Discussion Paper 1177

Patterns, Types, and Bayesian Learning*

Matthew O. Jackson Ehud Kalai Rann Smorodinsky*

MEDS, Kellogg Graduate School of Management
Northwestern University, Evanston, Il 60208-2009

Draft: January 1997

Abstract

Consider a probability distribution governing the evolution of a discrete-time stochastic process. Such a distribution may be represented as a convex combination of more elementary probability measures, with the interpretation of a two-stage Bayesian procedure. In the first stage, one of the measures is randomly selected according to the weights of the convex combinations (i.e., their prior probabilities), and in the second stage the selected measure governs the evolution of the stochastic process. Generally, however, the original distribution has infinitely many different such representations. Econometricians and economic agents may reach different insights about the process depending on the representation with which they start. This paper identifies one endogenous representation which is natural in the sense that its component measures are precisely the learnable probabilistic patterns.

* We thank Nabil Al-Najjar, Ehud Lehrer, and Meir Smorodinsky for helpful conversations, and Martin Northway for expositional comments on an earlier draft. Financial support under NSF grants SBR 9515421 and SBR 9507912 is gratefully acknowledged.

1 Introduction

The concept of a discrete-time stochastic process is a major modeling tool in decision theory and economics. Since the probability distribution governing the evolution of such a process can be highly complex, researchers often represent it as a convex combination of simpler distributions. Such a representation can result in a better understanding of a process and to better predictions about its evolution. The purpose of this paper is to identify such natural representations.

Convex representations arise when we deal with Bayesian models. For instance, a probability distribution that can be written as $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$, may be thought of as a two stage process. In a preliminary stage, μ_1 or μ_2 is selected with prior probabilities α and $1-\alpha$, respectively, and then in a second stage the selected distribution governs the evolution of the process. Even if the original distribution μ is complex, it may be that the components μ_1 and μ_2 are simple.

The multi-arm bandit problem is a well-known example from decision theory, which has many applications in economics. (For example, see Rotschid (1974), Banks and Sundaram (1992), and Bergemann and Välimäki (1996).) In this example, repeated uses of an arm, or activity, result in a stochastic sequence of payoffs. The agent must decide in each period whether to choose one activity (rather than other ones), and holds subjective beliefs about its future payoff sequences. These beliefs are described by a probability distribution over payoff streams. In many studies using this model, the agent assumes that the real distribution governing the system is one of many possible stationary distributions, but does not know which one. If one assigns prior probabilities to the underlying possible distributions, then the agent's beliefs may be represented as a convex combination of stationary distributions. Notice that while the agent's overall beliefs may not be stationary, since he updates them as he observes payoffs generated by the arm, these beliefs may be represented as a convex combination of stationary distributions.

Multiperson versions of such decision problems also arise in repeated games with incomplete information. (See Aumann and Maschler (1995).) Each player in such a game chooses a repeated game strategy according to his realized (Harsanyi (1967)) type, and the evolution of play is governed by the distribution induced by the vector of strategies of the realized types. To

an observer who does not know the realized types, the overall distribution is the convex combination induced by his prior beliefs of the likelihoods of the various types.

More generally, in Bayesian statistics, when the statistician or econometrician analyzes some discrete time stochastic process, she considers a set of models that may describe the evolution of the system. Starting with prior probabilities over these possible models, which are updated by Bayes' rule after every observation, she obtains an overall probability distribution for the evolution of the system. This econometrician's primitives are these models with their prior probabilities.

However, it is possible that other econometricians, with different models and prior probabilities, will all arrive at the same overall probability distribution for the evolution of the system. If they all make the same predictions, is it still possible that one's primitives are better than another's? Two criteria seem interesting: simplicity and learnability. The econometrician who obtains his overall distribution by building on simpler models may gain a better understanding of the process. And the econometrician who builds on models that become recognizable with time may learn to make better predictions.

Starting with an arbitrary probability distribution for a discrete-time finite-state stochastic process, our goal is to identify its natural representations. We want the component distributions of a representation to be simple, in order to explain the nature of the distribution, and useful, to improve predictive ability. For this purpose, we single out component distributions which we call patterns, and show that the distributions of well behaved processes may be represented as convex combinations of these patterns. Moreover, if one wants these patterns to be learnable, then every process has essentially a unique representation by patterns. These learnable patterns may be interpreted as the conditional distributions induced by the tail field of the process.

To explain the notions of patterns and learnability let us consider an example of a process consisting of infinite sequences of H and T's one obtains when studying repeated coin tosses.

Patterns. There are obviously many ways to define a notion of pattern.[1]

[1] Recent examples of other notions of pattern, drastically different from ours, can be found in Sonsino (1997), Fudenberg and Levine (1995), and Sargent (1993). Marimon's (1995) survey on learning includes discussion of the uses of the terms 'pattern' and 'pattern

For our purposes, we need to define probabilistic patterns, and the notion we adopt fits well into the mathematics of representations. A pattern is a probability distribution for the process, governing, for example, infinite sequences of H and T's, which has this property: given any finite initial segment of outcomes h, and a late time t, the conditional probabilities of events after t given the initial segment h, are essentially the same as the probabilities of these events without the knowledge of the initial segment h. In other words, a person who knows the pattern may disregard any initial information when she predicts late events. The following examples should help us understand and motivate this definition.

A *deterministic cyclical pattern*, say the infinite sequence $H, H, T, H, H, T, \ldots$, fits our definition of a pattern by viewing it as a Dirac measure that assigns a probability of 1 to this particular sequence. A person who knows this pattern will call the outcome correctly at any time while ignoring whatever initial information is given to him.

A *stationary stochastic pattern*, where the process is determined by i.i.d. tosses of a coin, with a fixed known probability θ for H, fits the definition of a pattern. Again, a person knowing this pattern will assign a probability of θ to H at every time and will disregard any initial history of outcomes in making this prediction.

