

Decision Makers as Statisticians:
*Diversity, Ambiguity and Learning**

Nabil I. Al-Najjar[†]

First draft: October 2006

This version: April 2007

Preliminary Draft

Abstract

I study individuals who use 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.

For updated versions and related work, visit:

<http://www.kellogg.northwestern.edu/faculty/alnajjar/htm/index.htm>

* I thank Lance Fortnow, Ehud Kalai, Peter Klibanoff, Nenad Kos, Mallesh Pai, and Jonathan Weinstein.

[†] Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston IL 60208.

Contents

1	Introduction	1
1.1	Uniform Learning and its Implications	2
1.2	Beliefs and Decisions	3
1.3	Robust vs. Bayesian Inference	5
2	Uniform Learning and Consistency with Empirical Evidence	7
2.1	Basic Setup	7
2.2	Motivation and Intuition	8
2.3	Uniform Learning, Statistical Models and Beliefs	11
2.4	Uniform Learning and Vapnik-Chervonenkis Theory	13
2.5	Over-fitting	15
2.6	Implications	16
2.7	Modeling Assumptions and Generalizations	17
3	Large Sample Theory	17
3.1	Exact Learning	18
3.2	Continuous outcome space	19
3.3	Discrete State Space	20
3.4	Probabilistic Closure	21
3.5	Learning Strategies	22
3.6	Discussion: <i>Bayesian Beliefs and Updating</i>	23
4	Diversity, Ambiguity and Decision Making	25
4.1	A Decision Theoretic Framework	25
4.2	Diversity of Beliefs	28
4.3	Ambiguity	30
4.4	Choice over Statistical Experiments	30
A	Proofs	31
A.1	Strategic Product Measures	31
A.2	Proof of Theorem 2: Exact Learning	31
A.3	Proof of Theorem 3: Complete Learning in continuous outcome spaces	34
A.4	Proof of Theorems 4 and 7: Failure of complete learning in discrete outcome spaces	35
A.5	Miscellaneous Proofs	40

*“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.² Major equilibrium theories in economics—be it Nash, sequential, or rational expectations equilibrium—all require beliefs to coincide with the true data generating process, reducing any disagreements in beliefs to differences in information.³ On the other hand, there is no shortage of examples in the sciences, business, politics or war 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 these phenomena, I study individuals facing the most classical of statistical learning problems, that of drawing inferences from i.i.d. data. These individuals are viewed as (classical) 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.

An essential feature of the analysis is that decision makers are assumed to be as rational as anyone can reasonably be. But no degree of rationality can eliminate the constraints inherent in statistical inference—any more than rationality can be expected to eliminate objective constraints such as lack of information. The methodology advocated in this paper is to study rational

¹G. Miller: “Trends and debates in cognitive psychology,” *Cognition*, vol. 10, pp. 215-25.

²The points made in this paragraph are not new. But being mainly part of the folklore of the literature, they are hard to trace to original references. The point on the contrast between the subjectivist and equilibrium methodologies is adapted from Hansen and Sargent (2001). An excellent recent exposition of the difficulty with the subjectivist approach is Gilboa, Postlewaite, and Schmeidler (2006).

³In games with incomplete information, this also requires the common prior assumption which dominates both theoretical and applied literatures.

individuals confronting environments in which learning is hard, rather than appeal to cognitive limitations or behavioral biases.

The remainder of the Introduction is organized in three independent subsection that can be skipped without loss of continuity.

1.1 Uniform Learning and its Implications

What makes learning hard? It is intuitive that two individuals, with common and extensive experience driving on US highways, will agree on which side of the road other drivers will use, or the likelihood that they stop at a red light. 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 other examples suggest that some learning problems are harder 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 learning from a sequence of coin flips. This is so regardless of how ‘complicated’ the event, the true distribution, or the outcome space are. Few would consider that learning from a sequence of i.i.d. coin flips hard.

Focusing on learning the probabilities of single events misses the point: decision making is about choosing from sets of feasible acts, and thus requires knowledge of the probabilities of *families of events*. The problem facing a decision maker is therefore that of using one set of data to uniformly learn a family of events, a problem that is radically different from learning ‘one event at a time.’ The theory of uniform learning is the formal framework that studies how hard learning is as a function of the complexity of the set of underlying events. This theory potentially provides a learning-based explanation for what makes some decision problems inherently more difficult than others.

The theory of uniform learning occupies a central role in modern statistics yet relatively unknown to economists. In Section 2 I use a simple model of belief formation to highlight the main findings of this theory in the case of finite data. Each individual chooses a statistical model consisting of a family of events whose probabilities are to be estimated from the data.

I identify a number of fairly direct implications of economic interest. First, beliefs are under-determined by empirical evidence: individuals may hold different models, and thus different beliefs, even though they know each

others' models and see the same data. Second, holding the statistical model fixed and increasing the number of observations, each individual becomes increasingly confident in his model and its implications. Empirical evidence by itself is not sufficient to force an individual to rethink a previously chosen model. Third, as the number of observations increases while for other events ambiguity about the true probabilities persists. On events determined by the individual's statistical model, his beliefs are essentially single-valued and, supplemented by any reasonable decision model, his choice behavior relative to these events approximate that of a standard Bayesian decision maker. Fourth, coarsening and categorization are necessary for learning. The pervasiveness of categorization in decision making is beyond dispute and does not, to me at least, 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 about fine details? If we take the view that decision makers are statisticians then categorization is necessary to draw secure inferences and avoid over-fitting the data. No appeal to computational complexity or cognitive limitations is needed.

I then turn to large sample theory, where the main formal contribution of this paper lies. The known theory of uniform learning focuses almost exclusively 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. At a basic methodological level, the notion of equilibrium itself—*e.g.*, Nash or rational expectations equilibrium—is meant to model qualitative insights about steady-state or long-run behavior. This is either done explicitly⁴ or, more often, implicitly in the use and interpretation of equilibrium notions. In practice, asymptotic models often enhance tractability and provide clearer intuition. So in Section 3 I extend the intuition of uniform learning to a steady-state model and derive some of its implications.

1.2 Beliefs and Decisions

Any individual's statistical model defines a *belief correspondence* that maps observations to the set of beliefs (probability measures) consistent with empirical evidence. With finite data, this correspondence is of course not single-

⁴As, for example, in Fudenberg and Levine (1993)'s paper on the learning foundation of Nash and self-confirming equilibria.

valued. More interesting is the fact that the measures consistent with empirical evidence will essentially agree on the probabilities of some events, but will typically wildly disagree on the probability others. I refer to the phenomenon that the data does not pin down all probabilities as *statistical ambiguity*. The qualifier ‘statistical’ is meant to distinguish it from subjectively derived attitude towards and perception of ambiguity. Statistical ambiguity simply reflects the inability to pin down the probability of some events due to the constraints imposed by statistical inference.

In economic applications, beliefs and tastes are combined to explain choice behavior. The specific way in which they are combined is of paramount importance, but orthogonal to questions like where beliefs come from or what makes them ‘reasonable.’ For this reason, the analysis reported here is not tied to any particular model of decision making (although in any given application a specific one is appealed to). All that is required is that individuals’ beliefs be consistent with empirical evidence.

In Section 4 I use a simple decision making framework to examine whether learning considerations are likely to lead rational decision makers to hold common beliefs. One of the clearest statements of one side of the argument is provided by 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:⁵

“[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 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 views, it may be a good idea to have in mind an explicit model that explains how “probabilities should be based on information.”

⁵Page numbers refer to the 1972 edition, Savage (1972).

My claim is that, when viewed as statisticians, it is perfectly natural for individuals to hold different (sets of) 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

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. It is well-known that the classic Savage (1954) framework precludes this concern for security. As a result many subsequent works attempted to capture the desire to draw secure or robust inferences. These include the classic works on ambiguity by 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 maxmin regret or other robustness criteria as found, for instance, in Manski (2004).

