Econometrica, Vol. 77, No. 5 (September, 2009), 1339–1369

DECISION MAKERS AS STATISTICIANS: DIVERSITY, AMBIGUITY, AND LEARNING

BY NABIL I. AL-NAJJAR[1]

I study individuals who use frequentist models to draw uniform inferences from independent and identically distributed data. The main contribution of this paper is to show that distinct models may be consistent with empirical evidence, even in the limit when data increases without bound. Decision makers may then hold different beliefs and interpret their environment differently even though they know each other's model and base their inferences on the same evidence. The behavior modeled here is that of rational individuals confronting an environment in which learning is hard, rather than individuals beset by cognitive limitations or behavioral biases.

KEYWORDS: Learning, statistical complexity, belief formation.

The crowning intellectual accomplishment of the brain is the real world—Miller (1981).

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. Most equilibrium concepts in economics—be it Nash, sequential, or rational expectations equilibrium—require beliefs to coincide with the true data generating process. As a result, disagreements and differences in beliefs are reduced to differences in information.[2] 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, namely inference from independent and identically distributed (i.i.d.) data. These individuals are modeled as classical, frequentist statisticians concerned with drawing uniform inferences that do not depend on prior beliefs. The main contribution of the paper is to show that distinct models can be consistent with the same empirical evidence, even asymptotically when data increases without bound. Individuals may then hold different beliefs and interpret their environment differently

[1]I am grateful to a co-editor and four referees for extensive and thoughtful feedback that substantially improved the paper. I also thank Drew Fudenberg, Ehud Kalai, Peter Klibanoff, Nenad Kos, Charles Manski, Pablo Schenone, and Jonathan Weinstein for their comments. I owe a special debt to Lance Fortnow and Mallesh Pai without whom this project would not have even started.

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

 DOI: 10.3982/ECTA7501

even though they know each other's 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 approach advocated in this paper is to model rational individuals as seeking uniform, distribution-free inferences in environments where learning is hard. No appeal to computational complexity, cognitive limitations, or behavioral biases is made.

What makes learning hard? It is intuitive that two individuals with common experience driving on U.S. highways will agree on which side of the road other drivers will use. It is far less obvious that two nutritionists, 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 event in an i.i.d. setting is equivalent to 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 difficult problem of using one sample to learn the probabilities of a *family of events* simultaneously. In this paper, I use the theory of uniform learning, also known as Vapnik–Chervonenkis theory, as the formal framework to model intuitive concepts like "a learning problem is hard" or "a set of events is statistically complex."[3]

In Section 2, I introduce the idea that decision makers use frequentist *models* to interpret evidence.[4] Theorem 1 identifies the essential tension between the amount of data available and the richness, or statistical complexity, of the set of events evaluated by the decision maker. Learning is straightforward in settings like repeated i.i.d. coin flips, where data is abundant and the set of alternatives to choose from is narrowly defined. In this case, frequentist, Bayesian, and just about any other sensible inference agree.

More interesting are situations where data is scarce relative to the statistical complexity of the set of alternatives being evaluated. Learning is hard in the impact-of-diet-on-health problem because we are concerned with learning

[3]This theory occupies a central role in modern statistics, but is relatively unknown to economic theorists. Two exceptions I am aware of are Kalai (2003) and Salant (2007). Section 2.5 provides a brief, self-contained exposition.

[4]The term "model" in this paper is used to refer both to the models used by decision makers to learn from their environments and to our formal description of that environment. The intended meaning will be clear from the context.

about many events simultaneously—namely how different diets affect individuals with different characteristics. In this case, a decision maker compensates for the scarcity of data by limiting inference to a statistically simple family of events.[5] Beliefs, which are pinned down only on a subset of events, are thus *statistically ambiguous*. As a result, different individuals with different models may draw different inferences and hold different beliefs based on the same data.

In Section 3, I turn to asymptotic properties of uniform learning as data increases without bound. On a practical level, large sample theories permit greater tractability and clearer intuitions. Another motivation is that equilibrium notions in economics are usually interpreted as capturing insights about steady-state or long-run behavior. A theory of learning in which statistical ambiguity is nothing more than a passing phenomenon will have little to say about steady-state behavior.[6]

Theorem 3 shows that the known theory of uniform learning has no bite in the limit. Specifically, in standard outcome spaces, which I take to be complete separable metric spaces with countably additive probabilities, all statistical ambiguity disappears in the limit. In these spaces, the tension between the availability of data and statistical complexity disappears. This is at odds with the central role this tension plays in finite settings in distinguishing between simple and hard learning problems.

I argue that the asymptotic elimination of statistical ambiguity in standard outcome spaces is a consequence of implicit structural restrictions these spaces impose. For example, the Borel events on [0, 1] are defined in terms of a topology that embeds a notion of similarity between outcomes. By restricting learning to Borel events, we in effect overcome statistical ambiguity through a substantive similarity assumption that should be made explicit, rather than built into the mathematical structure of the model.

To model a structure-free environment, I consider an arbitrary set of outcomes with the algebra of all events and all finitely additive probability distributions. Like the finite outcome case, this model is free from any inductive biases involving notions of distance, ordering, or similarity. In Theorem 4, I show that statistical ambiguity persists in the form of a set of probability measures representing beliefs that are not contradicted by data. Finally, in Section 4, I show how this set of beliefs can be integrated into standard models of decision making.

[5]A Bayesian decision maker, on the other hand, draws inferences about all events (by updating), but what he learns is highly sensitive to his prior.

[6]This point appears in Bewley (1988), who introduced the notion "undiscoverability" to capture the idea of stochastic processes that cannot be learned from data. His model and analysis are quite different from what is reported here.

2. UNIFORM LEARNING AND CONSISTENCY WITH EMPIRICAL EVIDENCE

2.1. *Basic Setup*

A decision maker uses i.i.d. observations to learn about the unknown probability distribution on a set of outcomes. My focus is on statistical inference and belief formation; decision making is discussed in Section 4.

To better convey the motivation, this section focuses on finite settings, where both the set of outcomes and the amount of data are finite.

BASIC MODEL—Finite Outcome Spaces $(X_f, 2^{X_f}, \mathcal{P}_f)$: X_f is a finite set, the set of events is the set of all subsets 2^X_f, and $\mathcal{P}_f$ is the set of all probability measures.

The decision maker bases his inference on repeated i.i.d. samples from $P \in \mathcal{P}_f$. Formally, let S denote the set of all infinite sequences of elements in X_f, interpreted as outcomes of infinite sampling. Under P, i.i.d. sampling corresponds to the product probability measure P^∞ on $(S, \mathcal{S})$, where $\mathcal{S}$ is the σ-algebra generated by the product topology. For an element $s = (x_1, \ldots) \in S$, let s^t denote the finite sample that consists of the first t observations from s. When discussing finite outcome spaces, I will assume that data is limited to finite samples of t observations.

2.2. *Informal Motivation and Intuition*

A decision maker is interested in learning the probabilities of events $A \subset X_f$. This decision maker does not have a prior belief about the probabilities of various events in X_f, but seeks instead uniform, that is, distribution-free, inferences about these probabilities.[7]

To be more specific, the decision maker observes a sequence $s^t = (x_1, \ldots, x_t)$ drawn i.i.d. from the true but unknown distribution P.[8] Define the empirical frequency of the event A relative to s^t by

$$\nu^t(A, s) \equiv \frac{\#\{i : x_i \in A, i \leq t\}}{t}, \tag{1}$$

where # denotes the cardinality of a finite set. The weak law of large numbers implies that the probability of any event can be estimated uniformly over all distributions when t is large. This can be stated formally as:[9]

[7]A discussion of the difficulties with the Bayesian procedure of starting with a prior and updating it using the data can be found in the working paper version of this paper.

[8]Many decision problems may be usefully modeled as stationary, while some nonstationary problems become stationary in a richer outcome space. In any event, if the underlying distribution is nonstationary, then one would expect learning to be even harder and for reasons quite distinct from those we wish to emphasize here. In particular, failure of learning would hold a fortiori in nonstationary settings where the object to be learned is constantly changing.

[9]This follows directly from Billingsley (1995, p. 86) applied to the indicator function of A.

LEMMA 1: *For every $\varepsilon > 0$ there is an integer t such that*

$$\forall A \subset X_f,\ \forall P \in \mathcal{P}_f: \quad P^\infty\{s : |P(A) - \nu^t(A, s)| < \varepsilon\} > 1 - \varepsilon. \tag{2}$$

Note that the sample size t in the lemma is independent of P, A, and $\#X_f$ (in fact, the lemma also holds for infinite outcome spaces). Inference about any single event is equivalent to learning from independent coin tosses, so it is not meaningful to talk about an event A being simple or complicated if all we are concerned about is learning the probability of A in isolation.

