Optimal Strategy under Unknown Stochastic Environment — Nonparametric Lob-Pass Problem

Kazuyuki Hiraoka, and Shun-ichi Amari, *Fellow, IEEE*

Abstract

The bandit problem consists of two factors, one being exploration or the collection of information on the environment and the other being the exploitation or taking benefit by choosing the optimal action in the uncertain environment. It is necessary to choose only the optimal actions for the exploitation, while the exploration or collection of information requires to take a variety of (non-optimal) actions as trials. Hence, in order to obtain the maximal cumulative gain, we need to compromise the exploration and exploitation processes. We treat a situation where our actions change the structure of the environment, of which a simple example is formulated as the lob-pass problem by Abe and Takeuchi. Usually, the environment is specified by a finite number of unknown parameters in the bandit problem, so that the information collection part is to estimate their true values. The present paper treats a more realistic situation of nonparametric estimation of the environment structure which includes an infinite number (a functional degrees) of unknown parameters. The asymptotically optimal strategy is given under such a circumstance, proving that the cumulative loss can be made of the order $O(t^{\epsilon})$ where ϵ is an arbitrarily small constant ($\epsilon > 0$) and t is the number of trials, in contrast with the optimal order $O(\log t)$ in the parametric case.

Index Terms—bandit problem, stochastic game, optimal strategy, nonparametric estimation, stochastic approximation.

K. Hiraoka is with the Department of Information Engineering, University of Tokyo, Tokyo 113, Japan.

S. Amari is with the Department of Information Engineering, University of Tokyo, Tokyo 113, Japan. He is also with Riken Frontier Research Program on Brain Information Processing, The RIKEN, Wako, Japan.

I Introduction

It is desirable to choose the best decision or reaction in the stochastic environment on which our knowledge is limited. A simple idea is first to collect information on the environment through our actions and then to choose the best action based on the estimated structure of the environment. In order to collect information efficiently, it is more useful to choose a variety of actions than to choose the single optimal action. Hence, we need to compromise the information collecting process (exploration) and decision process (exploitation) in order to maximize the cumulative gain in the long-run decisions. A typical problem combining these two is the classic two- or multi-armed bandit problem [1][2][3]. This is regarded as an on-line learning problem of rational choice.

The situation may be a little more complex in the real world in the sense that the environment is not fixed but changes depending on our actions [4]. The lob-pass problem formulated by Abe and Takeuchi [5] is a simple but a typical example. They considered a model of tennis play in which a player has two alternatives "lob" and "pass". The opponent is not in a fixed state but his state changes depending on the player's action. Let s be the rate of lob in the past trials of the player. It takes on the real value, $0 \leq s \leq 1$. We assume that the opponent's strategy depends on s so that his state is determined by s. The behavior of the opponent is characterized by two functions $f_L(s)$ and $f_P(s)$ which are unknown to the player and are to be estimated through games. The $f_L(s)$ and $f_P(s)$ are the probabilities that a lob and a pass, respectively, are successful to get a point when the opponent is in state s.

Abe and Takeuchi [5] studied the problem from the point of view of computational learning theory, where $f_L(s)$ and $f_P(s)$ are supposed to be linear functions including unknown parameters. So exploration (information collection) is necessary to estimate the unknown parameters. The point is that the best information collection is not derived from the best choice of alternatives. When the s satisfies $f_L(s) > f_P(s)$ $[f_L(s) < f_P(s)]$, the instantaneous optimal strategy is to choose a lob [a pass]. However, this choice is not optimal for estimating the parameters of the unknown curves $f_L(s)$ and $f_P(s)$. Abe, Takeuchi and Amari [6] proved that there exists a strategy for which the cumulative loss grows in the order $\log t$ when the number t of trials is large. Such a strategy is also shown to be realizable by a simple greedy algorithm [7]. This is the asymptotically optimal achievable strategy in the sense of order, because the amount of the Fisher information is at most of order t. Hence, the parameters can be estimated with an accuracy of $O(1/\sqrt{t})$ so that the instantaneous loss cannot be made smaller than $1/t$ in order.

The present paper studies the nonparametric case where the functions $f_L(s)$ and $f_P(s)$ are completely unknown except that $f_L(s)$ $[f_P(s)]$ is a monotonically decreasing [increasing] smooth convex function. This is a nonparametric estimation problem from the statistical point of view where the unknown parameter specifying the environment has infinite degrees (function degrees) of freedom.

We show that the Fisher information degenerates in this case so that the ordinary asymptotic theory of statistics cannot be applied. We construct a strategy of which the cumulative loss grows of order $O(t^{\epsilon})$ for an arbitrarily small $\epsilon > 0$. This is arbitrarily close to the theoretical lower bound $O(\log t)$ in order. This is a best asymptotic result since the above $O(\log t)$ bound is not achievable in the nonparametric case. We use a method similar to the stochastic approximation [8][9][10][11] in order to overcome the difficulty arising from the nonparametric situation.

Our result is asymptotic in the sense that the number t of trials is very large. It is static, since we assume the matching condition [5] which guarantees that a stationary strategy eventually gives the optimal minimax solution in the asymptotic sense. Without this condition, a nonstationary strategy can have a better performance. This is one of the interesting cases to be solved in future.

When the player knows that the number T of total trials is fixed and finite, he needs to maximize the cumulative loss until T and does not care after that. In this case, another interesting transient situation arises. This transient case is studied by another paper [12] under the condition that $f_L(s)$, $f_P(s)$ and T are known.