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?

Studying belief formation separately from decision making, as done in this paper, may seem like a serious violation of the Bayesian strictures. Historically, however, Savage seemed to have conceived his framework as normative, as a way to define rational behavior in situations involving uncertainty that is otherwise silent on the question of belief formation. Thus, he writes (1967, p. 307) that the subjectivist view of probability “seeks to distinguish between coherent behavior and blunder, or demonstrable incoherence in the face of uncertainty” and it is best thought of as a tool “by which a person can police his own potential decisions for incoherency.”

A common retort is that Bayesian theory already provides a theory of belief formation via the celebrated de Finetti theorem. The need for a separate model of belief formation is obviated, so the argument goes, by assuming a decision maker with exchangeable beliefs and who updates his prior using Bayes rule. The effectiveness of this as ‘learning’ and ‘belief formation’

procedure is a well-entrenched part of the Bayesian folklore, but it is also a myth. A classic theorem by Freedman (1965), detailed in Section 3.6, shows that Bayesian posteriors are “generically” erratic in a very strong sense whenever the outcome space is infinite.⁶ 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, largely 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.”

The desire to draw secure inferences leads individuals to statistical models that are coarser than the true model. This idea is new, the most prominent of its uses being the representation of incomplete information as partitions of the underlying state space. The same idea also appears in models of ‘bounded rationality,’ where individuals who suffer from bounded memory or bounded recall, say, are modeled as having coarse partitions of their environment. The similarity between the two uses is superficial, however. 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.

Although statistical models in this paper coarsen the outcome space, they profoundly differ from information partitions. Here individuals know each other’s information while they do not, by definition, in models of incomplete information. And the coarsening implied by the use of statistical models is not a ‘mistake,’ but a normatively acceptable response to the objective constraints of statistical inference. Statistical constraints apply to everyone regardless of intelligence, memory, or computational abilities.

⁶Freedman’s result that holds when the outcome space is the set of integers was generalized by Feldman (1991) to complete separable outcome spaces.

This paper provides a purely statistical model of complexity motivated by difficulty of learning a family of events rather than by difficulty of describing any of these event. This is in contrast to the language-based models I explored in earlier works, including Al-Najjar (1999) and Al-Najjar, Anderlini, and Felli (2006). Roughly, these papers attempt to model events that can be assigned probabilities, but that cannot be described *ex ante* relative to a given language. The learning- and language-based approaches are potentially complementary and may be related in some subtle ways.

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) . Decision making is examined in Section 4. Here we focus exclusively on beliefs, generically denoted by (lower case) $p \in \mathcal{P}$.

Three models of X, Σ, \mathcal{P} will be considered:

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} 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} the set of finitely additive probabilities.

The distinction between the second and third models will be unimportant in this section, but will be discussed thoroughly in Section 3.

To lighten 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 X from the fixed true distribution P . Formally, let S denote the set of all

infinite sequences of elements $s = (x_1, \dots)$ in X , interpreted as outcomes of infinite sampling from that space. 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.⁷

The case of finite number of observations t is modeled, as usual, by conditioning on the first t coordinates of an infinite sequence s .

2.2 Motivation and Intuition

We will ultimately be concerned with decision models for evaluating acts $f : X \rightarrow \mathcal{R}$. Section 4

Consider following stylized investment example: an investor faces “investment opportunities” randomly drawn from a set Y of such opportunities. Writing $I = \{buy, sell\}$ for the two possible decisions to be made, an investment rule is any function $f : Y \rightarrow I$ that places investment opportunities into one of two categories. For the purpose of this example, we shall assume that the space of potential investment opportunities Y is finite, but potentially very large.

An environment is a probability distribution $\bar{\mu}$ and categorization rule \bar{f} , both unknown to the decision maker. Together, $\bar{\mu}$ and \bar{f} define a probability distribution P on the outcome space $X = Y \times I$.

The problem of the decision maker is to formulate an investment plan f based on observations of past outcomes. We may also introduce specific payoff function that determines how decision maker’s rewards are determined by f and P , but this is not essential here. It suffices to assume that the decision maker’s goal is to learn P or, even more narrowly, the \bar{f} component of it since investment decisions can be made contingent on y . Learning about P is based on a sequence of observations $s^t = (x_1, \dots, x_t)$ drawn i.i.d. from P .

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

- *X is of the form $Y \times \{0, 1\}$ for some set of “instances” Y ; and*
- *The only feasible acts $f : X \rightarrow \mathcal{R}$ are ones that satisfy: For every*

⁷Most readers are likely to be familiar with these standard concept in the cases X_f and X_c . Section A.1 provides the requisite background in general enough terms as to cover the less familiar case of (X_d, Σ) .

$y \in Y$ precisely one of the two values $f(y, 1)$ and $f(y, 0)$ is equal to 1 and the other is zero.

Two key assumptions underlie the analysis:

- *Stationarity*: the decision maker faces a stationary problem (P is unchanging). Many decision problems may be usefully modeled as stationary, while some non-stationary problems become stationary in a richer outcome space. In any event, failure of learning would hold a fortiori in a non-stationary settings where the object to be learned is constantly changing.
- *Robustness*: the decision maker is agnostic about the true underlying distribution, and thus seeks distribution-free inferences. I discuss relaxations of this assumption in Section []

In our investment example, the assumption of stationarity may be interpreted as consisting of two parts: (a) the description of each investment opportunity is comprehensive so no potentially relevant factor is omitted; and (b) the economic fundamentals of what makes a company or a stock profitable are stable. An investor observes past track record of past investment opportunities and how they turned out. The assumption that this investor has no prior knowledge of P simply says that he lacks any theory (e.g., basic economics or finance) that puts a priori restrictions on the true distribution, which is a good metaphor for a technical investor.

For an arbitrary set of outcomes X and event $A \subset X$ define its empirical frequency relative to a sample s^t of length t as:

$$\nu^t(A, s^t) \equiv \frac{\#\{s^t \cap A\}}{t} \quad (1)$$

An application of your favorite version of the weak law of large numbers will ensure that $\nu^t(A, s^t)$ is a good approximation for the true probability when t is large. For example, by Chebyshev's inequality one has:

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

Stronger bounds exist, but this is not the point here. Rather, the key fact to note is that probabilities can be estimated uniformly well regardless of the event or the distribution:

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

It is irrelevant whether the event A is complicated or simple, the outcome space (X, Σ) is finite or infinite, ... etc. The inference about the probability of any single event is not more complicated than that of finding the probability of heads in independent coin flips. This, perhaps, is one reason for the commonly held intuition that “people eventually learn.”

But choice involves, almost by definition, the *evaluation of many acts simultaneously*, and consequently learning the probabilities of the family of events used in defining these acts, from one single draw of data.

To make the above intuition more precise, for each event $A \subset X$, define the set of ϵ -good samples for A as:

$$Good_{\epsilon, P}^t(A) \equiv \left\{ s : |P(A) - \nu^t(A, s^t)| < \epsilon \right\}$$

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

Inequality 2 simply asserts that

$$P^\infty [Good_{\epsilon, P}^t(A)] > 1 - \epsilon.$$

If the only bet a decision maker faces is on a single event A , we are done. But if he has to decide whether to bet on one event rather than another, the decision maker needs to be confident that the sample s^t is representative for both events simultaneously. In general, in comparing I events, A_1, \dots, A_I , the most one can say is that

$$P^\infty [\cap_i Good_{\epsilon, P}^t(A_i)] > 1 - I\epsilon. \tag{3}$$

This conclusion quickly becomes useless as the number of events I increases.