Notice, however, that a *randomly selected stationary distribution* results in a distribution which is *not* a pattern. For example, consider tossing repeatedly an i.i.d. coin whose probability of heads, θ, is unknown and was selected by a uniform distribution on the interval (0, 1). The first 100 outcomes of the selected coin inform an observer about the likelihood of the selected parameter θ, which is important information for predicting the outcome in any future period. In accordance with our results, while this distribution itself is not a pattern, it may be represented as a convex combination of continuously many stochastic patterns, i.e. the ones generated by all possible realized coins with known parameters.

A *heavy coin* is one which starts in the first period in a fixed position, say H, and in each subsequent period stays in the same position as in the previous period with probability 1-ϵ, but turns over to the other position with probability ϵ, for some known small positive number ϵ. The distribution generated by such a coin is a pattern consistent with the asymptotic defi-

---

recognition' in the literature.

nition given above. For each initial sequence of outcomes, the probabilities assessed for events occuring sufficiently far in the future are essentially the same whether or not these assessments condition on the initial segment. It is such patterns that lead us to an asymptotic definition. An initial segment may be relevant in making predictions at early times: for example, initially it is important to know the last position of the coin, but after enough time has passed the last position of the coin becomes less important to a person knowing the pattern.

We should emphasize that our analysis is not limited to stationary processes, as the example of the heavy coin illustrates. Our analysis may be applied to any process for which the outcome of some given finite horizon is asymptotically independent of events far into the future.

Learnability. Under the definition of patterns suggested above, any deterministic infinite sequence of H and T's is a pattern. If one knows the sequence, i.e. the pattern, she needs no initial information to correctly predict the outcome in any time t. This means that any distribution μ has a trivial representation by deterministic patterns: We simply assign to every infinite sequence its Dirac measure and use the original distribution μ as a prior over these measures. Unfortunately, these very fine patterns are not recognizable, in that even a long time observer may never be able to identify them. Consider, for example, the stationary stochastic distribution μ associated with the repeated toss of a fair coin (i.e., with θ=.5). While this distribution is decomposable into a combination of deterministic distributions that put weight one on each different infinite H,T sequence; an observer cannot learn to identify which sequence was realized. After every history of outcomes a Bayesian observer would assign all continuation sequences the same probability as she would without knowing the initial segment. Thus no information is revealed about the realized pattern, except for that part already revealed. In other words, this representation by patterns offers no help in predicting the future of the process.

In contrast with the above example where the type of coin is known, consider the earlier example of a repeated coin toss where θ (the coin's probability of heads) is unknown and has a uniform prior of possible parameters between 0 and 1. In this case, the realized pattern (i.e., the coin) is learnable. After observing the outcomes for a long time, one obtains a fairly sharp posterior distribution about the realized coin and is able to predict the

probabilities of the next outcome with a high degree of accuracy.

Therefore we may define a representation by patterns to be learnable if, for each realized pattern, a long-time observer of the evolution of the process assigns conditional probabilities to the values of the next state that are nearly the same as the ones assigned by a person who knows the realized pattern. In other words, an observer will learn to call probabilities correctly, just as if she had learned the realized pattern.

Relationship to de Finetti's Theorem. The celebrated de Finetti Theorem suggests a well-known example of a decomposition to learnable patterns. Illustrated in the context of repeated coin tosses, 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, when 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 patterns, which are stationary in his case.

Our main result is similar to de Finetti's, except that we replace the exchangeability condition with the weaker condition of asymptotic mixing. Our conclusion is therefore weaker: we obtain a representation by learnable patterns, which are not necessarily stationary, however.

For example, suppose that the probability assigned to every initial segment h of H's and T's is given by $\mu(h) = (2^c + 3^c)/2^{2n+1}$, where n is the length of h and c is the number of H's appearing in odd periods plus the number of T's appearing in even periods; then this is not an exchangeable process since switching the positions of adjacent H and T changes the value of c and therefore the probability of h. Thus, de Finetti's theorem does not apply. Nevertheless the process can be decomposed into two learnable patterns, μ_S and μ_A, described as follows:

The first pattern, μ_S, is our standard distribution obtained from repeated independent tosses of a fair coin. The second pattern, μ_A, may be described as the distribution of tosses that are independent across periods in which a 0.75 coin is tossed in even periods, whereas a 0.25 coin is tossed in odd periods. It is easy to see that we will obtain the probabilities of histories presented above when we randomly choose one of these two distributions

with equal prior probabilities and apply it to generate the sequence. Moreover, each of these two distributions is a probabilistic pattern that when be accurately predicted without using the initial information.

In addition, however, these patterns are learnable. For almost every evolution of the process, after a sufficiently long time an observer will make one of two types of predictions regarding the probabilities of oncoming H's. She will either predict them all to be 0.5, or she will predict them to be 0.75 for odd periods and 0.25 for even periods; she will do this just as if she had known the realized pattern of the two possible ones. Again, it is important to emphasize that the observer does not need to know the decomposition we presented: she will be lead to it as a consequence of simple Bayesian updating of the original distribution.

2 An Overview

Let Ω be a given set and $\{F_t\}_{t=1}^{\infty}$ be a sequence of finite σ fields on Ω. Let F be the σ-field on Ω generated by $\{F_t\}_{t=1}^{\infty}$ i.e., $F = \sigma\left(\cup_{t=1}^{\infty} F_t\right)$. Note that F is countably generated. Denote $H_t = \sigma\left(\cup_{j=1}^{t} F_j\right)$. $\{H_t\}_{t=1}^{\infty}$ is a filtration on (Ω, F). Let $\mathcal{P}(\Omega, F)$ be the set of all probability measures on (Ω, F). Throughout the paper we shall treat Ω, $\{F_t\}_{t=1}^{\infty}$, and $\mu \in \mathcal{P}(\Omega, F)$ as fixed.

Representations

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

1. $\theta \to \mu_\theta(S)$ is 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 obvious examples are as follows:

Let Θ consist of a single point θ and $\mu_\theta = \mu$.

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

Notice that, with a representation, we can think of a random draw of ω according to μ as equivalently first choosing θ by λ and then choosing ω by μ_θ. In other words, a representation consists of a prior λ over (Θ, B) and a collection of posteriors, μ_θ.

