Merging with a Set of Probability Measures : A Characterization*

Yuichi Noguchi

Departmant of Economics

Kanto Gakuin University

1-50-1 Mutsu-ura-higashi, Kanazawa-ku

Yokohama 236-8501

Japan

E-mail: ynoguchi@kanto-gakuin.ac.jp

February 21, 2005

Abstract

We give a characterization of a *set* of probability measures with which a prior "weakly merges." For that purpose we introduce the concept of "conditioning rules" which represent *regularities* of probability measures, and then we define a probability measure "eventually generated" by a family of conditioning rules. Then we

*Very preliminary. Comment welcome.

show that a set of probability measures is learnable, i.e., all probability measures in the set are weakly merged with by a prior, if and only if the set is included in a set of probability measures eventually generated by a *countable* family of conditioning rules. We also demonstrate that quite similar results obtain for "almost weak merging." In addition we argue that our characterization is associated with the impossibility result in Nachbar (1997) and (2004).

1 Introduction

Bayesian learning is one of learning procedures that have been extensively studied in game theory. Especially, Kalai and Lehrer (1993) introduce a learning notion called *merging* to repeated games, and show that if every player's prior belief merges with the probability measure induced by the players' true strategies, then the path of play converges to a Nash equilibrium path. Merging requires that the updated forecast (or belief) about any future events be eventually accurate; the future events include *infinite* future ones such as tail events. However, when players discount future payoffs in a repeated game, the merging property is too enough for obtaining convergence to Nash equilibrium: any information about infinite future events is not useful for the discounting players. Accordingly, Kalai and Lehrer (1994) propose a weaker notion of merging called "weak merging." Weak merging means that the updated forecast about any *finite period* future event is eventually accurate. Furthermore, Lehrer and Smorodinsky (1997) and Sandroni (1998) argue that the weak merging property is sufficient to deal with learning to play Nash equilibrium. Since then, the literature has mainly focused on weak merging (Nachbar (1997) and (2004), Jackson, Kalai and Smorodinsky (1999), Foster and Young (2001), and so on).

In this paper we consider a characterization of a *set* of probability measures with which a prior (belief) *weakly merges*. This problem is especially important in the context of repeated games. In a repeated game players sequentially interact each other so as to be quite uncertain about their opponents strategies at the beginning of the game. Therefore, it is natural to start with assuming that each player does not know his opponents charateristics except that they play behavior strategies. In this situation the player would want to use a prior that weakly merges with *as many* opponents strategies *as possible*. Nonetheless it is not difficult to show that there is *no* prior that weakly merges with *all* opponents strategies. Then, a foundamental question is addressed: what is a *learnable set* of opponents strategies, i.e., a set of opponent strategies with which a prior could weakly merges? Characterizing a learnable set is certainly helpful in clarifying the possibilities of Bayesian learning in repeated games. For example, as Nachbar (1997) and (2004) show, some *diversity* property of learnable sets may be related to the impossibility of learning to play Nash equilibrium. Later we will argue that our characterization is associated with the generality of Nachbar's impossibility result.

To characterize a learnable set we introduce the concept of "conditioning rules" which represent *regularities* of probability measures; it is originally introduced by Noguchi (2000) to capture the regularity of a behavior strategy. For instance, a second order Markov measure has the regularity that current probabilities of states are always determined by the last two period realized states. In other words, the current probabilities are conditioned on the last two period states. A conditioning rule captures such a conditioning property of a probability measure, so that the regularity of any probability measure is arbitrarily approximated by some conditioning rule. Further, we define a probability measure "eventually generated" by a family (or a set) of conditioning rules. This means that the regularity of the probability measure is (arbitrarily) approximated by one in the family *from some period on.* As for the above Markov case, a measure eventually generated by a second order Markov conditioning rule means that current probabilities of states (with respect to the measure) are determined by the last two period states from some period on. The main conclusion of this paper is that a set of probability measures eventually generated by a *countable* family of conditioning rules plays a canonical role in characterizing a learnable set. More specifically, a set of probability measures is learnable, i.e., all measures in the set are weakly merged with by a prior, if and only if the set is included in a set of probability measures eventually generated by a countable family of conditioning rules. Put differently, Bayesian learning can eventually make accurate predictions against all probability measures eventually generated by a countable family of conditioning rules, but it cannot do so against more than those.

We provide two main results to obtain the conclusion. First, we show that *any* prior *cannot* weakly merges with more probability measures than those eventually generated by a countable family of conditioning rules. Therefore, *any* learnable set must be included in a set of measures eventually generated by a countable family of conditioning rules. Second and more importantly, we show that for *any* countable family of conditioning rules, there exists a prior such that the prior weakly merges with all probability measures in the set of those eventually generated by the family. This implies that if a set of probability measures is included in a set of those eventually generated by a countable family of conditioning rules, then the set is learnable. Therefore we conclude that a learnable set is characterized by a countable family of conditioning rules. Furthermore, we demonstrate that quite similar results obtain for "almost weak merging."

The point of this paper is how to form or construct a prior for our purpose. As Gilboa, Postlewaite, and Schmeidler (2004) point out, "Bayesian learning means nothing more than the updating of a given prior. It does not offer any theory, explanation, or insight into the process by which prior beliefs are formed." Indeed, there have not been given many ideas of constructing nontrivial priors. As such, we make use of an insight obtained from a study of another learning procedure called *conditional smooth fictitious play*: we construct a prior on the basis of *conditional empirical frequencies*. Specifically, our constructing prior is simply a slight modification of a belief formation process for conditional smooth fictitious play in Noguchi (2003). Although the process is based on a quite simple intuitive story of individual learning behavior, it is sufficiently powerful as to eventually make accurate predictions against as many probability measures as possible, as will be shown. Furthermore, in order to prove that the prior works, we use a different mathematical technique than those in previous work: theory of *large deviations* which gives precise probability evaluations about rare events.

Previous work has mainly explored conditions on relations between a prior and a (true) probability measure for the prior to (weakly) merge with the measure (Blackwell and Dubins (1962), Kalai and Lehrer (1993), Lehrer and Smorodinsky (1997), Sandroni (1998), and so on). The main reason is that if we interpret priors as players' prior beliefs and a probability measure as that induced by players' true strategies, then those conditions are also conditions for convergence to Nash equilibrium. Clearly those conditions are helpful in considering learnable sets. For example, the absolute continuity condition implies that any countably many measures are merged with by any prior that puts a weight on each of those measures. Nonetheless those conditions are not easy to check in many cases, so that we take a different approach to characterize a learnable set.

This paper is organized as follows. In Section 2 we describe the basic model and several concepts. In Section 3, which is the main part of this paper, we give a characterization of a set of probability measures with which a prior weakly merges. In addition, we also show that similar results obtain for almost weak merging. In Section 4 we apply our results to repeated games, and discuss its implications including one to Nachbar's impossibility result. In Section 5 we conclude with giving several remarks.

2 The Model and Concepts

2.1 The basic model and notations

Let S be a finite set. We write H_T for the $T-$fold Cartesian product of S: $H_T := \times_{t=1}^{t=T} S$. That is, H_T is the set of all finite histories with time length T. Let H denote the set of all finite histories including the null history h_0, i.e., $H := \bigcup_{t=0}^{\infty} H_t$, where $h_0 := \emptyset$ and $H_0 := \{\emptyset\}$. A finite history is denote by h. When we emphasize the time length of a finite history, we write h_T for a finite history up to time T: $h_T := (s_1, \cdots, s_T)$. Let $\mathbf{H}_\infty$ designate the set of all infinite histories: $\mathbf{H}_\infty := \times_{t=1}^{t=\infty} S$. An infinite history is denoted by $h_\infty := (s_1, s_2, \cdots)$. If a finite history h is an initial segment of a finite history h', then it is denoted by $h \leq h'$. When $h \leq h'$ and $h \neq h'$, it is designated by $h < h'$. Similarly, if h is an initial segment of an infinite history h_∞, then it is denoted by $h < h_\infty$. We assume the standard measurable structure on $\mathbf{H}_\infty$. Let $\mathcal{F}_T$ denote the minimum $\sigma-$algebra including all cylinder sets based on finite histories with time length T: $\mathcal{F}_T := \sigma(\{\mathbf{C}_h \mid h \in H_T\})$

where $\mathbf{C}_h := \{h_\infty \mid h < h_\infty\}$; $\mathcal{F}_T \subset \mathcal{F}_{T+1}$ for all T. $\{\mathcal{F}_T\}_T$ is the standard filtration. Let $\mathcal{F}$ be the minimum $\sigma-$algebra including all cylinder sets based on finite histories: $\mathcal{F} := \sigma(\bigcup_{t=1}^{\infty}\{\mathbf{C}_h \mid h \in H_t\})$. Let μ denote a probability measure on $(\mathbf{H}_\infty, \mathcal{F})$. Especially, when a probability measure is considered as a prior, it is denoted by $\tilde{\mu}$.

2.2 Weak merging

In this paper we mainly focus on weak merging. Weak merging requires that the updated forecast about any *finite period* future event be eventually accurate.

Definition 1 *A prior $\tilde{\mu}$ weakly merges with a probability measure μ if for all $k \geq 1$,*

$$\lim_{T\to\infty} \sup_{\mathbf{A}\in\mathcal{F}_{T+k}} \mid \tilde{\mu}(\mathbf{A} \mid \mathcal{F}_T) - \mu(\mathbf{A} \mid \mathcal{F}_T) \mid= 0,\ \mu - a.s.$$

Let $\mu(s \mid h_T)$ denote the probability of s at time $T+1$ conditional on a realized past history h_T up to time T. Then it is important to note that $\tilde{\mu}$ weakly merges with μ if and only if the *one period ahead* forecast is eventually correct (see Lehrer and Smorodinsky (1996a)): for $\mu-$almost all h_∞ and all $s \in S$

$$\lim_{T\to\infty} \mid \tilde{\mu}(s \mid h_T) - \mu(s \mid h_T) \mid= 0.$$

The purpose of this paper is to characterize a *set* of probability measures with which a prior weakly merges, so that we shall define merging with a set of probability measures.

Definition 2 *We say that a prior $\tilde{\mu}$ weakly merges with a set M of probability measures if $\tilde{\mu}$ weakly merges with all probability measures in M. Moreover, we say that M is weakly merged if there exists a prior $\tilde{\mu}$ such that $\tilde{\mu}$ weakly merges with M.*

2.3 Conditional probability systems

We make use of *conditional probability systems.* A conditional probability system (abbreviated to CPS) is a mapping from the set H of finite histories to the set $\Delta(S)$ of probability distributions over S.[1] It is denoted by $f : H \to \Delta(S)$. Then, it follows from Kolmogorov's extension theorem (see Shiryaev (1984)) that for all f there exists a *unique* probability measure μ_f such that $\mu_f(s \mid h) = f(h)[s]$ for all $s \in S$ and all $h \in H$. Conversely, it is easy to see that for all probability measures μ there exists a CPS f_μ such that $f_\mu(h)[s] = \mu(s \mid h)$ for all $s \in S$ and all $h \in H$.[2] The correspondence allows us to focus on conditional probability systems, instead of probability measures. Indeed, $\tilde{\mu}$ weakly merges with μ if and only if for $\mu-$almost all h_∞

$$\lim_{T\to\infty} \|f_{\tilde{\mu}}(h_T) - f_\mu(h_T)\| = 0$$

where $\|\cdot\|$ is the maximum norm: $\|x\| := \max_s \mid x[s] \mid$.

