, L_Y \right), \tag{10}$$

because of the fact that $Q_A - Q_B = Q'_A - Q'_B$ in Eqs. (7) and (8). Here, Y is A if $X = B$, or B if $X = A$. In the form (10), the first two terms denote the information of success (no light stimulation) in a branch, while the third term denotes the information of failures (light stimulation) of the other branch. The weight parameter w can be interpreted as the contribution weight from the other branch.

How does this parameter w affect the model's performance? The performance of the model is evaluated in termes of the "accuracy rate"; accuracy rate is defined as the rate of correct (higher probability) selections made until t. Figure 3 shows the average accuracy rates for the models with $P_A = 0.4$ and $P_B = 0.6$ (circle), $P_A = 0.4$ and $P_B = 0.7$ (square), $P_A = 0.4$ and $P_B = 0.8$ (triangle up), $P_A = 0.3$ and $P_B = 0.6$ (diamond), and $P_A = 0.45$ and $P_B = 0.6$ (triangle down), respectively. The horizontal axis denotes

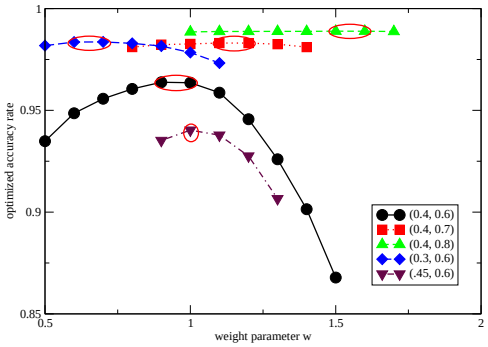

Figure 3: Optimized average accuracy rate of the TOW model for $P_A = 0.4$ and $P_B = 0.6$ (circle), $P_A = 0.4$ and $P_B = 0.7$ (square), $P_A = 0.4$ and $P_B = 0.8$ (triangle up), $P_A = 0.3$ and $P_B = 0.6$ (diamond), and $P_A = 0.45$ and $P_B = 0.6$ (triangle down), respectively.

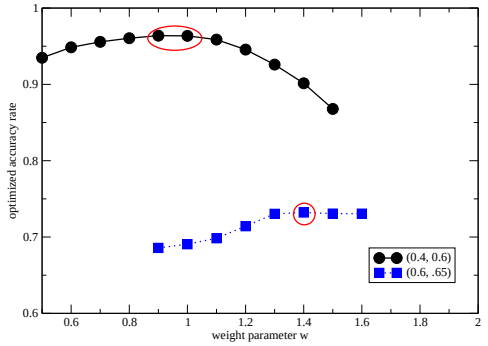

Figure 4: Optimized average accuracy rate of the TOW model for $P_A = 0.4$ and $P_B = 0.6$ (circle), and $P_A = 0.6$ and $P_B = 0.65$ (square), respectively.

the weight parameter w, and the vertical axis denotes the average accuracy rate at the number of selections = 500 for 1,000 samples of the TOW model. At each weight parameter w, the learning parameter μ was optimized in order to obtain the highest average accuracy rate. The elliptic curve on each line denotes its peak. The optimal ws (peaks in Fig. 3) depend on reward probabilities. This fact suggests that the performance of the TOW model can be further enhanced. We can summarize the dependence as follows: (I) The optimal w is 1.0 if reward probabilities have a symmetry (the mean value of reward probabilities is 0.5.). (II) If the mean value is larger (smaller) than 0.5, the optimal w is also larger (smaller) than 1.0. (III) The shift of the optimal w from 1.0 is proportional to the deviation of the mean value of reward probabilities from 0.5.

From Fig. 3, the parameter $w = 1.0$ is the best choice except for the cases in which the problem is difficult ($|P_A - P_B|$ is small) and does not have the symmetry ($P_A + P_B \neq 1$). We call these cases "non-symmetric difficult problems." Figure 4 shows the average accuracy rate for such case, namely the model with $P_A = 0.6$ and $P_B = 0.65$ (square).

In the same way, the elliptic curve on each line denotes its peak. In this case ($P_A = 0.6$ and $P_B = 0.65$), the peak is about 0.732 at $w = 1.4$. This value is not a little larger than 0.690 at $w = 1.0$ (original TOW). Therefore, we have to improve the TOW model so that the model can exhibit better performance even for such cases. It is easy to develop the model which can exhibit the best performances if we know the reward probabilities P_A and P_B. However, we can use only estimates for those probabilities, such as $q_X = \frac{N_X - L_X}{N_X}$, $(X = A \text{ or } B)$.

