Preliminary

THE UNBEARABLE FLIGHTINESS OF BAYESIANS: GENERICALLY ERRATIC UPDATING[‡]

MAXWELL B. STINCHCOMBE

Abstract. A decision maker tries to learn the distribution of an observed, utility relevant, independent and identically distributed (iid) sequence of random variables. The random variables have infinite support, and the decision maker learns by updating their prior distribution on the set of distributions of the sequence. For a generic set of priors, Bayesian updating and the corresponding optimizing behavior are wildly erratic.

1. Introduction

In solving dynamic maximization problems, one can either form a complete contingent plan that maximizes expected utility or one can update the prior distribution and then maximize using the posterior distribution. I apply this principle to the study of long run optimal behavior in commonly used class of dynamic programming problems, those with additively separable expected utility representations of preferences, and a utility-relevant, exogenous processes that is iid. The uncommon aspect of the problems under study is that the stochastic law for the exogenous variables is not known, it must learned by updating.

Optimal job search behavior in standard models has reservation wage policies that depend on the distribution of wages, savings, and the leisure-consumption tradeoff. Inventory models with fixed costs of ordering give rise to (s, S) policies that depend on the costs and distribution. When present actions and future realizations determine payoffs, forecasts become valuable. In these three cases, one may see many odd phenomena while the prior information is being replaced by data and the distribution of the exogenous randomness is being learned. One might hope that, as more data accumulates, the posterior distribution converges

Date: April 15, 2005.
[‡]Many thanks to Richard Boylan, Xiaohong Chen, Stephen Donald, Mark Machina, Haskell Rosenthal, Dan Slesnick, Hal White, and Paul Wilson for helpful conversations about this paper. Tom Wiseman and Jeff Ely helped with some difficult terminological issues, but I'm not sure they want to be thanked.

to the true distribution, and the corresponding optima choices converge to the best response to the true distribution.

This paper proves that for a generic set of priors, there is a probability 1 set of histories along which the decision maker becomes, infinitely often, nearly convinced of every element of a dense set distributions for the exogenous sequence. This means that they will choose, infinitely often, all of the actions that are rationalized by a dense set of beliefs. For example, in the job search models, with probability 1, arbitrarily high wages will be rejected, infinitely often, as not good enough, and arbitrarily low wages will be accepted, infinitely often.

Intuitively, this kind of updating arises because a prior over a set of distributions needs to encode relative likelihoods of infinitely many pairs, triplets, etc., of occurences. Generically, some of these relative likelihoods are arbitrarily close to 0. During any history in a set of realizations having probability 1, events with arbitrarily small relative likelihood will occur. The proof shows that this will lead to erratic updating behavior.

The present work can be understood in the context of two results. First, if the decision maker best responds to the empirical distribution, they will, with probability 1, eventually converge to best responding to the true distribution. An expected utility maximizer expects to beat this strategy. Second, suppose that the true distribution of the process is drawn according to some probability that is mutually absolutely continuous with respect to the decision maker's prior. Doob (1949) showed that the updated beliefs form a convergent martingale so that beliefs converge to point mass on the true distribution. The present work entertains the possibility that the decision maker's prior knowledge is not correct, and asks what we can say about the set of beliefs that the decision maker might have that would lead to finding the true distribution.

To specify what it means to be "nearly convinced of every element of a dense set of distributions," let G be an open set of possible distributions of the exogenous process. Beliefs **become nearly certain that G governs the process** if, for a set of histories having probability 1 and for every $\epsilon > 0$, the posterior probability assigned to G is greater than $1-\epsilon$ infinitely often. Since the empirical distribution converges to the true distribution, the (non-Bayesian) beliefs given by point mass on the empirical distribution become nearly certain that G governs if and only if G contains the true distribution. Beliefs are **erratic** if they become nearly

certain that every non-empty open G governs. Since different G can be disjoint, being erratic is a very strong form of failing to settle down. This paper proves:

Theorem. *If the true distribution has infinite support, then the set of erratic priors belongs to the complement of a countable collection of closed convex sets of priors, each of which has an empty interior.*

The next section examines the implications of this result in a three widely used, representative models, and gives some intuition and context for the result. The following two sections contain an examination of the interpretations of prevalence, the notion of genericity used here, as well as the formal statement and proof. The last section concludes.

Throughout, probabilities are countably additive, Borel probabilities, whether specified directly or through use of a density, probabilities are said to converge if and only if they converge weakly, and probabilities on sets of probabilities are defined on the σ-field generated by the topology of weak convergence.

2. Examples, Intuition, Empiricists, and Context

One can see implications of the result proved here in the context of inventory models, search models, and forecasting models. The implications are puzzling because the empiricist's strategy, best responding to the empirical distribution, converges to a best response to the true distribution with probability 1.

There are several intuition about the puzzle. First, when the prior distribution is widely spread, it may well have likelihood in neighborhoods of densities that "wiggle" sufficiently fast to 'track' or 'overfit' the data. Second, even among the priors for which the beliefs converge to something, the mapping from priors to limit points is extremely discontinuous. Third, even in the case of finitely supported observations, convergence of the posterior distribution can be arbitrarily slow. Finally, generically, the relative likelihoods of different realizations come arbitrarily close to 0 and ∞. Updating with extreme relative likelihoods can cause posterior distributions to jump, and this may matter, especially in the case of slow convergence.

From martingale convergence arguments (Doob, 1949), we know that beliefs converge if the true distribution is drawn according to any probability mutually absolutely continuous with the prior. Especially when the prior distribution has full support, this seems to be a strong argument for the consistency of updating.

The present work examines the possibility that the decision maker is mistaken about the process that generated the true distribution.

Though the results are somewhat more general, in the examples, the random variables, X_t, are assumed iid, taking values in a convex subset of $\mathbb{R}^n$ with non-empty interior, $n \geq 1$, and they have a continuous, strictly positive density, f, with respect to Lebesgue measure.

2.a. **Examples: Inventory, Search, Forecasting.** The following is a fairly standard example of an inventory model.

Example 1. *Each period, t, an iid demand, $X_t \geq 0$, is realized. Sales are $Y_t = \min\{X_t, I_t\}$ where $I_t \geq 0$ is the inventory on hand in period t. Any unfilled demand is lost. At the end of each period, as a function of previous history, a re-stocking decision, $R_t \geq 0$, is made, leading to inventories $I_{t+1} = (I_t - Y_t) + R_t$. Revenues in period t are $p \cdot Y_t$. Costs, C_t, have a storage component, $s \cdot (I_t - Y_t)$, and a re-ordering component, $(C + cR_t)1_{\{R_t > 0\}}$. The decision maker maximizes expected discounted profits, $E \sum_t \beta^t (pY_t - C_t)$.*

We assume that the X_t have a continuous strictly positive density f on an interval $[0, M]$ or $[0, \infty)$. For appropriate values of f, p, s, C, and c, it is profitable to run the inventory system at a positive level. If f is known, the optimal rule is an (s, S) inventory policy: order enough to get inventory up to S if it is below s. Or, if you prefer, $R_t^* = (S - (I_t - Y_t)) \cdot 1_{\{I_t - Y_t < s\}}$.

This paper studies the case in which f is not known and the decision maker has a prior distribution over the set of densities. The Theorem given above and the logic that proves the optimality of (s, S) policies for known f imply that, for every prior in a prevalent (generic) set, on a probability 1 set of histories, there are infinitely many points in time when inventory close to 0 is not replenished, and infinitely many points in time that huge purchases are made in the presence of a large inventory.