Remark 1 *If $\tilde{\mu}(A) = \tilde{\mu}'(A)$ for all $A \in \mathcal{F}$, then $\tilde{\mu}$ and $\tilde{\mu}'$ are identical as probability measures. However, it might be a case that $\tilde{\mu}(s \mid h) \neq \tilde{\mu}'(s \mid h)$ for some $h \in H$ with $\tilde{\mu}(\mathbf{C}_h)(= \tilde{\mu}'(\mathbf{C}_h)) = 0$ and some $s \in S$. In this paper, we implicitly assume that $\tilde{\mu}$ and $\tilde{\mu}'$ are different as priors in such a case.*

[1] $\Delta(S) := \{x \in \mathbb{R}^{\#S} \mid x[s] \geq 0$ for all $s \in S$ and $\sum_s x[s] = 1\}$, where $\#S$ denotes the cardinality of S, and $\mathbb{R}^{\#S}$ is the $\#S-$dimensional Euclidean space.

[2] A CPS f_μ corresponding to μ is not necessarily unique because if $\mu(h) = 0$, then $\mu(s \mid h)$ is arbitrarily taken. But $\mu(s \mid h)$ is uniquely determined for all $s \in S$ and all $h \in H$ with $\mu(h) > 0$. Thus, for any two f_μ and f'_μ corresponding to μ, $f_\mu(h)[s] = f'_\mu(h)[s]$ for all $s \in S$ and all $h \in H$ with $\mu(h) > 0$.

2.4 Conditioning rules

We introduce a key concept to characterize a learnable set: *conditioning rules*. A conditioning rule represents an (approximate) regularity of a conditional probability system or a probability measure. Formally, a conditioning rule is a *finite partition* of H, denoted by $\mathcal{P}$. An element of a conditioning rule $\mathcal{P}$ is called a *conditioning class* or simply a *class* in $\mathcal{P}$, denoted by β. Note that a class is considered as a *subset* of H because it is an element of a partition of H. In the following we often define a subset of H and call it a *class* by the abuse of language. For any CPS f, we may define its $\varepsilon-$*approximate* conditioning rule.

Definition 3 *A finite partition $\mathcal{P}_\varepsilon^f$ is called an $\varepsilon-$approximate conditioning rule of f, if for all $\beta \in \mathcal{P}_\varepsilon^f$ and all $h, h' \in \beta$, $\|f(h) - f(h')\| < \varepsilon$.*

The definition says that probability distributions (on S) after finite histories in each class β are almost the same. Note that any CPS f has its $\varepsilon-$approximate conditioning rule for all $\varepsilon > 0$.[3] For example, let $S = \{L, R\}$, and let f be a first-order Markov system such that $f(h_t) = (\frac{1}{3}, \frac{2}{3})$ when $s_t = L$, and $f(h_t) = (\frac{2}{3}, \frac{1}{3})$ when $s_t = R$. Then, let $\mathcal{P}^f := \{\beta_L, \beta_R\}$ where $\beta_L := \{h_t \in H \mid s_t = L\}$ and $\beta_R := \{h_t \in H \mid s_t = R\}$. $\mathcal{P}^f$ is an $\varepsilon-$approximate conditioning rule of f for all $\varepsilon > 0$.

[3] By the compactness of $\Delta(S)$, for all $\varepsilon > 0$ we may take a *finite* family $\{\Delta_j\}_{j=1}^m$ of subsets in $\Delta(S)$ such that (1) $\{\Delta_j\}_j$ covers $\Delta(S)$, i.e., $\bigcup_{j=1}^m \Delta_j = \Delta(S)$ and (2) those diameters are no more than ε, that is, $\sup_{\pi,\pi' \in \Delta_j} \|\pi - \pi'\| < \varepsilon$ for all j. Thus, for all CPS f and all $\varepsilon > 0$, an $\varepsilon-$approximate conditioning rule $\mathcal{P}_\varepsilon^f$ of f may be defined by the following equivalence relation on H:

$$h \sim_{\mathcal{P}_\varepsilon^f} h' \overset{def.}{\Leftrightarrow} \text{there exists } j \text{ such that } f(h), f(h') \in \Delta_j \text{ and } f(h), f(h') \notin \Delta_k \text{ for all } k < j.$$

Conversely, we may generate conditional probability systems from conditioning rules.

Definition 4 *We say that a CPS $f : H \to \Delta(S)$ is generated by a family Φ of conditioning rules, if for all $\varepsilon > 0$ there exists $\mathcal{P} \in \Phi$ such that $\mathcal{P}$ is an $\varepsilon-$approximate conditioning rule of f: for all $\beta \in \mathcal{P}$ and all $h, h' \in \beta$, $\|f(h) - f(h')\| < \varepsilon$.*

The definition says that for all $\varepsilon > 0$, the regularity of f is $\varepsilon-$approximated by a conditioning rule in Φ. For instance, let $S = \{L, R\}$ and $\mathcal{P}_1 = \{\beta_L, \beta_R\}$ where $\beta_L = \{h_t \in H \mid s_t = L\}$ and $\beta_R = \{h_t \in H \mid s_t = R\}$. Furthermore, let $\mathcal{P}_2 = \{\beta_e, \beta_o\}$ where $\beta_e = \{h_t \in H \mid t \text{ is odd}\}$ and $\beta_o = \{h_t \in H \mid t \text{ is even}\}$. Then, $f : H \to \Delta(S)$ is generated by $\{\mathcal{P}_1, \mathcal{P}_2\}$ if and only if either there exist $0 \leq p, q \leq 1$ such that for all $h \in \beta_L$, $f(h) = (p, 1-p)$, and for all $h \in \beta_R$, $f(h) = (q, 1-q)$, or there exist $0 \leq p', q' \leq 1$ such that for all $h \in \beta_e$, $f(h) = (p', 1-p')$, and for all $h \in \beta_o$, $f(h) = (q', 1-q')$. Note that *all* i.i.d. CPS's are generated by *any* (non-empty) family of conditioning rules.[4] Note also that *any* CPS f is generated by a *countable* family $\{\mathcal{P}^f_{\frac{1}{n}}\}_n$ of its $\frac{1}{n}-$approximate conditioning rules.

Similarly, we may define (approximate) conditioning rules of a probability measure μ.

Definition 5 *A finite partition $\mathcal{P}^\mu_\varepsilon$ is called an $\varepsilon-$approximate conditioning rule of μ if there exists a CPS f_μ corresponding to μ such that $\mathcal{P}^\mu_\varepsilon$ is an $\varepsilon-$approximate conditioning rule of f_μ.*

Definition 6 *We say that a probability measure μ is generated by a family Φ of conditioning rules if there exists a CPS f_μ corresponding to μ such that f_μ is generated by Φ. The set of all probability measures generated by Φ is denoted by $G(\Phi)$.*

[4] We say that a CPS f is i.i.d. if $f(h) = f(h')$ for all $h, h' \in H$.

As in the CPS case, all i.i.d. probability measures are generated by any (non-empty) family of conditioning rules. Also, any probability measure μ is generated by a *countable* family of conditioning rules.

Remark 2 *A useful fact is that any countable union of countable families is countable. For example, any countably many measures* $\{\mu_m\}_m$ *are generated by a countable family of conditioning rules because each* μ_m *is generated by a countable family* $\{\mathcal{P}_{\frac{1}{n}}^{f_{\mu_m}}\}_n$.

3 Characterization of a Learnable Set

3.1 Bounds of weak merging

In order to characterize a learnable set of probability measures, we need to slightly extend the generation of probability measures by conditioning rules. Let $\{\mathcal{P}_i\}_i$ be a *countable* family of conditioning rules.

Definition 7 *We say that a CPS* $f : H \to \Delta(S)$ *is eventually generated by* $\{\mathcal{P}_i\}_i$ *if for all* $\varepsilon > 0$ *there exist an index* i_0, *a* μ_f-*probability one set* $\mathbf{Z}_0$, *and a time function*[5] $T_0 : \mathbf{Z}_0 \to \mathbb{N}$ *such that for all* $\beta \in \mathcal{P}_{i_0}$ *and all* $h_T, h'_{T'} \in \beta$, *if there exist* $h_\infty, h'_\infty \in \mathbf{Z}_0$ *such that* $h_T < h_\infty$ *and* $T \geq T_0(h_\infty)$ *and* $h'_{T'} < h'_\infty$ *and* $T' \geq T_0(h'_\infty)$, *then* $\|f(h_T) - f(h'_{T'})\| < \varepsilon$.

The definition is rather complicated, but it simply says that for any $\varepsilon > 0$, the regularity of f is (almost surely) $\varepsilon-$approximated by one of conditioning rules $\{\mathcal{P}_i\}_i$ *from some period on.* Clearly any CPS generated by $\{\mathcal{P}_i\}_i$ is eventually generated by $\{\mathcal{P}_i\}_i$. Analogously, we may define the eventual generation of probability measures.

[5] $\mathbb{N}$ is the set of all natural numbers.

Definition 8 *We say that a probability measure μ is eventually generated by $\{\mathcal{P}_i\}_i$ if there exists a CPS f_μ corresponding to μ such that f_μ is eventually generated by $\{\mathcal{P}_i\}_i$. The set of all probability measures eventually generated by $\{\mathcal{P}_i\}_i$ is denoted by $EG(\{\mathcal{P}_i\}_i)$.*

As in the CPS case, any probability measure generated by $\{\mathcal{P}_i\}_i$ is eventually generated by $\{\mathcal{P}_i\}_i$: $G(\{\mathcal{P}_i\}_i) \subset EG(\{\mathcal{P}_i\}_i)$. More precisely it can be shown that $EG(\{\mathcal{P}_i\}_i)$ is strictly larger than $G(\{\mathcal{P}_i\}_i)$, i.e., $G(\{\mathcal{P}_i\}_i) \subsetneqq EG(\{\mathcal{P}_i\}_i)$.

Our first result is that the weak merging property is always bounded by a countable family of conditioning rules. In other words, any prior *cannot* merge with more probability measures than those eventually generated by a *countable* family of conditioning rules.

Proposition 1 *For any prior $\tilde{\mu}$, there exists a countable family $\{\mathcal{P}_i\}_i$ of conditioning rules such that $\tilde{\mu}$ never weakly merges with any $\mu \notin EG(\{\mathcal{P}_i\}_i)$.*

Proof. Fix any prior $\tilde{\mu}$. Let $f_{\tilde{\mu}}$ be a CPS corresponding to $\tilde{\mu}$. As noted in Subsection 2.4, for each n, we may take a $\frac{1}{n}-$approximate conditioning rule $\mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}$ of $f_{\tilde{\mu}}$. We shall show that $\tilde{\mu}$ never weakly merges with any $\mu \notin EG(\{\mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}\}_n)$. Take any probability measure $\mu \notin EG(\{\mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}\}_n)$. Then, there exists $\varepsilon_0 > 0$ such that for all n, all $\mu-$probability one sets $\mathbf{Z}$, and all time functions $T : \mathbf{Z} \to \mathbb{N}$, there exist $\beta \in \mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}$ and $h_T, h'_{T'} \in \beta$ such that $h_T < h_\infty$, $T \geq T(h_\infty)$, $h'_{T'} < h'_\infty$, and $T' \geq T(h'_\infty)$ for some $h_\infty, h'_\infty \in \mathbf{Z}$, and $\|f_\mu(h_T) - f_\mu(h'_{T'})\| \geq \varepsilon_0$.

