Decision Makers as Statisticians:

Diversity, Ambiguity and Robustness[*]

Nabil I. Al-Najjar[†]

First draft: October 2006
This version: October 2007

Abstract

I study individuals who use frequentist statistical models to draw *secure* or *robust* inferences from i.i.d. data. The main contribution of the paper is a steady-state model in which distinct statistical models are consistent with empirical evidence, even as data increases without bound. Individuals may hold different beliefs and interpret their environment differently even though they know each other's statistical model and base their inferences on identical data. The behavior modeled here is that of rational individuals confronting an environment in which learning is hard, rather than ones beset by cognitive limitations or behavioral biases.

[*] I thank Lance Fortnow, Drew Fudenberg, Ehud Kalai, Peter Klibanoff, Nenad Kos, Charles Manski, Mallesh Pai, and Jonathan Weinstein.

[†] Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston IL 60208.
e-mail: <al-najjar@northwestern.edu>.
Research page : http://www.kellogg.northwestern.edu/faculty/alnajjar/htm/index.htm

Contents

"The crowning intellectual accomplishment of the brain is the real world." [1]

1 Introduction

While classical subjectivist decision theory allows for virtually unlimited freedom in how beliefs are specified, this freedom is all but extinguished in economic modeling. Virtually all equilibrium concepts in economics—be it Nash, sequential, or rational expectations equilibrium—require beliefs to coincide with the true data generating process, reducing any disagreements in beliefs to differences in information.[2,3] On the other hand, there is no shortage of examples in the sciences, business or politics where the way individuals 'look at a problem' and 'interpret the evidence' is just as important in determining beliefs as the data on which these beliefs are based.

To capture this and other related phenomena, I study individuals facing the most classical of statistical learning problems, that of drawing inferences from i.i.d. data. These individuals are modeled as classical (frequentist) statisticians concerned with drawing *secure* or *robust* inferences. The main contribution of the paper is to show that distinct statistical models can be consistent with empirical evidence, even in a steady-state when data increases without bound. Individuals may then hold different beliefs and interpret their environment differently even though they know each other's statistical model and base their inferences on identical data.

Decision makers are assumed to be as rational as anyone can reasonably be. But rationality cannot eliminate the constraints inherent in statistical inference—any more than it can eliminate other objective constraints like lack of information. The methodology advocated in this paper is to study rational individuals confronting environments in which learning is hard, rather than appeal to cognitive limitations or behavioral biases.

[1] G. Miller: "Trends and debates in cognitive psychology," *Cognition*, 1981, vol. 10, pp. 215-25.

[2] In games with incomplete information, this also requires the common prior assumption which dominates both theoretical and applied literatures.

[3] The points made in this paragraph are not new. But being part of the folklore of the literature, they are hard to trace to original references. The contrast between the subjectivist and equilibrium methodologies is adapted from Hansen and Sargent (2001). For an exposition of the problems with the Bayesian methodology in statistical inference, see Efron (2005)'s presidential address to the American Statistical Society.

1.1 Uniform Learning and its Implications

What makes learning hard? It is intuitive that two individuals with common experience driving on US highways will agree on which side of the road other drivers will use. It is far less obvious that two nutritionists, even when exposed to a large common pool of data, will necessarily reach the same theories about the impact of diet on health. These, and countless other examples like them, suggest that some learning problems can be vastly more difficult than others. It is, however, not at all clear what this formally means: learning the probability of any single event in an i.i.d. setting reduces to the simple problem of learning from a sequence of coin flips. This is so regardless of how 'complicated' the event, the true distribution, or the outcome space is.

Focusing on learning probabilities 'one event at a time,' misses the point, however. Decision making is, by definition, about choosing from a family of feasible acts. From a learning perspective, this raises the radically different and more difficult problem of using one sample to learn the probabilities of a *family of events* simultaneously.

The theory of uniform learning is the formal framework that studies robust (*i.e.,* distribution-free) inference in this context.[4] In Section 2, I use this theory to introduce a simple model of belief formation where probabilities are estimated from empirical frequencies using a frequentist *statistical model*. Any such model gives rise to a *belief correspondence* that maps observations to a set of probability measures consistent with empirical evidence.

The individual chooses confidence levels in his estimates of various events. This choice is trivial when data is abundant and the set of alternatives to choose from is narrowly defined. An example is repeated i.i.d. coin flips or, more generally, a finite outcome space and data that asymptotically increases without bound. In this case, frequentist, Bayesian, and just about any other sensible inference all agree.

The more interesting case is situations characterized by *scarcity of data*. A key insight from the theory of uniform learning concerns the tension between the amount of data, and the 'richness' of the structure of acts the decision maker wants to evaluate. What makes a learning problem hard is not the amount of data per se, but this amount in relation to the 'statisti-

[4] This theory, also known as Vapnik-Chervonenkis theory, and its generalization, the theory of empirical processes, occupy a central role in modern statistics but are relatively unknown to economic theorists. The reason for this is clear and revealing: a Bayesian has no use for uniform learning.

cal complexity' of the alternatives considered. Thus, the impact-of-diet-on-health problem is hard because one is concerned with learning about many events simultaneously, namely how different diets impact individuals with different characteristics. The theory of uniform learning provides a formal framework to make sense of intuitive notions like a set of events is 'rich,' 'hard to learn' or 'statistically complex.'

With abundant data and a narrowly defined decision problem beliefs are (approximately) *determinate*: the set of measures consistent with empirical evidence collapses to a small (ϵ-) neighborhood and little scope for disagreement remains. But when data is scarce and the decision maker has to evaluate a rich set of options, the result is *statistical ambiguity*, in the sense that data is insufficient to pin down beliefs. In this case, different individuals with different statistical models may draw different inferences and hold wildly different beliefs even though they observe the same data and know each others' models.

An implication of this theory is a sort of 'law of conservation of confidence:' as the individual increases confidence in his estimates of the probability of some events, he inevitably decreases his confidence in others.[5] This trade-off has interesting implications for confidence-sensitive decision makers. First, as discussed earlier, beliefs may be under-determined by empirical evidence. Second, although ambiguity about the probability of some events disappears as the number of observations increases, ambiguity about others persists. On the statistically unambiguous events, the decision maker has Bayesian beliefs[6] but this is now a consequence of the learning model rather than an aspect of preferences. Third, coarsening and categorization are necessary for learning. The pervasiveness of categorization seems beyond dispute and does not require a model to establish. Why people categorize is less obvious and is potentially the more important question: do individuals categorize because of computational complexity, limited memory, lack of information? If decision makers are modeled as classical, frequentist statisticians, then categorization is necessary to draw secure inferences. Since the model is free of any presumed a priori structure, any implied categorization reflects individuals' attempt to make sense of their environment (hence the opening quote of this paper). No appeal to computational complexity or

[5] The reader may find it helpful to compare this with ordinary linear regression, where adding more regressors lowers the confidence in the estimated parameters. A key difference here is that our setting is non-parametric. The need for a non-parametric model is discussed at length in Section 3.4.

[6] That is, a single (additive) probability measure that is updated using Bayes rule.

cognitive limitations is made.

In Section 3, I turn to large sample theory, where the main technical contribution of this paper lies. The known theory of uniform learning primarily focuses on the case of finite data and has no bite in the limit, as the amount of data increases. This makes it unsuitable for use in most economic models. On a practical level, large sample theories permit greater tractability and clearer intuitions. More fundamentally, equilibrium notions in economics—*e.g.,* Nash or rational expectations equilibrium—are usually interpreted as capturing insights about steady-state or long-run behavior. A theory of learning in which diversity is nothing more than a temporary phenomenon would be difficult to reconcile with this steady-state interpretation.[7]

1.2 Beliefs and Decisions

The main concern of this paper is with belief formation, with questions like: where do beliefs come from? and what makes them 'reasonable?' An orthogonal, but equally important, question is: what decisions do individuals make given their beliefs? To answer this, a decision making model that combines beliefs and tastes into choices is needed. In Section 4, I introduce a simple framework to integrate uniform learning into standard models of decision making and their applications.

One issue I address using this framework is whether learning considerations lead rational decision makers to hold common beliefs. See Morris (1995) for a survey and synthesis. One of the clearest statements of one side of the argument is Aumann (1987, pp. 12-13):

> *"[T]he CPA expresses the view that probabilities should be based on information; that people with different information may legitimately entertain different probabilities, but there is no rational basis for people who have always been fed precisely the same information to do so."*

At the other end of the argument, Savage (1954) writes:[8]

> *"[I]t is appealing to suppose that, if two individuals in the same situation, having the same tastes and supplied with the same information, act reasonably, they will act in the same way. [....] Personally, I*

[7] This point, often under-appreciated, is discussed at length in Bewley (1988) who introduced the notion "undicoverability" to capture the idea of stochastic processes that cannot be learned from data. His model and analysis are, however, quite different from what is reported here.

[8] Page numbers refer to the 1972 edition, Savage (1972).

> *believe that [such agreement] does not correspond even roughly with reality, but, having at the moment no strong argument behind my pessimism on this point, I do not insist on it. But I do insist that, until the contrary be demonstrated, we must be prepared to find reasoning inadequate to bring about complete agreement. [...] It may be, and indeed I believe, that there is an element in decision apart from taste, about which, like taste itself, there is no disputing." (p. 7)*

In reconciling these conflicting views, it is a good idea to have in mind an explicit model that explains how "probabilities should be based on information." My claim is that, when viewed as statisticians, it is perfectly natural for individuals to hold different beliefs based on identical information. Their statistical models may be interpreted as Savage's "element in decision apart from taste, about which [...] there is no disputing."

1.3 Robust vs. Bayesian Inference

The reader imbued with the Bayesian paradigm may be bewildered by notions of learning and robustness that make no reference to prior beliefs, updating rules and the like. Besides, doesn't the standard Bayesian model already contain a theory of belief formation in the form of updating via Bayes rule?

Separating belief formation from decision making, as done in this paper, may seem like a serious violation of the Bayesian paradigm. Historically, however, Savage conceived his framework as normative, as a way to define rational behavior in situations involving uncertainty, but that is otherwise silent on the question of belief formation. Thus, he writes (1967, p. 307) that the subjectivist view of probability is best thought of as a tool "by which a person can police his own potential decisions for incoherency." This does not seem to commit even a Bayesian to any particular model of belief formation.

A common retort is that Bayesian theory already provides a theory of belief formation via de Finetti's theorem. The need for a separate model of belief formation is obviated, so the argument goes, by assuming a decision maker with exchangeable beliefs who updates his prior using Bayes rule. The effectiveness of this as a 'learning' and 'belief formation' procedure is deeply entrenched in the Bayesian folklore, but it is also a myth. A classic theorem by Freedman (1965), detailed in Section 5.5, shows that Bayesian posteriors are "generically" erratic in a very strong sense whenever the outcome space is infinite.[9] Although there is always room to quibble over the meaning of genericity of beliefs and probability laws, what seems beyond dispute is the impossibility of a general result establishing the consistency of Bayesian updating. In a survey of that literature, Diaconis and Freedman (1986, p. 14) write:

> *"Unfortunately, in high-dimensional problems, arbitrary details of the prior can really matter; indeed, the prior can swamp the data, no matter how much data you have."*

Our intuition, often naively derived from coins and urns, that data eventually swamps the priors is misleading. Freedman (1965) puts it quite vividly:

> *"[F]or essentially any pair of Bayesians, each thinks the other is crazy."*

If one substantially relaxes Bayesian theory, then there is not even a consensus on how to update beliefs to incorporate new evidence.[10] In summary, the erratic nature of Bayesian decision making and its inability to incorporate robustness suggest that one should not be quick to dismiss non-Bayesian inference as irrational.

1.4 Descriptive vs. Normative Interpretations

Readers who declare frequentist statisticians irrational will have a hard time not just with this paper, but with current statistical practice in all empirical fields of enquiry–which is overwhelmingly frequentist.[11] At a minimum, frequentist decision making is worth studying because it is *descriptively* important, and may therefore be a better approximation of how economic actors behave.

There are also reasons to think that a concern for robustness is *normatively* compelling. In his 1951 paper, Savage states that "the central problem of statistics is [..] to make reasonably secure statements on the basis of incomplete information." What applies to statisticians ought to apply just as well to economic actors. The fact that the classic Savage (1954) framework precludes concerns for security led to many subsequent attempts to reintroduce such concerns. These include Bewley's ((1986) and (1988)) studies

[9] Freedman's result holds when the outcome space is the set of integers. It was generalized by Feldman (1991) to complete separable outcome spaces, such as $[0, 1]$ or $\mathcal{R}^n$.

[10] Such as preferences not supported by a single prior, as in ambiguity models, for instance. See Machina (1989)'s classic survey of these issues, especially the resulting problem of dynamic (in-)consistency.

[11] See, for instance, Efron ((2005) and (1986)).

of Knightian uncertainty, the ambiguity models of Schmeidler (1989) and Gilboa and Schmeidler (1989), the macroeconomics literature on robustness and model uncertainty pioneered by Hansen and Sargent (*e.g.,* see their 2001 expository paper), and the econometrics literature that uses minimax regret or other robustness criteria as found, for instance, in Manski (2004).

Although classical frequentist methods dominate empirical studies in economics, they had negligible impact on economic *theorizing*. One area where frequentist-like procedures appear is the literature on learning in games, as in models of fictitious play, adaptive learning and regret matching. Another example is Kreps (1998)'s model of anticipated utility, aspects of which he attributes to the older literature on learning rational expectations.

Hansen and Sargent (2001, p. 215) argue that a fundamental aspect of the economic methodology is what might be termed *inferential symmetry*, namely that "the economist and the agents inside his model [be] on the same footing" and, in particular, that "economic agents share the modelers' doubts" and concerns for robustness. In light of this, it is puzzling that we choose frequentist methods to learn about the behavior of economic agents and their environments, yet assume that these very agents are Bayesians when learning about the same environment.

2 Uniform Learning and Consistency with Empirical Evidence

2.1 Basic Setup

A decision maker faces a set of outcomes X with a σ-algebra of events Σ and a true but unknown probability distribution P in $\mathcal{P}$, the set of probability measures on (X, Σ). Here we focus exclusively on statistical inference and belief formation; decision making is examined in Section 4.

