Discussion Paper No. 1228

Bayesian Representation of Stochastic Processes under Learning: de Finetti Revisited

Matthew O. Jackson Ehud Kalai Rann Smorodinsky

Revised: May 6, 1998

Math Center web site:
http://www.kellogg.nwu.edu/research/math

Bayesian Representation of Stochastic Processes under Learning: de Finetti Revisited*

Matthew O. Jackson Ehud Kalai Rann Smorodinsky

Revised: May 6, 1998

Abstract

A probability distribution governing the evolution of a stochastic process has infinitely many Bayesian representations of the form $\mu = \int_\Theta \mu_\theta d\lambda(\theta)$. Among these, a natural representation is one whose components (μ_θ's) are 'learnable' (one can approximate μ_θ by conditioning μ on observation of the process) and 'sufficient for prediction' (μ_θ's predictions are not aided by conditioning on observation of the process). We show the existence and uniqueness of such a representation under a suitable asymptotic mixing condition on the process. This representation can be obtained by conditioning on the tail-field of the process, and any learnable representation that is sufficient for prediction is asymptotically like the tail-field representation. This result is related to the celebrated de Finetti theorem, but with exchangeability

*This is a revision of the previously titled manuscript: "Patterns, Types and Bayesian Learning." We thank Nabil Al-Najjar, David Gilat, Ehud Lehrer, and Meir Smorodinsky for helpful conversations. Suggestions by Drew Fudenberg and three anonymous referees have led to substantial improvements in the paper. Financial support under NSF grants SBR 9515421 and 9507912 is gratefully acknowledged. Jackson is at HSS 228-77, Caltech, Pasadena CA 91125, (jacksonm@hss.caltech.edu); Kalai is at MEDS, KGSM, Northwestern University, Evanston, Il 60208, (kalai@nwu.edu); and Smorodinsky is at Industrial Engineering, Technion, Haifa 32000, Israel and MEDS, KGSM, Northwestern University, Evanston, Il 60208, (rann@ie.technion.ac.il).

weakened to an asymptotic mixing condition, and with his conclusion of a decomposition into iid component distributions weakened to components that are learnable and sufficient for prediction.

1 Introduction

Researchers in economics and decision theory often represent the probability distribution governing the the evolution of a stochastic process as a convex combination of the form $\mu = \int_\Theta \mu_\theta d\lambda(\theta)$. Such a representation describes a two stage Bayesian process. In the first stage nature chooses, according to the prior probability measure λ, one of the component measures μ_θ. In the second stage the selected μ_θ governs the evolution of the process. Zellner (1971), Rothschild (1974), and Aumann and Maschler (1995) present classical examples of this approach in Bayesian statistics, decision theory, and game theory, respectively.

It is easy to see, however, that a given μ can have infinitely many different representations. For instance, a researcher representing μ by $\mu = \int_\Theta \mu_\theta d\lambda(\theta)$ could have used an alternative representation $\mu = \int_{\Theta'} \mu_{\theta'} d\lambda'(\theta')$. Moreover, different representations may provide more or less convenient models of the same process.[1]

The purpose of this paper is to identify a specific natural representation of a probability distribution for a discrete-time finite-state stochastic process. We define a natural representation by imposing two requirements on the component distributions. First, we want the representation to be fine enough so that the component distributions are sufficient for (asymptotic) prediction, without needing to condition on additional information. Second, we want the representation to be coarse enough so that it does not consist of component distributions that are not learnable no matter how long the process is observed. As it turns out, such a natural representation exists, and this representation is essentially unique (in a sense to be made precise). Furthermore, this representation can be obtained by conditioning the original distribution on the tail field.

To make this discussion concrete, we first discuss a simple special case.

Example 1: A coin is chosen and then flipped an infinite number of times. The coin is not necessarily fair, i.e., it has a probability θ of turning up heads, 'H', and a probability $1-\theta$ of turn up tails, 'T', and θ is not necessarily

[1] For example, Nyarko (1996) argues that it is important for learning results in incomplete information games to be robust to equivalent reformulations of type spaces. He discusses examples which are not robust to such reformulations. In the language of this paper the reformulations are different representations of the process associated with the same game and strategies.

1/2. In fact, θ is chosen according to a uniform distribution over $[0,1]$. So we may think of this process as first choosing a coin, and then flipping it an infinite number of times. This process corresponds to a probability measure μ over infinite strings of 'H' and 'T's. Note that there are many different convex combinations of component distributions that we could use to represent μ. Let us discuss three. First, there is a representation which naturally corresponds to the description we gave for the process. That is, $\mu = \int_0^1 \mu_\theta d\theta$ where μ_θ corresponds to the measure induced by flipping a coin with parameter θ. From our perspective this will turn out to be a natural representation. Second, there is a representation of $\mu = 1/2\mu_{\text{low}} + 1/2\mu_{\text{high}}$, where μ_{low} corresponds to choosing a coin parameter by a uniform distribution over 0 to 1/2 and then flipping the coin, and μ_{high} to choosing a coin parameter by a uniform distribution over 1/2 to 1 and then flipping the coin. From our perspective this is too coarse a representation since the component distributions do not capture the relevant information about the realized coin that will be observed in the process. As we will make precise, this representation fails to be sufficient for prediction. Third, there is a representation of μ as a convex combination of Dirac measures, each giving weight one to some infinite sequence of heads and tails. Specifically, $\mu = \int_\Omega \delta_\omega d\mu(\omega)$ where each ω corresponds to a single infinite sequence of 'H' and 'T's and δ_ω is a degenerate measure with weight one on the sequence ω. From our perspective this extreme is too fine a representation because it captures information that an observer could never hope to learn. The implication of the main result of the paper for this example is that if one looks for a representation of μ that satisfies both sufficiency for prediction and learnability, then one recovers precisely the representation of μ as a convex combination of the coins.

Let us be a bit more explicit about the definitions of sufficiency for prediction, learnability (formal definitions appear in the Section 2), and our main results; and then discuss the relation of our results to de Finetti's theorem.

Sufficiency for Prediction. A component distribution is sufficient for prediction if the unconditional probabilities of late events are arbitrarily close to their probabilities when additionally conditioned on initial segments of the process. In other words, knowledge of the distribution alone, without any knowledge of the realizations of the process, is sufficient for making asymptotic predictions. A representation, $\mu = \int \mu_\theta d\lambda(\theta)$, is sufficient for

prediction if each of its component measures, μ_θ, is sufficient for prediction. So, the basic idea behind sufficiency for prediction is that the knowledge of the component distribution might be thought of as a statistic which is sufficient for the information an observer might learn from initial histories of the process for the purpose of making predictions of far-off events.

In Example 1 above, the representation of μ as coins with known parameters is sufficient for prediction. A forecaster who knows the parameter (θ) of the coin will make no use of the realized initial segments in assigning probability to the event H at future times t. Similarly, the representation of μ by Dirac measures is trivially sufficient for prediction. In contrast, the coarser representation, $\mu = 1/2\mu_{\text{low}} + 1/2\mu_{\text{high}}$, is not sufficient for prediction. For instance, under μ_{high}, the unconditional probability of H at any time t (including late times) is 3/4, while, the probability of H at time t conditional on an initial long segment that is very rich in H's is greater than 3/4.

Learnability. 'Learnability' has many interpretations. We follow the recent game theoretic Bayesian learning literature and define a representation to be learnable if a long term observer of the process, who starts only with knowledge of the original μ, by conditioning on past observations makes approximately the same predictions as a person who is additionally informed of the realized component distribution. In other words, the probabilities of future events conditional on history alone become arbitrarily close to the probabilities conditional on history and knowledge of the realized component distribution.

