

# Testable Forecasts

Luciano Pomatto\*

March 2020

## Abstract

Predictions about the future are commonly evaluated through statistical tests. As shown by recent literature, many known tests are subject to adverse selection problems and cannot discriminate between forecasters who are competent and forecasters who are uninformed but predict strategically.

We consider a framework where forecasters' predictions must be consistent with a *paradigm*, a set of candidate probability laws for the stochastic process of interest. The paper presents necessary and sufficient conditions on the paradigm under which it *is* possible to discriminate between informed and uninformed forecasters. We show that optimal tests take the form of likelihood-ratio tests comparing forecasters' predictions against the predictions of a hypothetical Bayesian outside observer. In addition, the paper illustrates a new connection between the problem of testing strategic forecasters and the classical Neyman-Pearson paradigm of hypothesis testing.

---

\*E-mail: [luciano@caltech.edu](mailto:luciano@caltech.edu) - Division of the Humanities and Social Sciences, Caltech, Pasadena, CA, 91125. I am grateful to the editor, Rani Spiegler, and two anonymous referees, as well as Nabil Al-Najjar, Kim Border, Andres Carvajal, Eddie Dekel, Federico Echenique, Ithzak Gilboa, Johannes Horner, Nicolas Lambert, Wojciech Olszewski, Mallesh Pai, Larry Samuelson, Alvaro Sandroni, Colin Stewart and Max Stinchcombe for their helpful comments. I thank the Cowles Foundation for Research in Economics, where part of this research was completed, for its support and hospitality.

# 1 Introduction

Forecasts are often formulated in terms of probability distributions over future events (e.g., “a recession will happen with 5% probability”). Probabilistic forecasts appear across a wide variety of economic and scientific activities, including the analysis of weather and climate (Gneiting and Raftery, 2005), aggregate output and inflation (Diebold, Tay and Wallis, 1997), epidemics (Alkema, Raftery and Clark, 2007), seismic hazard (Jordan et al., 2011), financial risk (Timmermann, 2000), demographic variables (Raftery et al., 2012) and elections (Tetlock, 2005), among others.<sup>1</sup>

One practical difficulty with probabilistic forecasts is that they cannot be falsified by casual observation but only through proper statistical tests. From an economic perspective, a key issue is that statistical tests aimed at evaluating forecasters can be subject to adverse selection. Consider, as an example, a forecaster who is asked to predict how a stochastic process of interest will evolve over time, and is evaluated by a test comparing her prediction against the realized sequence of outcomes. Suppose the forecaster can be either a *true expert*, who knows the actual distribution  $P$  generating the data and is willing to report it truthfully, or a *strategic forecaster*, who is uninformed about the stochastic process but is interested in passing the test in order to establish a false reputation of competence. Recent literature shows that many tests of interest cannot discriminate between the two.

In their seminal paper, Foster and Vohra (1998) examine the well-known calibration test.<sup>2</sup> They construct a randomized forecasting algorithm that allows to pass the test regardless of how data unfold, and without any knowledge about the distribution of the data generating process. By employing such an algorithm, an uninformed but strategic forecaster can completely avoid being discredited by the data, thus defeating the purpose of the test.

This surprising phenomenon is not restricted to calibration. Subsequent work emphasizes one critical feature of the calibration test: the fact that it is free of Type-I errors. For any possible true law  $P$  generating the data, where  $P$  is an arbitrary probability measure defined over sequences of outcomes, an expert who predicts

---

<sup>1</sup>Corradi and Swanson (2006) and Gneiting and Katzfuss (2014) review the literature on probabilistic forecasts.

<sup>2</sup>Consider a stochastic process that every day can generate two outcomes, say “rain” and “no rain.” A forecaster passes the calibration test if, roughly, for every  $p \in [0, 1]$ , the empirical frequency of rainy days computed over the days where the forecaster predicted rain with probability  $p$  is close to  $p$ .

according to  $P$  will pass the calibration test with high probability (Dawid, 1982). This remarkable property ensures that the test is unlikely to reject competent forecasters. But, as shown by Sandroni (2003) and Olszewski and Sandroni (2009), once incentives are taken into account, the same property leads to a general impossibility result for testing probabilistic predictions: *any* test that operates in finite time and is free of Type-I errors can be passed by a strategic but uninformed forecaster. This impossibility result has been further extended by Olszewski and Sandroni (2008) and Shmaya (2008), among many others.

Tests such as calibration, that are free of Type-I errors, do not impose any restriction on the unknown law  $P$  generating the process. Such a degree of agnosticism is all but common in economics and statistics. Indeed, most empirical studies posit that data are generated according to a specific model, often fully specified up to a restricted set of parameters. This paper takes a similar approach, and examines the problem of testing forecasters in the presence of a theory about the data generating process.

This paper considers a framework where it is known that the law generating the data belongs to a given set  $\Lambda$ , which represents a theory, or *paradigm*, about the phenomenon under consideration. Accordingly, forecasters are required to submit forecasts belonging to  $\Lambda$ , while predictions incompatible with the paradigm are rejected.