I consider three models, with the third not used until Section 3.3:

1. X_f is a finite set, $\Sigma = 2^{X_f}$ is the set of all subsets, and $\mathcal{P}$ is the set of all probability measures;

2. X_c is a complete separable metric space, $\Sigma = \mathcal{B}$ is the family of Borel sets, and $\mathcal{P}$ is the set of countably additive probability measures;

3. X_d is a countable set, $\Sigma = 2^{X_d}$ is the set of all subsets, and $\mathcal{P}$ is the set of finitely additive probabilities.

To simplify the notation, we will distinguish the probability spaces (as X_f, X_c or X_d) but not the sets of events Σ or probabilities $\mathcal{P}$ as they will always be clear from the context. Definitions, claims and interpretations relating to an outcome space X are meant to apply to all three possibilities listed above.

The decision maker bases his beliefs on repeated i.i.d. sampling from the fixed true distribution P on X. Finite samples of t observations are modeled by conditioning on the first t coordinates of an infinite sample $s = (x_1, \ldots)$. Formally, let S denote the set of all infinite sequences of elements in X, interpreted as outcomes of infinite sampling. I.i.d. sampling under P corresponds to the product probability measure P^∞ on $(S, \mathcal{S})$, where $\mathcal{S}$ is the σ-algebra generated by the product topology.[12]

2.2 Motivation and Intuition

This subsection focuses on the special class of *categorization problems* to introduce and motivate the main ideas.

Definition 1 *A decision maker faces a* categorization problem *if:*

- *X is of the form $Y \times \{a, b\}$ for some set of "instances" Y and two categories, a and b;*
- *The decision maker chooses an element of $\mathcal{F}$, the set of all functions $f : Y \to \{a, b\}$, interpreted as categorization rules;*
- *Given P, his payoff from f is $P\{(y, i) : f(y) = i, i = a, b\}$.*

As an example, consider a stylized investment problem where an investor faces randomly drawn "investment opportunities" from a set Y. A categorization rule is a contingent investment strategy $f : Y \to \{buy, sell\}$. Given such f, a correct categorization is made at $x = (y, i)$ if $f(y) = i$. His payoff is simply the probability of the set of outcomes where he 'gets it right:'

$$A_f \equiv \{(y, i) \in X : f(y) = i\}.$$

To convert this problem into a standard decision theoretic language, a natural state space is $\mathcal{P}$ and acts are functions from $\mathcal{P}$ to monetary payoffs in

[12] Most readers are familiar with these standard concepts in the cases X_f and X_c. Section A.1 provides the requisite background in general enough terms to cover the less familiar case of (X_d, Σ).

$[0,1]$.[13] We restrict attention to *categorization acts*:

$$\xi : \mathcal{P} \to [0,1]$$

such that for every P, $\xi(P) = P(A_f)$ for some $f \in \mathcal{F}$. For a Bayesian decision maker, this is a completely straightforward problem. He would have a belief π over $\mathcal{P}$ and chooses the investment rule that maximizes expected payoff given his belief.

We are interested in the case where no such belief is given as a primitive, but must rather be constructed from experience. Specifically, the decision maker observes a sequence $s^t = (x_1, \ldots, x_t)$ drawn i.i.d. from P. In our investment example, the stationarity of P may be interpreted as consisting of two parts: (a) the description of each investment opportunity is comprehensive enough that no potentially relevant factors are omitted; and (b) the economic fundamentals of what makes a company or a stock profitable are stable. If the underlying distribution is non-stationary, then one would expect learning to be even harder, and for reasons quite distinct from those we wish to emphasize here.

I will take the point of view that the decision maker is a frequentist who is concerned with obtaining secure inferences. Difficulties with the Bayesian procedure of starting with a prior and update it using the data were discussed in the Introduction and will be further elaborated in Section 5.5.

Define the empirical frequency of $A \subset X$ relative to a sample s^t:

$$\nu^t(A, s) \equiv \frac{\#\{(x_1, \ldots, x_t) \cap A\}}{t}.^{14} \tag{1}$$

An application of your favorite version of the weak law of large numbers ensures that $\nu^t(A_f, s)$ is a good estimate of $\xi(P) \equiv P(A_f)$ when t is large. For example, by Chebyshev's inequality one has, for any $f \in \mathcal{F}$:

$$P^\infty\{s : |P(A_f) - \nu^t(A_f, s)| < \epsilon\} > 1 - \frac{1}{4t\epsilon^2}.$$

[13] Assume risk neutrality throughout this example for simplicity.

[14] In the special structure of categorization problems, the data is in the form of a sequence $s^t = \{(y_r, i_r)\}_{r=1}^t$ of t instance-category pairs and the empirical frequency of A_f is

$$\nu^t(A_f, s^t) = \frac{\#\{f(y_r) = i_r\}}{t}.$$

In fact, probabilities can be estimated uniformly regardless of the event or the distribution:

$$\sup_{f\in\mathcal{F}} \sup_{P\in\mathcal{P}} P^{\infty}\{s : |P(A_f) - \nu^t(A, s)| < \epsilon\} > 1 - \frac{1}{4t\epsilon^2}. \tag{2}$$

It is irrelevant whether f is complicated or simple, the outcome space X is finite or infinite, ... etc. The inference about any single rule boils down to estimating the probability of one event, and this is formally equivalent to finding the probability of heads in independent coin flips. This may be one of the reasons behind the commonly held intuition that "people eventually learn."

But choice involves, almost by definition, *evaluating many acts simultaneously.* In the investment example, one has to evaluate the performance of as many candidate investment rules as possible in order to choose the best one. Taking a learning perspective, define the set of ϵ*-good samples* for f as:

$$Good^{t}_{\epsilon,P}(f) \equiv \Big\{s : \big|P(A_f) - \nu^t(A_f, s)\big| < \epsilon\Big\}$$

This is the set of samples on which the empirical frequency of A_f is a good estimate of its true probability. Here the parameter ϵ may be viewed as a measure of the confidence one has in this approximation.

Suppose now we are choosing among rules $\{f_1, \ldots, f_I\}$ and there is enough data to ensure that each event $Good^{t}_{\epsilon,P}(A_{f_i})$ has high probability. Then, by definition, we can confidently assess the performance of any given rule f_i. This, however, says little about how to confidently choose the best rule within the set $\{f_1, \ldots, f_I\}$ since this requires samples that are representative for each event $A_{f_1}, \ldots, A_{f_I}$ *simultaneously.* That is, one has to ensure that the probability:

$$P^{\infty}\left[\cap_i Good^{t}_{\epsilon,P}(A_{f_i})\right]. \tag{3}$$

is large. The fact that each event $Good^{t}_{\epsilon,P}(A_{f_i})$ is large guarantees only that the probability of the intersection is at least $1 - I\epsilon$, a conclusion that quickly becomes useless as the number of rules being compared increases.

The central issue is to determine for what class of events is uniform learning possible. This is illustrated in Figure 1 where the square represents the set of all samples of length t and the comparison is among three events $A_1, A_2,$ and A_3 in the outcome space X. In Figure 1(a) the events $Good^{t}_{\epsilon,P}(A_i),\ i = 1, 2, 3$, coincide so their intersection has probability $1 - \epsilon$. In this case, one has as much confidence in the joint evaluation of the three events as in each event simultaneously. Part (b) illustrates opposite case:

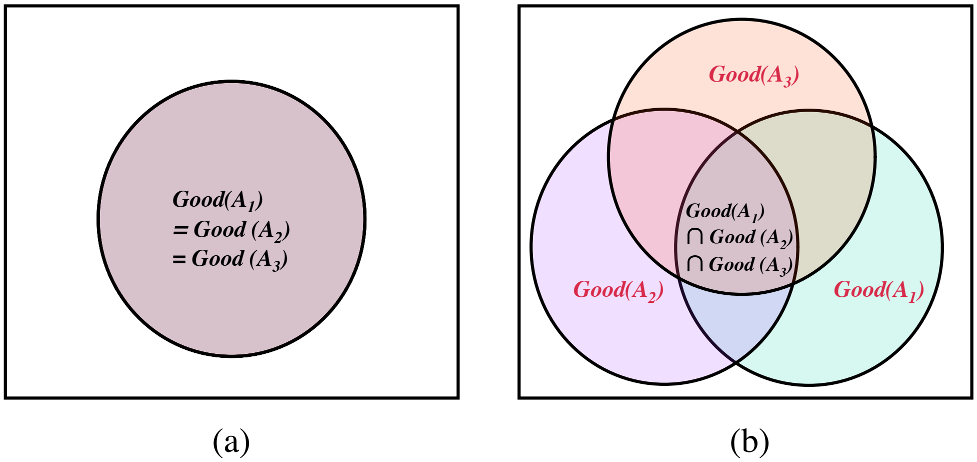

Figure 1: *Two examples of intersections of sets of samples*

(In each case the square represents the space X^t of samples of size t)

each event A_i gives rise to a set of representative samples $Good^{\,t}_{\epsilon,P}(A_i)$ which, by (2), has probability at least $1-\epsilon$. The problem is that these sets of samples stack up in such a way that their intersection has probability of only $1-3\epsilon$.

What determines whether a learning problem falls into type (a) or (b)? The answer is supplied by the beautiful and powerful theory of Vapnik and Chervonenkis (1971) (translated from an earlier paper in Russian).[15] Section 5.1 provides a brief and self-contained account.

2.3 Uniform Learning and Statistical Models

Definition 2 (Uniform Learnability) *A family of subsets $\mathcal{C} \subset 2^X$ is ϵ-*uniformly learnable by data of size t *if,*

$$\sup_{P\in\mathcal{P}} P^\infty \left\{ s : \sup_{A\in\mathcal{C}} \left|P(A) - \nu^t(A,s)\right| < \epsilon \right\} > 1-\epsilon. \tag{4}$$

$\mathcal{C}$ *is* uniformly learnable *if for every $\epsilon \in (0,1)$ there is t such that (4) holds.*

[15] For expositions of this theory, see Vapnik (1998) or Devroye, Gyorfi, and Lugosi (1996).

The crucial aspect of the definition is the location of $\sup_{A \in \mathcal{C}}$, indicating the requirement that the probability being evaluated in (4) is that of samples in which *all events* in $\mathcal{C}$ have their probabilities ϵ-close to their empirical frequencies.

Definition 3 (Statistical Models) *A triple* $(\mathcal{C}, \epsilon, t)$ *is a (feasible)* statistical model *whenever* $\mathcal{C}$ *is* ϵ*-uniformly learnable with data of size* t*.*

For each event A, think of $\nu^t(A, s)$ as a "point-estimate" of $P(A)$ and ϵ as denoting the boundaries of a confidence interval around $\nu^t(A, s)$. Extending this idea to all events, define

$$\mu^t_{\mathcal{C},\epsilon}(s) = \left\{ p \in \mathcal{P} : \sup_{A \in \mathcal{C}} \left| p(A) - \nu^t(A, s) \right| \leq \epsilon \right\}. \tag{5}$$

as the set of distributions *consistent with empirical evidence.* A probability measure that does not belong to $\mu^t_{\mathcal{C},\epsilon}$ is one that can be rejected with high confidence as inconsistent with the data. We suppress reference to $\mathcal{C}$ and ϵ, simply writing $\mu^t(s)$, whenever they are clear from the context.

We shall view the collection of events $\mathcal{C}$ and the degree of confidence ϵ as reflecting the decision maker's model of his environment. On the other hand, the amount of data t is an objective constraint that confronts the decision maker with a trade-off between confidence, measured by ϵ, and the scope of events $\mathcal{C}$ he can learn.

The problem with this logic is that feasibility of a statistical model, by itself, is a hopelessly weak criterion; it is, for instance, trivially satisfied when $\mathcal{C} = \emptyset$ or $\epsilon = 1$.[16] To obtain a meaningful theory, it is useful to introduce the following partial order on statistical models:

Definition 4 *A statistical model* $(\mathcal{C}', \epsilon', t')$ dominates *another model* $(\mathcal{C}, \epsilon, t)$ *if*

- $\mathcal{C} \subseteq \mathcal{C}'$*,* $\epsilon' \geq \epsilon$ *and* $t' \leq t$*.*

$(\mathcal{C}', \epsilon', t')$ strictly dominates $(\mathcal{C}, \epsilon, t)$ *if at least one of the above inequalities holds strictly.*

When considering decision makers' preferences over statistical models in Section 4.4, I will argue that it is normatively compelling that scarce data is not wasted. Thus, the most relevant statistical models must be maximal:

[16] In typical statistical learning theory applications, the family of events $\mathcal{C}$ is exogenously given, such as half intervals in [0,1], or rectangles in $\mathcal{R}^2$. I know of no instance in which the idea of maximality is used in that literature.

Definition 5 (Maximality) *A statistical model $(\mathcal{C}, \epsilon, t)$ is* maximal *if there is no feasible model $(\mathcal{C}', \epsilon', t')$ that strictly dominates it.*

A maximal model does not overlook any sharper inferences that could have been drawn using the same amount of data t.

For a finite outcome space X_f, the existence of a maximal model that dominates a given model $(\mathcal{C}, \epsilon, t)$ is straightforward. This is more delicate on infinite outcome spaces, but still true:

Theorem 1 *Fix t and suppose that $(\mathcal{C}, \epsilon, t)$ is a feasible statistical model. Then there is a maximal feasible statistical model $(\mathcal{C}', \epsilon', t')$ that dominates $(\mathcal{C}, \epsilon, t)$.*

2.4 Learning Complexity, Scarcity of Data and the Order of Limits

A key theme of this paper is that the desire to draw secure inferences from scarce data leads individuals to statistical models that are coarser than the true model. This is captured by the criterion of uniform learnability, which reflects the difficulty of learning when data is scarce relative to the set of options available to the decision maker. Recall that the weak law of large numbers implies that there is $\bar{t}$ such that for all $t > \bar{t}$

$$\sup_{A \in \mathcal{C}} \sup_{P \in \mathcal{P}} P^{\infty}\Big\{s : \big|P(A) - \nu^t(A, s)\big| < \epsilon\Big\} > 1 - \epsilon. \tag{6}$$

This bound pertains to a statistical experiment in which *a fresh sample is drawn for each event evaluated.* Each new event would require $\bar{t}$ new observations, a preposterous amount of data when the decision maker is comparing a large set of acts.