In the coins example, the representation of μ as coins with known parameters is learnable since an observer (who is not told the realized parameter of the coin) who observes a long history of outcomes will predict the probability of 'H' arbitrarily closely to the chosen parameter, as if he was told the parameter. Similarly, the coarser representation of $\mu = 1/2\mu_{\text{low}} + 1/2\mu_{\text{high}}$ is learnable. In contrast, the fully refined representation of μ by Dirac measures is not learnable. No matter how long a forecaster observes the process, he will not predict future realizations of H and T's as if he knew those realizations.

Our Main Results. The above discussion points out that learnability limits how 'fine' a representation can be and sufficiency for prediction limits how 'coarse' it can be. In the coins example we pointed out a representation which satisfies both of these requirements. More generally one would like to know if there always exists such a representation and, if so, what does the

class of all such representations look like. Our main results show for a certain class of mixing processes that there always exists a representation that is both learnable and sufficient for prediction, and that any such representation is asymptotically like the representation that one obtains by conditioning on the tail field. Thus, the conditions of learnability and sufficiency for prediction together identify representations corresponding to the tail field.

Relationship to de Finetti's Theorem. The celebrated de Finetti Theorem suggests an example of a learnable representation which is sufficient for prediction. Illustrated in the the space of infinite sequences of H and T's, de Finetti considers situations where the probability assigned to every initial finite sequence is exchangeable, i.e., the probability depends entirely on the number of H's and T's and not on their order in the sequence. De Finetti shows that the overall probability of such a process may be represented as a convex combination of distributions induced by repeated i.i.d. coin tosses, where the parameter of the coin is random. Thus, in the language of this paper, he represents an exchangeable distribution by a convex combination of learnable distributions that are sufficient for prediction.

Our representation result is similar to de Finetti's, except that we replace the exchangeability condition with a weaker condition of asymptotic reverse-mixing (which is loosely that conditional on sufficient observation additional far-off information does not significantly change the forecast of nearby events). Our conclusion is therefore weaker: we obtain a learnable representation by component distributions that are sufficient for prediction, but not necessarily i.i.d. across time.

The following example illustrates the importance of weakening exchangeability. The process in the example is not exchangeable but satisfies our mixing condition.

Example 2: Consider a finite state Markov chain with n states and an $n \times n$ transition matrix M. Suppose that M is not known, but that is randomly chosen according to a measure λ over possible $n \times n$ transition matrices. The measure μ governing the resulting stochastic process can be represented as $\mu = \int \mu_M d\lambda(M)$.[2] This process is not exchangeable (as, for instance, the probability that the process is in state 1 at both dates 1 and 2 can be quite

[2] An example of such a process arises in game theory when one considers an evolutionary process of the sort described by Kandori, Mailath, and Rob (1993), with an unknown mutation rate.

different from the probability that the process is in state 1 at both dates 1 and 10). Nevertheless, one would like a theorem that recovers $\mu = \int \mu_M d\lambda(M)$ as the natural representation of this process. A natural candidate for a condition to replace exchangeability would seem to be a mixing condition. However, note that this example does not satisfy standard mixing conditions. Initial events provide significant information concerning much later events, since initial realizations help one to estimate M, which influences the forecast late events. Therefore, initial events and much later events are not approximately independent. Nevertheless, the example does satisfy the asymptotic version of a (reverse) mixing condition that we define in this paper. Conditional on sufficient observation there is approximate independence between near and far events.

While exchangeability is too strong for most economic applications, asymptotic reverse mixing is significantly weaker and thus admits many new applications, as illustrated by the example above. Asymptotic reverse-mixing permits long run effects being generated by random early events. One way to test for such an effect is to see whether conditioning on far off events influences forecasts of nearby events. Asymptotic reverse mixing allows such forecasts to be influenced intially, but requires that conditional on sufficient observation of history, conditioning on far off events no longer influences forecasts of nearby events.

2 Definitions

Let $\{(\Omega_t, \mathcal{G}_t)\}_{t=1}^{\infty}$ be a sequence of finite state spaces and corresponding σ–fields. Let $\Omega = \times_{t=1}^{\infty}\Omega_t$, and let $\mathcal{F}$ be the σ-field on Ω generated by $\{\mathcal{G}_t\}_{t=1}^{\infty}$; i.e., $\mathcal{F} = \sigma\left(\cup_{t=1}^{\infty}\mathcal{G}_t\right)$, where $\mathcal{G}_t$ denotes the σ-field on Ω_t and its corresponding extension to Ω. Note that $\mathcal{F}$ is countably generated. Let $\mathcal{F}_t = \sigma\left(\cup_{j=1}^{t}\mathcal{G}_j\right)$. $\{\mathcal{F}_t\}_{t=1}^{\infty}$ is a filtration on $(\Omega, \mathcal{F})$. The notation $\mathcal{G}_t^{t'}$ denotes $\vee_{j=t}^{t'}\mathcal{G}_j$. Let Δ be the set of probability measures on $(\Omega, \mathcal{F})$. We treat Ω, $\{\mathcal{G}_t\}_{t=1}^{\infty}$, and $\mu \in \Delta$ as fixed.

Representations

Definition: A quadruple $(\Theta, \mathcal{B}, \lambda, (\mu_\theta)_{\theta\in\Theta})$ consisting of a probability space $(\Theta, \mathcal{B}, \lambda)$ and probability measures $\mu_\theta \in \Delta$ is a *representation* if $\forall S \in \mathcal{F}$

1. $\theta \to \mu_\theta(S)$ is Borel measurable, and
2. $\mu(S) = \int_\Theta \mu_\theta(S)\, d\lambda(\theta)$.

For any fixed μ, there are various representations from which to choose. Two extreme examples are

Θ consists of a single point θ and $\mu_\theta = \mu$, and

$\Theta = \Omega$, $\mathcal{B} = \mathcal{F}$, $\lambda = \mu$, and $\mu_\omega(S) = I_S(\omega)$ (where I_S is the indicator function, i.e., the Dirac measure on ω).

With a representation, we can think of a random draw of ω according to μ as equivalently first choosing a parameter θ by λ and then choosing ω by μ_θ. In other words, a representation consists of a prior λ over $(\Theta, \mathcal{B})$ and a collection of posteriors μ_θ over Ω.

We now provide precise definitions for sufficiency for prediction and learnability which will be used to identify a specific class of representations.

Sufficiency for Prediction

Definition: A measure $\tilde{\mu}$ *is sufficient for prediction* if for all t

$$\lim_n \sup_{A\in\mathcal{G}_n^\infty} |\tilde{\mu}(A|\mathcal{F}_t) - \tilde{\mu}(A)| = 0 \quad \tilde{\mu} - \text{a.e.}.$$

This definition states that conditioning on the filtration will not change the forecasts of far-off events by someone who already knows a measure $\tilde{\mu}$. We apply this definition to each μ_θ.

Definition: A representation $(\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ *is sufficient for prediction* if μ_θ *is sufficient for prediction* for λ-a.e. $\theta \in \Theta$.

To see the intuition behind the above definition (and especially the role of t), consider an agent who knows the transition probabilities of an irreducible, aperiodic Markov chain but is unfamilar with the history or current state of the chain (as in Example 2). Her forecast regarding the next period state may be incorrect (relative to someone who knows the history), yet the agent knows the pattern that the chain will follow asymptotically: her prediction about events on the far horizon are independent of the current state of the chain. In this case, knowing the transition probabilities (modeled as knowing θ) is sufficient for making predictions.

Learnability

Learnability is made precise by means of the notion of merging of measures originated by Blackwell and Dubins (1962). We use a weaker notion from Kalai and Lehrer (1994), which has proven to be useful in the Bayesian learning literature (e.g., Kalai and Lehrer (1993), Lehrer and Smorodinsky (1997), and Jackson and Kalai (1997)).