Example 2. *Each period, t, there is an iid wage offer, X_t. If accepted, one works for $T \geq 2$ periods (T can be random) at wage X_t, then becomes unemployed again, and the new offer X_{t+T} occurs in the following period. If rejected, a new X_{t+1} occurs in the following period. Wage income, Y_t, is the most recent accepted offer if employed during a period and is 0 otherwise. In each period, the decision maker consumes c_t and leisure, ℓ_t. The consumption c_t comes out of savings, S_t, or wage*

income, Y_t*,* $0 \leq c_t \leq S_t + Y_t$*, and* $S_{t+1} = (1+r)(S_t + Y_t - c_t)$*.* ℓ_t *is large in unemployed periods and smaller in employed periods. Utility (felicity) in period* t *is* $u(c_t, \ell_t)$*,* $u(\cdot, \cdot)$ *is concave and increasing, and the decision maker maximizes* $E \sum_t \beta^t u(c_t, \ell_t)$*.*

The easiest assumption on the stochastic X_t is that they have a continuous strictly positive density, f, on some interval $[a, b]$. If f is known, optimal policies involve a reservation wage that depends on the marginal rate of substitution between consumption and leisure in the utility function $u(\cdot, \cdot)$, negatively on the size of the savings account, positively on the riskiness of the X_t. Rothschild (1974) notes the lack of robustness of these search theoretic results, "Almost without exception, these results depend on the untenable assumption that searchers know the probability distribution from which they are searching." The present work shows a bit more.

Suppose that f is not known, and that the decision maker has any prior distribution in the prevalent set of priors here identified. The Theorem given above and the logic behind the optimality of reservation wages for known f imply that, for every $\epsilon > 0$, on a set of histories having probability 1, the searcher will infinitely often accept all wage offers greater than $a + \epsilon$, and will infinitely often reject any wage offer less than $b - \epsilon$ (where b is Bill Gates's annual income).

The need for good forecasting appears in the following class of models.

Example 3. *Each period* t*, an action* a_t *must be taken before the realization of a random* X_t*. The decision maker maximizes* $E \sum_t \beta^t u(a_t, X_t)$*.*

Suppose that $X_t = (R_t, P_t) \in \mathbb{R}^2_+$ is the amount of rain during a growing season and the selling price at harvest time, and that $a_t \geq 0$ is the amount of fertilizer applied. In this case, $u(a_t, X_t)$ might be of the form $P_t G(a_t, R_t) - p a_t$ for a concave growth function $G(\cdot, \cdot)$ and price p for fertilizer. Suppose that $X_t = (\tau_t, H_t) \in \mathbb{R}^2_+$ is the arrival time of the monsoon rains and the random maximal harvest level, and that a_t is the scheduled planting time. In this case, $u(a_t, X_t)$ might be of the form $H_t - r|\tau_t - a_t| \cdot 1_{\{\tau_t < a_t\}} - s|\tau_t - a_t| \cdot 1_{\{\tau_t > a_t\}}$, $r, s > 0$, so that one loses different amounts of the maximal harvest H_t depending on whether the monsoons arrive before or after planting.

In this class of problems, there are no (interesting) dynamics when the X_t are iid with known f — there is no need for forecasting because the optimal policy is stationary, $a_t \equiv a^*$ where a^* solves $\max_a \int u(a, x) f(x) dx$. This implies that,

when the decision maker has any prior in a prevalent subset of the possible priors, for each history in a set having probability 1, infinitely often, the decision maker will best respond to being nearly convinced that the true distribution is in the ϵ-ball aroung every density, g, in a dense set of densities. In the fertilizer example, the decision maker will infinitely often apply only enough fertilizer for desert conditions, and will infinitely often apply enough fertilizer for daily monsoon rains.

2.b. **Intuition.** There are four pieces of intuition for why the main result might be true.

First, when the prior distribution is widely spread, it may well have likelihood in neighborhoods of densities that "wiggle" sufficiently fast to track the data. This would result in updating "overfitting" the data. Second, even among the priors for which the beliefs converge to something, the mapping from priors to limit points is extremely discontinuous. Third, even in the case of finitely supported observations, convergence of the posterior distribution can be arbitrarily slow, essentially because the prior need not put much weight in a neighborhood of the true distribution. When the set of possible distributions is large, one might guess that "not having much weight in the neighborhood" happens often. Finally, generically, priors assign arbitrarily high and arbitrarily low relative likelihoods to different events, and updating uses these comparative likelihoods.

2.b.1. *Wiggles and overfitting.* If the prior distribution is full support, then the neighborhoods of arbitrarily variable densities receive positive weight. For any given realization of draws (i.e. for any given data set), there may well be a 'wiggly' density in the support that looks 'just like' the data. This leads to an 'overfit,' especially if the wiggly density has relatively high prior likelihood.

2.b.2. *A strong sense of discontinuity.* We now show that, for every strictly positive density, g, there is a dense set of priors, $M(g)$, with the property that, no matter what evidence accumulates, beliefs converge to g. Seeing why this is true helps demonstrate no more than plausibility. I emphasize that the priors in this example are **not** the priors that appear in the result, they are but an expository device.

Throughout, we suppose that X_t is an iid sequence of random variables having a continuous, strictly positive density, f, on an interval $[a, b]$. Let P_f^∞ denote

the associated product measure on the sequence space $[a,b]^{\mathbb{N}}$. Let $\mathcal{Z}$ denote the set of continuous densities on $[a,b]$. Thus, $\varphi \in \mathcal{Z}$ iff $\varphi \geq 0$ is continuous and satisfies $\int_{[a,b]} \varphi(x)\,dx = 1$. A prior distribution is a probability μ, belong to the set of probabilities on $\Delta(\mathcal{Z})$. After a t-length history h^t, let $\mu(\cdot|h^t) \in \Delta(\mathcal{Z})$ be the associated posterior distribution.

Let $A \subset \mathcal{Z}$ be the set of continuous densities with holes in their support, that is, densities for which the interior of $\varphi^{-1}(0)$ is non-empty. Note that A is weakly dense in $\mathcal{Z}$. For any strictly positive $g \in \mathcal{Z}$, let $M(g)$ be the set of priors putting positive mass on g and on finitely many points in A. Thus, any $\nu \in M(g)$ is of the form $= \alpha_0\delta_g + \sum_{i=1}^{I} \alpha_i\delta_{a_i}$ where each $a_i \in A$, each $\alpha_i > 0$, and $\sum_{i=0}^{I} \alpha_i = 1$.

Notice that the prior ν has finite support. Further, some of its support points, the a_i, do not have full support. The result here concerns full support priors on the set of densities with full support.

Lemma 1. *For every g, $M(g)$ is dense in $\Delta(\mathcal{Z})$. Further, on a set of histories having P_f^∞-mass 1, for every $\mu \in M(g)$, $\mu(\cdot\big|h^t)$ converges to δ_g.*

Proof: $M(g)$ is dense because any set of probabilities with finite support in a dense set is dense. Fix $\mu \in M(g)$. Let $\varphi_1, \ldots, \varphi_m$ be the points in $\text{supp}(\mu) \cap A$. For each φ_i, let $\mathcal{O}_i$ be a non-empty open subset of $\varphi_i^{-1}(0)$. The waiting time till φ_i is contradicted is a geometric random variable; the waiting time till all of the φ_i are contradicted is the maximum of a finite number of geometric random variables; and this is finite with P_f^∞-probability 1. Thus, for large t, $\mu(\cdot\big|h^t) = \delta_g$ with P_f^∞-probability 1. ∎

2.b.3. *Arbitrarily slow convergence.* Suppose that there are two possible observations, H and T, and that the true probability of each is $\frac{1}{2}$. Suppose that the prior has a density with respect to Lebesgue measure that is of proportional to $f(x - \frac{1}{2})$ where $f(r) \downarrow 0$ very quickly as $r \downarrow 0$. By choosing $f(\cdot)$ appropriately, the posterior distribution will converge to the true distribution as slowly as one wishes.

2.b.4. *Generically extreme relative likelihood ratios.* A prior encodes the likelihoods and relative likelihoods of the X_t landing in infinitely many different sets. We will see that a prior having full support is a generic condition. Provided the prior has full support, the relative likelihoods of disjoint events are necessarily unbounded and arbitrarily close to 0. Since the posterior probabilities involve dividing by the probability of *a priori* unlikely events, it is at least possible that

updating can move the posterior by a huge amount. There is a slightly more detailed way to look at this.