II Statement of the problem

The definitions, notations and conditions on the structure of the lob-pass problem are formally given in this section. Let x_t denote the choice of the player at time t $(t = 1, 2, \cdots)$, $x_t = 1$ implying that a "lob" is chosen and $x_t = 0$ implying that a "pass" is chosen. The outcome of the trial at time t is denoted by z_t, $z_t = 1$ implying that the player wins and $z_t = 0$ implying that he loses at time t. The state s_t of the opponent at time t is the past lob rate of the player,

$$s_t = \frac{1}{t-1}\sum_{i=1}^{t-1} x_i \quad (t \geq 2), \tag{1}$$

and s_1 is arbitrary. The player determines x_t, that is, a lob or a pass, based on the results of past trials $\{(x_1, s_1, z_1), \cdots, (x_{t-1}, s_{t-1}, z_{t-1})\}$.

The outcome z_t is a binomial random variable depending on x_t and s_t,

$$z_t = \begin{cases} 1, & \text{with probability } f_L(s_t) \\ 0, & \text{with probability } 1 - f_L(s_t) \end{cases}$$

when a lob is chosen $(x_t = 1)$, and

$$z_t = \begin{cases} 1, & \text{with probability } f_P(s_t) \\ 0, & \text{with probability } 1 - f_P(s_t) \end{cases}$$

when a pass is chosen $(x_t = 0)$. The characteristic curves $f_L(s)$ and $f_P(s)$ are unknown except that they satisfy regularity conditions to be stated soon (Figure 1). The goal is to choose x_t, $t = 1, 2, \cdots$, so as to maximize the expectation of the cumulative gain, that is, the number of wins,

$$G_t = E\sum_{i=1}^{t} z_i, \tag{2}$$

where E denotes the expectation and t is assumed to be sufficiently large.

Let us define

$$w(r, s) = rf_L(s) + (1-r)f_P(s), \tag{3}$$

which is the expectation of z by the mixed strategy of taking a lob with probability r and a pass with probability $1-r$ when the state is s. The expected cumulative gain is then written as

$$G_t = E\sum_{i=1}^{t} w(x_i, s_i). \tag{4}$$

The following conditions are imposed on the curves $f_L(s)$ and $f_P(s)$.

Condition 1 (smoothness, monotone and nontriviality conditions) *For $0 \le s \le 1$, the functions $f_L(s)$ and $f_P(s)$ are twice continuously differentiable and monotone,*

$$f'_L(s) < 0, \quad f'_P(s) > 0, \tag{5}$$

satisfying

$$f_L(0) > f_P(0), \quad f_L(1) < f_P(1) \tag{6}$$

at the end points $s = 0, 1$.

Condition 2 (concavity condition) *The two functions are concave,*

$$f''_L(s) \le 0, \quad f''_P(s) \le 0. \tag{7}$$

The meaning of the monotone condition is clear: The winning probability $f_L(s)$ $[f_P(s)]$ by a lob [pass] decreases [increases] as the player tries lobs more frequently in the past. The concavity condition is also natural. To explain it, we define the mixed strategy of choosing a lob with probability r and a pass with probability $1-r$. Then the concavity condition implies the following: Taking the mixed strategy with lob probability r in $2T$ trials is better than dividing it into two parts of taking the mixed strategy with probability $r-a$ $(0 < a < r)$ in the first T trials and then taking that with probability $r+a$ in the second T trials, provided s is kept constant. The latter strategy is easier for the opponent.

Let us define

$$w^*(r) = w(r, r) \tag{8}$$

which is the eventual winning probability of the player when he takes the stationary mixed strategy with probability r for long runs, since s_t converges to r in this case.

Theorem 1 *Under conditions 1 and 2, there is a unique s^* such that the stationary strategy with s^* is optimal among all the stationary strategies, that is*

$$w^*(s^*) \geq w^*(r) \tag{9}$$

for all r.

Proof: We have

$$\frac{d}{dr}w^*(r) = f_L(r) - f_P(r) + r(f_L' - f_P') + f_P',$$
$$\frac{d^2}{dr^2}w^*(r) = 2(f_L' - f_P') + rf_L'' - (1-r)f_P'' < 0,$$

where $'$ denotes the differentiation. Hence, $w^*(r)$ is concave and $dw^*/dr(0) > 0$, $dw^*/dr(1) < 0$, so that $w^*(r)$ has a unique maximum s^*. □

The present paper studies how the asymptotically optimal stationary strategy can be learned under the uncertainty situation that f_L and f_P are unknown. To this end, we assume

Condition 3 (matching condition)

$$f_L(s^*) = f_P(s^*). \tag{10}$$

This condition is also assumed in Abe and Takeuchi [5], and Kilian, Lang, and Pearlmutter [7]. We can show that, when the matching condition does not hold, there exists a non-stationary strategy better than the optimal stationary strategy. For example, a mixed strategy whose lob-probability r_t is slowly oscillating around s^* is better than the stationary strategy, when the amplitude and period are adequately chosen.

