Patterns, Types, and Bayesian Learning*

Matthew O. Jackson Ehud Kalai Rann Smorodinsky

Revised: November 17, 1997

Abstract

Bayesian Statisticians, decision theorists, and game theorists often use Bayesian representatiopns to describe the probability distribution governing the evolution of a stochastic process. Generally, however, one given distribution has infinitely many different Bayesian representations. This paper identifies natural, endogenous representations whose component distributions are learnable and follow patterns. Any given distribution that satisfies an asymptotic mixing condition has a unique, up to an equivalence class, natural Bayesian representation which can be obtained by conditioning on the tail-field of the process. This result follows a parallel to de Finetti's theorem, but with exchangeability weakened to asymptotic mixing which admits many more applications.

*We thank Nabil Al-Najjar, David Gilat, Ehud Lehrer, and Meir Smorodinsky for helpful conversations, and Martin Northway for expositional comments on an earlier draft. Suggestions by Drew Fudenberg and three anonymous referees have led to substantial improvements in the paper. Financial support under NSF grants SBR 9515421 and SBR 9507912 is gratefully acknowledged. Jackson is at HSS 228-77, Caltech, Pasadena CA 91125, (jacksonm@hss.caltech.edu); Kalai is at MEDS, Kellogg Graduate School of Management Northwestern University, Evanston, Il 60208, (kalai@nwu.edu); and Smorodinsky is at Industrial Engineering, Technion, Haifa 32000, Israel (rann@ie.technion.ac.il).

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 distributions μ_θ. In the second stage the selected distribution 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 distribution 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 to identify the 'patterns' that are present in the process. Second, we want the representation to be coarse enough so that it does not contain details that are not learnable no matter how long the process is observed. As it turns out, such natural representations exist, and moreover all such representations are equivalent (in a sense to be made precise). Furthermore, such a natural representation can be obtained by conditioning the original measure 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 $1/2$. In fact, θ is chosen according to a uniform distribution over $[0,1]$. So we

[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.

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. Notice that there are several convex combinations of component measures that we could use to represent μ. 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 degenerate representation of μ as μ (so this reprentation is $\mu = \mu$). From our perspective this is too coarse a representation since it does 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 provide patterns. Third, there is a representation of μ as a convex combination of Dirac measures where each Dirac measure gives weight one to some infinite sequence of heads and tails. This representation is $\mu = \int_\Omega \delta_\omega d\mu(\omega)$ where each ω corresponds to a single infinite sequence of 'H' and 'T's and δ_ω is a degenerate distribution with weight one on the sequence ω. From our perspective this other 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 puts the two ideas of patterns and learnability together and applies them to the measure μ, then one recovers precisely the representation of μ as the coins.

Let us be a bit more explicit about the definitions of patterns, 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.

Patterns. There are obviously many ways to define a notion of pattern. [2] In this paper, we think of a pattern as a measure whose knowledge enables a forecaster to make predictions that would not be influenced by any additional knowledge of past history of the process. A component measure follows a pattern if the unconditional probabilities of late events are arbitrarily close to their probabilities when additionally conditioned on initial segments of the process. A representation provides patterns if each of its component measures follows a pattern.

[2] Recent definitions of pattern, quite different from ours, can be found in Sonsino (1997), Fudenberg and Levine (1995), and Sargent (1993). Marimon (1997) discusses the uses of the terms 'pattern' and 'pattern recognition' in the literature.

In Example 1 above, the representation of μ as coins with known parameters provides patterns. 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 trivially provides patterns. In contrast, the coarsest representation of μ as μ does not provide patterns. The unconditional probability of H at time t is $1/2$, while, for instance, conditional on an initial long segment which is rich in H's the probability of H at time t is greater than $1/2$.

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 distribution 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 coarse representation of μ as μ itself is trivially learnable. The representation of μ as coins with known parameters is also 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. In contrast, the fully refined representation of μ by Dirac measures which each place probability one on some particular infinite sequence of H and T's 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 providing patterns 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 respresentations look like. Our main results show for a certain class of mixing processes that there always exists a representation that is both learnable and provides patterns, and any such representation is equivalent (in a sense to be defined) to the representation one obtains by conditioning on the tail field. Thus, the conditions of learnability and following patterns

together identify representations corresponding to the tail field.