Updating a prior, μ, to a posterior, $\mu(\cdot|h^t)$, after partial history $h^t = (x_1, \ldots, x_t)$ is done using the formula

$$\mu(B\big|h^t) := \frac{\int_B \Pi_{\tau=1}^t \varphi(x_\tau)\, \mu(d\varphi)}{\int_{\mathcal{Z}} \Pi_{\tau=1}^t \varphi(x_\tau)\, \mu(d\varphi)},\ B \subset \mathcal{Z}. \tag{1}$$

Let $G_i = B_\epsilon(g_i)$ be a disjoint pair of ϵ-balls around densities $g_i \in \mathcal{Z}$, $i = 1, 2$. There is a probability 1 set of histories such that $\limsup_t \mu(G_i|h^t) = 1$, $i = 1, 2$. Since the G_i are disjoint, we also know that $\liminf_t \mu(G_i|h^t) = 0$, $i = 1, 2$. Almost all histories have the property that the ratio in (1) moves almost as high and as low as possible infinitely often.

For $\mu(G_i|h^t)$ to be close to 1, we must have μ-most of the higher values of the function $\varphi \mapsto \Pi_{i=1}^t \varphi(x_t)$ concentrated in the ϵ-ball around g_i. Since the X_t are iid f, for large t, we (loosely) expect that $\Pi_{i=1}^t \varphi(x_t)$ takes the value $\int_{[a,b]^t} \Pi_{\tau=1}^t \varphi(x) \Pi_{\tau=1}^t \varphi(x) dx_1 \cdots dx_t$. With μ being full support, the set of φ having very small values in the neighborhood of some/many of the x_t's is strictly positive. For some h^t then, the denominator in (1) will be quite small, and the ratio quite sensitive.

2.c. **The Empiricist's Strategy.** The paradox of erratic updating is that the posterior beliefs, by not settling down, contrast sharply with a major implication of the maintained iid assumption, that the empirical cdf does settle down. It might seem that such behavior has negative implications for the optimality of dynamic expected utility maximization in iid contexts. Let us examine this argument in a version of the need for forecasting seen in Example 3.

Suppose that at each t, a decision maker chooses $a_t \in \{a, b\}$ after having observed the realizations of random numbers, $X_0, X_1, \ldots, X_{t-1}$, in $[0, 1]$ with density f (with X_0 any number in $[0, 1]$). After a_t is chosen, X_t is realized and observed, and the decision maker receives reward $R_t = u(a_t, X_t) \in [0, 1]$. Corresponding to any infinite sequence $((a_t), (x_t))_{t=1}^\infty$ of actions and realizations is the sequence of rewards $\tilde{R} = (R_t)_{t=1}^\infty = (u(a_t, x_t))_{t=1}^\infty$. Sequences of rewards are evaluated using the Bernoulli utility function $V_\beta(\tilde{R}) := (1 - \beta) \sum_t \beta^{t-1} R_t$, $0 < \beta < 1$. The decision maker's preferences over distributions on possible reward sequences are represented by the expected utility function $E^\mu V_\beta(\tilde{R})$ where "E^ν" indicates integration with respect to the measure ν.

Let $\tilde{R}^e$ be the achievable "empiricist" sequence of rewards associated with choosing a_t to best respond to the empirical distribution of the history $h^{t-1} = X_0, \ldots, X_{t-1}$. Let $\tilde{R}^\mu$ be the achievable sequence of rewards associated with choosing a_t to best respond to the posterior $\mu(\cdot|h^{t-1})$. Suppose that a is not a dominant action, i.e. there exists a neighborhood of some $x^\circ \in [0,1]$ such that $u(a,x) < u(b,x)$ for all x in the neighborhood, but that a is the strict best response to f, i.e. $\overline{r} = \int u(a,x)f(x)\,dx > \underline{r} = \int u(b,x)f(x)\,dx$.

Since the true distribution of the sequence of X_t's is P_f^∞, the "objective" maximizing strategy is $a_t \equiv a$, delivering an "objective" expected utility $\overline{r}$. By the Glivenko-Cantelli theorem, with P_f^∞ probability 1, the entries in the vector $\tilde{R}^e$ are $\overline{r}$, in expectation, with at most finitely many exceptions. By contrast, generically, $\tilde{R}^\mu$ is $\underline{r}$ infinitely often. This seems to indicate that patient optimizers will prefer $\tilde{R}^e$ to $\tilde{R}^\mu$, that the consistency embodied in the empiricist strategy leads to higher payoffs, at least for patient decision makers.

Let $\delta_f \in \Delta(\mathcal{Z})$ denote point mass on the distribution f. Formalizing the argument about patient decision makers would require the inequality

$$E^{\delta_f} V_\beta(\tilde{R}^e) > E^{\delta_f} V_\beta(\tilde{R}^\mu),\ \beta \in (\beta^\circ, 1) \text{ for some } \beta^\circ < 1, \tag{2}$$

and it is here that the fallacy appears clearly. It is quite easy to give generic μ's and f in the support of μ satisfying (2). However, an expected utility maximizer chooses actions bearing in mind a wide range of possibilities, knowing full well that this may involve the "wrong" action being chosen from time to time. Evaluating a course of action under δ_f, that is, under certainty about the true distribution is, from the decision maker's point of view, an entirely irrelevant exercise.

2.d. **Context.** There is a very strong consistency result available for Bayesian updating. Suppose, as above, that the decision maker has a prior μ, in $\Delta(\mathcal{Z})$. Suppose further than ν is mutually absolutely continuous with respect to μ, and that some $f \in \mathcal{Z}$ is drawn according to ν. In this context, Doob (1949) showed that the updated beliefs, $\mu(\cdot|h^t)$, is an almost everywhere convergent martingale, and that the limit of the martingale is almost everywhere δ_f.

This can be interpreted as saying that, if the decision maker is "more or less right" about the process(es) that generate f, they will eventually learn f by observing the X_t. Combined with the result here, this means that the set of

ν that are mutually absolutely continuous with μ is a shy set. It also provides another interpretation of the work here.

The present work entertains the possibility that the decision maker's prior knowledge is not correct. Suppose instead of f being drawn according to ν, we imagine that f is determined by some process outside of the decision maker, and that μ represents the beliefs part of the preferences of an expected utility maximizer. With this distinction between the process(es) generating f and the beliefs of the decision maker, it makes sense to ask what happens if the decision maker is mistaken.

3. The Prevalent Properties of Probabilities

This section provides a brief introduction to shyness and prevalence, and to its extension to convex subsets. This tool is here expanded, and then applied to the set of probabilities on a linear space, specifically, to the set of priors, $\Delta(\mathcal{Z})$, that are under study.

When discussing "smallness" of sets in a vector space $\mathfrak{X}$, $\mathfrak{X}$ will always denote an infinite dimensional, locally convex, topological vector space that is also a complete separable metric (csm) space.[1] When discussing "smallness" of subsets of a convex subset of $\mathfrak{X}$, e.g. $\Delta(\mathcal{Z})$, C will always denote a convex subset of $\mathfrak{X}$ that is topologically complete in the relative topology.[2]

3.a. **Large and Small Subsets of $\mathfrak{X}$.**

There are two main notions of small subsets available for $\mathfrak{X}$, a topological and a measure theoretic notion. The complement of a "small" subset is a "large" set.