A concern for scarcity of data means that data is not so abundant that one can generate samples at will. The uniform learning criterion

$$\sup_{P \in \mathcal{P}} P^{\infty}\Big\{s : \sup_{A \in \mathcal{C}} \big|P(A) - \nu^t(A, s)\big| < \epsilon\Big\} > 1 - \epsilon \tag{7}$$

models a decision maker who gets *one shot at sampling t observations.* The scarcity of data forces the decision maker to restrict attention to a narrower class of events $\mathcal{C}$.

In summary, modeling environments where learning is hard is an illusive goal because no event, when taken in isolation, is ever hard to learn. Rather, complexity in learning is a property of *families of events* and arises only when the scarcity of data is taken seriously.

Whether a decision problem is complex or not depends on the relationship between the amount of data available and the richness of the set of events considered. This is sharply illustrated in the following theorem:

Theorem 2 *Let $X = X_f$ be a finite outcome space with cardinality n. Then:*

1. *For any given n and $\epsilon > 0$ there is $\bar{t}$ such that 2^{X_f} is uniformly learnable with data of size $t \geq \bar{t}$; and*

2. *Given any t, $\epsilon > 0$ and $\alpha > 0$ there is n such that $\#X_f > n$ implies:*
$$\frac{\#\mathcal{C}}{\#2^{X_f}} < \alpha$$
for any $\mathcal{C}$ that is ϵ-uniformly learnable with data of size t.

The problem is one of order of limits: Holding the finite set of outcomes fixed, taking the amount of data to infinity guarantees uniform learning of the powerset. On the other hand, holding t large but fixed, the set of events and acts that can be uniformly learned shrinks down to zero as the size of the outcome space increases. This is even when t is large enough to guarantee learning the probability of any event in isolation.

If the outcome space is 'small,' as in the case of coin flips, the side of the road drivers are likely to use and so on, then case 1 of the theorem is relevant. Things differ dramatically when the outcome space is vast. Consider, for example, the problem of evaluating the impact of diet on health. If there are z_1 binary attributes that define an individual's characteristics, z_2 binary attributes that define diet characteristics, and z_3 binary attributes that define health consequences, then the cardinality of the finite outcome space is $2^{z_1+z_2+z_3}$. For entirely conservative values of, say, $z_1 + z_2 + z_3 = 50$, the cardinality of the set of events is the incomprehensibly large number $2^{2^{50}}$. For individuals to reach agreement on the probability of all events through learning is more in the realm of fantasy, even by the standard of idealized economic models of decision making.[17]

When part 2 of the theorem is relevant, as in the last example, individuals seeking robust inferences from a large but limited pool of data either restrict

[17] The reader may be amused by the fact that complete 0.01-agreement will require, using (17), a *minimum* t that exceeds the estimated number of minutes since the Big Bang.

the scope of theories $\mathcal{C}$, the set of models they consider $\mathcal{P}$, or both. The key point is that such restrictions must precede empirical evidence. It should therefore not be surprising that rational individuals entertain ambiguity and disagreements even when facing identical information.

2.5 Event-Dependent Confidence

The model presented above is unnecessarily restrictive in that it rules out trading off confidence across events. Consider, for example, a choice between a bet that pays 1 on an event A and zero otherwise, and another that pays 100 on B and -50 otherwise. One would expect a decision maker to demand greater precision about his estimate of $P(B)$ than $P(A)$.

The formalism used so far assumes a common confidence interval size ϵ for all events. This is done for simplicity, and a generalization can be readily made. Instead of a single ϵ applied to a family of events $\mathcal{C}$, confidence is now represented by

- A function $\zeta : \Sigma \to [0, 1]$, representing an event-dependent (size of) confidence interval;
- A constant $\delta \in [0, 1]$ representing a confidence level.

Uniformly learnable would then mean:

$$\sup_{P \in \mathcal{P}} P^{\infty}\Big\{ s : \forall A, \ \big| P(A) - \nu^t(A, s) \big| < \zeta(A) \Big\} > \delta. \tag{8}$$

Our more restrictive formulation $(\mathcal{C}, \epsilon, t)$ is one characterized by $\delta = \epsilon$ and a distinguished family of events $\mathcal{C}$ such that $\zeta(A) = \epsilon$ for each $A \in \mathcal{C}$. The arguments of the paper go through, with appropriate modifications, under the more general model.

3 Large Sample Theory

I now turn to the asymptotic properties of uniform learning as the amount of data increases. There are at least three reasons why large outcome spaces are important:

- *Robustness:* A natural question is: would the conclusions of the analysis eventually disappear as more data accumulates?

- *Tractability:* Asymptotic models can be considerably simpler and yield sharper intuitions.
- *Applicability:* Equilibria in economic and game theoretic models are often viewed as steady-states that arise as limits of learning processes.

3.1 Exact Learning

As the decision maker is given more data, he can sharpen his statistical model by either decreasing ϵ or increasing the range of events $\mathcal{C}$ he learns about. We formalize this using the notion of learning strategy:

Definition 6 *A* learning strategy *is a sequence $\sigma \equiv \{(\mathcal{C}_t, \epsilon_t, t)\}_{t=1}^{\infty}$ of statistical models satisfying:*

- $\epsilon_t \to 0$*;*
- $\mathcal{C}_t \subseteq \mathcal{C}_{t+1}$ *for every t;*
- $\mathcal{C}_t$ *is a maximal ϵ_t-uniformly learnable family of events by data of size t.*

The learning strategy is simple *if there is $\bar{t}$ such that $\mathcal{C}_t = \mathcal{C}_{t+1}$ for every $t \geq \bar{t}$.*

The idea is that, as the decision maker is given larger sets of data, the set of feasible statistical models increases. His choice from the larger set of models may either involve increasing confidence or enlarging $\mathcal{C}$. Simple strategies involve increasing confidence while holding the set of events $\mathcal{C}$ constant.

Given a learning strategy $\sigma = \{(\mathcal{C}_t, \epsilon_t, t)\}_{t=1}^{\infty}$, the set of *beliefs consistent with empirical evidence* is:

$$\mu_\sigma(s) \equiv \left\{ p : \forall t \ \sup_{A \in \mathcal{C}_t} \left| p(A) - \nu^t(A, s) \right| \leq \epsilon_t \right\}.$$

The next theorem shows that on a ‘typical’ sample, any $p \in \mu_\sigma(s)$ should assign to any event A a probability equal to the true probability $P(A)$, and thus $\mu_\sigma(s)$ has a very clean structure on most samples:

Theorem 3 (Exact Learning) *Fix any learning strategy $\{(\mathcal{C}_t, \epsilon_t, t)\}$ and write $\mathcal{C} = \cup_t \mathcal{C}_t$. Then for any $P \in \mathcal{P}$:*

$$\mu_\sigma(s) = \left\{ p : p(A) = P(A), \forall A \in \mathcal{C} \right\}, \qquad P^{\infty} - a.s.$$

In particular, $\mu_\sigma(s)$ is a convex set of probability measures, almost surely.

The main challenge in proving this result is to show that it holds for finitely additive P on X_d, as required in Section 3.3 below. Note also that agreement on $\mathcal{C}$ may lead to agreement on events outside $\mathcal{C}$. See Section 4.3.

The theorem justifies the following straightforward definition:

Definition 7 *Beliefs are* asymptotically determinate *if there is a learning strategy σ such that for every P,*

$$\mu_\sigma(s) = P, \qquad P^\infty - a.s.$$

That is, the only belief consistent with empirical evidence is the true distribution.

The phenomenon of most interest to us is beliefs that are not determinate. An easy consequence of Theorem 2 is that this cannot be achieved if X is finite. To model environments where learning is hard, data is scarce, and beliefs are indeterminate, we need to turn to infinite outcome spaces.

3.2 Continuous Outcome Spaces

An obvious candidate for an infinite outcome space is X_c, a complete separable metric space with the Borel σ-algebra $\mathcal{B}$. $\mathcal{P}$ is the set of all (countably additive) probability measures on $(X_c, \mathcal{B})$. We begin with a general, and discouraging, result:

Theorem 4 *If X is a complete metric space then beliefs are asymptotically determinate via a simple learning strategy.*

A prototypical example illustrating the theorem is:

Example 1 *Let $X = [0,1]$ and $\mathcal{C}$ be the class of half intervals: $[0,r]$ or $(r,1]$ where r is any number in $[0,1]$. Then this is an uncountable collection of events that is uniformly learnable.*

There are two distinct learning principles at play in this example:

- *Statistical learning:* by the classic Glivenko-Cantelli Theorem, the empirical distribution functions converge to the distribution function uniformly almost surely, so in the limit the probability of each half interval is known without error;
- *Deduction:* once the probabilities of events in $\mathcal{C}$ are known one can use the axioms of probability to deduce the probabilities of events outside $\mathcal{C}$. In this example, this leads to deducing the probabilities of all Borel events. The theorem shows that this intuition generalizes.

The argument underlying the theorem shows that complete learning in the limit can be achieved using an exceedingly simple class of events, similar in their simplicity to the half intervals. It is difficult to think of bounded rationality reasons that would prevent a decision maker from using such simple learning procedure.

The example is disturbing in another way, namely that it reveals a rather sharp disconnect with the finite outcome space/finite data model. It is easy to find examples of finite outcome space and finite data in which complete learning does not occur. Yet this cannot occur in the settings covered in Theorem 4. This is an artifact of the mathematical structure of X_c that distorts the learning problem by imposing strong restrictions on Σ and $\mathcal{P}$. This leads to the model of the next section in which complete learning cannot occur, reflecting more faithfully the phenomenon found in finite-finite models.

3.3 Discrete Outcome Spaces

Here we consider X_d to be countable; Σ is the set of *all subsets* of X; and $\mathcal{P}$ the set of all *finitely* additive probability measures on Σ. This space of outcomes is discrete in the sense that there is no extraneous metric or measurable structure that restricts the set of events.

Theorem 5 *Beliefs in the discrete outcome space (X_d, Σ) are not asymptotically determinate.*

In fact, for any learning strategy σ, there is P such that $\mu_\sigma(s) \neq \{P\}$, P-a.s.[18]

It is worth noting that the scope of disagreement asserted in the theorem can be substantial, as shown in the following corollary to its proof:

Corollary 6 *For any uniformly learnable $\mathcal{C}$ and any $\alpha \in (0, 0.5]$ there is a pair of probability measures λ and γ that agree on $\mathcal{C}$, yet $|\lambda(B) - \gamma(B)| = \alpha$ for uncountably many events B.*

To interpret these results, recall the earlier discussion that there are two learning principles: a statistical principle, under which probabilities are deduced from data, and a deductive principle, under which probabilities of some events are deduced from knowledge of the probabilities of others.

[18] This is a stronger claim than just saying that beliefs are not asymptotically determinate, which would have only required that $\mu_\sigma(s) \neq \{P\}$ on some set A with $P(A) > 0$.

Statistical inference works here just like it did in continuous outcome spaces. What is different is that there is no longer a uniformly learnable $\mathcal{C}$ such that the probability of all events can be deduced from knowledge of the probability of events in $\mathcal{C}$.

The proof uses a combinatorial result that bounds the "size" of uniformly learnable classes of events. Passing to the limit is quite delicate because, among other things, the cardinality of a family of events is not a useful measure of its learning complexity.[19] The proof uses a novel argument in which a (finitely additive) measure λ can be perturbed without changing the probability it assigns to events in $\mathcal{C}$.

3.4 Modeling Choices and Generalizations

3.4.1 The Absence of Presumed Structure

A finite outcome space X_f is free from any presumed a priori structure, such as notions of distance, ordering, or similarity between elements. I view a structure-free model as an essential backdrop to any study that seeks to shed light on how individuals model their environment. No one disputes that cognitive structures, like orderings and similarity, are essential in decision making. But to explain why these structures look the way they do, one should avoid letting extraneous presumed structures surreptitiously contaminate the analysis. In a structure-free model, like X_f, individuals end up using orderings and similarity in the form of a statistical model to facilitate learning and to make sense of empirical evidence (hence the opening quote of this paper).

In infinite outcome spaces, the counterpart of X_f is the discrete space X_d, which admits all subsets and all probability measures as legitimate. By contrast, the continuous outcome space X_c comes loaded with structural assumptions. When $X_c = [0, 1]$, for instance, a similarity function in the form of a metric is implied, limiting the range of events, acts, and probabilities used. This accounts for the learning result, Theorem 4, that stands in stark contrast to what happens in large but finite outcome spaces.

3.4.2 Stationarity

The model assumes that the decision maker faces a stationary problem (P is unchanging). Many decision problems may be usefully modeled as sta-

[19] The example in footnote 25 illustrates that knowledge of the probabilities of a countable family may be sufficient to determine the probabilities of all Borel sets.

tionary, while some non-stationary problems become stationary in a richer outcome space. In any event, failure of learning would hold a fortiori in non-stationary settings where the object to be learned is constantly changing.

3.4.3 Robustness

Robustness enters via our assumption that the decision maker is completely ignorant about the true model P, and that he seeks inferences robust to this model uncertainty. This may be viewed as too extreme. In defense of this requirement, consider:

- The model can accommodate the introduction of prior knowledge that narrows down the set of possible distributions. The qualitative insights generalize if we limit the decision maker's model uncertainly to some $\mathcal{P}^\circ \subsetneq \mathcal{P}$, provided this is a rich enough set of distributions.

- But one must then ask where does knowledge of $\mathcal{P}^\circ$ come from? The requirement to be robust to all distributions helps delineate the boundary between empirically-grounded and extra-factual sources of knowledge.

In the investment example, allowing all P's may correspond to a technical investor with no prior theory (*e.g.,* basic economics or finance) that puts a priori restrictions on the true distribution. If the investor were to incorporate theories from macroeconomics or finance, say, he will presumably be able to reduce $\mathcal{P}$ to some smaller set $\mathcal{P}^\circ$. But despite decades of extensive and commonly shared evidence, even the best theories these fields have to offer leave ample room for model uncertainty. This is seen daily in well-publicized conflicting policy recommendations, forecasts, and investment strategies.