Suppose that $\tilde{\mu}$ weakly merges with μ. Then, there exists a $\mu-$probability one set $\mathbf{Z}_0$ such that for all $h_\infty \in \mathbf{Z}_0$ there exists $T_0(h_\infty)$ such that for all $T \geq T_0(h_\infty)$, $\|f_{\tilde{\mu}}(h_T) - f_\mu(h_T)\| < \frac{\varepsilon_0}{4}$. On the other hand, letting $n_0 \geq \frac{4}{\varepsilon_0}$, it follows from the previous paragraph that for n_0, $\mathbf{Z}_0$ and $T_0 : \mathbf{Z}_0 \to \mathbb{N}$, there exist $\beta \in \mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n_0}}$ and $h_T, h'_{T'} \in \beta$ such that $h_T < h_\infty$,

$T \geq T_0(h_\infty)$, $h'_{T'} < h'_\infty$, and $T' \geq T_0(h'_\infty)$ for some $h_\infty, h'_\infty \in \mathbf{Z}_0$, and $\|f_\mu(h_T) - f_\mu(h'_{T'})\| \geq \varepsilon_0$. These deduce that $\|f_{\tilde{\mu}}(h_T) - f_{\tilde{\mu}}(h'_{T'})\| \geq \frac{\varepsilon_0}{2}$ for $h_T, h'_{T'} \in \beta$. But then, since $\beta \in \mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n_0}}$ and $h_T, h'_{T'} \in \beta$, $\|f_{\tilde{\mu}}(h_T) - f_{\tilde{\mu}}(h'_{T'})\| < \frac{1}{n_0} \leq \frac{\varepsilon_0}{4}$. This is a contradiction. Thus, $\tilde{\mu}$ does not weakly merge with μ. ■

3.2 Frequency-based prior

In this subsection we show a more important result: for *any* countable family of conditioning rules, we construct a prior that weakly merges with all probability measures that are eventually generated by the family. This, together with our first result (Proposition 1), implies that merging with a set of probability measures is characterized by a countable family of conditioning rules.

To obtain a desired prior we use the method of constructing a belief formation process for conditional smooth fictitious play in Noguchi (2003): a prior is defined on the basis of *conditional empirical frequencies*; such a prior will be called *frequency-based*.

Proposition 2 *For any countable family $\{\mathcal{P}_i\}_i$ of conditioning rules, there exists a frequency-based prior $\tilde{\mu}_F$ such that $\tilde{\mu}_F$ weakly merges with all $\mu \in EG(\{\mathcal{P}_i\}_i)$.*

In order to prove Proposition 2, we prepare a mathematical lemma. Recall that a subset of H is also called a class, denoted by β. If a realized history $h_{T-1} \in \beta$, then we say that β is *active at time T* or time T is a *$\beta-$active period*. Let $\mathcal{T}^\beta_n(h_\infty)$ denote the calendar time of the nth $\beta-$active period in h_∞; $\mathcal{T}^\beta_n(h_\infty) < \infty$ means that β is active at least n times in h_∞. Let $\mathbf{d}^\beta_n(h_\infty)$ designate a vector in which each coordinate $\mathbf{d}^\beta_n(h_\infty)[s]$ is the number of times that s has occurred in the first n $\beta-$active periods along h_∞.

The next lemma extends a basic fact of *large deviations* to a conditional case. The lemma says that if the probabilities of a state s have common upper and lower bounds in active periods of a given class, then the probability that the frequency of s occurring in the first n active periods of that class is *not* between those bounds decreases *exponentially* in the sample size n.

Lemma 1 *Let β be any subset of H. Suppose that a probability measure μ and a state s satisfy the following assumption:*

$$\text{For all } h \in \beta,\ l \leq \mu(s \mid h) \leq L$$

where l and L are nonnegative numbers. Then, for all $\varepsilon > 0$ and all $n = 1, 2, \cdots,$

$$\mu(\mathcal{T}_n^\beta < \infty, \frac{\mathbf{d}_n^\beta[s]}{n} \leq l - \varepsilon \text{ or } \frac{\mathbf{d}_n^\beta[s]}{n} \geq L + \varepsilon) \leq 2\exp(-2n\varepsilon^2)$$

where $\mathbf{d}_n^\beta[s]$ is the $s-$coordinate of $\mathbf{d}_n^\beta(h_\infty)$.

Proof. See Noguchi (2003). ■

Without loss of generality, we may assume that $\{\mathcal{P}_i\}_i$ is ordered in fineness: $\mathcal{P}_i \leq \mathcal{P}_{i+1}$ for all i.[6] In order to construct a frequency-based prior $\tilde{\mu}_F$, let us start with determining the *prior sample size* n_0^β for each class $\beta \in \bigcup_i \mathcal{P}_i$ on the basis of Lemma 1, where $\bigcup_i \mathcal{P}_i$ is the set of all classes in $\{\mathcal{P}_i\}_i$. Suppose $\beta \in \mathcal{P}_i$. Then, choose any positive integer n_0^β such that

$$n_0^\beta \geq \frac{i^2}{2}(i - \log(\frac{1 - \exp(-2i^{-2})}{\#\mathcal{P}_i})).$$

[6]Partition $\mathcal{P}$ is finer than partition $\mathcal{Q}$ if for all $\beta \in \mathcal{P}$ there exists $\alpha \in \mathcal{Q}$ such that $\beta \subset \alpha$. It is denoted by $\mathcal{Q} \leq \mathcal{P}$. We say that class β is finer than class α, or α is coarser than β, if $\beta \subset \alpha$.

This inequality is equivalent to $\#\mathcal{P}_i \cdot \sum_{n=n_0^\beta}^{\infty} \exp(-2ni^{-2}) \leq \exp(-i)$. Taking larger n_0^β means collecting more prior samples for class β, so that we may make the probability of wrong prediction *exponentially* smaller.

Next we introduce functions $i : H \to \mathbb{N}$ and $\beta : H \to \bigcup_i \mathcal{P}_i$. Intuitively, $i(\cdot)$ and $\beta(\cdot)$ describe a forecaster's selection of classes in $\{\mathcal{P}_i\}_i$: when h_T is a realized past history, the forecaster uses $\beta(h_T) \in \mathcal{P}_{i(h_T)}$ as a *category* at time $T+1$. Then, the forecaster collects prior samples for $\beta(h_T)$ and observed samples in the *past* periods of using $\beta(h_T)$ as a category, obtains the empirical distribution of those samples, and uses the distribution as the forecast at time $T+1$; having obtained enough prior samples for class $\beta(h_T)$ enables the forecaster to make accurate predictions from the *first* period (to the last period) of using $\beta(h_T)$ as a category. In the following we shall define $i(\cdot)$ and $\beta(\cdot)$, categories, prior samples, and then prior $\tilde{\mu}_F$.

To give formal definitions of $i(\cdot)$ and $\beta(\cdot)$, we also define three other functions $m : H \to \mathbb{N}$, $n : H \to \mathbb{N}$, and $\alpha : H \to \bigcup_i \mathcal{P}_i$. Roughly speaking, $m(h_T)$ is a maximum index of past used conditioning classes that are equal to or coarser than $\beta(h_T)$. Let $n(h_T)$ be the number of times that a currently employed class has been used as a category and $\alpha(h_T)$ be a finer class that will be employed next. We recursively define $i(\cdot)$, $\beta(\cdot)$, $m(\cdot)$, $n(\cdot)$ and $\alpha(\cdot)$ as follows:

- $i(h_0) := 1$ and $\beta(h_0) := \beta$, where $h_0 \in \beta$ and $\beta \in \mathcal{P}_1$. Further, let $m(h_0) := 1$, $n(h_0) := 0$, and $\alpha(h_0) := \beta$, where $h_0 \in \beta$ and $\beta \in \mathcal{P}_{m(h_0)+1}$.

• Suppose that $i(h_t)$, $\beta(h_t)$, $m(h_t)$, $n(h_t)$, and $\alpha(h_t)$ are defined for $0 \leq t \leq T-1$. Let

$$m(h_T) := \max\{i(h_t) \mid h_t < h_T, h_T \in \beta(h_t)\},$$

$$\alpha(h_T) := \beta, \text{ where } h_T \in \beta \text{ and } \beta \in \mathcal{P}_{m(h_T)+1},$$

$$n(h_T) := \#\{h_t \mid h_t < h_T, i(h_t) = m(h_T), \alpha(h_T) \subset \beta(h_t)\},$$

where if $\{i(h_t) \mid h_t < h_T, h_T \in \beta(h_t)\} = \emptyset$, then let $m(h_T) = 1$. Then, define $i(h_T)$ and $\beta(h_T)$ as follows:

$$i(h_T) := \begin{cases} m(h_T) + 1, \text{ if } n(h_T) \geq n_0^{\alpha(h_T)} \\ m(h_T), \qquad \text{otherwise} \end{cases}$$

where $n_0^{\alpha(h_T)}$ is the prior sample size for conditioning class $\alpha(h_T)$. Finally let $\beta(h_T) := \beta$, where $h_T \in \beta$ and $\beta \in \mathcal{P}_{i(h_T)}$.

The inequality in the definition of $i(h_T)$ is a switching criterion such that if $n_0^{\alpha(h_T)}$ samples are obtained as prior samples for $\alpha(h_T)$, then a forecaster switches to a finer class $\alpha(h_T)$; otherwise, he keeps employing a used class.

A category is represented by a pair of index and class, i.e., (i, β), where $\beta \in \mathcal{P}_i$. Given a finite history h_{T-1}, we say that time T is an *effective period of category* (i, β) or category (i, β) is *effective at time* T if $i(h_{T-1}) = i$ and $\beta(h_{T-1}) = \beta$. Note that given any h_∞, each period has exactly one effective category. Note also that any category (i, β) with $i \geq 2$ has its (unique) *predecessor* (i_p, β_p) such that $i_p = i - 1$ and $\beta \subset \beta_p \in \mathcal{P}_{i_p}$; by the definitions of $i(\cdot)$ and $\beta(\cdot)$, category (i, β) can be effective only after (i_p, β_p) has been effective n_0^β times.

Next we define the *prior samples* $d_0^{(i,\beta)}$ and the *prior sample size* $n_0^{(i,\beta)}$ for each category (i, β). For all $(1, \beta)$, we may take $d_0^{(1,\beta)}$ as an arbitrary $\#S-$dimensional vector with positive integer components and its sum being n_0^β: $d_0^{(1,\beta)}[s]$ is a positive integer for all $s \in S$ and $n_0^{(1,\beta)} := \sum_s d_0^{(1,\beta)}[s] = n_0^\beta$. For all categories (i,β) with $i \geq 2$, let $d_0^{(i,\beta)}$ consist of observed samples in the *first* n_0^β effective periods of its predecessor (i_p, β_p): each component $d_0^{(i,\beta)}[s]$ is the number of times that s has occurred in the first n_0^β effective periods of (i_p, β_p). Thus $d_0^{(i,\beta)}$ is *history-dependent* for $i \geq 2$. Let $n_0^{(i,\beta)} := \sum_s d_0^{(i,\beta)}[s] = n_0^\beta$.