[1] This class of spaces includes (but is not limited to) $\mathbb{R}^r$ as well as the spaces that appear in most of the theory of non-parametric regression theory: separable Banach spaces such as the $L^p(\Omega, \mathcal{F}, P)$ spaces, $1 \leq p < \infty$, $\mathcal{F}$ countably generated; $C(X)$, the continuous functions with the sup norm when X is compact; $C(X)$ with the topology of uniform convergence on compact sets when X is locally compact and separable (e.g. $X = \mathbb{R}^r$); the Sobolev spaces $S_m^p(\mathbb{R}^r, \mu)$ defined as the metric completion of $C_m^p(\mathbb{R}^r, \mu)$, the space of m times continuously differentiable functions, $m \geq 0$, on $\mathbb{R}^r$ having finite norm $\|f\|_{p,m,\mu} = \sum_{|\alpha| \leq m} \left[\int |D^\alpha f(x)|^p \, d\mu(x)\right]^{\frac{1}{p}}$, $p \in [1, \infty)$, μ a Borel probability measure on $\mathbb{R}^r$; and $C^m(X)$, the space of m times continuously differentiable functions on a compact X with the norm $\sum_{|\alpha| \leq n} \max_{x \in X} |D^\alpha f(x)|$. From Rudin (1973, Theorem 1.24, p. 18), the topology on $\mathfrak{X}$ can be metrized by a translation invariant metric $d(\cdot, \cdot)$, that is, $d(x, y) = d(x + z, y + z)$ for all $x, y, z \in \mathfrak{X}$. Whenever a metric on $\mathfrak{X}$ is in use, it is translation invariant.

[2] A subset of a metric space is **topologically complete** if there exists a complete metric inducing the topology.

The topological notion of smallness is called *meagerness* is due to Baire (1899, §59-61, pp. 65-67). The measure theoretic notion of smallness is called *Haar zero* sets is due to Christensen (1972, 1974, Ch. 7). A description and use of an earlier, and slightly more restrictive definition of small sets can be found in Aronszjan (1976), it's equivalence with several other definitions is covered in Csörnyei (1999). The properties of Haar zero sets and several applications were more thoroughly investigated under the name of *shy* sets by Hunt, Sauer and Yorke (HSY, 1992), who especially applied these techniques to the study of the generic behavior of dynamical systems. There are subtle and difficult problems in extending shyness to a definition of non-generic for subsets of convex subsets of vector spaces that are themselves shy, e.g. spaces of probability measures. These problems were discovered and resolved by Anderson and Zame (2001).

3.b. **Meager and Residual Sets.** A closed set with no interior seems small.

Definition 1. *A set S is* **nowhere dense** *if its closure has no interior. A set S is* **meager** *if it can be expressed as a countable union of nowhere dense sets. A set E is* **residual** *or* **Baire large** *if it is the complement of a meager set.*

Baire large sets are, equivalently, the countable intersection of open dense sets. The countable union of meager sets is meager, the countable intersection of Baire large sets is a Baire large set. Baire's Theorem shows that residual sets are dense, and to some extent this justifies thinking of residual sets as being "large" or "generic". Baire large sets can have Lebesgue measure 0 and seem quite small in $\mathbb{R}^k$ ($k < \infty$ throughout).

Example 1. *Let q_n be an enumeration of the vectors in $\mathbb{R}^k$ with rational coordinates. For any rational $\epsilon > 0$, let E_ϵ be the union of open balls centered at q_n, $\cup_n B(q_n, \epsilon/2^n)$. E_ϵ is an open dense subset of $\mathbb{R}^k$ having Lebesgue measure less than ϵ. The set $E = \cap_\epsilon E_\epsilon$ is a Baire large set having Lebesgue measure* 0.

3.c. **Shy and Prevalent Sets.** For $\mathbb{R}^k$, we have the following.

Lemma 2. *For a universally measurable $S \subset \mathbb{R}^k$, the following are equivalent*

a. $\Lambda^k(S) = 0$ where Λ^k is k-dimensional Lebesgue measure,

b. $P(S) = 0$ where P is a probability with an everywhere positive density with respect to Λ^k, and

c. there exists a compactly supported probability η such that $\eta(S+x)=0$ for all $x \in \mathbb{R}^k$.

Proof: The equivalence of (1) and (2) is immediate. The following proof of the equivalence of (1) and (3) is directly from Hunt, Sauer, and Yorke (1992), and is reproduced here for completeness.

If $\Lambda^k(S)=0$, take $\eta = U^k$, the uniform distribution on $[0,1]^k$. If there is a compactly supported η such that $\eta(S+x)\equiv 0$, then $\int_{\mathbb{R}^k} \eta(S+x)\, d\Lambda^k(x) = 0$, so that $\int_{\mathbb{R}^k}\left[\int_{\mathbb{R}^k} 1_{(S+x)}(y)\, d\eta(y)\right]\, d\Lambda^k(x)=0$, implying $\int_{\mathbb{R}^k}\left[\int_{\mathbb{R}^k} 1_{(S+x)}(y)\, d\Lambda^k(x)\right]\, d\eta(y) = 0$. Since $1_{(S+x)}(y) \equiv 1_{(S-y)}(x)$ and $\int_{\mathbb{R}^k} 1_{(S-y)}(x)\, d\Lambda^k(x) = \Lambda^k(S-y)$, we have $\int_{\mathbb{R}^k} \Lambda^k(S-y)\, d\eta(y) = 0$. Since Λ^k is translation invariant and non-negative, $\Lambda^k(S-y)\equiv \Lambda^k(S)\geq 0$, $\int_{\mathbb{R}^k}\Lambda^k(S)\, d\eta(y)=0$, implying that $\Lambda^k(S)=0$. ∎

In $\mathbb{R}^k$, this ties together the Lebesgue measure definition of smallness, (a), a probabilistic interpretation of smallness, (b), and a translation invariance definition, (c). As we will see, the first two have no direct generalizations to $\mathfrak{X}$, but the translation invariance condition does generalize.

Let $L:\mathbb{R}^k \to V$ be continuous and linear, and let U^k be the uniform distribution on $[0,1]^k$. Taking η to be the $L(U^k)$, the image law of U^k under L, is so useful that it merits a special name.

Definition 2 (Christensen. HSY)**.** *A subset S of a universally measurable $S' \subset \mathfrak{X}$ is* **shy** *if there exists a compactly supported probability η such that $\eta(S'+x)=0$ for all x. S is* **finitely shy** *if η can be taken the continuous linear image of U^k for some k. The complement of a (finitely) shy set is a* (**finitely**) **prevalent** *set.*

From HSY (1992, Facts $2'$ and $3''$), no S containing an open set can be shy in $\mathfrak{X}$ so that prevalent sets are dense, and countable unions of shy sets are shy, equivalently, countable intersections of prevalent sets are prevalent

3.d. **Shy Subsets of Convex Sets.** For a convex $C \subset \mathbb{R}^k$, the appropriate definition of shy subsets of C uses $\mathbf{aff}\,(C)$, the smallest affine subspace containing C, and the lower dimensional Lebesgue measure on $\mathbf{aff}\,(C)$.

Example 4. *For $x \neq y \in \mathbb{R}^2$, let $C = \{\alpha x + (1-\alpha)y : \alpha \in [0,1]\}$ and $S' = \{\alpha x + (1-\alpha)y : \alpha \in [\frac{1}{2},1]\} \subset C$ so that $dim\,(C) = 1 < 2 = k$. S' is a shy subset of $\mathbb{R}^2$ so that S' can be expressed as the intersection of a shy set and C, meaning that S' would be shy if we used that definition. Here $\mathbf{aff}\,(C) = \{\alpha x + (1-\alpha)y : \alpha \in \mathbb{R}\}$, and both C and S' have positive Lebesgue measure in $\mathbf{aff}\,(C)$, so that S' is* ***not*** *a shy subset of C.*

An example directly relevant to this paper demonstrates that the affine subspace approach does not generally work in $\mathfrak{X}$.