Definition: Consider ν and $\tilde{\nu} \in \Delta$. $\tilde{\nu}$ *merges* with ν if $\forall\ \epsilon > 0$, ℓ, and $\nu - a.e.\ \omega \in \Omega \quad \exists T$ such that for all $t \geq T$

$$\sup_{n \geq t, A \in \mathcal{G}_n^{n+\ell}} |\tilde{\nu}(A|\mathcal{F}_t) - \nu(A|\mathcal{F}_t)| < \epsilon.$$

If $\tilde{\nu}$ merges with ν, then eventually forecasts provided by $\tilde{\nu}$ regarding any finite horizon events at arbitrary times in the future will approach the "true" forecasts provided by ν.

This definition appears to be stronger than the definition of merging that appears in Kalai and Lehrer (1994), where $\sup_{n \geq t, A \in \mathcal{G}_n^{n+\ell}}$ would be replaced by $\max_{A \in \mathcal{G}_{t+1}}$. However, it is proven in the appendix (see Lemma 1) that the two definitions are equivalent.

The notion of merging formalizes what we mean by a learnable representation:

Definition: A representation $(\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ is *learnable* if μ merges with μ_θ for λ-a.e. $\theta \in \Theta$.

A representation is learnable if an observer of the filtration will eventually make predictions *as if* she had been informed about the realized parameter θ.

Thus, given what the observer has learned through the filtration, knowledge of θ has become redundant in that it would not change the observer's forecast. This means that the observer learns the distribution μ_θ (at least along the realized path), which is different from learning θ as we illustrate in Section 5.

Asymptotic Reverse-Mixing

The following example shows that one cannot have a theory of learnability and sufficiency for prediction that applies to all stochastic processes, as for some stochastic processes learnability and suffiency for prediction are incompatible. The distribution in the example continually brings in new, yet unlearned information.

Example 3: Let $\Omega_t = \{0, 1\}$. Consider μ generated as follows: Partition the set of periods $\mathcal{N} = \cup_{i=1}^{\infty} N_i$, where the N_i's are defined by letting N_1 be $\{1, 4, \ldots, n^n, \ldots\}$, and then N_2 is enumerated by renumbering the remaining $\mathcal{N} \setminus N_1$ and taking the corresponding slots $\{1, 4, \ldots, n^n, \ldots\}$ (so N_2 works out to be $\{2, 6, \ldots, n^n + n, \ldots\}$), and so on. Let $\Theta = [0, 1]^{\mathcal{N}}$ be the parameter space. Given $\theta = (\theta_1, \ldots, \theta_i, \ldots)$, μ_θ is the measure representing a sequence of independent coin flips where the probability of heads at time t is given by θ_i when $t \in N_i$. Assume that the prior λ used to select a $\theta \in \Theta$ has the property that the selection of the component θ_i is independent of other components θ_j for $j \neq i$. This means that if we do not know the entire infinite length θ, then no matter how long we wait, there will be new, independent coins used in future periods that we will have no useful information about. In fact, this happens on a non-trivial (positive density) set of periods. Thus, there will always be periods in which the forecast of an agent who has only observed history will differ from that of an agent who knows the information of θ from the representation. Finally, note that in this example any representation that is sufficient for prediction must make predictions as if one knew the entire sequence of coins and therefore will fail to be learnable in a similar manner.

3

This example illustrates the problem that there may be clear patterns associated with the sequences that arise from the filtration, and yet there is always important information that cannot be learned from any finite history: the information one needs in order to make predictions, is always contained further in the future. The asymptotic reverse mixing condition defined next rules out such chaotic distributions and guarantees that the patterns identifiable from the filtration are learnable.

Definition: μ is *asymptotically reverse-mixing* if for any $\delta > 0$ and μ-a.e. ω there exists T such that for any $t \geq T$

$$\overline{\lim}_n \overline{\lim}_\ell \max_{A \in \mathcal{G}_{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega) - \mu(A|\mathcal{F}_t \vee \mathcal{G}_n^{n+\ell})(\omega) \right| < \delta.$$

The definition of asymptotic reverse-mixing requires that conditioning on events far in the future of the process does not significantly alter predictions about nearby events, conditional on sufficient observation of the initial realization of the process. Thus, asymptotic reverse-mixing allows for lasting initial effects and dependence, but requires that eventually short run events have no lasting effect on the process in that they are approximately independent of far future events.

Our definition of asymptotic mixing is similar to standard definitions of ϕ-mixing (see Billingsley (1968)), but differs in two important respects. First, since we are interested in measures conditional on some observations (and are not restricting attention to stationary processes), we only require a mixing condition to hold conditional on sufficient information. Hence, the name 'asymptotic' and the role of T in the definition. Second, the condition is stated in terms of probabilities of nearby events being approximately independent of conditioning on far future events, whereas other mixing conditions are often stated the other way around (and this type of independence can be asymmetric). We need such a condition since our representations involve conditioning on the knowledge of θ, which when sufficient for prediction turns

[3] A similar example appears in Al-Najjar (1998) to demonstrate a chaotic asset market which is impossible to model with linear patterns (factor structure as in the Arbitrage Pricing Theory).

out to be equivalent to knowing limiting future patterns, captured through $\mathcal{G}_n^{n+t}$ in the definition above.

The μ described in Example 3 is not asymptotically reverse-mixing. For any T one can find a date $t > T$ where a new coin is brought in and so conditioning on additional future observations can significantly impact the forecasts.

3 The Main Theorems

We now state the main results of the paper.

Let $\mathcal{F}^{\text{tail}}$ denote the tail σ-field, $\mathcal{F}^{\text{tail}} = \cap_{j=1}^{\infty}\sigma(\cup_{t=j}^{\infty}\mathcal{G}_t)$. Let $\overline{\mathcal{F}}^{\text{tail}}$ denote the representation induced by the tail field: $(\Omega, \mathcal{F}, \mu, \mu_\omega^{\mathcal{F}^{\text{tail}}})$, where $\mu_\omega^{\mathcal{F}^{\text{tail}}}$ denotes the measure conditional on $\mathcal{F}^{\text{tail}}$ at ω (i.e., fixing versions of conditional expectations, $\mu_\omega^{\mathcal{F}^{\text{tail}}}(A) = E(I_A|\mathcal{F}^{\text{tail}})_\omega$, where I denotes the indicator function). It is shown in the appendix (using results of Dellacherie and Meyer (1978) and Stinchcombe (1990)) that this is in fact a representation, and in particular that $\mu_\omega^{\mathcal{F}^{\text{tail}}}(\cdot)$ is a probability measure for μ-a.e. ω.

The following results show that the tail field precisely captures the asymptotic information that an observer will learn through the filtration. This is stated in three pieces. First, the tail field is sufficient for prediction. Second, the tail field is learnable. Third, any representation which satisfies these properties is equivalent to the tail field.

Theorem 1: $\overline{\mathcal{F}}^{\text{tail}}$ is sufficient for prediction.

Theorem 2: $\overline{\mathcal{F}}^{\text{tail}}$ is learnable if, and only if, μ is asymptotically reverse-mixing.

Collecting these two results, it follows that $\overline{\mathcal{F}}^{\text{tail}}$ is learnable and is sufficient for prediction if, and only if, μ is asymptotically reverse-mixing.

The above results suggest that the representation based on the tail field is the 'natural' representation we were searching for. Note however, that there may be other learnable representations that are sufficient for prediction. Nevertheless, as we show below all such representations are asymptotically like the tail field, in that after sufficient time and conditional on any observation

their component measures behave like the component measures in the tail field representation.

Representations Asymptotically like the Tail Field Representation

As our definitions of sufficiency for prediction and learnability are asymptotically based, there may be many representations which are learnable and sufficient for prediction, and yet are differentiated in how their component measures describe finite horizon behavior. A simple example illustrates this point and motivates the definition of equivalence to follow.