Relationship to de Finetti's Theorem. The celebrated de Finetti Theorem suggests an example of a representation by learnable patterns. Illustrated in the context of repeated coin tosses where the parameter of the coin is unknown, de Finetti considers situations where the probability assigned to every initial finite sequence of H's and T's, 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 coin tosses, where the parameter of the coin is random and will correspond to the long run frequency distribution of H's versus T's. Thus, in the language of this paper, he represents an exchangeable distribution by a convex combination of learnable distributions that follow patterns.

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

While exchangeability is too strong for most economic applications, asymptotic mixing is substantially weaker and covers a significant number of applications. While asymptotic mixing permits long run effects being generated by random early events, it precludes the possibility of having lasting effects generated by arbitrarily late events. For example, any (infinite) convex combination of Markov chains is asymptotically mixing.

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.

As we discussed in the introduction, our interest is to identify "natural" representations from a specific point of view. An observer of the filtration $\{\mathcal{F}_t\}_{t=1}^\infty$ may update μ in a Bayesian manner over time. We would like to have the representation precisely capture what the observer will learn over time. That is, we would like to be able to say that what the observer can learn from the filtration is essentially equivalent to simply being told which θ has been drawn. In most cases of interest, the observer will learn more than the trivial representation (where Θ is a singleton) and less than the complete representation (where $\Theta = \Omega$).

We now provide precise definitions for following patterns and learnability.

Patterns

Definition: A measure $\tilde{\mu}$ follows a pattern if for all t

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

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 $(\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ provides patterns if μ_θ follows a pattern for λ-a.e. $\theta \in \Theta$.

In order to see the intuition behind the above definitions of patterns (and especially the role of n in the above definitions), 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. Her forecast regarding the next period state may be incorrect, 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) means that one knows the patterns of μ.

Learnability

Learnability is made precise by means of the notion of merging of measures originated by Blackwell and Dubins (1962). We use a weaker definition 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 (1995)).

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

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

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 $\max_{n, A \in \mathcal{G}_{t+n}^{t+n+\ell}}$ 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.

We use the notion of merging to formalize 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 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 is the sense in which the representation is "learned." This is different from requiring that an observer come to identify θ, and we illustrate the differences in Section 4.

Asymptotic Mixing

The following example shows that one cannot have a theory of learnability that applies to all measures. We restrict our attention to measures μ that are asymptotically mixing (to be defined shortly). This restriction is motivated by the following example of a measure which continually brings in new, yet unlearned information.

Example 2: Let $\Omega_t = \{0, 1\}$. Consider μ generated as follows: Partition the set of periods $\mathbb{N} = \cup_{i=1}^{\infty} N_i$, 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 $\mathbb{N} \setminus N_1$ 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 $\Theta = [0, 1]^{\mathbb{N}}$ 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 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 (with asymptotic density greater than zero) 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 provides patterns must make predictions as if one knew the sequence of coins and therefore will fail to be learnable in a similar

manner.

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. Below, we define a notion of asymptotic mixing in order to restrict attention to "non-chaotic" measures. This, as we shall see, guarantees that the patterns identifiable from the filtration are learnable.

Let $A_i^k(\omega)$ denote the atom in $\mathcal{G}_i^k$ containing ω, so

$$A_i^k(\omega) = \cap_{\{A \in \mathcal{G}_i^k : \omega \in A\}} A.$$

Definition: Given $\delta > 0$, ν is δ-mixing if for ν-a.e. ω, all ℓ, and all $A \in \mathcal{G}$ [1]

$$\overline{\lim}_n \left| \frac{\mu(A_n^{n+\ell}(\omega) | \cap A)}{\mu(A_n^{n+\ell}(\omega))} - 1 \right| < \delta.$$

Definition: μ is asymptotically mixing if for any $\delta > 0$ and μ-a.e. ω there exists T such that for any $t \geq T$ $\mu(\cdot|\mathcal{F}_t)$ is δ-mixing. [3]