If we could know the exact value of s^*, the optimal strategy at t is to choose a lob when $s_t \leq s^*$ and to choose a pass when $s_t > s^*$. The asymptotic average winning probability by the optimal stationary strategy is obviously $w^*(s^*) = w(s^*, s^*)$. Hence, we define the mean instantaneous loss of a trial sequence $\{(x_i, s_i, z_i)\ ;\ i = 1, 2, \cdots\}$ by

$$l_i = w(s^*, s^*) - Ew(x_i, s_i) \tag{11}$$

and the cumulative loss by

$$L_t = \sum_{i=1}^{t} l_i. \tag{12}$$

Our purpose is to find the strategy which minimizes L_t. The problem is said to be parametric, when the unknown functions f_L and f_P are specified by an unknown finite-dimensional vector parameter $\boldsymbol{\theta}$, $f_L(r) = f_L(r, \boldsymbol{\theta})$, $f_P(r) = f_P(r, \boldsymbol{\theta})$. In this case, an estimate $\hat{s}_t^*$ of s^* is obtained by using the estimate $\hat{\boldsymbol{\theta}}_t$ obtained from the past t trials. In the parametric case, we can show by calculating the Fisher information that the maximum likelihood estimator satisfies

$$\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\| \sim O_p\left(1/\sqrt{t}\right) \tag{13}$$

where O_p denotes the stochastic order. This implies that the best estimator $\hat{s}_t^*$ is deviated from s^* by $O(1/\sqrt{t})$. On the other hand, the loss based on the estimate $\hat{s}_t^*$ is of order

$$O\left(|\hat{s}_t^* - s^*|\right) = O_p\left(1/t\right), \tag{14}$$

because s^* is the maximum of a convex differentiable function $w^*(s)$ and $|\hat{s}_t^* - s^*| \propto \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|^2$. Hence, we can choose a strategy satisfying

$$l_t = O\left(1/t\right) \tag{15}$$

for a large t, and we have a strategy with

$$L_t = O(\log t) \tag{16}$$

via statistical estimation (Abe and Takeuchi [5]) or by a greedy algorithm (Kilian, Lang, and Pearlmutter [7]). On the other hand, there are no strategies better than this, because such a better strategy contradicts the Cramér-Rao bound of statistical estimation.

The present nonparametric case is different from the parametric case.

Theorem 2 *For any strategy, the cumulative loss is strictly larger than* $\log t$ *in order,*

$$\frac{L_t}{\log t} \to \infty \quad (t \to \infty). \tag{17}$$

Sketch of the proof: There exists no $\sqrt{t}$-consistent estimators, that is, no estimators satisfying

$$E|\hat{\theta}_t - s^*|^2 = O\left(1/t\right) \tag{18}$$

in the present case. This is shown by calculating the effective Fisher information. As is shown in Appendix B, the effective Fisher information degenerates in the present case. Statistical estimation under such a case is studied in Kawanabe and Amari [13]. Refer to Amari and Kawanabe [14] and Bickel et al.[15] for the nonparametric or semiparametric estimation. Since we have the strict inequality $l_t > O\left(1/t\right)$ in order, we have $L_t/\log t \to \infty$ as $t \to \infty$. □

III Optimal strategy

The following main theorem shows that there exists a strategy whose cumulative loss is arbitrarily close to the unachievable lower bound $O(\log t)$ in order.

Theorem 3 *For any* $\epsilon > 0$, *there exists a strategy whose cumulative loss is of order* $O(t^\epsilon)$,

$$L_t = O(t^\epsilon). \tag{19}$$

We prove the theorem by constructing such a strategy. This strategy is similar to the Robbins-Monro procedure in the stochastic approximation method [8]. A rough sketch of the strategy is like this: The player tries to make his lob rate r_t equal to the estimated optimal value $\hat{s}^*$, where the true s^* satisfies the relation $f_L(s^*) = f_P(s^*)$. When r_t converges to s^*, so does s_t. To this end, he estimates the current error $F(s_t) = f_P(s_t) - f_L(s_t)$ at s_t, and modifies the estimate $\hat{s}^*$ of s^* depending on the above estimated error. Then he plays some trials with the new lob ratio $\hat{s}^*$, and these procedures are repeated.

Now we will describe the strategy formally. Let us divide a sequence of trials into sessions. Let the number of trials in the n-th session be

$$M_n = \lceil n^\beta \rceil, \quad \beta > 0.5 \tag{20}$$

where $\lceil n^\beta \rceil$ denotes the smallest integer not less than n^β. Therefore, the first session consists of 1 trial, the second session consists of $\lceil 2^\beta \rceil$ trials, and so on. The n-th session begins with the t_n-th trial in the whole sequence, where

$$t_n = 1 + \sum_{j=1}^{n-1} M_j = O(n^{\beta+1}), \tag{21}$$

and the n-th session ends with the $t'_n = (t_n + M_n - 1)$-th trial (Figure 2). The parameter $\beta > 0.5$ is chosen arbitrarily, but it will be shown that, the larger β is, the better the asymptotic performance of the strategy is.

The first session: The first session consists of only one trial $t = 1$. We choose a lob, $x_1 = 1$.

Before the n-th session ($n \geq 2$) starts, we decide the lob rate R_n in the n-th session based on the results in the $(n-1)$-th session. It converges to the optimal value s^* to be searched for. The initial value R_1 is arbitrary. Roughly speaking, R_n is an estimator of s^* recursively determined from R_{n-1} and the results in the $(n-1)$-th session, to be shown soon.