Choice involves, almost by definition, evaluating many acts *simultaneously*. To appreciate this point, define the set of *ε-good samples* for an event A as

$$\text{Good}^t_{\varepsilon,P}(A) \equiv \{s : |P(A) - \nu^t(A, s)| < \varepsilon\}.$$

This is the set of representative samples for the event A, that is, those samples on which the empirical frequency of A is close to the true probability.

Suppose now that the decision maker is choosing between bets $f_i, i = 1, \ldots, I$, with f_i paying 1 if the event A_i occurs and 0 otherwise. To make the problem interesting, use Lemma 1 to assume t large enough so that $P^\infty[\text{Good}^t_{\varepsilon,P}(A_i)] > 1 - \varepsilon$ for each A_i. This says that we can accurately estimate the expected payoff of each bet f_i, but says little about accurately comparing these expected payoffs. The latter is possible only at samples that are representative for all of the events $A_1, \ldots, A_I$ *simultaneously*. That is, what we need is for the probability

$$P^\infty\left[\bigcap_i \text{Good}^t_{\varepsilon,P}(A_i)\right] \tag{3}$$

to be large. Our assumption that $P^\infty[\text{Good}^t_{\varepsilon,P}(A_i)] > 1 - \varepsilon$ for each A_i only ensures that the probability of the intersection in (3) is at least $1 - I\varepsilon$, a conclusion that quickly becomes useless as the number of events being compared increases.

Roughly, a family of events $\{A_1, \ldots, A_I\}$ is *statistically simple* if the sets of samples $\text{Good}^t_{\varepsilon,P}(A_i)$, $i = 1, \ldots, I$, overlap so that if each event has high probability, then so would their intersection. In this case, the amount of data needed to learn the entire family is not larger than what is needed to learn any one of its members in isolation. By contrast, a family of events $\{A_1, \ldots, A_I\}$ is *statistically complex* if the intersection $\bigcap_i \text{Good}^t_{\varepsilon,P}(A_i)$ has low probability, even though there is enough data to guarantee that each set of samples $\text{Good}^t_{\varepsilon,P}(A_i)$ has probability at least $1 - \varepsilon$. In this case, learning the entire family requires considerably more observations than what would have been sufficient to learn any one of its members.

What determines whether a family of sets is statistically simple or complex? The answer is supplied by the beautiful and powerful theory of Vapnik and Chervonenkis (1971) that characterizes the learning complexity of a family of events. Section 2.5 provides a brief account of this theory.

2.3. *Uniform Learning*

DEFINITION 1—Uniform Learnability: A family of subsets $\mathcal{C} \subset 2^{X_f}$ is ε-*uniformly learnable* by data of size t, $\varepsilon > 0$, if

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

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

The crucial aspect of the definition is that learning is uniform over the events, so $\sup_{A \in \mathcal{C}}$ is inside the probability statement. This comes at the expense of limiting the scope of learning to a subset of events $\mathcal{C}$. The probability being evaluated in (4) is that of samples that are representative for *all events in* $\mathcal{C}$ *simultaneously*, that is, samples at which the empirical frequency of each event $A \in \mathcal{C}$ is close to its true probability. This suggests the following definition:

DEFINITION 2: A (*feasible*) *model* is a triple $(\mathcal{C}, \varepsilon, t)$, where $\mathcal{C}$ 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 of ε as the size of a confidence interval around $\nu^t(A, s)$. Extending this intuition, we define

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

as the set of distributions *consistent with empirical evidence*. A probability measure that does not belong to $\mu^t_{\mathcal{C},\varepsilon}$ is one that can be rejected with high confidence as inconsistent with the data.

In a model $(\mathcal{C}, \varepsilon, t)$ we shall interpret the collection of events $\mathcal{C}$ and the degree of confidence ε as reflecting the decision maker's model of his environment. The amount of available data t, on the other hand, is an objective constraint.

The feasibility of a model, by itself, is a hopelessly weak criterion; it is, for instance, trivially satisfied when $\mathcal{C} = \emptyset$ or $\varepsilon = 1$. It is normatively compelling to think of the decision maker as selecting models that are *maximal* in the sense that they do not overlook additional inferences that could have been drawn using the same amount of data t. Concerns about maximality are orthogonal to the main results of this paper. The interested reader will find a formal treatment of these ideas in the working paper version.

2.4. *Learning, Scarcity of Data, and the Order of Limits*

A central question of this paper is "What might lead individuals to adopt a model $(\mathcal{C}, \varepsilon, t)$ that involve a coarse representation of the true environment?"

Here, "coarse" means a model where $\mathcal{C} \subsetneq 2^{X_f}$. Our aim is to answer this question without appealing to bounded rationality or behavioral biases. We envision instead a frequentist decision maker with no prior beliefs who desires to draw uniform inferences from limited data. When data is scarce, the criterion of uniform learnability captures the intuition that a decision maker may limit the richness of the set of events he draws inference about, possibly to a set much smaller than the power set 2^{X_f}.[10] Formally,

THEOREM 1:

(i) *For every X_f and $\varepsilon > 0$, there is $\bar{t}$ such that 2^X_f is ε-uniformly learnable with data of size $t \geq \bar{t}$.*

(ii) *For every t, $\varepsilon > 0$ and $\alpha > 0$, there is $\bar{n}$ such that $\#X_f > \bar{n}$ implies*

$$\frac{\#\mathcal{C}}{\#2^X_f} < \alpha$$

for any $\mathcal{C}$ that is ε-uniformly learnable with data of size t.

Part (i) reflects a setting where data is plentiful: one fixes the finite outcome space X_f and then, taking the amount of data to infinity, guarantees uniform learning of the power set. Part (ii) reflects situations where data is scarce relative to the richness of X_f. In this case, the set of events that can be uniformly learned is a small fraction of the set of all events.

To further clarify these points, we note that the bound (2) corresponds to a statistical experiment in which a *new* sample of t observations is drawn to evaluate each event A, potentially requiring a preposterous amount of data when the number of events to be evaluated is large. The uniform learning criterion (4), on the other hand, requires the set of representative samples to be the same for all events in $\mathcal{C}$. It therefore corresponds to a statistical experiment in which inference is based on *one shot at sampling t observations*. When data is scarce, this forces the decision maker to restrict attention to a narrower, statistically simple family of events.

[10]A numerical example may help the reader appreciate the difficulty: suppose 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, so 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 $2^{2^{50}}$. While learning the probability of any single event may require only a manageable amount of data, uniformly learning the probability of all events would require an amount of data that is in the realm of fantasy—even by the standard of idealized economic models. For example, with $\varepsilon = 0.01$, using (A.8) in the Appendix, a *lower bound* on the required number of observations is 3.5×10^{15}, which is of the same order of magnitude as the estimated number of minutes since the Big Bang (roughly, 7.35×10^{15}).

To sum up, when data is scarce, individuals seeking uniform inference restrict the scope of the events and acts they consider.[11] For example, an investor may rely on macroeconomic or finance theories to restrict the distributions of returns. But despite decades of extensive and commonly shared evidence, even the best theories in these fields leave ample room for disagreement, as seen daily in conflicting policy recommendations, forecasts, and investment strategies. In environments like these, one is left with considerable freedom to choose which set of nonfactual, theoretically based restrictions to impose. It should therefore not be surprising that rational individuals may disagree even when facing identical information.

2.5. *Vapnik–Chervonenkis Theory*

Uniform learning can be given an elegant and insightful characterization using the theory of Vapnik and Chervonenkis. Since the concepts that follow can be defined for outcome spaces of any cardinality, let X be an arbitrary (possibly infinite) set.

The key concept of the theory is the *shattering capacity* of a family of sets $\mathcal{C}$. In the remainder of the paper, assume that $\mathcal{C}$ is closed under complements. Define the nth *shatter coefficient* of such $\mathcal{C}$ to be

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

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

Clearly, $s(\mathcal{C}, n) \leq 2^n$. The *Vapnik–Chervonenkis* (*or VC*) *dimension* of $\mathcal{C}$ is

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

If there is no such n, we write $V_{\mathcal{C}} = \infty$. In words, the VC dimension is the largest cardinality n such that there exists a set of n points that can be shattered by $\mathcal{C}$. The central result in statistical learning theory is given by the following theorem.

VC THEOREM: *A family of events $\mathcal{C} \subset 2^X$ is uniformly learnable if and only if it has finite VC dimension*.