The condition of δ-mixing guarantees that short run events do not have significant effects on the infinite future of the process. In a sense, initial realizations and far future realizations are approximately independent. Asymptotic mixing, on the other hand, allows for lasting initial effects and dependence, but requires that eventually (conditional on sufficient observation of the initial realization of the process) short run events be 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 several 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, and hence the name 'asymptotic' and the role of T in the definition. Second, we do not need the mixing condition to hold exactly at any time, but only approximately, and hence the role of δ in the definition. Third, the condition is stated in terms of

[3] For a given t, interpret $\mathcal{G}_1$ in the definition of δ-mixing to be $\mathcal{G}_{t+1}$.

a ratio of probabilities while mixing conditions are usually stated in terms of differences. This ensures that far-off events are approximately independent of nearby events, and vice versa, whereas a condition such as ϕ-mixing lacks this symmetry. While our results only require a weaker definition in terms of differences (see the proof of Theorem 2), the above condition is easier to interpret.

The μ described in Example 2 is not asymptotically mixing. For any T one can find a date $t > T$ where a new coin is brought in and so conditioning on additional observations can significantly impact the forecasts of some far off events.

Equivalent Representations

Before proceeding to state the main results, we provide one more definition. Given that our definitions of patterns and learnability are asymptotically based, there can turn out to be many representations which differ in the finite horizon behavior that their component measures describe, but which are all learnable and provide patterns. A simple example illustrates this point and motivates the definition of equivalence to follow.

Example 3: 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 provides patterns, but the same is also true of the trivial representation of μ as itself. These two representations are equivalent as defined below. First, some auxillary definitions.

Definition: The asymptotic distance between $\nu \in \Delta$ and $\nu' \in \Delta$, denoted $d(\nu, \nu')$, is defined by

$$d(\nu, \nu') = \sup_\ell \overline{\lim}_n \max_{A \in \mathcal{G}_n^{n+\ell}} |\nu(A) - \nu'(A)|.$$

Thus, two measures have asymptotic distance 0 if the disagreement in the probability that they assign to far off events is asymptotically 0, regardless of the length of the cylinder that the events fall in.

Given a representation $\overline{\Theta} = (\Theta, \mathcal{B}, \lambda, (\mu_\theta))$, define $f_{\overline{\Theta}} : \Delta \times (0,1) \to \Theta$ by

$$f_{\overline{\Theta}}(\nu, \epsilon) = \{\theta \mid d(\nu, \mu_\theta) < \epsilon\}.$$

The set $f_{\overline{\Theta}}(\nu, \epsilon)$ captures the set of parameters θ that map (asymptotically) into measures near some ν.

Definition: Two representations $\overline{\Theta}_1$ and $\overline{\Theta}_2$ are asymptotically equivalent [4], denoted $\overline{\Theta}_1 \sim \overline{\Theta}_2$, if $\lambda_1(f_{\overline{\Theta}_1}(\nu, \epsilon)) = \lambda_2(f_{\overline{\Theta}_2}(\nu, \epsilon))$ for all $\nu \in \Delta$ and $\epsilon > 0$.[5]

The essence of the equivalence relationship is that one does not care about the particular labels θ or θ', to the extent that the associated measures μ_θ and $\mu_{\theta'}$ lead to the same (asymptotic) predictions. Two representations are equivalent if they assign the same probabilities to ending up close to each ν. The following example shows that the above definition treats as equivalent representations which are trivial relabelings of each other.

Example 4: Consider any μ. A valid representation is to have $\Theta = \{\theta_1, \theta_2\}$ and $\mu_{\theta_1} = \mu_{\theta_2} = \mu$. One may interpret this representation as having nature flip a coin to choose between using μ_{θ_1} or μ_{θ_2}, even though these are identical measures. Another representation is to have $\Theta = \{\theta_0\}$ and $\mu_{\theta_0} = \mu$. These two representations are asymptotically equivalent.

3 The Main Theorems

We now state the main results of the paper.

[4] We thank an anonymous referee for suggesting a change in our former definition of equivalence that led us to this definition. It has substantially simplified the paper.