Finally we define a prior $\tilde{\mu}_F$. Given a realized past history h_T up to the last date, a category, say (i,β), is effective at time $T+1$: $i = i(h_T)$ and $\beta = \beta(h_T)$. Then we obtain observed states in the past effective periods of category (i,β) which is represented by a vector $d_T^{(i,\beta)}$: each component $d_T^{(i,\beta)}[s]$ is the number of times that s has occurred in the past effective periods of (i,β). Let $n_T^{(i,\beta)}$ denote the sample size for (i,β) up to time T: $n_T^{(i,\beta)} := \sum_s d_T^{(i,\beta)}[s]$. We define the conditional empirical distribution $D_T^{(i,\beta)}$ on category (i,β) up to time T: $D_T^{(i,\beta)} := d_T^{(i,\beta)} + d_0^{(i,\beta)} \diagup\, n_T^{(i,\beta)} + n_0^{(i,\beta)}$. Then we use $D_T^{(i,\beta)}$ as the forecast at time $T+1$. Accordingly we define frequency-based CPS f_F as follows: for all $h_T \in H$,

$$f_F(h_T) := D_T^{(i,\beta)},$$

where $i = i(h_T)$ and $\beta = \beta(h_T)$. Then let $\tilde{\mu}_F := \mu_{f_F}$.

The following lemma will be used to prove Proposition 2. It insists that a forecaster uses finer and finer classes as categories, as time proceeds.

Lemma 2 *For all h_∞, $\lim_{T\to\infty} i(h_T) = \infty$.*

Proof. See Appendix A. ■

We are in a position to prove Proposition 2.

Proof of Proposition 2: Without loss of generality we may assume that $\mathcal{P}_i \leq \mathcal{P}_{i+1}$ for all i. Let f_F be a frequency-based CPS for $\{\mathcal{P}_i\}_i$. Fix any $\mu \in EG(\{\mathcal{P}_i\}_i)$, and let f_μ be a CPS corresponding to μ. Then, it suffices to show that for $\mu-$almost all h_∞, $\|f_F(h_T) - f_\mu(h_T)\| \to 0$ as $T \to \infty$:

$$\mu(\bigcap_{n=1}^{\infty} \bigcup_{T_0=1}^{\infty} \bigcap_{T \geq T_0} \{h_\infty \mid \|f_F(h_T) - f_\mu(h_T)\| < \frac{1}{n}\}) = 1.$$

Equivalently, we have only to show that for all n,

$$\mu(\bigcap_{T_0=1}^{\infty} \bigcup_{T \geq T_0} \{h_\infty \mid \|f_F(h_T) - f_\mu(h_T)\| \geq \frac{1}{n}\}) = 0.$$

Since $\mu \in EG(\{\mathcal{P}_i\}_i)$, we obtain a $\mu-$probability one set $\mathbf{Z}_0$, i.e., $\mu(\mathbf{Z}_0) = 1$ such that for any $\varepsilon > 0$ there exist i_0 and $T_0 : \mathbf{Z}_0 \to \mathbb{N}$ such that for all $\beta \in \mathcal{P}_{i_0}$ and all $h_T, h'_{T'} \in \beta$, if there exist $h_\infty, h'_\infty \in \mathbf{Z}_0$ such that $h_T < h_\infty$, $h'_{T'} < h'_\infty$, $T \geq T_0(h_\infty)$ and $T' \geq T_0(h'_\infty)$, then $\|f_\mu(h_T) - f_\mu(h'_{T'})\| < \varepsilon$.

Step 1: Let $\varepsilon := \frac{1}{3n}$. Since $\mathcal{P}_i \leq \mathcal{P}_{i+1}$ for all i, we may take $i_0 \geq 3n$. For all $\beta \in \mathcal{P}_{i_0}$, define a class $\hat{\beta}$ as follows: $h_T \in \hat{\beta}$ if and only if $h_T \in \beta$ and $h_T < h_\infty$ and $T \geq T_0(h_\infty)$ for some $h_\infty \in \mathbf{Z}_0$. Then, for all $\beta \in \mathcal{P}_{i_0}$, let $L^\beta[s] := \sup_{h \in \hat{\beta}} f_\mu(h)[s]$ and $l^\beta[s] := \inf_{h \in \hat{\beta}} f_\mu(h)[s]$; note that $L^\beta[s] - l^\beta[s] \leq \varepsilon$ for all s. Furthermore, for all categories (i, β), we define a class $\gamma(i, \beta)$ as follows: $h_T \in \gamma(i, \beta)$ if and only if $(i(h_T), \beta(h_T)) = (i, \beta)$ or (i_p, β_p) and $h_T < h_\infty$ and $T \geq T_0(h_\infty)$ for some $h_\infty \in \mathbf{Z}_0$. Since $\mathcal{P}_i \leq \mathcal{P}_{i+1}$ for all i, for each category (i, β) with $i \geq i_0 + 1$, there exists a unique class $\alpha \in \mathcal{P}_{i_0}$ such that $\gamma(i, \beta) \subset \hat{\alpha}$; then, let $L^{(i,\beta)}(s) := L^\alpha[s]$ and $l^{(i,\beta)}[s] := l^\alpha[s]$ for all s. Moreover, it follows from the definition of f_F and Lemma 2 that for all $h_\infty \in \mathbf{Z}_0$ and all $i \geq i_0 + 1$, there exists T_0 such that for all $T \geq T_0$, $f_F(h_T) = D_T^{(j,\beta)} = \mathbf{d}_n^{\gamma(j,\beta)}(h_\infty)/n$ for some (j, β) with $j \geq i$ and some $n \geq n_0^\beta$.

Step 2: For all categories (i, β) with $i \geq i_0 + 1$, let

$$\mathbf{B}_n^{(i,\beta)} := \{h_\infty \mid \mathcal{T}_n^{\gamma(i,\beta)} < \infty, \exists s \in S(\frac{\mathbf{d}_n^{\gamma(i,\beta)}[s]}{n} \geq L^{(i,\beta)}[s] + \frac{1}{i} \text{ or } \frac{\mathbf{d}_n^{\gamma(i,\beta)}[s]}{n} \leq l^{(i,\beta)}[s] - \frac{1}{i})\}.$$

Then, from Step 1 it follows that for all $j \geq i_0 + 1$,

$$\bigcap_{T_0=1}^{\infty} \bigcup_{T \geq T_0} \{h_\infty \mid \|f_F(h_T) - f_\mu(h_T)\| \geq \frac{1}{n}\} \bigcap \mathbf{Z}_0 \subset \bigcup_{i \geq j} \bigcup_{\beta \in \mathcal{P}_i} \bigcup_{n \geq n_0^\beta} \mathbf{B}_n^{(i,\beta)}.$$

Step 3: By Step 1 and Lemma 1, $\mu(\mathbf{B}_n^{(i,\beta)}) \leq 2(\#S)\exp(-2ni^{-2})$. Also, by the definition of n_0^β's, $\#\mathcal{P}_i \cdot \sum_{n=n_0^\beta}^{\infty} \exp(-2ni^{-2}) \leq \exp(-i)$ for all $\beta \in \mathcal{P}_i$ and all i. These imply that

$$\begin{aligned}
\mu(\bigcup_{i \geq j} \bigcup_{\beta \in \mathcal{P}_i} \bigcup_{n \geq n_0^\beta} \mathbf{B}_n^{(i,\beta)}) &\leq \sum_{i \geq j} \sum_{\beta \in \mathcal{P}_i} \sum_{n \geq n_0^\beta} 2(\#S)\exp(-2ni^{-2}) \\
&\leq 2(\#S) \sum_{i \geq j} \#\mathcal{P}_i \sum_{n \geq n_0^\beta} \exp(-2ni^{-2}) \\
&\leq 2(\#S) \sum_{i \geq j} \exp(-i) \\
&\leq 2(\#S)(1 - \exp(-1))^{-1} \exp(-j).
\end{aligned}$$

From this inequality and the set inclusion in Step 2 it follows that for all $j \geq i_0 + 1$

$$\mu(\bigcap_{T_0=1}^{\infty} \bigcup_{T \geq T_0} \{h_\infty \mid \|f_F(h_T) - f_\mu(h_T)\| \geq \frac{1}{n}\} \bigcap \mathbf{Z}_0) \leq 2(\#S)(1 - \exp(-1))^{-1} \exp(-j).$$

Thus, letting $j \to \infty$, we have

$$\mu(\bigcap_{T_0=1}^{\infty} \bigcup_{T \geq T_0} \{h_\infty \mid \|f_F(h_T) - f_\mu(h_T)\| \geq \frac{1}{n}\} \bigcap \mathbf{Z}_0) = 0.$$

Since the complement of $\mathbf{Z}_0$ is of probability zero, i.e., $\mu(\mathbf{Z}_0^c) = 0$, the above equality implies the desired result:

$$\mu(\bigcap_{T_0=1}^{\infty} \bigcup_{T \geq T_0} \{h_\infty \mid \|f_F(h_T) - f_\mu(h_T)\| \geq \frac{1}{n}\}) = 0.$$

■

Propositions 1 and 2 induce a characterization of merging with a set of probability measures by a countable family of conditioning rules, which is the main purpose of this paper.

Theorem 1 *A set M of probability measures is weakly merged if and only if all probability measures in M are eventually generated by some countable family $\{\mathcal{P}_i\}_i$ of conditioning rules, that is, there exists a countable family $\{\mathcal{P}_i\}_i$ of conditioning rules such that $M \subset EG(\{\mathcal{P}_i\}_i)$.*

Proof. Suppose that a set M of probability measures is weakly merged. Then, there exists a prior $\tilde{\mu}$ such that $\tilde{\mu}$ weakly merges with all μ in M. But then, Proposition 1 states that there exists a countable family $\{\mathcal{P}_i\}_i$ of conditioning rules such that $\tilde{\mu}$ never weakly merges with any $\mu \notin EG(\{\mathcal{P}_i\}_i)$. Therefore, $M \subset EG(\{\mathcal{P}_i\}_i)$.

Conversely, suppose that there exists $\{\mathcal{P}_i\}_i$ such that $M \subset EG(\{\mathcal{P}_i\}_i)$. Then, Proposition 2 states that there exists a prior $\tilde{\mu}_F$ such that $\tilde{\mu}_F$ weakly merges with all $\mu \in EG(\{\mathcal{P}_i\}_i)$. Thus, $\tilde{\mu}_F$ weakly merges with all $\mu \in M$ because $M \subset EG(\{\mathcal{P}_i\}_i)$. Therefore, M is weakly merged. ■

3.3 Almost weak merging

Lehrer and Smorodinsky (1996b) introduce a weaker notion of merging: *almost weak merging*. Almost weak merging says that the updated forecast about any finite period future event is accurate *in almost all periods*. Let us call a set D of positive integers a *dense* sequence of periods if $\lim_{T\to\infty} \#(D \bigcap \{1, \cdots, T\}) \diagup T = 1$, where $\#$ denotes the cardinality of a set.

Definition 9 *A prior $\tilde{\mu}$ almost weakly merges with a probability measure μ if for all $\varepsilon > 0$, all $k \geq 1$, and $\mu-$almost all h_∞ there exists a dense sequence D of periods such that for all $T \in D$*

$$\sup_{\mathbf{A} \in \mathcal{F}_{T+k}} \mid \tilde{\mu}(\mathbf{A} \mid \mathcal{F}_T) - \mu(\mathbf{A} \mid \mathcal{F}_T) \mid < \varepsilon.$$

Notice that a prior $\tilde{\mu}$ almost weakly merges with a probability measure μ if and only if for all $\varepsilon > 0$ and $\mu-$almost all h_∞ there exists a dense sequence D of periods such that for all $T \in D$, $\|f_{\tilde{\mu}}(h_T) - f_\mu(h_T)\| < \varepsilon$. It is also equivalent to the following: for all $\varepsilon > 0$ there exist a $\mu-$probability one set $\mathbf{Z}$ and a family $\{T_m\}_m$ of time functions, i.e., $T_m : \mathbf{Z} \to \mathbb{N}$, such that (1) for all $h_\infty \in \mathbf{Z}$ and all m, $\|f_{\tilde{\mu}}(h_{T_m}) - f_\mu(h_{T_m})\| < \varepsilon$, and (2) $\lim_{T\to\infty} \frac{N_T(h_\infty)}{T} = 1$ for all $h_\infty \in \mathbf{Z}$, where $N_T(h_\infty) := \{m \mid T_m(h_\infty) \leq T\}$.