The following interpretation is perhaps revealing. The weak law of large numbers in 2 pertains to a statistical experiment in which a fresh draw of a new sample is made in testing each event. The experiment underlying 3, on the other hand, is one where the decision maker gets one shot at sampling t observations, and uses this information to evaluate all events A_i simultaneously. This formalizes the idea that data is scarce: there may be enough data to evaluate each event in isolation, but not enough to evaluate with uniform degree of confidence many events simultaneously. Data is not so abundant that one can generate as many samples as there are acts to evaluate.

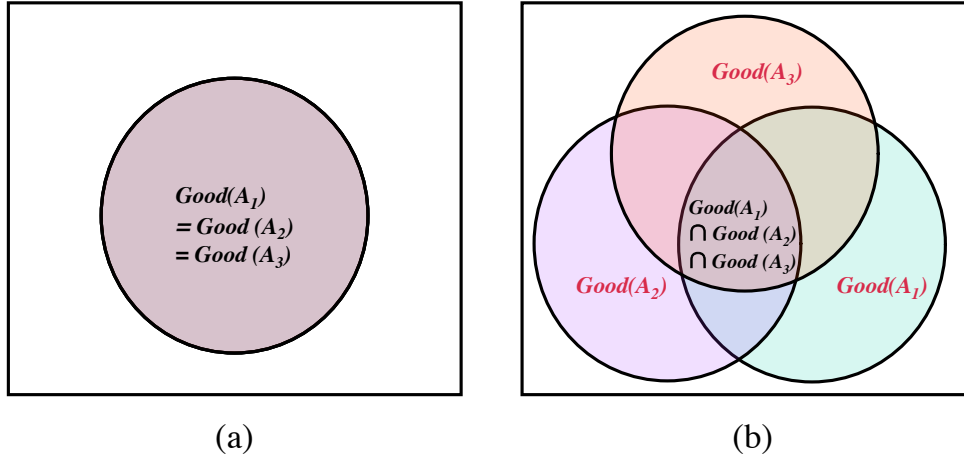


Figure 1: *Two examples of intersections of sets of samples*

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

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 between three events A_1, A_2 , and A_3 in the outcome space X . In Figure 1(a) the events $Good_{\epsilon, \mathcal{P}}^t(A_i)$, $i = 1, 2, 3$, 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: each event A_i gives rise to a set of representative samples $Good_{\epsilon, \mathcal{P}}^t(A_i)$ which, by 2, has probability at least $1 - \epsilon$. The problem is that these sets of samples stack up in such a way their intersection has probability of only $1 - 3\epsilon$.

What determines whether a learning problem falls into type (a) or (b)? The answer is contained in the beautiful and powerful theory of Vapnik and Chervonenkis (1971) (translated from an earlier paper in Russian).⁸

2.3 Uniform Learning, Statistical Models and Beliefs

The idea that data is scarce is made formal in the following definition:

⁸For textbook expositions of this theory, see Vapnik (1998) or Devroye, Györfi, and Lugosi (1996).

Definition 2 (Uniform Learnability) *A family of events $\mathcal{C} \subset \Sigma$ in an outcome space X is ϵ -uniformly learnable by data of size t if,*

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

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

The crucial aspect of the definition is the location of the quantifier “sup.”
 $\sup_{A \in \mathcal{C}}$
The probability being evaluated in 4 is that of samples in which *all events in \mathcal{C}* have their probabilities close to the empirical frequencies computed from the sample. This is to be contrasted with the weaker statement:

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

As noted earlier, this corresponds to the experiment where one gets a fresh sample for each event evaluated, while 4 corresponds to an experiment where all inference must be drawn from one single sample.

Uniform learnability by itself is a hopelessly weak a criterion, since it is trivially satisfied by taking \mathcal{C} to be empty or containing a few events.⁹ It is normatively sensible to require:

Definition 3 (Maximality) *A set of events \mathcal{C} is maximal ϵ -uniformly learnable by data of size t if there is no $\mathcal{C}' \supsetneq \mathcal{C}$ for which 4 holds.*

This says that \mathcal{C} does not leave out any event A that, given the same amount of data t , can be added without undermining the desired level of confidence in $\{A \cup \mathcal{C}\}$. Normatively, this is a weak form of monotonicity on a decision maker’s preferences over experiments, requiring him not to be wasteful in his use of data: If given \mathcal{C} and sample size t , the decision maker can also learn A “for free,” then maximality says he should.

Most of the results do not depend on maximality. Its main role is rather normative as it corresponds to the monotonicity of the preference over statistical experiments.

For a finite outcome space X_f , given a class of events \mathcal{C} , existence of a family of events $\mathcal{C}' \supseteq \mathcal{C}$ with the same VC-dimension is immediate, since

⁹In 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.

there is only finitely many events. Existence of such collections is more delicate on infinite spaces where it is at least conceivable that one can indefinitely enrich a class of events \mathcal{C} ϵ -uniformly learnable by data of size t . The following lemma resolves this issue:

Proposition 1 *Fix t and suppose that \mathcal{C} is a class of events that is ϵ -uniformly learnable by data of size t . Then there is a maximal class $\mathcal{C}' \supset \mathcal{C}$ with the same property.*

Definition 4 (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 .*

We will view the amount of data t as an objective constraint on the set of statistical models the decision maker can choose from. By contrast, the events \mathcal{C} the decision maker expresses interest in testing and the desired degree of confidence ϵ are both objects of choice.¹⁰

An agent behaving as classical statistician formulates a statistical model $(\mathcal{C}, \epsilon, t)$. Corresponding to such model is the set of distributions *consistent with empirical evidence*:

$$\mu^t(s) = \left\{ p \in \mathcal{P} : \sup_{A \in \mathcal{C}} |p(A) - \nu^t(s)(A)| \leq \epsilon \right\}.^{11} \quad (5)$$

This is set of all probability measures that are close to the empirical frequencies on events in \mathcal{C} . We shall interpret $\mu^t(s)$ as the set of probability measures the decision maker views as “consistent with empirical evidence” given his statistical model.

2.4 Uniform Learning and Vapnik-Chervonenkis Theory

Uniformly learnable classes have an elegant characterization using the theory of Vapnik-Chervonenkis. The fundamental concept 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

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

¹⁰We will limit ourselves to an ϵ that is fixed across events. The model can be modified to incorporate the general case where $\epsilon : \mathcal{C} \rightarrow [0, 1]$.

¹¹We suppress the dependence of μ^t on \mathcal{C} for notational convenience.

Here, $\#\{C \cap \{x_1, \dots, x_m\} : C \in \mathcal{C}\}$ is the number of subsets of the sample that can be obtained by intersecting the sample with some event in \mathcal{C} . We may interpret $s(\mathcal{C}, m)$ as a measure of the complexity of 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,

$$\sup_{P \in \mathcal{P}} P^\infty \left\{ s : \sup_{A \in \mathcal{C}} |\nu^t(s)(A) - P(A)| > \epsilon \right\} < K t^{V_{\mathcal{C}}} e^{-t\epsilon^2/32}, \quad (6)$$

where K is some constant. Tighter bounds are available, perhaps under mild assumptions; the above version is most useful for our purposes. See Devroye, Györfi, 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).

The inequality above provides an upper bound on the amount of data needed to ϵ -uniformly learn the probabilities of a class of events. The following is a lower bound on the amount of data needed:

$$t \geq \frac{V_{\mathcal{C}} - 1}{32\epsilon} \quad (7)$$

This was shown by Ehrenfeucht, Haussler, Kearns, and Valiant (1989) (see also Devroye, Györfi, and Lugosi (1996, Section 14.5)).