For the purpose of this paper, paradigms admit multiple interpretations. A paradigm can be seen as a summary of pre existing knowledge about the problem. It can also represent the set of restrictions imposed on the data-generating process by a scientific theory. It can, alternatively, be interpreted as a normative standard to which forecasters' predictions must conform in order to qualify as useful. Classic examples of paradigms include the classes of i.i.d., Markov or stationary distributions. In this paper, in order to make the analysis applicable to a broad class of environments, no a priori restrictions are imposed over paradigms (beyond measurability).

A paradigm  $\Lambda$  is *testable* if it admits a test with the following three features. First, it is unlikely that the test will reject a true expert who knows the correct law in  $\Lambda$ . Second, for any possible strategy that a forecaster might employ to misrepresent her knowledge, there is a law belonging to  $\Lambda$  under which the forecaster will fail the test with high probability. Hence, strategic forecasters are not guaranteed to avoid rejection. Third, the test returns a decision (acceptance or rejection) in finite time.

A crucial question, then, is which paradigms are testable and, if they are, by using what tests. The existing literature has presented instances of testable classes

of distributions (see, among others, Olszewski and Sandroni, 2009, and Al-Najjar, Sandroni, Smorodinsky and Weinstein, 2010). However, reasonably general conditions under which a paradigm is testable are not known.

The first step of the analysis is a general characterization of testable paradigms. The result is formulated by taking the perspective of a hypothetical Bayesian outside observer. Given a paradigm  $\Lambda$ , consider, for the sake of illustration, an analyst, consumer or statistician who is uncertain about the odds of the data generating process, and who is sophisticated enough to express a prior probability  $\mu$  over the set of possible laws. The prior assigns probability 1 to the paradigm. It is shown that  $\Lambda$  is testable if and only if there exists at least one prior  $\mu$  such that the observer, by predicting according to the prior, is led to forecasts that are incompatible with any law in the paradigm. Formally, testability is equivalent to the existence of a prior  $\mu$  over the paradigm such that the law  $\int_{\Lambda} P d\mu(P)$  obtained by averaging with respect to the prior is sufficiently distant, in the appropriate metric (the total-variation distance), from every law  $P$  in the paradigm.

Based on this characterization we show that given any testable paradigm, it is without loss of generality to restrict the attention to standard likelihood-ratio tests. For every testable paradigm  $\Lambda$  there exists a finite likelihood-ratio test that is unlikely to reject a true expert and cannot be manipulated. Such tests are constructed as follows. First, the test creates a fictitious Bayesian forecaster. This forecaster is obtained by placing a sufficiently “uninformative” prior  $\mu$  over the paradigm. Actual forecasters are then evaluated by comparing their predictions to the forecasts generated by the test. A forecaster passes the test if only if the realized sequence of outcomes was, ex-ante, deemed more likely by the agent than by the fictitious Bayesian forecaster.

The results suggests a simple, and perhaps intuitive, criterion for identifying competent forecasters: a predictor is recognized as knowledgeable if her forecasts results more accurate, in likelihood-ratio terms, than the predictions of a Bayesian endowed with an uninformative prior.

The third main result of the paper shows that likelihood-ratio tests are in, a proper sense, optimal. The result is based on a novel ordering over tests. A test  $T$  is evaluated by the worst-case probability of passing the test an uninformed forecaster can guarantee herself, where the worst-case is computed over all possible laws for the data-generating process. A test  $T_1$  is *less manipulable than*  $T_2$  if such worst-case probability is lower under  $T_1$  than under  $T_2$ . So, less manipulable tests are more

effective at screening between informed and uninformed experts. Theorem 3 shows that for any paradigm, and controlling for sample size and for the level of Type-I error, there exists a likelihood-ratio test that is less manipulable than any other test. As explained in the main text, the result is related to the celebrated Neyman-Pearson lemma and highlights a novel connection between the problem of testing strategic forecasters and the theory of hypothesis testing.

In sum, the analysis of this paper provides a foundation for likelihood-ratio tests as a general methodology for testing probabilistic predictions under adverse selection.

Section 4 studies several examples of paradigms: Markov processes, Mixing processes, paradigms defined by moments inequalities, and maximal paradigms. For each example, we provide conditions such that the paradigm under consideration is testable. Section 5 discusses extensions and provides further comments on the related literature.

## 1.1 Related Literature

Foster and Vohra (2011) and Olszewski (2015) survey the literature on testing strategic forecasters.<sup>3</sup> In this section, we comment on those papers that are closer to the present work.

Likelihood-ratio tests appear in Al-Najjar and Weinstein (2008) as a method for comparing the predictions of two forecasters under the assumption that at least one of them is informed. Examples of testable paradigms appear in Olszewski and Sandroni (2009), who also extend the result of Sandroni (2003) to finite tests where the paradigm is convex and compact. Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010) study the set of laws that have a learnable and predictable representation, a class of distributions introduced by Jackson, Kalai and Smorodinsky (1999). They show that the paradigm is testable.