Example 5. *Let $\mathfrak{X}$ be the set of countably additive, finite, signed measures on $2^{\mathbb{N}}$, $C = \Delta(\mathbb{N}) \subset \mathfrak{X}$ to be the probability measures on $2^{\mathbb{N}}$. $\Delta(\mathbb{N})$ is a convex, finitely shy subset of $\mathfrak{X}$ even though $\mathbf{aff}(\Delta(\mathbb{N})) = \mathfrak{X}$. To se that $\Delta(\mathbb{N})$ is finitely shy, let η be the uniform distribution on the line L joining the 0 measure and any point mass, δ_n. For any $x \in \mathfrak{X}$, $L \cap (\Delta(\mathbb{N}) + x)$ contains at most one point, so that $\eta(\Delta(\mathbb{N}) + x) = 0$.*

Working from the "outside" of C, that is, with $\mathbf{aff}(C)$, is not appropriate in $\mathfrak{X}$. Anderson and Zame (2001) give a definition of shy subsets of convex sets that works from the "inside." For any $c \in C$, C convex, and any $\epsilon \in (0,1)$, the set $\epsilon C + (1-\epsilon)c$ is a convex subset of C. This is a version of C that is shrunk toward c. For a measurable $C \subset \mathfrak{X}$, $\Delta^K(C)$ denotes the compactly supported probability measures on C. Recall the maintained assumption that C is a convex subset of $\mathfrak{X}$ that is topologically complete in the relative topology.

Definition 3 (Anderson and Zame)**.** *A subset S of a universally measurable $S' \subset C$ is* **shy relative to** C **at** $c \in C$*, or simply* ***shy at*** *c if C is clear from context, if for all neighborhoods U_c of c, and all $\epsilon > 0$, there exists a $\eta \in \Delta^K(C)$ satisfying*

1. *a support condition, $\eta(U_c \cap [\epsilon C + (1-\epsilon)c]) = 1$, and*
2. *a translation invariance condition, $(\forall x \in \mathfrak{X})[\eta(S' + x) = 0]$.*

S is ***shy*** *if it is shy at all $c \in C$. S is* **finitely shy relative to** C *if there exists a if $\eta \in \Delta^K(C)$ that is the continuous affine image of U^k for some k such that $(\forall x \in \mathfrak{X})[\eta(S' + x) = 0]$. The complement of a (finitely) shy set is a* (**finitely**) **prevalent** *set.*

Anderson and Zame (2001) demonstrate the following:

Fact 0: If S is shy at some $c \in C$, then it is shy.
Fact 1: Every subset of a shy set is shy.
Fact 2: If S is shy in C, then for all $x \in \mathfrak{X}$, $S + x$ is shy in $C + x$.
Fact 3: The countable union of shy sets is shy.
Fact 4: No relatively open subset of C is shy in C.
Fact 5: If $\mathfrak{X} = \mathbb{R}^n$ and $\text{int}(C) \neq \emptyset$, then $S \subset C$ is shy iff $\lambda(C) = 0$.
Fact 6: If S is finitely shy, then it is shy.

For convex S, finite shyness is related to the existence of an algebraic interior.

Lemma 3. *A convex $S \subset C$ is a finitely shy subset of C iff it has has empty algebraic interior relative to C.*

Proof: The algebraic interior of $\overline{\mathbf{co}}\, S$ relative to C is empty iff there exists a point $x \in C$ with the property that no line through x intersects $\overline{\mathbf{co}}\, S$ in more than a single point. One point subsets of lines are Lebesgue null sets. ∎

3.e. **Co-Shy Sets.** Some small, convex sets of probabilities are not shy.

Example 6. *Let $C = \Delta(\mathbb{R})$ be the set of countably additive, Borel probability measures on $\mathbb{R}$. For any $r \in \mathbb{R}$, let δ_r be the probability assigning mass 1 to the set $\{r\}$. For any $\epsilon \in (0,1)$, the closed, convex set $S_{\epsilon,r} = \epsilon C + (1-\epsilon)\delta_r$ has non-empty algebraic interior, empty relative interior, and is not shy.*

To see why $S_{\epsilon,r}$ is not shy, set $c = \delta_r$ and consider a candidate η to prove shyness. η must satisfy the support condition, $\eta(U_c \cap [\epsilon C + (1-\epsilon)c]) = 1$, and therefore assigns mass 1 to $S' = S_{\epsilon,r}$. Taking $x = 0$, we have a violation of the translation invariance condition, $(\forall x \in \mathfrak{X})[\eta(S' + x) = 0]$.

$S_{\epsilon,r}$ can be re-written as $\{\mu \in C : \mu(\{r\}) \geq (1-\epsilon)\}$. Since r is a single point in a vector space, the set of probabilities assigning mass to that point is intuitively small. More generally, for $\nu \in C$, the set $S_{\epsilon,\nu} = \epsilon C + (1-\epsilon)\nu$ cannot be shy, and $S_{\epsilon,\nu} = \{\mu \in C : \forall A \subset \mathbb{R},\ \mu(A) \geq (1-\epsilon)\nu(A)\}$.

To deal with such problems, we introduce a slightly larger class of "small" sets.

Definition 4. *A set $S' \subset C$ is **co-shy in C** if it is a subset of a universally measurable S where $S = T \cup V$, T a shy set and V a countable union of closed, convex sets with empty relative interior.*

Directly from the definition and the properties of shy sets, we have

Co-Fact 1: Every subset of a co-shy set is co-shy.
Co-Fact 2: If S is co-shy in C, then for all $x \in \mathfrak{X}$, $S + x$ is co-shy in $C + x$.
Co-Fact 3: The countable union of co-shy sets is co-shy.
Co-Fact 5: If $\mathfrak{X} = \mathbb{R}^n$ and $\operatorname{int} C \neq \emptyset$, then $S \subset C$ is co-shy iff $\lambda(C) = 0$.

Lemma 4 (Co-Fact 4)**.** *No relatively open subset of C is co-shy in C.*

Proof: There are two steps, showing that C is not co-shy in itself, and showing that this gives the desired result.

Step 1: C is not co-shy in itself.

Let $V = \cup_{m\in\mathbb{N}} V_m$ be a countable collection of closed convex subsets of C having empty relative interior. It is sufficient to show that $T^\circ := C \setminus V$ is not shy.

[Fill in, seems tricky]

Step 2: No relatively open subset of C is co-shy in C.

Since $\mathfrak{X}$ is locally convex, $Q \subset C$ is relatively open iff it contains the intersection of a convex open U and C. Let $C' = U \cap C$. Suppose, for the purposes of establishing a contradiction, that Q is co-shy in C. This implies that C' must be co-shy in C. This in turn implies that C' can be expressed as $[T \cup V] \cap C'$. This implies that C' is co-shy in itself, contradicting Step 1. ∎

In Banach spaces, the classes of shy and co-shy sets need not be equal.

Example 7 (Borwein and Noll)**.** *Let $\mathfrak{X} = c_0$ be the set of sequences converging to 0 with the sup norm. $\mathfrak{X}_+$ is convex, closed, and has empty interior, hence is co-shy. Suppose that $K \subset \mathfrak{X}$ is compact. For each n, let x_n be the minimum of 0 the n'th component of any element of K. By compactness, x_n is finite and $x_n \to 0$. Let $x = (-x_1, -x_2, \dots)$ and observe that $K + x \subset \mathfrak{X}_+$. Therefore $\mathfrak{X}_+$ contains a translate of any compact set, hence is not shy.*

Matouškova and Stegall (1996) show that the previous kind of example can only arise in non-reflexive Banach spaces. Specifically, their Theorem 6 shows that a separable Banach space is non-reflexive iff it contains a closed convex set with no interior that contains a translate of every compact set. This is some comfort.

3.f. **Prevalent Properties of Probabilities.** When M is a finite set, the set of full support probabilities in $\Delta(M)$ is prevalent. The analogous result for more general X also holds.