Our interest in the sequel is to identify the "right" representation from a specific point of view. An observer of the filtration $\{H_t\}_{t=1}^{\infty}$ 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 $\Theta = \Omega$).

However, let us be more precise about what we meant by the "right" representation.

Learning

First, since the representation is supposed to show what an observer of the filtration comes to know, it must be something that contains only information eventually available through the filtration. This 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 important in the Bayesian learning literature (e.g., Kalai and Lehrer (1993), Lehrer and Smorodinsky (1995), and Jackson and Kalai (1995)).

Definition: Let $\hat{\nu} \in \mathcal{P}(\Omega, F)$. We say that $\hat{\nu}$ *merges* with ν if $\forall\ \epsilon > 0$ and

$$\nu - a.e.,\ \omega \in \Omega \quad \exists T = T(\epsilon, \omega) \text{ such that for all } t \geq T$$

$$\sup_{A \in H_{t+1}} |\tilde{\nu}(A|H_t) - \nu(A|H_t)| < \epsilon.$$

If $\tilde{\nu}$ merges with ν, then eventually forecasts provided by $\tilde{\nu}$ regarding short horizon events will approach the "true" forecasts provided by ν.

Using the notion of merging, we may now formalize what we mean by a learnable representation:

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

Our definition of learnability states that 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 θ is "learned."

Patterns

The "right" representation should satisfy two considerations. First, as above, the right representation should not embody anything which could not be known by an observer of the filtration. Second part it should not embody anything less than what would be known by an observer of the filtration, at least asymptotically. That is captured in the following definitions.

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

$$\lim_{n \to \infty} \sup_{A \in \vee_{i=n}^{n+l} F_{t+i}} |\tilde{\mu}(A|H_t) - \tilde{\mu}(A)| = 0 \quad \tilde{\mu} - \text{a.e.}.$$

This definition captures the notion that observing the filtration (asymptotically) suggests nothing new to someone who already knows the measure $\tilde{\mu}$. We apply this definition to each μ_θ.

Definition: A representation $(\Theta, B, \lambda, (\mu_\theta))$ of a measure $\mu \in \mathcal{P}(\Omega, F)$ *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, consider an agent who knows the transition probabilities of an irreducible Markov chain but is unfamilar with the 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 onthe 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.

Our Goal

With the above definitions in hand, we now may be more specific about the main result that we pursue in this paper: our goal to identify the representations which are learnable *and* provide patterns.

The remainder of the paper is structured as follows. In the next section we show that, instead of working directly with representations, we can equivalently work with σ-fields. We then provide the characterization of the representations which are learnable *and* provide patterns. We close with brief remarks on possible applications and extensions of the result. Proofs of the results are presented in an appendix.

3 Representation by σ–Fields

Although the definitions above are intuitive, it is easier to describe representations by associated σ fields. This section shows that there is no loss of generality in doing so. Roughly, the information represented by Θ may be equivalently represented by a σ field, provided that Θ contains no superfluous information.

As the parameter set Θ is essentially arbitrary, one may naturally define an equivalence relation between two representations which are just renamings of each other.

In defining the equivalence relation, we use $\otimes$ to denote the product measure of $\lambda \times \mu_\theta$ over $(\Theta, B) \times (\Omega, F)$. That is,

$$\otimes(A \times C) = \int_{\theta \in A} \mu_\theta(C) d\lambda(\theta)$$

Note that the marginal of ϕ on Θ, denoted $\phi_\Theta(\cdot)$, is equal to λ, and the marginal on Ω, denoted $\phi_\Omega(\cdot)$, is equal to μ.

In the following we use $\overline{\Theta}_1$ to represent $(\Theta_1, B_1, \lambda_1, (\mu_{\theta_1})_{\theta_1 \in \Theta_1})$ and $\overline{\Theta}_2$ to represent $(\Theta_2, B_2, \lambda_2, (\mu_{\theta_2})_{\theta_2 \in \Theta_2})$.

Definition: $\overline{\Theta}_1$ is *weakly finer* than $\overline{\Theta}_2$, denoted $\overline{\Theta}_1 \succeq_1 \overline{\Theta}_2$, if there exists $f : B_2 \to B_1$ such that $\phi_1(f(A) \times C) = \phi_2(A \times C)$ for all $(A \times C) \in B_2 \times F$.

Definition: $\overline{\Theta}_1$ is *equivalent* to $\overline{\Theta}_2$, denoted $\overline{\Theta}_1 \sim_1 \overline{\Theta}_2$, if $\overline{\Theta}_1 \succeq_1 \overline{\Theta}_2 \succeq_1 \overline{\Theta}_1$.

For the sake of simplicity we confine our discussion to representations such that B is countably generated. We refer to these as *countably generated representations*. This is a weak restriction, however, given that F itself is countably generated.

Consistency

Before proceeding further, let us note that one may introduce extraneous information through a representation. This is illustrated in the following example.

Example: Consider any μ. A valid representation is to have $\Theta = \{H, T\}$ and $\mu_H = \mu_T = \mu$. One may interpret this representation as having nature flip a coin to choose between using μ_H or μ_T, even though this is the same measure in either case.

In the example above, information about θ is completely extraneous, from our perspective. This motivates the definition of consistency, below (see Diaconis and Freedman (1986)).

Definition: The representation $(\Theta, B, \lambda, (\mu_\theta)_{\theta \in \Theta})$ is *consistent* if Θ is a topological space and for λ-a.e. $\theta \in \Theta$ $\phi_\Theta(\cdot \mid H_t)$ weakly converges to δ_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 turns out to be 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: Λ consistent, but not learnable, representation: $\Omega = \Theta = [0,1]$, $B = F$, and μ is the uniform distribution. At each date t, the observation H_t is the first t digits of the binary expansion of ω.

Example: Λ learnable, but not consistent, representation: (as presented before the definition of consistent) $\Theta = \{H, T\}$ and $\mu_H = \mu_T = \mu$.

Note that the first example 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.

Representation by σ-Fields

We proceed to identify with each representation a σ-field corresponding to the information captured by the representation.