The n-th session ($n \geq 2$): The n-th session consists of $M_n = \lceil n^\beta \rceil$ trials from $t = t_n$ to $t = t'_n = t_n + M_n - 1$. Through this session, the player choose lobs $\lceil R_n M_n \rceil$ times and passes $M_n - \lceil R_n M_n \rceil$ times. In order to avoid large fluctuations, the distribution of these lobs and passes should be uniform in M_n trials in the n-th session so that x_t is decided as

$$x_t = \begin{cases} 1, & \text{if} \quad \sum_{i=t_n}^{t-1} x_i < R_n(t - t_n) \\ 0, & \text{if} \quad \sum_{i=t_n}^{t-1} x_i \geq R_n(t - t_n) \end{cases} \tag{22}$$

for $t_n \leq t \leq t'_n$.

The lob rate R_{n+1} for the next session is decided as follows. Since the target value s^* satisfies $F(s^*) = 0$, where

$$F(s) = f_P(s) - f_L(s), \tag{23}$$

we modify the current R_n to give a closer approximation to s^*. To this end, we estimate the value of F at the current s_t by

$$\hat{F}_n = \hat{f}_{P,n} - \hat{f}_{L,n}, \tag{24}$$

where the estimates $\hat{f}_{P,n}$ and $\hat{f}_{L,n}$ are the empirical winning rates of lobs and passes in the n-th session given by

$$\begin{aligned} \hat{f}_{L,n} &= \frac{\sum_{M_n} x_i z_i}{\sum_{M_n} x_i}, & (25) \\ \hat{f}_{P,n} &= \frac{\sum_{M_n} (1 - x_i) z_i}{\sum_{M_n} (1 - x_i)}. & (26) \end{aligned}$$

Here $\sum_{M_n}$ means $\sum_{i=t_n}^{t'_n}$. By using this estimate, when $\hat{F}_n$ is positive (negative), $\tilde{R}_{n+1}$ is put smaller (larger) than R_n, as

$$\tilde{R}_{n+1} = R_n - a_n \hat{F}_n, \tag{27}$$

where

$$a_n = \frac{1}{4C} \frac{M_n}{t'_n} \sim O(1/n). \tag{28}$$

Here, the constant C needs to be chosen such that it satisfies

$$\frac{f_P(s) - f_L(s)}{s - s^*} \geq C, \tag{29}$$

for all $s \neq s^*$. The existence of such a $C > 0$ is guaranteed from the assumptions. When we do not know such a C, we need to estimate it through trials. This is in the same situation as the stochastic approximation method. In order to avoid the case that the number of lobs or passes are too small to collect minimum required information about $f_L(s_t)$ and $f_P(s_t)$, when $\tilde{R}_{n+1}$ is too close to 0 or 1, we modify $\tilde{R}_{n+1}$ as

$$R_{n+1} = \begin{cases} (n+1)^{-0.5}, & \text{when } \tilde{R}_{n+1} \leq (n+1)^{-0.5} \\ \tilde{R}_{n+1}, & \text{when } (n+1)^{-0.5} < \tilde{R}_{n+1} < 1 - (n+1)^{-0.5} \\ 1 - (n+1)^{-0.5}, & \text{when } \tilde{R}_{n+1} \geq 1 - (n+1)^{-0.5} \end{cases} \tag{30}$$

for $n \geq 4$, and $R_2 = R_3 = R_4 = 1/2$. This trick forces the player to choose both lobs and passes at least $\lfloor n^{\beta - 0.5} \rfloor$ times in the n-th session ($n \geq 5$). Then the estimation error for $\hat{F}_n$ is guaranteed to converge to 0 as n increases. Note that the forced number of this "information collecting trials" $\lfloor n^{\beta - 0.5} \rfloor$ is much smaller than the length of the whole session $M_n = \lceil n^\beta \rceil$ if n is very large. Moreover, this forced correction occurs with a probability tending to 0 since $0 < s^* < 1$ and $R_n \to s^*$. So it can be neglected practically.

IV Analysis of the strategy

Theorem 3 is proved by analyzing the performance of the proposed strategy. First of all, we study the behavior of the state s_t of the opponent under this strategy. Let us define S_n to be the past lob ratio s_t at the head of the n-th session $t = t_n$,

$$S_n = s_{t_n}. \tag{31}$$

Then, in the n-th session,

$$\begin{aligned} s_t &= \frac{1}{t-1}\left((t_n-1)S_n + \lceil R_n(t-t_n)\rceil\right) \\ &= S_n + \frac{t-t_n}{t-1}(R_n - S_n) + \frac{\xi_t}{t-1} \end{aligned} \tag{32}$$

for $t_n < t \le t_{n+1}$, where ξ_t is the term due to rounding off $R_n(t-t_n)$ to an integer, satisfying $0 \le \xi_t \le 1$. In particular,

$$S_{n+1} = S_n + \frac{M_n}{t'_n}(R_n - S_n) + \frac{\xi_n^{(1)}}{t'_n}, \tag{33}$$

where $\xi_n^{(1)} = \xi_{t_{n+1}}$ satisfies $0 \le \xi_n^{(1)} \le 1$. From (32), it is also proved that s_t never goes away more than $2/(t_n-1)$ from the interval between S_n and S_{n+1} for $t_n \le t \le t_{n+1}$,

$$s_t \in I_n = \begin{cases} \left[S_n - \frac{2}{t_n-1},\quad S_{n+1} + \frac{2}{t_n-1}\right] & (\text{if } S_n \le S_{n+1}) \\ \\ \left[S_{n+1} - \frac{2}{t_n-1},\quad S_n + \frac{2}{t_n-1}\right] & (\text{if } S_n > S_{n+1}). \end{cases} \tag{34}$$