Lemma 5. *Let C be a convex subset of $\mathfrak{X}$ that is topologically complete in the relative topology. The full support distributions are a prevalent subset of $\Delta(C)$, and the non-atomic probabilities are shy in $\Delta(X)$.*

Proof: Fix a non-empty open $G \subset C$. Let $H = H(G)$ be the set of μ such that $\mu(G) = 0$. Pick full support ν_1 and ν_2 with the property that $\nu_1(G) \neq \nu_2(G)$. Because $\nu_1(G) \neq \nu_2(G)$, for any finite, signed measure x, the function $\alpha \mapsto [\alpha\nu_1(G) + (1-\alpha)\nu_2(G) + x(G)]$ is affine with a non-zero slope. Therefore, the set of $\alpha \in [0,1]$ such that $\alpha\nu_1 + (1-\alpha)\nu_2 + x \in H$ contains at most 1 point. Therefore, $H(G)$ is finitely shy. Let $\{G_n : n \in \mathbb{N}\}$ be a basis for the topology of C. The set of distributions in $\Delta(C)$ that fail to be full support is $\cup_n H(G_n)$, the countable union of shy sets, hence shy. The second statement is Lemma 2 in Stinchcombe (2001). ∎

3.g. **Interpretational Issues.** In $\mathbb{R}^k$, Lemma 2 ties together Lebesgue measure, a probabilistic interpretation, and a translation invariant property of the smallness of a set S. Lebesgue measure fails to extend to $\mathfrak{X}$ because there is no translation invariant measure on $\mathfrak{X}$ assigning positive mass to every open set. If there were, it would have to assign equal, and strictly positive mass to every open ball $B(x, \epsilon/4)$. Since $\mathfrak{X}$ is infinite dimensional, every $B(y, \epsilon)$ contains countably many disjoint balls with radius $\epsilon/4$, and the measure assigned to every open set would therefore be infinite.

Probability measures on csm spaces are tight, that is, for every $\epsilon > 0$, there is a compact set, F_ϵ, with $P(F_\epsilon) > 1 - \epsilon$. Probabilistic interpretations fail to extend to $\mathfrak{X}$ directly because the tightness of any probability P on $\mathfrak{X}$ implies that $P(S) = 1$ for S being the countable union of compact, hence shy, sets. Probabilistic interpretations of shyness also fail to be approximately true.

Let Y_i be an independent and identically distributed (iid) sequence of random variables distributed P. Suppose that that $r_n \to 0$, and that $N_n \to \infty$. A point $x \in \mathfrak{X}$ is (r_n, N_n)-**lonely** if $P^\infty(A(x)) = 0$ where $A(x) = [A_n(x) \text{ i.o.}]$, $A_n(x) = \{d(Y_i, x) < r_n \text{ for some } i \leq N_n\}$. In other words, the x is lonely if, with probability 1, $B(x, r_n)$ eventually receives no more visits from $Y_1, \ldots, Y_{N_n}$. Stinchcombe (2001) shows that, no matter how slowly r_n goes to 0 or how quickly N_n goes to ∞, a prevalent set of points are (r_n, N_n)-lonely.

4. The Generic Inconsistency

Fix an infinite, locally compact, complete, separable metric (csm) space (X, d) with $\mathcal{X}$ denoting Borel sigma-field. (A space is locally compact if every point has a neighborhood with compact closure. The spaces $\mathbb{R}^\ell$ and $\mathbb{N}$ are locally compact, infinite dimensional topological vector spaces are not.) $\Delta(X)$ denotes the set of (countably additive, Borel) probabilities on $\mathcal{X}$. An iid sequence of draws, $(Y_n)_{n\in\mathbb{N}}$, is made according to a distribution $\theta \in \Delta(X)$, and θ^∞ denotes the corresponding product distribution on $X^{\mathbb{N}}$. Prior beliefs, μ, are points in $\Delta(\Delta(X))$, the set of distributions on the set of distributions on X. Both $\Delta(X)$ and $\Delta(\Delta(X))$ are csm's in the weak* topology.

Define $\theta_\mu \in \Delta(X)$ by $\theta_\mu(E) = \int_{\Delta(X)} \theta(E)\, d\mu(\theta)$ for $\mu \in \Delta(\Delta(X))$. If $\theta_\mu \not\gg \theta$, the choice of version of the conditional probabilities matters quite sharply because, with continuous random variables, one typically updates after observing a null set. To relate μ to the updated, versions of μ conditional on finite histories

of draws, one must make a whole system of coordinated choices of versions. This will be done by assuming that the μ's under study put mass 1 on the dense set of θ's having continuous densities with respect to some full support reference measure. Throughout, a simple reference case has X countable and discrete, in which case the σ-finite reference measure, λ, can be taken to be counting measure. In this reference case, all probabilities have densities with respect to λ so that densities can effectively disappear from the analysis.

In use are the following assumptions and notation.

A. λ is a full support, σ-finite reference measure on $\mathcal{X}$. $C(X)$ is the set of continuous functions on X. $C_+^\lambda \subset C(X)$ is the set of non-negative f such that $\int_X f\, d\lambda = 1$, $C_{++}^\lambda \subset C_+^\lambda$ is the set of strictly positive f. Each $f \in C_+^\lambda$ is associated with a probability $\theta_f \in \Delta(X)$ defined by $\theta_f(A) = \int_A f\, d\lambda$. When X is countable and discrete, $C_+^\lambda = \Delta(X)$.

B. Both C_{++}^λ and C_+^λ are G_δ's in the csm $\Delta(X)$, implying that there are complete separable metrics, d_{++} and d_+ inducing the weak* topology. (A G_δ is a countable intersection of open sets. The relative topology on any G_δ in a csm can be metrized with a complete separable metric. It is easy to give explicit metrics making C_+^λ and C_{++}^λ into csm's.)

C. $\mathbb{M}^\lambda \subset \Delta(\Delta(X))$ denotes the set of probabilities on probabilities $\Delta(C_+^\lambda)$, while $\mathbb{M}_{++}^\lambda$ denotes $\Delta(C_{++}^\lambda)$. $\mathbb{M}_B^\lambda$ are those for which Bayes updating using densities will never involve division by 0, formally,

$$\mathbb{M}_B^\lambda = \{\mu \in \mathbb{M}^\lambda : \forall (x_1, \ldots, x_t)\ \mu(\{f : \Pi_{i=1}^t f(x_i) > 0\}) > 0\,\}.$$

From the definitions, $\mathbb{M}_{++}^\lambda \subset \mathbb{M}_B^\lambda \subset \mathbb{M}^\lambda$. It can be shown that $\mathbb{M}_{++}^\lambda$ is a convex, topologically complete, prevalent subset of the convex csm $\mathbb{M}^\lambda$, and $\mathbb{M}_B^\lambda$ is a G_δ, hence topologically complete.

Assuming that $\mu \in \mathbb{M}_B^\lambda$, updating after partial history $h^t = (x_1, \ldots, x_t) \in X^t$ is done using the values of the densities at h^t and the prior, μ,

$$\mu(B\big|h^t) := \frac{\int_B \Pi_{i=1}^t f(x_t)\, d\mu(f)}{\int_{C_+^\lambda} \Pi_{i=1}^t f(x_t)\, d\mu(f)},\ B \subset C_+^\lambda. \tag{3}$$

Definition 5. *For any $\theta \in \Delta(X)$, Cons(θ) $\subset \mathbb{M}_B^\lambda$ denotes the set of μ in $\mathbb{M}_B^\lambda$ that are* **consistent for** *θ, that is, the set of beliefs that satisfy $\mu(\cdot\big|h^t) \to_{w^*} \delta_\theta$ θ^∞-a.e. A pair (μ, θ) is* **erratic**, **wildly inconsistent**, **fickle**, *or* **faddish**, *written $\mu \in \mathrm{err}(\theta)$, if for all non-empty open subsets G of $\Delta(X)$, $\limsup_t \mu(G\big|h^t) = 1$ θ^∞-a.e.*

Being erratic is a very strong form of failing to be consistent.