[11] Al-Najjar and Pai (2008) study coarse decision making along these lines, spelling out in greater details the relationship between uniform learning and the problem of overfitting. Their paper then applies the framework to cognitive phenomena, like rules of thumb, categorization, linear orders, and satisficing, that appear anomalous from a Bayesian perspective.

A version of this result appeared in Vapnik and Chervonenkis (1971).[12] For a textbook treatment, see Theorems 12.5 and 13.3 in Devroye, Gyorfi, and Lugosi (1996).

A consequence of the theorem, stated as Equation (A.7) in the Appendix, relates the speed of learning $\mathcal{C}$ to its VC dimension. For a family of events $\mathcal{C}$ to have a small VC dimension means that it is not "too rich" to be uniformly learned. But the cardinality of a family $\mathcal{C}$ has at best a tangential relationship to its statistical complexity. For example, the family of half-intervals appearing in Example 1 is uncountable yet has a VC dimension of 2 and thus is easy to learn.

3. LARGE SAMPLE THEORY

I now turn to the asymptotic properties of uniform learning as the amount of data increases to infinity. There are at least three reasons why large sample theory is important. First, large samples make it possible to provide sharp definitions of concepts like statistical ambiguity and indeterminacy of beliefs. Second, one would like to know whether statistical ambiguity is robust in the limit as the amount of available data increases. Finally, equilibria in economic and game theoretic models are often viewed as steady states that arise as limits of learning processes.

To introduce scarcity of data and statistical ambiguity in the limit, I consider two models of infinite outcome spaces.

MODEL 1—Continuous Outcome Spaces $(X_c, \mathcal{B}, \mathcal{P}_c)$: X_c is a complete separable metric space, the set of events is the family of Borel sets $\mathcal{B}$, and $\mathcal{P}_c$ is the set of countably additive probability measures on $\mathcal{B}$.

A prototypical example is the continuum $[0, 1]$ with the usual metric topology.[13]

MODEL 2—Discrete Outcome Spaces $(X_d, 2^{X_d}, \mathcal{P}_d)$: X_d is an arbitrary infinite set with the discrete topology, the set of events is the set of all subsets 2^X_d, and $\mathcal{P}_d$ is the set of finitely additive probability measures on 2^{X_d}.

These spaces are "discrete" in the sense that there is no extraneous metric or measurable structure that restricts the set of events or probabilities.

As before, samples are drawn according to the product probability measure P^∞ on $(S, \mathcal{S})$, where $\mathcal{S}$ is the σ-algebra generated by the product topology

[12] An English translation of an earlier paper in Russian.

[13] Any continuous outcome space is, in a sense, equivalent to a subset of $[0, 1]$ with the metric topology, hence the use of the term "continuous" in describing these spaces. See Royden (1968, Theorem 8, p. 326) and the proof of Theorem 3.

on the set of infinite samples S.[14] Lemma 1, the concepts of uniform learning, shattering, the VC dimension and the VC Theorem all apply without change to infinite outcome spaces, assuming finite samples.

3.1. *Exact Learning and Statistical Ambiguity*

To minimize repetition, in this subsection, I use $(X, \Sigma, \mathcal{P})$ to stand for either the continuous or the discrete outcome model.

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

DEFINITION 3: A *learning strategy* is a sequence $\{(\mathcal{C}_n, \varepsilon_n, t_n)\}_{n=1}^{\infty}$ of models that satisfy the following conditions:

(i) $\varepsilon_n \to 0$.

(ii) $\mathcal{C}_n \subseteq \mathcal{C}_{n+1}$ for every n.

(iii) $\mathcal{C}_n$ is an ε_n-uniformly learnable family by data of size t_n.

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

As more data become available, the set of models that can be uniformly learned increases. Simple strategies increase confidence while holding $\mathcal{C}$ constant.

Given a learning strategy $\sigma = \{(\mathcal{C}_n, \varepsilon_n, t_n)\}_{n=1}^{\infty}$ and infinite sample s, the set of *beliefs consistent with empirical evidence* is

$$\mu_\sigma(s) \equiv \left\{ p : \forall n, \lim_{t \to \infty} \sup_{A \in \mathcal{C}_n} |p(A) - \nu^t(A, s)| = 0 \right\}.$$

The next theorem is a law of large numbers for limiting beliefs: on a "typical" sample, any probability distribution $p \in \mu_\sigma(s)$ assigns to each event $A \in \bigcup_n \mathcal{C}_n$ a probability equal to its true probability:

THEOREM 2—Exact Learning: *Fix any learning strategy* $\sigma = \{(\mathcal{C}_n, \varepsilon_n, t_n)\}_{n=1}^{\infty}$ *and write* $\mathcal{C}_\sigma = \bigcup_n \mathcal{C}_n$. *Then for any* $P \in \mathcal{P}$,

$$\mu_\sigma(s) = \{p : p(A) = P(A), \forall A \in \mathcal{C}_\sigma\}, \quad P^\infty\text{-}a.s.^{15} \tag{6}$$

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

[14]These are standard concepts in the case of X_c. Appendix A.1 provides the requisite background to cover the less familiar case of $(X_d, 2^{X_d})$.

[15]When P is only finitely additive, the notation P^∞ denotes the strategic product of P. See Appendix A.1 for details.

The main challenge in proving this result is to show that it holds for finitely additive probabilities, as required in Section 3.3 below.

Knowledge of the probabilities of events in $\mathcal{C}_\sigma$ may have implications for events outside $\mathcal{C}_\sigma$. For instance, if we know the probability of two disjoint events $A, B \in \mathcal{C}_\sigma$, then we can unambiguously deduce the probability of the event $A \cup B$ even if it did not belong to $\mathcal{C}_\sigma$. To make this formal, call a function $p : \mathcal{C}_\sigma \to [0, 1]$ a *partial probability* if it is the restriction to $\mathcal{C}_\sigma$ of some probability measure p' on Σ.[16] We can now give a formal definition of statistical ambiguity:

DEFINITION 4: An event $A \in \Sigma$ is (statistically) unambiguous relative to $\mathcal{C}_\sigma$ if, for any partial probability p on $\mathcal{C}_\sigma$, and any two extensions p' and p'' of p to Σ, $p'(A) = p''(A)$. Let $\mathcal{C}_\sigma^\star$ denote the set of all statistically unambiguous events (relative to $\mathcal{C}_\sigma$).[17]

In light of the definition and Theorem 2, we may therefore conclude that

$$\mu_\sigma(s) = \{p : p(A) = P(A),\ \forall A \in \mathcal{C}_\sigma^\star\}, \quad P^\infty\text{-a.s.} \tag{7}$$

Determinacy of beliefs can be defined in terms of the existence of learning strategies that eliminate statistical ambiguity:

DEFINITION 5: Beliefs are (asymptotically) determinate if there is a learning strategy σ such that $\mathcal{C}_\sigma^\star = \Sigma$. That is, under the strategy σ, for every $P \in \mathcal{P}$,

$$\mu_\sigma(s) = \{P\}, \quad P^\infty\text{-a.s.} \tag{8}$$

The question we turn to next is the determinacy of beliefs in the continuous versus the discrete model.

3.2. *Determinacy of Beliefs in Continuous Outcome Spaces*

The following theorem shows that no meaningful indeterminacy persists in the continuous model in the limit:

THEOREM 3: *In the continuous model* $(X_c, \mathcal{B}, \mathcal{P}_c)$ *beliefs are determinate via a simple learning strategy*.

[16] A more direct condition defining partial probabilities was identified by Horn and Tarski (1948). See also Bhaskara Rao and Bhaskara Rao (1983, Definition 3.2.2).

[17] While $\mathcal{C}_\sigma$ need not have any particular structure, $\mathcal{C}_\sigma^\star$ is easily seen to be a λ-system, that is, a family of events closed under complements and *disjoint* unions (Billingsley (1995)). For example, the set $\mathcal{C}$ of half-intervals in Example 1 is not closed under unions, but $\mathcal{C}^\star = \mathcal{B}$. The importance of λ-systems in the study of ambiguity was, to my knowledge, first pointed out by Zhang (1999).

That is, there is always a simple strategy that can "learn" the true distribution. The following example illustrates the theorem when $X_c = [0, 1]$:

EXAMPLE 1: Let $X_c = [0, 1]$, $\mathcal{B}$ the Borel sets on $[0, 1]$, and $\mathcal{P}$ the set of countably additive probabilities on $\mathcal{B}$. Consider the set $\mathcal{C}$ of half-intervals $[0, r]$, $r \in [0, 1]$, and their complements. Let σ be the simple strategy with $\mathcal{C}_n = \mathcal{C}$ for each n. Then:

(i) $\mathcal{C}$ is uniformly learnable.

