# Counterfactual Predictions

Wojciech Olszewski* and Alvaro Sandroni†

December 5, 2006

**Abstract**

The difficulties in properly anticipating key economic variables may encourage decision makers to rely on experts' forecasts. The experts' forecasts, however, may not be accurate. So, their forecasts must be empirically tested. This may induce experts to forecast strategically to pass the test.

A test can be ignorantly passed if a false expert, with no knowledge of the data generating process, can pass the test. Standard tests, if they are unlikely to reject correct forecasts, can be ignorantly passed. Tests that cannot be ignorantly passed must necessarily make use of future predictions (i.e., predictions based on data not yet realized at the time the forecasts are rejected). Such tests cannot be run if, as it is customary, forecasters only report the probability of next period's events given the observed data. This result shows that it is difficult to dismiss false, but strategic, experts. This result also suggests an important role of counterfactual predictions in the empirical testing of forecasts.

*Department of Economics, Northwestern University, 2003 Sheridan Road Evanston IL 60208

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

## 1. Introduction

Expectations of future events have long been recognized as a significant factor in economic activity (see Pigou (1927)). However, the processes by which agents form their beliefs remain largely unknown. The difficulties in anticipating key economic variables may encourage decision makers to rely on experts' forecasts. If informed, a professional forecaster can reveal the probabilities of interest to the decision makers. If uninformed, the forecaster (henceforth called Bob) may mislead the decision makers. Hence, it is important to check the quality of experts' forecasts. Assume that a tester (named Alice) tests Bob's forecasts empirically.

A standard test determines observable histories that are (jointly) unlikely under the null hypothesis that Bob's forecasts are correct. These data sequences are deemed inconsistent with Bob's forecasts and, if observed, lead to a rejection of the forecasts. This methodology is unproblematic if the forecasts are reported honestly. The main difficulty is that Bob, even if uninformed, might be capable of strategically manipulating Alice's test (i.e., capable of producing forecasts that will not be rejected by Alice's test, regardless of how the data turns out to be realized in the future).

There is limited purpose in running a test that can be manipulated when the forecaster is strategic. Even in the extreme case that the forecaster has no knowledge regarding the data generating process, the outcome of the test will almost inevitably support the hypothesis that the forecasts are correct. Hence, the uninformed expert would not fear having his forecasts discredited by the data.

Consider a standard calibration test that requires the empirical frequencies of an outcome (say 1) to be close to $p$ in the periods that 1 was forecasted with probability near $p$. Foster and Vohra (1998) show that the calibration test can be manipulated. So, it is possible to produce forecasts that, in the future, will prove to be calibrated, no matter which sequence of data is eventually observed. In contrast, Dekel and Feinberg (2006) and Olszewski and Sandroni (2006) show the existence of an empirical test that does not reject the forecasts of an informed expert and can reject the forecasts of an uniformed expert.[1]

The tests proposed by Dekel and Feinberg (2006) and Olszewski and Sandroni (2006a) require Bob to deliver, at period zero, an entire theory of the stochastic process. By definition, a theory must tell Alice, from the outset, all the forecasts

[1] The existence of such a test was first demonstrated by Dekel and Feinberg (2006) under the continuum hypothesis. Subsequently, Olszewski and Sandroni (2006) constructed a test with the required properties (therefore dispensing with the continuum hypothesis).

for the next period, conditional on any possible data set. Typically, a forecaster does not announce an entire theory but, instead, only publicizes a forecast in each period, according to the observed data. Dekel and Feinberg (2006) argued that asking for a theory at period zero may have been an important feature that enabled them to prove the existence of their test. Hence, a natural issue to consider is whether there exists a nonmanipulable test that does not require an entire theory, but rather uses only the forecasts made along the observed histories.

Assume that Bob, before any data is observed, delivers to Alice an entire theory of the stochastic process. Let's say that a test does not make use of future predictions if whenever a theory $f$ is rejected at some history $s_t$ (observed at period $t$) then another theory $f'$, that makes the exact same predictions conditional on any data set at or before period $t-1$, must also be rejected at history $s_t$. Now assume that instead of delivering an entire theory, Bob announces a forecast each period according to the observed data. Then, Alice cannot run a test that uses future predictions. So, we restrict attention to tests that do not use future predictions.

A statistical test is *regular* if it rejects the actual data generating process with low probability and it makes no use of future predictions. A statistical test *can be ignorantly passed* if it is possible to strategically produce theories that are unlikely to be rejected on any future realizations of the data.[2]

We show that any regular statistical test can be ignorantly passed. This result shows that it is difficult to prevent the manipulation of empirical tests. Experts have incentives to be strategic and the data will not show that their forecasts were strategically produced to pass the test. This holds even under the extreme assumptions that the tester has arbitrarily long data sets at her disposal and the strategic forecaster knows nothing about the data generating process.

## 2. Related literatures

### 2.1. Counterfactual predictions

Counterfactual predictions have a significant function in several literatures. In game theory, beliefs off the play path are relevant in determining whether an equilibrium satisfies refinements such as perfection. Psychologists are interested

[2]We allow the uninformed expert to produce theories at random at period zero. Hence, the term "unlikely" here refers to the expert's randomization and not to the possible realizations of the data.

in the direct impact on welfare of "want if" concerns (see Medvec, Madey, and Gilovich (1995)). Counterfactual predictions such as "what would be the salary of this woman if she were a man" are often made as an output of a statistical model. However, the use of a future prediction as an input to a statistical model is unusual.[3] Consider a future prediction such as "if it rains tomorrow then it will also rain the day after tomorrow." It is diffcult to test this prediction today because we have no data on it. So, it is counter-intuitive to make any use of this prediction today (and not the day after tomorrow) to determine the forecaster's type. Nevertheless, our results suggest a useful role for future predictions in the testing of forecasts.

### 2.2. Risk and uncertainty

An important distinction in economics is between risk and uncertainty.[4] Risk refers to the case in which the available information can be properly summarized by probabilities, uncertainty refers to the case in which it cannot. In our model, Bob, if informed, faces risk. Alice and Bob, if uninformed, face uncertainty.[5]

As is well-known, the distinction between risk and uncertainty cannot be made within Savage's (1954) axioms. The large literature on uncertainty often produces alternative axiomatic foundations where this distinction can be made (See, among others, Bewley (1986), Casadesus-Masanell et al. (2000), Epstein (1999), Ghirardato et al. (2004), Gilboa (1987), Gilboa and Schmeidler (1989), Klibanoff et al. (2005), Maccheroni et al. (2006), Olszewski (2006), Schmeidler (1989), Siniscalchi (2005), and Wakker (1989)). Unlike most of this literature, our objective here is not to provide a representation theorem for decisions under uncertainty nor to empirically test Savage's axioms, but rather to show how specific strategies can be used to effectively reduce or eliminate uncertainty.

[3] The use of counterfactual predictions is controversial (e.g., the literature of counterfactual history is seen as useful by some and as fantasies by others, see Fogel (1967) and McAfee (1983))

[4] The distinction is traditionally attributed to Knight (1921). However, LeRoy and Singell (1987) argue that Knight did not have in mind this distinction. Ellsberg (1961), in a well-known experiment, demonstrated that this distinction is empirically significant.

[5] This is significantly different from the case in which the tester is well, albeit imperfectly, informed. We refer the reader to Crawford and Sobel (1982) for a classic model of information transmission and to Morgan and Stocken (2003) and Sørensen and Ottaviani (2006) (among others) for cheap-talk games between forecasters and decision-makers. We also refer the reader to Dow and Gorton (1997), Ehrbeck and Waldmann (1996), Laster, Bennett and Geoum (1999) and Trueman (1988) (among others) for models in which professional forecasters have incentives to report their forecasts strategically.

### 2.3. Empirical tests of rational expectations

The rational expectations hypothesis has been subjected to extensive empirical testing. The careful examination of Keane and Runkle ((1990), (1998)) failed to reject the hypothesis that professional forecasters' expectations are rational (i.e., that the forecasts coincide with the correct probabilities).[6] In this literature, the forecasts are assumed to be reported honestly and nonstrategically. So, the connection between our paper and this literature is tenuous. In addition, unlike most statistical models, we make no assumptions on how the data might evolve. These differences in the basic assumptions are partially due to the differences in objectives. The main purpose of our paper is not to test forecasts, but rather to demonstrate the properties that empirical tests must satisfy to be nonmanipulable.

### 2.4. Testing strategic experts

As mentioned in the introduction, the calibration test can be ignorantly passed. In fact, strong forms of calibration tests can be ignorantly passed (see Fudenberg and Levine (1999), Hart (2005), Hart and Mas-Colell (2001), Lehrer (2001), Lehrer and Solan (2003), Kalai, Lehrer and Smorodinsky (1999), and Rustichini (1999) for related results). Sandroni (2003) considers a class of tests that can be ignorantly passed. However, severe restrictions are imposed on this class of tests.[7] We also refer the reader to Cesa-Bianchi and Lugosi (2006) and Vovk and Shafer (2006) for related results and to the recents paper of Al-Najjar and Weinstein (2006) and Feinberg and Stuart (2006) on comparing different experts and to Fortnow and Vohra (2006) on testing experts with computational bounds.

So far, the literature has produced classes of tests that can be ignorantly passed. The contribution of this paper is to show a complete impossibility result: *no* regular test can feasibly reject a potentially strategic expert. These results (combined with the results of Dekel and Feinberg (2006) and Olszewski and Sandroni (2006a)) provide a definite separation between the cases in which the expert delivers an entire theory and the case in which the expert delivers a forecast each period.

[6] See Lowell (1986) for other results on empirical testing of forecasts.

[7] The required conditions on the class of tests are so restrictive that significant effort is required to demonstrate that even the calibration test satisfies it.

## 3. Basic Set-Up

In each period one outcome, 0 or 1, is observed.[8] Before any data is observed, an expert, named Bob, announces a theory that must be tested. Conditional on any *t-history* of outcomes $s_t \in \{0,1\}^t$, Bob's theory claims that the probability of 1 in period $t+1$ is $f(s_t)$.

To simplify the language, we identify a theory with its predictions. That is, theories that produce identical predictions are not differentiated. Hence, we define a *theory* as an arbitrary function that takes as an input any finite history and returns as an output a probability of 1. Formally, a theory is a function

$$f : \{s_0\} \cup S_\infty \longrightarrow [0,1],$$

where $S_\infty = \bigcup_{t=1}^{\infty} \{0,1\}^t$ is the set of of all finite histories and $s_0$ is the null history.

A tester, named Alice, tests Bob's theory empirically. So, given a potentially long string of data, Alice must reject or not reject Bob's theory. Hence, a *test* $T$ is an arbitrary function that takes as an input a theory $f$ and returns, as an output, a set $T(f) \subseteq S_\infty$ of finite histories considered to be inconsistent with the theory $f$. So, Alice rejects Bob's theory $f$ if she observes data that belongs to $T(f)$.[9] Formally, a test is a function

$$T : F \to \bar{S},$$

where $F$ is the set of all theories and $\bar{S}$ is the set of all subsets of $S_\infty$.[10]

The timing of the model is as follows: at period zero, Alice selects her empirical test $T$. Bob observes the test $T$ and selects his theory $f$ (also at period zero).[11]

[8] It is immediate to extend the results to the case where there are finitely many possible outcomes in each period.

[9] Instead of a test, Alice could offer a contract to Bob in which Bob's reward is higher when his theory is not rejected by the data (see Olszewski and Sandroni (2006b)).

[10] We assume that $s_t \in T(f)$ implies that $s_m \in T(f)$ whenever $m \geq t$ and $s_t = s_m \mid t$ (i.e., $s_t$ are the first $t$ outcomes of $s_m$). That is, if a finite history $s_t$ is considered inconsistent with the theory $f$, then any longer history $s_m$ whose first $t$ outcomes coincide with $s_t$ is also considered inconsistent with the theory $f$.

For simplicity, we also assume that $s_t \in T(f)$ whenever $s_m \in T(f)$ for some $m > t$ and every $s_m$ with $s_t = s_m \mid t$. That is, if any $m$-history that is an extension of a finite history $s_t$ is considered inconsistent with the theory $f$, then the history $s_t$ is itself considered inconsistent with the theory $f$.

[11] The results of these paper can be extended to the case that Alice selects her test at random. It suffices to assume that Bob properly anticipates the odds that Alice selects each test.

In period 1 and onwards, the data is revealed and Bob's theory is either rejected or not rejected by Alice's test at some point in the future.

Bob can be an informed expert who honestly reports to Alice the data generating process. However, Bob may also be an uninformed expert who knows nothing about the data generating process. If so, Bob tries to strategically produce theories with the objective of not being rejected by the data. Alice anticipates this and wants a test such that Bob, if uniformed, cannot manipulate. Both the uninformed expert and Alice face uncertainty: they do not have any knowledge on the data generating process.

Although Alice tests Bob's theory using a string of outcomes, we do not make any assumptions on the data generating process (such as a Markov process, a stationary process, or some mixing condition). This lack of assumptions over the data generating process distinguishes our model from standard statistical models and hence it requires some explanation. It is very difficult to demonstrate that any key economic variable (such as inflation or GDP) follows any of these well-known processes. At best, such assumptions can be tested and rejected. More importantly assume that, before any data was observed, Alice knew that the actual process belonged to a parametrizable set of processes (such as independent, identically distributed sequences of random variables) then she could infer, almost perfectly, the actual process from the data. Alice could accomplish all of this without Bob. Therefore, the lack of assumptions over the data generating process adds an element of coherence into a model of a forecaster and a tester.

Given that Bob must delivers an entire theory, Alice knows, at period zero, Bob's forecast conditional on any finite history. At period $m \in N$, Alice observes the data $s_m \in \{0,1\}^m$. Let $s_t = s_m \mid t$ be the first $t$ outcomes of $s_m$. Let $f_{s_m} = \{f(s_t),\ s_t = s_m \mid t,\ t = 0,..,m\}$ be a *sequence of the actual forecasts* made up to period $m$, if $s_m$ is observed. Clearly, if Bob were required to produce only a forecast each period then Alice would observe at period $m$ only $f_{s_m}$ and $s_m$.

### 3.1. Example

We now consider an example of an empirical test. Let $J_t(s_t)$ be the $t$-th outcome of $s_t$. Then,

$$R_1(f, s_m) = \frac{1}{m}\sum_{t=1}^{m}[f(s_{t-1}) - J_t(s_t)]$$

marks the difference between the average forecast of 1 and the empirical frequency of 1.

Alice could reject the theory $f$ on all sufficiently long histories such that the average forecast of 1 did not become sufficiently close to the empirical frequency of 1. That is, fix $\eta > 0$ and a period $\bar{m}$. Bob's theory $f$ is rejected on any history $s_m$ (and longer histories $s_k$ with $s_m = s_k \mid m$) such that

$$|R_1(f, s_m)| \geq \eta \text{ and } m \geq \bar{m}. \tag{3.1}$$

The test defined above (henceforth called an $R_1-$test) is notationally undemanding and can be used to exemplify general properties of empirical tests. Given $\varepsilon > 0$ a pair $(\eta, \bar{m})$ can be chosen such that if the theory $f$ is correct (i.e., if the predictions made by $f$ coincide with the data generating process), then $f$ will not be rejected with probability $1 - \varepsilon$ (i.e., (2.1) occurs with probability less than $\varepsilon$). Hence, if Bob announces the data generating process, it is unlikely that he will be rejected.

At period $m$, the $R_1-$tests reject or do not reject a theory based on the sequence of the actual forecasts made up to period $m-1$, $f_{s_{m-1}}$, and the available data, $s_m$. Thus, the $R_1-$tests do not use predictions for which there is no data.

Now assume that Bob is a false expert who knows nothing about the data generating process. Assume that, at period zero, Bob announces a theory $f$ that satisfies:

$$\begin{array}{rll} f(s_t) = 1 & \text{if} & R_1(f, s_t) < 0; \\ f(s_t) = 0.5 & \text{if} & R_1(f, s_t) = 0; \\ f(s_t) = 0 & \text{if} & R_1(f, s_t) > 0. \end{array} \tag{3.2}$$

It is immediate to see that if $R_1$ is negative at period $t$ then, no matter whether 0 or 1 is realized at period $t+1$, $R_1$ increases. Conversely, if $R_1$ is positive at period $t$ then, no matter whether 0 or 1 is realized at period $t+1$, $R_1$ decreases. So, $R_1$ approaches zero as the data unfolds. Therefore, if $\bar{m}$ is sufficiently large, Bob can pass this test without any knowledge of the data generating process.

The $R_1-$tests may seem weak and a proof that some of them can be passed without any relevant knowledge seemingly confirms this intuition. However, the stronger calibration tests of Lehrer (2001) and Foster and Vohra (1998) can also be passed without any knowledge of the data generating process.

## 4. Properties of Empirical Tests

Any theory $f$ uniquely defines the probability of any set $A \subseteq S_\infty$ of finite histories (denoted by $P^f(A)$). The probability of each finite history $s_m$ is just the product

$$\prod_{t=1}^{m} h^f(s_t) \tag{4.1}$$

where $s_t = s_m \mid t$, $h^f(s_t) := f(s_{t-1})$ if $J_t(s_t) = 1$ and $h^f(s_t) := 1 - f(s_{t-1})$ if $J_t(s_t) = 0$.

**Definition 1.** *Fix $\varepsilon \in [0,1]$. A test $T$ does not reject the truth with probability $1-\varepsilon$ if for any $f \in F$*

$$P^f(T(f)) \leq \varepsilon.$$

A test does not reject the truth if the actual data generating process is unlikely to rejected. So, if Bob is an informed expert and announces his theory honestly, then he will not be rejected with high probability. Hence, an informed expert has good incentives to report his theory honestly.

Two theories $f$ and $f'$ are *equivalent until period $m$* if $f(s_t) = f'(s_t)$ for any $t-$history $s_t$, $t \leq m$. So, two theories are equivalent until period $m$ if they make the same predictions up to and at period $m$.

**Definition 2.** *A test $T$ does not make use of future predictions if, given any pair of theories $f$ and $f'$ that are equivalent until period $m$, $s_t \in T(f)$, $t \leq m$, implies $s_t \in T(f')$.*

A test does not make use of future predictions if, whenever a theory $f$ is rejected at an $m-$history $s_m$, another theory $f'$, that makes exactly the same predictions as $f$ until period $m$, must also be rejected at $s_m$.

If, as is customary in professional forecasting, Bob is only required to produce a forecast each period, then, at period $m$, Alice observes only the actual predictions $f_{s_m}$ and the data $s_m$. Hence, her test cannot make use of future predictions.[12] However, if Bob is required to deliver an entire theory at period zero then Alice's test could, in principle, make use of future predictions because she knows in advance how Bob's predictions would be conditioned on the data.

[12] At period $m$, the data $s_m$ is available and the realized predictions are $f_{s_m} = \{f(s_t),\ s_t = s_m \mid t,\ t = 0., ..., m\}$. The predictions $f(s'_t)$, $s'_t \neq s_m \mid t$ are called parallel predictions: $f(s'_t)$ is based on information $s'_t$ that was not observed at period $t$. The predictions $f(s_n)$, $n \geq m+1$, are not yet realized at period $m$ and, hence, called future predictions.

**Definition 3.** *A regular $\varepsilon-$test does not make use of future predictions and does not reject the truth with probability $1-\varepsilon$.*

In section 5, we show that standard statistical tests do not make use of future predictions. This is to be expected, because statistical tests are meant to use data and no data is yet available for future predictions.

Bob is not restricted to select a theory deterministically. He may randomize when selecting his theory at period 0.[13] Let a random generator of theories $\zeta$ be a probability distribution over the set $F$ of all theories. Given any finite history $s_t \in \{0,1\}^t$ let

$$\zeta(s_t) := \zeta(\{f \in F : s_t \in T(f)\})$$

be the probability that $\zeta$ selects a theory that will be rejected if $s_t$ is observed.[14]

**Definition 4.** *A test $T$ can be ignorantly passed with probability $1-\varepsilon$ if there exists a random generator of theories $\zeta$ such that for all finite histories $s_t \in S_\infty$*

$$\zeta(s_t) \leq \varepsilon.$$

The random generator $\zeta$ may depend on the test $T$, but not on any knowledge of the actual data generating process. If a test can be ignorantly passed, Bob can randomly select theories that, with probability $1-\varepsilon$ (according to Bob's randomization device), will not be rejected, no matter what data is observed. Alice has no reason to run a test that can be ignorantly passed if the forecaster is potentially strategic. Even in the extreme case that Bob completely ignores the data generating process, the test will almost certainly fail to reject his theory, *no matter how the data unfolds.*

## 5. Main Result

**Proposition 1.** *Fix $\varepsilon \in [0,1]$ and $\delta \in (0, 1-\varepsilon]$. Any test $T$ that does not reject the truth with probability $1-\varepsilon$ and does not make use of future predictions can be ignorantly passed with probability $1-\varepsilon-\delta$.*

[13] Given that Bob (perhaps) randomizes only once at period zero, Alice cannot tell whether the theory she just received was produced deterministically or at random.

[14] This definition requires a measurability provision on the sets $\{f \in F : s_t \in T(f)\}$. We will restrict attention to random generators of theories $\zeta$ for which sets of this form are measurable.

Proposition 1 shows a fundamental limitation of regular statistical tests. Any regular tests can be ignorantly passed. If Alice cannot make use of future predictions (e.g., only actual predictions are announced by the forecaster), she can only run regular tests and so she has no reason to run any test when confronted with a potentially strategic expert. These tests will not reveal whether the expert is uninformed. This result holds even if Alice possesses unboundedly large data sets and the fraudulent forecaster knows nothing about the data generating process.

Assume, for the moment, that Alice offers a formal contract to Bob defined by a regular $\varepsilon-$ test. In this contract, Bob receives a high payoff $h$ if his theory is not rejected and a low payoff payoff $l$ if it is rejected. By proposition 1, this contract is worth (approximately) the same to a completely informed expert as to a completely uninformed expert (and so, presumably, as to partially informed experts as well). Hence, Alice faces adverse selection and moral hazard problems that are unmitigated by contracts. None of these contracts can feasibly screen informed from uninformed experts. So, agents might anticipate that fraudulent forecasts will not be dismissed. In the absence of an effective exogenous check on the quality of the forecasts (that the data were supposed to provide), either decision makers will not consult professional forecasters or fraudulent formation of forecasts will become a wide-spread practice.

The difficulty pointed out in proposition 1 is difficult to circumvent if Alice has no access to future predictions because the results holds for any regular empirical test. Moreover, the result also holds for all future realizations of the data and so it requires no knowledge over the data generating process. The only requirement in proposition 1 is that Bob knows the regular tests that Alice uses. However, even this requirement can be relaxed. It suffices to assume that Bob properly anticipates the odds that Alice selects each regular test.

### 5.1. Intuition of Proposition 1

Fix a regular $\varepsilon-$test $T$. This test, as every test, is a limit of tests $T_m$, $m = 1, 2, ...$, such that $T_m$ makes the decision whether to reject a theory or not in period $m$ or earlier. Consider the following zero-sum game between Nature and Bob: Nature's pure strategy is an infinite sequence of outcomes. Bob's pure strategy is a theory. Bob's payoff is one if his theory is never rejected (by the test $T_m$) and zero otherwise. Both Nature and Bob are allowed to randomize.

By the assumption that $T$ does not reject the truth with probability $1-\varepsilon$, for every mixed strategy of Nature, there is a pure strategy for Bob (to announce the

theory $f$ that coincides with Nature's strategy) that gives him a payoff of $1-\varepsilon$ or higher. Hence, if the conditions of Fan's (1954) MinMax are satisfied, there is a (mixed) strategy $\zeta_m$ for Bob that ensures him a payoff arbitrarily close to $1-\varepsilon$, no matter what strategy Nature chooses. In particular, for any history $s_t \in S_\infty$ that Nature can select, Bob's payoff is arbitrarily close to $1-\varepsilon$.

Fan's MinMax theorem requires Nature's strategy space to be compact and the payoff function to be lower semi-continuous with respect to Nature's strategy. The topology that makes Nature's strategy space compact is the weak$-*$ topology. The assumption that $T_m$ makes the decision in period $m$ or earlier, guarantees the lower semi-continuity of the payoff function.[15]

A limit, $\zeta$, of (a subsequence) of these mixed strategies $\zeta_m$, $m = 1, 2, ...$, exists because $\Delta\Delta(\Omega)$ is compact in the weak$-*$ topology. However, $\zeta$ does not necessarily guarantee that Bob's theory will not be rejected with probability $1-\varepsilon$, no matter which data is observed. To this end, one must use a specific sequence of tests $T_m$; in particular, the sets of the form $\{f \in F : s_m \in T_m(f)\}$ must be open in the weak$-*$ topology. The assumption that $T$ makes no use of future predictions is critical for the construction of a sequence of test $T_m$, $m = 1, 2, ...$, with this property.

## 6. Empirical tests

The purpose of this section is to show that the assumptions of proposition 1 are satisfied by standard statistical models. We do not explicitly analyze every statistical model ever produced, but provide a few simple examples; however, we hope that these examples suffices because the arguments that we put forward seem to be general.

Asymptotic tests are common in statistics. These tests work as if Alice eventually had an infinite string of data and could decide whether or not to reject Bob's theory at infinity. Naturally, asymptotic tests can be approximated by tests that can reject theories in finite time (as defined in Section 2). In this section, we present a few examples of common asymptotic tests. We show that they can be approximated by regular tests.

[15] If Bob's payoff depended on the test $T$ instead of $T_m$, then the payoff function would not necessarily be lower semi-continuous.

Fix $\delta \in (0, 0.5)$. Given a theory $f$, let $f_\delta$ be an alternative theory defined by

$$f_\delta(s_t) = \begin{cases} f(s_t) + \delta & \text{if} \quad f(s_t) \leq 0/5; \\ \\ f_\delta(s_t) - \delta & \text{if} \quad f(s_t) > 0/5.^{16} \end{cases}$$

A straightforward martingale argument shows that, $P^f-$almost surely,

$$\frac{P^{f_\delta}(s_t)}{P^f(s_t)} \underset{t\to\infty}{\longrightarrow} 0.^{17}$$

That is, under the null hypothesis (that $P^f$ is the data generating process), the likelihood of $P^{f_\delta}$ becomes much smaller than the likelihood of $P^f$. The likelihood test rejects theory $f$ in favor of the alternative theory $f_\delta$ if the likelihood ratio

$$\frac{P^{f_\delta}(s_t)}{P^f(s_t)}$$

does not approach zero.

Let $R(f)$ be the set of infinite histories such that the likelihood ratio does not approach zero. Say that a test $T$ is harder than the likelihood test if $R(f) \subseteq T(f)$. So, rejection by the likelihood test implies rejection by the test $T$.

**Proposition 2.** *Given $\varepsilon > 0$, there exists a regular $\varepsilon-$test $T$ that is harder than the likelihood test.*

By proposition 1, the test $T$ can be ignorantly passed with probability $1 - \varepsilon$. Hence, by proposition 2, the likelihood test can be ignorantly passed with arbitrarily high probability. This is a surprising result because, without any knowledge of the data generating process, it is not obvious whether the theory $f$ or the alternative theory $f_\delta$ will eventually produce a higher likelihood. However, a false expert can produce theories that, no matter which data is realized, will prove in the future (with arbitrarily high chance) to generate a much higher likelihood than the alternative theories.

Of course, the unexpected result is proposition 1. proposition 2, in contrast, is a natural finding. An intuition is as follows: $P^f$-almost surely, the likelihood ratio

[17] Naturally, we refer here to the probability measure $P^f$ defined on the space of infinity histories. See the appendix for a precise definition.

approaches zero. Hence, with arbitrarily high probability, the likelihood ratio must remain small if the string of data is long enough. Hence, the test $T$ rejects the theory $f$ whenever the likelihood ratio is not small and the string of data long. By construction, the $T$ test is harder than the likelihood test and does not reject the truth with high probability. Moreover, the test $T$ does not make use of future predictions because the likelihood ratio depends only on the forecasts made along the observed history.

The basic idea in proposition 2 is not limited to the likelihood test. Other asymptotic tests can also be associated with harder regular tests. We conclude this section with the analysis of calibration tests. Let $\mathcal{I}_{t-1}$ be an indicator function that depends on the data up to period $t-1$ (i.e., $s_{t-1}$) and the predictions made up to period $t-1$ (i.e., $f(s_k)$, $s_k = s_{t-1} \mid k$, $k \leq t-1$). For example, $\mathcal{I}_{t-1}$ can be equal to 1 if $f(s_{t-1}) \in [\frac{j}{n}, \frac{j+1}{n}]$ for some $j < n$ and zero otherwise. Alternatively, $\mathcal{I}_{t-1}$ can be equal to 1 if $t$ is even and zero if $t$ is odd. Consider an arbitrary countable collection $\mathcal{I}^i = (\mathcal{I}^i_0, ..., \mathcal{I}^i_{t-1}, ...)$, $i \in N$, of indicator functions. The calibration test requires that for all $i \in N$

$$\frac{1}{m}\sum_{t=1}^{m}[f(s_{t-1}) - J_t(s_t)]\mathcal{I}^i_{t-1} \underset{m\to\infty}{\longrightarrow} 0. \tag{6.1}$$

These calibration tests require a match between average forecasts and empirical frequencies on specific subsequences. These subsequences could be, as in Foster and Vohra (1998), those in which the forecasts are near $p \in [0, 1]$. Then the test requires that the empirical frequencies of 1 be close to $p$ in the periods that followed a forecast of 1 that was close to $p$. Alternatively, these subsequences could also be, as in Lehrer (2001), periods in which a certain outcome was observed. In general, the calibration test rejects a theory $f$ if (5.2) does not hold.

**Proposition 3.** *Given $\varepsilon > 0$, there exists a regular $\varepsilon-$test $T'$ that is harder than the calibration test.*

The intuition of proposition 3 is the same as that of proposition 2. A sophisticated law of large numbers shows that, under the null hypothesis that $P^f$ is the data generating process, almost surely, the calibration scores in 5.2 eventually approach zero. Hence, with arbitrarily high probability, these calibration scores must remain small if the string of data is long enough. The test $T'$ rejects the theory $f$ whenever the calibration scores are not small and the string of data is long. By construction, the $T'$ test is harder than the calibration test and does not

reject the truth with high probability. Moreover, the test $T'$ does not make use of future predictions because the calibration scores depend only on the forecasts made along the observed history.

By Propositions 1 and 3, the calibration tests can be ignorantly passed with arbitrarily high probability. Hence, a false expert can produce forecasts that, in the future, once the data is revealed, will prove to be calibrated. This result combines the Foster and Vohra (1998) result (where the indicator function depends only on the forecasts) and the Lehrer (2001) result (where the indicator function depends only on the data). However, the examples presented here (likelihood and calibration tests) are just illustrations of the general point that several statistical tests can be associated with harder regular tests.

## 7. Conclusion

Strategic manipulation of tests is difficult to prevent. An expert can strategically produce forecasts that, once the data is revealed, will not be rejected by any given regular empirical test. This holds even under the extreme assumptions that the expert knows nothing about the data generating process, and that the tester has unbounded data at her disposal.

However, if forecasters must deliver an entire theory of a stochastic process, then tests that make use of future predictions can be employed. Some of these tests can dismiss false experts without dismissing informed experts. These results suggest the necessity of providing theories for a successful screening of correct forecasts from strategically produced forecasts.

## 8. Proofs

We use the following terminology: Let $\Omega = \{0,1\}^\infty$ be the set of all *paths*, i.e., infinite histories. Given a path $s$, let $s \mid t$ be the first $t$ coordinates of $s$. A *cylinder* with base on $s_t \in \{0,1\}^t$ is the set $C(s_t) \subset \{0,1\}^\infty$ of all infinite extensions of $s_t$. We endow $\Omega$ with the topology that comprises of unions of cylinders with finite base. 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 $\Im$ is the $\sigma$-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 $\Delta(\Omega)$ the set of all probability measures on $(\Omega, \Im)$. We endow $\Delta(\Omega)$ 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). Let $\Delta\Delta(\Omega)$ be the set of probability measures on $\Delta(\Omega)$. We endow $\Delta\Delta(\Omega)$ also with the weak$-*$ topology. It is well-known that $\Delta\Delta(\Omega)$ is a compact metric space. It is also well known that there is a 1-1 correspondence between theories $f \in F$ and probability measures $P \in \Delta(\Omega)$, which assigns to every theory $f$ the measure $P^f$ uniquely determined by (4.1). We refer to $P^f \in \Delta(\Omega)$ as the probability measure associated with the theory $f \in F$.

**Definition 5.** *A test $T$ is called finite if for every theory $f$ there exists a number $m \in N$ such that $s_t \in T(f)$, where $t > m$, if and only if $s_t \mid m \in T(f)$.*

**Definition 6.** *A test $T$ does not reject an informed expert with probability $1-\varepsilon$ if for every theory $f \in F$ there exists a theory $\widetilde{f} \in F$ such that*

$$P^f(T(\widetilde{f})) \leq \varepsilon.$$

**Definition 7.** *A test $T_1$ is harder than the test $T_2$ if for any $f \in F$, $s_t \in T_2(f)$ implies that $s_t \in T_1(f)$.*

**Definition 8.** *A set $F' \subseteq F$ is $\delta-$dense in $F$, $\delta > 0$, if for every theory $g \in F$ there exists a theory $f \in F'$ such that*

$$\sup_{A \in \Im} \left| P^f(A) - P^g(A) \right| < \delta.$$

### 8.1. Proof of Proposition 1

We will use the following lemmas. Stronger versions of the first two lemmas appear in Olszewski and Sandroni (2006) as proposition 1 and Lemma 1, respectively.

**Lemma 1.** *Fix $\varepsilon \in [0,1]$ and $\delta \in (0, 1-\varepsilon]$. Let $T$ be a finite test that does not reject an informed expert with probability $1-\varepsilon$. Then, the test $T$ can be ignorantly passed with probability $1-\varepsilon-\delta$.*

Let $X$ be a metric space. Recall that a function $u : X \to R$ is *lower semi-continuous* at an $x \in X$ if for every sequence $(x_n)_{n=1}^{\infty}$ converging to $x$:

$$\forall_{\varepsilon>0} \quad \exists_{\overline{n}} \quad \forall_{n \geq \overline{n}} \quad u(x_n) > u(x) - \varepsilon.$$

The function $u$ is lower semi-continuous if it is lower semi-continuous at every $x \in X$.

**Lemma 2.** *Let $U \subset \Delta(\Omega)$ be an open set. Then the function $H : \Delta\Delta(\Omega) \to [0,1]$ defined by*

$$H(\zeta) = \zeta(U)$$

*is lower semi-continuous.*

Let $\gamma$ be a sequence of positive numbers $(\gamma_t)_{t=1}^{\infty}$. Given $\gamma$, let $R$ be a sequence of finite sets $(R_t)_{t=1}^{\infty}$ such that $R_t \subset (0,1)$, $t \in N$, and

$$\forall_{x\in[0,1]}\ \exists_{r\in R_t} \quad |x-r| < \gamma_t.$$

Given $R$, let $(F_m)_{m=1}^{\infty}$ be a sequence of subsets of $F$ defined by

$$F_m = \left\{ f \in F : \forall_{t=0,\ldots,m}\ \forall_{s_t\in\{0,1\}^t \text{ (or } s_t=s_0 \text{ if } t=0)} \quad f(s_t) \in R_{t+1} \right\}.$$

Given that $R$ is well defined given $\gamma$ and $(F_m)_{m=1}^{\infty}$ is well defined given $R$, it follows that $(F_m)_{m=1}^{\infty}$ is well defined given $\gamma$.

**Lemma 3.** *For every $\delta > 0$ there exists a sequence of positive numbers $\gamma$ such that $F_m$ is $\delta-$dense in $F$ for every $m = 1, 2, ...$*

**Proof:** For now consider an arbitrary $\gamma$. Given $g \in F$ take $f \in F_m$ such that

$$\forall_{t=0,\ldots,m}\ \forall_{s_t\in\{0,1\}^t \text{ (or } s_t=s_0 \text{ if } t=0)} \quad |f(s_t) - g(s_t)| < \gamma_t$$

and

$$\forall_{t=m+1,\ldots}\forall_{s_t\in\{0,1\}^t} \quad f(s_t) = g(s_t).$$

We shall show that there exists a sequence $(\gamma_t)_{t=1}^{\infty}$ such that

$$\left|P^f(C(s_r)) - P^g(C(s_r))\right| < \frac{\delta}{2} \tag{8.1}$$

for every cylinder $C(s_r)$. Indeed,

$$\left|P^f(C(s_q)) - P^g(C(s_q))\right| = \left|h^f(s_1)\cdot\ldots\cdot h^f(s_q) - h^g(s_1)\cdot\ldots\cdot h^g(s_q)\right|,$$

where $q = \min\{r, m+1\}$, and

$$\left|h^f(s_1)\cdot\ldots\cdot h^f(s_q) - h^g(s_1)\cdot\ldots\cdot h^g(s_q)\right| \leq$$

$$\leq (h^g(s_1)+\gamma_1)\cdot ...\cdot \left(h^g(s_q)+\gamma_q\right) - h^g(s_1)\cdot ...\cdot h^g(s_q) \leq$$
$$\leq \left[\prod_{t=1}^{q}(1+\gamma_t)-1\right] \leq \left[\prod_{t=1}^{\infty}(1+\gamma_t)-1\right],$$

where $s_k = s_r \mid k$ for $k < r$. The first inequality follows from the fact that

$$|(a_1+b_1)\cdot ...\cdot (a_q+b_q) - a_1\cdot ...\cdot a_q| \leq (a_1+|b_1|)\cdot ...\cdot (a_q+|b_q|) - a_1\cdot ...\cdot a_q \quad (8.2)$$

for any sets of numbers $a_1, ..., a_q > 0$ and $b_1, ..., b_q$; apply (8.2) to $a_k = h^g(s_k)$ and $b_k = h^f(s_k) - h^g(s_k)$, $k = 1, ..., q$. The second inequality follows from the fact that the function

$$(a_1+b_1)\cdot ...\cdot (a_q+b_q) - a_1\cdot ...\cdot a_q$$

is increasing in $a_1, ..., a_q$ for any sets of positive numbers $a_1, ..., a_q$ and $b_1, ..., b_q$.

So, (8.1) follows if we take a sequence $(\gamma_t)_{t=1}^{\infty}$ such that

$$\prod_{t=1}^{\infty}(1+\gamma_t) < 1+\frac{\delta}{2}.$$

We shall show now that a slightly stronger condition

$$\prod_{t=1}^{\infty}(1+2\gamma_t) < 1+\frac{\delta}{4}$$

guarantees that $\left|P^f(U) - P^g(U)\right| < \delta/2$ for every union of cylinders $U$, not only for every single cylinder.

Indeed, suppose first that there is an $n$ such that $U$ is a union of cylinders with base on $s_t$ with $t \leq n$. Since every cylinder with base on $s_t$ can be represented as the union of two cylinders with base on $s'_{t+1} = (s_t, 0)$ and $s''_{t+1} = (s_t, 1)$ respectively, the set $U$ is the union of a family of cylinders $\mathcal{C}$ with base on histories of length $n$. Thus,

$$\left|P^f(U) - P^g(U)\right| \leq \sum_{C(s_n)\in C} \left|P^f(C(s_m)) - P^g(C(s_m))\right|$$

$$\leq \sum_{C(s_n)\in C} [(h^g(s_1)+\gamma_1)\cdot ...\cdot (h^g(s_n)+\gamma_n) - h^g(s_1)\cdot ...\cdot h^g(s_n)] \leq$$
$$\leq \sum_{s_n\in\{0,1\}^n} [(h^g(s_1)+\gamma_1)\cdot ...\cdot (h^g(s_n)+\gamma_n) - h^g(s_1)\cdot ...\cdot h^g(s_n)] =$$

$$= \sum_{s_n \in \{0,1\}^n} (h^g(s_1) + \gamma_1) \cdot ... \cdot (h^g(s_n) + \gamma_n) - 1 =$$

$$= \sum_{s_{n-1} \in \{0,1\}^{n-1}} (h^g(0) + \gamma_1) \cdot (h^g(0, s_1) + \gamma_2) \cdot ... \cdot (h^g(0, s_{n-1}) + \gamma_n) +$$

$$+ \sum_{s_{n-1} \in \{0,1\}^{n-1}} (h^g(1) + \gamma_1) \cdot (h^g(1, s_1) + \gamma_2) \cdot ... \cdot (h^g(1, s_{n-1}) + \gamma_n) - 1 \leq$$

$$\leq [h^g(0) + \gamma_1 + h^g(1) + \gamma_1] \cdot$$

$$\cdot \max \left\{ \begin{array}{c} \sum_{s_{n-1} \in \{0,1\}^{n-1}} (h^g(0, s_1) + \gamma_2) \cdot ... \cdot (h^g(0, s_{n-1}) + \gamma_n), \\ \sum_{s_{n-1} \in \{0,1\}^{n-1}} (h^g(1, s_1) + \gamma_2) \cdot ... \cdot (h^g(1, s_{n-1}) + \gamma_n) \end{array} \right\} - 1 =$$

$$= [1 + 2\gamma_1] \cdot$$

$$\cdot \max \left\{ \begin{array}{c} \sum_{s_{n-1} \in \{0,1\}^{n-1}} (h^g(0, s_1) + \gamma_2) \cdot ... \cdot (h^g(0, s_{n-1}) + \gamma_n), \\ \sum_{s_{n-1} \in \{0,1\}^{n-1}} (h^g(1, s_1) + \gamma_2) \cdot ... \cdot (h^g(1, s_{n-1}) + \gamma_n) \end{array} \right\} - 1.$$

We can estimate each sum in this last display in a similar manner to that we have used to estimate $\sum_{s_n \in \{0,1\}^n} (h^g(s_1) + \gamma_1) \cdot ... \cdot (h^g(s_n) + \gamma_n)$; we can continue in this fashion to conclude that

$$\left| P^f(U) - P^g(U) \right| \leq \left[ \prod_{t=1}^{n} (1 + 2\gamma_t) - 1 \right] < \frac{\delta}{4}.$$

Now, suppose that $U$ is the union of an arbitrary family of cylinders $\mathcal{C}$. Represent $U$ as

$$U = \bigcup_{n=1}^{\infty} U_n$$

where $U_n$ is the union of cylinders $C \in \mathcal{C}$ with base on $s_t$ with $t \leq n$. Since the sequence $\{U_n : n = 1, 2, ...\}$ is ascending, $\left| P^f(U) - P^f(U_n) \right| < \delta/8$ and $|P^g(U_n) - P^g(U)| < \delta/8$ for large enough $n$. Thus,

$$\left| P^f(U) - P^g(U) \right| \leq \left| P^f(U) - P^f(U_n) \right| +$$

$$+ \left| P^f(U_n) - P^g(U_n) \right| + |P^g(U_n) - P^g(U)| < \delta/2.$$

Finally, observe that $\left| P^f(A) - P^g(A) \right| < \delta$ for every $A \in \Im$. Indeed, take a set $U \supset A$, which is a union of cylinders, such that $\left| P^f(U) - P^f(A) \right|, |P^g(U) - P^g(A)| < \delta/4$. Since $\left| P^f(U) - P^g(U) \right| < \delta/2$,

$$\left| P^f(A) - P^g(A) \right| \leq \left| P^f(U) - P^f(A) \right| +$$

$$+\left|P^f(U)-P^g(U)\right|+\left|P^g(U)-P^g(A)\right|<\delta.$$

■

**Proof of Proposition 1** Define the test $T_m$ by

$$\begin{array}{ll} s_t\in T_m(f) & \text{if } t<m \text{ and } s_t\in T(f) \text{ or } t\geq m \text{ and } s_t\mid m\in T(f);\\ s_t\notin T_m(f) & \text{otherwise.}\end{array}$$

By assumption, $T$ does not reject the truth with probability $1-\varepsilon$. It implies that $T_m$ does not reject the truth with probability $1-\varepsilon$ because the test $T$ is harder than the test $T_m$. Since $T$ does not make use of future predictions, $T_m$ does not make use of future predictions either.

Define the test $T'_m$ by

$$\begin{array}{ll} s_t\notin T'_m(f) & \text{if } f\in F_m \text{ and } s_t\notin T_m(f);\\ s_t\in T'_m(f) & \text{otherwise.}\end{array}$$

By Lemma 3 (applied to $\delta/4$), $T'_m$ is a finite test that does not reject the informed expert with probability $1-\varepsilon-\delta/4$. It follows from Lemma 1 (applied to $\varepsilon'=\varepsilon+\delta/4$ and $\delta'=\delta/4$) that there exists a random generator of theories $\zeta_m\in\Delta\Delta(\Omega)$ such that for all finite histories $s_t\in S_\infty$,

$$\zeta_m(\{f\in F: s_t\notin T'_m(f)\})\geqslant 1-\varepsilon-\delta/2.$$

Notice now that the test $T'_{m+1}$ is harder than the test $T'_m$. Thus,

$$\zeta_m(\{f\in F: s_t\notin T'_k(f)\})\geqslant 1-\varepsilon-\delta/2 \text{ for all } m\geq k.$$

By the compactness of $\Delta\Delta(\Omega)$, there exists a convergent subsequence of the sequence $(\zeta_m)_{m=1}^\infty$, also indexed by $m$, with a limit $\zeta\in\Delta\Delta(\Omega)$, i.e., $\zeta_m\underset{m\to\infty}{\longrightarrow}\zeta$ (in the weak$-^*$ topology).

Fix an arbitrary finite history $s_t\in S_\infty$. We shall show that the set of all theories $f\in F$ with the property that $s_t\in T'_k(f)$ is open. To this end observe that we can assume without loss of generality that $t=k$. Indeed, for $t>k$, $s_t\in T'_k(f)$ if and only if $s_t\mid k\in T'_k(f)$; for $t<k$, $s_t\in T'_k(f)$ if and only if $s_k\in T'_k(f)$ for every $s_k$ with $s_k\mid t=s_t$, and there is only a finite number of such extensions $s_k$.

By the definition of $T'_k$, $s_k\in T'_k(f)$ for every $f\notin F_k$. Take now any $f\in F_k$. Since $T_k$ does not make use of future predictions, $T'_k$ does not make use of future

predictions either. Thus, it depends only on the predictions made by a theory $f$ up to period $k$ whether a history $s_k$ belongs to $T'_k(f)$.[18] Since $f \in F_k$, there is only a finite number of possible predictions

$$\left\{f(\widetilde{s}_0)\right\} \cup \left\{f(\widetilde{s}_1) : \widetilde{s}_1 \in \{0,1\}^1\right\} \cup ... \cup f(\widetilde{s}_k) : \widetilde{s}_k \in \{0,1\}^{k-1}\Big\}$$

that the theory $f$ can make up to period $k$. The set of possible predictions can be divided into two subsets, say $A$ and $B$, such that if a theory makes predictions from $A$, then $s_k \in T'_k(f)$, and if a theory makes predictions from $B$, then $s_k \notin T'_k(f)$. Thus, $\{f \in F : s_k \notin T'_k(f)\}$ consists of the theories that make predictions from the set $B$. By the finiteness of $B$, the set $\{f \in F : s_k \notin T'_k(f)\}$ is closed in in the weak$-^*$ topology.

Therefore, every set of the form $\{f \in F : s_k \in T'_k(f)\}$ is open in the weak$-^*$ topology. It follows now from Lemma 2 that $\xi(\{f \in F : s_t \in T'_k(f)\})$ is a lower semi-continuous function of $\xi \in \Delta\Delta(\Omega)$. Hence, there is an $\overline{m} \in N$ such that if $m \geq \overline{m}$ then

$$\zeta_m(\{f \in F : s_t \in T'_k(f)\}) \geq \zeta(\{f \in F : s_t \in T'_k(f)\}) - \delta/2,$$

and so $\zeta(\{f \in F : s_t \notin T'_k(f)\}) \geq 1 - \varepsilon - \delta$. Given that $T'_k$ is harder than $T_k$,

$$\zeta(\{f \in F : s_t \notin T_k(f)\}) \geq 1 - \varepsilon - \delta.$$

Let $\chi_k : F \longrightarrow \{0,1\}$ be the indicator function that is equal to 1 if $s_t \notin T_k(f)$ and zero otherwise. The last inequality can be written as

$$\int \chi_k d\zeta \geq 1 - \varepsilon - \delta.$$

Moreover, $\chi_k \downarrow \chi$ as $k$ goes to $\infty$, where $\chi : F \longrightarrow \{0,1\}$ is the indicator function equal to 1 when $s_t \notin T(f)$ and zero otherwise. By the monotone converge theorem,

$$\int \chi d\zeta \geq 1 - \varepsilon - \delta$$

which means that $\zeta(\{f \in F : s_t \notin T(f)\}) \geq 1 - \varepsilon - \delta$.■

[18] This is the only place in the proof, where we refer to the assumption that the test $T$ does not make use of counterfactual future predictions. However, this assumption is essential here. If a test $T'_k$ makes use of counterfactual future predictions, then the set $\{f \in F : s_t \in T'_k(f)\}$ is typically not open in the weak$-^*$ topology.

### 8.2. Proof of Propositions 2 and 3

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, let the sequence of conditional variances $VAR^P\{X_i \mid \Im_{i-1}\}$ be 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 4.** *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 $\delta > 0$,

$$P\left(\left\{s \in \Omega : \max_{1 \leq m \leq k} |S_m(s)| > \delta\right\}\right) \leq \frac{Var(S_k)}{\delta^2} \leq \frac{kM}{\delta^2}.^{19}$$

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,

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

[19] Shiryaev (1996) shows this result for independent random variables, but it's extension to martingales is well-known.

Hence,

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

■

In the proofs below we will identify the set $T(f)\subseteq S_\infty$ of finite histories considered to be inconsistent with the theory $f$ with the set

$$\bigcup_{s_t\in T(f)} C(s_t)$$

of infinite extensions of histories from $T(f)$, which can be interpreted as the set of infinite histories considered to be inconsistent with the theory $f$.

**Proof of Proposition 2** Let

$$Z_t(s)=\log\left(\frac{h^{f_\delta}(s_t)}{h^f(s_t)}\right),\ s_t=s\mid t.$$

Then, for some $\eta>0$ and for some $M>0$,

$$E^{P^f}\left\{Z_t\mid\Im_{t-1}\right\}<-\eta \text{ and } VAR^{P^f}\left\{Z_t\mid\Im_{t-1}\right\}<M.$$

The first inequality (on conditional expectation) follows directly from Smorodinsky (1971), Lemma 4.5, page 20, and Lehrer and Smorodinsky (1996), Lemma 2. The second inequality (on conditional variance) follows directly from the fact that $h^{f_\delta}(s_t)\in[\delta,1-\delta]$ and the fact that the functions $-plog(p)$ and $p\left(log(p)\right)^2$ are bounded on $[0,1]$.

Let $X_i=Z_i-E^{P^f}\left\{Z_i\mid\Im_{i-1}\right\}$. Let $j\in N$ be a natural number such that $\frac{1}{j}<\frac{\eta}{4}$. Let $\bar{m}(j,\,\varepsilon)$ be defined as in Lemma 4. The test $T$ is defined by

$$C(s_m)\subseteq T(f) \text{ if } \sum_{k=1}^{m}Z_k(s)>-m\frac{\eta}{2} \text{ whenever } m\geq\bar{m}(j,\varepsilon) \text{ and } s_m=s\mid m.$$

Note that

$$\left\{s\in\Omega:\forall_{m\geq\bar{m}(j,\varepsilon)}\quad\left|\frac{1}{m}\sum_{k=1}^{m}\left(Z_k(s)-E^{P^f}\left\{Z_k\mid\Im_{k-1}\right\}(s)\right)\right|\leq\frac{1}{j}\right\}\subseteq(T(f))^c$$

because

$$\frac{1}{m}\sum_{k=1}^{m}Z_k(s)\leq\frac{1}{j}-\eta<-\frac{\eta}{2}$$

implies

$$\sum_{k=1}^{m} Z_k(s) < -t\frac{\eta}{2}.$$

By Lemma 4,

$$P^f\left((T(f))^c\right) > 1-\varepsilon.$$

Hence, the test $T$ does not reject the truth with probability $1-\varepsilon$. By definition, the test $T$ does not use future predictions (in fact, whether the test $T$ rejects a theory or not at $s_t$ depends only on the forecasts $f(s_k)$, $s_k = s_t \mid k$, $k < t$). Finally, notice that

$$\sum_{k=1}^{t} Z_k(s) = \log\left(\frac{P^{f_\delta}(s_t)}{P^f(s_t)}\right),\ s_t = s \mid t.$$

Hence, if $s \notin T(f)$ then

$$\log\left(\frac{P^{f_\delta}(s_t)}{P^f(s_t)}\right) \underset{t\to\infty}{\longrightarrow} -\infty$$

which implies that $s \notin R(f)$.■

**Proof of Proposition 3** Let

$$X_t^i(s) = [f(s_{t-1}) - J_t(s_t)]\mathcal{I}_{t-1}^i,\ s_t = s \mid t,$$

and

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

Let now $\varepsilon_{j,i}$, $(j,i) \in N^2$, be such that $\varepsilon_{j,i} > 0$ and

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \varepsilon_{j,i} < \varepsilon.$$

Given that $E^{P_f}\{X_t^i \mid \Im_{t-1}\} = 0)$ and $VAR^P\{X_i \mid \Im_{i-1}\}$ are uniformly bounded, let $\bar{m}(j,\varepsilon_{j,k})$ be defined as in Lemma 4. The test $T'$ is defined by

$$C(s_m) \subseteq T'(f) \text{ if } \left|Y_m^i(s)\right| > \frac{1}{j} \text{ whenever } m \geq \bar{m}(j,\varepsilon_{j,i}).$$

By Lemma 4,

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} P^f\left(\left\{s \in \Omega : \left|Y_m^i(s)\right| > \frac{1}{j} \text{ for some } m \geq \bar{m}(j,\varepsilon_{j,i})\right\}\right) < \varepsilon.$$

So,

$$P^f\left((T'(f))^c\right) > 1-\varepsilon.$$

Hence, $T'$ does not reject the truth with probability $1-\varepsilon$. By definition, the test $T$ does not use future predictions. Finally, notice that $s \notin T'(f)$ implies

$$\left|Y_m^i(s)\right| \leq \frac{1}{j} \text{ for all } m \geq \bar{m}(j, \varepsilon_{j,i}) \text{ and } (j,i) \in N^2.$$

Hence, for all $i \in N, |Y_m^i(s)| \underset{m\to\infty}{\longrightarrow} 0$.■

## References


[1] Al-Najjar, N. and J. Weinstein (2006) "Comparative Testing of Experts," mimeo.

[2] Bewley, T. (1986) "Knightian Decision Theory, Part *1*," Cowles Foundation, Yale University.

[3] Casadesus-Masanell, R., P. Klibanoff, and E. Ozdenoren (2000), "Maxmin Expected Utility over Savage Acts with a Set of Priors," *Journal of Economic Theory* **92**, 35-65.

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

[5] Crawford, V. and J. Sobel (1982) "Strategic Information Transmission," *Econometrica*, **50**, 1431—1452.

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

[7] Dow, J. and G. Gorton (1997) "Noise Trading, Delegated Portfolio Management, and Economics Welfare," *Journal of Political Economy*, **105**, 1024-1050.

[8] Ehrbeck, T. and R. Waldmann (1996) "Why Are Professional Forecasters Biased? Agency versus Behavioral Explanations," *Quarterly Journal of Economics*, **111**, 21-40.

[9] Ellsberg, D. (1961) "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*, **75**, 643-669.

[10] Epstein, L. (1999) "A Definition of Uncertainty Aversion," *The Review of Economic Studies*, **66**, 579-608.

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

[12] Feinberg, Y. and C. Stewart (2006) "Testing Multiple Forecasters," mimeo.

[13] Fogel, R. (1967) "The Specification Problem in Economic History," *Journal of Economic History*, 27, 283-308.

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

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

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

[17] Gilboa, I. (1987) "Expected Utility with Purely Subjective Non-Additive Probabilities," *Journal of Mathematical Economics*, **16**, 65-88.

[18] Ghirardato, P., F.Maccheroni, and M. Marinacci (2004) "Differentiating Ambiguity and Ambiguity Attitude," *Journal of Economic Theory*, **118**, 133–173.

[19] Gilboa, I. and D. Schmeidler (1989) "Maxmin Expected Utility with A Non-Unique Prior," *Journal of Mathematical Economics*, **18**, 141–153.

[20] Hart, S. (2005) "Adaptative Heuristics," *Econometrica*, **73**, 1401 - 1430.

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

[22] Keane, M. and D. Runkle (1990) "Testing the Rationality of Price Forecasts: New Evidence from Panel Data," *American Economic Review*, **80**, 714-735.

[23] Keane, M. and D. Runkle (1998) "Are Financial Analysts' Forecasts of Corporate Profits Rational?" *Journal of Political Economy*, **106**, 768-805.

[24] Klibanoff, P., M. Marinacci, and S. Mukerji (2005), "A Smooth Model of Decision Making under Ambiguity," *Econometrica*, **73**, 1849-1892.

[25] Knight, F. (1921): *Risk, Uncertainty and Profit*, Houghton Mifflin, Boston.

[26] Laster, D., P. Bennett, and I. Geoum (1999) "Rational Bias in Macroeconomic Forecasts," *Quarterly Journal of Economics*, **114**, 293-318.

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

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

[29] Lehrer, E. and R. Smorodinsky (1996) "Compatible Measures and Merging," *Mathematics of Operations Research*, **21**, 697-706.

[30] LeRoy, S. and L. Singell (1987) "Knight on Risk and Uncertainty," *Journal of Political Economy*, **95**, 394-406.

[31] Lowell, M. (1986) "Tests of the Rational Expectations Hypothesis," *American Economic Review*, **76**, 110-154.

[32] Maccheroni, F., M. Marinacci, and A. Rustichini (2006), "Ambiguity Aversion, Robustness, and the Variational Representation of Preferences", forthcoming *Econometrica*.

[33] Medvec, V., S. Madey, and T. Gilovich (1995) "When Less is More: Counterfactual Thinking and Satisfaction among Olympic Medalists," *Journal of Personality and Social Psychology*, **69**, 603-610.

[34] McAfee, P. (1983) "American Economic Growth and the Voyage of Columbus," *American Economic Review*, **73**, 735-739.

[35] Morgan, J. and P. Stocken (2003) "An Analysis of Stock Recommendations," *RAND Journal of Economics*, **34**, 380-391.

[36] Olszewski W. (2006) "Preferences over Sets of Lotteries," forthcoming *Review of Economic Studies*.

[37] Olszewski W. and A. Sandroni (2006a) "Strategic Manipulation of Empirical Tests," mimeo.

[38] Olszewski, W. and A. Sandroni (2006b), "Contracts and Uncertainty," forthcoming *Theoretical Economics*.

[39] Pigou, A. (1927): *Industrial Fluctuations*, McMillam, London.

[40] Rustichini, A. (1999), "Minimizing Regret: The General Case," *Games and Economic Behavior*, **29**, 244-273.

[41] Savage, L. (1954): *The Foundations of Statistics*, Wiley, New York.

[42] Schmeidler, D. (1989) "Subjective Probability and Expected Utility Without Additivity," *Econometrica*, **57**, 571–587.

[43] Siniscalchi, M. (2006) "A Behavioral Characterization of Plausible Priors," *Journal of Economic Theory* **128**, 91-135.

[44] Sørensen, P. and M. Ottaviani (2006) "The Strategy of Professional Forecasting," *Journal of Financial Economics*, **81**, 441-466.

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

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

[47] Smorodinsky, M. (1971): *Ergodic Theory, Entropy*, Lecture Notes in Mathematics, Springer-Verlag.

[48] Trueman, B. (1988) "A Theory of Noise Trading in Securities Markets," *Journal of Finance*, **18**, 83-95.

[49] 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.

[50] Wakker, P. (1989) "Continuous subjective expected utility with non-additive probabilities," *Journal of Mathematical Economics* **18**, 1-27.