Let $G \subset F$ be a sub-σ-field. Denote by $\overline{G}$ the quadruple $(\Omega, G, \mu, (\mu^*_G)_{\omega\in\Omega})$ where $\mu^*_G(A) = E(1_A|G)(\omega)\ \forall\ A \in F$. (Fix a version of the conditional expectation.) In general, $\mu^*_G(\cdot)$ may not be a probability distribution. However, in our case, we need only consider G's (as we show below) that are countably generated, in which case $\mu^*_G(\cdot)$ is a probability distribution (see Theorem A from Stinchcombe (1989) in our appendix).

Let $\mathcal{F}^*$ be the set of all countably generated sub-σ-fields of F.

Lemma 1: If $(\Theta, B, \lambda, (\mu_\theta)_{\theta\in\Theta})$ is a consistent, countably generated representation then there exists $G \in \mathcal{F}^*$ such that $(\Theta, B, \lambda, (\mu_\theta)_{\theta\in\Theta}) \sim_1 \overline{G}$, and $\overline{G}$ is consistent.

Lemma 1 allows us to work with σ fields directly. We now show that the equivalence relationship $\sim_1$ between representations may be replaced with an equivalence relationship $\sim_2$ between σ fields.

Definition: $G \sim_2 G'$ if for all $A \in G$ there exists $B \in G'$ such that $\mu(A \Delta B) = 0$, and vice versa.

Lemma 2: If $G, G' \in \mathcal{F}^*$, then $G \sim_2 G'$ if and only if $\overline{G} \sim_1 \overline{G}'$.

We use $[G]$ and $[\overline{G}]$ to denote the equivalence classes under $\sim_2$ and $\sim_1$, respectively.

Given that it will often be useful to work directly with σ-field representations, we rewrite the definitions of patterns and learnability for σ field representations.

Remark: If $G \in \mathcal{F}^*$, then $\overline{G}$ is learnable if and only if for μ-a.e. ω and $\forall\ \epsilon > 0$ and $\mu^{\omega}_{\overline{G}}$-a.e. $\omega' \in \Omega\ \exists\ T$ such that $\forall\ t \geq T$

$$\sup_{A \in H_{t+1}} |\mu^{\omega}_{\overline{G}}(A|H_t) - \mu(A|H_t)| < \epsilon. \qquad (*)$$

Given this remark, we say that $G \in \mathcal{F}^*$ is *learnable* if $(*)$ is satisfied.

Remark: If $G \in \mathcal{F}^*$, then $\overline{G}$ provides patterns if and only if $\mu^{\omega}_{\overline{G}}$ *follows a pattern* for μ-a.e. ω.

Given this remark, we say that $G \in \mathcal{F}^*$ *provides patterns* if $\mu^{\omega}_{\overline{G}}$ *follows a pattern* for μ-a.e. ω.

Mixing

Before proceeding to our main results, we provide a few more definitions. We will restrict our attention to measures μ that are asymptotically mixing. This restriction is motivated by the following example.

Example: Let $\Omega = \{0,1\}^{\mathbb{N}}$. Let H_t be the field of all cylinders of length t, $F = \vee_{t=1}^{\infty} H_t$. Consider a measure μ generated as follows: Partition the set of periods $\mathbb{N}$ to $\mathbb{N} = \cup_i^{\infty} N_i$, where each N_i has a positive density in $\mathbb{N}$. 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. 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. Al-Najjar (1996a) uses a similar example to demonstrate a chaotic asset market which is impossible to model with a finite factor structure (as in the Arbitrage Pricing Theory).

This example illustrates the problem that there may be clear patterns associated with the sequences that may 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 farther in the future. We use the notion of asymptotic mixing (defined below) to restrict attention to "non-chaotic" measures. This, as we shall see, assures that the patterns identifiable from the filtration are learnable.

We now formalize the notion of asymptotic mixing.

Definition: Let G_1 and G_2 be two finite subfields of F. We say that G_1 is δ-*independent* of G_2 with respect to to a given measure μ, denoted $G_1 \perp^{\delta} G_2$, if $\sum_{A \in G_1} |\mu(A|C) - \mu(A)| < \delta$ for all $C \in G_2$ except on a set of atoms C whose union has measure less than δ.

Smorodinsky (1971) shows that if $G_1 \perp^{\epsilon} G_2$, then $G_2 \perp^{\sqrt{2\epsilon}} G_1$. Thus, in an asymptotic sense (as $\epsilon \to 0$) the relation $\perp^{\epsilon}$ is symmetric.

Definition: A measure μ is δ-*mixing* if there exists a sequence $\epsilon_{n,m} \in I\!R$ such that $\epsilon_{n,m} \to 0$ as $n \to \infty$ for each m, and such that for all $n \in N$ $\vee_{t=1}^{m} F_t \perp^{\delta+\epsilon_{n,m}} \vee_{t=n}^{n+m} F_t$.

The interpretation of δ-mixing is that any finite horizon information may influence the forecasts regarding events in the (finite) long-run future by no more than (asymptotically) δ. μ is *mixing* if it is 0-mixing.

We shall confine our discussion to measures which are asymptotically mixing in the following sense:

Definition: μ is *asymptotically mixing* if for μ-a.e. $\omega \in \Omega$, and for any $\delta > 0$, there exists $T = T(\delta, \omega)$ such that $\mu(\cdot|H^t)$ is δ-mixing.

4 The Tail Field

We are now ready to state the main result of the paper.

Let F^{tail} denote the tail σ-field, $F^{\text{tail}} = \cap_{j=1}^{\infty}\sigma(\cup_{t=j}^{\infty}F_t)$.

In the sequel we will abuse the notation $[F^{\text{tail}}]$ to represent an arbitrary, countably-generated, sub σ-field which is equivalent (by $\sim_2$) to the tail field.

The main result is that the tail field precisely captures the asymptotic information that an observer will learn through the filtration. This result 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 1: If μ is asymptotically mixing, then $[F^{\text{tail}}]$ is learnable.

Theorem 2: $[F^{\text{tail}}]$ provides patterns.