Example 4: Consider a process where a coin is flipped at each date. From time 2 on, a fair coin (probability 1/2 of heads) is flipped. At time 1 (and only time 1), either a coin with probability 2/3 of heads is flipped, or a coin with probability 1/3 of heads is flipped. The choice of the coin at time 1 is made by a flip of a fair coin. Thus, in fact μ corresponds to the process by which a fair coin is flipped at every date. However, a valid representation is to have $\Theta = \{\theta_1, \theta_2\}$ and μ_{θ_1} have probability 2/3 of heads at date 1, and 1/2 thereafter, and μ_{θ_2} have probability 1/3 of heads at date 1, and 1/2 thereafter; and to have $\lambda(\theta_1) = \lambda(\theta_2) = 1/2$. This representation is learnable and sufficient for prediction, but the same is also true of the trivial representation of μ as itself. These two representations are asymptotically like the tail field representation, as defined below.

Definition: $\overline{\Theta}$ is *asymptotically like* $\overline{\mathcal{F}}^{\text{tail}}$ if for λ-a.e. θ and for every ℓ

$$\lim_K \sup_{t,n:t+n\geq K} \max_{A\in\mathcal{G}_{t+n}^{t+n+\ell}} |\mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_t)(\omega)| = 0$$

for μ_θ-a.e. ω.

Note that $\overline{\mathcal{F}}^{\text{tail}}$ is asymptotically like $\overline{\mathcal{F}}^{\text{tail}}$, so that such representations exist.[4]

The essence of the definition of asymptotic likeness is that one does not care about the particular labels θ *to the extent* that the associated measures

[4]This is not quite as obvious as it seems. Applying the definition of asymptotically like to $\overline{\mathcal{F}}^{\text{tail}}$ leads to comparisons of the sort $|\mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega) - \mu_{\omega'}^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega)|$ where ω' may differ from ω. This is handled by Lemma 3 in the appendix.

μ_θ lead to the same predictions as the corresponding measures obtained by conditioning on tail field. The role of K in the definition ensures that the given representation provides approximately the same predictions both for very far off events conditional on an arbitrary history, and also on nearby events conditional on sufficient observation of history. The definition does not require that a given representation provide similar predictions for nearby events conditional on a short history. This captures the idea that the given representation may contain some additional finite information, but may not contain additional arbitrarily long run information relative to the tail field.

We show in the appendix (Lemma 5) that one can rewrite the above definition of asymptotic likeness in terms of two comparisons that are analogs to the comparisons made in the definitions of learnability and sufficiency for prediction. More precisely, we show that:

$\overline{\Theta}$ is asymptotically like $\overline{\mathcal{F}}^{\text{tail}}$ if and only if for λ-a.e. θ, any ℓ and μ_θ-a.e. ω

$$\overline{\lim}_t \sup_{n \geq t, A \in \mathcal{G}_n^{n+\ell}} |\mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_t)(\omega)| = 0$$

and for λ-a.e. θ, any fixed t' and ℓ, and μ_θ-a.e. ω

$$\overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_{t'})(\omega) - \mu_\theta(A|\mathcal{F}_{t'})(\omega)| = 0.$$

With these comparisons in hand, one can establish the learnability and sufficiency for prediction of a given representation $\overline{\Theta}$, from the corresponding properties of $\overline{\mathcal{F}}^{\text{tail}}$, as captured in the following theorem (whose proof appears in the appendix).

Theorem 3: If μ is asymptotically reverse-mixing, then a representation of μ, $\overline{\Theta}$, is learnable and is sufficient for prediction, if and only if $\overline{\Theta}$ is asymptotically like $\overline{\mathcal{F}}^{\text{tail}}$.

The following corollary summarizes our results:

Corollary 1: $\overline{\mathcal{F}}^{\text{tail}}$ is sufficient for prediction, and is learnable if and only if μ is asymptotically reverse-mixing. Moreover, if μ is asymptotically reverse-mixing, then a representation of μ, $\overline{\Theta}$, is learnable and sufficient for prediction, if and only if it is asymptotically like the tail field representation.

4 Discussion of Learnability

Example 4 shows that there may be asymptotically equivalent representations that have different sets of parameters mapping into the same (or asymptotically similar) measures. One implication of this is that even though an observer may learn the distribution and thus to forecast as if he or she knows the parameter θ, the observer may never be able to identify θ. This clarifies the scope of the 'learnable' condition and distinguishes it from another condition which is known as 'consistency' (see Diaconis and Freedman (1986)).

Definition: The representation $(\Theta, \mathcal{B}, \lambda, (\mu_\theta)_{\theta\in\Theta})$ is *consistent* if Θ is a topological space and for λ-a.e. $\theta \in \Theta$ the posterior probability measure on Θ conditional on $\mathcal{F}_t$ [5] weakly converges to the Dirac measure on θ as $t \to \infty$, μ_θ-a.e.

Consistency says that observing the filtration allows one to narrow in on the parameter θ, in the weak sense of convergence. This is quite different from learning the distribution associated with θ and being able to make predictions as if one knew θ. Example 4 shows that there are representations that are learnable but not consistent. Example 5, below, shows that there are representations that are consistent but not learnable. Therefore consistency and learnability are different notions, neither weaker than the other.

Example 5: A consistent, but not learnable, representation: $\Omega = \Theta = \{0,1\}^\infty$, $\mathcal{B} = \mathcal{F}$, and μ corresponds to a measure representing independent flips of a fair coin (where "heads" is represented as 1 and "tails" as 0). Let μ_θ be the Dirac measure on $\theta = \omega$, and set $\lambda = \mu$.

Note that Example 5 shows that the weak convergence in the definition of consistency allows the observer to place weight 0 on the "true" θ all along the sequence.

[5] Let ϕ denote the product measure $\lambda \times \mu_\theta$, so for $B \in \mathcal{B}$ and $E \in \mathcal{F}$ $\phi(B \times E) = \int_{\theta\in B} \mu_\theta(E) d\lambda(\theta)$. The the posterior referred to is $\phi_\Theta(\cdot|\mathcal{F}_t)$, the marginal ϕ_Θ conditional on $\mathcal{F}_t$.

5 Additional Remarks

Our analysis may be used to identify the natural models that an econometrician or a statistician could learn by observing a stochastic process. The arbitrage pricing theory (APT) model is an example in which the factor structure underlying a stochastic process of security prices is drawn from the data.

The representation identified in this paper may also be useful in assessing the value of information obtained from observation of a stochastic process. The representation tells one in advance what patterns the observer is likely to learn and with what probabilities. This is precisely the information an observer needs in order to compute the expected benefit of observing the process.

It would be useful to obtain additional results connecting our representations to specific attractive alternatives. For example, Theorem 3 provides a characterization of representations that are learnable and sufficient for prediction, and one might want to refine this class to representations with the least redundancy, e.g. where different parameter values imply different asymptotic distributions. This would mean adding consistency to the conditions of a desired representation.

Other types of connections between learnability and representations can also be examined. A new paper by Sandroni and Smorodinsky (1998) explores conditions on a set of measures that are sufficient for there to exist a measure which merges with all the measures in the set.

A recent paper by Al-Najjar (1996) considers continuum economies where agents may be indexed by the interval $[0,1]$. Associated with each agent is a random variable representing some action or characteristic. Al-Najjar considers decomposing the uncertainty in such an economy into 'aggregate states' and 'micro-states', where an observer of a random sample of agents may learn the correlation pattern in the aggregate states, but not the micro states. His aggregate states bear an intuitive similarity to our parameters θ. Al-Najjar's work differs in the extent to which states are broken down. His decomposition is driven by independent residuals (conditional on the aggregate states), while ours driven by learning and is thus based on the asymptotic reverse-mixing notion. Nevertheless, there may be interesting connections between decompositions in cross-sectional and time series models.