We will evaluate the mean square deviation of R_n and S_n from s^* by

$$V_n = E\left[(R_n - s^*)^2 + (S_n - s^*)^2\right] \tag{35}$$

and see how fast V_n converges to 0. To this end, we introduce the following recursive equation

$$\begin{aligned} \begin{pmatrix} R_{n+1} - s^* \\ S_{n+1} - s^* \end{pmatrix} = \begin{pmatrix} R_n - s^* \\ S_n - s^* \end{pmatrix} &+ A_n \begin{pmatrix} R_n - s^* \\ S_n - s^* \end{pmatrix} \\ &+ \text{small error term,} \end{aligned}$$

where A_n is a matrix. This equation is derived by evaluating $\hat{F}_n$ (See Appendix A). By evaluating the eigenvalues of A_n, we have the following lemma whose proof is also given in Appendix A.

Lemma 1 *Under the assumption of Theorem 3, the proposed strategy satisfies*

$$V_n \le O(n^{-\beta}) \tag{36}$$

for large n.

Then the following lemma is proved from Lemma 1.

Lemma 2 *The cumulative loss L_t for the proposed strategy is not greater than $O(t^{\frac{1}{\beta+1}})$:*

$$L_t = E\sum_{i=1}^{t}\Big\{w(s^*, s^*) - w(x_i, s_i)\Big\} \leq O(t^{\frac{1}{\beta+1}}). \tag{37}$$

Proof: The n-th session consists of $M_n = O(n^\beta)$ trials. Let us divide it again to $O(n^{\beta/2})$ subsessions so that each subsession consists of $O(n^{\beta/2})$ trials.

First, we evaluate the scattering of the lob rate r and the state s_i in the subsession. Let us consider the k-th subsession in the n-th session. This subsession contains $m_{n,k} = O(n^{\beta/2})$ trials. As to r, the lob rate $r_{n,k} = \sum^{n,k} x_i/m_{n,k}$ in the subsession satisfies

$$|r_{n,k} - R_n| \leq O(n^{-\beta/2}) \tag{38}$$

because of (22), where $\sum^{n,k}$ denotes the summation over the trials $\{i\}$ in the k-th subsession in the n-th session. Hence

$$E(r_{n,k} - s^*)^2 \leq O(n^{-\beta}) \tag{39}$$

is obtained from Lemma 1. As to s_i, define $s_{n,k}^{\min}$ and $s_{n,k}^{\max}$ to be the minimum and maximum value of s_i in this subsession. They satisfy

$$s_{n,k}^{\max} - s_{n,k}^{\min} \leq \frac{m_{n,k}}{t_n}|R_n - S_n| + O\Big(\frac{1}{t_n}\Big) \tag{40}$$

from (32), and so

$$E[s_{n,k}^{\max} - s_{n,k}^{\min}] \leq O\left(\frac{m_{n,k}}{t_n}\sqrt{V_n}\right) \leq O(n^{-\beta-1}) \leq O(n^{-\beta}). \tag{41}$$

Furthermore,

$$E(s_{n,k}^{\min} - s^*)^2 \leq O(n^{-\beta}) \tag{42}$$

follows from (34) and Lemma 1.

Now we will evaluate the cumulative loss in the subsession. By the definition of $w(r,s)$ (3) and the monotone condition (5),

$$w(x_i, s_i) \geq w(x_i, s_{n,k}^{\min}) - f'_{L,\max}(s_{n,k}^{\max} - s_{n,k}^{\min}) \tag{43}$$

for the trial i in the subsession, where $f'_{L,\max}$ is a constant defined as

$$f'_{L,\max} = \max_r |f'_L(r)| < \infty. \tag{44}$$

This implies

$$\sum\nolimits^{n,k} w(x_i, s_i) \geq m_{n,k}\Big\{w(r_{n,k}, s_{n,k}^{\min}) - f'_{L,\max}(s_{n,k}^{\max} - s_{n,k}^{\min})\Big\} \tag{45}$$

because $w(r,s)$ is a linear function for r. From the Taylor's theorem, there exists a constant W such that

$$|w(r,s) - w(s^*,s^*)| \leq W\Big\{(r-s^*)^2 + (s-s^*)^2\Big\} \tag{46}$$

for arbitrary r,s since $w(s^*,s^*)$ is an extremum of the function $w(r,s)$. Therefore the cumulative loss in the subsession is evaluated as

$$\begin{aligned} \sum\nolimits^{n,k} l_i &= \sum\nolimits^{n,k} E\Big[w(s^*,s^*) - w(x_i,s_i)\Big] \\ &\leq m_{n,k} E\Big[w(s^*,s^*) - w(r,s_{n,k}^{\min}) + f'_{L,\max}(s_{n,k}^{\max} - s_{n,k}^{\max})\Big] \\ &\leq O(n^{-\beta/2}) \end{aligned} \tag{47}$$

because of (39) (41) (42) (46).

The n-th session consists of $O(n^{\beta/2})$ such subsessions. Thus the cumulative loss $\tilde{L}_n$ in the n-th session is

$$\tilde{L}_n = \sum_{i=t_n}^{t'_n} l_i \leq O(1). \tag{48}$$

Then we get

$$L_{t'_n} \leq \sum_{k=1}^{n} \tilde{L}_k \leq O(n). \tag{49}$$

Since L_t is monotone increasing in t and $n = O(t^{\frac{1}{\beta+1}})$ for $t_n \leq t < t'_n$,

$$L_t \leq L_{t'_n} \leq O(n) = O(t^{\frac{1}{\beta+1}}). \tag{50}$$

□

Theorem 3 is immediately proved by putting $\beta = (1/\epsilon) - 1$ in Lemma 2.