Theorem. *For any full support $\theta \in \Delta(X)$, $\mathrm{err}(\theta)$ is the complement of a countable union of closed convex sets with empty interior in $\mathbb{M}_B^\lambda$.*

Proof: Fix a full support θ and a countable collection of $f_n \in C_{++}^\lambda$ such that the $\theta_n := \theta_{f_n}$ are dense in $\Delta(X)$, and $\theta_n \neq \theta$ for all $n \in \mathbb{N}$.

Abuse notation with $f_n(h^t) := \Pi_{i=1}^t f_n(x_i)$ for partial histories $h^t = (x_1, \ldots, x_t)$. Let $U_{n,m}$ be a nested sequence of open neighborhoods of θ_n not containing θ in their closure and having diameter less than $1/m$. Let $v_{n,m} : \Delta(X) \to [0,1]$ be a continuous function taking the value 1 on $U_{n,m}$ and 0 on the complement of $U_{n,m-1}$. For each h^t, define the continuous function $V_{n,m}(\cdot, h^t)$ on $\mathbb{M}_B^\lambda$ by

$$V_{n,m}(\mu, h^t) = \left(\int_{C_+^\lambda} v_{n,m}(f) f(h^t)\, d\mu(f)\right) / \left(\int_{C_+^\lambda} f(h^t)\, d\mu(f)\right).$$

This is the expected value of $v_{n,m}$ conditional on h^t when beliefs are μ. If posterior beliefs along a sequence of histories h^t converge to δ_{θ_n}, then $\lim_t V_{n,m}(\mu, h^t) = 1$.

For $\epsilon > 0$ and $t \in \mathbb{N}$, define $S_{n,m}(\epsilon, t) = \{\mu \in \mathbb{M}_B^\lambda : \int_{X^\infty} V_{n,m}(\mu, h^t)\, d\theta^\infty(h^t) \leq \epsilon\}$. By continuity, $S_{n,m}(\epsilon, t)$ is a closed subset of $\mathbb{M}_B^\lambda$.

Outline:

1. For all $\epsilon > 0$ and for all t, $S_{n,m}(\epsilon/2, t) \subset \overline{\mathbf{co}}\, S_{n,m}(\epsilon/2, t) \subset S_{n,m}(\epsilon, t)$. This intermediate result leads to
2. For all T, $\bigcap_{t \geq T} S_{n,m}(\epsilon, t)$ is shy, equivalently, $\bigcup_{t \geq T} S_{n,m}(\epsilon, t)^c$ is prevalent.
3. $\mathrm{err}(\theta) = \bigcap_{\epsilon,n,m,T} \bigcup_{t \geq T} S_{n,m}(\epsilon, t)^c$ (the intersection taken over rational ϵ in $(0,1)$). Since the intersection of countably many prevalent sets is prevalent, this and the second step complete the proof.

Details:

1. For all $\epsilon > 0$ and all t, $S(\epsilon/2, t) \subset \overline{\mathbf{co}}\, S(\epsilon/2, t) \subset S(\epsilon, t)$.
   Since each $S(\epsilon, t)$ is closed, showing $\mathbf{co}\, S(\epsilon/2, t) \subset S(\epsilon, t)$ is sufficient. Pick $\mu, \mu' \in S(\epsilon/2, t)$ and $0 < \alpha < 1$. Let $\mu\alpha\mu' = \alpha\mu + (1-\alpha)\mu'$. What must be shown is $\int m(\mu\alpha\mu', h^t)\, d\theta^\infty(h^t) \leq \epsilon$. For numbers $s, s' > 0$ and $r, r' \geq 0$ $\frac{\alpha r + (1-\alpha) r'}{\alpha s + (1-\alpha) s'} \leq \max\{\frac{r}{s}, \frac{r'}{s'}\} \leq \frac{r}{s} + \frac{r'}{s'}$. This delivers

$$\begin{aligned}
&\int m(\mu\alpha\mu', h^t)\, d\theta^\infty(h^t) \\
&\quad = \int \frac{\int v_{n,m}(f) f(h^t)\, d\mu\alpha\mu'(f)}{\int f(h^t)\, d\mu\alpha\mu'(f)}\, d\theta^\infty(h^t) \\
&\quad \leq \int \frac{\int v_{n,m}(f) f(h^t)\, d\mu(f)}{\int f(h^t)\, d\mu(f)}\, d\theta^\infty(h^t) + \int \frac{\int v_{n,m}(f) f(h^t)\, d\mu'(f)}{\int f(h^t)\, d\mu'(f)}\, d\theta^\infty(h^t) \\
&\quad \leq \epsilon/2 + \epsilon/2 = \epsilon,
\end{aligned}$$

completing the proof of the first step.

2. For all T, $\bigcap_{t\geq T} S_{n,m}(\epsilon,t)$ is co-shy.
   We have $\bigcap_{t\geq T} S_{n,m}(\epsilon,t) \subset \bigcap_{t\geq T} \overline{\mathbf{co}}\, S_{n,m}(\epsilon,t) \subset \bigcap_{t\geq T} S_{n,m}(2\epsilon,t)$. by the definition of $\overline{\mathbf{co}}$ and Step 1. The set $\bigcap_{t\geq T} \overline{\mathbf{co}}\, S_{n,m}(\epsilon,t)$ is convex and closed. Lemma 1 shows that it has no topological interior.
3. $\mathrm{err}(\theta) = \bigcap_{\epsilon,n,m,T} \bigcup_{t\geq T} S_{n,m}(\epsilon,t)^c$ (the $\epsilon \in (0,1)$ and rational).
   Since the θ_n are dense in $\Delta(X)$ and the diameters of the $U_{n,m}$ converge to 0, every non-empty open U contains a $U_{n,m}$. Therefore, $\mu \in \bigcap_{\epsilon,n,m,T} \bigcup_{t\geq T} S_{n,m}(\epsilon,t)^c$ iff for all rational ϵ in $(0,1)$, all non-empty open U, and all T, there exists a $t \geq T$ such that $\mu \notin S_{n,m}(\epsilon,t)$, that is, iff $\mu \in \mathrm{err}(\theta)$. ∎

Some technical comments:

A. Because the set of full support μ's is prevalent in $\mathbb{M}^\lambda$, the set of full supported elements of $\mathrm{err}(\theta)$ is prevalent. This means that Theorem **??** does <u>not</u> arise because of some generalized failure of support conditions.

B. In a similar vein, since $\mathbb{M}^\lambda_{++} \subset \mathbb{M}^\lambda_B \subset \mathbb{M}^\lambda$ and $\mathbb{M}^\lambda_{++}$ is a prevalent subset of $\mathbb{M}^\lambda$, $\mathrm{err}(\theta) \cap \mathbb{M}^\lambda_{++}$ is a prevalent subset of $\mathbb{M}^\lambda_{++}$. Also, if $f \in C^\lambda_{++}$, θ_f is full support and $\mathrm{err}(\theta_f)$ is prevalent. Theorem **??** does <u>not</u> arise because the full support θ need be outside the set of probabilities supporting μ.

C. The continuity of the densities can be weakened — the result holds if the set of densities being considered are continuous with respect to a metric for which: (a) λ is still full support, and (b) the Borel σ-field is still $\mathcal{X}$.

D. When $X \subset \mathbb{R}$, the Glivenko-Cantelli theorem tells us that θ^∞-a.e., the empirical cdf converges uniformly to the cdf of θ. Generically, Bayes estimators behave much differently, not converging to the true θ nor to anything else. When $X = \mathbb{N}$, Freedman (1965) shows that a Baire large set of (μ,θ) pairs in $\Delta(\Delta(\mathbb{N})) \times \Delta(\mathbb{N})$ are erratic. This uses a "Fubini" theorem for Baire sets. Anderson and Zame (2001, Example 4, p. 57) show that no such Fubini result is available for prevalent sets.