That the tail field is learnable (in the asymptotic mixing case) and provides patterns relative to μ suggests that it gives us the right representation of the learnable uncertainty. To make this statement convincing, we show that any field which is learnable and which provides the patterns relative to μ must be informationally equivalent to the tail field.

Our notion of informational equivalence is based on the forecasts provided by a field, i.e., two fields are informationally equivalent if the provide the same asymptotic forecasts.

Definition G_1 and G_2 are *asymptotically informationally equivalent* (denoted $G_1 \sim_3 G_2$) if for every n and μ-a.e. ω

$$\lim_t \sup_{A\in\vee_{i=1}^{n}F_{t+i}} |\mu^{\omega}_{G_1}(A) - \mu^{\omega}_{G_2}(A)| = 0.$$

The equivalence relation $\sim_3$ is coarser than $\sim_2$ (and thus $\sim_1$).

Lemma 3: Given $G, G' \in \mathcal{F}^*$, if $G \sim_2 G'$, then $G \sim_3 G'$.

We may now provide the main result which characterizes the representations that are learnable and provides patterns.

Theorem 3: If μ is asymptotically mixing, then G is learnable and provides patterns if, and only if, $G \sim_3 F^{\text{tail}}$.

5 Additional Remarks

As we mentioned in the introduction, our analysis may be used to identify the natural models that an econometrician or a econometrician 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 long observation of a stochastic process. The representation tells one in advance what patterns she is likely to learn and with what probabilities. This is precisely the information she will need in order to compute the expected additional 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 an equivalence class of representations under $\sim_3$, and one might want to refine this class to representations with the least redundancy, e.g. where different parameter values imply different distributions.

A recent paper by Al-Najjar (1996b) 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.

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.

References

Allen, B. [1983]. "Neighboring Information and Distributions of Agents' Characteristics under Uncertainty." *Journal of Mathematical Economics*, Vol. 12, pp. 63-101.

Al-Najjar, N. [1996a]. "Factor Structures and Arbitrage Pricing in Large Asset Markets," forthcoming: *Journal of Economic Theory*.

Al-Najjar, N. [1996b]. "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.

Banks, J.S. and R.K. Sundaram [1992]. "Denumerable Armed Bandits," *Econometrica*, Vol. 60, pp. 1071-1096.

Bergemann, D. and J. Välimäki [1996]. "Learning and Strategic Pricing," *Econometrica*, Vol. 64, pp. 1125–1150.

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.

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. 14, pp. 1–26.

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

Harsanyi, J.C. [1967-68]. "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," mimeo: Northwestern University.

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 [1994], "Repeated Large Games with Incomplete Information," forthcoming: *Games and Economic Behavior*.

Marimon, R. [1995], "Learning from Learning in Economics," prepared for the 7th World Congress of the Econometric Society.

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.

Appendix: Proofs

The following results will be useful in proving Lemma 1.

Let $AT_G(\omega) = \cap_{\{A \in G | \omega \in A\}} A$. (See definition 3.2.4 of Stinchcombe (1989).)

Theorem A: [Stinchcombe (1989)] If $G \in \mathcal{F}^*$, then there exist versions of $E(1_A|G)$ for all $A \in F$ such that $\mu^{\omega}_G(A) \equiv E(1_A|G)$ is a probability measure μ-a.e.. Furthermore, $\mu^{\omega}_G(AT_F(\omega)) = 1$ for μ-a.e. ω.

Lemma 9: If $G \subset F$, then $\overline{G}$ is a consistent representation.

Proof of Lemma 9: We first need to show that $\overline{G}$ is a representation. This is straightforward as the two conditions of the definition of representation are as follows:

1. $\omega \to \mu^{\omega}_G(S)$ is the same mapping as $\omega \to E(1_S|G)(\omega)$, which by definition of the conditional expectation is Borel measurable.

2. $\int \mu^{\omega}_G(S) d\mu(\omega) = \int E(1_S|G)(\omega) d\mu(\omega) = \int 1_S(\omega) d\mu(\omega) = \mu(S)$.

The fact that G is consistent then follows from Theorem A and the fact that $\{H_t\}_{t=1}^{\infty}$ generates F. ∎

Proof of Lemma 1: By consistency we know that λ-a.e. $\phi_\Theta(\cdot \mid H_t)$ weakly converges to δ_θ μ-a.e.. Thus, we can generate a function $h : \Omega \to \Theta$ by $h(\omega) = \theta$ such that $\phi_\Theta(\cdot \mid H_t)(\omega) \to \delta_\theta$. h is defined μ-a.e. and is measurable. Extend h in an arbitrary way to be defined on all of Ω. Consider the collection $G = \{h^{-1}(A) : A \in B\}$. It is straightforward to show that $G \in F$, G is countably generated (recall that B is countably generated), and G is a σ-algebra, i.e., $G \in F^*$. We now show that $\Theta \succeq_1 G$. By consistency it is clear that $\mu_\theta(h^{-1}(A)) = 1$ for almost all $\theta \in A$, and $\mu_\theta(h^{-1}(A)) = 0$ for almost all $\theta \in A^c$. Define $f(A) = h^{-1}(A)$, so

$$\begin{aligned}
\phi_2(f(A) \times D) &= \int_{\omega \in f(A)} \mu^{\omega}_G(D)\, d\mu(\omega) \\
&= \int_{\omega \in h^{-1}(A)} E(1_D|G)(\omega) d\mu(\omega) \\
&= \int_{\omega \in h^{-1}(A)} 1_D(\omega)\, d\mu(\omega) = \mu(h^{-1}(A) \cap D) \\
&= \int_{\theta \in \Theta} \mu_\theta(h^{-1}(A) \cap D)\, d\lambda(\theta)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\theta\in A} \mu_\theta(h^{-1}(A)\cap D)d\lambda(\theta) + \int_{\theta\in A^c} \mu_\theta(h^{-1}(A)\cap D)d\lambda(\theta)\\
&= \int_{\theta\in A} \mu_\theta(D)d\lambda(\theta) + \int_{\theta\in A^c} 0 d\lambda(\theta)\\
&= o_1(A\times D).
\end{aligned}
$$