V Conclusion

The asymptotically optimal strategy of the lob-pass problem is presented under the nonparametric environment. The nonparametric (or semiparametric) estimation is a most attractive area in modern statistics. The Fisher information in this situation degenerates to 0, so that the ordinary asymptotic theory of parametric or nonparametric statistics cannot be applied to this problem. Nevertheless, we proved that the cumulative loss of the proposed strategy is of order $O(t^\epsilon)$ for an arbitrarily small $\epsilon > 0$, which is as close as $O(\log t)$ of the parametric case. The best strategy is attained by a stationary mixed strategy under the matching condition which we assumed.

Many interesting problems remain to be studied in future. One is to study an oscillatory strategy of cheating the opponent to obtain a larger cumulative gain. Another one is to obtain the maximum cumulative gain when the game ends in a fixed finite time T. The problem can also be generalized in a various way where the computational efficiency might play an important role.

Acknowledgment

It is our pleasure to thank Dr. N. Abe, the NEC C&C Research Laboratory, for valuable discussions. We also thank Prof. S. Yoshizawa for encouragements. The present research is partly supported by Grant-in-Aid of the Priority Research Area on Higher Information Processing in the Brain, Ministry of Education and Culture, and the Real World Computing Program, MITI, Japan.

Appendices

A

In this appendix we give the proof of Lemma 1. Through the n-th session ($n \geq 5$), the player is forced to choose lobs at least $\lfloor n^{\beta-0.5} \rfloor$ times. Since the estimator $\hat{f}_{L,n}$ is the arithmetic mean of independent binomial random variables taking values 0 and 1, and since the variance of such a variable is not greater than 1/4 in general, the variance of $\hat{f}_{L,n}$ is bounded as

$$\mathrm{Var}[\hat{f}_{L,n}|D_{n-1}] \leq \frac{1}{4}\frac{1}{n^{\beta-0.5}-1}. \tag{51}$$

Here $D_n = \{(x_1, s_1, z_1), \cdots, (x_{t'_n}, s_{t'_n}, z_{t'_n})\}$, and $\mathrm{Var}[\,\cdot\,|D_{n-1}]$ denotes the conditional variance that the data D_{n-1} of past trials is known. Similarly

$$\mathrm{Var}[\hat{f}_{P,n}|D_{n-1}] \leq \frac{1}{4}\frac{1}{n^{\beta-0.5}-1} \tag{52}$$

is satisfied for passes. Thus

$$\mathrm{Var}[\hat{F}_n|D_{n-1}] \leq \frac{1}{2}\frac{1}{n^{\beta-0.5}-1}. \tag{53}$$

On the other hand, from the mean value theorem and the range of s_t shown in (34), $E[\hat{F}_n|D_{n-1}]$ is decomposed into three terms,

$$\begin{aligned} E[\hat{F}_n|D_{n-1}] = F(S_n) \quad &+ \quad F'_{\max}\frac{M_n}{t'_n}(R_n - S_n)\xi_n^{(2)} \\ &+ \quad F'_{\max}\frac{2}{t_n - 1}\xi_n^{(3)}, \end{aligned}$$

where

$$\begin{aligned} &F'_{\max} \equiv \max_{0\leq s\leq 1} F'(s) < \infty, \qquad (54) \\ &0 \leq \xi_n^{(2)} \leq 1, \\ &-1 \leq \xi_n^{(3)} \leq 1. \end{aligned}$$

$$\tag{55}$$

Putting

$$\eta_n = \sqrt{2(n^{\beta-0.5}-1)}(\hat{F}_n - E[\hat{F}_n|D_{n-1}]), \tag{56}$$

we get $E[\eta_n|D_{n-1}] = 0$ and $\mathrm{Var}[\eta_n|D_{n-1}] \leq 1$. The estimator $\hat{F}_n$ of $F(S_n)$ is then rewritten as

$$\begin{aligned} \hat{F}_n &= F(S_n) + F'_{\max}\frac{M_n}{t'_n}(S_n - R_n)\xi_n^{(2)} \\ &\quad + F'_{\max}\frac{2}{t_n - 1}\xi_n^{(3)} + \frac{1}{\sqrt{2(n^{\beta-0.5} - 1)}}\eta_n. \end{aligned} \tag{57}$$

From the smoothness condition, there exists two constants C, C' such that

$$0 < C \leq C' < \infty, \tag{58}$$

$$C \leq \frac{F(s)}{s - s^*} \leq C' \qquad \text{(for all } s \neq s^*\text{)}. \tag{59}$$

We define

$$C_n \equiv \begin{cases} \dfrac{F(S_n)}{S_n - s^*} & \text{(for } S_n \neq s^*\text{)} \\ F'(s^*) & \text{(for } S_n = s^*\text{)}. \end{cases} \tag{60}$$