Babaioff, Blumrosen, Lambert and Reingold (2010) consider a principal-agent model where the principal offers a monetary contract with the intent of discriminating between informed and uninformed experts. They show, quite surprisingly, that screening is possible if and only if the true law is restricted to a non-convex set of distributions. This paper follows the literature on testing strategic experts where transfers are absent and the forecaster expected payoff is the probability of passing

---

<sup>3</sup>Contributions to the literature that are not included in the surveys include Sandroni and Shmaya (2014), Al-Najjar, Pomatto, and Sandroni (2014), Feinberg and Lambert (2015), and Kavalier and Smorodinsky (2017).

the test chosen by the tester. As a consequence, the two papers arrive at different conclusions. In particular, there exist non-convex paradigms that are not testable, and convex paradigms that are testable.<sup>4</sup>

Likelihood-ratio tests play an important role in Stewart (2011). Stewart proposes a framework where the tester is a Bayesian endowed with a prior over laws and the forecaster is evaluated according to a likelihood-ratio test against the predictions induced by the prior. In the current paper the tester is not assumed to be Bayesian. Instead, the existence of an appropriate prior which allows to construct a nonmanipulable likelihood-ratio test is shown to be a property that is intrinsic to all testable paradigm. The relation between this paper and Stewart (2011) is discussed more in detail in Section A.6 in the Appendix.

## 2 Basic Definitions

In each period an outcome from a finite  $X$  is realized, where  $|X| \geq 2$ . A *path* is an infinite set of outcomes and  $\Omega = X^\infty$  denotes the set of all paths. Time is indexed by  $n \in \mathbb{N}$ , and for each path  $\omega = (\omega_1, \omega_2, \dots)$  the corresponding finite history of length  $n$  is denoted by  $\omega^n$ . That is,  $\omega^n$  is the set of paths that coincide with  $\omega$  in the first  $n$  periods. We denote by  $\mathcal{F}_n$  the algebra generated by all histories of length  $n$  and by  $\mathcal{B}$  the  $\sigma$ -algebra generated by  $\bigcup_n \mathcal{F}_n$ . The set of paths  $\Omega$  is endowed with the product topology, which makes  $\mathcal{B}$  the corresponding Borel  $\sigma$ -algebra. We denote by  $\Delta(\Omega)$  the space of Borel probability measures on  $\Omega$ . Elements of  $\Delta(\Omega)$  will be interchangeably referred to as *laws* or *distributions*. The space  $\Delta(\Omega)$  is endowed with the weak\* topology and the corresponding Borel  $\sigma$ -algebra.<sup>5</sup> The same applies to the space  $\Delta(\Delta(\Omega))$  of Borel probability measures over  $\Delta(\Omega)$ . Given a measurable subset  $\Gamma \subseteq \Delta(\Omega)$ ,  $\Delta(\Gamma)$  is the set of Borel probability measures on  $\Delta(\Omega)$  that assign probability 1 to  $\Gamma$ .

---

<sup>4</sup>The paradigm studied in Al-Najjar, Sandroni, Smorodinsky and Weinstein (2010) is convex, but testable. Consider a binary process that in each period can take two values,  $x$  or  $y$ . The paradigm of all distributions such that the probability of observing  $x$  in the first period is restricted to be in  $[0, 0.25) \cup (0.75, 1]$  is not convex and not testable.

<sup>5</sup>A sequence  $(P_n)$  in  $\Delta(\Omega)$  converges to  $P$  in the weak\* topology if and only if  $\mathbb{E}_{P_n}[\phi] \rightarrow \mathbb{E}_P[\phi]$  for every continuous function  $\phi : \Omega \rightarrow \mathbb{R}$ . Given a measure  $P$ ,  $\mathbb{E}_P$  denotes the expectation operator with respect to  $P$ .

## 2.1 Empirical Tests

A *forecaster* announces a law  $P \in \Delta(\Omega)$ , under the claim that  $P$  describes how the data will evolve. A *tester* is interested in evaluating this claim using a statistical test.

**Definition 1** A *test* is a measurable function  $T : \Omega \times \Delta(\Omega) \rightarrow [0, 1]$ .

A test  $T$  compares the realized path  $\omega$  with the reported law  $P$ . The law is accepted if  $T(\omega, P) = 1$  and rejected if  $T(\omega, P) = 0$ . Values strictly between 0 and 1 describe randomized tests where the forecaster is accepted with probability  $T(\omega, P)$ .<sup>6</sup> The timing is as follows: (i) At time 0, the tester chooses a test  $T$ ; (ii) The forecaster announces a law  $P$ ; (iii) Nature generates a path  $\omega$ ; and (iv)  $T$  reports acceptance or rejection.

A test  $T$  is *finite* if for every law  $P$  there exists a time  $n_P$  such that  $T(\cdot, P)$  is measurable with respect to  $\mathcal{F}_{n_P}$ . That is, a law  $P$  is accepted or rejected as a function of the first  $n_P$  observations, where  $n_P$  is deterministic and known ex-ante. Throughout the paper we restrict the attention to finite tests. A relevant special case is given by the class of *non-asymptotic* tests, where there exists a single deadline  $N$  such that  $n_P \leq N$  for every  $P$ . While the main focus will be on asymptotic tests, in Section 5.1 we show that many of the results extend without difficulties to non-asymptotic tests.