4 Diversity, Ambiguity and Decision Making

4.1 A Decision Theoretic Framework

Learning leads to a compact convex set of probability measures $\mu^t(s)$ and $\mu_\sigma(s)$ consistent with empirical evidence. These are purely statistical constructs that impose constraints on beliefs, but otherwise orthogonal to how beliefs are incorporated into choice. To do so requires an explicit decision theoretic framework.

Here I use a simple formulation based on Gajdos, Hayashi, Tallon, and Vergnaud (2006)'s model of how objective information can be incorporated into a subjective setting. Fix a finite set of consequences Z. An act is a function of the form:

$$\xi : \mathcal{P} \to \Delta(Z).$$

Here, $\mathcal{P}$ is interpreted as the set of states and $\Delta(\mathcal{P})$ as an individual's beliefs about these states. For example, ξ may be induced by a categorization rule, as detailed in Section 2.2. Endow $\mathcal{P}$ and $\Delta(\mathcal{P})$ with their weak* topologies,[20] and let $\mathcal{K}$ be the set of all compact and convex subsets of $\Delta(\mathcal{P})$.

Consider now a decision maker with objective information that the true distribution π over the states space $\mathcal{P}$ lies in some compact convex set $\Pi \subseteq \Delta(\mathcal{P})$. Gajdos, Hayashi, Tallon, and Vergnaud (2006) proposed that this decision maker evaluates an act ξ according to:

$$U(\xi) = \min_{\pi \in \varphi(\Pi)} \int_X u \circ \xi(P) \; d\pi(P) \tag{9}$$

where

- u is a vNM utility function; and
- $\varphi : \mathcal{K} \to \mathcal{K}$ maps objective information, in the form of a set of measures Π, to a subjective set of measures $\varphi(\Pi)$, and satisfies

$$\varphi(K) \subseteq K, \quad \forall K \in \mathcal{K}. \tag{10}$$

They provide preference axioms, extending those of Gilboa and Schmeidler (1989), that characterize this representation. Unfortunately, their setup includes assumptions of technical nature that make their preference characterization inapplicable to our problem.[21] Here I use their functional form; verifying whether their representation holds with these technical assumptions removed will be undertaken in future work.

The main innovation here is the specific source of objective information proposed, namely statistical inference from repeated sampling.

Definition 8 (Frequentist restrictions on subjective beliefs)

[20] The topology on $\Delta(\mathcal{P})$ is generated by sets of the form: $\{\pi \in \Delta(\mathcal{P}) : \alpha < \pi(\mathcal{E}) < \beta\}$ where $\mathcal{E} \subset \mathcal{P}$ and $0 \leq \alpha < \beta \leq 1$.

[21] Namely that $\mathcal{P}$ must be countable and $\Delta(\mathcal{P})$ consists of measures of finite support.

- Infinite samples: *Given a strategy* σ *and a sample* s*, the set* $\Pi_\sigma(s) \subseteq \Delta(\mathcal{P})$ *of* beliefs consistent with empirical evidence *consists of all probability measures* $\pi \in \Delta(\mathcal{P})$ *that put mass 1 on* $\mu_\sigma(s)$*.*

- Finite samples: *Given a statistical model* $(\mathcal{C}, \epsilon, t)$ *and a sample* s*, the set* $\Pi^t_\sigma(s) \subseteq \Delta(\mathcal{P})$ *of* beliefs consistent with empirical evidence *consists of all probability measures* $\pi \in \Delta(\mathcal{P})$ *that put mass at least* $1 - \epsilon$ *on* $\mu_{(\mathcal{C},\epsilon,t)}(s)$*.*

Below I focus exclusively on $\Pi_\sigma(s)$ since the case of finite data $\Pi^t_\sigma(s)$ is quite similar. It is straightforward to verify that the mapping

$$s \mapsto \Pi_\sigma(s)$$

is a correspondence assigning to each sample a compact convex set of probability measures. The sets of beliefs $\Pi_\sigma(s)$ do not vary arbitrarily with data. Rather, they all share the property that there is a family of events, independent of s, on which any two measures in $\Pi_\sigma(s)$ must agree.[22]

The functional form generally expressed in (9) above now becomes:

$$U(\xi; s) = \min_{\pi \in \varphi_s(\Pi_\sigma(s))} \int_X u \circ \xi \;\; d\pi. \tag{11}$$

The inclusion condition (10), which now becomes:

$$\varphi_s\left(\Pi_\sigma(s)\right) \subseteq \Pi_\sigma(s) \;\; \forall s,$$

says that the decision maker cannot be completely delusional: he must put no weight on probabilities that are *securely* rejected by available evidence.

If beliefs are asymptotically determinate, as in the finite or continuous outcome spaces, $\mu_\sigma(s)$ is a singleton measure $\bar{P}$. In this case, inclusion forces the decision maker to hold the belief $\delta_{\bar{P}}$ that puts unit mass on $\bar{P}$ in almost all samples. The utility function in (11) implies that decision maker will behave exactly as a Bayesian, almost surely.

If $\mu_\sigma(s)$ is non-degenerate (as in the discrete model, or in a finite outcome space with limited data), then objective information cannot pin down a single distribution P. This leaves the decision maker the freedom to entertain many possible beliefs π; all learning does is to restrict these beliefs to $\Pi_\sigma(s)$. To evaluate acts, as in (11), the decision maker transforms the objective information $\Pi_\sigma(s)$ into a subjective set of measures $\varphi_s\left(\Pi_\sigma(s)\right)$. Two polar cases of such transformation are worth noting:

[22] Lehrer (2005) makes a similar point in a very different context.

- *Bayesian Belief Selection:* φ_s is single-valued.
- *Maximally Ambiguous Beliefs:* φ_s is the identity ($\varphi(\Pi) = \Pi$ for all $\Pi \in \mathcal{K}$).

Note that, in evaluating acts, the taste component u is assumed to be independent of the sample s. Samples only provide information, so it makes sense that they only impact beliefs. On the other hand, we allow φ_s to vary with s to reflect subjective elements of how the decision maker interprets ambiguous objective information. When beliefs are asymptotically determinate, the inclusion condition (10) makes this freedom superfluous. But when empirical evidence is not sufficient to reduce $\mu_\sigma(s)$ to a singleton, the decision maker's subjective "inferences" and interpretations of the evidence may well vary from sample to sample. He may potentially be influenced by unmodeled heuristics, misconceptions, over-confidence, biases involving superstitions, or by reading patterns in otherwise randomly generated numbers. One may debate whether or not doing so is "rational," but such debate would have to appeal to criteria that go beyond the minimalist approach adopted here.

I conclude by noting that the analysis of this paper is not wedded to any particular model of how beliefs are incorporated into decision making. The above model is attractive for its simplicity and tractability, but the main points on the role of uniform learning and the incorporation of objective information could have just as easily been made using smooth ambiguity preferences (Klibanoff, Marinacci, and Mukerji (2005)) or variational preferences (Maccheroni, Marinacci, and Rustichini (2006)).

4.2 Diversity of Beliefs

Should individuals who have observed a large, common pool of data hold the same beliefs? We consider the case of infinite data for simplicity. Similar definitions and argument hold in the finite data case.

Consider two individuals, $i = 1, 2$ who:

1. Face the same unknown environment P;
2. Observe the same infinite sequence of data s;
3. Have identical vNM utility u;
4. Each has a learning strategy σ^i;

5. φ_s^i is single-valued for every s.

Conditions 1-3 are obvious; they rule out differences in the environment, data, or tastes as sources of disagreement. Decision makers who differ on these dimensions are, not surprisingly, likely to disagree on how to evaluate acts, even with infinite data. Condition 4 is definitional. Condition 5 is necessary to even define what "holding the same beliefs" means.

Under these conditions, each individual is a subjective utility maximizer, with a single-valued $\pi_s^i \equiv \varphi_s^i\big(\Pi(s)\big)$ that varies with the sample. We say that two individuals *almost always hold a common belief* if for every P, $\pi_s^1 = \pi_s^2,\ P^\infty - a.s.$

Theorem 7 *Two individuals almost always hold a common belief if either*

- *Beliefs are asymptotically determinate; or*
- *The two individuals use the same learning strategy σ and the functions φ_s^1 and φ_s^2 are equal almost surely.*

The theorem, whose proof is straightforward, highlights the sort of conditions needed for common beliefs to obtain. If beliefs are asymptotically determinate, then $\mu_\sigma(s)$, and hence $\Pi(s)$, is single-valued. Inclusion then forces the two individuals to hold identical beliefs, and evaluate acts identically. Under these assumptions, the only sources of differences in behaviors are differences in tastes or information about future outcome realizations.

But if beliefs are asymptotically indeterminate, then differences in the σ^i's and φ_s^i's can no longer be ignored. Assume first that the two individuals choose identical learning strategies. Then they both have the same objective information $\Pi(s)$ and this set of beliefs exhausts all available statistical evidence. This leaves room for the subjective mappings φ_s^i to play a role. These mappings summarize individuals' reliance on unmodeled heuristics, intuitions or biases to figure out how to assign probabilities to events not pinned down by $\Pi(s)$.

The subjective mappings φ_s^i that drive disagreement are not shaped by evidence, nor are they subject to learning. It would seem rather remote that individuals end up with identical φ_s^i's on their own.

The second, and potentially more radical, source of disagreement is difference in learning strategies. With common learning strategies, individuals must agree on the probability of some events. But when their learning strategies are allowed to differ, then it may so happen that, in a rich enough problem, there is no agreement on any non-trivial event. Again, there is no reason to expect that learning or players' rationality should lead their strategies to merge over time.

The upshot is that, in environments where learning is hard, conditions that guarantee common beliefs are exceedingly demanding because different individuals can look at the same problem differently. One should therefore not be surprised, and indeed should expect, that different individuals with access to a large identical pool of data can reach different conclusions.

4.3 Statistical Ambiguity and Probabilistic Closure

One may interpret the multiplicity of distributions consistent with empirical evidence (*i.e.,* the fact that $\mu_\sigma(s)$ is non-singleton) as indicative of *statistical ambiguity.* Although $\mu_\sigma(s)$ is a purely statistical construct, it does bear on whether ambiguity-sensitive behavior arises. The following result is straightforward:

Theorem 8 *An individual displays ambiguity averse behavior almost surely if:*

- *Beliefs are not asymptotically determinate; and*
- φ_s *is set-valued almost surely.*

There is a large literature on the role of ambiguity in decision making that builds on the insights of Schmeidler (1989) and Gilboa and Schmeidler (1989). This literature, which is too vast for even a cursory review here, characterizes ambiguity aversion in terms of axioms on choice behavior. This paper takes a complementary approach: I put forth an explicit learning model and a decision model, within which ambiguity aversion may or may not arise, then ask *"what must be true about a given environment to cause ambiguity averse behavior arise or vanish?"*

The next natural question is: *"what are the events on which all ambiguity disappears in the limit?"* To make this more formal, fix a learning strategy $\sigma = \{(\mathcal{C}_t, \epsilon_t, t)\}_{t=1}^{\infty}$ and write $\mathcal{C} = \cup_t \mathcal{C}_t$. By Theorem 3, the decision maker learns the probability of all events in $\mathcal{C}$ exactly. Two problems arise: first, $\mathcal{C}$ need not have any particular structure; for example, it need not be closed under complements, unions ... etc. For example, the class of events in Example 1 is not an algebra. Second, agreement on the probabilities of events in $\mathcal{C}$ may lead to agreement on events outside $\mathcal{C}$. For instance, if we know the probability of two events A, $B \in \mathcal{C}$ and these events are *disjoint*, then we can unambiguously deduce the probability of the event $A \cup B$, even if it does not belong to $\mathcal{C}$.

These issues point to the need of formally defining the class of events whose probabilities can be unambiguously determined in the limit. Call a function $p : \mathcal{C} \to [0, 1]$ a *partial probability* if it is the restriction to $\mathcal{C}$ of some probability measure p' on Σ.[23]

Definition 9 *An event $A \in \Sigma$ has* unambiguous probability given $\mathcal{C}$ *if, for any partial probability p on $\mathcal{C}$, and any two extensions p', p'' of p to Σ, $p'(A) = p''(A)$.*

The set $\mathcal{C}^\star$ of all such events will be referred to as the probabilistic closure *of $\mathcal{C}$.*

Obviously, $\mathcal{C} \subseteq \mathcal{C}^\star$ with equality obtaining if $\mathcal{C}$ is an algebra of events in X_d, or a σ-algebra in X_c. What can be said about the structure of $\mathcal{C}^\star$ in general?

There is an increasing agreement in the ambiguity literature that unambiguous events need not form an algebra, but only a λ-system. This observation was first made by Zhang (1999), and subsequently elaborated in many papers, in particular Epstein and Zhang (2001). A λ-system is a family of events closed under complements and *disjoint* unions, but not necessarily arbitrary unions or intersections (Billingsley (1995)).

Theorem 9 *Fix any uniformly learnable family $\mathcal{C}$:*

1. *$\mathcal{C}^\star$ is a λ-system;*

2. *$\mathcal{C}^\star$ may be strictly larger than the smallest λ-system containing $\mathcal{C}$;*

3. *$\mathcal{C}^\star$ need not be an algebra.*

Part (1) is evident. Part (2) is known; an example in de Finetti (1974) illustrates the point. A version of his example, which also illustrates part (3), is reproduced in the appendix.

[23] A more direct condition defining partial probabilities on arbitrary families of sets was identified by Horn and Tarski (1948). See Bhaskara Rao and Bhaskara Rao (1983, Definition 3.2.2).

4.4 Choice over Statistical Models

So far we have examined choice over acts, taking the prior choice of a learning strategy σ as given. Different strategies of course lead to different information structures, begging the question: how a decision maker chooses among alternative learning strategies?

For expositional convenience, restrict attention to statistical models $(\mathcal{C}, \epsilon, t)$ with finite data. Intuitively, a decision maker chooses the model that is most helpful in evaluating the acts of interest to him. One can define a binary relationship $\sqsupseteq$ on statistical models that reflects this preference and introduce standard assumptions like completeness and transitivity. It also makes sense to require $\sqsupseteq$ to be monotone increasing in $\mathcal{C}$ and t, and decreasing in ϵ. This just says that the decision maker prefers, other things held fixed, more data, more events and higher confidence.

