数理解析研究所講究録 1215 巻 2001 年 53-63

Toward Sophisticated Learning

Yuichi Noguchi*
Hitotsubashi University

May 10, 2001

Abstract

We summarize a recent development of the study of sophisticated learning with emphasis on learning to optimize.

1 Introduction

In the last decade, the theory of learning in games has intensively been studied. Many researchers have investigated learning dynamics including fictitious play and replicator dynamic. However, those learning rules are naive in the sense that those cannot adjust to even simple patterns, so that some researchers recently consider sophisticated learning as an important issue to study. It is natural to think that Bayesian learning is a candidate of sophisticated learning. In fact, Kalai and Lehrer (1993) study Bayesian learning in infinitely repeated games and show that the grain of truth condition implies Bayesian learners eventually play a Nash equilibrium path. This approach, however, has two problems. First, it is unclear how the grain of truth condition or other similar conditions is related to sophisticated learning. Second, as Nachbar (1996) shows, Bayesian learning players with the same degree of sophistication may fail to satisfy the grain of truth condition each other.

Another approach is proposed by Foster and Vohra (1993): they introduce a concept of calibrated learning and construct a calibrated learning rule. Their leaning rule is explained by the following example: suppose that a

*Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan. E-mail: ynoguchi@econ.hit-u.ac.jp

player or forecaster, who follows the learning rule, predicts a probability of rain, say p_t, at the beginning of every period. Now fix any p and pick up periods in which a player predicts $p_t = p$. Then, empirical distributions of rain in those periods always converge to p: the prediction is *empirically* correct. The Foster and Vohra learning rule is *calibrated* to empirical data in this sense. They consider repeated games where every player takes the leaning rule, and show that the path converges to a correlated equilibrium. Inspired by their work, Fudenberg and Levine (1995) and (1999) introduce a concept of the universal consistency and show that smooth fictitious play has the universal consistency property. In the following, I summarize Fudenberg and Levine's work and then explain Noguchi (1999) that develops their work to study the problem of learning to optimize against many regularities.

2 The Model

We consider one player who plays an infinitely repeated game against an opponent, where he may observe only a past history of actions in each period; the opponent may be a machine, nature, or consist of multiple players. The player's (resp. opponent's) pure action in a stage game is denoted by a (resp. y) and a set of all player's pure actions (resp. all opponent's pure actions) is finite, denoted by A (resp. Y). A set of probability distributions over S is denoted by $\Delta(S)$. $\lambda \in \Delta(A)$ (resp. $\pi \in \Delta(Y)$) denotes the player's mixed action (resp. the opponent mixed action). $u(\lambda, \pi)$ is written for the player's expected utility of a stage game. Histories of the repeated game consist of sequences of actions by the player and his opponent. A finite history with time length T is denoted by $h_T = (a_1, y_1, \cdots, a_T, y_T)$ and an infinite history is denoted by $h_\infty = (a_1, y_1, a_2, y_2, \cdots)$; we define $h_0 = \emptyset$. Let H denote a set of all finite histories and H_∞ denote a set of all infinite histories.

Our player takes an action on the basis of a past history in every period, so that his strategy may be represented by a behavior strategy $\sigma : H \to \Delta(A)$. A set of all player's behavior strategies is denoted by $\mathcal{B}_P$. We assume that the player does not know about any opponent characteristic except that the opponent plays a behavior strategy and takes an action in Y at each period. An opponent behavior strategy is denoted by ρ and a set of all opponent behavior strategies is denoted by $\mathcal{B}_O$.