Accordingly, we define the *almost* generation of a CPS by a family of conditioning rules.

Definition 10 *We say that a CPS $f : H \to \Delta(S)$ is almost generated by $\{\mathcal{P}_i\}_i$ if for all $\varepsilon > 0$ there exist an index i_0, a μ_f-probability one set $\mathbf{Z}_0$, and a family $\{T_m\}_m$ of time functions such that* (1) *for all $\beta \in \mathcal{P}_{i_0}$ and all $h_T, h'_{T'} \in \beta$, if there exist $h_\infty, h'_\infty \in \mathbf{Z}_0$ and m, m' such that $h_T < h_\infty$ and $T = T_m(h_\infty)$ and $h'_{T'} < h'_\infty$ and $T' = T_{m'}(h'_\infty)$, then $\|f(h_T) - f(h'_{T'})\| < \varepsilon$, and* (2) $\lim_{T\to\infty} \frac{N_T(h_\infty)}{T} = 1$ *for all $h_\infty \in \mathbf{Z}_0$, where $N_T(h_\infty) := \max\{m \mid T_m(h_\infty) \leq T\}$.*

The definition says that for any $\varepsilon > 0$, the regularity of f is (almost surely) $\varepsilon-$approximated by one of the conditioning rules *in almost all periods*. We shall define the almost generation of probability measures.

Definition 11 *We say that a probability measure μ is almost generated by $\{\mathcal{P}_i\}_i$ if there exists a CPS f_μ corresponding to μ such that f_μ is almost generated by $\{\mathcal{P}_i\}_i$. The set of all probability measures almost generated by $\{\mathcal{P}_i\}_i$ is denoted by $AG(\{\mathcal{P}_i\}_i)$.*

Obviously $EG(\{\mathcal{P}_i\}_i) \subsetneqq AG(\{\mathcal{P}_i\}_i)$. We show that any prior cannot almost weakly merge with more probability measures than those that are almost generated by a countable family of conditioning rules.

Proposition 3 *For any prior $\tilde{\mu}$, there exists a countable family $\{\mathcal{P}_i\}_i$ of conditioning rules such that $\tilde{\mu}$ never almost weakly merges with any $\mu \notin AG(\{\mathcal{P}_i\}_i)$.*

Proof. See Appendix B. ■

For any countable family of conditioning rules, we construct a prior $\tilde{\mu}'_F$ that almost weakly merges with all probability measures almost generated by the family. For that purpose, we have only to modify the switching criterion in the definition of $i(\cdot)$ as follows:

$$i(h_T) := \begin{cases} m(h_T)+1, \text{ if } n(h_T) \geq n_0^{\alpha(h_T)} \text{ and } \frac{\sum_{i=1}^{M(h_T)+1} \sum_{\beta \in \mathcal{P}_i} n_0^\beta}{n(h_T)} < \frac{1}{m(h_T)} \\ m(h_T), \qquad \text{otherwise} \end{cases}$$

All other things are exactly the same as in the weak merging case. In the definition of $i(\cdot)$ the switching criterion consists of two inequalities: the first one is the same as in the weak merging case, and the second inequality is added so that for almost all categories, the prior sample size is negligible relative to the number of effective periods; this fact is useful for proving Proposition 4.

Proposition 4 *For any countable family $\{\mathcal{P}_i\}_i$ of conditioning rules, there exists a frequency-based prior $\tilde{\mu}'_F$ such that $\tilde{\mu}'_F$ almost weakly merges with all $\mu \in AG(\{\mathcal{P}_i\}_i)$.*

Proof. See Appendix B. ■

Notice that $\tilde{\mu}'_F$ also weakly merges with $EG(\{\mathcal{P}_i\}_i)$; the proof is quite similar to that of Proposition 2. Therefore, $\tilde{\mu}'_F$ *not only* almost weakly merges with $AG(\{\mathcal{P}_i\}_i)$ *but also* weakly merges with $EG(\{\mathcal{P}_i\}_i)$.

Corollary 1 *For any countable family $\{\mathcal{P}_i\}_i$ of conditioning rules, there exists a prior $\tilde{\mu}'_F$ such that $\tilde{\mu}'_F$ not only almost weakly merges with $AG(\{\mathcal{P}_i\}_i)$ but also weakly merges with $EG(\{\mathcal{P}_i\}_i)$.*

Finally, Propositions 3 and 4 give us a characterization of almost weak merging with a set of probability measures, as in the weak merging case.

Theorem 2 *A set M of probability measures is almost weakly merged if and only if there exists a countable family $\{\mathcal{P}_i\}_i$ of conditioning rules such that $M \subset AG(\{\mathcal{P}_i\}_i)$.*

Proof. Similar to the proof of Theorem 1. ■

4 Application to Repeated Games

4.1 Basic observation

Let us apply our results to an infinitely repeated game of two players; the argument may be easily extended to the n player case, and our results about almost weak merging may also be applied in the same way. Suppose that each player may observe a past history of actions at each period. Player $i(=1,2)$ takes an action a^i in a finite set A_i at every period. A history of the repeated game is a sequence of actions by players 1 and 2. Notations h_T, h_∞, H, $\mathbf{H}_\infty$, $h \leq h'$, $h < h'$, and $h < h_\infty$ have the same meanings as in Subsection 2.1. Player i's behavior strategy is denoted by $\sigma_i : H \to \Delta(A_i)$. Let Σ_i denote the set of all player i's behavior strategies. We write $\mu(\sigma_1, \sigma_2)$ for the probability measure (on $\mathbf{H}_\infty$) induced by playing σ_1 and σ_2.

Since player $i(\neq j)$'s *prior belief* about player j's behavior is identified with a player j's behavior strategy, let $\tilde{\sigma}_j : H \to \Delta(A_j)$ designate a player i's prior. Note that given a player i's strategy σ_i, $\mu(\sigma_i, \tilde{\sigma}_j)$ weakly merges with $\mu(\sigma_i, \sigma_j)$ if and only if for $\mu(\sigma_i, \sigma_j)-$almost all h_∞,

$$\lim_{T \to \infty} \|\tilde{\sigma}_j(h_T) - \sigma_j(h_T)\| = 0.$$

Note also that merging with an opponent true strategy may *depend on* a player own behavior in a repeated game. If a prior weakly merges with an opponent strategy for *all* player's strategies, we say that the prior weakly merges with the opponent strategy.

Definition 12 *We say that a player i's prior $\tilde{\sigma}_j$ weakly merges with a set M_j of player j's strategies if for all $\sigma_i \in \Sigma_i$ and all $\sigma_j \in M_j$, $\mu(\sigma_i, \tilde{\sigma}_j)$ weakly merges with $\mu(\sigma_i, \sigma_j)$. It is said that a set M_j of player j's strategies is weakly merged if there exists a player i's prior $\tilde{\sigma}_j$ such that $\tilde{\sigma}_j$ weakly merges with M_j.*

A behavior strategy may be generated by a countable family of conditioning rules, as in the CPS case.

Definition 13 *We say that a player j's strategy σ_j is generated by a countable family $\{\mathcal{P}_n\}_n$ of conditioning rules if for all $\varepsilon > 0$ there exists n_0 such that for all $\beta \in \mathcal{P}_{n_0}$ and* *all* $h, h' \in \beta$, $\|\sigma_j(h) - \sigma_j(h')\| < \varepsilon$. *The set of all player j's strategies generated by* $\{\mathcal{P}_n\}_n$ *is denoted by* $G_j(\{\mathcal{P}_n\}_n)$.

Noting that merging in a repeated game may depend on a player's own behavior, we define the eventual generation of behavior strategies.

Definition 14 (1) *We say that a player j's strategy* σ_j *is eventually generated by a countable family* $\{\mathcal{P}_n\}_n$ *of conditioning rules with a player i's strategy* σ_i *if for all* $\varepsilon > 0$ *there exists an index* n_0, *a* $\mu(\sigma_i, \sigma_j)-$*probability one set* $\mathbf{Z}_0$, *and a time function* $T_0 : \mathbf{Z}_0 \to \mathbb{N}$ *such that for all* $\beta \in \mathcal{P}_{n_0}$ *and all* $h_T, h'_{T'} \in \beta$, *if there exist* $h_\infty, h'_\infty \in \mathbf{Z}_0$ *such that* $h_T < h_\infty$, $T \geq T_0(h_\infty)$, $h'_{T'} < h'_\infty$, *and* $T' \geq T_0(h'_\infty)$, *then* $\|\sigma_j(h_T) - \sigma_j(h'_{T'})\| < \varepsilon$. *The set of all player j's strategies eventually generated by* $\{\mathcal{P}_n\}_n$ *with* σ_i *is denoted by* $EG_j(\{\mathcal{P}_n\}_n, \sigma_i)$.

(2) *We say that a player j's strategy* σ_j *is eventually generated by a countable family* $\{\mathcal{P}_n\}_n$ *of conditioning rules if* σ_j *is eventually generated by* $\{\mathcal{P}_n\}_n$ *with all* σ_i. *The set of all player j's strategies eventually generated by* $\{\mathcal{P}_n\}_n$ *is denoted by* $EG_j(\{\mathcal{P}_n\}_n)$.

Obviously, $EG_j(\{\mathcal{P}_n\}_n) = \bigcap_{\sigma_i} EG_j(\{\mathcal{P}_n\}_n, \sigma_i)$ and $G_j(\{\mathcal{P}_n\}_n) \subsetneqq EG_j(\{\mathcal{P}_n\}_n)$. As in the CPS case, *all* i.i.d. strategies are generated by *any* (non-empty) family of conditioning rules;[7] thus $G_j(\{\mathcal{P}_n\}_n)$ is always *uncountable*. Also, *any* strategy σ_j is generated by a *countable* family of conditioning rules. Thus any countably many strategies $\{\sigma_j^n\}_n$ are generated by a countable family of conditioning rules. In general, even $G_j(\{\mathcal{P}_n\}_n)$ is much larger than previously known learnable sets; $G_j(\{\mathcal{P}_n\}_n)$ is also neither parametrized in any finite dimensional space nor compact (in the sup norm). Propositions 5 and 6 correspond to Propositions 1 and 2 respectively.

[7] We say that σ_j is i.i.d. if $\sigma_j(h) = \sigma_j(h')$ for all $h, h' \in H$.