5. Concluding Remarks

These remarks concern whether or not genericity analysis is sensible in the present context, some connections to the theory of learning, and some possible explanations of fads, bubbles, and other seeming oddities.

5.a. **Does it Make Sense?** A genericity analysis becomes nonsensical if the wrong setting is chosen — if the statistically relevant cases are two dimensional,

then a three dimensional genericity analysis hides more than it reveals. The first challenge then, is to understand the results found here in this light. Perhaps I find that the set of consistent priors is small because I am working in the set of all priors, and the set of all priors is an implausibly large space.

Supporting this point of view are some very attractive classes of priors for which consistency can be guaranteed. They essentially involve the existence of a finite dimensional pattern relating different observations to each other. A leading case is in Freedman (1963) and Diaconis and Freedman (1986a), who discuss the class of **tail-free** priors. Recall that it is, in good part, the near 0 relative likelihoods of rare events that drives the inconsistency.

When $X = \mathbb{N}$, each $\theta \in \Delta(X)$ is specified by the countably many numbers $\theta(n)$. Let $S_k(\theta) = \sum_{n \leq k} \theta(n)$. Picking a θ according to μ gives rise to countably many random variables $Y_k = (1 - S_{k-1}(\theta))^{-1}\theta(k)$. Following Freedman (1963), a prior μ is **tail-free** if $\mu(S_k < 1) = 1$ for all k, and there exists a K such that the random vector $(\theta_k)_{k=1}^K$ and the random variables $Y_{K+1}, Y_{K+2}, \ldots$ are mutually independent (see Ferguson (1973) for a wide set of applications of these ideas). With tail-free priors, large observations in $\mathbb{N}$ have no information about the smaller observations, and there is essentially only a finite dimensional set of relations between the different observations. The extent to which these intuitions generalize to more general X is somewhat unclear. However, the wiggles in densities that overfit can be thought of as happening along linearly independent dimensions in the space of densities.

However, it is not the intuitive or finite dimensional aspect of the set of priors that is always at work in giving inconsistency. Arnold *et al.* (1984) and Diaconis and Freedman (1986b) for very natural settings (competing risks and location estimators respectively) in which Bayes estimators are not consistent. In sum, limiting the set of priors can, in many cases, provide computationally tractable, parametrized statistical models. As a general model of optimizing behavior, such a step is clearly unsatisfactory. Further, even in these parametrized models, Bayesian updating can be inconsistent. These considerations lead (me) to the conclusion that the genericity of inconsistent updating is not an artifact of the wrong setting being chosen, but reflects something more fundamental about the models we use.

5.b. **Some Connections to Learning.** Bayesian updating and optimization in the face of uncertainty are intimately tied, nowhere more so than in the theory of learning. Nachbar's (1997) crucial result for infinitely repeated games is that, when combined with optimization, Bayesian updating of priors about other players' repeated game strategies often leads the players to play strategies that others were certain were not going to be played. Generic inconsistency implies that for interesting single agent games, Bayesian updating is "objectively" sub-optimal. An open question is how much real difference this makes, after all, "infinitely often" need not mean "a non-vanishing fraction of the time."

Consider, for example, Example 3, in which, during each period t, an action a_t must be taken before the realization of a random X_t, and the decision maker maximizes $E\sum_t \beta^t u(a_t, X_t)$. Let $a_t^* = a^*(h_{t-1})$ be optimal given beliefs after h_{t-1}, and let a^t be optimal given the true distribution. It is conceivable that $\forall \mu \in \Delta^{++}(\mathcal{Z}_{++})$ and for all $\epsilon > 0$, $\limsup_t P^\infty(h : |u(a_t^*, X_t) - u(a_t^f, X_t)| > \epsilon) = 0$. Establishing the truth or falsity of this conjecture is a topic for future research, and the work of Lijoi *et. al.* (2004) may be relevant.

5.c. **Fads, Bubbles, and Other Oddities.** Consider again Example 3, but now suppose that the set of rationalizable actions is infinite. Another way to understand the oddity of erratic updating is that as new observations on the same process arrive, beliefs will wander arbitrarily far away from the historical record infinitely often, and actions will follow them.

If a large population of people behaves in so erratic a fashion, one is tempted to look for a model of irrationality. Responses against irrational explanations have included a number of fully rational models of bank runs, informational cascades, bubbles, or exogenous sources of variability such as sunspots. One point to be taken from the present result is that one will, generically, observe a huge variety of rational behavior, even in the quite limited case of iid observations and unchanging, time-separable utility functions.

References

Aronszjan, N. (1976). "Differentiability of Lipschitzian mappings between Banach spaces," *Studia Mathematica*, LVII, 147-190.

Anderson, R. M. and W. R. Zame (2001). "Genericity with Infinitely Many Parameters," *Advances in Theoretical Economics*: Vol. 1: No. 1, Article 1.

Arnold, B. C., P. L. Brockett, W. Torrez, and A. L. Wright (1984). "On the Inconsistency of Bayesian Non-Parametric Estimators in Competing Risk/Multiple Decrement Models," *Insurance, Mathematics and Economics*, **3**, 49-55.

Baire, R. (1899). "Sur les Fonctions de Variables Réelles," *Annali di Matematica Pura ed Applicada* Series 3, Vol. 3, 1-122.

Borwein, J. M. and D. Noll (1994). "Second Order Differentiability of Convex Functions in Banach Spaces," *Trans. Amer. Math. Soc.* **342** 43-81.

Christensen, J. P. R. (1972). "On Sets of Haar Measure Zero in Abelian Polish Groups," *Israel J. Math.* **13**, 255-260.

— (1974). *Topology and Borel Structure.* Amsterdam: North-Holland Publishing Company.

Csörnyei, M. (1999). "Aronszajn null and Gaussian null sets coincide," *Israel J. Math.* **111**, 191-201.

Diaconis, P. and D. Freedman (1986a). "On the Consistency of Bayes Estimates," *Annals of Statistics* **14**(1), 1-26.

— (1986b). "On Inconsistent Bayes Estimates of Location," *Annals of Statistics* **14**(1), 68-87.

Doob, J. L. (1949). "Application of the theory of martingales," in *Le Calcul des Probabilités et ses Applications*, 22-28. Colloques Internationaux du Centre National de la Recherche Scientifique, Paris.

Ferguson, T. (1973). "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics* **1**, 209-230.

Freedman, D. (1963). "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case I," *Annals of Mathematical Statistics* **34**, 1386-1403.

Freedman, D. (1965). "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II," *Annals of Mathematical Statistics* **36**, 454-456.

Hunt, B. R., T. Sauer, and J. A. Yorke (1992). "Prevalence: A Translation-Invariant 'Almost Every' on Infinite-Dimensional Spaces," *Bulletin (New Series) of the American Mathematical Society* **27**, 217-238.

Lijoi, A., I. Prünster and S. G. Walker (2004). "Extending Doob's consistency theorem to nonparametric densities," *Bernoulli* **10**(4), 651-663.

Matouškova, E. and C. Stegall (1996). "A Characterization of Reflexive Banach Spaces," *Proceedings of the American Mathematical Society*, **124**(4), 1083-1090.

Nachbar, J. (1997). "Prediction, Optimization, and Learning in Repeated Games," *Econometrica*, **65**(2), 275-309.

Rothschild, M. (1974). "Searching for the Lowest Price When the Distribution of Prices is Unknown," *Journal of Political Economy*, 82(4).

Rudin, W. (1973). *Functional Analysis*. McGraw-Hill, New York.

Stinchcombe, M. (2001). "The Gap Between Probability and Prevalence: Loneliness in Vector Spaces," *Proceedings of the American Mathematical Society* **129**, 451-457.

— (2005). Strong Approximations to Infinite Games with Type-Dependent Strategies, working paper, Department of Economics, University of Texas at Austin.

Department of Economics, University of Texas, Austin, TX 78712-1173 USA, e-mail: maxwell@eco.utexas.edu