3 Universal Consistency

A word "consistency" has several definitions in the learning literature. We first describe a definition of consistency in statisitics and then introduce that in the economic context. In statistics, the consistency means that prediction *averages* asymptotically coincide with empirical frequencies. For example, consider sequences of two possible states, rain and no rain. Let $\mu_t(R)$ be a prediction of rain at the beginning of period t and D_t an empirical frequency of rain up to period t. Then, the consistecy criterion requires that

$$\frac{\sum_{s=1}^{t} \mu_s}{t} - D_t \to 0, \text{ as } t \to \infty, \ a.s.$$

This criterion is unsatisfactory because the criterion cannot deal with regulariteis. Suppose that the weather is very cyclical, for example, it is rain in even periods and it is no rain in odd periods. Let a prediction always give $50-50$ probability. That is, the prediction is so naive that it does not learn the weather simple pattern. However, it is clear that the prediction passes the consistency criterion. Note that this criticism may be applied to the Foster and Vohra learning rule.

Fudenberg and Levine (1995) introduce the universal consistency in the economic sense (precisely its definition dates back to Hannan (1959)).

Definition 1 *A behavior strategy σ is said to be $\varepsilon-$universally consistent, if for all opponent strategies ρ*

$$\lim_{T\to\infty} \sup V(D(h_T)) - \frac{1}{T}\sum_{s=1}^{T} u(a_s, y_s) < \varepsilon, \ a.s.,$$

where $D(h_T)$ is an empirical distribution of all opponent actions up to period T and $V(D(h_T)) = \max\{u(\lambda, D(h_T)) \mid \lambda \in \Delta(A)\}$.

We need to give several remarks. First, the criterion is concerned with *optimization*, not prediction. Second, the criterion is universal in the sense that it requires *single* behavior strategy pass the criterion for *all* opposing behavior strategies. Third, a target in the criterion is a maximum payoff against an *empirical* distribution up to the current period, so that realized average payoffs eventually becomes at least as high as the minmax payoff in a stage game. Those mean that the criterion may be interpreted as a *safety*

criterion in the sense that the player's strategy does not perform very poorly against any opposing behavior strategy.

Fudenberg and Levine (1995) show that smooth fictitious play, that is, a smooth approximate best response against an empirical distribution of past opponent actions, is universally consistent.

Proposition 1 *(Fudenberg and Levine (1995)) For any $\varepsilon > 0$, a smooth fictitious play is ε−universally consistent.*

It is important to note that a very simple strategy such as smooth fictitious play has the property. Unfortunately, this criterion may be criticized by the same argument as in the statistical definition: it does not assure a plausible optimality against even very simple patterns. To understand it, consider a repeated matching pennies game where an opponent takes an alternating strategy:

$$H, T, H, T, H, T, H, T, H, T, H, T, H, T, H, T, H, T, \cdots$$

Then, the universal consistency criterion only assures that player's average payoffs may attain the minmax payoff 0. However, when the player is a little smart, he may recognize the simple pattern and try to adjust to it, so that he must earn much more payoffs than 0. It means that the universal consistency is too weak to evaluate sophisticated learning behaviors.

4 Conditional Universal Consistency

Fudenberg and Levine have been aware of the weakness, and then they propose "conditional" universal consistency (1999). They show that *conditional* smooth fictitious play has the conditional universal consistency property under mild assumptions. Let us explain conditional (smooth) fictitious play. It is just a (smooth approximate) best response to conditional empirical distributions. For example, suppose that a player consider samples should be separated according to even and odd periods. It may be represented by a *classification rule* $\mathcal{R}$ that is defined as a partition of H. In this example, $\mathcal{R} = \{\gamma_e, \gamma_o\}$, where $\gamma_e = \{h_t \mid t \text{ is even}\}$ and $\gamma_o = \{h_t \mid t \text{ is odd}\}$. Then, conditional fictitious play on $\mathcal{R}$ is a (smooth approximate) best response to an empirical distribution of past odd period samples in any odd peirod and a best reponse to an empirical distribution of past even period samples in any