As for the converse, $(G \geq_1 \Theta)$ define $h : \Theta \to \Omega$ in any way so that $h \circ h$ is the identity. This definition can be made almost everywhere since the image of h has probability 1. Take $C \in G$. By construction there exists $A \in B$ such that $C = h^{-1}(A)$, i.e., $h^{-1}(C) = h^{-1}(h^{-1}(A)) = A$. Then
$o_1(h^{-1}(C)\times D) = o_1(A\times D) = o_2(h^{-1}(A)\times D) = o_2(C\times D)$. ∎

Proof of Lemma 2: We begin by showing that $\sim_2$ is stronger. Consider the map $h_1 : G \to G'$ defined arbitrarily by $\mu(h_1(A)\Delta A) = 0$. As $\sim_2$ is satisfied, this is well defined. Similarly, define the map $h_2 : G' \to G$. Consider any $(A, D) \in G\times F$.

$o_1(A\times D) = \int_A \mu^{\omega}_G(D)d\mu = \int_A E(1_D|G)(\omega)d\mu = \int_A 1_D(\omega)d\mu = \int_{h_1(A)} 1_D(\omega)d\mu = \int_{h_1(A)} \mu^{\omega}_{G'}(D)d\mu = o_2(h_1(A)\times D)$.

Similarly, $o_1(h_2(A)\times D) = o_2(A\times D)$, and so $\overline{G} \sim_1 \overline{G}'$.
For the other direction, assume that $\overline{G} \sim_1 \overline{G}'$ and define h_1 and h_2 accordingly. For any $A \in G$

$$
\begin{aligned}
\mu(A) &= \int_A 1_A(\omega)d\mu = \int_A E(1_A\,|\,G_1)(\omega)d\mu =\\
&= \int_A \mu^{\omega}_{G_1}(A)d\mu = \int_{h_1(A)} \mu^{\omega}_{G_2}(A)d\mu = \int_{h_1(A)} E(1_A\,|\,G_2)(\omega)d\mu =\\
&= \int_{h_1(A)} 1_A(\omega)d\mu = \mu(A\cap h_1(A)).
\end{aligned}
$$

Therefore:
(a) $\mu(A\cap \overline{h_1(A)}) = 0$.
Applying a symmetric argument to $h_1(A)$ we get:
(b) $\mu(h_1(A)\cap \overline{h_2(h_1(A))}) = 0$.
Also

$$
\begin{aligned}
\mu(A \cap h_2(h_1(A))) &= \int_A 1_{h_2(h_1(A))}(\omega)d\mu =\\
&= \int_A E(1_{h_2(h_1(A))}|G_1)(\omega)d\mu =
\end{aligned}
$$

$$= \int_A \mu^{\omega}_{G_2}(h_2(h_1(A)))\,d\mu = \int_{h_1(A)} \mu^{\omega}_{G_2}(h_2(h_1(A)))d\mu =$$
$$= \int_{h_2(h_1(A))} \mu^{\omega}_{G_2}(h_2(h_1(A)))d\mu = \int_{h_2(h_1(A))} E(1_{h_2(h_1(A))}|G_1)(\omega)d\mu =$$
$$= \int_{h_2(h_1(A))} 1_{h_2(h_1(A))}(\omega)d\mu = \mu(h_2(h_1(A)))$$

which implies
(c) $\mu(h_2(h_1(A)) \cap A) = 0$.
Combining (b) and (c):

$$\begin{aligned}\mu(h_1(A) \cap A) &= \mu(h_1(A) \cap A \cap h_2(h_1(A))) + \\ &+ \mu(h_1(A) \cap A \cap \overline{h_2(h_1(A))}) \leq \\ &\leq \mu(A \cap h_2(h_1(A))) + \mu(h_1(A) \cap \overline{h_2(h_1(A))}) = 0\end{aligned}$$

which together with (a) shows that for any $A \in G_1$, $h_1(A) \in G_2$ satisfies $\mu(A \Delta h_1(A)) = 0$. By symmetric arguments for every $B \in G_2$, $\mu(B \Delta h_2(B)) = 0$. ∎

Proof of Lemma 3: We show the stronger result that $\sup_{t, A \in F_t} |\mu^{\omega}_{G_1}(A) - \mu^{\omega}_{G_2}(A)| = 0$ μ a.e.. As there are countably many sets in $\cup_{t=1}^{\infty} F_t$, it is sufficient to show that for any given t and $A \in F_t$ $\mu^{\omega}_{G_1}(A) - \mu^{\omega}_{G_2}(A) = 0$ μ-a.e.. We proceed by way of contradiction. Suppose that there exist $A, B \in F$ such that $\mu^{\omega}_{G_1}(A) \neq \mu^{\omega}_{G_2}(A)$ for all $\omega \in B$ and $\mu(B) > 0$. Without loss of generality we may assume that B is of the form $\{\omega \mid \mu^{\omega}_{G_1}(A) > \alpha > \mu^{\omega}_{G_2}(A)\}$. Denote $B_1 = \{\omega \mid \mu^{\omega}_{G_1}(A) > \alpha\}$, $B_2 = \{\omega \mid \mu^{\omega}_{G_2}(A) < \alpha\}$. Obviously, $B_1 \in G_1$ and $B_2 \in G_2$. Denote by $h_1 : G_1 \to G_2$ and $h_2 : G_2 \to G_1$ the isomorphism generators of $\sim_1$. Let $B_i^* = h_i(B_i)$: i.e., $B_1^* \in G_2$ and $B_2^* \in G_1$. Without loss of generality (see Proof of Lemma 2), we assume that $\mu(B_1 \Delta B_1^*) = \mu(B_2 \Delta B_2^*) = 0$. As $B = B_1 \cap B_2$ we have $\mu(B \Delta (B_1 \cap B_2^*)) = \mu(B \Delta (B_1^* \cap B_2)) = 0$. Also by the properties of h_1, h_2, $\mu((B_1 \cap B_2^*) \Delta h_1(B_1 \cap B_2^*)) = 0$ and $\mu((B_2 \cap B_1^*) \Delta h_2(B_2 \cap B_1^*)) = 0$. We are now ready to reach a contradiction as follows:

$$\begin{aligned}\alpha\ \mu(B) &< \int_B \mu^{\omega}_{G_1}(A)d\mu = \\ &= \int_{B_1 \cap B_2^*} \mu^{\omega}_{G_1}(A)d\mu = \int_{h_1(B_1 \cap B_2^*)} \mu^{\omega}_{G_2}(A)d\mu = \int_{B_2 \cap B_1^*} \mu^{\omega}_{G_2}(A)d\mu = \\ &= \int_B \mu^{\omega}_{G_2}(A)d\mu < \alpha\ \mu(B)\end{aligned}$$ ∎

