Testing theories with learnable and predictive representations ☆

Nabil I. Al-Najjar [a,*], Alvaro Sandroni [a,b], Rann Smorodinsky [c], Jonathan Weinstein [a]

[a] *Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL 60208, United States*
[b] *Department of Economics, University of Pennsylvania, United States*
[c] *Davidson Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel*

Abstract

We study the problem of testing an expert whose theory has a learnable and predictive parametric representation, as do standard processes used in statistics. We design a test in which the expert is required to submit a date T by which he will have learned enough to deliver a sharp, testable prediction about future frequencies. We show that this test passes an expert who knows the data-generating process and cannot be manipulated by a uninformed one. Such a test is not possible if the theory is unrestricted.
© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Forecasting is central to all aspects of economic activity. Consumers, firms, investors and governments base their decisions on forecasts of variables such as demand, cost, stock prices and indices, GDP, unemployment and interest rates. The prevailing practice in economic modeling sidesteps the problem of forecasting by envisioning economic agents who know the underlying data-generating process. In practice, however, decision makers almost never know the true process, and must therefore either learn from data and/or seek the advice of 'experts.' Entire industries, such as management consulting and the countless forms of financial advising and analysis, are dedicated to the business of supplying such forecasts. Looking beyond economic activities, self-declared 'experts' thrive in all environments fraught with ambiguity, volunteering predictions on subjects ranging from war and politics, to diet and climate change.

Can forecasts be tested? Can we tell apart true experts from charlatans? Cures from quacks? Science from pseudo-science? These questions have been the focus of a growing number of recent papers. The purported expert in this setting announces, at the beginning of each period, a history-dependent probabilistic forecast for that period's outcome. A (pure) *forecasting strategy*, or *theory*, for this expert is a function that assigns to each finite history of past outcomes an element of $\Delta(A)$, the simplex over the (finite) set of outcomes A. A test τ takes as input the expert's theory and the actual sequence of realized outcomes, and decides whether to *accept* or *reject* the expert.

This literature's most robust—and startling—finding is that all reasonable tests can be manipulated. Manipulation means that a strategic expert has a mixed strategy that is guaranteed to pass the test with high probability, regardless of how the data is generated. The key assumption leading to manipulation is that an expert who reports any data-generating process μ is guaranteed to pass the test with high μ-probability. It is important for this result that the expert is free to submit any probability distribution whatsoever.

One way around the impossibility result, then, is to restrict the theories the expert is allowed to submit. Restrictions, however, should not be arbitrary. Ideally, they should be aligned with normative standards, such as those typically expected of scientific theories and statistical models. Two such appealing standards are that theories should be:

1. *Learnable:* the expert should be able to refine his initial theory by learning from data; and
2. *Predictive:* the expert does not need to keep learning forever; eventually, he will have learned enough so that new evidence will have a small effect on his predictions about the distant future.

We follow the formalization of these ideas introduced in [7]. The main result of this paper is to show that the class of theories that are learnable and predictive, denoted $\mathcal{P}^\star$, can be tested. Specifically, we construct a test τ that passes any theory in $\mathcal{P}^\star$ and is not manipulable.

To make the notions of learnability and predictiveness precise, we first express any theory μ as a probability distribution λ on a set of parameters Θ, with each $\theta \in \Theta$ indexing a stochastic process μ_θ.[1] Here, λ represents prior beliefs about the likelihood of various values of the true parameter θ. As observations accumulate, the expert refines his initial theory by updating his beliefs over the parameters. We consider only parametric representations that are *learnable* in

---

[1] The classic example is de Finetti's representation of exchangeable processes with outcomes in $\{0, 1\}$ as a two stage lottery. Here, $\Theta = [0, 1]$ and μ_θ is i.i.d. with mean θ.

the sense that, as data accumulate, the expert is eventually able to forecast as if he knew the true parameter θ to any desired degree of precision.

The restriction to learnable parametrizations is, by itself, vacuous since any μ has a trivial "learnable" representation with a singleton $\Theta = \{\theta\}$ and $\mu_\theta = \mu$. The force of the model comes from requiring that the parameters generate sharp enough predictions about future realizations to be tested against new evidence.

The key assumption is that the theory μ has a *learnable and predictive* parametrization. The property of predictiveness, introduced and characterized by Jackson, Kalai and Smorodinsky [7],[2] requires that, given a parameter θ and an integer t, the outcomes of the next t periods do not improve predictions of outcomes in the distant future. This formalizes the intuition that the θ is indeed a parameter in the sense that, as in statistical models, it summarizes all that can be learned from the data. The requirement that a representation is learnable and predictive then says that the theory has been decomposed into its 'smallest' learnable components. Any coarser representation overlooks patterns that could have been learned, while finer representations contain parameters whose values are impossible to learn.

We show that the class of learnable and predictive theories, denoted $\mathcal{P}^\star$, is testable. Specifically, there is a test τ such that: (1) any expert who knows the true $\mu \in \mathcal{P}^\star$ can pass; and (2) the test is non-manipulable: for any randomization φ over theories by a manipulating expert there is $\mu \in \mathcal{P}^\star$ such that the expert is rejected with arbitrarily high probability.

Our ability to construct such a test hinges on a key property of learnable and predictive theories, which is a consequence of Lemma A.2. For any theory in $\mathcal{P}^\star$ there is a date T and, as a function of the observations up to time T, a distribution $\alpha \in \Delta(A)$ and subsequent dates $m_1 < m_2 < \cdots$ such that the time-T conditional forecast of the outcome at these dates is approximately i.i.d. with distribution α. Our test is based on a classical hypothesis test of these i.i.d. conditional predictions.