Proposition 5 *For any player i's prior $\tilde{\sigma}_j$, there exists a countable family $\{\mathcal{P}_n\}_n$ of conditioning rules such that for all $\sigma_i \in \Sigma_i$ and all $\sigma_j \notin EG_j(\{\mathcal{P}_n\}_n, \sigma_i)$, $\mu(\sigma_i, \tilde{\sigma}_j)$ does not weakly merge with $\mu(\sigma_i, \sigma_j)$.*

Proof. As in the CPS case, $\tilde{\sigma}_j$ has its $\frac{1}{n}-$approximate conditioning rule $\mathcal{P}^{\tilde{\sigma}_j}_{\frac{1}{n}}$ for all n. The rest of the argument is the same as the proof of Proposition 1. ■

Proposition 6 *For any countable family $\{\mathcal{P}_n\}_n$ of conditioning rules, there exists a frequency-based prior $\tilde{\sigma}^F_j$ such that for all $\sigma_i \in \Sigma_i$ and all $\sigma_j \in EG_j(\{\mathcal{P}_n\}_n, \sigma_i)$, $\mu(\sigma_i, \tilde{\sigma}^F_j)$ weakly merges with $\mu(\sigma_i, \sigma_j)$.*

Proof. The construction of $\tilde{\sigma}^F_j$ is just the same as that of a frequency-based CPS f_F except that $D^{(i,\beta)}_T$ is the conditional empirical distribution of *player j's actions*. ■

Note that $\tilde{\sigma}^F_j$ is exactly a belief formation process of conditional (smooth) fictitious play in Noguchi (2003); taking a (smooth approximate) *myopic* best response to $\tilde{\sigma}^F_j$ is just conditional (smooth) fictitious play (see Fudenberg and Levine (1999)). Thus, in the case that a player takes a myopic best response to his belief, conditional fictitious play is interpreted as a Bayesian learning procedure. Surprisingly Propositions 5 and 6 imply that from the weak merging point of view, the learning performance of conditional (smooth) fictitious play is *better than* or at least *as good as* that of *any other* Bayesian learning procedure.[8]

Corollary 2 *For any prior $\tilde{\sigma}_j$, there exists a frequency-based prior $\tilde{\sigma}^F_j$ such that $\tilde{\sigma}^F_j$ always weakly merges with all strategies that $\tilde{\sigma}_j$ could merges with: for all strategies σ_j, if $\mu(\sigma_i, \tilde{\sigma}_j)$*

[8] In addition, conditional *smooth* fictitious play also has a sophisticated no-regret property *simultaneously* while Bayesian learning procedures do not have a no-regret property. See Noguchi (2003).

weakly merges with $\mu(\sigma_i, \sigma_j)$ for some σ_i, then $\mu(\sigma_i, \tilde{\sigma}_j^F)$ weakly merges with $\mu(\sigma_i, \sigma_j)$ for all σ_i.

Propositions 5 and 6 also entail a characterization of a set of behavior strategies with which a prior weakly merges.

Theorem 3 *A set M_j of player j's strategies is weakly merged if and only if there exists a countable family $\{\mathcal{P}_n\}_n$ of conditioning rules such that $M_j \subset EG_j(\{\mathcal{P}_n\}_n)$.*

Proof. The proof is similar to that of Theorem 1. ■

4.2 Regular strategies

As a typical example of a learnable set we may consider the set of *regular strategies*. Regular strategies have strong regularities in the sense that those conditioning rules are determined by computer algorithms. Precisely, a function $\Lambda : H \times H \to \{0, 1\}$ is called a characteristic function of a partition $\mathcal{P}$ if for all $h, h' \in H$, $h \sim_{\mathcal{P}} h' \Leftrightarrow \Lambda(h, h') = 1$. A conditioning rule $\mathcal{P}$ is said to be *regular* if its characteristic function is *(Turing machine) computable*. Let Φ^R designate the set of all regular conditioning rules. Note that Φ^R is countable because Turing machines are countable. A strategy $\sigma_j : H \to \Delta(A_j)$ is called *regular* if σ_j is generated by Φ^R. Most of practical strategies, including all i.i.d. strategies, all (Turing machine) computable pure strategies,[9] all Markov strategies of all orders, equilibrium strategies in Folk Theorems,[10] and so on, are regular. Regular strategies may

[9] For example, see Nachbar and Zame (1996) for computable pure strategies.

[10] For example, equilibrium strategies in Fudenberg and Maskin (1991) are regular, provided that players' discount factors, players' payoffs in a stage game, and the target values of averaged discounted payoff sums are computable numbers.

also be interpreted as a generalization of computable pure strategies to *mixed* strategies. Indeed, any computable pure strategy is generated by some regular conditioning rule. Let Σ_j^R denote the set of all player j's regular strategies: $\Sigma_j^R := G_j(\Phi^R)$. Σ_j^R is also so large that it cannot be finite dimension parameterized nor compact (in the sup norm). Since regular conditioning rules are countable, we may apply Proposition 6 to Σ_j^R.

Corollary 3 *There exists a frequency-based prior $\tilde{\sigma}_j^F$ such that $\tilde{\sigma}_j^F$ weakly merges with Σ_j^R.*

Remark 3 *The union of Φ^R and any countable family $\{\mathcal{P}_i\}_i$ of conditioning rules is also countable. Thus we may always assume that Σ_j^R is included in a learnable set.*

4.3 Implication to Nachbar's impossibility result

Nachbar (1997) and (2004) show that in a large class of games, *if* all players' learnable sets are so diverse and symmetric that they equally include various strategies, *then* some player's optimizing strategy to his belief does *not* belong to his opponent's learnable set. Do our results strengthen or weaken the validity of Nachbar's impossibility result? As a positive fact, it is straightforward to see that the product of the sets of strategies generated by $\{\mathcal{P}_n\}_n$, $G_1(\{\mathcal{P}_n\}_n) \times G_2(\{\mathcal{P}_n\}_n)$ satisfies the learnability and diversity conditions in Nachbar (2004). Thus the impossibility holds for $G_1(\{\mathcal{P}_n\}_n) \times G_2(\{\mathcal{P}_n\}_n)$: some player i's optimizing strategy to his belief does *not* belong to his opponent's learnable set $G_j(\{\mathcal{P}_n\}_n)$. Nonetheless, since our results show that $EG_i(\{\mathcal{P}_n\}_n)$ includes any learnable set, what we really want to know is whether the impossibility still holds for $EG_1(\{\mathcal{P}_n\}_n) \times EG_2(\{\mathcal{P}_n\}_n)$. Unfortunately it is not clear whether $EG_1(\{\mathcal{P}_n\}_n) \times EG_2(\{\mathcal{P}_n\}_n)$ satisfies one of the diversity conditions, that is, condition CS (1) in Nachbar (2004); loosely speaking, CS (1) requires that if a *mixed* strategy belongs to $EG_i(\{\mathcal{P}_n\}_n)$, then some *pure* strategy in the support of the mixed strategy also belong to $EG_i(\{\mathcal{P}_n\}_n)$. If the impossibility holds for $EG_1(\{\mathcal{P}_n\}_n) \times EG_2(\{\mathcal{P}_n\}_n)$, then it may be a strong evidence for the generality of Nachbar's impossibility; otherwise, one may doubt its generality.

5 Concluding Remarks

5.1 Bayesian prior

One might think that a prior belief $\tilde{\mu}$ should be defined by an appropriate Bayesian representation. The representation form clearly exhibits that the prior $\tilde{\mu}$ puts a positive weight on every learnable probability measure μ_θ or at least an arbitrary small neighborhood of it. Typically it may be the form $\tilde{\mu} = \int_\Theta \mu_\theta d\lambda(\theta)$, where Θ is a parameter set and λ is a probability distribution over Θ. Motivated by the thought, Jackson et al. (1999) pursue a *canonical* Bayesian representation. From this point of view, we may ask whether our constructing prior $\tilde{\mu}_F$ (or $\tilde{\mu}'_F$) has such a Bayesian representation. One might want to apply the result in Jackson et al. (1999) to $\tilde{\mu}_F$. According to their result, $\tilde{\mu}_F$ must satisfy the *asymptotically reverse-mixing* condition so as to have a (unique) canonical Bayesian representation. However, it does not seem immediate to check whether $\tilde{\mu}_F$ satisfies the condition or not.

5.2 Merging

We have only explored (almost) weak merging. We may ask the same question about merging: what is a set of probability measures with which a prior merges? Sandroni and Smorodinsky (1999) investigate relations between merging and weak merging, and show that merging requires weak merging with a fast speed in almost all periods. Their result leads us to guess that in general, a merged set of probability measures may be much smaller than a weakly merged set. But the problem is beyond the scope of this paper.

6 Appendix A

Proof of Lemma 2: Suppose that $\liminf_{T\to\infty} i(h_T) < \infty$ for some h_∞. It means that $i(h_{T_k}) = i_0$ for infinitely many T_k. Since $\mathcal{P}_{i_0}$ has only finite classes, it in turn implies that there exists $\beta_0 \in \mathcal{P}_{i_0}$ such that $\beta(h_{T_l}) = \beta_0$ for infinitely many T_l; since $\{T_l\}_l$ is a subsequence of $\{T_k\}_k$, $i(h_{T_l}) = i_0$ for all T_l. It, together with the definition of $m(\cdot)$, implies that $m(h_{T_l}) = i_0$ for all T_l. Thus there exists a unique class $\alpha_0 \in \mathcal{P}_{i_0+1}$ such that $\alpha_0 \subset \beta_0$ and $\alpha(h_{T_l}) = \alpha_0$ for all T_l. But then, by the definition of $n(\cdot)$, $n(h_{T_l}) \to \infty$ as $l \to \infty$. This means that for some l, $n(h_{T_l}) \geq n_0^{\alpha_0} = n_0^{\alpha(h_{T_l})}$, so that by the definition of $i(\cdot)$, $i(h_{T_l}) = m(h_{T_l}) + 1 = i_0 + 1$. This is a contradiction. ■

7 Appendix B

Given a history h_T, let $n_T^{(i,\beta)}$ be the number of times that category (i,β) has been effective up to time T, and let Γ_T denote the set of categories that have been effective up to time T. Then, we obtain the following lemma, which will be used to prove Proposition 4. Lemma

3 (1) says the same as Lemma 2, and Lemma 3 (2) insists that for almost all categories, the prior sample size is negligible relative to the number of effective periods.

Lemma 3 (1) $\lim_{T\to\infty} i(h_T) = \infty$ *for all* h_∞.

(2) *For all* h_∞, $\lim_{T\to\infty} \sum_{(i,\beta)\in\Gamma_T} \frac{n_T^{(i,\beta)}}{T} \frac{n_0^{(i,\beta)}}{n_T^{(i,\beta)}} = 0$.

Proof. (1) Suppose that $\liminf_{T\to\infty} i(h_T) < \infty$ for some h_∞. It means that $i(h_{T_k}) = i_0$ for infinitely many T_k. Since $\mathcal{P}_{i_0}$ has only finite classes, it in turn implies that there exists $\beta_0 \in \mathcal{P}_{i_0}$ such that $\beta(h_{T_l}) = \beta_0$ for infinitely many T_l; since $\{T_l\}_l$ is a subsequence of $\{T_k\}_k$, $i(h_{T_l}) = i_0$ for all T_l. It, together with the definition of $m(\cdot)$, implies that $m(h_{T_l}) = i_0$ for all T_l. Thus there exists a unique class $\alpha_0 \in \mathcal{P}_{i_0+1}$ such that $\alpha_0 \subset \beta_0$ and $\alpha(h_{T_l}) = \alpha_0$ for all T_l. But then, by the definition of $n(\cdot)$, $n(h_{T_l}) \to \infty$ as $l \to \infty$. This means that for some l, $n(h_{T_l}) \geq n_0^{\alpha_0} = n_0^{\alpha(h_{T_l})}$ and

$$\frac{\sum_{i=1}^{m(h_{T_l})+1} \sum_{\beta\in\mathcal{P}_i} n_0^\beta}{n(h_{T_l})} = \frac{\sum_{i=1}^{i_0+1} \sum_{\beta\in\mathcal{P}_i} n_0^\beta}{n(h_{T_l})} < \frac{1}{i_0} = \frac{1}{m(h_{T_l})}$$

so that by the definition of $i(\cdot)$, $i(h_{T_l}) = m(h_{T_l}) + 1 = i_0 + 1$. This is a contradiction.