(ii) Agreement on $\mathcal{C}$ implies agreement on all Borel sets.

By Theorem 2, any $p \in \mu_\sigma(s)$ must agree with the true P on $\mathcal{C}$ almost surely. Therefore p and P define identical distribution functions, hence identical probability measures on $\mathcal{B}$.[18]

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

• *Statistical Learning:* The set of half-intervals in $[0, 1]$ is uniformly learnable. This is the classical Glivenko–Cantelli theorem.[19]

• *Deduction:* The half-intervals are sufficient to determine beliefs on all Borel events.

The theorem generalizes the intuition in Example 1 by showing that any complete separable metric space contains a uniformly learnable family that determines beliefs in the limit. As shown in the proof, a belief-determining family can be found whose structure is similar to that of half-intervals. It is difficult to think of bounded rationality reasons that would prevent a decision maker from using simple learning procedures like these.

Theorem 3 reveals that continuous outcome spaces fail to capture the limiting behavior in finite settings, where indeterminacy of beliefs is natural. In the next subsection, I will argue that the conclusion of Theorem 3 is an artifact of the structure of X_c which distorts the learning problem by restricting the sets of permissible events and distributions. These restrictions are artificial in the sense that they have no counterparts in finite models.

3.3. *Indeterminacy of Beliefs in Discrete Outcome Spaces*

In this section, I consider asymptotic learning in the discrete outcome space $(X_d, 2^{X_d}, \mathcal{P}_d)$. First we need the following definition:

[18]Note that $\mathcal{B}$ itself is not uniformly learnable. This can be easily seen from the fact that $\mathcal{B}$ has infinite VC dimension. What matters for eliminating disagreements in the limit is that there is a uniformly learnable family (the subintervals) that is sufficient to determine beliefs on $\mathcal{B}$.

[19]The Glivenko–Cantelli theorem states that the empirical distribution function converges to the true distribution function uniformly almost surely. This theorem follows from the Vapnik–Chervonenkis theorem by noting that the half intervals have VC dimension of 2. To see this, any pair of points $x_1, x_2 \in X_f$ can be shattered by $\mathcal{C}$, so $V_\mathcal{C} \geq 2$. Given any set of three points $x_1 < x_2 < x_3$, intersections with elements of $\mathcal{C}$ generate the sets $\{x_1\}, \{x_3\}, \{x_1, x_2\}$, and $\{x_2, x_3\}$, but no intersection can generate the singleton set $\{x_2\}$. Since no set with three points can be shattered, we have $V_\mathcal{C} = 2$.

DEFINITION 6: Beliefs are (asymptotically) indeterminate if there exists P such that for every learning strategy σ,

$$\mu_\sigma(s) \neq \{P\}, \quad P^\infty\text{-a.s.}$$

Indeterminacy is stronger than the negation of determinacy in two ways. First, the quantifiers are reversed: one can find a single "difficult-to-learn" distribution P that cannot be identified from the data regardless of the learning strategy used. Second, the failure to identify P occurs with probability 1, rather than just with positive probability. The relationship with statistical ambiguity is that if beliefs are indeterminate, then there are (statistically) ambiguous events under any learning strategy.

THEOREM 4: *Beliefs are statistically indeterminate in any discrete outcome space* $(X_d, 2^{X_d}, \mathcal{P}_d)$.

To compare Theorems 3 and 4, note first that statistical inference in X_d works just like it did in continuous outcome spaces. What changes here is that beliefs on 2^{X_d} are no longer determined by a uniformly learnable $\mathcal{C}$. The proof builds on a fundamental combinatorial result, known as Sauer's lemma, that bounds the cardinality of uniformly learnable families in finite outcome spaces. This result cannot be directly used here because we must consider infinite families of events where information about cardinality is not very useful. This necessitates a more delicate indirect argument in which finitely additive probabilities are used in an essential way.

The scope of disagreement asserted in the theorem can be substantial:

COROLLARY 1: *Given any discrete outcome space* $(X_d, 2^{X_d}, \mathcal{P}_d)$, *uniformly learnable* $\mathcal{C}$, *and* $\alpha \in (0, 0.5]$, *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.

3.4. *The Role of Finite Additivity*

We use infinite outcome spaces and infinite data to gain new insights into settings with finite outcomes and scarce data. The contrast between Theorems 3 and 4 thus reflects that continuous models fail, and discrete models succeed, as idealizations of finite settings.

Why Asymptotic Learning Is Easy in Continuous Models

In the continuous outcome space $(X_c, \mathcal{B}, \mathcal{P}_c)$, the amount of data tends to infinity, suggesting that learning is easier than in finite settings. On the other hand, $\mathcal{B}$ contains infinitely many events, suggesting that learning should be harder and statistical ambiguity more severe. The conclusion of Theorem 3

that statistical ambiguity always disappears in the limit may therefore seem puzzling.[20]

The puzzle is explained by noting that the continuous model is loaded with structural assumptions and inductive biases. Although they may appear as innocuous regularity conditions, these assumptions and biases substantively drive (and, in my view, mislead) our intuition. I illustrate with two prototypical examples. Consider first the case $X_c = [0, 1]$ with the Borel sets generated by the usual metric topology. Here, a decision maker can eliminate statistical ambiguity in the limit by first learning the probabilities of the half-intervals and then using them to deduce those of the remaining Borel events. Non-Borel events are cast out as illegitimate, thus simplifying the learning problem by limiting the range of events the decision maker is able to contemplate.

Consider next the case where X_c is countable with the discrete topology. In this case, the set of events $\mathcal{B}$ is 2^{X_c}, so no event is a priori ruled out. Here, the mismatch with the finite-outcome-space intuition is that countable additivity requires probability distributions to be concentrated on negligible subsets of the outcome space.

To make this precise, let $\{x_1, \ldots\}$ be an arbitrary enumeration of X_c. Fix a small $\alpha > 0$ and define the (random) integer:

$$N(s) = \min_n \{\nu(\{x_1, \ldots, x_n\}, s) > 1 - \alpha\}.$$

This is the smallest integer n such that the empirical distribution ν concentrates $1 - \alpha$ mass on the finite set $\{x_1, \ldots, x_n\}$.

The family $\mathcal{C} = \{\{x_1, \ldots, x_n\}; n = 1, 2, \ldots\}$ of initial segments has a finite VC dimension and thus is uniformly learnable.[21] Using the VC theorem, for any $\varepsilon > 0$ there is $\bar{t}$ such that for all $P \in \mathcal{P}_c$ and $t \geq \bar{t}$,

$$P^\infty\{s : P(\{1, \ldots, x_{N(s)}\}) > 1 - \alpha - \varepsilon\} > 1 - \varepsilon.$$

In words, without prior knowledge of P (other than that it belongs to $\mathcal{P}_c$), the decision maker can determine from finite-sample information the integer $N(s)$, and hence the initial segment $\{x_1, \ldots, x_{N(s)}\}$ on which the true distribution is concentrated. Once this initial segment is known, the problem all but reduces to one with a *fixed* finite set of outcomes. Increasing the amount of data beyond $\bar{t}$ corresponds (approximately) to case (i) of Theorem 1, where the set of outcomes is fixed but data increase without bound. This conflicts with the intuition, formalized in case (ii) of that theorem, that scarcity of data can be important when the set of outcomes is finite but rich enough.

[20]Commenting on Theorem 4, a referee noted that "one might have conjectured a possibility result, presumably because intuitions live in metric spaces."

[21]This can be shown using an argument similar to that appearing in footnote 19.

The Finite-Outcome-Space Motivation

There is little doubt that individuals rely on cognitive devices, such as ordering or similarity, to organize information and guide learning when data is scarce (hence the opening quote of this paper). But to understand why these cognitive devices look the way they do, our model of an outcome space should act as a neutral backdrop against which they may arise as objects of choice, rather than being built into the primitives. Finite outcome spaces represent one such class of models as they embed no a priori inductive biases like notions of distance, ordering, or similarity.

It would be odd to build into the primitives of a finite model a distinguished family of events as the only legitimate ones to consider or to restrict attention to distributions that place most of their mass on a small fraction of the total number of outcomes. Yet this is what the mathematical structure of events $\mathcal{B}$ and distributions $\mathcal{P}_c$ impose in the continuous model. These restrictions limit the scope of statistical ambiguity and diversity of beliefs by fiat. By contrast, the discrete space $(X_d, 2^{X_d}, \mathcal{P}_d)$, just like finite outcome spaces, is free from any such a priori structures.