[5] To see that $f_{\overline{\Theta}}(\nu, \epsilon)$ is a $\mathcal{B}$-measurable set, note that $f_{\overline{\Theta}}(\nu, \epsilon) = \cup_k \cap_{n>k} \cap_\ell \cap_{A \in \mathcal{G}_n^{n+\ell}} \{\theta : |\mu_\theta(A) - \nu(A)| < \epsilon\}$. Since $\overline{\Theta}$ is a representation the map $\theta \to \mu_\theta(A)$ is Borel measurable for any $A \in \mathcal{F}$, and so $\{\theta : |\mu_\theta(A) - \nu(A)| < \epsilon\}$ is a $\mathcal{B}$-measurable set for any $A \in \mathcal{F}$.

Theorem 1: If μ is asymptotically mixing, then there exists a representation that is learnable and provides patterns. Moreover, if $\overline{\Theta}_1$ and $\overline{\Theta}_2$ are representations that are both learnable and provide patterns, then $\overline{\Theta}_1 \sim \overline{\Theta}_2$.

Theorem 1 shows that learnability and providing patterns are not mutually exclusive. It also shows that any two such representations are asymptotically equivalent. The following results give more detailed insight into the class of representations that are learnable and provide patterns.

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: $(\Omega, \mathcal{F}, \mu, \mu^{\mathcal{F}^{\text{tail}}}_{\omega})$, where $\mu^{\mathcal{F}^{\text{tail}}}_{\omega}$ denotes the measure conditional on $\mathcal{F}^{\text{ tail}}$ at ω (i.e., the Radon–Nikodym derivative with respect to $\mathcal{F}^{\text{ tail}}$ at ω) . It is shown in the appendix (using a result of Dellacherie and Meyer (1978) and Stinchcombe (1990)) that this is in fact a representation.

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 learnable. Second, the tail field provides patterns in the sense that any finite horizon information is redundant given knowledge of the tail field. Third, any representation which satisfies these properties is equivalent to the tail field.

Theorem 2: If μ is asymptotically mixing, then $\overline{\mathcal{F}}^{\text{tail}}$ is learnable.

Theorem 3: $\overline{\mathcal{F}}^{\text{tail}}$ provides patterns.

That the tail field is learnable (in the asymptotic mixing case) and provides patterns relative to μ suggests that it gives us a "natural" representation of the learnable uncertainty. Theorem 1 then implies that any representation which is learnable and which provides patterns must be asymptotically equivalent to the tail field. The following Corollary summarizes the results.

Corollary 1: If μ is asymptotically mixing, then $\mathcal{F}^{\overline{\text{tail}}}$ is learnable and provides patterns. Moreover, if $\overline{\Theta}$ is learnable and provides patterns, then $\overline{\Theta} \sim \overline{\mathcal{F}}^{\text{tail}}$.

Both learnability and providing patterns play important roles in Theorem 1 and Corollary 1. Learnability limits the representation from being too fine and providing patterns limits the representation from being too coarse. To emphasize the role of providing patterns recall Example 1 from the introduction where a coin with a probability θ of heads is chosen (where θ is uniformly distributed on [0,1]) and then is flipped an infinite number of times. Consider the representation of $\mu = \frac{1}{2}\mu_{low} + \frac{1}{2}\mu_{high}$, where μ_{low} corresponds to a similar process where θ is uniformly distributed on [0,.5] where μ_{high} corresponds to a similar process of where θ is uniformly distributed on [.5,1]. Here, μ is asymptotically mixing and this representation is learnable. However, neither μ_{low} nor μ_{high} follow patterns.

One might ask to what extent converses to the results hold. Lemma 4 in the appendix shows that with an appropriate weakening of the asymptotic mixing condition, the tail field is learnable if and only if the asymptotic mixing condition is satisfied. Thus, there is a converse to Theorem 2 and the first sentence of the corollary. This leaves open the question of whether or not there is converse of the first sentence of Theorem 1: if the (weakened) asymptotic mixing condition is not satisfied is there no representation which is learnable and provides patterns? We do not have an answer to this question. Finally, one can ask whether the converse to the second sentence of the corollary is true. If μ is asymptotically mixing and $\Theta \overline{\sim} \mathcal{F}^{\overline{\text{tail}}}$, is it true that $\overline{\Theta}$ is learnable and provides patterns? The answer to this is no, as illustrated by the following example.