## 2.2 Strategic Forecasting

The forecaster can be of two possible types. A *true expert* (or informed forecaster) knows the law governing the data generating process and is willing to report it truthfully. A *strategic* (or uninformed) *forecaster* does not possess any relevant knowledge about the data generating process. Her goal is simply to pass the test. Strategic forecasters can produce their predictions using mixed strategies. Formally, a *strategy* is a randomization over laws  $\zeta \in \Delta(\Delta(\Omega))$ .

The next example shows how a standard likelihood-ratio test can be manipulated by strategic forecasters.

**Example 1.** (*A manipulable likelihood-ratio test*) The test is specified by a time  $n$  and a probability measure  $Q \in \Delta(\Omega)$  with full support. The law  $Q$  serves as a

---

<sup>6</sup>Except for Theorem 3 below, none of the results are affected by restricting the attention to non-randomized tests.

benchmark against which the forecaster is compared. Given a forecast  $P$  and a path  $\omega$ , the test returns 1 if

$$\frac{P(\omega^n)}{Q(\omega^n)} > 1 \tag{1}$$

and 0 otherwise, where, as defined above,  $\omega^n$  is the set of paths that coincide with  $\omega$  in the first  $n$  periods. Thus, the forecaster passes the test if and only if the realized history is more likely under the forecast  $P$  than under the benchmark  $Q$ . The test can be manipulated using the following simple strategy. For each history  $\omega^n$  of length  $n$ , consider the measure  $P_{\omega^n} = Q(\cdot | \Omega - \omega^n)$  obtained by conditioning  $Q$  on the complement of  $\omega^n$ . It satisfies

$$P_{\omega^n}(\omega^n) = 0 \text{ and } P_{\omega^n}(\tilde{\omega}^n) > Q(\tilde{\omega}^n) \text{ for all } \tilde{\omega}^n \neq \omega^n.$$

Let  $\zeta$  be the mixed strategy that randomizes uniformly over all measures of the form  $P_{\omega^n}$ . Given a history  $\omega^n$ , a forecaster using strategy  $\zeta$  will pass the test as long as the law she happens to announce is different from  $P_{\omega^n}$ . This is an event that under  $\zeta$  has probability greater or equal than  $1 - 2^{-n}$ . So, no matter how the data will unfold, even for  $n$  relatively small, the forecaster is guaranteed to pass the test with high probability.

The test in Example 1 does not assume any structure on the data-generating process. In this example, the freedom of announcing any law allows the uninformed predictor to manipulate the test. We will see how appropriate restrictions on the domain of possible laws will allow even simple likelihood-ratio tests to screen between informed and uninformed forecasters.

### 2.3 Testable Paradigms

The tester operates under a theory, or *paradigm*, about the data generating process. In this paper a theory is identified with the restrictions it imposes over the law of the observed process. Formally, a paradigm is a measurable set  $\Lambda \subseteq \Delta(\Omega)$ , with the interpretation that the data are generated according to some unknown law belonging to  $\Lambda$ . Beyond measurability, no assumptions are imposed on  $\Lambda$ .

A paradigm can be defined in many ways. For instance, it can express statistical independence between different variables (“the outcome  $\omega_n$  realized at time  $n$  is independent from the outcome realized at time  $n+365$ ”) or it might reflect assumptions



about the long run behavior of the process (“ $P$  is ergodic”). Additional examples will be discussed in Section 4.

Given a paradigm, a basic property we would like a test to satisfy is to not reject informed experts.

**Definition 2** Given a paradigm  $\Lambda$ , a nonrandomized test  $T$  *accepts the truth with probability at least  $1 - \epsilon$*  if for all  $P \in \Lambda$  it satisfies

$$P(\{\omega : T(\omega, P) = 1\}) \geq 1 - \epsilon. \quad (2)$$

A test that accepts the truth is likely to not reject an expert who reports the actual law of the data generating process. As shown by Olszewski and Sandroni (2009), any finite test that accepts the truth with respect to the unrestricted paradigm  $\Lambda = \Delta(\Omega)$  can be manipulated: Given a finite test  $T$  that satisfies property (2) for all  $P \in \Delta(\Omega)$ , there exists a strategy  $\zeta$  such that

$$\zeta(\{P : T(\omega, P) = 1\}) \geq 1 - \epsilon \text{ for all paths } \omega \in \Omega.$$

Thus, the strategy allows the forecaster to completely avoid rejection. The result motivates the next definition.

**Definition 3** Given a paradigm  $\Lambda$ , a non-randomized test  $T$  is  $\epsilon$ -*nonmanipulable* if for every strategy  $\zeta$  there is a law  $P_\zeta \in \Lambda$  such that

$$(P_\zeta \otimes \zeta)(\{(\omega, P) : T(\omega, P) = 1\}) \leq \epsilon.$$

The notation  $P_\zeta \otimes \zeta$  stands for the independent product of  $P_\zeta$  and  $\zeta$ . A test  $T$  is  $\epsilon$ -nonmanipulable if for any strategy  $\zeta$  there is a law  $P_\zeta$  in the paradigm such that the forecaster is rejected with probability greater than  $1 - \epsilon$ . Thus, no strategy can guarantee a strategic forecaster more than an  $\epsilon$  probability of passing the test.