Then C_n satisfies

$$C \leq C_n \leq C' \quad \text{(for } n = 1, 2, \ldots\text{)}. \tag{61}$$

With these symbols, the renewal of (R_n, S_n) is written as

$$\begin{aligned} \begin{pmatrix} \tilde{R}_{n+1} - s^* \\ S_{n+1} - s^* \end{pmatrix} &= \begin{pmatrix} R_n - s^* \\ S_n - s^* \end{pmatrix} + A_n \begin{pmatrix} R_n - s^* \\ S_n - s^* \end{pmatrix} \\ &\quad + \begin{pmatrix} -a_n F'_{\max}\dfrac{2}{t_n - 1}\xi_n^{(3)} - a_n \dfrac{\eta_n}{\sqrt{2(n^{\beta-0.5} - 1)}} \\ \dfrac{1}{t'_n}\xi_n^{(1)} \end{pmatrix}, \end{aligned} \tag{62}$$

where

$$A_n = \begin{pmatrix} a_n F'_{\max}\dfrac{M_n}{t'_n}\xi_n^{(2)} & -a_n\left[C_n + F'_{\max}\dfrac{M_n}{t'_n}\xi_n^{(2)}\right] \\ \dfrac{M_n}{t'_n} & -\dfrac{M_n}{t'_n} \end{pmatrix}. \tag{63}$$

Since $a_n = M_n/(4Ct'_n) = O(1/n)$, the eigenvalues λ_n^+ and λ_n^- of A_n is

$$\lambda_n^{\pm} = \frac{1}{2}\frac{M_n}{t'_n}\left(-1 \pm \sqrt{1 - (C_n/C)}\right) + o\left(\frac{M_n}{t'_n}\right) \tag{64}$$

for large n.

Note that R_n is always closer to s^* than $\tilde{R}_n$ is for sufficiently large n, because the strict inequality $0 < s^* < 1$ is obtained from the nontriviality condition (6), and the restricted bounds $n^{-0.5}$ and $1 - n^{-0.5}$ for R_n converges to 0 and 1 respectively when n is large. Thus we get

$$\begin{aligned} V_{n+1} &\leq |1 + \lambda_n^+|^2 V_n + \frac{2(1 + 2a_n F'_{\max}\xi_n^{(3)})}{t_n - 1}|1 + \lambda_n^+|\sqrt{V_n} \\ &\quad + \frac{1 + (2a_n F'_{\max}\xi_n^{(3)})^2}{(t_n - 1)^2} + \frac{a_n^2}{2(n^{\beta-0.5} - 1)} \end{aligned}$$

$$\leq \left(|1+\lambda_n^+|^2 + \frac{(1+2a_n F'_{\max}\xi_n^{(3)})}{t_n - 1}|1+\lambda_n^+|\right) V_n$$
$$+ \left(\frac{1+2a_n F'_{\max}\xi_n^{(3)}}{t_n - 1} + \frac{1+(2a_n F'_{\max}\xi_n^{(3)})^2}{(t_n-1)^2} + \frac{a_n^2}{2(n^{\beta-0.5}-1)}\right). \tag{65}$$

Here

$$|1+\lambda_n^+|^2 \leq 1 - \frac{M_n}{t'_n} + o\left(\frac{M_n}{t'_n}\right), \tag{66}$$

because $\sqrt{1-(C_n/C)}$ is an imaginary number. Omitting higher order terms,

$$V_{n+1} \leq (1 - \frac{M_n}{t'_n})V_n + \frac{1}{t_n} \tag{67}$$

is obtained, where $M_n/t'_n = O(1/n)$ and $t_n = O(n^{\beta+1})$. This inequality implies the conclusion $V_n \leq O(n^{-\beta})$. (See the technique of Remark 2.5.1 in [9]. Consider $U_n = \varphi_n = n^{\beta-\delta}V_n$, where $\delta > 0$ is an arbitrary small number.) □

B

In this appendix we calculate the effective Fisher information to prove Theorem 2. When the conditions 1, 2, 3 are satisfied, we rewrite functions $f_L(s)$ and $f_P(s)$ in the following form by singling out linear terms,

$$\begin{aligned} f_L(s,s^*) &= \alpha - \beta(1-s^*)(s-s^*) + k_L(s-s^*), \quad (68)\\ f_P(s,s^*) &= \alpha + \beta(s^*)(s-s^*) + k_P(s-s^*). \quad (69)\end{aligned}$$

Here, the two curves intersect at $s = s^*$, so that s^* is the unknown parameter which is of interest. According to the common usage of symbols in statistics, we rewrite s^* as θ in the following. The constants $\alpha > 0$, $\beta > 0$ are unknown but we have no interest in their values. Such parameters are called nuisance parameters. The higher-order terms k_L and k_P are smooth functions satisfying

$$\begin{gathered} k_L(0) = k_P(0) = 0,\\ k'_L(0) = k'_P(0) = 0, \\ k''_L(0) < 0, \quad k''_P(0) < 0. \end{gathered} \tag{70}$$

Let $q(x)$ be the probability of choosing $x = 0, 1$. Then, the joint probability of (x, z) depends on the state s of the opponent and functions f_L and f_P, or equivalently the parameters α, β, $\theta = s^*$, k_P, k_L. We write the logarithm of the probability $P(x, z \mid s, \theta, \alpha, \beta, k_L, k_P)$ as

$$l(x, z \mid s, \theta, \alpha, \beta, k_L, k_P) = \log P(x, z \mid s, \theta, \alpha, \beta, k_L, k_P). \tag{71}$$