Proof of Theorem 1: The proof of Theorem 1 relies on the following lemmas.

Lemma 6: If $S_1 \supset S_2 \supset S_3 \ldots$ is a decreasing (with respect to inclusion) sequence

of σ-fields, then for any integrable random variable X

$$E(X|S_i) \underset{i\to-\infty}{\longrightarrow} E(X|\cap_{i=1}^{\infty} S_i) \ \mu\text{-a.e.}$$

Lemma 6 is a straightforward application of the convergence theorem for reversed martingales (see, e.g., Theorem 35.9 in Billingsley, third edition).

Lemma 7: If μ is ϵ-mixing, then for any $A \in F_1$

$$\left|\mu(A) - E(1_A|F^{\text{tail}})(\omega)\right| \leq \epsilon.$$

Proof of Lemma 7: For any $n, m \in N$

$$\begin{aligned}\left|\mu(A) - E(1_A|F^{\text{tail}})(\omega)\right| \ \leq \ & \left|\mu(A) - E(1_A|\vee_{j=n}^{n+m} F^j)\right| + \\ & + \left|E(1_A|\vee_{j=n}^{n+m} F^j) - E(1_A|F^{\text{tail}})\right|. \qquad (1)\end{aligned}$$

According to the definition of ϵ-mixing for any $k \in N$, one could take n large enough so the first summand on the right side of (6) is less or equal to $\frac{k+1}{k}\epsilon$; i.e.,

$$\left|\mu(A) - E(1_A|\vee_{j=n}^{n+m} F^j)\right| < \frac{k+1}{k}\epsilon \qquad (2)$$

We turn to treat the second summand. Note that for any n, $E\left(1_A|\vee_{j=n}^{n+m} F^j\right)$ is a bounded martingale with respect to the filtration $(\vee_{j=n}^{n+m} F^j)_{m=1}^{\infty}$. Therefore it converges to $E(1_A|\vee_{t=n}^{\infty} F^t)$. By applying Egorov's on a set A_k satisfying $\mu(A_k) > 1 - \frac{1}{k}$, for any given n one can choose m large enough such that

$$\left|E(1_A|\vee_{t=n}^{n+m} F^j) - E(1_A|\vee_{t=n}^{\infty} F^t)\right| < \frac{\epsilon}{k} \qquad (3)$$

As $\vee_{t=n}^{\infty} F^t \searrow_{n\to-\infty} F^{\text{tail}}$, by Lemma 6 we have $E(1_A|\vee_{t=n}^{\infty} F^t)$ converges a.e. to $E(1_A|F^{\text{tail}})$. So (by similar arguments) one can take n large enough to satisfy

$$\left|E(1_A|\vee_{t=n}^{\infty} F^t) - E(1_A|F^{\text{tail}})\right| < \frac{\epsilon}{k} \qquad (4)$$

on a set B_k satisfying $\mu(B_k) > 1 - \frac{1}{k}$. Finally, Equations (2), (3), (4) and (5) give us

$$\left|\mu(A) - E(1_A|F^{\text{tail}})\right| < \frac{k+3}{k}\epsilon$$

on $A_k \cap B_k$. Note that $\mu(A_k \cap B_k) > 1 - \frac{2}{k}$. As this holds for all $k \in N$, the result is proven. ∎

Proof of Theorem 1: $\forall \epsilon > 0$ and μ-a.e. $\omega \in \Omega$ there exists n_0 such that for all $n > n_0$, $\mu_n(\cdot)$ (defined as $\mu(\cdot|P_n(\omega))$ is ϵ-mixing. By Lemma 5

$$\sup_{A \in F^{n+1}} \left| \mu_n(A) - E(1_A | F^{\text{tail}} \vee F_n) \right| < \epsilon,$$

which is exactly the required condition for learnability. ∎

Proof of Theorem 2: Since $\mu^{\omega}_{F^{\text{tail}}}$ has a trivial tail for μ-a.e. ω,[2] the theorem follows from the fact that for any measure ν, mixing is equivalent to having a trivial tail. (See Smorodinsky (1971).) ∎

Proof of Theorem 3: The following lemmas are useful in proving Theorem 3.

Lemma 8 $AT_G(\omega) = AT_G(\omega')$ implies $\mu^{\omega}_G = \mu^{\omega'}_G$.

Proof: Look at arbitrary $\omega_1, \omega_2 \in \Omega$ such that $AT_G(\omega_1) = AT_G(\omega_2)$. For an arbitrary set $B \in F$, denote $\alpha = E(1_B \mid G)(\omega_1)$ and $C = \{\omega : E(1_B \mid G)(\omega) = \alpha\}$. Obviously, $C \in G$. Definitely $\omega_1 \in C$ and so $AT(\omega_1) \subset C$. But $\omega_2 \in AT(\omega_2) = AT(\omega_1) \subset C$ and so $E(1_B \mid G)(\omega_2) = \alpha$. As ω_1, ω_2 and B were chosen arbitrarily we are finished. ∎

Lemma 10: $\tilde{\nu}$ merges with ν if, and only if, $\forall \epsilon > 0$ and ν-a.e. $\omega \in \Omega$ $\exists T = T(\epsilon, \omega)$ such that $\forall t \geq T$

$$\sup_{A \in F_{t+r}} |\tilde{\nu}(A|H_t) - \nu(A|H_t)| < \epsilon$$

Proof:[3] In the following, we make use of Lemma 1 from Kalai and Lehrer (weak and strong merging). For the reader's convenience, we restate this as Lemma 11 below.