As discussed by Olszewski and Sandroni (2009b), nonmanipulable tests can screen out uninformed forecasters. To elaborate, assume that a forecaster who opts not to participate in the test receives a payoff of 0, while a forecaster announcing a law  $P$  obtains a payoff that depends on the outcome of the test. If  $P$  is accepted then she is recognized as knowledgeable and gets a payoff  $w > 0$ . If the law is rejected then she is discredited and incurs a loss  $l < 0$ . Assume, in addition, that an uninformed

forecaster chooses in accordance with the maxmin criterion of Wald (1950) and Gilboa and Schmeidler (1989), where each strategy  $\zeta$  is evaluated according to the minimum expected payoff with respect to a set of laws. If such a set equals the paradigm, then for each strategy  $\zeta$  the expected payoff is<sup>7</sup>

$$\inf_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta} [wT + l(1 - T)] \quad (3)$$

If  $\epsilon$  is sufficiently small, then the value (3) is negative and so the optimal choice for a strategic forecaster is to not take the test. Therefore, given a test that accepts the truth with probability at least  $1 - \epsilon$  and is  $\epsilon$ -nonmanipulable, a true expert finds profitable to participate in the test, while for an uninformed expert it is optimal not to participate.<sup>8</sup>

Definitions 2 and 3 extend immediately to possibly randomized tests. Given a paradigm  $\Lambda$ , a test  $T$  *accepts the truth with probability at least  $1 - \epsilon$*  if for every  $P \in \Lambda$  it satisfies  $\mathbb{E}_P[T(\cdot, P)] \geq 1 - \epsilon$ . The test is  *$\epsilon$ -nonmanipulable* if for every strategy  $\zeta$  there is a law  $P_\zeta \in \Lambda$  such that  $\mathbb{E}_{P_\zeta \otimes \zeta}[T] \leq \epsilon$ . The next definition summarizes the properties introduced so far.

**Definition 4** Given  $\epsilon > 0$ , a paradigm  $\Lambda$  is  *$\epsilon$ -testable* if there is a finite test  $T$  such that:

1.  $T$  accepts the truth with probability at least  $1 - \epsilon$ ; and
2.  $T$  is  $\epsilon$ -nonmanipulable.

A paradigm  $\Lambda$  is *testable* if it is  $\epsilon$ -testable for every  $\epsilon > 0$ .

### 3 Main Results

It will be useful, in what follows, to take the perspective of a Bayesian outside observer (e.g. an analyst, a voter, or a statistician) who is interested in the problem at hand and uncertain about the odds governing the data generating process. The uncertainty perceived by the observer is expressed by a prior probability  $\mu \in \Delta(\Gamma)$ , where  $\Gamma \subseteq \Delta(\Omega)$  is the set of laws the observer believes to be possible. We focus on

---

<sup>7</sup>In what follows,  $\mathbb{E}_{P \otimes \zeta}$  denotes the expectation with respect to  $P_\zeta \otimes \zeta$ .

<sup>8</sup>Section 8.3 considers a different specification where uninformed forecasters are less conservative and, in (3), the worst case scenario is taken with respect to a neighborhood of laws in the paradigm.

the case where  $\Gamma$  coincides with (or is close to) the paradigm  $\Lambda$ , so that the observer and the tester have compatible views. If asked to make forecasts about the future, the observer would predict according to the probability measure defined as

$$Q_\mu(E) = \int_\Gamma P(E) d\mu(P) \quad \text{for all } E \in \mathcal{B}. \quad (4)$$

The definition (4) follows the standard approach in Bayesian statistical decision theory of defining, from the prior  $\mu$ , a probability measure over the sample space  $\Omega$  by averaging with respect to the prior.<sup>9</sup>

### 3.1 Characterization

The next result characterizes testable paradigms. Given laws  $P$  and  $Q$ , let  $\|P - Q\| = \sup_{E \in \mathcal{B}} |P(E) - Q(E)|$  denote the (normalized) total-variation distance between the two measures. Given a paradigm  $\Lambda$ , its closure with respect to the weak\* topology is denoted by  $\bar{\Lambda}$ .

**Theorem 1** *A paradigm  $\Lambda$  is testable if and only if for every  $\epsilon > 0$  there exists a prior  $\mu \in \Delta(\bar{\Lambda})$  such that  $\|Q_\mu - P\| \geq 1 - \epsilon$  for all  $P \in \Lambda$ .*

Consider an outside observer whose prior assigns probability 1 to (the closure of)  $\Lambda$ . The result compares the observer's forecasts with the paradigm. Two polar cases are of interest. If  $Q_\mu \in \Lambda$ , then the observer's prediction cannot be distinguished, ex-ante, from the prediction of an expert who announced  $Q_\mu$  knowing it was the true law of the process. Theorem 1 is concerned with the opposite case, where the prediction  $Q_\mu$  is far from *any* possible law  $P$  in the paradigm. It shows that a paradigm is testable if and only if there is some observer whose uncertainty about the data generating process leads her to predictions that are incompatible (in the sense of being far with respect to the total-variation distance) with respect to any law in the paradigm.