even period. In a general case, a class in a classification rule $\mathcal{R}$ is called a category, denoted by γ; if a past history $h_T \in \gamma$, we say that γ is active in period $T+1$. We assume that each γ has prior samples represented by a vector d^0_γ; $n^0_\gamma = \sum d^0_{\gamma,y}$ is called a prior sample size for γ. Thus, when a past history h_{T-1} belongs to γ, a player considers that a current period T is $\gamma-$active. He collects observed samples $d_\gamma(h_{T-1})$ in past $\gamma-$active periods and plays a (smooth approximate) best resposne to an empirical distribution $\tilde{D}_\gamma(h_{T-1})$ of prior and observed samples conditional on γ: $\tilde{D}_\gamma(h_{T-1}) = \frac{d_\gamma(h_{T-1})+d^0_\gamma}{n_\gamma(h_{T-1})+n^0_\gamma}$, where $n_\gamma(h_{T-1}) = \sum_y d_{\gamma,y}(h_{T-1})$; $D_\gamma(h_T)$ denotes an empirical distribution only of observed samples in $\gamma-$active periods.

Proposition 2 *(Fudenberg and Levine (1999)) Suppose that a classification rule* $(\mathcal{R}, (d^\gamma_0)_{\gamma\in\mathcal{R}})$ *satisfies the following two assumptions:*

$$(1)\ \lim_{T\to\infty} \frac{K^{\mathcal{R}}(h_{T-1})}{T} = 0 \text{ for all } h_\infty \in H_\infty, \text{ and } (2)\ \sup_{\gamma\in\mathcal{R}} n^\gamma_0 < \infty$$

where $K^{\mathcal{R}}(h_{T-1})$ *is the number of categories that have been active up to period* T. *Then, for any* $\varepsilon > 0$, *a conditional smooth fictitious play on* $\mathcal{R}$ *satisfies the conditional universal consistency: for any opposing behavior strategy* ρ

$$\lim_{T\to\infty} \sup \| \frac{1}{T} \sum n^\gamma_T V(D_\gamma(h_T)) - \sum_{s=1}^{T} u(a_s, y_s) \| < \varepsilon, \text{ a.s.}$$

where n^γ_T *is the number of times that* γ *is active up tp period* T *and* $V(D_\gamma(h_T)) : \max\{u(\lambda, D_\gamma(h_T)) \mid \lambda \in \Delta(A)\}$.

However, the conditional universal consistency is still weak to evaluate a sophisticated player's behavior, although it is very useful to prove propositions in the next section: the target $\frac{1}{T}\sum_{\gamma\in\mathcal{R}} n^\gamma_T V(D_\gamma(h_T))$, a maximum average payoff against conditional empirical distributions on $\mathcal{R}$, might not be an upper bound to time average payoff, even if an opponent plays a simple regular strategy. The following example explains it:

Example 1 *Suppose that a player's payoff matrix is:*

	L	M	R
U	3	3	3
D	3	3	3

Let the player's classification rule be $\mathcal{R} = \{\gamma_{RL}, \gamma_M\}$, *where* $\gamma_{RL} = \{h_t \mid y_t = R \text{ or } L\}$ *and* $\gamma_M = \{h_t \mid y_t = M\}$. *Assume, however, that his opponent follows a cyclical behavior* $LMRLMRLMR\cdots$. *Even though a conditional smooth fictitious play on* $\mathcal{R}$ *is supposed to perform well according to the conditional universal consistency criterion, the conditional fictitious play on* $\mathcal{R}$ *generates only time-average utility* $\frac{2}{3}0 + \frac{1}{3}3 = 1$ ***by proposition 2. If the player is smart, then he could receive*** $\frac{1}{3}3 + \frac{1}{3}3 + \frac{1}{3}3 = 3$ ***as his time-average utility by switching to a cyclical behavior,*** $UDUUDUUDU\cdots$. ***It is hard to believe that the smart player takes the conditional fictitious play on*** $\mathcal{R}$ ***forever.***

The example suggests (1) we need a stronger criterion of time-average utility especially for regular opponent strategies and (2) if a player wants to learn his opponent strategy, then his classification rule should be more sophisticated (or finer) than a conditioning rule of the opponent strategy in some sense.