The following is perhaps the best known, and first discovered, class of finite VC-dimension. It is the class of events appearing in well-known Glivenko-Cantelli Theorem and it vividly illustrates the theory:

Example 1 *Let $X = [0, 1]$ and \mathcal{C} to 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 with $V_{\mathcal{C}} = 2$.¹²*

¹²The class \mathcal{C} has finite VC-dimension but it is clearly not maximal; for instance, one may add uncountably many new events, with r irrational, without affecting its VC-dimension. Lemma 1 ensures that \mathcal{C} can be enlarged to a maximal class of the same dimension, and the new class is therefore uniformly learnable. The fact that \mathcal{C} is not necessarily maximal illustrate the point even more sharply.

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 $V_{\mathcal{C}} = 2$, 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 $V_{\mathcal{C}} = 2$.

The class \mathcal{C} is simple from a learning stand point even though it contains an uncountable collection of events. On the other hand, for any integer v , there is a *finite* collection of events with VC-dimension v . This shows that “complexity” from a learning stand point has little to do with cardinality per se.

Note also that \mathcal{C} is not an algebra. The following example illustrates that closing \mathcal{C} by taking the algebra it generates is not innocuous as far as learning is concerned.

Example 2 $X = [0, 1]$ with \mathcal{C}' the algebra generated by \mathcal{C} . Then $V_{\mathcal{C}'} = \infty$. In particular, if \mathcal{B} is the set of Borel subsets of X , then $V_{\mathcal{B}} = \infty$.

To verify the first claim, note that any element of any sample $x_1 < \dots < x_m$ of finite size can be obtained by intersecting with a set of the form $(r, r']$, and all such sets belong the algebra generated by \mathcal{C} . Since the algebra is closed under finite unions, we can obtain any subset of the sample by intersecting with the appropriate set in \mathcal{C}' . Finally, passing to a larger family of events (the Borel sets) can only increase the shattering capacity of the original family.

2.5 Over-fitting

One interpretation of uniform learning is as a criterion to avoid over-fitting the data. To illustrate this, consider the categorization problem introduced earlier. Let \mathcal{F} be a class of categorization rules $f : Y \rightarrow I$ and suppose that $(\bar{\mu}, \bar{f})$ is the true environment. Assume here that Y is infinite. A decision maker wants to find the best approximation of the true categorization \bar{f} in the sense of, say, minimizing

$$\min_{f \in \mathcal{F}} \sum_y \bar{\mu}(y) m(f, \bar{f})(y),$$

where $m(f, \bar{f})(y)$ is 1 if f and \bar{f} match at y and 0 otherwise.

A decision maker who chooses a very rich \mathcal{F} (e.g., the set of all functions $f : Y \rightarrow I$) will be able to fit any sample of data x_1, \dots, x_t . This decision maker can rationalize everything but learns nothing.

This is obvious if X is infinite, so one natural response is that, realistically, X is finite in which case only always learns in the limit as data accumulates. This, however, takes a rather liberal attitude towards data constraints. Consider 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 X_f is $2^{z_1+z_2+z_3}$. For entirely conservative value of, say, $z_1+z_2+z_3 = 50$, the cardinality of the set of events is so astoundingly large that complete learning, while guaranteed in the limit, is closer to a fantasy than to the even idealized setting of rational economic decision making.

The concept of the VC-dimension of a set of categorization rules \mathcal{F} ¹³ captures this over-fitting problem. If one views \mathcal{F} as a set of hypotheses representing a decision maker's theory of potential explanations of the data, then 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).

From the perspective of decision making, a rational decision maker with too rich a class of hypotheses \mathcal{F} over-fits the data, a problem he should recognize and guard against. In this sense, uniform learnability provides a normative criterion that a rational individual concerned about learning should adhere to. In the diet example, with a fixed amount of data t , a rational decision maker must choose a theory that necessarily restrict what explanations he can draw inferences on.

2.6 Implications

Incomplete.

- *Indeterminacy of beliefs and statistical ambiguity*

¹³This is defined as the VC-dimension of the family of sets $\{(y, i) : f(y) = i\}$ as y and f range over Y and \mathcal{F} , respectively.

- *Increasing confidence in the ‘wrong’ model*
- *Distinct models consistent with empirical evidence*
- *Coarsening and categorization*

2.7 Modeling Assumptions and Generalizations

Readers who deem irrational decision makers acting as classical statisticians will have a hard time not just with this paper, but with current statistical practice, which is overwhelmingly classical rather than Bayesian. See, for instance, Efron ((2005) and (1986)).

non-uniform confidence function

Robustness: This can be defended on many grounds, including that the decision maker may doubt that he has the right model. The model can accommodate the introduction of prior knowledge that narrows down the set of possible distributions. But one must then ask where these restrictions come from. One way to think of the role of this assumption is that it enables us to delineate the boundary between empirically-grounded and extra-factual sources of knowledge.

3 Large Sample Theory

Here we study the asymptotic properties of uniform learning as the amount of data available to the decision maker becomes large. Such setting carries a number of important advantages:

- *Robustness:* The asymptotic properties of the finite-data model are not entirely clear or even convincing: Given a conclusion for particular value of t , one would wonder what happens if there is a little more data?
- *Tractability:* An infinite model can be considerably simpler and yields sharper intuitions. For instance, many results in Statistics and their applications (the law of large numbers, de Finetti’s theorem etc) are immensely simpler and more transparent in a context of infinite samples.

As we have seen in the last section, fixing a finite outcome space X_f and taking t to infinity always leads to full learning, but this masks the

difficulties arising in the uniform learning of the probability of a family of events. In conducting asymptotic analysis, we therefore need to consider the case of an infinite outcome space.

3.1 Exact Learning

Given a *uniformly learnable* family of events \mathcal{C} , the corresponding set of distributions *consistent with empirical evidence* (given \mathcal{C}) is:

$$\mu(s) \equiv \left\{ p : \forall \epsilon > 0 \limsup_{t \rightarrow \infty} \sup_{A \in \mathcal{C}} |p(A) - \nu^t(s)(A)| \leq \epsilon \right\}.^{14}$$

Our intuition would be that, on a typical sample, any $p \in \mu(s)$ should assign to any event A a probability close to the true probability $P(A)$. Of course, on an arbitrary sample s , the probability measures in $\mu(s)$ can be just about anything. The following theorem shows that $\mu(s)$ has a very clean structure on most samples:

Theorem 2 (Exact Learning) *For any uniformly learnable \mathcal{C} and $P \in \mathcal{P}$*

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

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

That is, almost surely, beliefs consistent with empirical evidence in the limit are precisely those coinciding with the true probability law P on all the events in \mathcal{C} .

Among other things, this theorem makes it possible to have a much cleaner definition of what it means for empirical evidence to determine beliefs:

Definition 5 *Beliefs are asymptotically determinate if there is a uniformly learnable \mathcal{C} such that for every P ,*

$$\mu(s) = P, \quad P^\infty - a.s.$$

That is, in the limit, almost surely, the only belief consistent with empirical evidence is the true distribution. It is immediate, and hardly surprising, that beliefs are asymptotically determinate in the finite outcome space X_f since there are finitely many events and unbounded amount of data. Clearly, to be interesting, a model with asymptotically infinite data should be one with an infinite number of outcomes. This is what we turn to next.

¹⁴Again, we suppress the dependence of μ on \mathcal{C} for notational convenience.

3.2 Continuous outcome space

An obvious starting point is to take the outcome space X_c to be 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 3 *Beliefs are asymptotically determinate if X is a complete metric space.*

To get the intuition underlying the theorem, consider Example 1 and note that there are two distinct principles in play. First, the classic Glivenko-Cantelli Theorem asserts that the empirical distribution functions converge to the distribution function uniformly almost surely. In our context, this is just saying that empirical frequencies converge uniformly over \mathcal{C} . Second, any countably additive probability measure on the half intervals has a unique extension to all Borel sets. The first principle is statistical, asserting that a class of events is uniformly learnable, while the second has to do with the potential of knowledge of the probabilities of events in a class \mathcal{C} to pin down the probabilities of all events.