The Foundational Case for Finite Additivity

Do finitely additive probabilities have a meaningful interpretation?[22] Paradoxically, in the axiomatic foundations of decision theory, it is the requirement of countable additivity that is viewed as questionable and demands justification. Savage and de Finetti held that the fundamental axioms from which subjective probability is derived only imply finite additivity. Savage's celebrated axiomatization, as well as many subsequent ones, was cast in a finitely additive setting. Countable additivity of subjective probability is an assumption to be introduced for expedience, not foundational considerations.[23]

de Finetti and Savage's insistence on finite additivity is not an expression of a desire for technical generality or idiosyncratic modeling taste. Rather, it

[22]Another concern is whether finitely additive models are tractable. They are certainly not as tractable as countably additive probabilities. However, the widespread misperception that none of the classical results of probability theory applies to them is just that—a misperception. In Appendix A.1 and in Al-Najjar (2007), I indicate that much of the classical theory applies once natural technical conditions are imposed.

[23]de Finetti's (1974, p. 123) view is reflected in the following quote: "Suppose we are given a countable partition into events E_i, and let us put ourselves into the subjectivistic position. An individual wishes to evaluate the p_i: he is free to choose them as he pleases [...] Someone tells him that in order to be coherent he can choose the p_i in any way he likes, so long as the sum $= 1$ (it is the same thing as in the finite case, anyway!).

The same thing?!!! You must be joking, the other will answer. In the finite case, this condition allowed me to choose the probabilities to be all equal, or slightly different, or very different; in short, I could express any opinion whatsoever. [Now] I am obliged to pick "at random" a convergent series which, however I choose it, is in absolute contrast to what I think. If not, you call me incoherent! In leaving the finite domain, is it I who has ceased to understand anything, or is it you who has gone mad?"

reflects the methodological separation between (a) what constitutes structural assumptions about the choice setting and (b) the feasibility constraints facing the decision maker in a particular choice problem. As an example, take an outcome space X that has the cardinality of the continuum. In Savage's model the set of acts $\mathcal{F}$ is the set of *all* functions that map points in X to consequences. Suppose, for whatever reason, that the decision maker wants to introduce a metric structure based on a linear order, perhaps to incorporate a notion of similarity between outcome, or some other concerns. This can be formalized as a choice of a bijection $\phi : X \to [0, 1]$ that imports the metric topology of $[0, 1]$ onto X. It would then be natural to consider the restriction to the set of acts $\mathcal{F}_\phi$ that are measurable with respect to the Borel structure implied by ϕ. In Savage's theory, the selection of a specific ϕ to represent, say, a notion of similarity between outcomes is modeled as the constraint that the decision maker must choose from the feasible set $\mathcal{F}_\phi$. The de Finetti–Savage case for finite additivity is that one should not confuse constraints like $\mathcal{F}_\phi$ with the structure of the choice problem where all acts are permitted. This structure is invariant, while constraints are not.

A common argument used to justify the removal of non-Borel sets from considerations is that they cannot be described in terms of finite sets of intervals and their limits.[24] In Savage's theory, describability is not a primitive but a constraint like any other. For example, one may find a linear order on $[0, 1]$, and the describability constraints it implies, intuitive. This paper takes a different point of view: structures like linear orders are devices decision makers use to facilitate learning. While linear orders may seem natural or canonical structures when describing prices or quantities, it is just as easy to think of examples where no obvious a priori structures exist: What is a natural linear order on a set of players in a large game, on the set of diets or medical conditions, or on past experiences with presidential elections or military contests?[25] When we model these problems using finite outcomes, we *choose* to introduce structures like orders, metrics, or similarities, since without them learning would be impossible. But it would seem unreasonable to have these structures appear as part of the primitives.

4. DIVERSITY, AMBIGUITY, AND DECISION MAKING

The main concern of this paper is with belief formation; that is, with questions like, "Where do beliefs come from and what makes them 'reasonable?" An orthogonal, but equally important, question is, "What decisions would individuals make given their beliefs?" Here, I sketch how uniform learning may be integrated into standard models of decision making.[26]

[24] For a formal model of undescribability, see Al-Najjar, Anderlini, and Felli (2006).

[25] To put this in perspective, there are $52! \approx 8 \times 10^{67}$ possible linear orders in a deck of 52 cards, roughly the number of atoms in a typical galaxy.

[26] The working paper version contains a more formal and detailed discussion.

An informal outline may be helpful. Many models in the literature represent beliefs as sets of probability measures to reflect decision makers with insufficient knowledge to form precise probabilistic beliefs. The set of probabilities in these models is usually derived axiomatically and *interpreted* as capturing the decision maker's limited understanding of his environment. This paper proceeds in a different direction: I use an explicit model of learning to derive a set of probability measures $\mu_\sigma(s)$ consistent with empirical evidence; I then combine this objective information with subjective decision making criteria to produce choice behavior.

To minimize repetition, we continue to use $(X, \Sigma, \mathcal{P})$ to stand for either the continuous outcome or the discrete outcome model. I also limit attention to infinite samples to streamline the discussion. Our focus will be on acts of the form

$$f : X \to \mathcal{R},$$

where we interpret f to be valued in utils in order to abstract from the decision maker's risk attitude.

BEWLEY'S INCOMPLETE PREFERENCES CRITERION:

$$f \succsim^*_{\sigma,s} g \iff \int_X f\,dP \geq \int_X g\,dP \quad \forall P \in \mu_\sigma(s). \tag{9}$$

Bewley (1986) axiomatized the behavior of a decision maker whose preference may be incomplete. His representation consists of a set of probability measures K and the criterion that f is preferred to g if and only if f yields higher expected payoff under any $P \in K$. Criterion (9) coincides with Bewley's when $K = \mu_\sigma(s)$.

Bewley's model is sometimes informally interpreted as[27] (i) the set K is a set of "objective distributions" representing the decision maker's information and (ii) the decision maker prefers f to g if and only if f has higher expected payoff under any objective distribution. Although intuitively appealing, this interpretation has no formal basis in Bewley's setup and axioms. The set K in his model is derived axiomatically from the decision maker's preference and need not have any objective interpretation.

Using the framework of this paper, we can formally interpret the set of measures $\mu_\sigma(s)$ as resulting from a learning process. When beliefs are determinate, $\mu_\sigma(s)$ collapse to a single measure P, in which case learning is complete and so is the preference $\succsim_{\sigma,s}$. By contrast, when beliefs are indeterminate, $\succsim^*_{\sigma,s}$ is

[27]See, for instance, Bewley (1988) and, more recently, Gilboa, Maccheroni, Marinacci, and Schmeidler (2008).

necessarily incomplete. Learning provides a motivation for what makes events (un)ambiguous and sheds light on how $\mu_\sigma(s)$ varies with samples.[28]

THE MAXIMIN EXPECTED UTILITY CRITERION:

$$f \succsim^{\circ}_{\sigma,s} g \quad \Longleftrightarrow \quad \inf_{P\in\mu_\sigma(s)} \int_X f\,dP \geq \inf_{P\in\mu_\sigma(s)} \int_X g\,dP. \tag{10}$$

This is the functional form introduced by Gilboa and Schmeidler (1989) with $\mu_\sigma(s)$ substituting for their subjectively derived set of measures. Gajdos, Hayashi, Tallon, and Vergnaud (2008) provided an axiomatic model of how objective information, in the form of a set of measures, can be incorporated into the subjective maximin expected utility setting. The difference is that the set $\mu_\sigma(s)$ in our case has a specific motivation in terms of frequentist learning, a motivation lacking in these authors' more abstract formulation.[29]

If beliefs are asymptotically determinate, then $\mu_\sigma(s)$ is a singleton, ambiguity disappears, and the decision maker behaves exactly as a Bayesian. The framework of this paper makes it possible to relate the persistence of ambiguity to the failure of learning to pin down a unique distribution.

THE BAYESIAN CRITERION:

$$f \succsim^{\bullet}_{\sigma,\varphi,s} g \quad \Longleftrightarrow \quad \int_X f\,dP \geq \int_X g\,dP,$$

where $P = \varphi(\mu_\sigma(s))$ and φ is a selection from the correspondence $s \mapsto\!\!\!\rightarrow \mu_\sigma(s)$.

Here the decision maker selects an element of the set of measures $\mu_\sigma(s)$ and behaves as a Bayesian given this selection. This amounts to selecting a Bayesian completion of the incomplete preference $\succsim^{*}_{\sigma,s}$ in the Bewley formulation (9).

We can then shed some light on the question: Should individuals who have observed a large, common pool of data hold the same beliefs? A commonly held view is that differences in opinions are due only to differences in information. This is best expressed by Aumann (1987, pp. 12–13):

> 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.