Part of our interest in the problem studied here arose from thinking about

Bayesian updating in the context of a game where a player is faced with an opponent playing an unkown strategy. If, for instance, players choose finite automata to play for them then the resulting process will be asymptotically reverse-mixing and so our results would apply. In this sense a learnable representation that is sufficient for prediction provides an alternative endogenous definition of types to the exogenous notions already in the literature (e.g., Harsanyi (1967-68) and Mertens and Zamir (1985)).[6] This perspective can be explored in more detail.

Finally, one may consider roughly the reverse of the question we have analyzed: that is, given types (which may incorporate some posterior beliefs about such things as patterns), one may examine conditions under which there are well-defined priors, or even common priors, consistent with the types. Recent papers by Samet (1996ab) address such questions.

[6]See Nyarko (1996) for some discussion of equivalent reformulations of type space, and Bergin (1992) for more discussion of posteriors and type distributions in the context of games.

References

Al-Najjar, N. [1996], "Aggregation and the Law of Large Numbers in Economies with a Continuum of Agents," CMSEMS wp no. 1160, Northwestern University.

Al-Najjar, N. [1998], "Factor Analysis and Arbitrage Pricing in Large Asset Economies," *Journal of Economic Theory*, Vol. 78, pp. 231-262.

Aumann, R.J., and M.B. Maschler [1995], *Repeated Games with Incomplete Information*, MIT Press: Cambridge.

Bergin, J. [1992], "Player Type Distributions as State Variables and Information Revelation in Zero Sum Games with Discounting," *Mathematics of Operations Research*, Vol. 17, pp. 640-656.

Billingsley, P. [1968], *Convergence of Probability Measures*, Wiley: New York.

Billingsley, P. [1979], *Probability and Measure*, Wiley: New York, (third edition).

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

Blackwell, D. and L. Dubins [1975], "On Existence and Non-existence of Proper, Regular Conditional Distributions," *Annals of Probability*, Vol. 3, pp. 741–752.

Delacherie, C. and P. A. Meyer [1978], *Probabilities and Potential*, North Holland: Amsterdam.

Diaconis, P. and D. Freedman [1986], "On the Consistency of Bayes Estimates," *Annals of Statistics*, Vol. 11, pp. 1-26.

Fudenberg, D. and D. Levine [1995], "Conditional Universal Consistency," mimeo.

Harsanyi, J. [1967], "Games with Incomplete Information Played by Bayesian Players, Parts I, II, and III," *Management Science*, Vol. 14, pp. 159-182, 320-334, 486-502.

Jackson, M. and E. Kalai [1997], "Social Learning in Recurring Games," *Games and Economic Behavior*, Vol. 21, pp. 102-134.

Kalai, E. and E. Lehrer [1993], "Rational Learning Leads to Nash Equilibrium," *Econometrica*, Vol. 61, pp. 1019-1045.

Kalai, E. and E. Lehrer [1994], "Weak and Strong Merging of Opinions," *Journal of Mathematical Economics*, Vol. 23, pp. 73–86.

Kandori, M., G. Mailath, and R. Rob [1993], "Learning, Mutation, and Long Run Equilibria," *Econometrica*, Vol. 61, pp. 27–56.

Lehrer, E. and R. Smorodinsky [1997], "Repeated Large Games with Incomplete Information," *Games and Economic Behavior*, Vol. 18, pp. 116-134.

Lehrer, E. and R. Smorodinsky [1997b], "Learning and Merging." In Ferguson, Shapley and McQueen, *Statistics, Probability, and Game Theory: Papers in Honor of David Blackwell*, IMS Lecture Notes Monograph Series, Vol. 30.

Marimon, R. [1997], "Learning from Learning in Economics." In D. Kreps and K. Wallis, *Advances in Economics and Econometrics: Theory and Applications, 7th World Congress of the Econometric Society, Volume 1*, Econometric Society Monographs, Cambridge University Press.

Mertens, J-F. and S. Zamir [1985], "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, Vol. 14, pp. 1-29.

Nyarko, Y. [1996], "Bayesian Learning and Convergence to Nash Equilibria without Common Priors," mimeo NYU.

Rothschild, M. [1974], "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, Vol. 9, pp. 185-202.

Samet, D. [1996a], "Looking Backwards, Looking Inwards: Priors and Introspection," mimeo.

Samet, D. [1996b], "Common Priors and Markov Chains," mimeo.

Sandroni, A. and R. Smorodinsky [1998], "Rates of Merging and Learnability," mimeo.

Sargent, T. [1993], *Bounded Rationality in Macroeconomics*, Oxford: Oxford University Press.

Sonsino, D. [1997], "Learning to Learn, Pattern Recognition and Nash Equilibrium," *Games and Economic Behavior*, Vol. 18, pp. 286–331.

Smorodinsky, M. [1971], *Ergodic Theory, Entropy*, Lecture Notes in Mathematics edited by A. Dold and B. Eckmann, Springer Verlag: Berlin.

Stinchcombe, M. [1990], "Bayesian Information Topologies," *Journal of Mathematical Economics*, Vol. 19, pp. 233-253.

Zellner, A. [1971], *An Introduction to Bayesian Inference in Econometrics*, J. Wiley: New York.

Appendix: Proofs

We begin by showing that merging according to the definition of Kalai and Lehrer (1994) is equivalent to the definition in this paper.

Lemma 1: Consider $\nu, \nu' \in \Delta$. If $\forall\ \epsilon > 0$ and ν -a.e. $\omega\ \exists T = T(\epsilon, \omega)$ such that for all $t \geq T$

$$\max_{A \in \mathcal{G}_{t+1}} |\tilde{\nu}(A|\mathcal{F}_t) - \nu(A|\mathcal{F}_t)| < \epsilon,$$

then, $\forall\ \epsilon > 0$, ℓ, and ν -a.e. $\omega\ \exists T$ such that for all $t \geq T$

$$\sup_{n \geq t, A \in \mathcal{G}_n^{n+\ell}} |\tilde{\nu}(A|\mathcal{F}_t) - \nu(A|\mathcal{F}_t)| < \epsilon.$$

Proof:[7] In the following, we make use of a lemma from Kalai and Lehrer (1994), which is stated below.

Lemma (Kalai and Lehrer): Let g_t be a sequence of measurable functions that converge ν-a.e. to $g \neq 0$. For every $\epsilon > 0$ there is a time t_0 such that $\nu(\{\omega \mid \nu(C_t \mid \mathcal{F}_{t-1})(\omega) > \epsilon \text{ for at least one } t \geq t_0\}) < \epsilon$ where

$$C_t = \left\{\omega : |\frac{g_s(\omega)}{g(\omega)} - 1 \mid > \epsilon \text{ for some } s \geq t\right\}.$$

We apply the lemma to the following sequence of indicator functions

$$g_t(\omega) = I_{\{\omega \mid \forall A \in \mathcal{G}_{t+1} |\tilde{\nu}(A|\mathcal{F}_t) - \nu(A|\mathcal{F}_t)| < \epsilon\}}$$

In this case, $g_t(\omega) \to 1$ for μ-a.e. ω and C_t is

$$C_t = \left\{\omega \mid \sup_{n \geq t, A \in \mathcal{G}_{n+1}} \mid \tilde{\nu}(A \mid \mathcal{F}_n) - \nu(A \mid \mathcal{F}_n) \mid > \epsilon\right\}$$

and its complement, denoted C_t^c, is

$$C_t^c = \left\{\omega \mid \sup_{n \geq t, A \in \mathcal{G}_{n+1}} \mid \tilde{\nu}(A \mid \mathcal{F}_n) - \nu(A \mid \mathcal{F}_n) \mid \leq \epsilon\right\}.$$

[7] We thank Ehud Lehrer for this proof.