Example 5: Consider a process where a fair coin is flipped at each date. Thus, μ assigns probability $\frac{1}{2}$ to H and T at each date, independently of all other dates. In this example, the tail field representation is equivalent to the trivial one where $\mu = \mu$. However, we can also construct another representation which is equivalent to the tail field (and thus the trivial representation), but which is not learnable. Consider the sparse sequence of dates $\{t_n\}$ where $t_n = n^n$. Let the parameter space, Θ, be the unit interval. We think of each $\theta \in \Theta$ as an infinite sequence of 0's and 1's (e.g., corresponding to the

dyadic expansion of θ). The measure μ_θ is constructed as follows: for any date $t \neq t_n - 1$ for any $n \geq 2$ a fair coin is flipped. At dates $t_n - 1$ for some $n \geq 2$, μ_θ puts weight one on $\omega_{t_n-1} = \omega_{t_{n-1}}$ if the n-th entry of θ is 1 (and if the n-th entry of θ is 0, then μ_θ puts weight one on $\omega_{t_n-1} \neq \omega_{t_{n-1}}$). Thus, knowing θ tells an observer something about the flips on the particular sparse sequence of dates: namely whether or not the flip on some date in the sparse sequence matches the flip immediately preceeding the next date in the sparse sequence. The possible values of θ partition the statespace.

For each θ, it is easy to see that μ_θ has distance zero from μ itself. To see this, note that given any finite cylinder length ℓ, past some date (in particular, date t_n such that $(n+1)^{n+1} - 1 - n^n > \ell$) the information in θ has no impact on forecasts of cylinders of length ℓ. As μ_θ is equivalent to μ for all θ, it follows that the representation is equivalent to the tail field representation. However, for any θ, μ_θ is not learnable since on each date $t_{n+1}-1$, knowledge of θ and what occurred on date t_n determines the outcome on date $t_{n+1} - 1$.

One further thing to note about the example is that the representation induced in this example provides patterns since the information in any finite horizon does not influence distant calculations.

This example suggests that one needs to strengthen the definition of equivalence in order to make sure that $\overline{\Theta} \sim \overline{\mathcal{F}^{\text{tail}}}$ only if $\overline{\Theta}$ is learnable and provides patterns. [6]

4 Discussion of Learnability

Examples 3 and 4 show 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 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)).

[6] We conjecture that the converse would hold if one modifies the asymptotic distance to $d(\nu, \nu') = \sup_\ell \overline{\lim}_K \sup_{t,n:t+n\geq K} \max_{A\in\mathcal{G}_{t+n}^{t+n+\ell}} \max_{B\in\mathcal{F}_t:\nu(B)\nu'(B)>0} |\nu(A|B) - \nu'(A|B)|$. Our stated results still hold with this stronger (but admittedly more cumbersome) definition of distance.

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 $\mathcal{F}_t$ [7] 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 being able to make predictions as if one knew θ, as we point out in the following examples. They show that consistency and learnability are different notions, neither weaker than the other.

Example 6: A consistent, but not learnable, representation: $\Omega = \Theta = [0, 1]$, $\mathcal{B} = \mathcal{F}$, and μ is the uniform distribution. At each date t, the observation $\mathcal{F}_t$ is the first t digits of the binary expansion of ω.

Example 7: A learnable, but not consistent, representation: (as in example 4) $\Theta = \{\theta_1, \theta_2\}$ and $\mu_{\theta_1} = \mu_{\theta_2} = \mu$.

Note that Example 6 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 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

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

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 1 and Corollary 1 provide an equivalence class of representations that are learnable and provide patterns, 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.

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 mixing notion. Nevertheless, there may be interesting connections between decompositions in cross-sectional and time series models.

The word 'types' appears in the title, but has not appeared in the paper. 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 mixing and so our results would apply. In this sense the representation by probabilistic patterns 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)). [8] 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

[8] See Nyarko (1996) for some discussion of equivalent reformulations of type space.

types. Recent papers by Samet (1996ab) address such questions.

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.

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

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 [1995], "Social Learning in Recurring Games," forthcoming: *Games and Economic Behavior.*

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.

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.