If beliefs are asymptotically determinate, $\mu_\sigma(s)$ is a singleton and the selection $\varphi(\mu_\sigma(s))$ is unique. In this case, learning forces all individuals who observe

[28]In particular, any two measures $P, P' \in \mu_\sigma(s)$ must agree on $\mathcal{C}^\star_\sigma$ almost surely. Lehrer (2005) made a similar point in a very different context.

[29]Gajdos et al. (2008) axiomatized a more general form where the inf in (10) is taken over a subset of $\mu_\sigma(s)$.

a common pool of data to hold identical beliefs in the limit. But if beliefs are asymptotically indeterminate, then two individuals who observe the same data may hold different beliefs either because their subjective φ's differ or because they use different learning strategies. In both cases, they draw different inferences from the same evidence, even in the limit.

APPENDIX: PROOFS

A.1. *Strategic Product Measures*

Defining sampling for a continuous outcome space X_c is standard: we take as the sample space Ω the product $X_c \times X_c \times \cdots$ endowed with the Borel σ-algebra generated by the product topology. In the discrete case X_d, on the other hand, we must appeal to concepts and results that may be unfamiliar to some readers. Here we give each coordinate the *discrete* topology and define the sample space as the product $\Omega = X_d \times X_d \times \cdots$ with the product topology. As in the countably additive case, we take as the 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_d. 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 need not be uniquely defined.

Dubins and Savage (1965) dealt with this problem in their book on stochastic processes by introducing 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 X_d has a unique extension to a strategic product λ^∞ on 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's (1976) constructions because they are not essential for what follows. For the purpose of the present paper, what the reader should bear in mind is (a) the concept of strategic products is a natural restriction (for example, all product measures in the countably additive setting are strategic) and (b) 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 2*

The theorem is standard when the outcome space is finite or continuous. The main challenge is to provide arguments that do not require countable additivity. To avoid repetition, this proof applies to an outcome space X that stands

for either X_c or X_d with the corresponding structures. Also, the notation P^∞ will always denote the strategic product of P (which, in the case of a countably additive P, coincides with the usual product).

LEMMA A.1: *Fix any uniformly learnable $\mathcal{C}$ and probability measure P. Then*:

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

PROOF: From (A.7) we have that for every $P\in\mathcal{P}$ and $\varepsilon>0$,

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

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

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

Take a sequence $\varepsilon_n\downarrow 0$ and note that each event

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

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

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

hence

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

For a family of sets $\mathcal{C}$ and sample s, define

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

This is just the counterpart of $\mu_\sigma(s)$ for a single family of events $\mathcal{C}$. For an arbitrary sample, this set of measures can be badly behaved or even empty. The following lemma characterizes it on a typical sample:

LEMMA A.2: *For any uniformly learnable $\mathcal{C}$ and probability measure P, we have, P^∞-a.s.,*

$$\mu_{\mathcal{C}}(s) = \left\{ p : \sup_{A \in \mathcal{C}} |p(A) - P(A)| = 0 \right\}. \tag{A.1}$$

PROOF: Lemma A.1 states that the event

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

has P^∞-probability 1. Thus, in the argument below, we restrict attention to samples s in this event. For any such s, given $\varepsilon > 0$, we have $\sup_{A \in \mathcal{C}} |\nu^t(A, s) - P(A)| < \varepsilon$ for all large t.

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

$$\begin{aligned} \sup_{A \in \mathcal{C}} |p(A) - P(A)| &\leq \sup_{A \in \mathcal{C}} \big[|p(A) - \nu^t(A, s)| + |\nu^t(A, s) - P(A)| \big] \\ &\leq \sup_{A \in \mathcal{C}} |p(A) - \nu^t(A, s)| + \sup_{A \in \mathcal{C}} |\nu^t(A, s) - P(A)| \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

so p is in the right-hand side of (A.1).

Conversely, if p belongs to the set in the right-hand side of (A.1), then fixing $\alpha > 0$ and taking t large enough, we have

$$\begin{aligned} \sup_{A \in \mathcal{C}} |p(A) - \nu^t(A, s)| &\leq \sup_{A \in \mathcal{C}} |p(A) - P(A)| + \sup_{A \in \mathcal{C}} |P(A) - \nu^t(A, s)| \\ &\leq \varepsilon + \alpha. \end{aligned}$$

Since α is arbitrary, the conclusion follows. *Q.E.D.*

PROOF OF THEOREM 2: For a learning strategy $\{(\mathcal{C}_n, \varepsilon_n, t_n)\}_{n=1}^\infty$ and integer $\bar{n}$, we note that

$$\mu_\sigma(s) = \bigcap_{\bar{n}=1,2,\ldots} \mu_{\mathcal{C}_{\bar{n}}}(s).$$

Any event of the form

$$\left\{ s : \mu_{\mathcal{C}_{\bar{n}}}(s) = \left\{ p : \sup_{A \in \mathcal{C}_{\bar{n}}} |p(A) - P(A)| = 0 \right\} \right\}$$

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

$$\bigcap_{\tilde{n}=1,2,\ldots} \left\{ s : \mu_{\mathcal{C}_{\tilde{n}}}(s) = \left\{ p : \sup_{A \in \mathcal{C}_{\tilde{n}}} |p(A) - P(A)| = 0 \right\} \right\}$$

also has P^∞-probability 1. From this it follows that

$$P^\infty \left\{ s : \bigcap_{\tilde{n}=1,2,\ldots} \mu_{\mathcal{C}_{\tilde{n}}}(s) = \left\{ p : \sup_{A \in \mathcal{C}_{\tilde{n}}} |p(A) - P(A)| = 0 \right\} \right\} = 1. \qquad \textit{Q.E.D.}$$

A.3. *Proof of Theorem 3*

This is essentially a consequence of two facts: (i) all complete separable metric spaces are "equivalent" to a Borel subset of [0, 1] and (ii) 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), there is a Borel subset $B \subset [0, 1]$ and a measurable bijection $\phi : X_c \to B$ such that ϕ^{-1} is also measurable. For each $r \in [0, 1]$ define $A_r = \phi^{-1}([0, r])$ and let $\mathcal{C} = \{A_r : r \in [0, 1]\}$. That is, the collection $\mathcal{C}$ mimics the structure of half-intervals in [0, 1]. Note that these sets need not preserve the geometric properties of the half-intervals (e.g., connectedness). They are, however, nested: $A_r \subsetneq A_{r'}$ whenever $r < r'$. It is easy to verify that the family of sets $\mathcal{C}$ has VC dimension of 1.[30]

Consider any simple learning strategy $\{(\mathcal{C}_n, \varepsilon_n, t_n)\}_{n=1}^\infty$ with a constant $\mathcal{C}_n = \mathcal{C}$. Using Theorem 2, we have

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

Fix any sample path s for which the above holds and fix $p \in \mu_\sigma(s)$.

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 $\tilde{p}(A) \equiv p(\phi^{-1}(A))$ and $\tilde{P}(A) \equiv P(\phi^{-1}(A))$. 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 Theorem 4*

Fix a discrete space $(X_d, 2^{X_d}, \mathcal{P}_d)$ and let $(X'_d, 2^{X'_d}, \mathcal{P}'_d)$ be a subspace, where X'_d is an infinite subset of X_d and $\mathcal{P}'_d$ is the set of all finitely additive prob-

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

abilities that put unit mass on X'_d. To show that beliefs on X_d are indeterminate, it suffices to display a subspace X'_d on which they are. The strategy I follow is to focus on an increasing sequence $\{X_N\}_{N=1}^{\infty}$ of finite subsets of X_d and prove indeterminacy for the outcome space $X'_d \equiv \bigcup X_N$. Since this procedure is applicable in any (infinite) outcome space X_d, to avoid redundant notation, *assume for the remainder of the proof that X_d is countable*.

I start with following proposition which establishes the result for a single uniformly learnable family $\mathcal{C}$. The general case will follow as a corollary:

PROPOSITION A.1: *There is a finitely additive probability measure λ on the discrete outcome space $(X_d, 2^{X_d})$ such that for every uniformly learnable family of events $\mathcal{C}$, there are uncountably many distinct (finitely additive) probability measures that agree with λ on $\mathcal{C}$.*

The proof proceeds in three steps: (i) Construct a "nice" finitely additive probability measure λ on $(X_d, 2^{X_d})$. (ii) Given any $\mathcal{C}$, construct a class of perturbations of the density of λ with the property that they leave λ unaffected on $\mathcal{C}$. (iii) Show that each such perturbation defines a finitely additive probability measure distinct from λ.