Lemma 11: Let g_t be a sequence of measurable functions that converge ν-a.e. to $k \neq 0$. For every $\epsilon > 0$ there is a time T such that

$$\nu(\{\omega \mid \nu(C_t \mid H_{t-1}(\omega)) > \epsilon \text{ for at least one } t \geq T\}) < \epsilon$$

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

[3] We thank Ehud Lehrer for this proof.

where

$$C_t = \left\{ \omega' \mid \frac{g_s(\omega)}{k(\omega)} - 1 \mid \ > \epsilon \text{ for some } s \geq t \right\}.$$

We use Lemma 11 with respect to the following sequence of functions:

$$g_t(\omega) = I(\{\omega \mid \exists\ A \in F_{t+1} \mid \tilde{\nu}(A \mid H_t) - \nu(A|H_t) \mid > \epsilon\})$$

In this case, the set C_t can be rewritten as

$$C_t = \left\{ \omega \mid \sup_{l, A \in F_{t+l+1}} \mid \tilde{\nu}(A \mid H_{t+l}) - \nu(A \mid H_{t+l}) \mid > \epsilon \right\}$$

and

$$\overline{C_t} = \left\{ \omega \mid \sup_{l, A \in F_{t+l+1}} \mid \tilde{\nu}(A \mid H_{t+l}) - \nu(A \mid H_{t+l}) \mid \leq \epsilon \right\}$$

By Lemma 11, there exists t_0 such that $t > t_0$ implies

$$\nu\{\omega \mid \nu(C_{t+1} \mid H_t) < \epsilon \forall t > t_0\} > 1 - \epsilon$$

Substituting $\overline{C}_{t+1}$ for C_{t+1} we get

$$\nu\{\omega \mid \nu(C_{t+1} \mid H_t) > 1 - \epsilon \forall t > t_0\} > 1 - \epsilon.$$

For convenience we shall denote the above set by D_{t_0}, i.e., $D_{t_0} = \{\omega \mid \tilde{\nu}(C_{t+1} \mid H_t) > 1 - \epsilon$ for all $t > t_0\}$. By the assumption of Lemma 10, $\exists t_1 = t_1(\omega, \epsilon)$ such that $t > t_1$ implies that for any l and $A \subset F_{t+l+1}$

$$\begin{gathered}
|\tilde{\nu}(A \mid H_t) - \nu(A \mid H_t)| \\
= \tilde{\nu}(D_{t_0})\nu(C_{t+l} \mid D_{t_0}) \mid \tilde{\nu}(A \mid H_t \cap D_{t_0} \cap C_{t+l}) - \nu(A \mid H_t \cap D_{t_0} \cap C_{t+l}) \mid \\
+\nu(D_{t_0})\nu(C_{t+l} \mid D_t) \mid \tilde{\nu}(A \mid H_t \cap D_{t_0} \cap C_{t+l}) - \nu(A \mid H_t \cap D_{t_0} \cap C_{t+l}) \mid \\
+\nu(D_{t_0})\tilde{\nu}(A \mid H_t \cap D_{t_0}) - \nu(A \mid H_t \cap D_{t_0}) \leq \\
\leq (1 - \epsilon)(1 - \epsilon)\epsilon + (1 - \epsilon)(\epsilon)(2) + (\epsilon)(2) \leq \\
\leq 5\epsilon \blacksquare
\end{gathered}$$

Lemma 10 states that merging assures that predictions of all short run events will eventually be accurate, whether they occur in the near or far future.

To prove Theorem 3, we show that if G_1 and G_2 are both learnable and provide patterns, then $G_1 \sim_3 G_2$.
By Lemma 10, Lemma 2 in Lehrer and Smorodinsky (1996), (*), and learnability, for any m and μ-a.e. ω

$$\lim_t \sup_{r, A \in \vee_{n=1}^m F_{t+r+n}} |\mu^{\omega}_{G_i}(A|H_t) - \mu(A|H_t)| = 0, \ \ \mu_{G_i}\text{-a.e.},$$

Since G_1 provides patterns, it follows that there exists a set C_1 with $\mu(C_1) = 1$ such that for all m and $\omega \in C_1$

$$\lim_t \sup_{r, A \in \vee_{n=1}^m F_{t+r+n}} |\mu(A|H_t)(\omega') - \mu^{\omega}_{G_1}(A)| = 0$$

for any $\omega' \in B_1(\omega)$ where $\mu^{\omega}_{G_i}(B_1(\omega)) = 1$. By Theorem A we may assume, without loss of generality, that $B_1(\omega) \subset AT_{G_1}(\omega)$. By Lemma 8, one can write $\mu^{\omega'}_{G_i}$ instead of $\mu^{\omega}_{G_i}$ in the previous equation. Let $D_1 = \cup_{\omega \in C_1} B_1(\omega)$. Then $\mu(D_1) = 1$, and for all m and $\omega' \in D_1$

$$\lim_t \sup_{r, A \in \vee_{n=1}^m F_{t+r+n}} |\mu(A|H_t)(\omega') - \mu^{\omega'}_{G_i}(A)| = 0.$$

We can find a similar D_2 corresponding to G_2 so that $\mu(D_2) = 1$, and for all m and $\omega' \in D_2$

$$\lim_t \sup_{r, A \in \vee_{n=1}^m F_{t+r+n}} |\mu(A|H_t)(\omega') - \mu^{\omega'}_{G_i}(A)| = 0.$$

Combining the two previous equations, let $D = D_1 \cap D_2$. Then $\mu(D) = 1$, and for all m and $\omega' \in D$

$$\lim_t \sup_{A \in \vee_{n=1}^m F_{t+n}} |\mu^{\omega'}_{G_1}(A) - \mu^{\omega'}_{G_2}(A)| = 0 \blacksquare$$