By the lemma above, there exists t_0 such that $t > t_0$ implies

$$\nu\left(\{\omega \mid \nu(C_{t+1} \mid \mathcal{F}_t) < \epsilon \ \forall t > t_0\}\right) > 1 - \epsilon$$

This implies

$$\nu\left(\left\{\omega \mid \nu(C_{t+1}^c \mid \mathcal{F}_t) > 1 - \epsilon \ \forall t > t_0\right\}\right) > 1 - \epsilon$$

Let $D_0 = \{\omega \mid \tilde{\nu}(C_{t+1}^c \mid \mathcal{F}_t) > 1 - \epsilon \text{ for all } t > t_0\}$. By the assumption of Lemma 1, $\exists T(\omega, \epsilon)$ such that $t > T$ implies that for all $n \geq t$ and $A \in \mathcal{G}_{n+1}$

$$\begin{aligned}
&|\tilde{\nu}(A \mid \mathcal{F}_t) - \nu(A \mid \mathcal{F}_t)| \\
&\leq \tilde{\nu}(D_0)\nu(C_{t+1}^c \mid D_0) \ | \tilde{\nu}(A \mid \mathcal{F}_t \cap D_0 \cap C_{t+1}^c) - \nu(A \mid \mathcal{F}_t \cap D_0 \cap C_{t+1}^c) | \\
&+\nu(D_0)\,\nu(C_{t+1} \mid D_0) \ | \tilde{\nu}(A \mid \mathcal{F}_t \cap D_0 \cap C_{t+1}) - \nu(A \mid \mathcal{F}_t \cap D_0 \cap C_{t+1}) | \\
&+\nu(D_0^c)\,|\tilde{\nu}(A \mid \mathcal{F}_t \cap D_0) - \nu(A \mid \mathcal{F}_t \cap D_0^c)| \\
&\leq (1-\epsilon)(1-\epsilon)\epsilon + (1-\epsilon)(\epsilon) + (\epsilon) \\
&\leq 3\epsilon
\end{aligned}$$

Then, by Remark 5 in Lehrer and Smorodinsky (1997b), we can add an arbitrary ℓ to obtain the desired conclusion. ∎

Next, as promised in the text, the fact that $\overline{\mathcal{F}}^{\text{tail}}$ is a representation is established by noting that $\mu_\omega^{\mathcal{F}^{\text{tail}}}$ can be chosen to be a probability measure μ-a.e., according to the following theorem.

Given $\mathcal{H}$ which is a sub σ-field of $\mathcal{F}$, let $AT_{\mathcal{H}}(\omega) = \cap_{\{A \in \mathcal{H} \mid \omega \in A\}} A$. (See definition 3.2.4 of Stinchcombe (1990).)

Theorem A: [Stinchcombe (1990)[8]] If $\mathcal{H}$ is a sub σ-field of $\mathcal{F}$, then there exist versions of $E(1_A|\mathcal{H})$ for all $A \in \mathcal{F}$ such that $\mu_\omega^{\mathcal{H}}(A) \equiv E(1_A|\mathcal{H})$ is a probability measure μ-a.e.. Furthermore, if $\mathcal{H}$ is countably generated then $\mu_\omega^{\mathcal{H}}(AT_{\mathcal{H}}(\omega)) = 1$ for μ-a.e. ω.

Let us now turn to proving Theorems 1-3.

[8]Stinchcombe assumes Blackwell measurability of the underlying probability space - while we have not made that assumption here. Consult Dellacherie and Meyer (1978) pages III-70, 71, 79, to see that this result holds in our setting.

Useful results are that if $\mathcal{H}$ and $\mathcal{H}'$ are equivalent sub σ-fields of $\mathcal{F}$ (that is, for every $A \in \mathcal{H}$ there exists $B \in \mathcal{H}'$ with $\mu(A \Delta B) = 0$ and vice versa), then $\mu_\omega^{\mathcal{H}} = \mu_\omega^{\mathcal{H}'}$, μ-a.e., and that for every sub σ-field of $\mathcal{F}$, there exists an equivalent countably generated sub σ-field of $\mathcal{F}$ (see Stinchcombe (1990) section 2.4 and Lemma 3.2.2). These two facts imply that we can find a countably generated sub σ-field of $\mathcal{F}$, $\mathcal{H}$ such that $\mu_\omega^{\mathcal{H}} = \mu_\omega^{\mathcal{F}^{\text{tail}}}$ and $\mu_\omega^{\mathcal{H}}(AT_{\mathcal{H}}(\omega)) = 1$ for μ-a.e. ω. (We offer this construction since the tail-field is not necessarily countably generated. See Blackwell and Dubins (1975).)

The following Lemmas are useful.

Lemma 2: For any $\mathcal{H}$, sub σ-field of $\mathcal{F}$, $AT_{\mathcal{H}}(\omega_1) = AT_{\mathcal{H}}(\omega_2)$ implies $\mu_{\omega_1}^{\mathcal{H}} = \mu_{\omega_2}^{\mathcal{H}}$.

Proof: Look at arbitrary $\omega_1, \omega_2 \in \Omega$ such that $AT_{\mathcal{H}}(\omega_1) = AT_{\mathcal{H}}(\omega_2)$. For an arbitrary set $B \in \mathcal{F}$, denote $\alpha = E(1_B \mid \mathcal{H})(\omega_1)$ and $H = \{\omega \mid E(1_B \mid \mathcal{H})(\omega) = \alpha\}$. Obviously, $H \in \mathcal{H}$. Definitely $\omega_1 \in H$ and so $AT_{\mathcal{H}}(\omega_1) \subset H$. Since $\omega_2 \in AT_{\mathcal{H}}(\omega_2) = AT_{\mathcal{H}}(\omega_1) \subset H$, it follows that $E(1_B \mid \mathcal{H})(\omega_2) = \alpha$. As B was chosen arbitrarily, it follows that $E(1_B \mid \mathcal{H})(\omega_2) = E(1_B \mid \mathcal{H})(\omega_1)$ for any B and so $\mu_{\omega_1}^{\mathcal{H}} = \mu_{\omega_2}^{\mathcal{H}}$. ∎

Lemma 3: Consider $\mathcal{H}$, a countably generated sub σ-field of $\mathcal{F}$, and let $A(\omega) = \{\omega' | \mu_\omega^{\mathcal{H}} = \mu_{\omega'}^{\mathcal{H}}\}$. There exists X with $\mu(X) = 1$ such that $\mu_\omega^{\mathcal{H}}(A(\omega)) = 1$ for all $\omega \in X$.

Proof: Note that $A(\omega) \in \mathcal{F}$, since $\mathcal{F}$ is countably generated. Note also that $AT_{\mathcal{H}}(\omega) \subset A(\omega)$. (This follows from Lemma 2 since $\omega' \in AT_{\mathcal{H}}(\omega)$ implies $AT_{\mathcal{H}}(\omega') = AT_{\mathcal{H}}(\omega)$, and Lemma 2 then implies that $\mu_\omega^{\mathcal{H}} = \mu_{\omega'}^{\mathcal{H}}$). By Theorem A, $\mu_\omega^{\mathcal{H}}(AT_{\mathcal{H}}(\omega)) = 1$ μ-a.e., which implies that $\mu_\omega^{\mathcal{H}}(A(\omega)) = 1$ μ-a.e.. ∎

Proof of Theorem 1: Since $\mu_\omega^{\mathcal{F}^{\text{tail}}}$ has a trivial tail for μ-a.e. ω,[9] the theorem follows from the fact that for any measure ν, having a trivial tail implies that ν is K-mixing. (See page 39, second point in the proof of theorem 7.9, of Smorodinsky (1971).) In our context, this implies that for μ-a.e. ω and any t and $C \in \mathcal{F}_t$

$$\lim_n \sup_{A \in \mathcal{G}_n^\infty} |\mu_\omega^{\mathcal{F}^{\text{tail}}}(A \cap C) - \mu_\omega^{\mathcal{F}^{\text{tail}}}(A)\mu_\omega^{\mathcal{F}^{\text{tail}}}(C)| = 0.$$

[9] ν has a trivial tail if $\nu(A) \in \{0, 1\}$ for all $A \in \mathcal{F}^{\text{tail}}$.

which divided through by $\mu_\omega^{\mathcal{F}^{\text{tail}}}(C)$, (when $\mu_\omega^{\mathcal{F}^{\text{tail}}}(C) > 0$) provides the desired conclusion. ∎

Proof of Theorem 2: First, we show that if μ is asymptotically reverse-mixing, then $\overline{\mathcal{F}}^{\text{tail}}$ is learnable.