3.2. Improved TOW Model

We investigated performances of several improved models which were found by using heuristic method, and eventually found the two best forms. If we substitute $w = 1.0 + \gamma D$ to Eq. (9), we can obtain the following equation:

$$Q_X(t) = \mu \left(N_X - 2\, L_X \, - \gamma\, D\, L_X \right). \tag{11}$$

Here, γ is a parameter, and D is the deviation from the symmetry defined as follows:

$$D = \frac{1}{2}\left(q_A + q_B - 1 \right), \tag{12}$$

$$= \frac{1}{2}\left(1 - \frac{L_A}{N_A} - \frac{L_B}{N_B} \right). \tag{13}$$

Computer simulation studies showed that $\gamma = 1.0$ is the best for its performance.

If we substitute $\gamma = 2$ and $D = (1/2 - L_X/N_X)$ to Eq. (11), instead of Eq. (13), we can obtain the following simplified form:

$$Q_X(t) = \mu \left(N_X - 2\, L_X\, (3/2 - L_X/N_X) \right). \tag{14}$$

In fact, these two improved models, Eqs.(11, 13) and (14), can exhibit better performances. For example, in the case of $P_A = 0.6$ and $P_B = 0.65$, the average accuracy rate at the number of selections = 500 is 0.708, which is larger than 0.690 in the original TOW model although the value is still smaller than 0.732 at $w = 1.4$. In the other cases except for "non-symmetric difficult problems," these models exhibit almost the same performances as the original TOW model.

4. Conclusions and Discussions

We improved the "tug-of-war (TOW) model" which conducts unique parallel searches using many nonlocally correlated search agents. The conservation law entails a "nonlocal correlation" among the branches, i.e., volume increment in one branch is immediately compensated by volume decrement(s) in the other branch(es). This nonlocal correlation was shown to be useful for decision making in the case of a dilemma. In our previous reports [1, 2, 3], the average accuracy rate of the model is higher than those of well-known algorithms such as the modified ϵ-greedy algorithm and modified softmax algorithm. Moreover, the model flexibly adapts to changing environments, a property essential for living organisms surviving in uncertain environments.

In this study, we investigated performances of the extended version of the TOW model for the two-armed bandit problem. We added the weight parameter w to the original TOW model, and show that the optimized weight parameters depend on reward probabilities. This implied that the TOW model can be improved for better performance. Using heuristic method, we developed improved TOW models which can exhibit better performances regardless of the reward probabilities. The improved models were the best improvements among those we have ever examined. However, whether a better improvement will be possible is still open problem.

This TOW model is applicable to the Monte-Carlo tree search which is used in algorithms for the "game of GO" [5, 6]. We believe that the variant of the TOW model will become one of the best promising approaches to develop the effective algorithm due to its parallelism and nonlocal correlation between search agents.

References

[1] S. -J. Kim, M. Aono, M. Hara, "Tug-of-war model for two-bandit problem," In: C. Calude, et al. (Eds.), *Unconventional Computation, Lecture Notes in Computer Science* vol.5715, Springer, p.289, 2009.

[2] S. -J. Kim, M. Aono, M. Hara, "Tug-of-war model for multi-armed bandit problem," In: C. Calude, et al. (Eds.), *Unconventional Computation, Lecture Notes in Computer Science* vol.6079, Springer, pp.69–80, 2010.

[3] S. -J. Kim, M. Aono, M. Hara, "Tug-of-war model for the two-bandit problem: Nonlocally-correlated parallel exploration via resource conservation," *BioSystems* vol.101, pp.29–36, 2010.

[4] M. Aono, Y. Hirata, M. Hara, K. Aihara, "Resource-competing oscillator network as a model of amoeba-based neurocomputer," In: C. Calude, et al. (Eds.), *Unconventional Computation, Lecture Notes in Computer Science* vol.5715, Springer, pp.56–69, 2009.

[5] L. Kocsis, C. Szepesvári, "Bandit based monte-carlo planning," In: J. G. Carbonell, et al. (Eds.), *17th European Conference on Machine Learning, Lecture Notes in Artificial Intelligence* vol.4212, Springer, pp.282–293, 2006.

[6] S. Gelly, Y. Wang, R. Munos, O. Teytaud, "Modification of UCT with patterns in monte-carlo Go," *RR-6062-INRIA*, pp.1-19, 2006.