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 $t \geq T$

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

Proof:[9] 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 $\nu(\{\omega \mid \nu(C_t \mid \mathcal{F}_{t-1})(\omega) > \epsilon$ for at least one $t \geq t_0\}) < \epsilon$ where

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

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

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

and its complement, denoted C_t^c, is

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

[9] 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

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

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

$$\begin{gathered}
|\tilde{\nu}(A \mid \mathcal{F}_t) - \nu(A \mid \mathcal{F}_t)| \\
\leq \tilde{\nu}(D_0)\nu(C^c_{t+1} \mid D_0) \ |\, \tilde{\nu}(A \mid \mathcal{F}_t \cap D_0 \cap C^c_{t+1}) - \nu(A \mid \mathcal{F}_t \cap D_0 \cap C^c_{t+1}) \mid \\
+\nu(D_0)\, \nu(C_{t+1} \mid D_0) \ |\, \tilde{\nu}(A \mid \mathcal{F}_t \cap D_0 \cap C_{t+1}) - \nu(A \mid \mathcal{F}_t \cap D_0 \cap C_{t+1}) \mid \\
+\nu(D^c_0)\, |\tilde{\nu}(A \mid \mathcal{F}_t \cap D_0) - \nu(A \mid \mathcal{F}_t \cap D^c_0)| \\
\leq (1-\epsilon)(1-\epsilon)\epsilon + (1-\epsilon)(\epsilon) + (\epsilon) \\
\leq 3\epsilon
\end{gathered}$$

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

Let us now turn to proving Theorems 1-3. We prove Theorem 1 by first proving Theorems 2 and 3 and next proving that if Θ is learnable and provides patterns then $\Theta \sim \mathcal{F}^{\text{tail}}$. Since $\sim$ is transitive, this establishes Theorem 1.

First, as promised in the text, we show that $\mathcal{F}^{\text{tail}}$ is a representation by showing that $\mu_\omega^{\mathcal{F}^{\text{tail}}}$ is a probability measure μ-a.e..

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) [10]] 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.

[10] 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.

Other 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 last two facts imply that we can find a countably generated sub σ-field of $\mathcal{F}$, $\mathcal{H}$ such that $\overline{\mathcal{H}} \sim \overline{\mathcal{F}}^{\text{tail}}$, $\mu_{\omega}^{\mathcal{H}} = \mu_{\omega}^{\mathcal{F}^{\text{tail}}}$ and $\mu_{\omega}^{\mathcal{H}}(AT_{\mathcal{H}}(\omega)) = 1$ for μ-a.e. ω. (This is done since the tail-field is not necessarily countably generated. See Blackwell and Dubins (1975).)

The following Lemmas will be 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 2: We prove the following lemma which is a stronger result.

Lemma 4: $\overline{\mathcal{F}}^{\text{tail}}$ is learnable if, and only if, for every $\delta > 0$ and μ-a.e. ω there exists T such that for all $t \geq T$

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

To see that asymptotic mixing implies the condition in Lemma 4, use Bayes' rule to rewrite asymptotic mixing to read that for any δ μ-a.e. ω there exists T such that for any $t \geq T$ and ℓ

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

Proof: First, we show that if the condition is satisfied then $\mathcal{F}^{\overline{\text{tail}}}$ 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, [11] 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

[11] 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 $\mu_\omega^{\mathcal{F}^{\text{tail}}}$.

$A \in \mathcal{G}_{t+1}$ it follows that $\mu(A|\mathcal{F}_t \vee \mathcal{G}_n^\infty))$ converges to $\mu(A|\mathcal{F}_t \vee \mathcal{F}^{\text{tail}})$ as $n \to \infty$, μ-a.e.. Thus, for μ-a.e. $\omega \in K$ there are infinitely many t such that

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

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

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

This is a contradiction.

Second, let us show the converse: if $\mathcal{F}^{\text{tail}}$ is learnable, then the condition of the lemma is satisfied.