5 Optimal Properties of Conditional Fictitious Play

Taking into account the weakness of the consistency, Noguchi (1999) introduces two strong criteria: time average optimality and conditioning class optimality. The first criterion requires that a player may eventually obtain almost as high time-average payoff as if he knew a *true* opposing behavior strategy. The second one insists that a player may eventually obtain almost as high payoff as if he knew a true opposing behavior strategy. Then, he construct an "optimal" classification rule for each of those criteria and shows that conditional smooth fictitious play on the optimal rule pass the strong criterion for all regular opposing strategies. We shall explain those criteria and optimal classification rules.

5.1 Time average optimality

Recall the defect of the consistency: a target might not be an upper bound to time average utility. We shall intorduce a strong criterion to resolve the problem: time-average optimality.

Definition 2 *A behavior strategy $\sigma : H \to \Delta(A)$ is called $\varepsilon-$optimal in the average sense for $S \subset \mathcal{B}$, if for all $\rho \in S$*

$$\lim_{T\to\infty} \sup \frac{1}{T}[\sum_{s=0}^{T-1} V(\rho(h_s)) - \sum_{s=1}^{T} u(a_s, y_s)] < \varepsilon, \ \mu_{(\sigma,\rho)} - a.s.$$

where $\mu_{(\sigma,\rho)}$ is a probability distribution on H_∞ generated by σ and ρ, and $V(\rho(h_s))$ is a maximum payoff against $\rho(h_s)$.

In this definition, we put average maximum payoff against a *true* opponent strategy as a target, instead of that against empirical distributions. Note that it is an asymptotic least upper bound of time average utility: for any player's strategy σ and opposing strategy ρ,

$$\lim_{T\to\infty} \inf \frac{1}{T}[\sum_{s=0}^{T-1} V(\rho(h_s)) - \sum_{s=1}^{T} u(a_s, y_s)] \geq 0, \ a.s.$$

and the equality holds *a.s.* for $\bar{\sigma}$ with $\bar{\sigma}(h_s) \in \arg\max\{u(\lambda, \rho(h_s)) \mid \lambda \in \Delta(A)\}$ for all $h_s \in H$. It is easily obtained by combining the strong law of large numbers and a fact that $V(\rho(h_s)) \geq u(\lambda, \rho(h_s))$ for all $\lambda \in \Delta(A)$. Hence, if time average utility is almost as high as the target, a player, who is concerned with time average utility, has no incentive to change his behavior, so that the weakness of the consistency is resolved. However, we cannot hope the best result for the time average optimality criterion if a player has no weakly dominant action in a stage game: there is no single strategy that is optimal in the average sense for all opposing strategies; when there is a dominant strategy, the optimality is alway obtained by playing the dominant action.

Proposition 3 *Assume that there is no weakly dominant action in a stage game. Then, for some $\varepsilon_0 > 0$ there exists no ε_0-optimal behavior strategy for all opponent behavior strategies.*

Although it is impossible to obtain the best result, it is worthwhile studying a player's optimal strategy for *many* opponent behavior strategies. Indeed, we may show that there exists a time average optimal strategy for all opposing strategies generated by a countable family of finitely conditioning

Definition 3 *(1) A conditioning rule is a partition of H, denoted by $\mathcal{P}$. An element of a conditioning rule is called a conditioning class, denoted by β. A conditioning rule $\mathcal{P}$ is said to be finite if the number of classes in $\mathcal{P}$ is finite. (2) A conditioning rule of a behavior strategy ρ is a partition, denoted by $\mathcal{P}_\rho$, that is generated by the following equivalent relation:*

$$h_t \sim \bar{h}_{\tilde{t}} \Leftrightarrow \rho(h_t) = \rho(\bar{h}_{\tilde{t}})$$

The mixed action ρ_β conditional on a class β is uniquely defined by $\rho_\beta = \rho(h_t)$, $h_t \in \beta$.