The remarkable aspect of the example is that complete learning obtains using an exceedingly simple class of events. Similarly, the class of events used in the theorem also has a simple structure, similar to that of the half intervals on $[0,1]$.¹⁵ It seems difficult to think of bounded rationality reasons that would prevent a decision maker from using such simple learning procedure.¹⁶

The example is disturbing in another way, namely that it reveals a rather sharp disconnect with the finite outcome space/finite data model. There, 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 3. In my view, 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.

¹⁵By simple we mean having a low VC-dimension.

¹⁶The choice of rational r 's in Example 1 is made to further emphasize that a countable family of events suffices to learn everything.

3.3 Discrete State Space

X_d is a countable set of outcomes; Σ is the set of all subsets of X ; and \mathcal{P} the set of all finitely additive probability measures on Σ .

Theorem 4 *Beliefs in the discrete outcome space (X_d, Σ) are not asymptotically determinate.*

That is, for any uniformly learnable family \mathcal{C} , $\mu(s)$ almost surely contains distinct probability measures that agree on \mathcal{C} , but disagree on some events outside \mathcal{C} .

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 5 *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} and uncountably many events B such that $|\lambda(B) - \gamma(B)| = \alpha$.*

To interpret these results, recall the discussion following the Example 1, namely that there are two principles used in deducing probabilities. First, the statistical principle that allows to deduce probabilities from data, by virtue of Theorem 2, is in effect here just like it was in the case of continuous outcome spaces. The difference has do to with the second principle that uses the axioms of probability to extend the probabilities of events in \mathcal{C} to events outside \mathcal{C} . It is the second principle that fails here. Very roughly, the proof relies on a well-known combinatorial result asserting that, for a class of subsets of a given finite set to be uniformly learnable, its cardinality must be much smaller than the powerset. Passing to the limit (which is neither obvious nor technically simple), the intuition is that the union of \mathcal{C}_t 's, \mathcal{C} , will be a sparse, leaving out many sets. This is then used to conclude that knowledge of the probabilities of events in \mathcal{C} (which is guaranteed by Theorem 2) is not enough to pin down the probability of all events in 2^{X_d} .

The delicate part of the argument has to do with understanding and qualifying what “sparse” means. One may be naively tempted to identify this with cardinality of the families of events. This, however, is a complete non-starter: In Example 1 knowledge of the probabilities of a countable family of events is sufficient to determine the probabilities of all Borel sets (which is uncountable), while in X_d there exist uniformly learnable \mathcal{C}_t 's that are uncountable. The correct notion of “sparseness” used in the arguments is therefore much more subtle. The next section sheds some light on that.

3.4 Probabilistic Closure

A uniformly learnable family \mathcal{C} need not have any particular structure (other than uniform learnability and maximality). In particular, it needs not be closed under any of the set theoretic operations (complements, intersections, ... etc). On the other hand, two probability measures $p, p' \in \mu(s)$ will generally agree on more than just \mathcal{C} . Beliefs, being probability measures, are additive so, for instance, knowledge of the probabilities of disjoint events $A, B \in \mathcal{C}$ completely determines the probability of $A \cup B$ for any probability measure in $\mu(s)$.

To make this formal, call a function $p : \mathcal{C} \rightarrow [0, 1]$ a *partial probability* if it is the restriction to \mathcal{C} of some probability measure p' on Σ . 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).

Definition 6 *An event $A \in \Sigma$ has unambiguous probability given \mathcal{C} if, for any partial probability measure p on \mathcal{C} , and any two extensions p', p'' of p to Σ , $p'(A) = p''(A)$.*

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

The question now is: *What can be said about the structure of \mathcal{C}^* ?* To motivate our answer, I first note 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) who use this idea to provide a behavioral definition of (un)ambiguous events. Formally, a λ -system is a family of events closed under complements and *disjoint* unions, but not necessarily arbitrary unions or intersections (Billingsley (1995)).

Theorem 6 *Fix any any uniformly learnable family \mathcal{C} :*

1. \mathcal{C}^* is a λ -system;
2. \mathcal{C}^* may be strictly larger than the smallest λ -system containing \mathcal{C} ;
3. \mathcal{C}^* need not be an algebra.

Part (1) is evident. Part (2) is known; an example in de Finetti (1974) illustrates the point. Here is a version of his example: 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, $\mathcal{C} \neq \mathcal{C}^*$: 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}^* . On the other hand, it is easy to see that $\mathcal{C}^* \neq 2^X$.¹⁷ Finally, these arguments show that the set $\{1, 2, 3\} \cap \{1, 3, 5\} = \{1\} \notin \mathcal{C}^*$, establishing part (3).¹⁸

Letting $\lambda(\mathcal{C})$ and $\Sigma_0(\mathcal{C})$ denote, respectively, the λ -system and algebra generated by an arbitrary family of events \mathcal{C} , we conclude that, in general,

$$\mathcal{C} \subsetneq \lambda(\mathcal{C}) \subsetneq \mathcal{C}^* \subsetneq \Sigma_0(\mathcal{C}).¹⁹$$

3.5 Learning Strategies

In Theorem 4 additional data refines beliefs only by tightening the bounds on the errors in estimating the probabilities of events in \mathcal{C} . A natural suggestion is that the additional data is used, in addition, to enlarge the set of events \mathcal{C} itself. This leads to the following definition:

Definition 7 *A sequence $\{(\mathcal{C}_t, \epsilon_t)\}_{t=1}^\infty$ is a learning strategy if $\epsilon_t \rightarrow 0$, $\mathcal{C}_t \subseteq \mathcal{C}_{t+1}$ for every t , and \mathcal{C}_t is a maximal ϵ_t -uniformly learnable by data of size t .*

A learning strategy is an idealized model of a decision maker who uses larger samples to both enrich the set of events and decrease ϵ . At this point

¹⁷Fix 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 $\{1\} \notin \mathcal{C}^*$.

¹⁸Note that, as an illustration of part (3), this is not a very convincing example when learning is introduced. Here the only maximally learnable family of events is the power set and so all events are eventually perfectly learned as data accumulates. I suspect that a better example, using the discrete model, is possible, but this has not been worked out for this version.

¹⁹The example establishes that the last two inclusions can be strict.

a reader imbued with the Bayesian paradigm should be cautioned not to misperceive a learning strategy as some sort of dynamic updating process: We do *not* have in mind a situation where the decision maker first gets a sample of size t , then subsequently updates his beliefs as a new $t + 1$ st data point is added. Rather, we are thinking of *distinct and independent statistical experiments* $(\mathcal{C}_t, \epsilon_t)$, $t = 1, 2 \dots$. Our goal is to analyze the limiting behavior of these experiments.

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

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

The next theorem shows that Theorem 4 is robust to allowing learning strategies (rather than just a fixed \mathcal{C}):

Theorem 7 *For any learning strategy, $\mu(s)$ almost surely contains distinct probability measures that agree on \mathcal{C} , but disagree on some events outside \mathcal{C} .*

3.6 Discussion: *Bayesian Beliefs and Updating*

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 Bayesianism and its detractors is well beyond the scope of this paper.²⁰ 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 has a belief on the state space S , the space of all infinite sequences of such draws. If this belief is exchangeable, meaning that he regards the outcomes at each stage as symmetric, then by the celebrated de Finetti Theorem²¹ this belief can be represented as a two-stage lottery where the first stage consists of a probability measure ν on the set of probability distributions P on X ,

²⁰See, 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." In his presidential address to the American Statistical Association, Efron (2005) he advocates a combination of frequentist and Bayesian ideas.