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

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

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

$$\lim_t \lim_n \max_{A \in \mathcal{G}^{t+1}} |\mu(A|\mathcal{F}_t)(\omega) - \mu(A|\mathcal{F}_t \vee \mathcal{G}_n^\infty)(\omega)| = 0.$$

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

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

which is the desired conclusion. ∎

Proof of Theorem 3: Since $\mu_\omega^{\mathcal{F}^{\text{tail}}}$ has a trivial tail for μ-a.e. ω, [12] 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 $C \in \mathcal{F}_t$

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

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

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

To prove Theorem 1, given Theorems 2 and 3 we show that if μ is asymptotically mixing and $\overline{\Theta}$ is learnable and provides patterns, then $\Theta \overline{\sim} \mathcal{F}^{\overline{\text{tail}}}$. The following Lemma is instrumental in that proof and is of interest in its own right.

Lemma 5: Consider $\mathcal{H}$ which is countably generated such that $\mu\ {}^{\mathcal{H}}_{\omega} = \mu^{\mathcal{F}^{\text{tail}}}_{\omega}$ for μ-a.e. ω. If μ is asymptotically mixing and $\overline{\Theta}$ is learnable and provides patterns, then for λ-a.e. θ

$$d(\mu^{\mathcal{H}}_{\omega}, \mu_{\theta}) = 0$$

for μ_{θ}-a.e. ω.

Proof of Lemma 5: Since $\overline{\Theta}$ is learnable it follows that for λ-a.e. θ and all ℓ

$$\lim_t \max_{n, A \in \mathcal{G}^{t+n+\ell}_{t+n}} |\mu_{\theta}(A|\mathcal{F}_t)(\omega) - \mu(A|\mathcal{F}_t)(\omega)| = 0, \ \ \mu_{\theta} - \text{a.e.}.$$

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

$$\lim_n \max_{A \in \mathcal{G}^{t+n+\ell}_{t+n}} |\mu_{\theta}(A|\mathcal{F}_t)(\omega) - \mu_{\theta}(A)| = 0 \ \ \mu_{\theta} - \text{a.e.}.$$

Combining the two previous equations, it follows that for λ-a.e. θ and all ℓ

$$\lim_t \lim_n \max_{A \in \mathcal{G}^{t+n+\ell}_{t+n}} |\mu(A|\mathcal{F}_t)(\omega) - \mu_{\theta}(A)| = 0 \ \ \mu_{\theta} - \text{a.e.}. \tag{5}$$

Similarly, we can show that for all $\omega' \in D$, where $\mu(D) = 1$, and all ℓ

$$\lim_t \lim_n \max_{A \in \mathcal{G}^{t+n+\ell}_{t+n}} |\mu(A|\mathcal{F}_t)(\omega) - \mu^{\mathcal{H}}_{\omega'}(A)| = 0 \tag{6}$$

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

Next, we show that $\mu(Y) = 1$ where $Y = \{\omega' \mid \omega' \in B(\omega')\}$. To see this, suppose to the contrary that $\mu(Y) < 1$. [13] Therefore, by the definition of representation there exists a set S' such that $\mu(S') > 0$ and $\mu^{\mathcal{H}}_{\omega}(Y) < 1$

[13] 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 $\{\omega : |\mu\big(A|\mathcal{F}_t\big)\big(\omega\big) - \mu^{\mathcal{H}}_{\omega'}\big(A\big)| < \frac{1}{k}\}$.

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 $\mu^{\mathcal{H}}_{\omega"}(Y^c) > 0$ (where Y^c is the complement of Y), $\mu^{\mathcal{H}}_{\omega"}(B(\omega")) = 1$, and $\mu^{\mathcal{H}}_{\omega"}(A(\omega")) = 1$. 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 $\mu^{\mathcal{H}}_{\omega'} = \mu^{\mathcal{H}}_{\omega"}$ which imply that $\omega' \in B(\omega')$, which is a contradiction.