(2) Let $i^*(h_T) := \max\{i(h_t) \mid t \leq T\}$, $t^*(h_T) := \min\{t \mid i(h_t) = i^*(h_T), t \leq T\}$, $n^*(h_T) := n(h_{t^*(h_T)})$, and $m^*(h_T) := m(h_{t^*(h_T)})$. Since $i(h_T) \to \infty$ as $T \to \infty$, $i^*(h_T) \to \infty$ as $T \to \infty$. Note that $i^*(h_T) = m^*(h_T) + 1$ because switching occurs at time $t^*(h_T) + 1$. Note also that if category (i, β) has been effective through time $T+1$, then $\beta \in \bigcup_{i=1}^{i^*(h_T)} \mathcal{P}_i$. Thus $\sum_{(i,\beta)\in\Gamma_T} n_0^{(i,\beta)} \leq \sum_{i=1}^{m^*(h_T)+1} \sum_{\beta\in\mathcal{P}_i} n_0^\beta$. Obviously $n^*(h_T) \leq T$. These induce that

$$\sum_{(i,\beta)\in\Gamma_T} \frac{n_T^{(i,\beta)}}{T} \frac{n_0^{(i,\beta)}}{n_T^{(i,\beta)}} = \frac{\sum_{(i,\beta)\in\Gamma_T} n_0^{(i,\beta)}}{T} \leq \frac{\sum_{i=1}^{m^*(h_T)+1} \sum_{\beta\in\mathcal{P}_i} n_0^\beta}{n^*(h_T)} < \frac{1}{m^*(h_T)} = \frac{1}{i^*(h_T) - 1}.$$

The first equality and the second inequality are obvious. The third inequality holds because switching occurs at time $t^*(h_T) + 1$ so that the switching criterion is passed. Since $i^*(h_T) \to \infty$ as $T \to \infty$, the desired result obtains. ■

Proof of Proposition 3: Fix any prior $\tilde{\mu}$. Let $f_{\tilde{\mu}}$ be a CPS corresponding to $\tilde{\mu}$. As noted in Subsection 2.4, for each n, we may take a $\frac{1}{n}-$approximate conditioning rule $\mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}$ of $f_{\tilde{\mu}}$. We shall show that $\tilde{\mu}$ never weakly merges with any $\mu \notin AG(\{\mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}\}_n)$. Take any probability measure $\mu \notin AG(\{\mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}\}_n)$. Then, there exists $\varepsilon_0 > 0$ such that for all n, all $\mu-$probability one sets $\mathbf{Z}$, and all families of time functions $\{T_m\}_m$, either there exist $\beta \in \mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n}}$ and $h_T, h'_{T'} \in \beta$ such that for some $h_\infty, h'_\infty \in \mathbf{Z}$ and some m, m', $h_T < h_\infty$ and $T = T_m(h_\infty)$ and $h'_{T'} < h'_\infty$ and $T' = T_{m'}(h'_\infty)$ and $\|f_\mu(h_T) - f_\mu(h'_{T'})\| \geq \varepsilon_0$, or $\lim_{T\to\infty} \inf \frac{N_T(h_\infty)}{T} < 1$ for some $h_\infty \in \mathbf{Z}$.

Suppose that $\tilde{\mu}$ almost weakly merges with μ. Then, for $\frac{\varepsilon_0}{4}$, there exist a $\mu-$probability one set $\mathbf{Z}_0$ and a family $\{T^0_m\}_m$ of time functions such that for all $h_\infty \in \mathbf{Z}_0$ and all m, $\|f_{\tilde{\mu}}(h_{T^0_m}) - f_\mu(h_{T^0_m})\| < \frac{\varepsilon_0}{4}$, and $\lim_{T\to\infty} \frac{N^0_T(h_\infty)}{T} = 1$ for all $h_\infty \in \mathbf{Z}_0$, where $N^0_T(h_\infty) := \max\{m \mid T^0_m(h_\infty) \leq T\}$. On the other hand, letting $n_0 \geq \frac{4}{\varepsilon_0}$, it follows from the previous paragraph that for n_0, $\mathbf{Z}_0$ and $\{T^0_m\}_m$, either there exist $\beta \in \mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n_0}}$ and $h_T, h'_{T'} \in \beta$ such that for some $h_\infty, h'_\infty \in \mathbf{Z}_0$ and some m, m', $h_T < h_\infty$, $T = T^0_m(h_\infty)$, $h'_{T'} < h'_\infty$, and $T' = T^0_{m'}(h'_\infty)$ and $\|f_\mu(h_T) - f_\mu(h'_{T'})\| \geq \varepsilon_0$, or $\lim_{T\to\infty} \inf \frac{N^0_T(h_\infty)}{T} < 1$ for some $h_\infty \in \mathbf{Z}_0$. Since $\lim_{T\to\infty} \frac{N^0_T(h_\infty)}{T} = 1$ for all $h_\infty \in \mathbf{Z}_0$, these induce that $\|f_{\tilde{\mu}}(h_T) - f_{\tilde{\mu}}(h'_{T'})\| \geq \frac{\varepsilon_0}{2}$ for $h_T, h'_{T'} \in \beta$. But then, since $\beta \in \mathcal{P}^{f_{\tilde{\mu}}}_{\frac{1}{n_0}}$ and $h_T, h'_{T'} \in \beta$, $\|f_{\tilde{\mu}}(h_T) - f_{\tilde{\mu}}(h'_{T'})\| < \frac{1}{n_0} \leq \frac{\varepsilon_0}{4}$. This is a contradiction. Thus, $\tilde{\mu}$ does not almost weakly merge with μ. ■

Proof of Proposition 4: Without loss of generality, we may assume that $\mathcal{P}_i \leq \mathcal{P}_{i+1}$ for all i. Let f'_F be a frequency-based CPS for $\{\mathcal{P}_i\}_i$. Fix any $\mu \in AG(\{\mathcal{P}_i\}_i)$. Suppose that $\tilde{\mu}'_F$ does not almost weakly merge with μ.

Step 1: On the one hand, since $\tilde{\mu}'_F$ does not almost weakly merge with μ, there exists $\varepsilon_0 > 0$ such that for any $\mu-$probability one set $\mathbf{Z}$, there exists $h_\infty \in \mathbf{Z}$ such that there exists a sequence D such that for all $T \in D$, $\|f'_F(h_T) - f_\mu(h_T)\| \geq \varepsilon_0$, and $\limsup_{T\to\infty} \#(D \bigcap \{1, \cdots, T\}) \diagup T > 0$.

Step 2: On the other hand, since $\mu \in AG(\{\mathcal{P}_i\}_i)$, for all $\varepsilon > 0$ there exist an index i_0, a $\mu-$probability one set $\mathbf{Z}_0$, and a family $\{T^0_m\}_m$ of time functions such that (1) for all $\beta \in \mathcal{P}_{i_0}$ and all $h_T, h'_{T'} \in \beta$, if there exist $h_\infty, h'_\infty \in \mathbf{Z}_0$ and m, m' such that $h_T < h_\infty$ and $T = T^0_m(h_\infty)$ and $h'_{T'} < h'_\infty$ and $T' = T^0_{m'}(h'_\infty)$, then $\|f_\mu(h_T) - f_\mu(h'_{T'})\| < \varepsilon$, and (2) $\lim_{T\to\infty} \frac{N^0_T(h_\infty)}{T} = 1$ for all $h_\infty \in \mathbf{Z}_0$, where $N^0_T(h_\infty) := \max\{m \mid T^0_m(h_\infty) \leq T\}$.

Step 3: Let $\varepsilon := \frac{1}{4}\varepsilon_0$. Then, it follows from Step 2 that for ε, there exist an index i_0, a probability one set $\mathbf{Z}_0$, and a family $\{T^0_m\}_m$ of time functions such that (1) and (2) hold. Since $\mathcal{P}_i \leq \mathcal{P}_{i+1}$ for all i, we may take $i_0 \geq \frac{4}{\varepsilon_0}$. For all $\beta \in \mathcal{P}_{i_0}$, define a class $\hat{\beta}$ as follows: $h_T \in \hat{\beta}$ if and only if $h_T \in \beta$ and $h_T < h_\infty$ and $T = T^0_m(h_\infty)$ for some $h_\infty \in \mathbf{Z}_0$ and some m. Then, for all $\beta \in \mathcal{P}_{i_0}$, let $L^\beta[s] := \sup_{h\in\hat{\beta}} f_\mu(h)[s]$ and $l^\beta[s] := \inf_{h\in\hat{\beta}} f_\mu(h)[s]$; note that $L^\beta[s] - l^\beta[s] \leq \varepsilon$ for all s. Furthermore, for all categories (i, β), we define a class $\gamma(i, \beta)$ as follows: $h_T \in \gamma(i, \beta)$ if and only if time $T + 1$ is an effective period of (i, β), i.e., $(i(h_T), \beta(h_T)) = (i, \beta)$, or time $T + 1$ is one of the first n^β_0 effective periods of (i_p, β_p) *and* $h_T < h_\infty$ and $T = T^0_m(h_\infty)$ for some $h_\infty \in \mathbf{Z}_0$ and some m. Since $\mathcal{P}_i \leq \mathcal{P}_{i+1}$ for all i, for each category (i, β) with $i \geq i_0 + 1$, there exists a unique class $\alpha \in \mathcal{P}_{i_0}$ such that $\gamma(i, \beta) \subset \hat{\alpha}$; let $L^{(i,\beta)}(s) := L^\alpha[s]$ and $l^{(i,\beta)}[s] := l^\alpha[s]$ for all s. Thus, for all $h \in \gamma(i, \beta)$

and all $s \in S$, $l^{(i,\beta)}[s] \le f_\mu(h)[s] \le L^{(i,\beta)}[s]$.

Step 4: For all categories (i,β) with $i \ge i_0 + 1$, let

$$\mathbf{B}_n^{(i,\beta)} := \{h_\infty \mid \mathcal{T}_n^{\gamma(i,\beta)} < \infty, \exists s \in S(\frac{\mathbf{d}_n^{\gamma(i,\beta)}[s]}{n} \ge L^{(i,\beta)}[s] + \frac{1}{i} \text{ or } \frac{\mathbf{d}_n^{\gamma(i,\beta)}[s]}{n} \le l^{(i,\beta)}[s] - \frac{1}{i})\}.$$

Then, from Step 2 and Lemma 1 it follows that for all categories (i,β) with $i \ge i_0 + 1$ and all n, $\mu(\mathbf{B}_n^{(i,\beta)}) \le 2(\#S)\exp(-2ni^{-2})$. Also, by the definition of n_0^β's, $\#\mathcal{P}_i \cdot \sum_{n=n_0^\beta}^{\infty} \exp(-2ni^{-2}) \le \exp(-i)$ for all $\beta \in \mathcal{P}_i$ and all i. These imply that for all $j \ge i_0+1$,

$$\begin{aligned} \mu(\bigcap_j \bigcup_{i\ge j} \bigcup_{\beta\in\mathcal{P}_i} \bigcup_{n\ge n_0^\beta} \mathbf{B}_n^{(i,\beta)}) &\le \mu(\bigcup_{i\ge j} \bigcup_{\beta\in\mathcal{P}_i} \bigcup_{n\ge n_0^\beta} \mathbf{B}_n^{(i,\beta)}) \\ &\le 2(\#S)(1-\exp(-1))^{-1}\exp(-j). \end{aligned}$$

Thus, letting $j \to \infty$, we have $\mu(\bigcap_j \bigcup_{i\ge j} \bigcup_{\beta\in\mathcal{P}_i} \bigcup_{n\ge n_0^\beta} \mathbf{B}_n^{(i,\beta)}) = 0$.