Suppose the contrary, so there exists C with $\mu(C) > 0$ such that for each $\omega \in C$ there exists B_ω and ϵ_ω such that $\mu_\omega^{\mathcal{F}^{\text{tail}}}(B_\omega) > 0$ and for every $\omega' \in B_\omega$ there are infinitely many t such that

$$\max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega') - \mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega') \right| > \epsilon_\omega.$$

Thus, there exists C' with $\mu(C') > 0$ and a common ϵ such that for each $\omega \in C'$ there exists B_ω such that $\mu_\omega^{\mathcal{F}^{\text{tail}}}(B_\omega) > 0$ and for every $\omega' \in B_\omega$ there are infinitely many t such that

$$\max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega') - \mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega') \right| > \epsilon. \tag{1}$$

By Lemma 3,[10] for μ-a.e. ω and $\mu_\omega^{\mathcal{F}^{\text{tail}}}$-a.e. ω', we can change $\mu_\omega^{\mathcal{F}^{\text{tail}}}$ to read $\mu_{\omega'}^{\mathcal{F}^{\text{tail}}}$, and so, without loss of generality, we rewrite the above inequality as

$$\max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega') - \mu_{\omega'}^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega') \right| > \epsilon. \tag{2}$$

Let K be the set of ω such that for infinitely many t

$$\max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega) - \mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega) \right| > \epsilon.$$

Notice that K is in $\mathcal{F}$ and $\cup_{\omega \in C'} B_\omega \subset K$ and so $\mu(K) > 0$ (since $\mu(C') > 0$ and $\mu_\omega^{\mathcal{F}^{\text{tail}}}(K) > 0$ for each $\omega \in C'$).

As $\mathcal{G}_n^\infty \searrow_{n \to -\infty} \mathcal{F}^{\text{tail}}$, by the convergence theorem for reversed martingales (see, e.g., Theorem 35.9 in Billingsley (1986), third edition) for any t and $A \in \mathcal{G}_{t+1}$ it follows that $\mu_\omega^{\mathcal{G}_n^\infty}(A|\mathcal{F}_t)(\omega)$ converges to $\mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega)$ as $n \to \infty$, μ-a.e.. Thus, for μ-a.e. $\omega \in K$ there are infinitely many t such that

$$\overline{\lim}_n \max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega) - \mu_\omega^{\mathcal{G}_n^\infty}(A|\mathcal{F}_t)(\omega) \right| > \frac{\epsilon}{2}.$$

[10] To be careful, note that (1) holds for μ-a.e. ω relative to $\mathcal{H}$ which is equivalent to $\mathcal{F}^{\text{tail}}$ and for which $\mu_\omega^{\mathcal{H}} = \mu_\omega^{\mathcal{F}^{\text{tail}}}$ for μ-a.e. ω. Thus, (2) follows for μ-a.e. $\omega \in C'$ relative to $\mu_\omega^{\mathcal{H}}$, and therefore also for μ-a.e. $\omega \in C'$ relative to $\mu_\omega^{\mathcal{F}^{\text{tail}}}$.

By the martingale convergence theorem, for μ-a.e. $\omega \in K$ there are infinitely many t such that

$$\overline{\lim}_n \overline{\lim}_\ell \max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega) - \mu(A|\mathcal{F}_t \vee \mathcal{G}_n^{n+\ell})(\omega) \right| > \frac{\epsilon}{4}. \tag{3}$$

This is a contradiction.

Second, let us show the converse: if $\overline{\mathcal{F}}^{\text{tail}}$ is learnable, then μ is asymptotically reverse-mixing.

Using an argument similar to that proceeding (2), learnability implies that for μ-a.e. ω

$$\overline{\lim}_t \max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega) - \mu_\omega^{\mathcal{F}^{\text{tail}}}(A|\mathcal{F}_t)(\omega) \right| = 0. \tag{4}$$

Again, noting that $\mathcal{G}_n^\infty \searrow_{n=-\infty} \mathcal{F}^{\text{tail}}$, it follows that for μ-a.e. ω

$$\overline{\lim}_t \overline{\lim}_n \max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega) - \mu_\omega^{\mathcal{G}_n^\infty}(A|\mathcal{F}_t)(\omega) \right| = 0.$$

Finally, by the martingale convergence theorem, for μ-a.e. ω

$$\overline{\lim}_t \overline{\lim}_n \overline{\lim}_\ell \max_{A \in \mathcal{G}^{t+1}} \left| \mu(A|\mathcal{F}_t)(\omega) - \mu(A|\mathcal{F}_t \vee \mathcal{G}_n^{n+\ell})(\omega) \right| = 0,$$

which is the desired conclusion. ∎

Proof of Theorem 3: Fix $\mathcal{H}$ which is countably generated and such that $\mu_\omega^{\mathcal{H}} = \mu_\omega^{\mathcal{F}^{\text{tail}}}$ for μ-a.e. ω.

Let us say that $\overline{\Theta}$ is asymptotically like $\mathcal{H}$ if for λ-a.e. θ and for every ℓ

$$\lim_K \sup_{t,n:t+n \geq K} \max_{A \in \mathcal{G}_{t+n}^{t+n+\ell}} \left| \mu_\omega^{\mathcal{H}}(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_t)(\omega) \right| = 0 \tag{5}$$

for μ_θ-a.e. ω.

Lemma 4: $\overline{\Theta}$ is asymptotically like $\overline{\mathcal{H}}$ if and only if it is asymptotically like $\overline{\mathcal{F}}^{\text{tail}}$.

Proof of Lemma 4: Consider $\overline{\Theta}$ which is asymptotically like $\overline{\mathcal{H}}$. Let $E_\theta^{\mathcal{H}}$ be the set for which (5) holds and A to be the set such that $\mu(A) = 1$ and $\mu_\omega^{\mathcal{H}} = \mu_\omega^{\mathcal{F}^{\text{tail}}}$ for $\omega \in A$.

Observation 1: If $\overline{\Theta}$ is a representation and $\mu(A) = 1$. then $\mu_\theta(A) = 1$ for λ-a.e. θ.

It follows from Observation 1 that $\mu_\theta(E_\theta^{\mathcal{H}} \cap A) = 1$ for λ-a.e. θ, which establishes that $\overline{\Theta}$ is asymptotically like $\overline{\mathcal{F}}^{\text{tail}}$. The converse follows from the same argument with respect to $\mathcal{F}^{\text{tail}}$. ∎

Lemma 5: $\overline{\Theta}$ is asymptotically like $\overline{\mathcal{H}}$ if and only if for λ-a.e. θ, any ℓ and μ_θ-a.e. ω

$$\overline{\lim}_t \sup_{n \geq t, A \in \mathcal{G}_n^{n+\ell}} |\mu_\omega^{\mathcal{H}}(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_t)(\omega)| = 0 \tag{6}$$

and for λ-a.e. θ, any fixed t' and ℓ, and μ_θ-a.e. ω

$$\overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu_\omega^{\mathcal{H}}(A|\mathcal{F}_{t'})(\omega) - \mu_\theta(A|\mathcal{F}_{t'})(\omega)| = 0. \tag{7}$$

Proof of Lemma 5: It follows directly that (5) implies (6) and (7).