Given a prior  $\mu$  with the above properties, it is possible to define an explicit non-manipulable test. In the next section we provide a direct construction of such a test, together with an intuition for the result. The intuition for why testability of a paradigm implies the existence of a prior that satisfies the conditions of Theorem

---

<sup>9</sup>In the literature,  $Q_\mu$  is often referred to as a *predictive* probability. Cerreia-Vioglio, Maccheroni, and Marinacci (2013) provide, under appropriate conditions on  $\Gamma$ , an axiomatic foundation for the representation (4).

1 can be sketched as follows. For a strategic forecaster, randomization is valuable because it allows to increase the probability of passing the test in the worst-case, across all possible distributions that belong to the paradigm. Naturally, to different strategies will correspond different worst-case distributions. For a given strategy  $\zeta$  it is irrelevant whether the worst-case is computed within the paradigm, or across the set of all distributions of the form  $Q_\mu$  for some prior  $\mu$ . This follows from the fact that the forecaster’s “payoff function” is given by the expectation of  $T$ , hence it is linear in the randomization  $\zeta$  and in the law  $P$ . However, as we show in the proof of Theorem 1, considering the set of laws  $Q_\mu$  that can be achieved by some prior  $\mu$  is important. In the proof we show that given a test, there exists a worst-case distribution  $Q_\mu$  that is common to all strategies. Intuitively, this worst-case distribution must not be within the paradigm, since otherwise a forecaster could simply announce it and pass the test. The result shows that, in a specific sense, it must be sufficiently far away from the paradigm.

Testability of a paradigm is a property which can be formulated as a lack of compactness and convexity. To illustrate this idea we associate to each paradigm  $\Lambda$  an index  $I(\Lambda)$  of its compactness and convexity. The definition is based on notions introduced in the context of general equilibrium theory by Folkmann, Shapley, and Starr (see Starr, 1969). Given a subset  $\Lambda \subseteq \Delta(\Omega)$ , let

$$I(\Lambda) = \sup_{Q \in \overline{\text{co}}(\Lambda)} \inf_{P \in \Lambda} \|Q - P\|$$

where  $\overline{\text{co}}(\Lambda)$  is the weak\*-closed convex hull of  $\Lambda$ .  $I$  satisfies  $0 \leq I(\Lambda) \leq 1$  by the definition of the total-variation distance. If  $I(\Lambda) = 0$ , then any law  $Q$  in the closed convex hull of the paradigm can be approximated with arbitrary precision by a law  $P$  in  $\Lambda$ . In this case, as shown by Olszewski and Sandroni (2009), any finite test that accepts the truth is manipulable.<sup>10</sup> In the opposite case, when  $I(\Lambda) = 1$ , one can find a law in the closed convex hull of  $\Lambda$  that has distance arbitrarily close to 1 from every law in the paradigm. The next result shows that this is true if and only if the paradigm is testable.

---

<sup>10</sup>The intuition behind the result can be sketched as follows. Finiteness of the test, together with compactness and convexity of  $\Lambda$ , allow to invoke Fan’s minmax theorem and establish the equality  $\min_{P \in \Lambda} \max_{\zeta} \mathbb{E}_{P \otimes \zeta}[T] = \max_{\zeta} \min_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta}[T]$ . If  $T$  accepts the truth with probability  $1 - \epsilon$  then the left-hand side is greater than  $1 - \epsilon$ . Hence there exists a strategy that passes the test with probability  $1 - \epsilon$  for every  $P \in \Lambda$ . Therefore, the paradigm is not testable.

**Corollary 1** *A paradigm  $\Lambda$  is testable if and only if it satisfies  $I(\Lambda) = 1$ .*

### 3.2 Nonmanipulable Tests

Next we study non-manipulable tests. By applying the characterization provided by Theorem 1, we show that given a testable paradigm it is without loss of generality to restrict the attention to simple likelihood-ratio tests:

**Theorem 2** *Let  $\Lambda$  be a testable paradigm. Given  $\epsilon > 0$ , let  $\mu \in \Delta(\bar{\Lambda})$  be a prior that satisfies  $\|Q_\mu - P\| > 1 - \epsilon$  for all  $P \in \Lambda$ . There exist positive integers  $(n_P)_{P \in \Lambda}$  such that the test defined as*

$$T(\omega, P) = \begin{cases} 1 & \text{if } P \in \Lambda \text{ and } P(\omega^{n_P}) > Q_\mu(\omega^{n_P}) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

*accepts the truth with probability at least  $1 - \epsilon$  and is  $\epsilon$ -nonmanipulable.*

Given a law  $P$ , the test reaches a decision after  $n_P$  observations, where  $n_P$  is known in advance. The forecaster passes the test if and only if the history realized at time  $n_P$  is strictly more likely under  $P$  than under the law  $Q_\mu$ . The prior  $\mu$  is required to be sufficiently “uninformative” so that the induced law  $Q_\mu$  is far from every law in the paradigm. As implied by Theorem 1, such a prior exists whenever the paradigm is testable.

The likelihood-ratio test is one of the most well-known statistical tests.<sup>11</sup> It is therefore reassuring that all testable paradigms can be unified under the same, canonical, family of tests.

The main idea behind the proof of Theorem 2 is to exploit a key relation between likelihood-ratio tests and the total-variation distance. To illustrate, let  $A^P$  be the set of paths where the law  $P \in \Lambda$  passes the test (5), and consider the difference in probability  $P(A^P) - Q_\mu(A^P)$ . It can be shown that by taking  $n_P$  large enough, this difference approximates the distance  $\|P - Q_\mu\|$  between the two measures. Hence, the event  $A^P$  must have probability higher than  $1 - \epsilon$  under  $P$ , so the test accepts the truth with high probability. In addition,  $A^P$  must have probability at most  $\epsilon$  under  $Q_\mu$ . Because this is true for every  $P$ , then, in the hypothetical scenario where the data were generated according to  $Q_\mu$ , a forecaster would be unlikely to pass the test

---

<sup>11</sup>See, for instance, Lehmann and Romano (2006) for an introduction to the likelihood-ratio test.

regardless of what law is announced and, therefore, regardless of whether or not she randomizes her prediction. It follows from this observation and from the fact that  $Q_\mu$  is a mixture of laws in the paradigm, that against every fixed randomization  $\zeta$  there must exist some law  $P_\zeta$  in the paradigm against which passing the test is unlikely. That is, the test cannot be manipulated.

### 3.3 The Optimality of Likelihood Tests: a Neyman-Pearson Lemma

Theorem 2 shows that simple likelihood-ratio tests can screen between informed and uninformed forecasters. However, it leaves open the possibility that such tests are inefficient in the number of observations they require. A natural question is whether there exist tests that for a fixed sample size can outperform likelihood-ratio tests in screening between experts and strategic forecasters. We now make this question precise by introducing a new ordering over tests.

**Definition 5** Let  $\Lambda$  be a paradigm. Given tests  $T_1$  and  $T_2$ , the test  $T_1$  is *less manipulable than*  $T_2$  if

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta} [T_1] \leq \sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta} [T_2]. \quad (6)$$

Consider a strategic forecaster who is confronted with a test  $T$  and must choose whether or not to undertake the test. As discussed in Section 2, an uninformed forecaster will participate only if the value  $\sup_{\zeta \in \Delta(\Lambda)} \inf_{P \in \Lambda} \mathbb{E}_{\zeta \otimes P} [T]$ , which is proportional to the maxmin expected payoff from taking the test, is sufficiently large. So, the left-hand side of (6) is proportional to the highest expected payoff a strategic forecaster can guarantee when facing test  $T_1$ .

The ranking (6) requires that any strategic forecaster who finds optimal not to participate in the test  $T_2$  must also find optimal not to participate in the test  $T_1$ . Hence, any uninformed forecaster who is screened out by the test  $T_2$  is also screened out by the test  $T_1$ . In other terms, a less manipulable test has a greater deterrent effect against strategic forecasters.

A comparison between tests is more informative when some variables, such as the required number of observations, are kept fixed. To this end, we call a collection  $(n_P)_{P \in \Lambda}$  of positive integers a collection of *testing times* if the map  $P \mapsto n_P$  is measurable. A test  $T$  is *bounded* by the testing times  $(n_P)_{P \in \Lambda}$  if  $T(\cdot, P)$  is a function

of the first  $n_P$  observations. The definition allows for the possibility that different predictions may need different sample sizes in order to be properly tested. Finally, given a class  $\mathcal{T}$  of tests, we say that a test  $T$  is *least manipulable in  $\mathcal{T}$*  if it belongs to  $\mathcal{T}$  and is less manipulable than any other test in the same class. We can now state the main result of this section.

**Theorem 3** *Fix a paradigm  $\Lambda$ , testing times  $(n_P)_{P \in \Lambda}$  and a probability  $\alpha \in [0, 1]$ . There exists a prior  $\mu^* \in \Delta(\bar{\Lambda})$ , thresholds  $(\lambda_P)_{P \in \Lambda}$  in  $\mathbb{R}_+$  and a test  $T^*$  such that:*

1.  $T^*(\omega, P) = 1$  if  $P \in \Lambda$  and  $P(\omega^{n_P}) > \lambda_P Q_{\mu^*}(\omega^{n_P})$ ;
2.  $T^*(\omega, P) = 0$  if  $P \notin \Lambda$  or  $P(\omega^{n_P}) < \lambda_P Q_{\mu^*}(\omega^{n_P})$ ; and
3.  $T^*$  is least manipulable in the class of tests that are bounded by  $(n_P)$  and accept the truth with probability at least  $\alpha$ .

Theorem 3 is a general result illustrating the optimality of likelihood-ratio tests. Given the number of data points  $n_P$  that the tester is willing to collect for each forecast  $P$ , and given a lower bound  $\alpha$  on the probability of accepting a true expert, there exists a likelihood-ratio test that is less manipulable than any other test that satisfies the same constraints.