A.4.1. *Constructing λ*

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

$$\eta_{N-1} < \frac{\eta_N}{N}, \quad \text{where} \quad \eta_N \equiv \#X_N.$$

This says that the cardinality of X_N increases rapidly with N. 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 $\{N : |x_N - x| < \varepsilon\}$ belongs to $\mathcal{U}$ for every $\varepsilon > 0$. 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.

For readers not familiar with these concepts, the idea is to define the probability of the event A, $\lambda(A)$, as a limit of the finite probabilities $\lambda_N(A)$. If the sequence $\{\lambda_N(A), N = 1, 2, \ldots\}$ converges, then the statement that $\lambda(A) \equiv \lim_{N\to\infty} \lambda_N(A)$ is equivalent to saying that the set of integers $\{N : |\lambda_N(A) - \lambda(A)| < \varepsilon\}$ is cofinite (i.e., complement of a finite set) for every $\varepsilon > 0$. That is, $\lambda_N(A)$ converges to $\lambda(A)$ means that the set of N's on which $\lambda_N(A)$ and $\lambda(A)$ are ε apart is small for all $\varepsilon > 0$, 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; 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.[31]

A.4.2. *Perturbations*

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

Let π be the unique *countably* additive product measure on $(V, \mathcal{V})$, assigning probability 0.5 to each of the events $\{s : s(x) = 1 + \varepsilon\}$. That is, π is constructed by taking equal probability i.i.d. randomizations for $s(x) \in \{1 - \varepsilon, 1 + \varepsilon\}$. Note that $(V, \mathcal{V}, \pi)$ is a standard countably additive probability space constructed using standard methods. The only finite additivity is in the measure λ.

Fix an arbitrary N. For any event $A \subset X_d$, we use A_N to denote the finite set $A \cap X_N$ 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.

If $\mathcal{C}$ has finite VC dimension v on X_d, then no subset of $v + 1$ points in X_d 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 $\mathcal{C}_N$ is at most v. A fundamental 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. That is, the cardinality of a family of subsets is polynomial in η_N with degree equal to its VC dimension.

To appreciate this bound, recall that X_N contains 2^{η_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 $\eta_N = 50$, if $\mathcal{C}$ has a VC dimension of 5, say, then the ratio of the number of events in $\mathcal{C}$ to the power set is no more than 5.5×10^{-7}.

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

This cardinality argument, while suggestive, does little for us in the limit: when the size of X_N goes to infinity and, holding v fixed, 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_d 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.

Since the perturbations are independent, Hoeffding's inequality (see, e.g., Devroye, Gyorfi, and Lugosi (1996, Theorem 8.1, p. 122)) implies that for any subset $A_N \in \mathcal{C}_N$,

$$\pi\left\{s: \frac{1}{\#A_N}\left|\sum_{x\in A_N} s(x) - \#A_N\right| > \alpha\right\} \leq 2e^{-2\#A_N\alpha^2}.$$

This inequality is not particularly useful without some bounds on $\#A_N$. So for $0 < \beta \leq 0.5$, let $\mathcal{C}_{N,\beta}$ denote the family of events $\{A_N : A = X_d$, or $A \in \mathcal{C}$ and $\lambda_N(A) > \beta\}$. Restricting attention to N's with $\frac{1}{N} < \beta$, we have

$$2e^{-2\#A_N\alpha^2} \leq 2e^{-2\beta\eta_N\alpha^2}.$$

Since, by Sauer's lemma, there are no more than $2(\eta_N)^v$ events in $\mathcal{C}_{N,\beta}$, we obtain

$$\pi(Z_{\alpha,\beta,N}) \leq 4(\eta_N)^v e^{-(2\beta\alpha^2)\eta_N},$$

where

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

Summing up over N, for fixed α and β, we obtain

$$\sum_{N=1}^{\infty} \pi(Z_{\alpha,\beta,N}) \leq 4\sum_{N=1}^{\infty}(\eta_N)^v e^{-(2\beta\alpha^2)\eta_N} < \infty.$$

By the Borel–Cantelli lemma (the usual version, since π is countably additive), the set $Z_{\alpha,\beta}$ of perturbations that belong to infinitely many of the $Z_{\alpha,\beta,N}$'s has π -measure 0. This implies (again using the countable additivity of π) that the event

$$Q \equiv \bigcap_{k=1}^{\infty}\bigcap_{k'=1}^{\infty}(Z_{\alpha=1/k,\beta=1/k'})^c \tag{A.2}$$

has π-measure 1. In particular, Q is not empty.

The preceding argument is the heart of the proof. Think of the 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 $\lambda(A) = \int_A s\,d\lambda$. Any such perturbation s defines a new probability measure $\gamma(A) \equiv \int_A s\,d\lambda$ that leaves the probability of A intact yet differs from A at least on the event $B \equiv \{x : s(x) = 1 - \varepsilon\}$. The proposition is proven by showing the existence of perturbations s that accomplish this not just with respect to a single event A, but all events in $\mathcal{C}$ simultaneously. This argument, which culminates in Appendix A.4.3, is founded on the material above.

The strategy is to draw, for each x, a value in $\{1 + \varepsilon, 1 - \varepsilon\}$ 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 π-almost surely neutral with respect to A. Since the intersection of countably many π-measure 1 sets has π-measure 1, this conclusion can be extended to any countable family of events $\{A_1, A_2, \ldots\}$. 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, and Borel sets are uncountable. A less direct and more subtle argument is needed.

Here, the assumption that $\mathcal{C}$ has finite VC dimension plays a critical role via 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 larger families of events, provided their cardinalities do not grow too quickly. Sauer's lemma delivers the slow rate of growth, asserting 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 has 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.

A.4.3. *Perturbed Measures*

For a fixed s and any event A, define

$$A^+ = \{x \in A : s(x) = 1 + \varepsilon\}$$

and

$$A^- = \{x \in A : s(x) = 1 - \varepsilon\}.$$

Although there is a well developed theory of integration with respect to finitely additive probabilities, for the purpose of this proof all we need is to define

$$\int_A s(x)\,d\lambda \equiv (1 + \varepsilon)\lambda(A^+) + (1 - \varepsilon)\lambda(A^-) \tag{A.3}$$

and

$$\begin{aligned}\int_A s(x)\,d\lambda_N &\equiv \frac{1}{\eta_N} \sum_{x \in A_N} s(x) \\ &= (1 + \varepsilon)\lambda_N(A^+) + (1 - \varepsilon)\lambda_N(A^-).\end{aligned}$$

LEMMA A.3: *For all $A \in \mathcal{C}$ and $s \in Q$,*

$$\int_A s(x)\, d\lambda = \lambda(A).$$

PROOF: Fix a set A with $\beta \equiv \lambda(A) > 0$ and an $\alpha > 0$. We deal with the case $\lambda(A) = 0$ separately. Belonging to Q implies that for all N large enough, $s \in Z_{\alpha,\beta,N}$; thus,

$$\max_{A_N \in \mathcal{C}_{N,\beta}} \frac{1}{\#A_N} \left| \sum_{x \in A_N} s(x) - \#A_N \right| < \alpha.$$

In this case, multiplying both sides by $\#A_N/\#\eta_N$, we get

$$\max_{A_N \in \mathcal{C}_{N,\beta}} \frac{1}{\#\eta_N} \left| \sum_{x \in A_N} s(x) - \#A_N \right| < \frac{\#A_N}{\eta_N}\alpha \leq \alpha.$$

Substituting in the definitions of $\int_A s(x)\, d\lambda_N$ and $\lambda_N(A)$, we have that for all large enough N,

$$\max_{A_N \in \mathcal{C}_{N,\beta}} \left| \int_A s(x)\, d\lambda_N - \lambda_N(A) \right| < \alpha. \tag{A.4}$$

From the definition (A.3) and the properties of ultrafilters, there is a subsequence $\{N_k\}$ such that[32]

$$\begin{aligned} \int_A s(x)\, d\lambda - \lambda(A) &\equiv \left[(1+\varepsilon)\, \mathcal{U}\text{-}\lim_{N\to\infty} \lambda_N(A^+) + (1-\varepsilon)\, \mathcal{U}\text{-}\lim_{N\to\infty} \lambda_N(A^-) \right] \\ &\quad - \mathcal{U}\text{-}\lim_{N\to\infty} \lambda_N(A) \\ &= \left[(1+\varepsilon) \lim_{k\to\infty} \lambda_{N_k}(A^+) + (1-\varepsilon) \lim_{k\to\infty} \lambda_{N_k}(A^-) \right] \\ &\quad - \lim_{k\to\infty} \lambda_{N_k}(A) \\ &= \lim_{k\to\infty} \left[\int_A s(x) d\lambda_{N_k} - \lambda_{N_k}(A) \right]. \end{aligned} \tag{A.5}$$

[32]To see this, fix a pair of sets A_1 and A_2 and an integer k. Then the set $U_{i,k} \equiv \{N : |\lambda(A_i) - \lambda_N(A_i)| < \frac{1}{k}\}$ belongs to $\mathcal{U}$ for $i = 1, 2$ and so does their intersection (since ultrafilters are closed under finite intersections). Pick $N_k \in U_{1,k} \cap U_{2,k}$. Repeating the process, we generate the desired sequence by picking $N_{k+1} > N_k$ in $U_{1,k+1} \cap U_{2,k+1}$.