Step 5: Let $\mathbf{E}_0 := \bigcup_j \bigcap_{i\ge j} \bigcap_{\beta\in\mathcal{P}_i} \bigcap_{n\ge n_0^\beta} (\mathbf{B}_n^{(i,\beta)})^c$, where $(\mathbf{B}_n^{(i,\beta)})^c$ is the complement of $\mathbf{B}_n^{(i,\beta)}$. From Steps 3 and 4, $\mu(\mathbf{E}_0 \bigcap \mathbf{Z}_0) = 1$. Thus, by Step 1, for $\mathbf{E}_0 \bigcap \mathbf{Z}_0$, there exists $h_\infty \in \mathbf{E}_0 \bigcap \mathbf{Z}_0$ such that there exists a sequence D such that $\limsup_{T\to\infty} \#(D \bigcap \{1,\cdots,T\})/T > 0$ and for all $T \in D$, $\|f'_F(h_T) - f_\mu(h_T)\| \ge \varepsilon_0$. Let $I_T^{(i,\beta)}$ denote the number of effective periods of category (i,β) in which $\|f'_F(h_T) - f_\mu(h_T)\| \ge \varepsilon_0$. Recall Γ_T is the set of categories that have been effective up to time T. Then, the above statement is equivalent to the following: $\limsup_{T\to\infty} \sum_{(i,\beta)\in\Gamma_T} \frac{n_T^{(i,\beta)}}{T} \frac{I_T^{(i,\beta)}}{n_T^{(i,\beta)}} > 0$. Letting $I_T(\delta) := \{(i,\beta) \mid \frac{I_T^{(i,\beta)}}{n_T^{(i,\beta)}} \ge \delta\}$, this in turn implies that there exists $\delta > 0$ such that for infinitely many T_k, $\sum_{(i,\beta)\in I_{T_k}(\delta)} \frac{n_{T_k}^{(i,\beta)}}{T_k} > \delta$.

Step 6: Since $h_\infty \in \mathbf{Z}_0$, $\lim_{T\to\infty} \frac{N_T^0(h_\infty)}{T} = 1$. Let $J_T^{(i,\beta)}$ be the number of times that for some m, time $T_m^0(h_\infty)$ is an effective period of (i,β) up to time T. Note that $\sum_{(i,\beta)\in\Gamma_T} J_T^{(i,\beta)} = N_T^0(h_\infty)$. Therefore, $\lim_{T\to\infty} \sum_{(i,\beta)\in\Gamma_T} \frac{n_T^{(i,\beta)}}{T} \frac{J_T^{(i,\beta)}}{n_T^{(i,\beta)}} = 1$. It means that for all $\epsilon > 0$, there exists k_0 such that for all $k \geq k_0$, $\sum_{(i,\beta)\in J_{T_k}(\epsilon)} \frac{n_{T_k}^{(i,\beta)}}{T_k} \geq 1-\epsilon$, where $J_{T_k}(\varepsilon) := \{(i,\beta) \mid \frac{J_T^{(i,\beta)}}{n_T^{(i,\beta)}} \geq 1-\epsilon\}$. Moreover, from Lemma 3 (2) it follows that $\lim_{T\to\infty} \sum_{(i,\beta)\in\Gamma_T} \frac{n_T^{(i,\beta)}}{T} \frac{n_0^{(i,\beta)}}{n_T^{(i,\beta)}} = 0$. This implies that for all $\epsilon > 0$, there exists k_1 such that for all $k \geq k_1$, $\sum_{(i,\beta)\in P_{T_k}(\epsilon)} \frac{n_{T_k}^{(i,\beta)}}{T_k} \geq 1-\epsilon$, where $P_T(\varepsilon) := \{(i,\beta) \mid \frac{n_0^{(i,\beta)}}{n_T^{(i,\beta)}} \leq \epsilon\}$.

Step 7: It follows from Steps 5 and 6 that for any $\epsilon > 0$ there exists k_2 such that for all $k \geq k_2$, $\sum_{(i,\beta)\in I_{T_k}(\delta)\bigcap J_{T_k}(\epsilon)\bigcap P_{T_k}(\epsilon)} \frac{n_{T_k}^{(i,\beta)}}{T_k} > \delta$. This, together with $\lim_{T\to\infty} i(h_T) = \infty$ from Lemma 3 (1), implies that for all $\epsilon > 0$, all T, and all $i \geq i_0 + 1$, there exist $\bar{T} \geq T$ and $(\bar{\imath}, \bar{\beta})$ with $\bar{\imath} \geq i$ such that (i) time $\bar{T}+1$ is an effective period of $(\bar{\imath}, \bar{\beta})$, i.e., $i(h_{\bar{T}}) = \bar{\imath}$ and $\beta(h_{\bar{T}}) = \bar{\beta}$, (ii) $\bar{T} = T_m^0(h_\infty)$ for some m, (iii) $\|f_F'(h_{\bar{T}}) - f_\mu(h_{\bar{T}})\| \geq \varepsilon_0$, (iv) $\frac{n_0^{(\bar{\imath},\bar{\beta})}}{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})}} \leq \epsilon$, and (v) $\frac{J_{\bar{T}}^{(\bar{\imath},\bar{\beta})}}{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})}} \geq 1-\epsilon$. Note that

$$f_F'(h_{\bar{T}}) = D_{\bar{T}}^{(\bar{\imath},\bar{\beta})} = \frac{d_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + d_0^{(\bar{\imath},\bar{\beta})}}{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + n_0^{(\bar{\imath},\bar{\beta})}} = \frac{n}{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + n_0^{(\bar{\imath},\bar{\beta})}} \frac{\mathbf{d}_n^{\gamma(\bar{\imath},\bar{\beta})}[s]}{n} + \frac{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + n_0^{(\bar{\imath},\bar{\beta})} - n}{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + n_0^{(\bar{\imath},\bar{\beta})}} \frac{d_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + d_0^{(\bar{\imath},\bar{\beta})} - \mathbf{d}_n^{\gamma(\bar{\imath},\bar{\beta})}[s]}{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + n_0^{(\bar{\imath},\bar{\beta})} - n}$$

where n is the number of times that for some m, time $T_m^0(h_\infty)$ is either an effective period of $(\bar{\imath}, \bar{\beta})$, or one of the first $n_0^{(\bar{\imath},\bar{\beta})}$ effective periods of its predecessor $(\bar{\imath}_p, \bar{\beta}_p)$ (up to time $\bar{T}$). Then, by (iv) and (v), $n \geq n_0^{(\bar{\imath},\bar{\beta})} (= n_0^{\bar{\beta}})$ and $\frac{n}{n_{\bar{T}}^{(\bar{\imath},\bar{\beta})} + n_0^{(\bar{\imath},\bar{\beta})}} \geq \frac{J_{\bar{T}}^{(\bar{\imath},\bar{\beta})}}{(1+\epsilon) n_{\bar{T}}^{(\bar{\imath},\bar{\beta})}} \geq \frac{1-\epsilon}{1+\epsilon}$, which means that $f_F'(h_{\bar{T}}) \approx \frac{\mathbf{d}_n^{\gamma(\bar{\imath},\bar{\beta})}[s]}{n}$. These, together with (ii) and (iii), imply that for some s,

$$\frac{\mathbf{d}_n^{\gamma(\bar{\imath},\bar{\beta})}[s]}{n} \geq L^{(\bar{\imath},\bar{\beta})}[s] + \frac{1}{\bar{\imath}}, \text{ or } \frac{\mathbf{d}_n^{\gamma(\bar{\imath},\bar{\beta})}[s]}{n} \leq l^{(\bar{\imath},\bar{\beta})}[s] - \frac{1}{\bar{\imath}}.$$

Thus, $h_\infty \in (\mathbf{E}_0)^c$ where $(\mathbf{E}_0)^c$ is the complement of $\mathbf{E}_0$. But then, $h_\infty \in \mathbf{E}_0$. This is a contradiction to $\mathbf{E}_0 \bigcap (\mathbf{E}_0)^c = \emptyset$. Therefore $\tilde{\mu}_F'$ almost weakly merges with μ. ■

References

[1] Blackwell, D and L. Dubins (1962). "Merging with Opinions with Increasing Information," Annals of Mathematical Statistics, 38, 882-886.

[2] Foster, D. and H.P. Young (2001). "On the Impossibility of Predicting the Behavior of Rational Agents," Proceedings of the National Academy of Sciences of the USA, 98, 12848–53.

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

[4] Fudenberg, D. and E.S. Maskin (1991). "On the Dispensability of Public Randomization in Discounted Repeated Games," Journal of Economic Theory, 53, 428-438.

[5] Gilboa, I., A. Postlewaite, and D. Schmeidler (2004). "Rationality of Belief Or: Why Bayesianism is neither Necessary nor Sufficient for Rationality," Mimeo. Tel Aviv Univ.

[6] Jackson, M.O., E. Kalai, and R. Smorodinsky (1999). "Bayesian Representation of Stochastic Processes under Learning: De Finetti Revisited," Econometrica, 67, 875-893.

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

[8] Kalai, E. and E. Lehrer (1994). "Weak and Strong Merging of Opinions," Journal of Mathematical Economics, 23, 73-86.

[9] Lehrer, E. and R. Smorodinsky (1996a). "Merging and Learning," in Statistics, Probability, and Game Theory: Papers in Honor of David Blackwell, edited by Ferguson, Shapley, and McQueen. IMS Lecture Notes-Monograph Series Vol. 30, Springer-Verlag, New York.

[10] Lehrer, E. and R. Smorodinsky (1996b). "Compatible Measures and Merging," Mathematics of Operations Research, 21, 697-706.

[11] Lehrer, E. and R. Smorodinsky (1997). "Repeated Large Games with Incomplete Information," Games and Economic Behavior, 18, 116-134.

[12] Nachbar, J.H. and W.R. Zame (1996). "Non-Computable Strategies and Discounted Repeated Games," Economic Theory, 8, 103-122.

[13] Nachbar, J.H. (1997). "Prediction, Optimization, and Learning in Repeated Games," Econometrica, 65, 275-309.

[14] Nachbar, J.H. (2004). "Beliefs in Repeated Games," Mimeo., Washington Univ.

[15] Noguchi, Y. (2000). "Optimal Properties of Conditional Fictitious Play," in Ph.D. dissertation at Harvard Univ.

[16] Noguchi, Y. (2003). "Optimal Properties of Conditional Adaptive Strategies," Mimeo., Hitotsubashi Univ.

[17] Sandroni, A. (1998). "Necessary and Sufficient Conditions for Convergence to Nash Equilibrium: The Almost Absolute Continuity Hypothesis," Games and Economic Behavior, 22, 121-147.

[18] Sandroni, A. and R. Smorodinsky (1999). "The Speed of Rational Learning," International Journal of Game Theory, 28, 199-210.

[19] Shiryaev, A.N. (1984). Probability. Springer-Verlag, New York.