An expert with a theory in $\mathcal{P}^\star$ learns in the sense that his conditional forecast α and dates $m_1, m_2, \ldots$ are all functions of the observations up to time T. Crucially, however, although the expert is allowed to learn from the data, he is not free to revise his theory arbitrarily. He must pre-specify a time T at which he learns enough to make sharp predictions.

In the literature, a test τ is *manipulable* if for every $\epsilon > 0$ the expert has a mixed strategy φ that passes the test on every sample path with probability $1 - \epsilon$. This strong criterion for manipulability is what makes impossibility results so striking. For our positive result, we introduce a stronger criterion than the negation of manipulability. First, for every expert's mixed strategy φ there is a path at which the expert's probability of passing is no more than ϵ. Second, conditional on the announced T, a fixed μ, that does not depend on the mixed strategy used, holds the expert to a low probability of passing. This says that announcing that T periods are sufficient for learning is a substantial commitment that a manipulative expert cannot get around.

1.1. Literature

The impossibility of designing non-manipulable tests was first established for the special class of *calibration tests* by Foster and Vohra [5].[3] Sandroni [15] showed that this result holds for *all* finite-horizon tests. These are tests that condition only on observations during a pre-specified

---

[2] What we call predictive is "sufficient for predictions" in their terminology.

[3] Their result was subsequently strengthened by various authors, including Fudenberg and Levine [6], Kalai, Lehrer and Smorodinsky [9], Lehrer [10] and Sandroni, Smorodinsky and Vohra [16].

finite number of periods T. Finite-horizon tests are restrictive because for every T there may exist theories that are sufficiently complex to require more than T periods to test. In the infinite horizon, some non-manipulable tests do exist, but these must lack certain desirable properties. For instance, Olszewski and Sandroni [12] showed that they cannot be future-independent and Shmaya [17] showed that they must condition on counter-factual predictions.[4]

Most relevant to our work is Olszewski and Sandroni [14] who show that, when the class of theories is unrestricted, a non-manipulable test cannot be an *acceptance* test. An acceptance test is one where, if the expert is accepted on an infinite path, there is an initial segment of the path on which he is accepted for any continuation. Rejection tests are defined similarly. Our test is both an acceptance test and a rejection test, *i.e.*, it always gives an answer in finite time. Such a test could not be effective without any restriction on the set of theories.

Note that while our test always answers in finite time, this advantage is mitigated by the fact that there is no uniform bound on the time required. In particular, the expert can delay rejection for any finite duration he wishes. The good news is that he must specify this delay at the start of time. This means that the tester would be free not to "hire" him if the delay is too long, though rejecting experts with long delays would lead to rejection of some true experts with theories in $\mathcal{P}^\star$. A practical-minded reader, who wishes to insist on tests that answer in a pre-specified finite time, may imagine replacing our test with a family of tests in which the time allowed goes to infinity and the set of accepted theories converges to $\mathcal{P}^\star$.

Olszewski and Sandroni [14] also construct a non-manipulable test by restricting the class of theories to a non-convex set of distributions.[5] Their restriction lacks the statistical motivation of learnability and predictiveness we have given earlier. We also argue in Sections 3.4 and 4.3 that convexity is a desirable property of a class of theories because it corresponds to allowing the expert to learn.

Dekel and Feinberg [2] and Olszewski and Sandroni [13] showed the existence of non-manipulable tests without restricting the class of theories. However, these are necessarily not acceptance tests, and thus may not return any outcome in finite time. In addition, as the results of Olszewski and Sandroni [12] and Shmaya [17] show, these tests are difficult to implement. In contrast, our test is a simple adaptation of a test frequently used in practice.

Al-Najjar and Weinstein [1] and Feinberg and Stewart [3] consider the problem of testing multiple experts when one expert knows the true process. As shown by Al-Najjar and Weinstein [1], the positive results in these papers do not circumvent the single-expert testing problem because there is no way to test if there is at least one expert who knows the truth.

Finally, Fortnow and Vohra [4] proved positive results based on algorithmic complexity of theories. Roughly, there is a test that cannot be manipulated by an expert who submits (algorithmically) simple forecasts.

2. Tests and their properties

2.1. Model and notation

- Fix a finite set A representing outcomes in any given period. For any set let $\Delta(\cdot)$ denote the set of probability distributions on that set.

---

[4] A test is future-independent if, whenever a theory f is rejected at some history h^T, the test rejects all other theories f' that make the same predictions as f until period $T + 1$. A test does not condition on counter-factual predictions if its decision to accept or reject depends only on the forecasts made along the actual history h^∞.

[5] The class they rule out consist of all theories that are sufficiently close to a given theory μ.

- Assume, for expository reasons, that $A = \{0, 1\}$. All of our analysis goes through with any finite A.
- The horizon is infinite, with time periods indexed by $t = 1, 2, \ldots$. Let H^∞ denote the set of infinite sequences of outcomes, which we shall also refer to as complete histories.[6]
- The set of histories up to and including time t is denoted H^t, with generic element h^t.
- At a history h^∞, the outcome at time t is denoted $a_t(h^\infty)$, or simply a_t when h^∞ is clear from the context.
- Let $\mathcal{F}^t_n$ denote the algebra of events determined by the outcomes between periods n and t. When $n = 0$, we simply write $\mathcal{F}^t$ instead of $\mathcal{F}^t_0$.
- Let $\mathcal{F}^\infty_n$ denote the σ-algebra of events determined by the outcomes at all times $t \geqslant n$. Again, we shall write $\mathcal{F}^\infty$ instead of $\mathcal{F}^\infty_0$.
- The set of (countably additive) probability measures on H^∞ is denoted $\mathcal{P}$ and is endowed with the weak topology.
- A *theory* is a probability measure μ in $\mathcal{P}$.

2.2. Tests

We shall define tests and their properties in a more general setting than typically done in the literature. In our setting, an expert communicates a message $\mathbf{m}$, assumed to be an element of some measurable space of messages $\mathcal{M}$. A test is any measurable function

$$\tau : \mathcal{M} \times H^\infty \to \{0, 1\}$$

with the interpretation that the test takes as input a message $\mathbf{m}$ and an infinite path h^∞ and delivers a verdict of either accepting or rejecting the expert (1 or 0, respectively).

This formulation is more general than what is common in the literature, where the standard assumption is $\mathcal{M} = \mathcal{P}$, so the expert reports a probability distribution as his message.[7] Our reason for the more general setting is expositional: making the set of messages explicit gives a better sense of how much information the test needs in order to render a verdict.

2.3. Properties of tests

First we introduce some notation. Define

$$V(\varphi, \mu) \equiv \int_{\mathbf{m}} \mu\{h^\infty \colon \tau(\mathbf{m}, h^\infty) = 1\} \, d\varphi(\mathbf{m}),$$

which is the expert's payoff, namely the probability of passing given that the expert uses the mixed strategy φ and the true process is μ. For notational convenience we will sometimes use the message $\mathbf{m}$ or the path h^∞ to indicate Dirac measures that put unit mass on $\mathbf{m}$ and h^∞ respectively.

Let $\mathcal{P}' \subseteq \mathcal{P}$ be an arbitrary set of theories.

---

[6] To minimize repetition, from this point on, all product spaces are endowed with the product topology and the Borel σ-algebra.

[7] This does not weaken our result. We could recast our test into the standard setting by having the expert report a probability measure and then converting it to a message as in the proof of Proposition 4.1.

Definition 1. A test τ *passes all theories in* $\mathcal{P}'$ with probability $1 - \epsilon$ if for every $\mu \in \mathcal{P}'$ there is a message $\mathbf{m}_\mu$ such that

$$V(\mathbf{m}_\mu, \mu) > 1 - \epsilon.$$

In the testing literature, this assumption is usually stated for $\mathcal{P}' = \mathcal{P}$ and $\mathbf{m}_\mu = \mu$. That is, no restriction on the class of permitted processes is made and the test passes with high probability an expert who reports the true process. This is the crucial substantive assumption responsible for the impossibility theorems. The goal of this paper is to identify reasonable restrictions on the class of theories to enable effective testing.

Definition 2. A test τ can be *ignorantly passed* with probability $1 - \epsilon$ if there is a mixed strategy $\varphi \in \Delta(\mathcal{M})$ such that, for every $\mu \in \mathcal{P}$,

$$V(\varphi, \mu) > 1 - \epsilon.$$

As indicated in the Introduction, the manipulability of a test is a strong and striking property in the context of impossibility theorems. Inevitably, this means that the negation of manipulability is a weak requirement: a test is not manipulable provided only that each mixed strategy fails on one sample path with probability greater than ϵ. Our definition of testability includes a notion of non-manipulability which is stronger than the negation of manipulability:

Definition 3. A set of theories $\mathcal{P}'$ is ϵ-*testable by a test* τ if

1. τ passes all theories in $\mathcal{P}'$ with probability $1 - \epsilon$; and
2. for every $\varphi \in \Delta(\mathcal{M})$ there is $\mu \in \mathcal{P}'$ such that $V(\varphi, \mu) < \epsilon$.

The set $\mathcal{P}'$ is *testable* if, for every $\epsilon > 0$, $\mathcal{P}'$ is ϵ-testable by some test τ.

Note that in the second part of the definition, we require that for any strategic expert, there is a $\mu \in \mathcal{P}'$ on which he fails. This means that merely knowing that the distribution is in $\mathcal{P}'$ is not enough to ensure that an expert passes; the expert must know the particular element of $\mathcal{P}'$. To satisfy both parts of this definition, $\mathcal{P}'$ must be neither too big nor too small, but in an intermediate range where knowing the particular μ is superior to knowing we are in the class $\mathcal{P}'$. We will not give a full classification of testable theories in this paper; our main result is to show that the class $\mathcal{P}^\star$ we are about to define is testable (and indeed, so are all classes lying between the smaller class of Dirac measures and $\mathcal{P}^\star$).

3. Learnable and predictive theories

3.1. Parametric representations of stochastic processes

We define the notions of learning and predictions in terms of parametric representations of stochastic processes. Formally,

Definition 4. A *parametric representation* of a stochastic process μ is a quadruple $(\Theta, \mathcal{B}, \lambda, (\mu_\theta)_{\theta \in \Theta})$ where $(\Theta, \mathcal{B}, \lambda)$ is a probability space and $(\mu_\theta)_{\theta \in \Theta}$ is a family of probability distributions on $(H^\infty, \mathcal{F}^\infty)$ such that for every event $S \in \mathcal{F}^\infty$

1. The function $\theta \mapsto \mu_\theta(S)$ is measurable; and
2. $\mu(S) = \int_\Theta \mu_\theta(S) \, d\lambda(\theta)$.

The notion of a parametric representation is not useful without further structure; in fact, every process always has two trivial parametrizations:

1. *The coarsest parametrization:* $\Theta = \{\theta\}$, $\mu_\theta = \mu$, and λ puts unit mass on θ;
2. *The finest parametrization:* $\Theta = H^\infty$, μ_{h^∞} is a Dirac measure on the infinite sequence of outcomes h^∞, and $\lambda = \mu$.

Next we discuss formal criteria for evaluating representations that will rule these out, namely learnability and predictiveness.

3.2. Learnability

A natural requirement to impose on a parametrization is for it to be "learnable." This means that, as evidence accumulates, the updated forecasts become close to those made if the true parameters were known. The idea of learnability is made formal using the notion of weak merging introduced by Kalai and Lehrer [8].[8] We follow Jackson, Kalai and Smorodinsky [7] and define:

Definition 5. A parametric representation $(\Theta, \mathcal{B}, \lambda, (\mu_\theta)_{\theta \in \Theta})$ of a stochastic process μ is *learnable* if for λ-a.e. θ: for every $\epsilon > 0$, non-negative integers l, and μ_θ-a.e. h^∞, there is T such that for all $t \geqslant T$

$$\sup_{\substack{n \geqslant t \\ S \in \mathcal{F}_n^{n+l}}} \left| \mu(S | \mathcal{F}^t) - \mu_\theta(S | \mathcal{F}^t) \right| < \epsilon.$$

For a fixed θ, the above simply says that μ *weakly merges* with μ_θ, in the sense of [8]. Note that weak merging does not imply that one learns the true θ, only that the forecasts made according to the conditional distribution and those made using μ_θ become close in the sense of weak merging.

The notion of learnability rules out parametrizations that are too fine. In particular, a consequence of the proof of Proposition 3.1 is that the finest parametrization introduced above is not learnable whenever μ has uncountable support.

3.3. Predictive representations

Learnability, by itself, is not a very demanding property. Any process has coarse parametrizations where knowledge of the true parameter is of little or no value. Effective testing requires representations in which parameters are fine enough to yield sharp predictions that can be tested.

The following is a key concept, introduced by Jackson, Kalai and Smorodinsky [7]:

Definition 6. A probability distribution p over $(H^\infty, \mathcal{F}^\infty)$ is *sufficient for predictions* if for all t

$$\lim_n \sup_{S \in \mathcal{F}_n^\infty} \left| p(S | \mathcal{F}^t) - p(S) \right| = 0.$$

---

[8] See [18] for characterizations and links to the literature.

A process p is sufficient for predictions if knowledge of the realizations in the near future is of little help in refining forecasts about the distant future. Examples include i.i.d. distributions and Dirac measures δ_{h^∞}. A less trivial example is irreducible Markov processes. Given any such process, learning the next few realizations will typically have some value in predicting realizations in the near future. However, realizations in the near future convey no useful information about the behavior of the process in the distant future. On the other hand, any non-i.i.d. exchangeable process, or mixture of irreducible Markov processes, is not sufficient for predictions.

Definition 7. A parametric representation $(\Theta, \mathcal{B}, \lambda, (\mu_\theta)_{\theta \in \Theta})$ is *predictive* if μ_θ is sufficient for predictions for λ-a.e. θ.

The key definition for our paper is:

Definition 8. $\mathcal{P}^\star$ denotes the class of all theories with a learnable and predictive representation.

Jackson, Kalai and Smorodinsky [7] proved that a stochastic process has a predictive and learnable representation if and only if it satisfies a condition called *asymptotic reverse mixing*. This provides a representation-free characterization of $\mathcal{P}^\star$. Roughly, it says that we eventually learn everything about the tail behavior that is relevant to predicting the near future. We refer the reader to their paper for a precise definition and discussion of this condition, and for an example of a distribution which is not in $\mathcal{P}^\star$.[9] In the present paper, we find it more useful to work with the characterization via representations.

3.4. How appealing is $\mathcal{P}^\star$?

The value of our approach depends on the appeal of the class of predictive theories, $\mathcal{P}^\star$. Here we provide some examples and considerations that can be helpful in illuminating this issue.

For a class of probability measures $\mathcal{Q}$, let $\text{co}\,\mathcal{Q}$ denote its convex hull. Consider the following classes of theories:

- *Deterministic theories, $\mathcal{P}_{Det}$.*
- *i.i.d. theories, $\mathcal{P}_{IID}$.*
- *Markov theories, $\mathcal{P}_{Mkv}$.*

A deterministic theory is a (Dirac) measure δ_{h^∞} that puts unit mass on a single path h^∞. Such a theory makes deterministic predictions about the outcome in each period. The definition of i.i.d. and Markov theories is standard. Note that by De Finetti's theorem, the closure of the convex hull of $\mathcal{P}_{IID}$ is the class of exchangeable distributions $\mathcal{P}_{Ex}$.

Our view is that any reasonable class of theories should include the three classes above. Indeed, any theory in one of these classes above is sufficient for predictions, and thus belongs to $\mathcal{P}^\star$.[10]

We also argue that it is desirable that the expert should be allowed to learn. This corresponds to accepting theories that are convex combinations of parameters. These represent prior beliefs that

---

[9] The distribution which we use in the proof of Theorem 4.2 is also outside $\mathcal{P}^\star$ and is very similar to their example.

[10] In this special case, the coarse parametrization with only one parameter is both learnable and predictive.

can potentially be refined as data accumulates. To illustrate this, consider an expert who predicts that the path will be either the deterministic sequence h_1^∞ or h_2^∞ with equal probability. This expert will eventually be able to make deterministic predictions after an initial period of learning. The general point is that we should not restrict the class of theories so it rules out experts who know that the true parameter is either θ or θ' but who need data to be able to tell which. The class of learnable and predictive theories, $\mathcal{P}^\star$, is indeed convex, while the three classes above are not. Note that $\mathcal{P}^\star$ is not subject to the impossibility results mentioned in the Introduction because it is convex but not compact.

In addition to being convex and containing important classes of distributions as special cases, $\mathcal{P}^\star$ is rich enough that the knowledge of a truthful expert cannot be imitated by someone who merely knows that the process is in $\mathcal{P}^\star$. This is in contrast with the class of exchangeable theories: If one knows that the process lies in $\mathcal{P}_{Ex}$, then there is a procedure that will always learn the true parameter. In particular, if μ is the exchangeable distribution obtained by uniformly randomizing on Θ, then the conditional distribution $\mu(\cdot\|h^T)$ will weakly merge with the true μ_θ as data accumulates. Restricting theories to be in $\mathcal{P}_{Ex}$ imposes so much structure that the value of the true parameter can be inferred from the data in a purely empirically manner. In this case there is no need to rely on an expert in the first place.

The next proposition shows that, by contrast, knowing merely that the process lies in $\mathcal{P}^\star$ is not enough to guarantee that one can eventually predict as well as an expert who knows the particular $\mu' \in \mathcal{P}^\star$.

Proposition 3.1. *There does not exist a belief $\mu \in \mathcal{P}$ that weakly merges with every μ in $\mathcal{P}^\star$.*

Proof. Note that every Dirac measure is in $\mathcal{P}^\star$, so it suffices to show that μ cannot weakly merge with every such measure. Abusing notation, denote the measure that puts unit mass on the path h^∞ by h^∞. We will say that μ ϵ-merges with h^∞ by time T if the statement in Definition 5 holds for ϵ and T, with $l = 1$.

Fix any $\epsilon < 0.5$. We claim that for each partial history h^T, μ ϵ-merges with h^∞ by time T for at most one continuation h^∞ of h^T. Indeed, μ is required in each period to assign probability greater than 0.5 to the realization along h^∞, and this can be true for at most one continuation. Since the set of partial histories is countable, while H^∞ is uncountable, there is some h^∞ with which μ does not ϵ-merge by time T for any T, completing the proof. $\quad\square$

Indeed, the test we develop in this paper can be passed by an expert who knows the true μ when $\mu \in \mathcal{P}^\star$, but not by an expert who merely knows that the process is some $\mu \in \mathcal{P}^\star$ without knowing which.

4. Main results

4.1. The test

We consider a family of tests, parametrized by the sharpness we require for the expert's predictions, δ, and the testing time, L. We first fix a sequence of infinite sets of integers $\{\mathcal{N}_T\}_{T=1}^\infty$ such that

1. $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$ for $i \neq j$;
2. $n > T$ for every $n \in \mathcal{N}_T$.

Given $\delta > 0$ and integer L, define the test $\tau(\delta, L)$ by these steps:

1. The expert submits a date T;
2. At date T, having observed a history h^T, the expert submits dates

$$m_1, \ldots, m_L \in \mathcal{N}_T$$

   and a number $\alpha \in [0, 1]$;
3. The expert passes if and only if:

$$\left| \frac{a_{m_1}(h^\infty) + \cdots + a_{m_L}(h^\infty)}{L} - \alpha \right| < \delta. \tag{1}$$

In terms of our formalism, the expert communicates a message that consists of a date T and $L + 1$ $\mathcal{F}^T$-measurable functions $\alpha(h^T), m_1(h^T), \ldots, m_L(h^T)$. The interpretation is that the expert uses the data of the first T periods to calibrate his initial theory. At the end of the T periods, and as a function of the history h^T, he specifies L dates $m_1, \ldots, m_L$ on which he predicts the outcomes to be approximately i.i.d. with mean α. The test then requires the actual mean to be close to α in the periods $m_1, \ldots, m_L$ announced by the expert.

Other aspects of this test, including the need for the partition $\{\mathcal{N}_T\}_{T=1}^\infty$, will be discussed in Section 4.3.

4.2. Main theorem

Theorem 1. $\mathcal{P}^\star$ *is testable.*

This theorem is a consequence of Proposition 4.1 and Corollary 1. Proposition 4.1 shows part (1) of the definition of testability (Definition 3):

Proposition 4.1. *For every $\epsilon, \delta > 0$ there exists $\bar{L}$ such that any expert who knows the truth $\mu \in \mathcal{P}^\star$ can pass the test $\tau(\delta, L)$ with probability at least $1 - \epsilon$ for every $L > \bar{L}$.*

The proof, found in Appendix A, uses lemmas that put uniform bounds on the times by which learnability and sufficiency for predictions start to have a bite. The conclusion then follows from a result of Lehrer [11] that establishes a law of large numbers for approximate i.i.d. sequences.

Next we turn to non-manipulability. We discuss the short proof of the following intermediate result in the next subsection.

Proposition 4.2. *For any $\epsilon > 0$ there exist δ and L such that, fixing the test $\tau(\delta, L)$, there is $\mu \in \mathcal{P}$ such that $V(\varphi, \mu) < \epsilon$ for every $\varphi \in \Delta(\mathcal{M})$.*

Proof. We construct $\mu \in \mathcal{P}$ such that for any φ, $V(\varphi, \mu) < \epsilon$. We define μ through a two-stage process. For each T select a number θ_T independently and uniformly from $[0, 1]$. Conditional on θ_T, the observations on the subset $\mathcal{N}_T$ are independent with mean θ_T. On the set $\mathcal{N} - \bigcup_{T=1}^\infty \mathcal{N}_T$, μ assigns unit mass to the constant 0.

Next we bound the expected payoff the expert can achieve conditional on announcing any T, h^T and α. Clearly, for every α the probability of $\theta_T \in [\alpha - 2\delta, \alpha + 2\delta]$ is at most 4δ. Furthermore, for every $\delta, \epsilon_2 > 0$ using Lemma A.3 we can pick L such that $f_L \notin [\theta_T - \delta, \theta_T + \delta]$ with

probability at most ϵ_2. Thus, $f_L \in [\alpha - \delta, \alpha + \delta]$ with probability at most $4\delta + \epsilon_2$. Since h^T and α were arbitrary, the probability that this expert passes is no more than $(4\delta + \epsilon_2)(1 - \epsilon_1) + \epsilon_1$. Since δ, ϵ_1 and ϵ_2 were arbitrary, we indeed have $V(\varphi, \mu) < \epsilon$. □

Proposition 4.2 does not directly deliver part (2) of definition of testability since the measure μ constructed is not in $\mathcal{P}^\star$. However, since Dirac measures are in $\mathcal{P}^\star$, the following corollary suffices to complete the proof of the theorem:

Corollary 1. *Let μ be chosen as in the proposition, with $\epsilon = v^2$. Then for every $\varphi \in \Delta(\mathcal{M})$, h^∞ satisfies $V(\varphi, h^\infty) < v$ with μ-probability at least $1 - v$, hence for uncountably many h^∞ (since μ is atomless).*

A variant of Proposition 4.2 states that once T is announced, there is an element $\mu \in \mathcal{P}^\star$ for which the expert can no longer pass (in Proposition 4.2 it is necessary that $\mu \notin \mathcal{P}^\star$). This means that announcing that T periods are sufficient for learning is a substantial enough commitment to prevent the expert from passing for all $\mu \in \mathcal{P}^\star$.

Proposition 4.3. *For every $\epsilon > 0$ and T there exists $\mu \in \mathcal{P}^\star$ such that $V(\varphi, \mu) < \epsilon$ for any $\varphi \in \Delta(\mathcal{M})$ such that $\varphi(T \leqslant \bar{T}) > (1 - \epsilon)$.*

Proof. Again we define μ through a two-stage process. First, for each $T \leqslant \bar{T}$ select a number θ_T independently and uniformly from $[0, 1]$. Then, for each $t \in \mathcal{N}_T$, with $T \leqslant \bar{T}$, the observation at time T is independent with mean θ_T. On the set $\mathcal{N} - \bigcup_{T=1}^{\bar{T}} \mathcal{N}_T$, μ puts unit mass on the constant 0. This distribution is clearly in $\mathcal{P}^*$ with parameter set $[0, 1]^{\bar{T}}$ and uniform measure. The remainder of the proof follows exactly as in the remainder of the proof of Proposition 4.2, with ϵ replaced by any $\epsilon' < \varphi(T \leqslant \bar{T}) - (1 - \epsilon)$. □

4.3. Discussion

The challenge in Proposition 4.2 is to construct a probability distribution for nature that foils the false expert's randomization. To gain some intuition, one may roughly think of the expert as randomizing over: (1) the *learning date T*; (2) the history-dependent subsequence of *testing dates $m_1, \dots, m_L$*; and (3) the *predicted frequency α*.

First, hold T fixed and consider the choice of testing dates. This choice is irrelevant against the μ specified in the proof of Proposition 4.2. Acceptance and rejection are determined in a manner that does not depend on the particular dates chosen because Nature's randomization is exchangeable across dates. Exchangeability renders the strategic expert's freedom to manipulate the testing dates useless. Also, the uniform choice of θ renders the freedom to choose α useless.

Consider next randomizations over the learning date T. By choosing μ to be independent across different elements of the partition $\{\mathcal{N}_T\}_{T=1}^\infty$, we guarantee that by any time T, the expert has actually learned nothing about the dates $\mathcal{N}_T$ on which he will be tested. This eliminates the possibility of manipulating the test by the choice of T.

An interesting feature of our test is how it relates to convexity. A key assumption underlying the impossibility theorems is the convexity of the set of distributions. Recall that the set $\mathcal{P}^\star$ is convex. However, for any fixed T, the subset $\mathcal{P}_T^\star \subset \mathcal{P}^\star$ consisting of theories with learning

date T is *not* convex.[11] Intuitively, this is because if we take a randomization over two theories μ and $\mu' \in \mathcal{P}_T^\star$, learning whether μ or μ is the true process may take more observations than T. Given h^T, the expert is required to commit to an element of the non-convex set of i.i.d. measures. The non-convexity is crucial to our argument; if the expert could announce a mixture, the proof would fail.

A natural question is the extent to which the requirement that the parametrization be sufficient for predictions can be relaxed. The essential consequence of this requirement is that, conditional on the parameter, forecasts are approximately i.i.d. with mean α on some infinite sequence of dates (Lemma A.2). This is a key element of our proof that a true expert can pass our test. Suppose we instead required only that processes can predict, as a function of the observations up to time T, whether or not α belongs to some subset $B \subsetneq [0, 1]$, but not the precise value of α. To make sure true experts could pass, we would have to modify the test to ask only whether $\alpha \in B$. A manipulating expert could then guarantee a payoff of 0.5 by randomizing over whether to predict $\alpha \in B$. The full power of sufficiency for predictions is not necessary for our results—we only need the consequence in Lemma A.2. But the condition in the conclusion of that lemma would be rather artificial as a primitive assumption, so we have maintained the more natural notion.

5. Concluding remarks

If the data-generating process is outside of $\mathcal{P}^\star$, our test may reject an expert who knows the true process. Whether we consider such rejection acceptable depends on our answer to the question: *What makes someone an expert?* In our model, a legitimate theory is one that sets a date by which learning is complete and sharp predictions about the future can be made. This accords with our intuition of legitimate scientific theories as frameworks for predicting outcomes in a given context based on information about the relevant details of that context. Just as it would be unreasonable to expect a theory to deliver sharp predictions without data, a legitimate theory should not be given unlimited freedom to adjust its predictions either. Rather, there must be a point at which this theory makes predictions that can be tested.

Appendix A. Proof of Proposition 4.1

We begin with two uniformization lemmas. The first provides a uniform version of the learnability condition:

Lemma A.1. *Suppose that $(\Theta, \mathcal{B}, \lambda, (\mu_\theta)_{\theta \in \Theta})$ is a learnable and predictive representation for $\mu \in \mathcal{P}^\star$. Then for every $\epsilon_1, \epsilon_2 > 0$ and integer l there is an integer $T = T(\epsilon_1, \epsilon_2, l)$ and a set $\Theta_1 \subset \Theta$ with $\lambda(\Theta_1) > 1 - \epsilon_1$ such that for all $t \geqslant T$ and $\theta \in \Theta_1$*

$$\sup_{\substack{n \geqslant t \\ S \in \mathcal{F}_n^{n+l}}} \left| \mu\left(S | \mathcal{F}^t\right) - \mu_\theta\left(S | \mathcal{F}^t\right) \right| < \epsilon_2.$$

Proof. Given ϵ_2 and l, define the function

$$\vartheta_{\epsilon_2, l} : \Theta \to \{0, 1, \dots, \infty\}$$

---

[11] It should be noted that the learning dates associated with a theory also depend on ϵ. We suppress this dependence here for expository clarity.

by letting $\vartheta_{\epsilon_2,l}(\theta)$ be the smallest integer for which the definition of merging holds with l and ϵ_2. Then this function is measurable. Thus, $\lambda(\vartheta_{\epsilon_2,l}^{-1}(\{0,1,\ldots,T\}))$ is well defined and converges to 1 as $T \to \infty$. The desired $T = T(\epsilon_1,\epsilon_2,l)$ is chosen so that

$$\lambda\big(\vartheta_{\epsilon_2,l}^{-1}(\{0,1,\ldots,T(\epsilon_1,\epsilon_2,l)\})\big) > 1 - \epsilon_1$$

and

$$\Theta_1 = \vartheta_{\epsilon_2,l}^{-1}(\{0,1,\ldots,T(\epsilon_1,\epsilon_2,l)\}). \qquad \square$$

Define:

- $\mathrm{cov}(a_{m_j}, a_{m_i} \| \theta)$ to be the covariance between the two 0–1 random variables a_{m_j} and a_{m_i} computed according to the probability measure μ_θ;
- $\mathrm{cov}(a_{m_j}, a_{m_i} \| \mathcal{F}^T)$ to be the covariance between a_{m_j} and a_{m_i} computed according to conditional distribution $\mu(\cdot \| \mathcal{F}^T)$.

The next lemma provides a uniform date at which the expert makes approximately i.i.d. predictions:

Lemma A.2. *Suppose that $(\Theta, \mathcal{B}, \lambda, (\mu_\theta)_{\theta \in \Theta})$ is a representation for a $\mu \in \mathcal{P}^\star$. Let $\epsilon_3 > 0$. Then for every increasing sequence of positive integers $\{\bar{m}_1, \ldots\}$ there is an increasing subsequence of positive integers $\{m_1, \ldots\}$, and a subset $\Theta_2 \subset \Theta$ with $\lambda(\Theta_2) > 1 - \epsilon_3$ such that for every $\theta \in \Theta_2$*

$$\lim_{j \to \infty} \sup_{0 \leqslant i < j} \big|\mathrm{cov}(a_{m_j}, a_{m_i} | \theta)\big| = 0.$$

Proof. Fix a sequence $\{r_j\}_{j=1}^\infty$ such that $\sum_{j=1}^\infty r_j < \epsilon_3$. Next, define a sequence of integers $\{m_j, \ldots\}$ inductively as follows: set $m_1 = \bar{m}_1$; for $j > 1$, define the function

$$\varrho_j : \Theta \times H^\infty \to \{0, 1, \ldots, \infty\}$$

by letting $\varrho_j(\theta, h^\infty)$ be the smallest integer in $\{\bar{m}_1, \ldots\}$ such that for every $m > \varrho_j(\theta, h^\infty)$

$$\big|\mu_\theta(a_m | \mathcal{F}^{m_{j-1}}) - \mu_\theta(a_m)\big| < r_j$$

(here, and throughout the proof, we write a_m instead of $a_m(h^\infty)$). The fact that μ is in $\mathcal{P}^\star$ guarantees that for λ-a.e. θ, ϱ_j is finite μ_θ-a.s.

Since the function ϱ_j is measurable, we can set m_j to be the smallest integer in $\{\bar{m}_1, \bar{m}_2, \ldots\}$ such that the set $\Theta_j \subset \Theta$ of θ such that $\varrho_j(\theta, h^\infty) < m_j$ satisfies $\lambda(\Theta_j) > 1 - r_j$. Then the set $\bar{\Theta} \equiv \bigcap_{j=1}^\infty \Theta_j$ has λ-probability at least $1 - \epsilon_3$.