²¹The classic reference is Hewitt and Savage (1955) which generalizes de Finetti's result.

denoted $\Delta(X)$, and in the second outcomes are generated i.i.d. with respect to P . In words, an exchangeable belief on a stochastic process is equivalent to a belief that the process is i.i.d. with unknown parameter P .

De Finetti's theorem is not a theory of learning, but an elegant and insightful representation of beliefs on symmetric experiments. Granted the de Finetti representation, one can ask whether a decision maker with prior belief ν on $\Delta(X)$ and who observes an infinite sample s learns the true distribution in the limit. Specifically, let $\nu^t(s)$ denote the posterior computed after observing the first t elements of the infinite sample s . The learning question is now whether, given a true distribution P , the posteriors converge to put unit mass on P .

A well known result, is that learning obtains if the true distribution P is drawn at random according to a probability measure $\hat{\nu}$ on $\Delta(X)$ that is mutually absolutely continuous with respect to ν .²²

Suppose that X is a complete separable metric space and endow both $\Delta(X)$ and $\Delta(\Delta(X))$ with the weak topology, and $\Delta(X) \times \Delta(\Delta(X))$ with the product topology. Interpret a typical element (P, ν) of this space as a true distribution P and a Bayesian belief ν on the set of distributions.

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 $\Delta(\Delta(X))$ infinitely often P -a.s.

This result was first shown when X is the set of integers by Freedman (1965) and later generalized to arbitrary complete separable metric spaces by Feldman (1991). The notion of genericity here is that of a residual set.²³ As 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 Q and 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.

²²See Feldman (1991) for references. The proofs of this result typically rely on the fact that the sequence of posteriors form a martingale under P .

²³*i.e.*, the complement of a countable union of closed and nowhere dense sets.

They suggest that building a compelling normative case for Savage-style 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}^* are Bayesian, although they are not arrived at in a Bayesian fashion.

Consider next the frequentist belief formation. By this I mean that for every event A and any sample s , beliefs are set equal to the empirical measure:

$$\mu^t(s)(A) = \nu^t(s)(A), \quad \forall A \text{ and } s.$$

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. On the other hand, a frequentist will put unit mass on the past observations, and so, equivalently, rule out as impossible those states 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. The notion of consistency with empirical evidence allows the empirical measure (the frequentist's beliefs), but also includes all those measures that cannot be incontrovertibly ruled out by evidence.

4 Diversity, Ambiguity and Decision Making

4.1 A Decision Theoretic Framework

Learning leads to a set of beliefs consistent with empirical evidence, here described as the compact convex set of probability measures $\mu^t(s)$ and $\mu(s)$. How these sets of beliefs are combined with tastes to produce choice requires an explicit decision theoretic framework. The learning model of this paper is not tied to any particular decision framework, the idea being that belief formation should be orthogonal to how these beliefs translate into actions.

Here I outline a particularly simple framework that is general enough for the points I want to make. The framework is that of Gajdos, Hayashi, Tallon, and Vergnaud (2006) which provides an axiomatic treatment of how to incorporate objective information into a subjective setting. The main innovation here is a specific source of objective information, namely statistical

inference from repeated sampling.

I directly use their functional form; the reader should refer to their paper for the underlying axioms and their motivation. Fix a finite set of consequences Z , and consider acts of the form $f : X \rightarrow \Delta(Z)$, and let \mathcal{K} be the set of all compact and convex subsets of \mathcal{P} . If decision maker is supplied with objective information that the true distribution lies in some set $\mathcal{P}' \subset \mathcal{P}$, he evaluates an act f according to:

$$U(f) = \min_{P \in \varphi(\mathcal{P}')} \int_X u \circ f dP \quad (8)$$

where u is a vNM utility function and $\varphi : \mathcal{K} \rightarrow \mathcal{K}$ is a function, independent of f , that maps objective information \mathcal{P}' to a subjective set of measures $\varphi(\mathcal{P}')$.

We shall assume that φ satisfies:

1. *Inclusion*: $\varphi(K) \subset K$ for every $K \in \mathcal{K}$; and
2. *Mixture linearity*: For any lottery that yields $K_1 \in \mathcal{K}$ with probability ϵ and $K_2 \in \mathcal{K}$ otherwise,

$$\varphi(\epsilon K_1 + (1 - \epsilon)K_2) = \epsilon\varphi(K_1) + (1 - \epsilon)\varphi(K_2).$$

Gajdos, Hayashi, Tallon, and Vergnaud (2006) identify behavioral conditions extending those of Gilboa and Schmeidler (1989) guaranteeing the existence of a unique function φ satisfying inclusion and mixture linearity. Unfortunately, their setup includes assumptions of technical nature that preclude its direct application to our problem, namely that X must be countable and \mathcal{P} consists of measures of finite support. Verifying whether their representation still holds with these technical assumptions removed is to be undertaken in a future revision of this paper.

This decision model can be motivated as follows. The decision maker believes he is facing a malevolent Nature that “chooses” the worst probability distribution in \mathcal{P}' . This decision maker is not so paranoid to think that Nature can choose any probability measure in \mathcal{P} . This extreme paranoia seems both descriptively implausible and normatively unwarranted. The idea, then, is that he has objective information in the form of a set of measures $\mathcal{P}' \subseteq \mathcal{P}$ that restricts Nature’s choice. Inclusion simply says that the decision maker is not so paranoid to doubt the information represented by \mathcal{P}' . Mixture linearity says that he satisfies a substitution axiom with respect to information structures.

4.1.1 Decision Framework: Infinite Samples

In our setting, information is supplied by learning: Nature “locks” its choice of the true P , reflecting a fundamental belief that the environment is stable, and the decision maker is supplied with i.i.d. draws from that P . In the case of infinite samples, this information takes the form of a set of measures $\mu(s)$.

In the case of infinite data, a sample s is drawn, determining a set of probability measures $\mu(s)$. The decision maker then evaluates acts according to:

$$U(f; s) = \min_{P \in \varphi_s(\mu(s))} \int_X u \circ f dP.$$

Note that we assume that the taste component of preferences u is independent of the sample s . The sample only provides information used in the formation of beliefs and there is no reason why it should impact taste over consequences. On the other hand, we allow φ_s to vary with s . While the restriction to $\mu(s) \subseteq \mathcal{P}$ is objective, the further restriction to $\varphi_s(\mu(s))$ may reflect subjective elements of how the decision maker interprets ambiguous objective information. When empirical evidence is not sufficient to reduce $\mu(s)$ to a singleton set, the decision maker’s “inferences” are subjective and 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.

4.1.2 Decision Framework: Finite Samples

Under a statistical model $(\mathcal{C}, \epsilon, t)$, the decision maker faces an objective lottery on sets of probability measures:

1. With probability $1 - \epsilon$, the sample s is representative, in the sense that $\sup_{A \in \mathcal{C}} |P(A) - \nu^t(s)(A)| < \epsilon$, in which case P is guaranteed to belong to $\mu^t(s)$.
2. With probability ϵ , the sample s may have little to do with the true P , in which case the entire set of probability measures \mathcal{P} is possible.²⁴

²⁴This glosses over the following small issue, namely observing the first t elements of the sample rules out those P 's that assign zero probability to the sample. The set of probability measures that assign positive probability to the finite sample is open; taking \mathcal{P} here amounts to taking the closure of that set.