From the fact that $\lambda(A) = \lim_{k\to\infty} \lambda_{N_k}(A)$, it follows that for all k large enough, $\mathcal{A}_{N_k} \in \mathcal{C}_{N_k,\beta}$. Combining (A.4) and (A.5), we obtain

$$\left| \int_A s(x)\, d\lambda - \lambda(A) \right| < \alpha.$$

The conclusion of the lemma follows since α was arbitrary. Finally, the case where $\lambda(A) = 0$ follows from the fact that the above applies to the complement of A, since $\lambda(A^c) = 1$, and the additivity of λ. *Q.E.D.*

To conclude the proof of the proposition, fix an $s \in Q$ and define

$$\gamma(A) \equiv (1+\varepsilon)\lambda(A^+) + (1-\varepsilon)\lambda(A^-).$$

As noted earlier, this is just the integral $\int_A s(x)\, d\lambda$ of the function s with respect to λ. 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 $\varepsilon \in [0, 1]$. Finally, note that $X_d \cap X_N \in \mathcal{C}_{N,1}$ for each N, so $\int s(x)\, d\lambda = 1$ and (A.3) imply $\gamma(X_d) = 1$.

That λ and γ coincide on $\mathcal{C}$ (hence necessarily on $\mathcal{C}^\star$) follows from Lemma A.3. 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 $B \equiv \{x : s(x) = 1 - \varepsilon\}$. From $\int s(x)\, d\lambda = 1$ and (A.3), we have $\lambda(B) = 0.5$, yet

$$\gamma(B) \equiv \int_B s(x)\, d\lambda = (1-\varepsilon)\lambda(B) \neq \lambda(B), \tag{A.6}$$

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 A.1. *Q.E.D.*

From Theorem 3, we know that this proof must break down somewhere if the outcome space 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 3 avoided?" For example, the construction of the perturbation s by i.i.d. sampling is not possible in an uncountable, complete, separable outcome space with countably additive probabilities. The reason is that a typical sample path s is nonmeasurable so 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 3.

PROOF OF THEOREM 4: Given a learning strategy $\{(\mathcal{C}_n, \varepsilon_n, t_n)\}_{n=1}^{\infty}$, index the events defined in (A.2) by n, writing each as Q_n to make explicit its dependence on $\mathcal{C}_n$. Consider now the event

$$\bigcap_{n=1}^{\infty} Q_n$$

and note that it must have π-probability 1. Let s be any element of this set. It is clear that the remainder of the argument in Appendix A.4.3 goes through unaltered. *Q.E.D.*

PROOF OF COROLLARY 1: From (A.6) and the fact that $\lambda(B) = 0.5$, we can write

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

so that

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

Varying ε within the interval $(0, 1]$ yields the desired conclusion. 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. *Q.E.D.*

A.5. *Proof of Theorem 1*

Writing $n = \#X_f$, the VC dimension of 2_f^X is n. The first claim follows from the fact that there is a constant K such that

$$\text{(A.7)} \qquad \sup_{P\in\mathcal{P}} P^\infty\left\{s : \sup_{A\in\mathcal{C}} |\nu^t(A, s) - P(A)| > \varepsilon\right\} < Kt^{V_\mathcal{C}} e^{-t\varepsilon^2/32}.$$

See Devroye, Gyorfi, and Lugosi (1996).[33]

For the second part, a lower bound on the amount of data needed was shown by Ehrenfeucht, Haussler, Kearns, and Valiant (1989)[34] to be

$$\text{(A.8)} \qquad t \geq \frac{V_\mathcal{C} - 1}{32\varepsilon}.$$

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

REFERENCES

AL-NAJJAR, N. I. (2007): "Finitely Additive Representation of L^p Spaces," *Journal of Mathematical Analysis and Applications*, 330, 891–899. [1353]

[33]For another take on the problem, see Pollard (1984). A characterization in terms of samples appears in Talagrand (1987).

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

AL-NAJJAR, N. I., AND M. PAI (2008): "Coarse Decision Making," Report, Northwestern University. [1346]

AL-NAJJAR, N. I., L. ANDERLINI, AND L. FELLI (2006): "Undescribable Events," *Review of Economic Studies*, 73, 849–868. [1354]

AUMANN, R. J. (1987): "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1–18. [1356]

BEWLEY, T. (1986): "Knightian Decision Theory: Part I," Discussion Paper 807, Cowles Foundation. [1355]

——— (1988): "Knightian Decision Theory and Econometric Inference," Discussion Paper 868, Cowles Foundation. [1341,1355]

BHASKARA RAO, K. P. S., AND M. BHASKARA RAO (1983): *Theory of Charges*. New York: Academic Press. [1349,1362]

BILLINGSLEY, P. (1995): *Probability and Measure* (Third Ed.). New York: Wiley-Interscience. [1342,1349]

DE FINETTI, B. (1974): *Theory of Probability*, Vols. 1 and 2. New York: Wiley. [1353]

DEVROYE, L., L. GYORFI, AND G. LUGOSI (1996): *A Probabilistic Theory of Pattern Recognition*. Berlin: Springer Verlag. [1347,1360,1362,1363,1367]

DUBINS, L. E., AND L. J. SAVAGE (1965): *How to Gamble if You Must. Inequalities for Stochastic Processes*. New York: McGraw-Hill. [1357]

EHRENFEUCHT, A., D. HAUSSLER, M. KEARNS, AND L. VALIANT (1989): "A General Lower Bound on the Number of Examples Needed for Learning," *Information and Computation*, 82, 247–261. [1367]

GAJDOS, T., T. HAYASHI, J.-M. TALLON, AND J.-C. VERGNAUD (2008): "Attitude Toward Imprecise Information," *Journal of Economic Theory*, 140, 27–65. [1356]

GILBOA, I., AND D. SCHMEIDLER (1989): "Maxmin Expected Utility With Nonunique Prior," *Journal Mathematical Economics*, 18, 141–153. [1356]

GILBOA, I., F. MACCHERONI, M. MARINACCI, AND D. SCHMEIDLER (2008): "Objective and Subjective Rationality in a Multiple Prior Model," Report, Collegio Carlo Alberto, Universita di Torino. [1355]

HORN, A., AND A. TARSKI (1948): "Measures in Boolean Algebras," *Transactions of the American Mathematical Society*, 64, 467–497. [1349]

KALAI, G. (2003): "Learnability and Rationality of Choice," *Journal of Economic Theory*, 113, 104–117. [1340]

LEHRER, E. (2005): "Partially-Specified Probabilities: Decisions and Games," Report, Tel-Aviv University. [1356]

MILLER, G. (1981): "Trends and Debates in Cognitive Psychology," *Cognition*, 10, 215–225. [1339]

POLLARD, D. (1984): *Convergence of Stochastic Processes*. Berlin: Springer Verlag. [1367]

PURVES, R. A., AND W. D. SUDDERTH (1976): "Some Finitely Additive Probability," *Annals of Probability*, 4, 259–276. [1357,1358]

——— (1983): "Finitely Additive Zero–One Laws," *Sankhyā Series A*, 45, 32–37. [1358,1360]

ROYDEN, H. L. (1968): *Real Analysis* (Second Ed.). New York: MacMillan. [1347,1360]

SALANT, Y. (2007): "On the Learnability of Majority Rule," *Journal of Economic Theory*, 135, 196–213. [1340]

SAUER, N. (1972): "On the Density of Families of Sets," *Journal of Combinatorial Theory*, 13, 145–147. [1362]

TALAGRAND, M. (1987): "The Glivenko–Cantelli Problem," *Annals of Probability*, 15, 837–870. [1367]

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*, 16, 264–280. [1343,1347]

ZHANG (1999): "Qualitative Probabilities on λ-Systems," *Mathematical Social Sciences*, 38, 11–20. [1349]

Dept. of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL 60208, U.S.A.; al-najjar@northwestern.edu.

Manuscript received October, 2007; final revision received February, 2009.