To see the converse, pick $\epsilon > 0$ and any θ such that both (6) and (7)hold (which is a set of λ measure 1). Pick any ℓ and any ω such that both (6) and (7)hold (which is a set of μ_θ measure 1). By (6) there exists K such that if $t \geq K$, then

$$\sup_{n \geq t} \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu_\omega^{\mathcal{H}}(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_t)(\omega)| < \epsilon.$$

By (7), for every t' there exists $K_{t'}$ such that if $n \geq K_{t'}$, then

$$\max_{A \in \mathcal{G}_{t+n}^{t+n+\ell}} |\mu_\omega^{\mathcal{H}}(A|\mathcal{F}_{t'})(\omega) - \mu_\theta(A|\mathcal{F}_{t'})(\omega)| < \epsilon.$$

Let $\overline{K} = K + \max_{t' < K} K_{t'}$ (which is well defined given the finite set of such t'). It then follows that for any t and n with $t + n \geq \overline{K}$

$$\max_{A \in \mathcal{G}_{t+n}^{t+n+\ell}} |\mu_\omega^{\mathcal{H}}(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_t)(\omega)| < \epsilon,$$

which establishes (5). ∎

Lemma 6: If μ is asymptotically reverse-mixing and $\overline{\Theta}$ is learnable and is sufficient for prediction, then $\overline{\Theta}$ is asymptotically like $\overline{\mathcal{H}}$.

Proof of Lemma 6: Assume that μ is asymptotically reverse-mixing and $\overline{\Theta}$ is learnable and is sufficient for prediction. Following Lemmas 4 and 5, it is sufficient to show that (6) and (7) hold.

We first prove that (7) holds. Since $\overline{\Theta}$ is learnable it follows that for λ-a.e. θ and all ℓ

$$\overline{\lim}_t \sup_{n \geq t, A \in \mathcal{G}_n^{n+\ell}} |\mu_\theta(A|\mathcal{F}_t)(\omega) - \mu(A|\mathcal{F}_t)(\omega)| = 0, \ \ \mu_\theta - \text{a.e.}. \tag{8}$$

Since $\overline{\Theta}$ is sufficient for prediction, it follows that for λ-a.e. θ and all t and ℓ

$$\overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu_\theta(A|\mathcal{F}_t)(\omega) - \mu_\theta(A)(\omega)| = 0 \ \ \mu_\theta - \text{a.e.}. \tag{9}$$

Note that (9) implies that for λ-a.e. θ and all t, t' and ℓ,

$$\overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu_\theta(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_{t'})(\omega)| = 0 \ \ \mu_\theta - \text{a.e.}. \tag{10}$$

Combining (8) and (10) it follows that for λ-a.e. θ and all ℓ and t'

$$\overline{\lim}_t \overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu(A|\mathcal{F}_t)(\omega) - \mu_\theta(A|\mathcal{F}_{t'})(\omega)| = 0 \ \ \mu_\theta - \text{a.e.}. \tag{11}$$

Similarly, we can show that for all $\omega' \in D$, where $\mu(D) = 1$, and all ℓ and t'

$$\overline{\lim}_t \overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu(A|\mathcal{F}_t)(\omega) - \mu_{\omega'}^{\mathcal{H}}(A|\mathcal{F}_{t'})(\omega)| = 0 \tag{12}$$

for $\omega \in B(\omega')$ where $\mu_{\omega'}^{\mathcal{H}}(B(\omega')) = 1$.

Next, we show that $\mu(Y) = 1$ where $Y = \{\omega' \mid \omega' \in B(\omega')\}$. To see this, suppose to the contrary that $\mu(Y) < 1$.[11] Therefore, by the definition of representation there exists a set S' such that $\mu(S') > 0$ and $\mu_{\omega'}^{\mathcal{H}}(Y) < 1$ for all $\omega \in S'$. Find $\omega'' \in S' \cap D \cap X$ where X is from Lemma 3 (these have a non-empty intersection since $\mu(S') > 0$ and $\mu(X) = \mu(D) = 1$). Then $\mu_{\omega''}^{\mathcal{H}}(Y^c) > 0$ (where Y^c is the complement of Y), $\mu_{\omega''}^{\mathcal{H}}(B(\omega'')) = 1$, and $\mu_{\omega''}^{\mathcal{H}}(A(\omega'')) = 1$. Consider $\omega' \in Y^c \cap B(\omega'') \cap A(\omega'')$. This implies that $\omega' \notin B(\omega')$, but also that $\omega' \in B(\omega'')$ and $\mu_{\omega'}^{\mathcal{H}} = \mu_{\omega''}^{\mathcal{H}}$ which imply that $\omega' \in B(\omega')$, which is a contradiction.

Thus, since $\omega' \in B(\omega')$ for almost every ω', it follows from (12) that for all ℓ, t' and all $\omega \in D \cap Y$ (where $\mu(D \cap Y) = 1$)

$$\overline{\lim}_t \overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu(A|\mathcal{F}_t)(\omega) - \mu_\omega^{\mathcal{H}}(A|\mathcal{F}_{t'})(\omega)| = 0. \tag{13}$$

[11] Note that Y is an $\mathcal{F}$-measurable set since it can be written as a countable combination of intersections and unions of sets of the form $\{\omega : |\mu(A|\mathcal{F}_t)(\omega) - \mu_\omega^{\mathcal{H}}(A|\mathcal{F}_{t'})(\omega)| < \frac{1}{k}\}$.

Thus, from (11), (13), and Observation 1 it follows that for λ-a.e. θ, all ℓ, any t', and μ_θ-a.e. ω

$$\overline{\lim}_t \overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu_\theta(A|\mathcal{F}_{t'})(\omega) - \mu_\omega^{\mathcal{H}}(A|\mathcal{F}_{t'})(\omega)| = 0. \tag{14}$$

Since the t in (14) is redundant, this establishes (7).

Next we establish (6). By learnability, and the argument preceeding (13), it follows that for any ℓ and all μ-a.e. ω

$$\overline{\lim}_t \sup_{n \geq t, A \in \mathcal{G}_n^{n+\ell}} |\mu(A|\mathcal{F}_t)(\omega) - \mu_\omega^{\mathcal{H}}(A|\mathcal{F}_t)(\omega)| = 0. \tag{15}$$

This combined with (8) and Observation 1 establishes (6). ∎

Lemma 7: If μ is asymptotically reverse-mixing and $\overline{\Theta}$ is asymptotically like $\overline{\mathcal{H}}$, then $\overline{\Theta}$ is learnable and is sufficient for prediction.

Proof of Lemma 7: First we show learnability. Again, (15) holds by the learnability of $\overline{\mathcal{F}}^{\text{tail}}$. This combined with (6) and Observation 1, establishes the learnability of $\overline{\Theta}$.

Next, we show that $\overline{\Theta}$ is sufficient for prediction. Since $\overline{\mathcal{F}}^{\text{tail}}$ is sufficient for prediction, it follows that for μ-a.e. ω and all t' and ℓ

$$\overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\mu_\omega^{\mathcal{H}}(A|\mathcal{F}_{t'})(\omega) - \mu_\omega^{\mathcal{H}}(A)(\omega)| = 0. \tag{16}$$

(Again the argument preceding (13) is invoked.) Then (16), (7), combined with Observation 1, establish that $\overline{\Theta}$ is sufficient for prediction. ∎