By mixture linearity, the decision maker treats this lottery as equivalent to the convex set of probability measures:

$$(1 - \epsilon)\mu^t(s) + \epsilon\mathcal{P}$$

His evaluation of acts is defined as:

$$U(f; s^t) = \min_{P \in \varphi_s[(1-\epsilon)\mu^t(s) + \epsilon\mathcal{P}]} \int_X u \circ f dP.$$

4.2 Diversity of Beliefs

Should individuals who have observed a large, common pool of data hold the same beliefs? To make this precise, assume that there are two decision makers, $i = 1, 2$ who:

1. Face the same unknown environment P ;
2. Observe the same data s ;
3. Have identical vNM utility u .

Let φ_s^i denote decision maker i 's subjective transformation of objective information, and let \mathcal{C}^i denote the statistical model of individual i and μ^i the corresponding set of distributions consistent with empirical evidence.²⁵

Assume that φ_s^i is singleton valued for every s , so each decision maker is subjective utility maximizer. Such decision makers are confidently picking a probability distribution $p_s^i \equiv \varphi_s(\mu^i(s))$ as function of the sample. How far does data restrict their choice? The inclusion requirement says that they cannot be completely delusional; each in fact learns the probability of a maximal class of events \mathcal{C}^{*i} .

This, however, exhausts the statistical evidence available to them, so beyond these classes of events, “they are on their own.” As I argued earlier, there is no reason why φ_s^i should not vary with s given i . A decision maker may use heuristics or fall for biases in reading the “tea leaves” to figure out how to assign probabilities to events outside \mathcal{C}^{*i} . Even less compelling is to assume that the individuals subjective transformations should agree.

²⁵ There are, of course, *practical* reasons to assume that individuals hold common beliefs, including the tractability and the discipline it affords. This paper has little to say about these justifications.

Disagreement can appear at two levels. First, assume that the two individuals use the same statistical model $\mathcal{C} = \mathcal{C}^i, i = 1, 2$. Let us examine in turn the 4 models considered in this paper:

Model A-Finite outcome space X_f and infinite data: In this case, beliefs are asymptotically determinate, so $p_s^1 = p_s^2 = \mu(s)$ almost surely. Complete learning and lack of disagreement are guaranteed in the limit, almost surely.

Model B-Continuous outcome space X_c and infinite data: Beliefs are asymptotically determinate by Theorem 3. Again, complete learning and lack of disagreement are guaranteed in the limit, almost surely.

The last two cases should be contrasted with, arguably, the most important case in practice:

Model C-Finite outcome space X_f and large finite data: In this case, whether approximate complete learning occurs and whether disagreements are small depends (as it should) on the cardinality of X_f and the amount of data t .

For example, if we are learning about the probability of a coin ($n = 2$) then one should expect approximate learning of the true probability with enough data. The analysis of this paper introduced a new limiting model that better captures the spirit of Model C in a way that the previous two do not:

Model D-Discrete outcome space X_d and infinite data: By Theorem 4 complete learning fails.

In Model D, two decision makers facing the same environment, have access to the same evidence and use identical statistical models will have their evaluation of events in \mathcal{C}^* completely pinned down by data. But empirical evidence leaves room for disagreement on the probabilities of events outside \mathcal{C}^* . This is reflected in the subjective component of preferences, φ_s^i that has, in principle, nothing to do with empirical evidence.

Finally a more radical source of disagreement is difference in the statistical models used by the decision makers. If $\mathcal{C}^i \neq \mathcal{C}^j$ there may be no agreement on any events.

4.3 Ambiguity

There is by now 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 is too vast to even attempt a cursory review of it here.

A key issue is whether learning eliminates ambiguity in the long run. To see how our model sheds light on this, consider a decision maker with maxmin expected utility preferences. This decision maker translates his inability to pin down the probability of some events by taking a pessimistic attitude, in the sense exhibited in 8.

In this settings, the conclusions are similar to those in the last subsection: In Models A and B one should expect all ambiguity to be eliminated in the limit. Empirical evidence simply swamps any initial uncertainty about the probabilities. On the other hand, in a rich finite outcome space and finite data, as in Model C, the elimination of ambiguity depends on the relationship between the richness of the outcome space and the amount of data available. Model D is the limiting case in which ambiguity does not disappear even in the limit when the outcome space is rich enough.

To my knowledge the only model incorporating learning in a model of ambiguity is Epstein and Schneider (2005). They show that ambiguity about the transition function of the stochastic process generating the data ensures that ambiguity does not disappear in the limit. However, their setting is sufficiently different that a direct comparison is rather difficult.

4.4 Choice over Statistical Experiments

Incomplete

A Proofs

A.1 Strategic Product Measures

Data is generated according to random draws from X . Formally, let $\mathcal{X} = X \times X \times \dots$ be the infinite product of X , and let \mathcal{B} denote the Borel σ -algebra on \mathcal{X} , where each coordinate is given the discrete topology.

Suppose we are given a finitely additive probability measure λ on X . We are interested in defining the product measure λ^∞ on \mathcal{X} uniquely so that natural disintegration operations continue to hold. If λ were countably additive, this is standard result. For finitely additive probabilities, Dubins and Savage (1965) introduced the concept of strategic products that generalizes the usual construction of product measures. In a classic paper, Purves and Sudderth (1976) showed quite generally that the Dubins and Savage procedure gives rise to a unique probability measure on the $(\mathcal{X}, \mathcal{B})$. Here, \mathcal{B} is constructed in the usual way, as the σ -algebra generated by finite cylinder sets.

With Purves and Sudderth (1976)'s result, many of the results found in the countably additive setting extend to finitely additive probabilities immediately or with little effort. These include the Borel-Cantelli lemma, the law of large numbers, the Glivenko-Cantelli theorem and the Kolmogorov 0-1 law. These results are assumed throughout the proofs below.

A.2 Proof of Theorem 2: Exact Learning

First we need the following lemma:

Lemma A.1 *Fix any (\mathcal{C}, ϵ) , we have:*

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

Proof: From 6 we have that for every $P \in \mathcal{P}$ and $\alpha > 0$

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

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

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

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

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

is a tail event. By the results of Purves and Sudderth (1983), P^∞ is countably additive on tail events. This implies:

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

hence:

$$P^\infty \left\{ s : \limsup_{t \rightarrow \infty} \sup_{A \in \mathcal{C}} |\nu^t(s)(A) - P(A)| = 0 \right\} = 1. \quad \blacksquare$$

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

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

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

$$\left\{ s : \limsup_{t \rightarrow \infty} \sup_{A \in \mathcal{C}} |\nu^t(s)(A) - P(A)| = 0 \right\} \quad (9)$$

has P^∞ -probability 1. Being in the event in 9 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 $\sup_{A \in \mathcal{C}} |p(A) - \nu^t(s)(A)| < \epsilon$ for all large enough t . Thus, for all large t , we have:

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

Since ϵ' was arbitrary, we conclude

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

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

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

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

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

Turning to the proof of Theorem 2, we prove a slightly stronger claim, covering the case of learning strategies (this is needed for Theorem 7 to make sense):

Theorem 2* *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 $P \in \mathcal{P}$*

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

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

Proof: We first note that:

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

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 the results of Purves and Sudderth (1983), P^∞ is countably additive on tail events. This implies

$$P^\infty \left\{ \bigcap_{t'=1,2,\dots} \bigcap_{\epsilon>0} \left\{ s : \mathcal{M}(\mathcal{C}_{t'}, \epsilon, s) = \left\{ p : \sup_{A \in \mathcal{C}_{t'}} |p(A) - P(A)| \leq \epsilon \right\} \right\} \right\} = 1. \text{ }^{26}$$

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 3: Complete Learning in continuous outcome spaces

This is essentially a consequence of the facts that: (1) all complete separable metric spaces are “equivalent” to the interval $[0,1]$; 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, \mathcal{B}) is Borel equivalent to a Borel subset of $[0,1]$.²⁷ That is, there is a Borel subset $B \subset [0,1]$ and a measurable bijection $\phi : X \rightarrow B$ 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 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 1.²⁸ From this it follows that for every (countably additive) probability distribution P :

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

²⁶These are countable intersections, which can be achieved by taking a sequence $\epsilon_n \downarrow 0$ if necessary.