Thus, since $\omega' \in B(\omega')$ for almost every ω', it follows from (5) that for all ℓ and all $\omega \in D \cap Y$ (where $\mu(D \cap Y) = 1$)

$$\lim_t \lim_n \max_{A \in \mathcal{G}^{t+n+\ell}_{t+n}} |\mu(A|\mathcal{F}_t)(\omega) - \mu^{\mathcal{H}}_{\omega}(A)| = 0. \tag{7}$$

Thus, from (4) and (6) it follows that for λ-a.e. θ, all ℓ, and μ_θ-a.e. ω [14]

$$\lim_t \lim_n \max_{A \in \mathcal{G}^{t+n+\ell}_{t+n}} |\mu_\theta(A) - \mu^{\mathcal{H}}_{\omega}(A)| = 0.$$

This implies that for λ-a.e. θ, and μ_θ-a.e. ω $d(\mu_\theta, \mu^{\mathcal{H}}_{\omega}) = 0$. ∎

Proof of Theorem 1: The following Lemma completes the proof of Theorem 1 since Theorems 2 and 3 demonstrate the existence of a learnable representation that provides patterns (noting that $\sim$ is transitive).

Lemma 6: If μ is asymptotically mixing and $\overline{\Theta}$ is learnable and provides patterns, then $\overline{\Theta} \sim \overline{\mathcal{F}^{\text{tail}}}$.

Proof of Lemma 6: Recall $\mathcal{H}$ which is countably generated such that $\mu^{\mathcal{H}}_{\omega} = \mu^{\mathcal{F}^{\text{tail}}}_{\omega}$ for μ-a.e. ω. We show that $\overline{\Theta} \sim \overline{\mathcal{H}}$. Consider any $\nu \in \Delta$ and $\epsilon > 0$.

$$\lambda(f_{\overline{\Theta}}(\nu, \epsilon)) = \int_{\theta \in f_{\overline{\Theta}}(\nu,\epsilon)} d\lambda(\theta).$$

Let $G_\theta = \{\omega \mid d(\mu^{\mathcal{H}}_{\omega}, \mu_\theta) = 0\}$. By Lemma 5, $\mu_\theta(G_\theta) = 1$ for λ-a.e. θ. Note that if $\theta \in f_{\overline{\Theta}}(\nu, \epsilon)$, then $G_\theta \subset f_{\overline{\mathcal{H}}}(\nu, \epsilon)$. Therefore, $\mu_\theta(f_{\overline{\mathcal{H}}}(\nu, \epsilon)) = 1$ for λ-a.e. $\theta \in f_{\overline{\Theta}}(\nu, \epsilon)$. Thus we can write

$$\lambda(f_{\overline{\Theta}}(\nu, \epsilon)) = \int_{\theta \in f_{\overline{\Theta}}(\nu,\epsilon)} \mu_\theta(f_{\overline{\mathcal{H}}}(\nu, \epsilon)) d\lambda(\theta).$$

[14] To see that this holds for μ_θ-a.e. ω, suppose that it did not hold for $\theta \in J$ where $\lambda(J) > 0$ and for some E_θ with $\mu_\theta(E_\theta) > 0$ for each $\theta \in J$. It follows that $E_\theta \cap D \cap X = \emptyset$ for $\theta \in J$ and so $\mu_\theta(D \cap X) < 1$ for all $\theta \in J$. This contradicts the fact that $\mu(D \cap X) = 1 = \int_\theta \mu_\theta(D \cap X) d\lambda(\theta)$.

Next, note that if $\theta \notin f_{\overline{\Theta}}(\nu, \epsilon)$, then $G_\theta \cap f_{\overline{\mathcal{H}}}(\nu, \epsilon) = \emptyset$. Since $\mu_\theta(G_\theta) = 1$ for λ-a.e. θ, this implies that $\mu_\theta(f_{\overline{\mathcal{H}}}(\nu, \epsilon)) = 0$ for λ-a.e. $\theta \notin f_{\overline{\Theta}}(\nu, \epsilon)$. Thus, we can write

$$\lambda(f_{\overline{\Theta}}(\nu, \epsilon)) = \int_\Theta \mu_\theta(f_{\overline{\mathcal{H}}}(\nu, \epsilon)) d\lambda(\theta).$$

Given that $\overline{\Theta}$ is a representation, the right hand side of the above equation is simply $\mu(f_{\overline{\mathcal{H}}}(\nu, \epsilon))$ and so

$$\lambda(f_{\overline{\Theta}}(\nu, \epsilon)) = \mu(f_{\overline{\mathcal{H}}}(\nu, \epsilon)),$$

which establishes Lemma 6. ∎