Beyond these weak requirements, there seems to be little additional structure on $\sqsupseteq$ that one would be compelled to impose *in general.* In specific contexts, it is not hard to think of idiosyncratic as well as social and competitive factors that shape the choice of a statistical model. Statistical practice offers insights into the process of model selection: practitioners design statistical models as a function of the hypotheses they want to test, intuitions about likely connections, conventions and so on. These are extra-factual considerations that lie outside our minimalist learning model. One can hope that a more ambitious model than the present one can account for patterns of diversity and conformity of individuals' models.

5 Implications and Connections

5.1 Uniform Learning and Vapnik-Chervonenkis Theory

Uniform learning can be given an elegant and insightful characterization using the theory of Vapnik and Chervonenkis. The key concept in this theory is that of *shattering capacity* of a family of events $\mathcal{C}$. Define the m^{th} *shatter coefficient* of a family of sets $\mathcal{C}$ to be

$$s(\mathcal{C}, m) = \max_{\{x_1, \ldots, x_m\} \subset X} \#\{C \cap \{x_1, \ldots, x_m\} : C \in \mathcal{C}\}.$$

Here, interpret $\{x_1, \ldots, x_m\}$ as a potential sample drawn from X. Then $\#\{A \cap \{x_1, \ldots, x_m\} : A \in \mathcal{C}\}$ is the number of subsets that can be obtained by intersecting the sample with some event in $\mathcal{C}$. The shatter coefficient $s(\mathcal{C}, m)$ is a measure of the complexity of a family of events $\mathcal{C}$.

Clearly, $s(\mathcal{C}, m) \leq 2^m$. The highest integer m at which this bound is achieved is called the *Vapnik-Chervonenkis (or VC-) dimension* of $\mathcal{C}$:

$$V_{\mathcal{C}} \equiv \max_m \{s(\mathcal{C}, m) = 2^m\}.$$

If there is no such m, we write $V_{\mathcal{C}} = \infty$.

A central result in statistical learning theory is that a class of events is uniformly learnable if and only if it has finite VC-dimension. In particular,

$$\sup_{P \in \mathcal{P}} P^{\infty} \left\{ s : \sup_{A \in \mathcal{C}} |\nu^t(A, s) - P(A)| > \epsilon \right\} < K\, t^{V_{\mathcal{C}}}\ e^{-t\epsilon^2/32}, \tag{12}$$

where K is some constant. While tighter bounds are available (perhaps under mild additional assumptions), the above version is the most useful for our purposes.[24]

For a family of events $\mathcal{C}$ to have a finite VC-dimension means that it is not too rich to be uniformly learned. A finite $\mathcal{C}$ obviously has finite VC-dimension, while the powerset has VC-dimension equal to the cardinality of the space. But the cardinality of a family $\mathcal{C}$ has at best a tangential (and often misleading) relationship to its statistical complexity.

The best-known class of finite VC-dimension is the half intervals appearing in Example 1, also known as the Glivenko-Cantelli class. This is an uncountable family of events that nevertheless has a VC-dimension of 2. Historically, this was the first class of events for which uniform learnability results were shown. This, and the striking generalization provided by Vapnik and Chervonenkis (1971), spawned the vast literature on the subject.

To see that this class has $V_{\mathcal{C}} = 2$, note that any pair of points $x_1,\ x_2 \in X$ can be shattered by $\mathcal{C}$, so $V_{\mathcal{C}} \geq 2$. Take any set of three points $x_1 < x_2 < x_3$, intersections with elements of $\mathcal{C}$ generate the sets $\{x_1\},\ \{x_3\},\ \{x_1, x_2\},\ \{x_2, x_3\}$, but no intersection can generate the singleton set $\{x_2\}$. Since no set with three points can be shattered, we have $V_{\mathcal{C}} = 2$.[25]

The next example shows that passing from a family $\mathcal{C}$ to the algebra it generates is not innocuous from a learning stand point:

[24] See Devroye, Gyorfi, and Lugosi (1996) and, for another take on the problem, Pollard (1984). A characterization in terms of samples drawn from a given subset of X is given in Talagrand (1987).

[25] Consider the class $\mathcal{C}' \subsetneq \mathcal{C}$ with identical definition as $\mathcal{C}$ but where t is restricted to be a rational number. Then $\mathcal{C}$ and $\mathcal{C}'$ have identical VC-dimension, even though $\mathcal{C}'$ is countable while $\mathcal{C}$ is not. This is another illustration that learning is only tangentially related to the cardinality of the events to be learned.

Example 2 $X = [0, 1]$ *with* $\mathcal{C}'$ *the algebra generated by the Glivenko-Cantelli class* $\mathcal{C}$*. Then* $V_{\mathcal{C}'} = \infty$*. In particular, the family of Borel subsets has infinite VC-dimension.*

To verify the claim, note first that the algebra generated by $\mathcal{C}$ contains all finite unions of intervals. Fix a finite set $\{x_1, \ldots, x_m\}$. It is clear that any subset of $\{x_1, \ldots, x_m\}$ can be expressed as the intersection of $\{x_1, \ldots, x_m\}$ with a finite union of intervals. This means that the algebra generated by $\mathcal{C}$ shatters a finite set $\{x_1, \ldots, x_m\}$ of size m for any m, and hence has infinite VC-dimension. To verify the last claim, note that passing to superset of events can only increase the VC-dimension.

5.2 Over-fitting and Falsifiability

The fundamental trade-off facing the decision maker in this paper is between the desire to learn the probability of as many events as possible, and the fear of over-fitting the data. To see the relationship to over-fitting, I focus on the special case of categorization problems.

Let $\mathcal{F}$ be a class of categorization rules $f : Y \to I$, with $I = \{0, 1\}$. A state is a probability distribution P on the set of outcomes $Y \times I$. Denote its marginal on Y by η_P (or just η, when P is clear). We make a few simplifying assumptions, which only sharpen our point:

- Y is finite;
- All P's have the same marginal η on Y;
- $\eta(y) > 0$ for every y;
- There is $f \in \mathcal{F}$ such that $P(i \,|\, y) = f(y)$ for every $y,\ i$.

The last condition says that the true category is deterministic conditional on y. Under these assumptions, every P may be identified with a true categorization, which we denote by $\bar{f}$.

Fix a small $\epsilon > 0$ and finite t, so the decision maker's statistical model reduces to the set of events $\{A_f : f \in \mathcal{F}^\circ\}$, where $\mathcal{F}^\circ$ is a subset of the set of all categorization rules $\mathcal{F}$.

As in Section 2.2, the decision maker evaluates categorization rules according to the probability of the event A_f that he 'gets it right:'

$$P(A_f) \equiv P\big\{(y, i) \in X : f(y) = i\big\}.$$

Under our simplifying assumptions, this takes a very simple form. Define $m(f, P)(y)$ to be 1 if f and $\bar{f}$ match at y and 0 otherwise. The goal of the decision maker is to match $\bar{f}$ as closely as possible, in the sense of solving:

$$\max_{f \in \mathcal{F}^\circ} \sum_y m(f, \bar{f})(y)\, \eta(y), \tag{13}$$

with η being fixed and known.

How should a decision maker constrain his statistical model $\mathcal{F}^\circ$? By choosing a 'rich' class $\mathcal{F}^\circ$, say the set of all categorization rules $\mathcal{F}$, he would be relaxing the constraint in (13), potentially producing a better choice of f. The problem is that $\mathcal{F}^\circ = \mathcal{F}$ is so rich that it perfectly fits any sample $s^t = \{(y_r, i_r)\}_{r=1}^t$ of t instance-category pairs. This decision maker can rationalize everything but learns nothing. In fact, all he learns from the sample is the values of $\bar{f}$ at the instances $y_1, \ldots, y_t$ but nothing about what to do elsewhere. This is the classic problem of over-fitting.

Learning is fundamentally about generalization, and this is possible only through ex ante restrictions on the set of admissible rules. The theory of uniform learning and the concept of VC-dimension formally delineate what sort of restrictions are needed in order to over-come this over-fitting problem. This is formally done by requiring that $\mathcal{F}^\circ$ be uniformly learnable, in the sense of (4).

One may view a statistical model $\mathcal{F}^\circ$ as a theory and each $f \in \mathcal{F}^\circ$ as an admissible hypothesis or explanation within that theory. The decision maker uses data to decide which explanation within those admissible under his theory $\mathcal{F}^\circ$ has the greatest empirical support. If $\mathcal{F}^\circ$ is very rich, then it can produce an explanation that fits any observed set of data. As a theory, $\mathcal{F}^\circ$ is not falsifiable. Conversely, the requirement that $\mathcal{F}$ has a finite (and, ideally, small) VC-dimension amounts to saying that $\mathcal{F}$ is falsifiable by some realizations of the data. The interpretation of VC theory in terms of falsifiability of theories is further elaborated on in Vapnik (1998) and Harman and Kulkarni (2007).

5.3 Statistical Models, Coarsening, and Information Partitions

A key theme of this paper is that the desire to draw secure inferences from scarce data leads individuals to coarser models of their environment. As an example, consider a categorization problem and a decision maker who chooses a statistical model $(\mathcal{C}, \epsilon, t)$ with small ϵ. Then an ambiguity-sensitive decision maker will strongly favor categorization acts measurable with respect to $\mathcal{C}$. For example, if $\mathcal{C}$ is the algebra generated by some partition $\{A_1, \ldots, A_L\}$ of X, then he will be inclined to choose acts that do not make fine distinctions between outcomes within any given A_l, effectively categorizing outcomes according to the partition $\{A_1, \ldots, A_L\}$.[26]

This coarsening is superficially similar to the representation of incomplete information as partitions of the underlying state space. There are profound differences, however. Information partitions model the *availability* of information, while the coarsening of the state space using statistical models captures constraints on information *processing*. In the model of this paper, individuals may draw different inferences even though they have identical information, know each other's statistical models, and have a common initial understanding of the structure of their environment.

An important and frequently raised concern about non-standard models of behavior is that they often boil down to a model of incomplete information. The Bayesian incomplete information paradigm has been a run-away success precisely because of its flexibility in incorporating a broad range of phenomena previously seen as impervious to analysis in terms of rational choice. The model of this paper differs from incomplete information models just as fundamentally as Bayesian and frequentist approaches to inference differ. The two modeling approaches ask different questions, raise different issues, and reach different conclusions.

5.4 Statistical Models vs. 'Bounded Rationality'

In his survey of the literature, Lipman (1995) describes 'boundedly rational' behavior as "choice that is imperfect in the sense that the output is often not the 'correct' one but is sensible in that it can be understood as an attempt by the agent to do reasonably well." A natural modelling approach is what Lipman refers to as *partitional models* where a decision maker is assumed to be constrained by a partition that reflects his coarse and limited understanding of the environment. The decision maker displays "boundedly rational" behavior in the sense that he chooses from a diminished set of acts (those measurable with respect to his partition). Lipman (1995) observes that much of the 'bounded rationality' literature is of the partitional variety, citing as examples models where decision makers use analogies and costly partitions, or suffer from memory limitations, bounded recall, computability constraints, among others.

[26] In general, $\mathcal{C}$ can have a more subtle structure than an algebra generated by a partition.

It is tempting to think of 'boundedly rational' behavior through the lenses of partitional, incomplete information models. This, I believe, is unfortunate. Lack of information is an objective constraint that limits what an agent can and cannot condition on. The constraints imposed by 'bounded rationality,' on the other hand, have to do with information *processing*, an object that is inherently more nebulous, constantly changing with learning, introspection and competitive pressures. I suspect this is one reason why 'bounded rationality' models are often perceived, fairly or not, as ad hoc.

Al-Najjar, Anderlini, and Felli (2006) explore a class of partitional models of *undescribable events.* Their goal is to model events that can be assigned probabilities, but that cannot be described ex ante relative to a *given* language. By contrast, the present paper explains why learning may cause decision makers to use statistical models that coarsely lump outcomes—independent of any language. The driving force here is the difficulty of learning a family of events rather than the difficulty of describing any single event.

From the perspective language-based models, this paper says that language, or any other cognitive construct that goes beyond the simple-minded counting of frequencies, should matter only when learning is hard. The learning- and language-based approaches are potentially complementary and may be related in some subtle ways.

5.5 Bayesian and Frequentist Beliefs

A true Bayesian would be bemused by the seemingly arbitrary use of frequencies in the learning model of this paper. The decades-old debate between Bayesians and their detractors is well beyond the scope of this paper.[27] There are many basic and well-known reasons why Bayesianism may be problematic, such as the lack of procedure to form priors. Here I elaborate on the intractability of learning in a Bayesian setting.

Suppose a decision maker faces an experiment in which random draws are taken from an outcome space X. As a good Bayesian, he should have a prior belief on the state space S, the space of all infinite sequences of such draws. If he regards the outcomes at each stage as symmetric, then his belief

[27] See for example Efron (1986)'s "Why isn't everyone a Bayesian?" which points out that Bayesianism "has failed to make much of dent in the scientific statistical practice" because objectivity in this practice is key and "by definition one cannot argue with a subjectivist." Efron (2005) advocates a combination of frequentist and Bayesian ideas.

is an exchangeable distribution. By the celebrated de Finetti Theorem[28] his prior is equivalent to a two-stage lottery where he first draws a P according to some probability measure ν on $\mathcal{P}$, then outcomes are generated i.i.d. according to P. In words, exchangeable beliefs must be i.i.d. with unknown parameter P.

De Finetti's theorem is an elegant representation of beliefs on symmetric experiments, but it is not a theory of learning. Granted de Finetti's representation, and using $\nu^t(s)$ to denote the posterior after t observations, the learning question is: given that data is generated according to P, would the posteriors converge to put unit mass on P?

If the true distribution P is compatible with the decision maker's beliefs ν [29] then he ends up learning the true P. A Bayesian, confident ν is the correct model, is convinced he will eventually learn. But what happens if his model is mis-specified?