Definition 4 *We say that a behavior strategy ρ is generated by a family of conditioning rules Ω, if a conditioning rule of ρ belongs to Ω, that is, $\mathcal{P}_\rho \in \Omega$.*

Proposition 4 *For any countable family of finitely conditioning rules $\{\mathcal{P}_i\}_{i=1}^{\infty}$ there exists a classification rule $\mathcal{R}_0$ such that for any $\varepsilon > 0$ some conditional smooth fictitious play on $\mathcal{R}_0$ is $\varepsilon-$optimal in the average sense for all behavior strategies generated by the family $\{\mathcal{P}_i\}_{i=1}^{\infty}$.*

This proposition implies that conditional smooth fictitious play is optimal in the avearge sense for all regular opposing strategies because conditioning rules of regular strategies are at most countable; regular strategies are those that have *computable* finitely conditioning rules (see Noguchi (1999) for its formal definition).

Instead of giving a rigorous proof, we will explain a basic idea of constructing an optimal rule $\mathcal{R}_0$. It is based on a given family $\{\mathcal{P}_i\}_{i=1}^{\infty}$; without loss of generality, we may assume that $\{\mathcal{P}_i\}_{i=1}^{\infty}$ is ordered by the fineness as partitions: $\mathcal{P}_i \prec \mathcal{P}_{i+1}$ for all i ($\mathcal{P}_i \prec \mathcal{P}_{i+1} \Leftrightarrow \forall \beta \in \mathcal{P}_{i+1} \exists \tilde{\beta} \in \mathcal{P}_i (\beta \subset \tilde{\beta})$). A key is that a player makes use of each $\mathcal{P}_i$ as a temporary classification rule in some periods. That is, $\mathcal{R}_0$ represents the following player behavior: at the first stage he starts with employing $\mathcal{P}_1$ as a temporary rule. But he has an incentive to change finer conditioning rules because a true conditioning rule might be finer than $\mathcal{P}_1$; the player could not receive the maximum average payoff against a true strategy if he got stuck in $\mathcal{P}_1$ and the true rule were finer than $\mathcal{P}_1$(recall example 1). The argument may be applied to any period and any conditioning rule he employs in that period: there is always possibility that a true conditioning rule might be finer than a current employed one. Therefore, he employs finer and finer rules as time proceeds. But he wants to change the rules *very slowly*. Delayed changes of temporary rules allow

the player to obtain enough samples for each rule, so that if his employing rule is *finer* than a true one, he could have a good prediction on the true opponent strategy and obtain the maximum average payoff to it. The player employs finer and finer rules, and eventually employed rules are finer than a true one. It means that the player may eventually obatin the maximum avearge payoff.

5.2 Conditioning class optimality

Even the time average optimality may be weak as a behavior criterion of a sophisticated player. The weakness is that the time average optimality criterion may ignore a performance of a player's strategy in active periods of a conditiong class whose frequency vanishes relative to all periods. Consider a repeated matching pennies game where a player always plays heads and his opponent plays the following regular strategy:

$$H, T, H, H, T, H, H, H, T, H, H, H, H, T, H, H, H, H, H, T, \cdots$$

That is, the frequency of playing tails diminishes *regularly* as time proceeds. The player suffers the worst outcome -1 whenever the opponent plays tails, although tails are played *regularly*. However, the player's strategy passes the time average optimality criterion for this opponent strategy because periods of the worst outcome asymptotically vanish relative to all periods. If the player is smart, then he would be aware of the regularity and very likely to play tails when his opponent plays tails. The example suggests that we need a stronger criterion in order to capture sophisticated behaviors: a criterion that assures an optimality in each conditioning class. We shall only give the definition of the conditioning class optimality for a case of finitely conditioned strategies, that is, strategies whose conditioning rules are finite (see Noguchi (1999) for a general case).