This is calculated as

$$\begin{aligned} l(x,z) &= \log q(x) \\ &\quad +\delta_1(x)\{z\log f_L + (1-z)\log(1-f_L)\} \\ &\quad +\delta_0(x)\{z\log f_P + (1-z)\log(1-f_P)\}, \end{aligned} \tag{72}$$

where

$$\delta_i(x) = \begin{cases} 0, & x \neq i, \\ 1, & x = i. \end{cases} \tag{73}$$

The family of probability distributions $\{P(x,z) \mid s, \theta, \alpha, \beta, k_L, k_P\}$ is called semiparametric, because it is specified by the parameter θ of interest and by the nuisance parameters α, β, k_L, k_P. The nuisance parameters include unknown functions k_L and k_P so that their dimensions are infinite or of function degrees of freedom. It is only recently that efficient estimation in the semiparametric model is fully analyzed [14][15]. Its asymptotic property is given by the efficient Fisher information which is the covariance matrix of the score of the parameter of interest projected to the orthogonal complement of the scores of the nuisance parameters. The score is the derivative of $l(x,z)$ with respect to parameters. Since k_L and k_P are functions, the derivatives with respect to these parameters are operators given by the Frechét derivatives.

Let us put

$$\begin{aligned} A_L &= -\frac{1}{1-f_L}, & A_P &= -\frac{1}{1-f_P}, \\ B_L &= \frac{1}{f_L(1-f_L)}, & B_P &= \frac{1}{f_P(1-f_P)}. \end{aligned} \tag{74}$$

The θ-score is given by

$$\begin{aligned} \frac{\partial}{\partial\theta} l(x,z) &= c_L(A_L + zB_L)\delta_1(x) \\ &\quad +c_P(A_P + zB_P)\delta_0(x) \end{aligned} \tag{75}$$

where

$$c_L = \beta(1+s-2\theta) - k_L'(s-\theta), \tag{76}$$

$$c_P = \beta(s-2\theta) - k_P'(s-\theta). \tag{77}$$

In order to calculate the nuisance scores, we put

$$k_L(s,t) = k_L(s) + ta_L(s), \tag{78}$$

$$k_P(s,t) = k_P(s) + ta_P(s) \tag{79}$$

where a_L and a_P are variations of k_L and k_P. Then, the nuisance scores in the directions of a_L and a_P are given by

$$\begin{aligned} \left.\frac{\partial}{\partial t} l\right|_{t=0} &= a_L(s-\theta)\{A_L + zB_L\}\delta_1(x) \\ &\quad +a_P(s-\theta)\{A_P + zB_P\}\delta_0(x). \end{aligned} \tag{80}$$

Since a_L and a_P are arbitrary functions satisfying $a_L(0) = a_P(0) = a'_L(0) = a'_P(0) = 0$, we can conclude that the θ-score is included in the linear space spanned by the nuisance scores, except at the point $s = \theta$. Since we do not know $\theta(= s^*)$, the Fisher information is degenerate. This implies that there exist no estimators $\hat{\theta}_t$ which have $\sqrt{t}$-consistency, that is, which satisfy

$$\lim_{t\to\infty} tE\left[(\hat{\theta}_t - \theta)^2\right] < \infty. \tag{81}$$

Theorem 2 is thus proved. □

References

[1] J. C. Gittins and D. M. Jones, "A dynamic allocation index for the sequential design of experiments," in *Progress in Satistics* (K. J.Gani and I. Vincze, eds.), pp. 241–266, North-Holland, Amsterdam, 1974.

[2] K. D. Glazebrook, "Optimal strategies for families of alternative bandit processes," *IEEE Trans. Automat. Control*, vol. AC-28, pp. 858–861, 1983.

[3] D. A. Berry and B. Fristedt, *Bandit Problems.* Chapman and Hall, 1985.

[4] R. J. Herrnstein, "Rational choice theory," *American Psychologist*, vol. 45, no. 3, pp. 356–367, 1990.

[5] N. Abe and J. Takeuchi, "The 'lob-pass' problem and an on-line learning model of rational choice," *Proceedings of the 1993 Workshop on Computational Learning Theory*, pp. 422–428, 1993.

[6] N. Abe, J. Takeuchi, and S. Amari, in preparation.

[7] J. Kilian, K. J. Lang and B. A. Pearlmutter, "Playing the matching-shoulder lob-pass game with logarithmic regret," *Proceedings of the 1994 Workshop on Computational Learning Theory*, 1994.

[8] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400–407, 1951.

[9] M. B. Nevel'son and R. Z. Has'minskiĭ, *Stochastic Approximation and Sequential Estimation.* Nauka, 1968 (in Russian; translated in Japanese by T. Kitagawa and K. Tajima, 1983).

[10] M. T. Wasan, *Stochastic Approximation.* Cambridge at the university press, 1969.

[11] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* No. 26 in Applied Mathematical Sciences, Springer-Verlag, 1978.

[12] K. Hiraoka, in preparation.

[13] M. Kawanabe and S. Amari, "Estimation of network parameters in semiparametric stochastic perceptron," *Neural Computation*, vol. 6, pp. 1244–1261, 1994.

[14] S. Amari and M. Kawanabe, "Information geometry of estimating functions in semiparametric statistical models," to appear.

[15] P. J. Bickel, C. A. J. Klassen, Y. Ritov and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore, 1993.

List of Figure Captions

Figure 1 is refered in the section II. Figure 2 is refered in the section III.

Figure 1:

Figure 2: