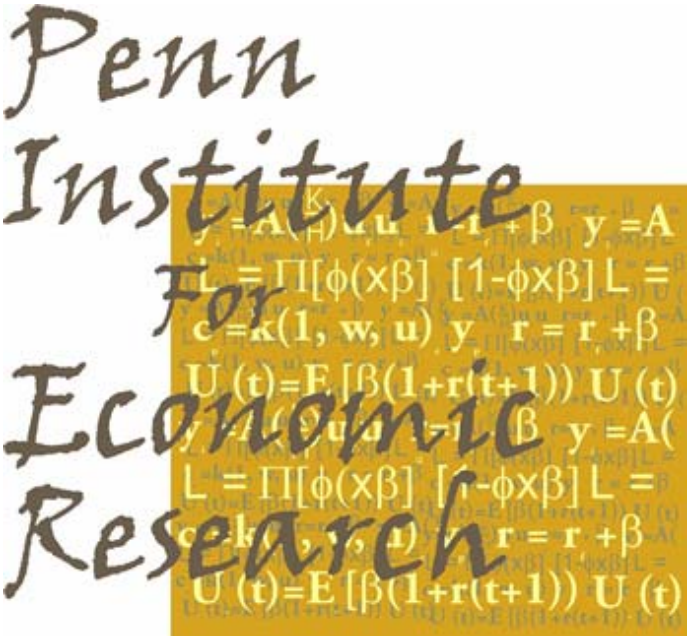

Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://www.econ.upenn.edu/pier

PIER Working Paper 08-015

"Strategic Manipulation of Empirical Tests"

by

Wojciech Olszewski and Alvaro Sandroni

http://ssrn.com/abstract=1133118

Strategic Manipulation of Empirical Tests*

Wojciech Olszewski† and Alvaro Sandroni‡

April 2008

Abstract

Theories can be produced by experts seeking a reputation for having knowledge. Hence, a tester could anticipate that theories may have been strategically produced by uninformed experts who want to pass an empirical test.

We show that, with no restriction on the domain of permissible theories, strategic experts cannot be discredited for an arbitrary but given number of periods, no matter which test is used (provided that the test does not reject the actual data-generating process).

Natural ways around this impossibility result include 1) assuming that unbounded data sets are available and 2) restricting the domain of permissible theories (opening the possibility that the actual data-generating process is rejected out of hand). In both cases, it is possible to dismiss strategic experts, but only to a limited extent. These results show significant limits on what data can accomplish when experts produce theories strategically.

*We thank the editor, Eilon Solan, an associate editor, and two referees for useful comments on an earlier draft of this paper. In particular, a comment from one of the referees improved our proof of proposition 1. We also thank Eddie Dekel, Yossi Feinberg, Nabil Al-Najjar, Dean Foster, Sandeep Baliga, Roger Myerson, and Rakesh Vohra for useful comments. All errors are ours.

†Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston IL 60208

‡Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia PA 19104, and Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston IL 60208

1. Introduction

The production and transmission of knowledge play a central role in economic activity (Hayek (1945)). Knowledge, however, is often structured as a theory which must be tested. A theory can be rejected if it makes a deterministic prediction that is not observed in the data. Nevertheless, in economics and several other disciplines, theories regularly make probabilistic forecasts that attach strictly positive probability to all outcomes. This leads to the basic question of how to test probabilistic theories. If a blunt contradiction between a theory and data is impossible, then the standard procedure is to employ large data sets so that any theory must attribute small probability to some events which, if observed, induce a rejection of the theory. It is, however, essential which low-probability events should be regarded as sufficiently incompatible with the theory to validate its rejection.

Assume that the problem at hand requires an understanding of a stochastic process which generates an outcome that can be either 0 or 1. Before any data is observed, a potential expert named Bob delivers a theory, defined as a probability measure P on the space of infinite histories. Bob may be an informed expert who truthfully reveals the data-generating process. Bob may also be a false expert who knows nothing about the data-generating process.

A tester named Alice tests Bob's theory P by selecting an event A_P (a set of outcome sequences) that she regards as consistent with the theory, and its complement A_P^c as inconsistent with it. Assume that P assigns high probability to A_P, i.e.,

$$P(A_P) \geq 1 - \varepsilon. \tag{1.1}$$

Then, if Bob's theory coincides with the data-generating process, it will not be rejected with probability $1 - \varepsilon$. When equation (1.1) is satisfied, we say that Alice's test accepts the data-generating process with probability $1 - \varepsilon$.

A vast effort has been devoted to suppling results that take the form of equation (1.1). These results (such as the law of large numbers, the law of iterated logarithmic, and the central limit theorem) relate the unobservable concept of a theory P with a potentially observable event A_P. Each of these findings can be used to define a test

that accepts the data-generating process. So, many tests are available to Alice.

Alice's task of finding a suitable test is related to the classical problem in statistics called a goodness-of-fit test problem: Given a process P and some data, a tester must determine if it is plausible that the data came from the process P. The tester must make sure that if P does, in fact, run the data, then P is not rejected. As usual in statistics, a small probability of an incorrect rejection is allowed (i.e., 1.1 must hold). However, it is essential for a test that it is capable of rejecting theories. There is only a limited purpose in running a test if we know from the outset that the theory will not be rejected.

The possibility of theory rejection is seemingly assured if, for any theory P, the complement of A_P is non-empty. Then, no matter which theory Bob announces, there is at least one path that, if realized, rejects Bob's theory. However, recent literature shows that, for several natural tests, a strategic expert can avoid rejection, *no matter how the data evolves in the future*. This can accomplished even if the test is such that for any theory P, the complement of A_P is non-empty. Assume that, before any data is observed, Bob uses a random device ζ to select his theory P. Suppose that for *any* sequence of outcomes Bob's theory P will not be rejected with arbitrarily high probability, according to Bob's randomization device ζ. No matter which data are realized, Alice will accept Bob's theory (unless Bob had an unlucky draw from ζ which is, by definition, nearly impossible). If such a device ζ can be constructed, the test is said to be *manipulable*. By definition, manipulable tests cannot dismiss strategic experts, even if the experts are completely uninformed (i.e., even if they have no knowledge over which process runs the data).

The calibration test requires the empirical frequency of 1 to be close to p in the periods in which 1 was forecasted with probability close to p. Foster and Vohra (1998) show that the calibration test can be manipulated. Several extensions of the calibration test have also been proven to be manipulable. (See, for example, Fudenberg and Levine (1999), Lehrer (2001), and Sandroni, Smorodinsky, and Vohra (2003).) Sandroni (2003), Vovk and Shafer (2005), Olszewski and Sandroni (2008)

and Shmaya (2008) show general classes of manipulable tests.[1]

Dekel and Feinberg (2006) and Olszewski and Sandroni (2008a) show the existence of a nonmanipulable test. However, these results do not determine how long it takes to discredit (uninformed) strategic experts. Let us say that rejection can be delayed for m periods if theories can be strategically generated at random in such a way that for any sequence of outcomes, the realized theory will not be rejected before period m with high probability according to the randomization device. We show that for any period m, and for any test that accepts the data-generating process with high probability, rejection can be delayed for m periods. Thus, Bob may not be able to sustain forever a false reputation for knowing the stochastic process, but he can maintain this false reputation within an arbitrarily long time horizon.

The main feature of this result is that *no* assumptions are placed on Alice's test (apart from its not rejecting the data-generating process). Even if Alice uses the nonmanipulable tests from Olszewski and Sandroni (2008a) and Dekel and Feinberg (2006), Bob can delay rejection for an arbitrarily long time, no matter which data are observed.

This impossibility result motivates an extreme recourse: we assume that the class of permissible theories is restricted and as a result some theories are excluded from the outset. The set of permissible theories constitutes a *paradigm*. We assume that a theory in the paradigm is likely to be accepted if it runs the data. Our impossibility results still hold as long as the paradigm is convex and compact (in the weak-* topology). We show that a strategic expert is likely to pass the test, no matter which process in the paradigm runs the data. The paradigm that comprises all theories is convex and compact. Other paradigms, like those consisting of exchangeable processes, are also convex and compact. So the main impossibility result holds when the expert is allowed to announce any theory, and also when the expert must announce a theory structured in the form of an exchangeable process.

[1]See Cesa-Bianchi and Lugosi (2006), Hart and Mas-Colell (2001), Lehrer and Solan (2003), Rustichini (1999), Olszewski and Sandroni (2008b), and Kalai, Lehrer and Smorodinsky (1999) for related work.

However, our impossibility result does not necessarily hold when the paradigm is not convex. We show an example of a topologically large paradigm and an empirical test, which accepts the data-generating process (provided that it is in the paradigm) and which is failed by a false expert in bounded time. No matter how the false expert randomizes, there exists at least one process in the paradigm such that, if the process runs the data, then the expert is likely to fail the test. In this paradigm, all theories are permitted except for those that are sufficiently close to a given theory f. Hence, if all that is known is that the data-generating process is in this paradigm, then the data-generating process cannot be inferred from the data. Hence, with no additional help, Alice cannot find out which process runs the data.

So far, we have implicitly assumed that Bob knows Alice's test (because Alice presents a test before Bob announces his theory). We consider zero-sum games in which Alice announces a test and Bob announces a theory simultaneously. Bob's payoff is 1 if his theory is accepted and 0 of his theory is rejected. We show that this game may have no equilibrium.

Turning to the case where Alice has unbounded data sets, the test in Olszewski and Sandroni (2008a) has the property that no matter how Bob randomizes, failure is inevitable on a set of outcome sequences that is topologically large. This suggests that strategic experts often fails this test. This depends, however, on how the word "often" is interpreted, because if we replace the topological interpretation with a measure-theoretic one, then we obtain almost the opposite result. We show that for any probability measure Q over outcome sequences, and for any test T that accepts the data-generating process with high probability, Bob can ensure that his randomly selected theory is unlikely to be rejected, on a set of outcome sequences whose $Q-$measure is as close to 1 as he wishes. Hence, even with unbounded data sets, strategic experts can be discredited only to a limited extent.

The paper is organized as follows: In section 2, we introduce our main concepts. In section 3, we show that strategic experts can delay rejection. In section 4, we explore some routes around the basic impossibility theorem. Section 5 concludes the paper. Proofs are in section 6.

2. Basic concepts

In each period one outcome, 0 or 1, is observed.[2] Let $\Omega = \{0,1\}^\infty$ be the set of all *paths*, i.e., infinite histories. A path $s \in \Omega$ is an *extension* of a history $s_t \in \{0,1\}^t$ if the first t outcomes of s coincide with the outcomes of s_t. In the opposite direction, let $s \mid t$ be the history $s_t \in \{0,1\}^t$ whose outcomes coincide with the first t outcomes of s. A *cylinder* with base on s_t is the set $C(s_t) \subset \{0,1\}^\infty$ of all infinite extensions of s_t. Let $\Im_t$ be the algebra that consists of all finite unions of cylinders with base on $\{0,1\}^t$. Denote by N the set of natural numbers. Let $\Im$ be the σ-algebra generated by the algebra $\Im^0 \equiv \bigcup_{t \in N} \Im_t$, i.e., $\Im$ is the smallest $\sigma-$algebra which contains $\Im^0$.

Let $\Delta(\Omega)$ be the set of all probability measures on $(\Omega, \Im)$. We endow Ω with the product topology (i.e., the topology that comprises unions of cylinders with a finite base) and $\Delta(\Omega)$ with the weak*-topology and with the $\sigma-$algebra of Borel sets (i.e., the smallest $\sigma-$algebra which contains all open sets in weak*-topology).[3] Let $\Delta\Delta(\Omega)$ be the set of probability measures on $\Delta(\Omega)$.

Before any data are observed, an expert named Bob announces a probability measure $P \in \Delta(\Omega)$ which (Bob claims) describes how Nature will generate the data. To simplify the language, we call a probability measure a *theory*. A tester named Alice tests Bob's theory empirically.

Definition 1. *A test is a function* $T : \Omega \times \Delta(\Omega) \to \{0,1\}$.

That is, a test is defined as an arbitrary function that takes as input a theory and a path, and returns a verdict that is 0 or 1. When the test returns a 1, it does not reject (or, simply, it accepts) the theory. When a 0 is returned, the theory is rejected.

[2] Our results generalize to any finite number of outcomes per period.

[3] The weak*-topology consists of all unions of finite intersections of sets of the form

$$\left\{Q \in \Delta(\Omega) : \left|E^P h - E^Q h\right| < \varepsilon\right\},$$

where E stands for the expected-value operator, $P \in \Delta(\Omega)$, $\varepsilon > 0$, and h is a real-valued and continuous function on Ω. See Rudin (1973).

Any test divides paths into those in $A_P \equiv \{s \in \Omega |\ T(s,P) = 1\}$, where the theory P is accepted; and those in A_P^c, where the theory is rejected. The set A_P is called the *acceptance set*, and its complement A_P^c is called the *rejection set.* We consider only tests T such that the acceptance sets A_P are $\Im-$measurable.

Definition 2. *A test T rejects a theory $P \in \Delta(\Omega)$ on a finite history $s_t \in \{0,1\}^t$ (denoted $T(s_t, P) = 0$) if $T(s,P) = 0$ for all paths s that extend s_t. Otherwise, the theory P passes the test on the history s_t, which is denoted $T(s_t, P) = 1$.*

A test thus rejects a theory on a finite history s_t if it rejects the theory on all paths s such that the first t outcomes of s coincide with the corresponding outcomes of s_t. Given any test T and a period m, let $T^m : \Omega \times \Delta(\Omega) \to \{0,1\}$ be the test such that

$$T_m(s,P) = 1 \text{ if and only if } T(s_m, P) = 1,\ s_m = s \mid m.$$

That is, T^m rejects theory P on s if and only if the test T rejects P on the first m observations of s. By definition, T^m rejects or accepts a theory at period m.

Some theories may be rejected out of hand (i.e., on all histories). The permissible theories constitute a set $\Lambda \subseteq \Delta(\Omega)$, called a *paradigm.* The excluded theories are those in Λ^c, the complement of Λ. A paradigm can consist of theories that can be efficiently computed or theories sufficiently different from a given theory (perhaps produced by another expert).[4]

Fix any $\varepsilon \in [0,1]$.

Definition 3. *A test T accepts any data-generating process in the paradigm Λ with probability $1-\varepsilon$ if for any $P \in \Lambda$,*

$$P\left(A_P\right) \geqslant 1 - \varepsilon.$$

[4]See Fortnow and Vohra (2006) for results on testing experts with computational bounds, and Al-Najjar and Weinstein (2008) and Feinberg and Stewart (2008) for results on testing multiple experts.

A test thus accepts any data-generating process in the paradigm Λ if any process in Λ that actually generates the data is likely to pass the test.

Bob is allowed to select his theory P at random. Before any data are observed, Bob may select a theory P according to a probability measure $\zeta \in \Delta\Delta(\Omega)$; we call ζ a *random generator of theories*.

Definition 4. *Fix a test T. Given a random generator of theories $\zeta \in \Delta\Delta(\Omega)$ and $\varepsilon \geq 0$, let $R_\zeta^\varepsilon \subseteq \Omega$ be the set of all paths $s \in \Omega$ such that*

$$\zeta\{P \in \Delta(\Omega) \mid T(s,P)) = 0\} \geq 1 - \varepsilon.$$

The set R_ζ^ε is called the *revelation set*; it comprises the paths on which the random generator of theories ζ fails the test with probability $1-\varepsilon$. If a test is such that, for some $\zeta \in \Delta\Delta(\Omega)$, R_ζ^ε is empty, then that test is said to be *manipulable* with probability ε. If a test accepts any data-generating process in $\Delta(\Omega)$ with probability $1-\varepsilon$, and that test is such that, for all $\zeta \in \Delta\Delta(\Omega)$, R_ζ^ε is non-empty, then the test is said to be $\varepsilon-$*effective*. So, if a test is $\varepsilon-$effective, then no matter how the expert randomizes, there exists at least one path that, if observed, rejects the expert with probability $1-\varepsilon$. Hence, effective tests are those for which it is feasible to reject an uninformed, but strategic, expert.

Given $\zeta \in \Delta\Delta(\Omega)$ and $P \in \Delta(\Omega)$, let $P\text{x}\zeta$ be product measure on $\Omega\text{x}\Delta(\Omega)$.

Definition 5. *Rejection by a test T and a paradigm Λ can be delayed for m periods with probability $1-\varepsilon$ if there exists a random generator of theories $\zeta \in \Delta\Delta(\Omega)$ such that for every theory $\tilde{P} \in \Lambda$,*

$$\tilde{P}\text{x}\zeta\{(s,P) \in \Omega x\Delta(\Omega) \mid T^m(s,P) = 1\} \geq 1 - \varepsilon.$$

Definition 6. *Rejection by a test T and a paradigm Λ can be arbitrarily delayed with probability $1-\varepsilon$ if it can be delayed for m periods, with probability $1-\varepsilon$, for every $m \in N$.*

If rejection can be arbitrarily delayed, then Bob can first choose an arbitrary period m and randomly select theories such that, with high probability (according to Bob's randomization), he will pass the test up to period m, no matter which process in the paradigm Λ runs the data. Conversely, rejection by a test T and paradigm Λ cannot be delayed for m periods with probability ε if for any random generator of theories $\zeta \in \Delta\Delta(\Omega)$, there exists a theory $\bar{P} \in \Lambda$ such that

$$\bar{P}\mathrm{x}\zeta\left\{(s,P) \in \Omega\mathrm{x}\Delta(\Omega) \mid T^m(s,P) = 1\right\} < \varepsilon.$$

3. Impossibility result

Proposition 1. *Fix $\varepsilon \in [0,1]$ and $\delta \in (0, 1-\varepsilon]$. Let Λ be a convex, compact paradigm. Let T be an arbitrary test that accepts any data-generating process in the paradigm Λ with probability $1-\varepsilon$. Then, rejection by the test T and the paradigm Λ can be arbitrarily delayed with probability $1-\varepsilon-\delta$.*

Proposition 1 shows that Bob can maintain a false reputation for knowing the data-generating process for an arbitrarily long time horizon. No matter which test Alice uses (as long as it accepts any data-generating process in a convex, compact paradigm Λ) and no matter which process in the paradigm Λ actually runs the data, Alice's test will accept Bob's theory within any given time frame. Now, assume that Bob decides on a very large number at period 0. Bob can pass Alice's test unless Alice's data set is so large that it contains more entries than the number Bob decided on at period 0.

The focus of the expert literature has been on the paradigm of all theories $\Delta(\Omega)$. It is well-known that $\Delta(\Omega)$ is convex and compact (in the weak*-topology). Hence, it follows from proposition 1 that if the test T passes any data-generating process with high probability, then rejection by the test T can be arbitrarily delayed with high probability, *no matter how the data evolves in the future.*

The set of all exchangeable processes is also convex and compact. Hence, even if the expert is restricted to announcing a theory structured in the form of an exchangeable process, and even if the expert is completely ignorant of which exchangeable

process runs the data, the expert can still arbitrarily delay rejection (provided that the test passes any exchangeable process that generates the data).

If we compare proposition 1 and the results in Dekel and Feinberg (2006) and Olszewski and Sandroni (2008a) which demonstrate the existence of effective tests, an interesting discontinuity is revealed. Suppose that Alice uses any effective test and Bob is an uninformed expert. By proposition 1, for any m, Bob can use a random generator of theories $\hat{\zeta}_m$ to delay rejection for m periods. A limit $\hat{\zeta}$ of (a subsequence of) these random generators of theories exists (because $\Delta\Delta(\Omega)$ is compact in the weak*-topology). Since the test is effective, the limit $\hat{\zeta}$ cannot delay rejection indefinitely. Moreover, it can also be shown that given any $\delta > 0$, $\hat{\zeta}$ does not delay rejection for m periods with probability δ, if m is sufficiently large. Thus, the difference between $\hat{\zeta}_m$ and $\hat{\zeta}$ becomes arbitrarily small as m increases, but a change from $\hat{\zeta}_m$ to $\hat{\zeta}$ triggers an abrupt change in delaying rejection; the rejection delay which was formerly assured on all paths with arbitrarily high probability, is no longer assured on at least some paths even with small probability. Therefore, to delay rejection, Bob must choose the random generator of theories in a very precise way.

3.1. Comparison with the literature

Proposition 1 differs fundamentally from the results we find in the existing literature. All previous contributions, place restrictions on which tests Alice can use (apart from not rejecting the data-generating process). Sandroni (2003) restricts attention to tests that accept any data-generating process *uniformly* with m data points. In addition, in this contribution, Alice is not allowed to use all the information available in a theory, but only the forecasts made along the observed history. This second restriction is also imposed in Vovk and Shafer (2005) and Shmaya (2008), and a similar restriction is imposed in Olszewski and Sandroni (2008) . As a result, Alice is not permitted to use several tests that would otherwise be available, such as those in Dekel and Feinberg (2006) or Olszewski and Sandroni (2008a). Thus, the results in the existing literature do not assure that Bob can delay rejection (as long as Alice's test accepts any data-generating process). This is demonstrated in proposition 1.

3.2. Sketch of the proof of proposition 1

The proof of proposition 1 relies on a result from Olszewski and Sandroni (2008a). Call a test *finite*, if for ever P, both the rejection set A_P^c and the acceptance set A_P are open. Olszewski and Sandroni (2008a) show that finite tests (that accept the data-generating process) can be manipulated.[5] Here is the intuition for this result:

Consider a finite test $\bar{T}$. Let $V : \Lambda \times \Delta(\Lambda) \longrightarrow [0,1]$ be a function defined by $V(P,\zeta) = E^P E^\zeta \bar{T}$, i.e., $V(P,\zeta)$ is the probability of the verdict 1 if P is the data-generating process and ζ is the random generator of theories used by Bob. By assumption, for every $P \in \Lambda$ there exists $\zeta_P \in \Delta(\Lambda)$ (a deterministic generator of theories that assigns probability 1 to P) such that $V(P,\zeta_P) \geq 1-\varepsilon$. Thus, if the conditions of Fan's minmax theorem are satisfied, then there exists as well $\zeta_{\bar{T}} \in \Delta\Delta(\Omega)$ such that $V(P,\zeta_{\bar{T}}) \geq 1-\varepsilon-\delta$ for every $P \in \Lambda$. The assumption that the test is finite guarantees that V is a continuous function of P. All other conditions of Fan's minmax theorem (Fan (1953)) are satisfied.

Now consider an arbitrary test T. The test T^m is finite for every period m, and accepts any data-generating process. Hence, for each test T^m, there exists a random generator of theories ζ_m that is likely to pass the test T^m, no matter which process in Λ generates the data. So, by the construction of the test T^m, rejection by the test T can be delayed for m periods.

4. Ways around the impossibility result

Proposition 1 shows that no test (that passes any data-generating process in a convex, compact paradigm Λ) can dismiss a strategic expert with bounded data sets. Proposition 1 is an impossibility result, which provides motivation for investigating ways to get around it. We consider three possible routes. In section 4.1, we consider nonconvex paradigms. In section 4.2 we relax the (implicit) assumption that Bob knows Alice's test. In section 4.3., we consider the case of unbounded data sets.

[5]Olszewski and Sandroni (2008a) prove that a slightly larger class of tests is manipulable. However, for the purposes of the present paper, we need to know only that finite tests are manipulable.

4.1. Non-convex paradigms

It is fairly easy to see that proposition 1 does not extend to the case of non-convex paradigms.

- Consider a paradigm $\Lambda_D = \{P_1, P_2\}$ with two theories. Suppose that there exist disjoint sets A_{P_1} and A_{P_2} satisfying (1.1), and comprising histories of length m. (It is straightforward to produce examples of two theories and two sets with these properties.) Let T be the test such that A_{P_1} and A_{P_2} are the acceptance sets of P_1 and P_2, respectively, and all other theories are rejected on all paths. Then, T accepts any data-generating process in paradigm Λ_D, and if the expert does not pick the actual data-generating process, he fails T with probability $1 - \varepsilon$. Also, every random generator of theories must assign less than a 0.5 chance to one of the theories in Λ_D. As a result, rejection by this test and paradigm Λ_D cannot be delayed for m periods with probability higher than $0.5 + 0.5\varepsilon$.

The paradigm Λ_D consisting of only two theories is in striking contrast with the paradigm $\Lambda = \Delta(\Omega)$ of all possible theories. In the latter case, rejection can be arbitrarily delayed. In the former case, it cannot. One might wonder whether this difference arises the fact that the paradigm Λ_D is small, whereas the paradigm $\Lambda = \Delta(\Omega)$ is large. Nevertheless, we will now construct a more elaborate example in which rejection cannot be delayed even though theories are restricted to a (topologically) large paradigm.

4.1.1. A large nonconvex paradigm

Given a theory $P \in \Delta(\Omega)$, a path $s \in \Omega$, and $s_t = s \mid t$ such that $P(C(s_t)) > 0$, let

$$f_0^P(s) \equiv P(C(1)), \text{ and } f_t^P(s) \equiv \frac{P(C(s_t, 1))}{P(C(s_t))}$$

be forecasts made along s.

Given $\delta \in (0,\, 0.25)$ and $m \in N$, let $\Lambda_{m,\delta} \subseteq \Delta(\Omega)$ be the paradigm of theories P such that for all paths $s \in \Omega$ such that $P(C(s_m)) > 0$,

$$\frac{1}{m}\sum_{t=1}^{m}(f_{t-1}^{P}(s) - 0.5)^2 > \delta.$$

The paradigm $\Lambda_{m,\delta}$ excludes theories forecasting 1 and 0 with near equal odds sufficiently often. Let $\bar{P}$ be the theory that always forecasts 1 with probability 0.5. Let $\beta = (K, m, \delta) \in N^2\mathrm{x}(0, 1]$ be a triple of parameters. Let T^β be the test

$$T^\beta(s, P) = \left\{ \begin{array}{ll} 1 & \text{if } P \in \Lambda_{m,\delta} \text{ and } P(C(s_m)) \geq K\bar{P}(C(s_m)),\ s_m = s \mid m; \\ 0 & \text{otherwise.} \end{array} \right\}$$

The test T^β passes a theory P from the paradigm $\Lambda_{m,\delta}$ on all histories along which P is found K times more likely than $\bar{P}$. In particular, $T^\beta(s, P) = 0$ if $P(C(s_m)) = 0$. Theories outside the paradigm are rejected. Let $\widetilde{\Lambda}_{m,\delta}$ be the set of theories $P \in \Lambda_{m,\delta}$ such that $P(C(s_m)) > 0$ for any cylinder $C(s_m)$, i.e., theories never predicting up to period m any outcome with certainty.

Proposition 2. *For any $\varepsilon > 0$, $\delta \in (0, .25)$, and $K \in N$, there exists a period $\bar{m} \in N$ such that if $m \geq \bar{m}$, then:*

1) The test T^β passes any data-generating process in paradigm $\Lambda_{m,\delta}$ with probability $1 - \varepsilon$.

2) Rejection by a test T^β and paradigm $\Lambda_{m,\delta}$ cannot be delayed for m periods with probability ε.

In addition, the set $\widetilde{\Lambda}_{m,\delta} \subset \Lambda_{m,\delta}$ is an open subset of $\Delta(\Omega)$, and given any theory $P \in \Delta(\Omega)$ and a neighborhood U of P, there exist $\hat{m} \in N$ and a theory Q such that $Q \in U \cap \widetilde{\Lambda}_{m,\delta}$ for every $m \geq \hat{m}$.

Assume that Alice will eventually have m data points at her disposal and that she tests Bob with the test T^β. Also assume that the data-generating process belongs to the paradigm $\Lambda_{m,\delta}$ (and that this is known to Alice and Bob). If informed, Bob knows

which process in the paradigm $\Lambda_{m,\delta}$ generates the data. But if Bob is uninformed, he does not know which process in the paradigm $\Lambda_{m,\delta}$ generates the data. Proposition 2 shows that if Bob is informed, he is likely to pass T^{β}. If uninformed, Bob cannot be assured that he will arbitrarily delay rejection, because no matter how he randomizes, for some theories inside the paradigm, he fails the test with high probability. In addition, proposition 2 shows that $\widetilde{\Lambda}_{m,\delta}$ is an open set and that any open set intersects $\widetilde{\Lambda}_{m,\delta}$ if m is sufficiently large. Hence, the sets $\Lambda_{m,\delta}$ become topologically large, if m becomes large. That is, even if Alice and Bob are relatively ill-informed (i.e., they only know that the data-generating process belongs to a topologically large set), then it is not possible for Bob, without additional information, to be nearly certain that rejection can be arbitrarily delayed.

Notice that if Bob announces the data-generating process, then Alice benefits from Bob's announcement. Given that she only knows that the actual process belongs to $\Lambda_{m,\delta}$, she would not be able to infer the process from the data without additional information.

4.1.2. Intuition of proposition 2

The paradigm $\Lambda_{m,\delta}$ excludes the theory $\bar{P}$ and other theories which forecast 1 and 0 with near equal odds sufficiently often. Hence, the paradigm $\Lambda_{m,\delta}$ excludes only a relatively small set of theories. It is therefore intuitive that the paradigm $\Lambda_{m,\delta}$ is a topologically large set. The intuition for parts 1 and 2 of proposition 2 is as follows:

Consider a theory P that belongs to the paradigm $\Lambda_{m,\delta}$. The histories to which P assigns a sufficiently higher likelihood than $\bar{P}$ does have a high probability, according to P. By definition, if Bob announces P, then he passes the test T^{β} on paths to which P assigns a sufficiently higher likelihood than $\bar{P}$ does. So if the actual data-generating process is P, then Bob is likely to pass T^{β}.

On the other hand, if Bob announces theory P, he fails the test on the histories to which $\bar{P}$ assigns a sufficiently higher likelihood than P does. The histories with this property have a high probability according to $\bar{P}$. Assume that $\bar{P}$ is the data-generating process. Then, no matter which theory P (in the paradigm) Bob

announces, he fails the test with high probability (according to $\bar{P}$). Theories outside the paradigm are rejected out of hand.

Hence, no matter which random generator of theories ζ Bob uses, he must announce a theory which fails the test with high probability (according to $\bar{P}$). By Fubini's theorem, there exists a history s such that Bob is likely to fail test T^β on s (according to ζ). Now consider a theory $\hat{P}$ in the paradigm $\Lambda_{m,\delta}$ that assigns high probability to s (for example, a Dirac measure centered at s). It follows that if $\hat{P} \in \Lambda_{m,\delta}$ is the data-generating process, then Bob is likely to fail test T^β.

4.2. Simultaneous moves

The analysis considered so far could be embedded in a game where Alice moves first and presents a test. After observing Alice's test, Bob announces a theory. Finally, Nature produces the data. This zero-sum game between Alice and Bob can be defined as follows: Fix any arbitrary $m \in N$. Given $\varepsilon > 0$, let $\Upsilon(\varepsilon)$ be the set of all tests that pass any data-generating process with probability $1 - \varepsilon$. Alice chooses a test $T \in \Upsilon(\varepsilon)$ and Bob chooses a random generator of theories $\zeta \in \Delta\Delta(\Omega)$. Bob's payoff is

$$V_m(\zeta, T) = \inf_{s_m \in \{0,1\}^m} \zeta \left\{ P \in \Delta(\Omega) \mid T(s_m, P) = 1 \right\}. \tag{4.1}$$

That is, Bob's payoff is the probability that his theory will pass the test at period m, computed under a worse-case scenario over the outcome sequences that Nature might produce. By proposition 1, if Alice moves first, then Bob can assure him a payoff close arbitrarily close to $1 - \varepsilon$. Now, assume that Alice can select a test at random by $\theta \in \Delta(\Upsilon(\varepsilon))$. An argument entirely analogous to the one presented in the proof of proposition 1 shows that for every mixed strategy of Alice, there exists a strategy of Bob that also assures him a payoff close to $1 - \varepsilon$. That is, for any $\theta \in \Delta(\Upsilon(\varepsilon))$ and $\delta \in (0, 1 - \varepsilon]$, there exists $\zeta \in \Delta\Delta(\Omega)$ such that

$$\inf_{s_m \in \{0,1\}^m} E^\theta \zeta \left\{ P \in \Delta(\Omega) \mid T(s_m, P) = 1 \right\} \tag{4.2}$$

is greater than $1 - \varepsilon - \delta$. Hence, if Bob correctly anticipates Alice's mixed strategy, then Alice cannot determine whether Bob has any relevant knowledge about the

data-generating process.

Now consider a zero-sum game in which (uninformed) Bob and Alice move simultaneously so that theories and tests are announced at the same time. Bob's payoffs are given by either (4.1) or (4.2), depending on whether Alice is allowed to randomize. Alice's pure strategies set is $\Upsilon(\varepsilon)$. This game may have no equilibrium.

Example 1. *Fix $\varepsilon = 5/8$ and $m = 2$. For any random generator of theories $\zeta \in \Delta\Delta(\Omega)$ there is a test T_ζ such that Bob's payoff $V_2(\zeta, T_\zeta)$ is smaller than or equal to* $2/8$.

The proof of this example is in section 6. By proposition 1, if Bob properly anticipates Alice's strategy, then he ensures himself at least a payoff close to $3/8$. By example 1, if Alice can properly anticipate Bob's strategy, then Bob gets at most $2/8$. Hence, the game has no equilibrium. In contrast to the case of a known probability distribution over tests, these results imply that Bob, if uninformed, cannot simultaneously pass all tests from $\Upsilon(5/8)$ with probability arbitrarily close to $3/8$.

4.3. Effectiveness bound

We return to the case in which Alice announces the test first, but we now assume that she has unbounded data sets at her disposal. By definition, if Bob uses a random generator of theories ζ, then he fails the test with probability $1-\varepsilon$, on the revelation set R^{ε}_{ζ}. These revelation sets are empty for some random generator of theories ζ in the case of the calibration test or other manipulable tests. In contrast, for the Dekel and Feinberg (2006) test, the revelation sets are uncountable (even for $\varepsilon = 0$) for any random generator of theories ζ. For this latter test, the revelation sets are large in a set-theoretic sense. And if we consider the test in Olszewski and Sandroni (2008a), the complement of the revelation sets (also for $\varepsilon = 0$) are Baire's first-category sets. For this test, then, the revelation sets are large in a topological sense.[6]

[6]First-category sets are often described as topologically small. However, there are other definitions of small sets that we have not examined here (see Anderson and Zame (2001) and Stinchcombe (2001)).

It would be desirable to extend these results by showing a test with revelation sets that are large in a measure-theoretic sense, i.e., revelation sets that are guaranteed to have nonnegligible measure according to a given probability measure. However, this is not possible, as we will now show.

Given a random generator of theories $\zeta \in \Delta\Delta(\Omega)$, let $A_\zeta^{1-\varepsilon} \subseteq \Omega$ be the set of paths such that

$$\zeta\{P \in \Delta(\Omega) : T(s,P) = 1\} > 1 - \varepsilon. \tag{4.3}$$

The set $A_\zeta^{1-\varepsilon}$ comprises the paths on which the random generator of theories ζ passes the test with probability greater than $1-\varepsilon$. These sets are called $\zeta-$*approval sets.*

Proposition 3. *Fix a probability measure $Q \in \Delta(\Omega)$. Fix also any real numbers $\varepsilon \in (0,1]$ and $\delta \in (0, 1-\varepsilon]$. Consider a test T that passes any data-generating process with probability $1-\varepsilon$. For every real number $\nu > 0$, there exists a random generator of theories $\zeta \in \Delta\Delta(\Omega)$ such that*

$$Q(A_\zeta^{1-\varepsilon-\delta}) \geq 1 - \nu. \tag{4.4}$$

Proposition 3 shows that given any probability measure, Bob can ensure that his theories will pass the test on approval sets that have high probability according to this probability measure. This result holds no matter which test Alice uses (and no matter how much data she has) provided that the test passes any data-generating process.

Here is a sketch of the proof of proposition 3: Consider an arbitrary test T that is likely to accept any data-generating process. The test T can be approximated by a finite test $\overline{T}$ such that the differences between the rejection sets A_P^c of T and $\overline{T}$, respectively, are small according to probability measure Q. Since $\overline{T}$ is a finite test, it is manipulable.

Any random generator of theories ζ that passes $\overline{T}$ on all paths can fail T only on specific paths s; namely, ζ must assign a large measure to probability measures P for which s belongs to the difference between the rejection sets A_P^c of T and $\overline{T}$. The set

of these paths s must, however, be small according to probability measure Q, by the property defining the test $\overline{T}$ and Fubini's theorem.

Remark 1. *It is straightforward to modify the proof of Proposition 3 to obtain this slightly stronger result: For every finite family of probability measures $\mathcal{Q}$, there exists a random generator of theories $\zeta \in \Delta\Delta(\Omega)$ satisfying (4.4) for every $Q \in \mathcal{Q}$.*

5. Conclusion

If an empirical examiner plans to reject false theories, then it must be possible for her to reject theories which are based on no relevant knowledge. Empirical tests with this property are nonmanipulable tests. Any test that accepts the data-generating process is susceptible to strategic manipulation for arbitrarily long periods of time. Even if a tester has arbitrarily large data sets at her disposal, she will only be able to discredit a strategic expert in a limited sense.

It is possible to arbitrarily delay rejection even if one knows from the outset that the data-generating process belongs to a convex, compact paradigm such as the class of all exchangeable processes. However, it may not be possible to arbitrarily delay rejection if theories forecasting 1 and 0 with near equal odds sufficiently often are excluded.

6. Proofs

The proofs apply an assertion from Olszewski and Sandroni (2008a). It is convenient to restate that assertion here:

Definition 7. *Finite tests of length m are defined by the property that for any theory $P \in \Delta(\Omega)$, the rejection set A_P^c is a union of cylinders with base on histories of length $t \leq m$. A test is called finite if it is a finite test on length m for some $m \in N$.*

Note that a finite test of length m can be equivalently defined by the property that for any theory $P \in \Delta(\Omega)$, the acceptance set A_P is a union of cylinders with base on histories of length $t \leq m$.

Proposition 5 from Olszewski and Sandroni (2008a). Fix any $\varepsilon \in (0,1]$ and $\delta \in (0, 1-\varepsilon]$. Let Λ be a convex, compact paradigm. Let T be an arbitrary test that accepts the data-generating process in the paradigm Λ with probability $1-\varepsilon$. Then, there exists a random generator of theories $\zeta \in \Delta\Delta(\Omega)$ such that for every theory $\tilde{P} \in \Lambda$,

$$\tilde{P}\mathrm{x}\zeta\left\{(s,P) \in \Omega\mathrm{x}\Delta(\Omega) \mid T(s,P)=1\right\} \geq 1-\varepsilon-\delta.^{7}$$

Proof of proposition 1: Fix any period m. By proposition 5 in Olszewski and Sandroni (2008a), there exists a random generator of theories $\zeta_{T,m}$ such that for every $\tilde{P} \in \Lambda$,

$$\tilde{P}\mathrm{x}\zeta_{T,m}\left\{(s,P) \in \Omega\mathrm{x}\Delta(\Omega) \mid T^m(s,P)=0\right\} \leq \varepsilon+\delta.$$

Let s_m be any finite history from $\{0,1\}^m$ such that $T(s_m,P)=0$. By definition, $C(s_m) \subseteq A_P^c$ and therefore $C(s_m) \subseteq (A_P^m)^c$. In other words, $T(s_m,P)=0$ implies that $T^m(s_m,P)=0$. Thus,

$$\left\{P \in \Delta(\Omega) : T(s_m,P)=0\right\} \subseteq \left\{P \in \Delta(\Omega) : T^m(s_m,P)=0\right\},$$

and so for every $\tilde{P} \in \Lambda$, $\tilde{P}\mathrm{x}\zeta_{T,m}\left\{(s,P) \in \Omega\mathrm{x}\Delta(\Omega) \mid T(s_m,P)=0\right\} \leq \varepsilon+\delta$.■

We now state and prove three lemmas that will be used in the proof of proposition 2.

Lemma 1. *There exists $\kappa \in \Re_+$ such that*

$$1)\ \kappa\left[p\log\left(\frac{p}{0.5}\right)+(1-p)\log\left(\frac{1-p}{0.5}\right)\right] \geq (p-0.5)^2 \text{ for every } p \in [0,1];$$

and

$$2)\ \kappa\left[0.5\log\left(\frac{0.5}{p}\right)+0.5\log\left(\frac{0.5}{1-p}\right)\right] \geq (p-0.5)^2 \text{ for every } p \in [0,1].$$

Proof: We shall first prove part 1. Let

$$E_1(p) = p \log\left(\frac{p}{0.5}\right) + (1-p)\log\left(\frac{1-p}{0.5}\right).$$

With some algebra, it follows that $E_1(p)$ is a positive function (on $[0,1]$) and zero if and only if $p = 0.5$. Taking l'Hospital's rule, twice it follows that

$$\frac{E_1(p)}{(p-0.5)^2} \underset{p\to 0.5}{\longrightarrow} 2.$$

Let

$$J(p) = \left\{ \begin{array}{ccc} 2 & if & p = 0.5 \\ \frac{E_1(p)}{(p-0.5)^2} & if & p \neq 0.5 \end{array} \right\}.$$

Hence, $J(p)$ is a continuous and strictly positive function (on $[0,1]$); in particular, $J(p)$ is a bounded away from zero on $[0,1]$.

The proof of part 2 is completely analogous to the proof of part 1.

$$E_2(p) = 0.5\log\left(\frac{0.5}{p}\right) + 0.5\log\left(\frac{0.5}{1-p}\right)$$

is also a positive function (on $[0,1]$) and zero if and only if $p = 0.5$. Moreover,

$$\frac{E_2(p)}{(p-0.5)^2} \underset{p\to 0.5}{\longrightarrow} 2.$$

■

Let E^P and VAR^P be the expectation and variance operator associated with $P \in \Delta(\Omega)$. Let $(X_i)_{i=1}^{\infty}$ be a sequence of random variables such that X_i is $\Im_i$-measurable and its expectation conditional on $\Im_{i-1}$ is zero (i.e., $E^P\{X_i \mid \Im_{i-1}\} = 0$). Moreover, the sequence of conditional variances $VAR^P\{X_i \mid \Im_{i-1}\}$ are uniformly bounded (i.e., $VAR^P\{X_i \mid \Im_{i-1}\} < M$ for some $M > 0$). We define

$$S_m := \sum_{i=1}^{m} X_i \text{ and } Y_m := \frac{S_m}{m}.$$

Lemma 2. *For every $\varepsilon' > 0$ and $j \in N$, there exists $\bar{m}(j,\varepsilon') \in N$ such that*

$$P\left(\left\{s \in \Omega : \forall_{m \geq \bar{m}(j,\varepsilon')} \quad |Y_m(s)| \leq \frac{1}{j}\right\}\right) > 1 - \varepsilon'.$$

Proof: By definition, S_m is a martingale. By Kolmogorov's inequality (see Shiryaev (1996), Chapter IV, §2), for any $\delta > 0$,

$$P\left(\left\{s \in \Omega : \max_{1\leq m\leq k} |S_m(s)| > \delta\right\}\right) \leq \frac{Var(S_k)}{\delta^2} \leq \frac{kM}{\delta^2}.^{8}$$

Let $M_n := \max_{2^n<m\leq 2^{n+1}} Y_m$. Then,

$$P\left(\left\{s \in \Omega : M_n(s) > \frac{1}{j}\right\}\right) \leq P\left(\left\{s \in \Omega : \max_{2^n<m\leq 2^{n+1}} |S_m(s)| > \frac{1}{j}2^n\right\}\right) \leq$$

$$\leq P\left(\left\{s \in \Omega : \max_{1\leq m\leq 2^{n+1}} |S_m(s)| > \frac{1}{j}2^n\right\}\right) \leq 2Mj^2\frac{2^n}{4^n} = 2Mj^2\frac{1}{2^n}.$$

Therefore,

$$\sum_{n=m^*}^{\infty} P\left(\left\{s \in \Omega : M_n(s) > \frac{1}{j}\right\}\right) \leq 2Mj^2 \sum_{n=m^*}^{\infty} \frac{1}{2^n} < \varepsilon' \text{ (for a sufficiently large } m^*).$$

Let $\bar{m}(j, \varepsilon') = 2^{m^*}$ for this sufficiently large m^*. By definition,

$$\left\{s \in \Omega : \forall_{m\geq\bar{m}(j,\varepsilon')} \quad |Y_m(s)| \leq \frac{1}{j}\right\}^c \subseteq \bigcup_{n=m^*}^{\infty} \left\{s \in \Omega : M_n(s) > \frac{1}{j}\right\}.$$

Hence,

$$P\left(\left\{s \in \Omega : \forall_{m\geq\bar{m}(j,\varepsilon')} \quad |Y_m(s)| \leq \frac{1}{j}\right\}\right) > 1 - \varepsilon'.$$

■

Given $P \in \Delta(\Omega)$ and $s \in \Omega$, let

$$h_{t-1}^P(s) = f_{t-1}^P(s)^{s^t}(1 - f_{t-1}^P(s))^{1-s^t},$$

i.e. $h_{t-1}^P(s)$ is the forecast associated with the actual outcome in period t. Let

$$Z_t(s) = \log\left(\frac{h_{t-1}^P(s)}{0.5}\right) \text{ and } \bar{Z}_t = Z_t - E^P\left\{Z_t \mid \Im_{t-1}\right\}.$$

Lemma 3. *Given $K \in N$, $\varepsilon > 0$, and $\delta > 0$, there exists $\bar{m}$ such that if $m \geq \bar{m}$, then for every $P \in \Lambda_{m,\delta}$,*

$$1)\ P\left\{\sum_{t=1}^{m} Z_t(s) \geq \log(K)\right\} \geqslant 1-\varepsilon \text{ and } 2)\ \bar{P}\left\{\sum_{t=1}^{m} Z_t(s) < \log(K)\right\} \geqslant 1-\varepsilon.$$

Proof: By definition,

$$\frac{1}{m}\sum_{t=1}^{m} Z_t = \frac{1}{m}\sum_{t=1}^{m} \bar{Z}_t + \frac{1}{m}\sum_{t=1}^{m} E^P\left\{Z_t \mid \Im_{t-1}\right\}.$$

By Lemma 1 (part 1),

$$\frac{1}{m}\sum_{t=1}^{m} E^P\left\{Z_t \mid \Im_{t-1}\right\} \geq \frac{1}{\kappa}\frac{1}{m}\sum_{t=1}^{m}(f_{t-1}^P(s)-0.5)^2.$$

So if $P \in \Lambda_{m,\delta}$, then

$$\frac{1}{m}\sum_{t=1}^{m} E^P\left\{Z_t \mid \Im_{t-1}\right\} > \frac{\delta}{\kappa}.$$

By definition,

$$\sum_{t=1}^{m} Z_t(s) \geq \log(K) \Longleftrightarrow \frac{1}{m}\sum_{t=1}^{m} \bar{Z}_t + \frac{1}{m}\sum_{t=1}^{m} E^P\left\{Z_t \mid \Im_{t-1}\right\} \geq \frac{\log(K)}{m}.$$

Take any $j \geq 2\kappa/\delta$ and $\bar{m}(j,\varepsilon)$ as defined in Lemma 2. Next, take $\hat{m} \geq \bar{m}(j,\varepsilon)$ such that if $m \geq \hat{m}$, then

$$\frac{\log(K)}{m} < \frac{\delta}{2\kappa}.$$

So if $m \geq \hat{m}$, then

$$P\left\{\sum_{t=1}^{m} Z_t(s) \geq \log(K)\right\} \geq P\left\{\frac{1}{m}\sum_{t=1}^{m} \bar{Z}_t > -\frac{\delta}{2\kappa}\right\} \geq P\left\{\frac{1}{m}\sum_{t=1}^{m} \bar{Z}_t > -\frac{1}{j}\right\},$$

and this last probability is greater than $1-\varepsilon$ by Lemma 2. This demonstrates the first part of Lemma 3. The proof of the second part of Lemma 3 is analogous to the proof of the first part. Let us define

$$W_t(s) = -Z_t(s) \text{ and } \bar{W}_t = W_t - E^{\bar{P}}\left\{W_t \mid \Im_{t-1}\right\}.$$

By definition,

$$\frac{1}{m}\sum_{t=1}^{m} W_t = \frac{1}{m}\sum_{t=1}^{m} \bar{W}_t + \frac{1}{m}\sum_{t=1}^{m} E^{\bar{P}}\left\{W_t \mid \Im_{t-1}\right\}.$$

By Lemma 1 (part 2),

$$\frac{1}{m}\sum_{t=1}^{m} E^{\bar{P}}\left\{W_t \mid \Im_{t-1}\right\} \geq \frac{1}{\kappa}\frac{1}{m}\sum_{t=1}^{m}(f_{t-1}^{P}(s) - 0.5)^2.$$

So if $P \in \Lambda_{m,\delta}$, then

$$\frac{1}{m}\sum_{t=1}^{m} E^{\bar{P}}\left\{W_t \mid \Im_{t-1}\right\} > \frac{\delta}{\kappa}.$$

By definition,

$$\sum_{t=1}^{m} Z_t(s) < \log(K) \Longleftrightarrow \sum_{t=1}^{m} W_t(s) > -\log(K) \Longleftrightarrow$$

$$\frac{1}{m}\sum_{t=1}^{m} \bar{W}_t + \frac{1}{m}\sum_{t=1}^{m} E^{\bar{P}}\left\{W_t \mid \Im_{t-1}\right\} > \frac{-\log(K)}{m}.$$

Take any $j \geq \kappa/\delta$ and $\bar{m}(j,\varepsilon)$ as defined in Lemma 2. If $m \geq \widetilde{m} \equiv \bar{m}(j,\varepsilon)$, then

$$\bar{P}\left\{\sum_{t=1}^{m} Z_t(s) < \log(K)\right\} \geq \bar{P}\left\{\frac{1}{m}\sum_{t=1}^{m} \bar{W}_t > -\frac{\delta}{\kappa}\right\} > 1-\varepsilon.$$

The proof is now concluded by defining $\bar{m}$ as $max\{\widetilde{m}, \hat{m}\}$.■

Lemma 4. *For any $\varepsilon > 0$, $\delta > 0$, and $K \in N$, there exists a period $\bar{m} \in N$ such that if $m \geq \bar{m}$, then the test T^{β} accepts a data-generating process in $\Lambda_{m,\delta}$ with probability $1-\varepsilon$. Moreover, if $m \geq \bar{m}$, then for any random generator of theories $\zeta \in \Delta\Delta(\Omega)$,*

$$\bar{P}x\zeta\left\{(s,P) \in \Omega x\Delta(\Omega) \mid T^{\beta}(s,P) = 1\right\} \leq \varepsilon. \tag{6.1}$$

Proof: By definition,

$$\log\left(\frac{P(C(s_m))}{\bar{P}(C(s_m))}\right) = \sum_{t=1}^{m} Z_t(s), s_m = s \mid m.$$

So

$$P\left\{s \in \Omega \mid T^{\beta}(s,P)=1\right\} = P\left\{\sum_{t=1}^{m} Z_t(s) \geq \log(K)\right\}.$$

By Lemma 3 (part 1), if $m \geq \bar{m}$, then T^{β} accepts any data-generating process in $\Lambda_{m,\delta}$ with probability $1-\varepsilon$.By Lemma 3 (part 2), if $m \geq \bar{m}$, then for every $P \in \Lambda_{m,\delta}$,

$$\bar{P}\left\{s \in \Omega \mid T^{\beta}(s,P)=0\right\} \geqslant 1-\varepsilon.$$

In addition, $T^{\beta}(s,P)=0$ if $P \in (\Lambda_{m,\delta})^{c}$. So, $E^{\bar{P}}\left\{T^{\beta}(\cdot\,,P)\right\} \leq \varepsilon$ for every $P \in \Delta(\Omega)$. It follows that given any $\zeta \in \Delta\Delta(\Omega)$,

$$E^{\bar{P}\mathrm{x}\zeta}\left\{T^{\beta}\right\} = E^{\zeta}E^{\bar{P}}\left\{T^{\beta}\right\} \leq \varepsilon.$$

Hence,

$$E^{\bar{P}\mathrm{x}\zeta}\left\{T^{\beta}\right\} = \bar{P}\mathrm{x}\zeta\left\{P \in \Delta(\Omega),\ s \in \Omega \mid T^{\beta}(s,P)=1\right\} \leq \varepsilon.$$

■

Proof of proposition 2: Part 1) of proposition 2 is shown in lemma 4. Part 2) of proposition 2 follows from lemma 4 because if (6.1) holds, then $E^{\bar{P}}E^{\zeta}\{T^{\beta}\} \leq \varepsilon$. So there must exist at least one path $\tilde{s} \in \Omega$ such that $E^{\zeta}\{T^{\beta}(\tilde{s},P)\} \leq \varepsilon$. It follows that the theories produced by ζ fail T^{β} with probability $1-\varepsilon$, provided that the data is given by $\tilde{s}$; or, equivalently, that $\tilde{s}$ is produced by the Dirac measure that assigns full measure to $\tilde{s}$. If $\delta < 0.25$, then any Dirac measure is in $\Lambda_{m,\delta}$.

The final part of proposition 2 can be shown as follows:

We shall show first that the set $\widetilde{\Lambda}_{m,\delta}$ is open. Take any probability measure $P \in \widetilde{\Lambda}_{m,\delta}$. For any $s \in \Omega$ and $t \leq m$, define functions $h_{s,t} : \Omega \to R$ by

$$h_{s,t}(r) = 1 \text{ if } r_t = s_t, \text{ and } h_{s,t}(r) = 0 \text{ otherwise.}$$

Thus, if

$$\forall_{s\in\Omega} \quad \left|E^{P}h_{s,t} - E^{Q}h_{s,t}\right| < \varepsilon_t, \tag{6.2}$$

then the measures assigned by P and Q to the cylinder $C(s_t)$ are closer than ε_t. It suffices to pick (sufficiently small) numbers ε_t recursively in order to guarantee

that any probability measure Q satisfying inequality (6.2) for all $t \leq m$ has the property that $f^P_{t-1}(s)$ and $f^Q_{t-1}(s)$ are arbitrarily close, which (by definition) implies that $Q \in \widetilde{\Lambda}_{m,\delta}$.

We shall now show that given any theory $P \in \Delta(\Omega)$ and any neighborhood U of P, there exist $\hat{m} \in N$ and a theory Q such that $Q \in U \cap \widetilde{\Lambda}_{m,\delta}$ for every $m \geq \hat{m}$. With no loss of generality, assume that $P(C(s_m)) > 0$ for any cylinder $C(s_m)$. Take continuous functions $h_1, ..., h_l : \Omega \to R$, and positive real numbers $\varepsilon_1, ..., \varepsilon_l$. It follows from the continuity of $h_1, ..., h_l$ and the compactness of Ω that there exists a (large enough) $k \in N$ such that

$$r_k = s_k \Longrightarrow \forall_{i=1,...,l} \quad |h_i(r) - h_i(s)| < \varepsilon_i.$$

Thus, if two probability measures P and Q have the property that

$$\forall_{s \in \Omega}\ P(C(s_k)) = Q(C(s_k)), \tag{6.3}$$

then

$$\forall_{i=1,...,l} \quad \left|E^P h_i - E^Q h_i\right| < \varepsilon_i.$$

Take a probability measure Q satisfying (6.3) and the following property: for any $s \in \Omega$ and $t = k+1, ..., m-1$,

$$f^Q_t(s) = q \text{ for some } q \in (\overline{\delta}, 1), \tag{6.4}$$

where

$$\left(\overline{\delta} - 0.5\right)^2 = \delta.$$

Then, by (6.4), $Q \in \Lambda_{m,\delta}$ if m is sufficiently large; by (6.3) and (6.4), $Q(C(s_m)) > 0$ for any cylinder $C(s_m)$, and so $Q \in \widetilde{\Lambda}_{m,\delta}$. Finally, by (6.3), Q belongs to the neighborhood of the probability measure P determined by $h_1, ..., h_l$ and $\varepsilon_1, ..., \varepsilon_l$.■

Proof of example 1: Fix any $\zeta \in \Delta\Delta(\Omega)$. Given any history $i \in \{0,1\}^2$, let $T^i \in \Upsilon(5/8)$ be the test such that $T^i(j, P) = 1$ for every $j \in \{0,1\}^2, j \neq i; T^i(i, P) = 1$ if $P(C(i)) \geq 5/8$, and $T^i(i, P) = 0$ if $P(C(i)) < 5/8$. Let $D_i \subseteq \Delta(\Omega)$ be the set of

theories P such that $P(C(i)) \geq 5/8$. Clearly, D_l and D_k are disjoint sets, $l \neq k$. So, for some $\bar{\imath} \in \{0,1\}^2$, $V_2(\zeta, T^{\bar{\imath}}) = \zeta(D_{\bar{\imath}}) \leq 2/8$.■

Proof of proposition 3: Recall the following well-known result: for any given probability measure $P \in \Delta(\Omega)$ and $\delta > 0$, any set $A \in \Im$ can be enlarged to an open set $U \supset A$ such that $P(U) < P(A) + \delta$ (see Ulam's Theorem, 7.1.4 in Dudley (1989)). It implies that for every probability measure P, the rejection set A_P^c can be enlarged to an open set B_P^c such that

$$P(B_P^c - A_P^c) \leq \frac{\delta}{4}. \tag{6.5}$$

Let B_P^n denote the union of those cylinders $C \subset B_P^c$ whose base has length n. For every P, if n is sufficiently large, then

$$Q(B_P^c - B_P^n) \leq \nu \cdot \frac{\delta}{2}.$$

Take any n with this property, and define a test $\overline{T}$ by

$$\overline{T}(s, P) = 0 \text{ iff } s \in B_P^n.$$

By (6.5), the test $\overline{T}$ accepts the data-generating process with probability $1 - \varepsilon - \delta/4$. So, by proposition 5 in Olszewski and Sandroni (2008a), there exists a random generator of theories $\zeta \in \Delta\Delta(\Omega)$ such that for every $s \in \Omega$,

$$\zeta\{P \in \Delta(\Omega) : \overline{T}(s, P) = 1\} > 1 - \varepsilon - \delta/2.$$

It follows that condition (4.3) (where ε is replaced with $\varepsilon + \delta$) may be violated only for paths s such that

$$\zeta\{P \in \Delta(\Omega) : s \in A_P^c - B_P^n\} > \delta/2. \tag{6.6}$$

Thus, as $A_P^c \subset B_P^c$, $\{s \in \Omega : (4.3)$ - but with $\varepsilon + \delta$ instead of ε - is violated$\} \subset \{s \in \Omega :$ (6.6) - but with $B_P^c - B_P^n$ instead of $A_P^c - B_P^n$ - is satisfied$\}$.

Let χ_S denote the indicator function of the set S, i.e., $\chi_S(s) = 1$ if $s \in S$, and $\chi_S(s) = 0$ otherwise. The measure Q of the latter (and therefore, also the former) set does not exceed

$$E^Q \chi_{\left\{s \in \Omega : \zeta\{P \in \Delta(\Omega) : s \in B_P^c - B_P^n\} > \delta/2\right\}} \leq E^Q \frac{2}{\delta} E^\zeta \chi_{\left\{(s,P) \in \Omega \times \Delta(\Omega) : s \in B_P^c - B_P^n\right\}}.$$

Indeed, this inequality follows from the fact that if for some s the set $\{P \in \Delta(\Omega) : s \in B_P^c - B_P^n\}$ has measure ζ greater than $\delta/2$, then

$$\chi_{\left\{s\in\Omega:\zeta\{P\in\Delta(\Omega):s\in B_P^c-B_P^n\}>\delta/2\right\}} = 1 < \frac{2}{\delta}E^{\zeta}\chi_{\left\{(s,P)\in\Omega\times\Delta(\Omega):s\in B_P^c-B_P^n\right\}};$$

and if for some s the set $\{P \in \Delta(\Omega) : s \in B_P^c - B_P^n\}$ has measure ζ no greater than $\delta/2$, then

$$\chi_{\left\{s\in\Omega:\zeta\{P\in\Delta(\Omega):s\in B_P^c-B_P^n\}>\delta/2\right\}} = 0 \leq \frac{2}{\delta}E^{\zeta}\chi_{\left\{(s,P)\in\Omega\times\Delta(\Omega):s\in B_P^c-B_P^n\right\}}.$$

By Fubini's Theorem,

$$E^Q\frac{2}{\delta}E^{\zeta}\chi_{\left\{(s,P)\in\Omega\times\Delta(\Omega):s\in B_P^c-B_P^n\right\}} = \frac{2}{\delta}E^{\zeta}E^Q\chi_{\left\{(P,s)\in\Delta(\Omega)\times\Omega:s\in B_P^c-B_P^n\right\}} \leq$$

$$\leq \frac{2}{\delta}E^{\zeta}\left(\nu\cdot\frac{\delta}{2}\right) = \nu.$$

■

References

[1] Anderson, R. and W. Zame (2001), "Genericity with Infinitely Many Parameters," *Advances in Theoretical Economics*, **1**, 1-62.

[2] Al-Najjar, N. and J. Weinstein (2008), "Comparative Testing of Experts," *Econometrica*, forthcoming.

[3] Cesa-Bianchi, N. and G. Lugosi (2006), *Prediction, Learning and Games*, Cambridge University Press.

[4] Dekel, E. and Y. Feinberg (2006), "Non-Bayesian Testing of a Stochastic Prediction," *Review of Economic Studies*, **73**, 893 - 906.

[5] Dudley, R.M. (1989), *Real Analysis and Probability*, Wadsworth Inc., Belmont, California.

[6] Fan, K. (1953), "Minimax Theorems," *Proceedings of the National Academy of Science U.S.A.*, **39**, 42-47.

[7] Feinberg, Y. and C. Stewart (2008), "Testing Multiple Forecasters," *Econometrica*, forthcoming.

[8] Fortnow, L. and R. Vohra (2006), "The Complexity of Forecast Testing," mimeo.

[9] Foster, D. and R. Vohra (1998), "Asymptotic Calibration," *Biometrika*, **85**, 379-390.

[10] Fudenberg, D. and D. Levine (1999), "An Easier Way to Calibrate," *Games and Economic Behavior*, **29**, 131-137.

[11] Hart, S. and A. Mas-Colell (2001), "A General Class of Adaptive Strategies," *Journal of Economic Theory*, **98**, 26-54.

[12] Hayek, F. (1945), "The Use of Knowledge in Society," *American Economic Review*, **35**, 519-530.

[13] Kalai, E., E. Lehrer, and R. Smorodinsky (1999), "Calibrated Forecasting and Merging," *Games and Economic Behavior*, **29**, 151-169.

[14] Lehrer, E. (2001), "Any Inspection Rule is Manipulable," *Econometrica*, **69**, 1333-1347.

[15] Lehrer, E. and E. Solan (2003), "No Regret with Bounded Computation Capacity," Tel Aviv University, mimeo.

[16] Olszewski, W. and A. Sandroni (2008), "Manipulability of Future-Independent Tests," *Econometrica*, forthcoming.

[17] Olszewski, W. and A. Sandroni (2008a), "A Nonmanipulable Test," *Annals of Statistics*, forthcoming.

[18] Olszewski, W. and A. Sandroni (2008b), "Falsifiability," mimeo.

[19] Rudin, W. (1973), *Functional Analysis*, McGraw-Hill.

[20] Rustichini, A. (1999), "Optimal Properties of Stimulus-Response Learning Models," *Games and Economic Behavior*, **29**, 244-273.

[21] Sandroni A. (2003), "The Reproducible Properties of Correct Forecasts," *International Journal of Game Theory*, **32**, 151-159.

[22] Sandroni, A., R. Smorodinsky, and R. Vohra (2003), "Calibration with Many Checking Rules," *Mathematics of Operations Research,* **28***, 141-153.*

[23] Shiryaev, A. (1996), *Probability*, Springer Verlag, New York Inc.

[24] Shmaya, E. (2008), "Many Inspections are Manipulable," *Theoretical Economics*, forthcoming.

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

[26] Vovk, V. and G. Shafer (2005), "Good Randomized Sequential Probability Forecasting is Always Possible," *Journal of the Royal Statistical Society Series B*, **67**, 747-763.