Definition 5 *A behavior strategy $\sigma : H \to \Delta(A)$ is called $\varepsilon-$optimal in the classwise sense for S, if for any opponent strategy and any conditioning class $\beta \in \mathcal{P}_\rho$*

$$\lim_{T\to\infty} \sup V(\rho_\beta) - \frac{1}{n_T^\beta} \sum_{\substack{h_s \in \beta \\ 0 \le s \le T-1}} u(a_{s+1}, y_{s+1}) < \varepsilon, \text{ if } n_T^\beta \to \infty, \text{ a.s.}$$

where n_T^β is the number of $\beta-$active periods up to period T and $V(\rho_\beta)$ is a maximum payoff against ρ_β.

This optimality criterion assures that a player may eventually obtain almost as high payoffs as if he knew a true opponent strategy, and the above problem is resolved.

Proposition 5 *For any countable family of finite conditioning rules* $\{\mathcal{P}_i\}_i$, *there exists a classification rule* $\mathcal{R}_1$ *such that for any* $\varepsilon > 0$, *some conditional smooth fictitious play strategy on* $\mathcal{R}_1$ *is optimal for all behavior strategies generated by the family, i.e.* $\mathcal{P}_\rho = \mathcal{P}_i \ \exists i$.

This criterion is stronger than the time average optimality and it would be the strongest as a myopic optimality criterion in the sense that the same proposition with a stronger criterion would not hold.

A basic idea of constructing $\mathcal{R}_1$ is very similar to that of the time average optimal rule $\mathcal{R}_0$. A difference is that a player switches *classwise*: if a class β in $\mathcal{P}_i$ obtains enough samples, then the class is changed to finer classes in $\mathcal{P}_{i+1}$, while any other class in $\mathcal{P}_i$ with few samples keeps to be employed, i.e. it is not switched to finer classes.

6 Concluding Remarks

We conclude by giving several remarks. First, we need to emphasize that the time average optimality and the conditioning class optimality (proposition 4 and 5) may be obtained by using the conditional universal consistency. This means that those optimalities may be attained *without the cost of the universal consistency.* Thus, single smooth fictitious play is not only optimal for many opposing strategies but also safe against all other ones. Second, we have focused on myopic optimality. However, nonmyopic cases are important in some economic situations. Noguchi (2001) extends the conditioning class optimality to a nonmyopic case and show that a nonmyopic version of conditional smooth fictitious play passes the nonmyopic optimality criterion for many strategies.

References

[1] Foster, D. and Vohra, R. (1993). "Calibrated Learning and Correlated Equilibrium," Mimeo, Wharton.

[2] Fudenberg, D. and Levine, D. (1995). "Universal Consistency and Cautious Fictitious Play," Journal of Economic Dynamics and Control 19, 1065-1089.

[3] Fudenberg, D. and Levine, D. (1999). "Conditional Universal Consistency," Games and Economic Behavior 29, 104-130.

[4] Hannan, J. (1959). "Approximation to Bayes Risk in Repeated Plays," in Contributions to the Theory of Games (M. Dresher, A.W. Tucker, and P. Wolfe, Eds.), Vol. 3, pp. 97-139, Princeton, NJ, Princeton Univ. Press.

[5] Kalai, E. and Lehrer, E. (1993). "Rational Learning Leads to Nash Equilibrium," Econometrica 61, 1019-1045.

[6] Nachbar, J. (1996). "Prediction, Optimization, and Learning in Repeated Games," Econometrica 65, 275-309

[7] Noguchi, Y. (1999). "Optimal Properties of Conditional Fictitious Play," Mimeo, Harvard University.

[8] Noguchi, Y. (2001). "Nonmyopic Optimality of Conditional Learning," Mimeo, Hitotsubashi University.