Suppose that X is a complete separable metric space and endow both $\mathcal{P}$ and $\Delta(\mathcal{P})$ with the weak topology, and $\mathcal{P} \times \Delta(\mathcal{P})$ with the product topology. Interpret a typical element (P, ν) of this space as a true distribution P and a Bayesian belief ν. The following is a startling result on the pathological nature of Bayesian updating: if X is any infinite complete, separable metric space, then for a generic choice of (P, ν) the sequence of posteriors $\nu^t(s)$ visits every open set in $\Delta(\mathcal{P})$ infinitely often P-*a.s.*

This was first shown by Freedman (1965) and later generalized by Feldman (1991). The notion of genericity here is that of a residual set.[30] As an illustration, take any distribution Q and any open neighborhood of the belief δ_Q that puts unit mass on Q. Then the Bayesian will put almost unit mass on that neighborhood, believing with near certainty that the process is driven by Q. For almost all samples s, this occurs infinitely often for every neighborhood of every Q. Diaconis and Freedman (1990) conclude that for a Bayesian in a higher dimensional setting, the prior swamps the data, rather than the other way around.

From a decision making stand point, these inconsistency results seem to undermine the *normative* case for forming beliefs via Bayesian updating. They suggest that building a compelling normative case for Savage-style

[28] The classic reference is Hewitt and Savage (1955) which generalizes de Finetti's result.

[29] Formally, ν is drawn at random according to a probability measure $\hat{\nu}$ on $\Delta(X)$ that is mutually absolutely continuous with respect to ν. See Feldman (1991) for references. A typical proof of this result relies on the fact that the sequence of posteriors forms a martingale under P.

[30] *i.e.,* the complement of a countable union of nowhere dense sets.

behavior that takes belief formation and learning seriously one should allow for non-Bayesian belief formation processes. In the model of this paper, given a statistical model $\mathcal{C}$, beliefs on the set of events $\mathcal{C}^\star$ are Bayesian, although they are not arrived at in a Bayesian fashion.

Consider, finally, the case of a *strict frequentist*, by which I mean a decision maker with belief given by:

$$\pi^t_{freq}(s) \equiv \delta_{\nu^t(s)}$$

where $\nu^t(s)$ is the empirical measure (1). When X is a complete separable metric space, the empirical measure converges to the true measure, so a frequentist will not suffer from the erratic belief formation inflicting the Bayesian.

Our definition of the set of distributions $\mu^t(s)$ consistent with empirical evidence is also based on empirical frequencies, and the empirical measures *always* belongs to $\mu^t(s)$. The difference is that $\mu^t(s)$ uses the empirical frequences only for the uniformly learnable family of events $\mathcal{C}$ that are part of the decision maker's statistical model. It is agnostic about the probabilities of events outside $\mathcal{C}$.[31] This is essential for our model for two reasons. First, a strict frequentist will put unit mass on the set of outcomes appearing in the sample, ruling out as impossible outcomes that did not appear in the sample. This is particularly striking when X is infinite (*e.g.,* [0,1]), in which case a frequentist cannot entertain the possibility that the true distribution is atomless. Even when X is finite but very large, a strict frequentist exposed to a realistic size sample will hold beliefs that would appear overly dogmatic and unreasonable. Second, the strict frequentist assigns probabilities regardless of concerns for confidence in these estimates. The notion of consistency with empirical evidence incorporates confidence via the requirement of uniform learning. Although the set $\mu^t(s)$ always includes the empirical measures, it also includes all other measures that cannot be incontrovertibly ruled out by evidence.

[31] Except in so far as they can be bounded by knowledge of probabilities of events in $\mathcal{C}$.

A Proofs

A.1 Strategic Product Measures

Defining sampling when X is complete separable metric space (*i.e.,* $X = X_c$) is standard: we take as sample space Ω the product $X \times X \times \cdots$ endowed with the Borel σ-algebra generated by the product topology.

The case of discrete X is, however, not standard and appeals to results likely to be unfamiliar to the reader. They are, however, appropriate generalizations of the usual constructions: Given $X = X_d$, we also define the sample space $\Omega = X \times X \times \cdots$. The product topology on this space is defined the usual way, where each coordinate is given the *discrete* topology. As in the countably additive case, we take as set of events the Borel σ-algebra generated by the product topology on Ω.

Suppose we are given a finitely additive probability measure λ on X. We are interested in defining the product measure λ^∞ on Ω. If λ happens to be countably additive, a standard result is that a countably additive λ^∞ can be uniquely defined. When λ is only finitely additive, the product measure λ^∞ need not be uniquely defined.

Dubins and Savage (1965) faced this problem in their book on stochastic processes and proposed the concept of *strategic products*. These are product measures that satisfy natural disintegration properties (trivially satisfied when λ is countably additive). In a classic paper, Purves and Sudderth (1976) showed that any finitely additive λ on a discrete X has a *unique* extension λ^∞ to the Borel σ-algebra on Ω.

I do not provide the details of the Dubins and Savage (1965) concept of strategic products or Purves and Sudderth (1976)'s constructions because they are not essential for what follows. For the purpose of the present paper, what the reader should bear in mind is: (1) the concept of strategic products is a natural restriction (for example, all product measures in the countably additive setting are strategic products); and (2) Purves and Sudderth's result permits extensions to the finitely additive setting of many of the major results in stochastic processes, including the Borel-Cantelli lemma, the strong law of large numbers, the Glivenko-Cantelli theorem and the Kolmogorov 0-1 law.

A.2 Proof of Theorem 3: Exact Learning

Lemma A.1 *Fix any uniformly learnable* $(\mathcal{C}, \epsilon)$*, we have:*

$$P^\infty\Big\{s : \lim_{t\to\infty} \sup_{A\in\mathcal{C}} \big|\nu^t(A,s) - P(A)\big| = 0\Big\} = 1.$$

Proof: From (12) we have that for every $P \in \mathcal{P}$ and $\alpha > 0$

$$\sum_{t=1}^{\infty} P^{\infty} \left\{ s : \sup_{A \in \mathcal{C}} |\nu^t(A,s) - P(A)| > \alpha \right\} < \infty.$$

As shown by Purves and Sudderth (1976), the Borel-Cantelli Lemma applies in the strategic setting. This implies:

$$P^{\infty} \left\{ s : \exists \bar{t} \; \forall t > \bar{t}, \; \sup_{A \in \mathcal{C}} |\nu^t(A,s) - P(A)| \leq \alpha \right\} = 1.$$

Take a sequence $\alpha_n \downarrow 0$, and note that each of the events:

$$\left\{ s : \exists \bar{t} \; \forall t > \bar{t}, \; \sup_{A \in \mathcal{C}} |\nu^t(A,s) - P(A)| \leq \alpha_n \right\}$$

is a tail event. Purves and Sudderth (1983) show that P^{∞} is countably additive on tail events, so:

$$P^{\infty} \bigcap_{n} \left\{ s : \exists \bar{t} \; \forall t > \bar{t}, \; \sup_{A \in \mathcal{C}} |\nu^t(A,s) - P(A)| \leq \alpha_n \right\} = 1,$$

hence:

$$P^{\infty} \Big\{ s : \lim_{t \to \infty} \sup_{A \in \mathcal{C}} \big|\nu^t(A,s) - P(A)\big| = 0 \Big\} = 1.$$

■

Lemma A.2 *For any uniformly learnable $\mathcal{C}$ and an $\epsilon > 0$, we have, P^{∞}-a.s.,*

$$\begin{aligned}
\mathcal{M}(\mathcal{C}, \epsilon, s) &\equiv \bigcup_{i=1}^{\infty} \bigcap_{t \geq i} \left\{ p : \sup_{A \in \mathcal{C}} \big|p(A) - \nu^t(A,s)\big| \leq \epsilon \right\} \\
&= \left\{ p : \; \exists \bar{t}, \; \forall t > \bar{t} \; \sup_{A \in \mathcal{C}} \big|p(A) - \nu^t(A,s)\big| \leq \epsilon \right\} \\
&= \Big\{ p : \sup_{A \in \mathcal{C}} |p(A) - P(A)| \leq \epsilon \Big\}.
\end{aligned}$$

Proof: Lemma A.1 states that the set of sample paths:

$$\Big\{ s : \lim_{t \to \infty} \sup_{A \in \mathcal{C}} \big|\nu^t(A,s) - P(A)\big| = 0 \Big\} \tag{14}$$

has P^∞-probability 1. Being in the event in (14) above implies that given any $\epsilon' > 0$ we have $\sup_{A' \in \mathcal{C}} \ |\nu^t(A', s) - P(A')| < \epsilon'$ for all large t. For the remainder, fix s to be any sample in this set.

If $p \in \mathcal{M}(\mathcal{C}, \epsilon, s)$ then $\sup_{A \in \mathcal{C}} \ |p(A) - \nu^t(A, s)| < \epsilon$ for all large enough t. Thus, for all large t, we have:

$$\begin{aligned}
\sup_{A \in \mathcal{C}} |p(A) - P(A)| &\leq \sup_{A \in \mathcal{C}} \Big[\big|p(A) - \nu^t(A,s)\big| + \big|\nu^t(A,s) - P(A)\big| \Big] \\
&\leq \sup_{A \in \mathcal{C}} \big|p(A) - \nu^t(A,s)\big| + \sup_{A' \in \mathcal{C}} \big|\nu^t(A',s) - P(A')\big| \\
&\leq \epsilon + \epsilon'.
\end{aligned}$$

Since ϵ' was arbitrary, we conclude

$$\sup_{A \in \mathcal{C}} |p(A) - P(A)| \leq \epsilon,$$

so $p \in \big\{p' : \sup_{A \in \mathcal{C}} |p'(A) - P(A)| \leq \epsilon \big\}$.

Conversely, if $p \in \big\{p : \sup_{A \in \mathcal{C}} |p(A) - P(A)| \leq \epsilon \big\}$ then, to show that $\sup_{A \in \mathcal{C}} \ |p(A) - \nu^t(A, s)| < \epsilon$ for all large t, we proceed similarly to the above argument:

$$\begin{aligned}
\sup_{A \in \mathcal{C}} \big|p(A) - \nu^t(A,s)\big| &\leq \sup_{A \in \mathcal{C}} \Big[\big|p(A) - P(A)\big| + \big|P(A) - \nu^t(A,s)\big| \Big] \\
&\leq \sup_{A \in \mathcal{C}} \big|p(A) - P(A)\big| + \sup_{A' \in \mathcal{C}} \big|\nu^t(A',s) - P(A')\big| \\
&\leq \epsilon + \epsilon',
\end{aligned}$$

and the conclusion follows from the fact that ϵ' was arbitrary. ■

Theorem 3 *Fix any learning strategy $\{(\mathcal{C}_t, \epsilon_t)\}_{t=1}^\infty$ and write $\mathcal{C} \equiv \cup_{t=1}^\infty \mathcal{C}_t$. Then for any $P \in \mathcal{P}$*

$$\mu_\sigma(s) = \Big\{ p : p(A) = P(A), \forall A \in \mathcal{C} \Big\}, \qquad P^\infty - a.s.$$

In particular, $\mu_\sigma(s)$ is a convex set of probability measures, almost surely.

Proof: We first note that:

$$\begin{aligned}
\mathcal{M}(s) &= \bigcap_{\substack{\mathcal{C}_{t'} \\ t'=1,2\ldots}} \ \bigcap_{\epsilon > 0} \ \bigcup_{i=1}^{\infty} \bigcap_{t \geq i} \left\{ p : \sup_{A \in \mathcal{C}_{t'}} \ \big|p(A) - \nu^t(A,s)\big| \leq \epsilon \right\} \\
&= \bigcap_{\substack{\mathcal{C}_{t'} \\ t'=1,2\ldots}} \ \bigcap_{\epsilon > 0} \mathcal{M}(\mathcal{C}_{t'}, \epsilon, s).
\end{aligned}$$

Note also that any event of the form:

$$\left\{ s : \mathcal{M}(\mathcal{C}', \epsilon, s) = \left\{ p : \sup_{A \in \mathcal{C}'} |p(A) - P(A)| \leq \epsilon \right\} \right\}$$

is a tail event and, by Lemma A.2, must have P^∞-probability 1. By Purves and Sudderth (1983)'s result that P^∞ is countably additive on tail events, we have

$$P^\infty \left\{ \bigcap_{\substack{\mathcal{C}_{t'} \\ t'=1,2\ldots}} \bigcap_{\epsilon > 0} \left\{ s : \mathcal{M}(\mathcal{C}_{t'}, \epsilon, s) = \left\{ p : \sup_{A \in \mathcal{C}_{t'}} |p(A) - P(A)| \leq \epsilon \right\} \right\} \right\} = 1.$$[32]

This is equivalent to the desired result, namely:

$$P^\infty \left\{ s : \mathcal{M}(s) = \left\{ p : \sup_{A \in \mathcal{C}} |p(A) - P(A)| \leq \epsilon \right\} \right\} = 1$$

(recall that $\mathcal{C} \equiv \cup_{t=1}^\infty \mathcal{C}_t$). ■

A.3 Proof of Theorem 4: Complete Learning in Continuous Outcome Spaces

This is essentially a consequence of two facts: (1) all complete separable metric spaces are "equivalent" to a Borel subset of $[0, 1]$; and (2) on $[0, 1]$ knowing the probabilities of half intervals is sufficient to determine the probability of all Borel sets. The technical details are as follows:

By Royden (1968, Theorem 8, p. 326) $(X = X_c, \mathcal{B})$ is Borel equivalent to a Borel subset of [0,1].[33] That is, there is a Borel subset $B \subset [0, 1]$ and a measurable bijection $\phi : X \to B$ such that ϕ^{-1} is also measurable. For each $r \in [0, 1]$ define $A_r = \phi^{-1}\big([0, r]\big)$ and let $\mathcal{C} = \{A_r : r \in [0, 1]\}$. That is, the collection $\mathcal{C}$ mimics the structure of half-intervals on $[0, 1]$. Note, however, that these sets need not preserve much of the geometric properties of half interval, such as connectedness. What they do preserve, however, is the fact that they are nested: $A_r \subsetneq A_{r'}$ whenever $r < r'$. It is easy to verify, then, that the family of sets $\mathcal{C}$ has VC-dimension of 1.[34] From this it follows that

[32] These are countable intersections, since we can take a sequence $\epsilon_n \downarrow 0$ if necessary.

[33] The interval [0,1] will always be understood as being endowed with the Borel σ-algebra.