Fixing $\theta \in \bar{\Theta}$, for all j

$$\big|\mu_\theta(a_{m_j} | \mathcal{F}^{m_{j-1}}) - \mu_\theta(a_{m_j})\big| < r_j$$

from which it follows that

$$\sup_{0 \leqslant i < j} \big|\mathrm{cov}(a_{m_j}, a_{m_i} | \theta)\big| \to 0. \qquad \square$$

The next lemma adapts a result of Lehrer [11] to establish a law of large numbers for asymptotically independent random variables:

Lemma A.3. *Let $g_1, g_2, \ldots$ be a sequence of bounded, 0-mean random variables such that*

$$\lim_{j \to \infty} \sup_{0 \leqslant i < j} \operatorname{cov}(g_j, g_i) = 0.$$

Then there is a subsequence $\{g_{i_l}\}$ such that the random variable

$$f_l \equiv \frac{g_{i_1} + \cdots + g_{i_l}}{l}$$

converges to 0 a.s. In particular, for every $\epsilon, \delta > 0$, there is L such that for every $L > \bar{L}$

$$\operatorname{Prob}\{f_L \in [-\delta, \delta]\} > 1 - \epsilon. \tag{2}$$

Proof. This follows directly from Lehrer [11, Theorem 2, p. 259]. His theorem delivers the conclusion under the assumption that the covariances do not grow very rapidly. This can be guaranteed by passing to a subsequence. $\quad\square$

We are now able to prove Proposition 4.1:

1. Use Lemma A.1 to pick $T = T(\epsilon_1, \epsilon_1, 0)$ satisfying the conclusions of that lemma.
2. Use Lemma A.2 to pick, as a function of ϵ_3 and h^T, a sequence of dates

   $$\{m_i\}_{i=1}^{\infty} \subset \mathcal{N}_T$$

   that satisfy the conclusion of that lemma.
3. Partition the interval $[0, 1]$ into K equal subintervals, and choose $0 \leqslant k' < K$ such that

   $$\frac{k'}{K} < E(a_{m_i} | \mathcal{F}^T) \leqslant \frac{k'+1}{K}$$

   for infinitely many m_i's. Let $\{m_i'\}_{i=1}^{\infty}$ denote the infinite subsequence of dates on which this occurs.
4. Now apply Lemma A.3 to

   $$g_{m_i} = a_{m_i} - E(a_{m_i} | \mathcal{F}^T)$$

   to extract a subsequence $\{m_l''\}_{l=1}^{\infty}$ and an integer L such that for every $L > \bar{L}$ (2) holds, *i.e.*, $f_L \in [-\delta, \delta]$ with probability $1 - \epsilon$.
5. The expert submits the dates $\{m_1'', \ldots, m_L''\}$ and sets $\alpha = (1/L) \sum_{i=1}^{L} E(a_{m_i} | \mathcal{F}^T)$. By the previous point, the conclusion of the proposition holds.

References

[1] N.I. Al-Najjar, J. Weinstein, Comparative testing of experts, Econometrica 76 (3) (2008) 541–559.
[2] E. Dekel, Y. Feinberg, Non-Bayesian testing of an expert, Rev. Econ. Stud. 73 (2006) 893–906.
[3] Y. Feinberg, C. Stewart, Testing multiple experts, Econometrica 76 (2008) 561–582.
[4] L. Fortnow, R. Vohra, The complexity of forecast testing, Econometrica 77 (2009) 93–105.
[5] D. Foster, R. Vohra, Asymptotic calibration, Biometrika 85 (2) (1998) 379–390.
[6] D. Fudenberg, D.K. Levine, An easier way to calibrate, Games Econ. Behav. 29 (1) (1999) 131–137.
[7] M.O. Jackson, E. Kalai, R. Smorodinsky, Bayesian representation of stochastic processes under learning: de Finetti revisited, Econometrica 67 (4) (1999) 875–893.
[8] E. Kalai, E. Lehrer, Weak and strong merging of opinions, J. Math. Econ. 23 (1994) 73–86.
[9] E. Kalai, E. Lehrer, R. Smorodinsky, Calibrated forecasting and merging, Games Econ. Behav. 29 (1) (1999) 151–159.

[10] E. Lehrer, Any inspection is manipulable, Econometrica 69 (5) (2001) 1333–1347.

[11] E. Lehrer, Approachability in infinite dimensional spaces, Int. J. Game Theory 31 (2) (2003) 253–268.

[12] W. Olszewski, A. Sandroni, Manipulability of future-independent tests, Econometrica 76 (2008) 1437–1466.

[13] W. Olszewski, A. Sandroni, A nonmanipulable test, Ann. Statist. 37 (2009) 1013–1039.

[14] W. Olszewski, A. Sandroni, Strategic manipulation of empirical tests, Math. Oper. Res. 34 (2009) 57–70.

[15] A. Sandroni, The reproducible properties of correct forecasts, Int. J. Game Theory 32 (1) (2003) 151–159.

[16] A. Sandroni, R. Smorodinsky, R. Vohra, Calibration with many checking rules, Math. Oper. Res. 28 (1) (2003) 141–153.

[17] E. Shmaya, Many inspections are manipulable, Theoretical Econ. 3 (2008) 367–382.

[18] S. Sorin, Merging, reputation, and repeated games with incomplete information, Games Econ. Behav. 29 (1–2) (1999) 274–308.