²⁷ $[0,1]$ will always be understood as having the Borel sets as measurable structure.

²⁸See Problem 13.15 of Devroye, Györfi, and Lugosi (1996, p. 231) for this obvious fact and its (slightly less obvious) converse.

Fix any sample path s such that $\sup_{A \in \mathcal{C}} |\lim_{t \rightarrow \infty} \nu^t(s)(A) - P(A)| = 0$. If $p \in \lim_{t \rightarrow \infty} \mu^t(s)$, then it follows from the definition of μ^t 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 $\tilde{p}(A) \equiv p(\phi^{-1}(A))$ 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 $p = P$.

A.4 Proof of Theorems 4 and 7: Failure of complete learning in discrete outcome spaces

We first prove the result when \mathcal{C} is a family of events of finite VC-dimension v . Given the proof it will be easy to show that the same holds when \mathcal{C} is of the form $\cup_{t=1}^{\infty} \mathcal{C}_t$ asserted in the theorem.

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}^* . The idea of the proof is: (1) Construct a “nice” finitely additive probability measure λ on \mathcal{C} ; (2) Obtain a perturbation \bar{s} of the density of λ that leaves its values on events in \mathcal{C} unaffected; (3) Show that applying this perturbation to λ yields a new finitely additive probability measure $\gamma \neq \lambda$ that, by construction, coincides with λ on \mathcal{C}^* .

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

Let \mathcal{U} be any free ultrafilter on the integers, and define

$$\lambda(A) = \mathcal{U}\text{-}\lim_{N \rightarrow \infty} \lambda_N(A),$$

where this expression means that for every $\epsilon > 0$ the set $\{N : |\lambda_N(A) - \lambda(A)| < \epsilon\}$ belongs to \mathcal{U} . Intuitively, λ is a “uniform” distribution on the integers. Note that λ is atomless (*i.e.*, assigns zero mass to each point) and purely finitely additive.

A.4.2 Perturbations

A *perturbation* is any function $s : X \rightarrow \{1 - \epsilon, 1 + \epsilon\}$. 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 events of the form $\{s : s(x) = 1 + \epsilon\}$ for some $x \in X$.

Let P be the standard *countably* additive product measure on the measure space (S, Σ) assigning probability 0.5 to each of the events $\{s : s(x) = 1 + \epsilon\}$, $x \in X$. Note that (S, Σ, P) is a standard countably additive probability space; 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 restriction of \mathcal{C} to 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, Györfi, 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 restrict how rich a family of events can be. For example, with 50 states ($N = 50$) the ratio of any family of VC-dimension 5, say, 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 families of events in X of VC-dimension 1 with uncountable cardinality (see Devroye, Györfi, 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 $\{A_N, A_N^c\}$ 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,

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

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, $Z_{\alpha N}$, $N = 1, 2, \dots$ is a sequence of independent events, for any fixed $\alpha > 0$. (This is reason why we insisted on using 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 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

$$Q_0 \equiv \bigcap_{k=1}^{\infty} (Z_{1/k})^c \tag{10}$$

also has P -measure 1.

In addition, using the law of large numbers, $P(Q_1) = 1$ where $Q_1 = \{s \in S : \lambda\{x : s(x) = 1 + \epsilon\} = \frac{1}{2}\}$ (this requires an argument, along the lines in Al-Najjar (2007)—but straightforward).

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

A.4.3 Perturbed Measures

Define the set function:

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

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 $\epsilon \in [-1, 1]$. 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 verify that the perturbed measure γ must differ from λ on some (in fact, many) events outside \mathcal{C}^* . Take the event $B \equiv \{x : s(x) = 1 - \epsilon\}$. Note that s was chosen so that this event does not belong to \mathcal{C} . Since $s \in Q_1$, we have $\lambda(B) = 0.5$. On the other hand,

$$\gamma(B) \equiv \int_B s(x) d\lambda = (1 - \epsilon) \lambda(B) \neq \lambda(B). \quad (11)$$

It only remains to show that λ and γ coincide on \mathcal{C} (hence necessarily on \mathcal{C}^*). Take any set $A \in \mathcal{C}$. If $\lambda(A) = 0$, then

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

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, assume that $\lambda(A) \geq 0.5$ (if not, take its complement and use additivity again). Then there is a subsequence N_k , $k = 1, 2, \dots$ such that $\lambda_{N_k}(A) \rightarrow \lambda(A)$ [this follows from the ultrafilter construction], hence for each k we have $A_{N_k} \in \mathcal{C}'_{N_k}$.

Now

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

Using the triangle inequality, we have:

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

from which we conclude that:

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

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

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

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

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

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

A.4.4 Proof of Corollary 5

From Eq. 11 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 countable.

A.4.5 Proof of Theorem 7

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 Eq. 10 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 Proposition 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

$\mathbb{C}^* \subset \mathbb{C}$ such that $\mathcal{C} \in \mathbb{C}^*$ and \mathbb{C}^* is linearly ordered by set inclusion. Define $\bar{\mathcal{C}} \equiv \cup_{\hat{\mathcal{C}} \in \mathbb{C}^*} \hat{\mathcal{C}}$.

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, \dots, x_m that can be shattered by $\bar{\mathcal{C}}$ but not by any class in \mathbb{C}^* . But shattering a finite sample requires only finitely many events. By the definition of $\bar{\mathcal{C}}$ there must be $\hat{\mathcal{C}} \in \mathbb{C}^*$ 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}^* is maximal, $\bar{\mathcal{C}} \in \mathbb{C}^*$. ■

References

- AL-NAJJAR, N. I. (1999): “Complexity as a barrier to competitive imitation,” MEDS Department, Kellogg School of Management, Northwestern University.
- (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.
- BHASKARA RAO, K. P. S., AND M. BHASKARA RAO (1983): *Theory of charges*. New York: Academic Press Inc.
- BILLINGSLEY, P. (1995): *Probability and measure*, Wiley Series in Probability and Mathematical Statistics. Third edn. New York: John Wiley & Sons Inc. 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*. Berlin: Springer.
- 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*. New York: McGraw-Hill Book Co.
- 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 M. SCHNEIDER (2005): “Learning Under Ambiguity,”
 University of Rochester.
- 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.
- FUDENBERG, D., AND D. K. LEVINE (1993): “Steady state learning and Nash equilibrium,” *Econometrica*, 61(3), 547–573.
- GAJDOS, T., T. HAYASHI, J.-M. TALLON, AND J.-C. VERGNAUD (2006):
 “Attitude toward Imprecise Information,” Universite Paris I, Pantheon-Sorbonne.
- GILBOA, I., A. POSTLEWAITE, AND D. SCHMEIDLER (2006): “Rationality of Belief, Or why Bayesianism is Neither Necessary Nor Sufficient for Rationality,” University of Pennsylvania.
- 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,” NYU.
- HARMAN, G., AND S. KULKARNI (2007): *Reliable Reasoning: Induction and Statistical Learning Theory*. MIT Press (Forthcoming).
- 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.
- MANSKI, C. F. (2004): “Statistical treatment rules for heterogeneous populations,” *Econometrica*, 72(4), 1221–1246.

- 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*. New York: MacMillan Publishing Co., Inc.
- 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*. New York: John Wiley & Sons Inc.
- (1967): “Difficulties in the theory of personal probability,” *Philosophy of Science*, 34, 305–10.
- (1972): *The foundations of statistics*. revised edn. New York: Dover Publications Inc.
- 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. New York: John Wiley & Sons Inc. 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): “Qualitative Probabilities on λ -systems,” *Mathematical Social Sciences*, 38, 11–20.