[34] See Problem 13.15 of Devroye, Gyorfi, and Lugosi (1996, p. 231) for this obvious fact and its (slightly less obvious) converse.

for every (countably additive) probability distribution P:

$$P^\infty \left\{ s : \sup_{A \in \mathcal{C}} \ |\lim_{t\to\infty} \nu^t(A,s) - P(A)| = 0 \right\} = 1.$$

Fix any sample path s such that $\sup_{A\in\mathcal{C}} \ |\lim_{t\to\infty} \nu^t(A,s) - P(A)| = 0$. If $p \in \lim_{t\to\infty} \mu^t_\sigma(s)$, then it follows from the definition of μ^t_σ that $p(A) = P(A)$ for every $A \in \mathcal{C}$.

To show that p and P are identical, we "transfer" p and P to the interval [0,1]. For every Borel set $A \subset [0,1]$, define $\tilde{p}(A) \equiv p\big(\phi^{-1}(A)\big)$ and $\tilde{P}(A) \equiv P\big(\phi^{-1}(A)\big)$. Then by Royden (1968, Proposition 1, p. 318) $\tilde{P}$ and $\tilde{p}$ are probability measures on [0,1] that agree on the values they assign to all half intervals, and thus must have the same distribution functions. From this, it follows that $\tilde{p} = \tilde{P}$, hence $p = P$ since ϕ is a Borel equivalence.

A.4 Proof of Theorems 5: Failure of Complete Learning in Discrete Outcome Spaces

We first prove the following weaker claim:

Proposition 10 *Beliefs in the discrete outcome space* (X_d, Σ) *are not asymptotically determinate by any simple learning strategy.*

We will show that there are two probability measures λ and γ that agree on $\mathcal{C}$ but disagree on some (in fact, many) events outside $\mathcal{C}^\star$. The proof proceeds in three steps: (1) Construct a "nice" finitely additive probability measure λ on $\mathcal{C}$; (2) Construct a class of admissible perturbations s of the density of λ with the property that they leave λ unaffected on $\mathcal{C}$; (3) Show that any admissible perturbation to λ yields a new finitely additive probability measure γ that differs from λ in the value it assigns to many sets.

A.4.1 Constructing λ

Let $\{X_N\}_{N=1}^\infty$ be an increasing sequence of finite subsets of X such that

$$\frac{\eta_N - \eta_{N-1}}{\eta_N} > 1 - \frac{1}{N}$$

where $\eta_N \equiv \#X_N$ denotes the cardinality of X_N. Note that this implies that $\eta_N > N\eta_{N-1}$. To avoid excessive repetition, in the remainder of the proof it will be understood that $N - 1 \geq 1$ whenever necessary.

Define the probability measure λ_N on 2^X by

$$\lambda_N(A) = \frac{\#(A \cap X_N)}{\#X_N}.$$

That is, $\lambda_N(A)$ is the frequency of the set A in X_N.

Let $\mathcal{U}$ be a free ultrafilter on the integers and for any sequence of real numbers x_N define the expression

$$\mathcal{U}\text{–}\lim_{N\to\infty} x_N = x$$

to mean that the set $\{N : |x_N - x| < \epsilon\}$ belongs to $\mathcal{U}$ for any for every $\epsilon > 0$. Then for any event A, define:

$$\lambda(A) \equiv \mathcal{U}\text{–}\lim_{N\to\infty} \lambda_N(A),$$

Intuitively, λ is a "uniform" distribution on the integers. It is immediate that λ is atomless (*i.e.,* assigns zero mass to each point) and purely finitely additive.

> **Comments:** For readers not familiar with these concepts, the idea is to define the probability of the event A, $\lambda(A)$, as limit of the finite probabilities $\lambda_N(A)$. If $\lambda_N(A), N = 1, 2\ldots$ converges, then the statement that $\lambda(A) \equiv \lim_{N\to\infty} \lambda_N(A)$ is equivalent to saying that the set of integers $\{N : |\lambda_N(A) - \lambda(A)| < \epsilon\}$ is cofinite (*i.e.,* complement of a finite set) for every $\epsilon > 0$. That is, "$\lambda_N(A)$ converges to $\lambda(A)$" *means* that the set of N's on which $\lambda_N(A)$ is within ϵ of $\lambda(A)$ is small for all $\epsilon > 0$, where 'small' here means finite.
>
> The notion of ultrafilter generalizes this intuition by identifying a collection of large subsets of integers $\mathcal{U}$. That $\mathcal{U}$ is free means that it contains all cofinite sets, and that it is 'ultra' means that each set of integers is either in $\mathcal{U}$ or its complement is. This immediately implies that the operation $\mathcal{U}$–lim generalizes the usual limit, and that any sequence must have a generalized $\mathcal{U}$–lim. Ultrafilters is a standard mathematical tool that generalizes limits by selecting convergent subsequences in a consistent manner.[35]

A.4.2 Perturbations

A *perturbation* is any function $s : X \to \{1 - \epsilon, 1 + \epsilon\}$, with $\epsilon \in [0, 1]$. Let S denote the set of all perturbations. Endow S with the σ-algebra Σ generated by the product topology, *i.e.,* the one generated by all sets of the form $\{s : s(x) = 1 + \epsilon\}$ for some $x \in X$.

[35] Bhaskara Rao and Bhaskara Rao (1983) provide formal definitions. Wikipedia has a nice article on the subject.

Let P be the *countably* additive product measure on the measure space (S, Σ) assigning probability 0.5 to each of the events $\{s : s(x) = 1 + \epsilon\}$, $x \in X$. That is, P is constructed by taking equal probability i.i.d. randomizations for $s(x) \in \{1-\epsilon, 1+\epsilon\}$. Note that (S, Σ, P) is a standard countably additive probability space constructed using standard methods (*e.g.,* Kolmogorov extension). The only finite additivity is in the measure λ on the index set X.

Fix an arbitrary N. Two events $A, B \subset X$ are X_N-*equivalent* (or simply equivalent, when N is understood) if $A \cap X_N = B \cap X_N$. We use A_N to denote the equivalence class of A and define $\mathcal{C}_N \equiv \{A_N : A \in \mathcal{C}\}$. That is, $\mathcal{C}_N$ is the appropriate 'projection' of $\mathcal{C}$ on X_N.

The key observation is that, $\mathcal{C}$ having finite VC-dimension v on all of X means that no subset of $v + 1$ points in X can be shattered by $\mathcal{C}$. Then, a fortiori, no subset of $v + 1$ points in X_N can be shattered by $\mathcal{C}$, so the VC-dimension of the family of events $\mathcal{C}_N$ in X_N is at most v.

A key combinatorial result, due to Sauer (1972) (see also Devroye, Gyorfi, and Lugosi (1996, Theorem 13.3, p. 218)) states that, given an outcome space of η_N points, any family of events of finite VC-dimension v cannot contain more than $2(\eta_N)^v$ events.

> **Comments:** To appreciate this bound, recall that X_N contains 2^N events in all, so an implication of Sauer's Lemma is that being of finite VC-dimension severely restricts how rich a family of events can be. For example, with 50 states ($N = 50$) if $\mathcal{C}$ has a VC-dimension of 5, say, then the ratio of the number of events in $\mathcal{C}$ to the powerset is no more than 5.5×10^{-7}.
>
> This cardinality argument, while suggestive, does little for us in the limit: when the size of X_N goes to infinity, even fixing v, both the cardinality of $\mathcal{C}$ and the power set go to infinity. In fact, it is possible to construct a family of events $\mathcal{C}$ in X of VC-dimension 1, yet $\mathcal{C}$ has uncountable cardinality (see Devroye, Gyorfi, and Lugosi (1996, Problem 13.14, p. 231)). This necessitates a more indirect approach than just "counting sets."

Let $\mathcal{C}'_N \subset \mathcal{C}_N$ denote the family of events $\{A_N : A \in \mathcal{C},\ \lambda(A) \geq \frac{1}{4}\}$. Since $\mathcal{C}$ is closed under complements, so is $\mathcal{C}_N$, hence for each $A \in \mathcal{C}$ at least one of the sets $\{A_N, A^c_N\}$ belongs to $\mathcal{C}'_N$.

Since the perturbations are chosen independently, we may apply the Chernov bound to conclude that, for any subset of X_N containing at least

$N/4$ points:

$$P\left\{s \in S : \frac{1}{\eta_N}\left|\sum_{x \in A_N - X_{N-1}} s(x) - \#(A_N - X_{N-1})\right| > \alpha\right\} \leq 2\, e^{-2\,\#(A_N - X_{N-1})\,\alpha}$$
$$\leq 2\, e^{-2\frac{\eta_N - \eta_{N-1}}{4}\alpha}$$
$$\leq 2\, e^{-2\frac{\eta_N}{8}\alpha}.$$

Since there are no more than $2(\eta_N)^v$ events in $\mathcal{C}'_N$, we obtain:

$$P(Z_{\alpha N}) \leq 4\,(\eta_N)^v\ e^{-\frac{\eta_N}{4}\alpha}$$

where

$$Z_{\alpha N} \equiv \left\{s \in S : \max_{A_N \in \mathcal{C}'_N} \frac{1}{\eta_N}\left|\sum_{x \in A_N - X_{N-1}} s(x) - \#(A_N - X_{N-1})\right| > \alpha\right\}.$$

By construction, $Z_{\alpha N},\ N = 1, 2 \ldots$ is a sequence of independent events, for any fixed $\alpha > 0$. (This is the reason why we use the sets $A_N - X_{N-1}$, rather than simply A_N. Had we used the latter, the $Z_{\alpha N}$'s will obviously be correlated.) Summing up, we obtain:

$$\sum_{N=1}^{\infty} P(Z_{\alpha N}) \leq 4 \sum_{N=1}^{\infty} (\eta_N)^v\ e^{-\frac{\eta_N}{4}\alpha} < \infty.$$

By the Borel-Cantelli Lemma (the usual version, since P is countably additive), the set Z_α of perturbations that belong to infinitely many of the $Z_{\alpha N}$'s has P-measure 0. This, in turn, implies that the event

$$Q_0 \equiv \bigcap_{k=1}^{\infty} (Z_{1/k})^c \tag{15}$$

also has P-measure 1. In addition, using the law of large numbers, $P(Q_1) = 1$ where $Q_1 = \left\{s \in S : \lambda\{x : s(x) = 1 + \epsilon\} = \frac{1}{2}\right\}$ (this follows from Al-Najjar (2007)–the argument in this case is in fact completely straightforward).

From the above it follows that $P(Q_0 \cap Q_1) = 1$ and so $Q_0 \cap Q_1$ is, in particular, non-empty.

> **Comments:** The above argument is delicate and is the heart of the proof. Think of an indicator function χ_A of an event A with $0 < \lambda(A) < 1$ as

its 'density' function with respect to the distribution λ. The idea is to perturb that density by tweaking it up and down by ϵ. Call a perturbation s *neutral with respect to* A if $\lambda(A) \equiv \int \chi_A \, d\lambda = \int s \cdot \chi_A \, d\lambda$. Any such perturbation s defines a new probability measure $\gamma(A) \equiv \int s \, \chi_A \, d\lambda$ that leaves the probability of A intact yet differs from A at least on the event $B \equiv \{x : s(x) = 1 - \epsilon\}$. I show the existence of perturbations s that accomplish this not just with respect to a single event A, but all events in $\mathcal{C}$ simultaneously.

The strategy is to draw, for each x, a value in $\{1 + \epsilon, 1 - \epsilon\}$ with equal probability and independently across the x's. It is straightforward to check that, given a single fixed event A, any draw s will be, $P-$almost surely, neutral with respect to A. Since the intersection of countably many P-measure 1 sets has P-measure 1, this conclusion can be extended to any countable family of events $\mathcal{A} = \{A_1, A_2, \ldots\}$. The trouble is in dealing with an uncountable family $\mathcal{C}$, a case that is essential for the theory since many standard classes like half intervals, half spaces, Borel sets, ... etc are uncountable. A less direct and more subtle argument is needed.

Here, the assumption that $\mathcal{C}$ is uniformly learnable (specifically, has finite VC-dimension) plays a critical role via the fundamental combinatorial result known as Sauer's lemma. It is well-known from the theory of large deviations that convergence in the (weak) law of large numbers is exponential in sample size. This implies that one can estimate the probabilities of an increasing family of events, provided the cardinality of this family does not increase too quickly. Sauer's lemma roughly states that a family with finite VC-dimension must have a cardinality that is polynomial in the size of the outcome space. The difficulty, of course, is that neither large deviations nor Sauer's lemma have much meaning in the limit, when t is infinite. In the proof I first project the (possibly uncountable) family $\mathcal{C}$ on the finite sets X_N, identify the (approximately) good perturbations, and bound their probabilities. Only then can I pass to the limit to obtain the sets Q_0 and Q_1.

A.4.3 Perturbed Measures

The desired candidate perturbation is any element of $Q_0 \cap Q_1$. Fixing one such s, define the set function:

$$\gamma(A) \equiv \int_A s(x) \, d\lambda, \qquad A \subset X.$$

We first verify that γ is a finitely additive probability measure. From the additivity of the integral, it immediately follows that γ is an additive set function. Positivity of γ follows as long as $\epsilon \in [0, 1]$. Finally, the fact that

$s \in Q_1$ implies

$$\begin{aligned}\gamma(X) &= \int_X s(x)\,d\lambda \\ &= (1+\epsilon)\,\lambda\{x : s(x) = 1+\epsilon\} + (1-\epsilon)\,\lambda\{x : s(x) = 1-\epsilon\} \\ &= \frac{1}{2}\,(1+\epsilon) + \frac{1}{2}\,(1-\epsilon) = 1.\end{aligned}$$

Next we show that λ and γ coincide on $\mathcal{C}$ (hence necessarily on $\mathcal{C}^\star$). Take any set $A \in \mathcal{C}$. If $\lambda(A) = 0$, then

$$\gamma(A) \equiv \int_A s(x)\,d\lambda \le (1+\epsilon)\,\lambda(A) = 0,$$

so γ agrees with λ on A. By the additivity of the integral, the same conclusion holds if $\lambda(A) = 1$.

Having disposed of this case, assume that $0 < \lambda(A) < 1$. Without loss of generality, let $\lambda(A) \ge 0.5$ (if not, take its complement and use additivity again). From the ultrafilter construction, there is a subsequence N_k, $k = 1, 2 \ldots$ such that $\lambda_{N_k}(A) \to \lambda(A)$, hence for each k we have $A_{N_k} \in \mathcal{C}'_{N_k}$.

Now

$$\begin{aligned}|\gamma(A) - \lambda(A)| &= \left| \mathcal{U}\text{–}\lim_{N\to\infty} \int_A s(x)\,d\lambda_N - \mathcal{U}\text{–}\lim_{N\to\infty} \lambda_N(A) \right| \\ &= \left| \lim_{k\to\infty} \int_A s(x)\,d\lambda_{N_k} - \lim_{k\to\infty} \lambda_{N_k}(A) \right| \\ &= \lim_{k\to\infty} \frac{1}{\eta_{N_k}} \left| \sum_{x\in A\cap X_{N_k}} s(x) - \#(A \cap X_{N_k}) \right| \\ &= \lim_{k\to\infty} \frac{1}{\eta_{N_k}} \left| \sum_{x\in A_{N_k}} s(x) - \#A_{N_k} \right|.\end{aligned}$$

Using the triangle inequality, we have:

$$\begin{aligned}\left| \sum_{x\in A_N} s(x) - \#A_N \right| &= \Bigg| \sum_{x\in A_N - X_{N-1}} s(x) + \sum_{x\in A_N\cap X_{N-1}} s(x) \\ &\qquad\qquad - \#(A_N - X_{N-1}) - \#(A_N \cap X_{N-1}) \Bigg| \\ &\le \left| \sum_{x\in A_N - X_{N-1}} s(x) - \#(A_N - X_{N-1}) \right| + \epsilon\eta_{N-1},\end{aligned}$$

from which we conclude that:

$$|\gamma(A)-\lambda(A)| \leq \lim_{k\to\infty} \frac{1}{\eta_{N_k}} \left| \sum_{x\in A_{N_k}-X_{N_k-1}} s(x) - \#(A_{N_k} - X_{N_k}) \right| + \epsilon \lim_{k\to\infty} \frac{\eta_{N_k-1}}{\eta_{N_k}}.$$

Fixing $\alpha > 0$ and using the fact that $s \in Q_0$ we have, for all large enough k,

$$\frac{1}{\eta_{N_k}} \left| \sum_{x\in A_{N_k}-X_{N_k-1}} s(x) - \#(A_N - X_{N-1}) \right| < \alpha.$$

The above, and the assumption that $\lim_{k\to\infty} \frac{\eta_{N_k-1}}{\eta_{N_k}} = 0$ imply that

$$|\gamma(A) - \lambda(A)| \leq \alpha.$$

Since α is arbitrary, it follows that $\gamma(A) = \lambda(A)$.

All that remains to prove is that the perturbed measure γ must differ from λ on some (in fact, many) events outside $\mathcal{C}^\star$. Take the event $B \equiv \{x : s(x) = 1 - \epsilon\}$. Since $s \in Q_1$, we have $\lambda(B) = 0.5$, yet

$$\gamma(B) \equiv \int_B s(x)\, d\lambda = (1-\epsilon)\, \lambda(B) \neq \lambda(B), \tag{16}$$

so $B \notin \mathcal{C}^\star$ since, by the earlier part of the argument, λ and γ coincide on $\mathcal{C}$. This completes the proof of Proposition 10. ■

Comments: From Theorem 4, we know that this proof must break down somewhere if X were a complete separable metric space with countably additive probabilities. A natural question is: at what stage was finite additivity needed and the implications of Theorem 4 avoided? The construction of the perturbation s by i.i.d. sampling is not possible in a complete, separable X with countably additive probabilities. The reason is that a typical sample path s is non-measurable and the perturbed measure $\gamma(A) = \int s \cdot \chi_A \, d\lambda$ cannot be meaningfully defined. Of course, I do not claim that finding s via random sampling is the only feasible procedure to construct perturbations, but only point out that this particular procedure breaks down in standard spaces–as it should, given Theorem 4.

A.4.4 Proof of Corollary 6

From (16) and the fact $\lambda(B) = 0.5$ we can write:

$$\gamma(B) = \lambda(B) - 0.5\,\epsilon$$

so

$$|\gamma(B) - \lambda(B)| = |0.5\,\epsilon|.$$

Varying ϵ within the interval (0,1] yields the desired conclusion. The fact that there are uncountably many such B's follows from the fact that the distribution on admissible perturbations is atomless, and hence its support must be uncountable.

A.4.5 Proof of Theorem 5

We now assume that $\mathcal{C} = \cup_{t=1}^{\infty}\mathcal{C}_t$ with $\mathcal{C}_t$ having finite VC-dimension. Index the events defined in (15) by t, writing it as Q_0^t to make explicit its dependence on $\mathcal{C}_t$. Consider now the event

$$\bigcap_{t=1}^{\infty} Q_0^t \ \cap \ Q_1$$

and note that it must have P^{∞}-probability 1. Let s be any element of this set. It is clear that the remainder of the argument in Section A.4.3 goes through unaltered.

A.5 Miscellaneous Proofs

Proof of Theorem 1: Let $\mathbb{C}$ denote the set of all classes of events containing $\mathcal{C}$ that are ϵ-uniformly learnable by data of size t. Then $\mathbb{C}$ is partially ordered by set inclusion. By Hausdorff maximal principle, in $\mathbb{C}$ there is a totally ordered chain containing $\mathcal{C}$. That is, there is a maximal set $\mathbb{C}^{\star} \subset \mathbb{C}$ such that $\mathcal{C} \in \mathbb{C}^{\star}$ and $\mathbb{C}^{\star}$ is linearly ordered by set inclusion. Define $\bar{\mathcal{C}} \equiv \cup_{\hat{\mathcal{C}} \in \mathbb{C}^{\star}} \hat{\mathcal{C}}$.

First we note that $\bar{\mathcal{C}}$ is ϵ-uniformly learnable by data of size t. For if this were not the case, then there is sample $x_1, \ldots, x_m$ that can be shattered by $\bar{\mathcal{C}}$ but not by any class in $\mathbb{C}^{\star}$. But shattering a finite sample requires only finitely many events. By the definition of $\bar{\mathcal{C}}$ there must be $\hat{\mathcal{C}} \in \mathbb{C}^{\star}$ that can also shatter the same sample, contradicting the assumption that $\hat{\mathcal{C}}$ is ϵ-uniformly learnable by data of size t. Since $\mathbb{C}^{\star}$ is maximal, $\bar{\mathcal{C}} \in \mathbb{C}^{\star}$. ■

Proof of Theorem 2: The VC-dimension of 2^{X_f} is n. The first claim follows directly from (12). For the second part, a lower bound on the amount of data needed was shown by Ehrenfeucht, Haussler, Kearns, and Valiant

(1989)[36] to be:

$$t \geq \frac{V_{\mathcal{C}} - 1}{32\epsilon}. \tag{17}$$

Applying this bound with $V_{\mathcal{C}} = n$, holding t and ϵ fixed, yields the result. ■

Example illustrating part 2 of Theorem 9: Take $X = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{C}$ to consist of X, the empty set, and all events of the form $\{x, x+1, x+2\}$, where $x \in X$ and addition is modulo 6. It is easy to verify that $\mathcal{C}$ is closed under complements and disjoint unions, so $\mathcal{C}$ itself is a λ-system. On the other hand, $\mathcal{C} \neq \mathcal{C}^{\star}$: if μ is any probability measure on 2^X, then

$$\mu(\{1, 2, 3\}) + \mu(\{3, 4, 5\}) - \mu(\{2, 3, 4\}) = \mu(\{1, 3, 5\}).$$

So the probability of the event $\{1, 3, 5\} \notin \mathcal{C}$ can be unambiguously determined, and so this event belongs to $\mathcal{C}^{\star}$.

On the other hand, $\mathcal{C}^{\star} \neq 2^X$: Fix any μ that assigns positive probability to each state. Consider vectors of the form $\bar{\alpha} = (\alpha, -0.5\alpha, -0.5\alpha, \alpha, -0.5\alpha, -0.5\alpha)$. Then for any appropriately chosen value for $\alpha > 0$, $\mu + \bar{\alpha}$ is a probability measure that assigns identical values as μ to events in $\mathcal{C}$ even though μ and $\mu + \bar{\alpha}$ differ at each state. This shows, in particular, that $\{1\} \notin \mathcal{C}^{\star}$. ■

[36] See also Devroye, Gyorfi, and Lugosi (1996, Section 14.5)).

References

Al-Najjar, N. I. (2007): “Finitely Additive Representation of L^p Spaces,” *Journal of Mathematical Analysis and Applications*, 330, 891–899.

Al-Najjar, N. I., L. Anderlini, and L. Felli (2006): “Undescribable Events,” *Review of Economic Studies*, 73, 849–68.

Aumann, R. J. (1987): “Correlated equilibrium as an expression of Bayesian rationality,” *Econometrica*, 55(1), 1–18.

Bewley, T. (1986): “Knightian Decision Theory: Part I,” Cowles Foundation Discussion Paper no. 807.

——— (1988): “Knightian Decision Theory and Econometric Inference,” Cowles Foundation Discussion Paper no. 868.

Bhaskara Rao, K. P. S., and M. Bhaskara Rao (1983): *Theory of charges.* Academic Press Inc., New York.

Billingsley, P. (1995): *Probability and measure*, Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edn., A Wiley-Interscience Publication.

de Finetti, B. (1974): *Theory of Probability, Vol. 1-2.* Wiley, New York.

Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition.* Springer, Berlin.

Diaconis, P., and D. Freedman (1986): “On the consistency of Bayes estimates,” *Ann. Statist.*, 14(1), 1–26.

Diaconis, P., and D. Freedman (1990): “On the uniform consistency of Bayes estimates for multinomial probabilities,” *Ann. Statist.*, 18(3), 1317–1327.

Dubins, L. E., and L. J. Savage (1965): *How to gamble if you must. Inequalities for stochastic processes.* McGraw-Hill Book Co., New York.

Efron, B. (1986): “Why Isn’t Everyone a Bayesian?,” *The American Statistician*, 40(1), 1–5.

——— (2005): “Bayesians, Frequentists, and Scientists.,” *Journal of the American Statistical Association*, 100(469), 1–6.

Ehrenfeucht, A., D. Haussler, M. Kearns, and L. Valiant (1989): “A general lower bound on the number of examples needed for learning,” *Inform. and Comput.*, 82(3), 247–261.

Epstein, L., and J. Zhang (2001): “Subjective Probabilities on Subjectively Unambiguous Events,” *Econometrica*, 69(2), 265–306.

Feldman, M. (1991): "On the generic nonconvergence of Bayesian actions and beliefs," *Econom. Theory*, 1(4), 301–321.

Freedman, D. A. (1965): "On the asymptotic behavior of Bayes estimates in the discrete case. II," *Ann. Math. Statist.*, 36, 454–456.

Gajdos, T., T. Hayashi, J.-M. Tallon, and J.-C. Vergnaud (2006): "Attitude toward Imprecise Information," *Journal of Economic Theory (Forthcoming)*.

Gilboa, I., and D. Schmeidler (1989): "Maxmin expected utility with nonunique prior," *J. Math. Econom.*, 18(2), 141–153.

Hansen, L., and T. Sargent (2001): "Acknowledging Misspecification in Macroeconomic Theory," *Monetary and Economic Studies (Special Edition)*, 19, 213–237.

Harman, G., and S. Kulkarni (2007): *Reliable Reasoning: Induction and Statistical Learning Theory.* MIT Press.

Hewitt, E., and L. J. Savage (1955): "Symmetric measures on Cartesian products," *Trans. Amer. Math. Soc.*, 80, 470–501.

Horn, A., and A. Tarski (1948): "Measures in Boolean algebras," *Trans. Amer. Math. Soc.*, 64, 467–497.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005): "A smooth model of decision making under ambiguity," *Econometrica*, 73(6), 1849–1892.

Kreps, D. (1998): "Anticipated Utility and Dynamic Choice," in *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures*, ed. by D. Jacobs, E. Kalai, and M. Kamien. Cambridge University Press.

Lehrer, E. (2005): "Partially-specified Probabilities: Decisions and Games," Tel-Aviv University.

Lipman, B. (1995): "Information Processing and Bounded Rationality: A Survey," *The Canadian Journal of Economics*, 28(1), 42–67.

Maccheroni, F., M. Marinacci, and A. Rustichini (2006): "Ambiguity Aversion, Malevolent Nature, and the Variational Representation of Preferences," *Econometrica*, 74, 1447–98.

Machina, M. (1989): "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty," *Journal of Economic Literature*, 27(4), 1622–1668.

MANSKI, C. F. (2004): "Statistical treatment rules for heterogeneous populations," *Econometrica*, 72(4), 1221–1246.

MORRIS, S. (1995): "The Common Prior Assumption in Economic Theory," *Economics and Philosophy*, 11, 227–253.

POLLARD, D. (1984): *Convergence of Stochastic Processes.* Springer-Verlag.

PURVES, R. A., AND W. D. SUDDERTH (1976): "Some finitely additive probability," *Ann. Probability*, 4(2), 259–276.

——— (1983): "Finitely additive zero-one laws," *Sankhyā Ser. A*, 45(1), 32–37.

ROYDEN, H. L. (1968): *Real Analysis, 2ed Edition.* MacMillan Publishing Co.,Inc., New York.

SAUER, N. (1972): "On the density of families of sets," *Journal of Combinatorial Theory*, 13, 145–147.

SAVAGE, L. J. (1951): "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 46(253), 55–67.

——— (1954): *The foundations of statistics.* John Wiley & Sons Inc., New York.

——— (1967): "Difficulties in the theory of personal probability," *Philosophy of Science*, 34, 305–10.

——— (1972): *The foundations of statistics.* Dover Publications Inc., New York, revised edn.

SCHMEIDLER, D. (1989): "Subjective Probability and Expected Utility Without Additivity," *Econometrica*, 57(3), 571–587.

TALAGRAND, M. (1987): "The Glivenko-Cantelli Problem," *Annals of Probability*, 15, 837–70.

VAPNIK, V. N. (1998): *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, A Wiley-Interscience Publication.

VAPNIK, V. N., AND A. Y. CHERVONENKIS (1971): "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities," *Theory of Probability and its Applications*, XVI, 264–80.

ZHANG (1999): "Qualtative Probabilities on λ-systems," *Mathematical Social Sciences*, 38, 11–20.