The result does not demand any assumptions on the paradigm, which is not required to be testable. Another difference with the test introduced in Theorem 2 is the use of law-specific thresholds  $\lambda_P$  which allow to adjust the probability of accepting a true expert as a function of the desired level  $\alpha$  of Type-I errors.<sup>12</sup>

The result is based on a novel connection between the problem of testing strategic forecasters and the statistical hypothesis testing literature. To illustrate this idea, consider the standard problem of testing a null hypothesis  $P_0$  against an alternative hypothesis  $P_1$ , where  $P_0$  and  $P_1$  are two given probability measures over paths. To be clear, in such a context a (possibly randomized) *hypothesis test* is a function  $\phi : \Omega \rightarrow [0, 1]$ , where  $\phi(\omega)$  is the probability of accepting  $P_0$  given the path  $\omega$ .

The test  $T^*$  is formally equivalent to a hypothesis test where the law  $P$  produced by the expert plays the role of the null hypothesis while the outside observer's

---

<sup>12</sup>The proof of Theorem 3 provides a complete description of the test  $T^*$  and illustrates how the thresholds and the prior  $\mu^*$  are computed. In the knife-edge case where  $P(\omega^{n_P}) = \lambda_P Q_{\mu^*}(\omega^{n_P})$  the test is randomized. The use of randomized tests greatly simplifies the analysis and allows the tester to achieve a probability of accepting a true expert that is exactly equal to  $\alpha$ .

prediction  $Q_{\mu^*}$  plays the role of the alternative. The crucial difference with the standard hypothesis testing framework is that the two “hypotheses”  $P$  and  $Q_{\mu^*}$  are not given exogenously:  $P$  is produced by a possibly strategic forecaster while  $Q_{\mu^*}$  is chosen by the tester.

The celebrated Neyman and Pearson lemma shows that given two hypotheses  $P_0$  and  $P_1$ , and given an upper bound on the probability of Type I error, there exists a likelihood-ratio test between  $P_0$  and  $P_1$  that minimizes the probability of Type II errors. The proof of Theorem 3 applies this fundamental result to the problem of strategic forecasters. The proof proceeds in two steps. First, the belief  $\mu^*$  is obtained as the solution of a nonlinear minimization problem over the space of priors. The test  $T^*$  is then defined by applying the Neyman-Pearson Lemma to each pair of laws  $P$  and  $Q_{\mu^*}$ . The key step is to show that because of the particular choice of  $\mu^*$ , a test which minimizes the probability of Type-II errors with respect to  $Q_{\mu^*}$  is also a test that is least manipulable.

## 4 Examples and Properties Related to Testability

In this section we analyze examples of paradigms. For each example, we provide conditions under which the paradigm under consideration is testable.

### 4.1 Markov Processes

We first consider Markov processes. The law of a Markov process is described by a transition probability  $\pi : X \rightarrow \Delta(X)$  and an initial probability  $\rho \in \Delta(X)$ . We denote by  $\Pi = \Delta(X)^X$  the set of all transition probabilities. Every pair  $(\rho, \pi)$  induces a Markov distribution  $P_{\rho, \pi} \in \Delta(\Omega)$ . We denote such a law by  $P_\pi$  whenever  $\rho$  is uniform.

Consider a Bayesian outside observer who is uncertain about the transition probability of the process and believes the true law to be  $P_\pi$  for some  $\pi$ . Let  $m$  be a Borel probability measure over  $\Pi$  that for every  $c \in [0, 1]$  and  $x, y \in X$  satisfies  $m(\{\pi : \pi(x)(y) = c\}) = 0$ . In particular,  $m$  is non-atomic.<sup>13</sup> By taking  $\pi$  to be distributed according to  $m$ , we obtain, implicitly, a prior  $\mu$  defined over the set of

---

<sup>13</sup>For instance, if  $X = \{x, y\}$  then we can define  $m$  by setting  $\pi(x)(y)$  and  $\pi(y)(x)$  to be independent and uniformly distributed over  $(0, 1)$ .



Markov distributions such that the resulting law  $Q_\mu$  satisfies

$$Q_\mu(E) = \int_{\Pi} P_\pi(E) dm(\pi) \text{ for all } E \in \mathcal{B}$$

The next result follows by applying standard asymptotic results for Markov processes.

**Proposition 1** *The prior  $\mu$  satisfies  $\|Q_\mu - P_{\rho,\pi}\| = 1$  for all Markov  $P_{\rho,\pi}$ . Hence, the paradigm of Markov distributions is testable.*

It follows from Theorem 2 that the paradigm of Markov distributions is testable by means of a likelihood-ratio test defined with respect to the law  $Q_\mu$ . Under this test, forecasters' predictions are compared against the predictions of a Bayesian who is endowed with a non-atomic prior over the true transition probabilities.

## 4.2 Asymptotic Independence

There is considerable interest, in the analysis of economic time series, in dependence conditions that go beyond independence. A common assumption is *mixing*, which expresses the idea that two events are approximately independent provided they occur sufficiently far apart in time. Mixing is a generalization of the i.i.d. assumption which has found applications in econometrics and in the forecasting literature.<sup>14</sup>

For every  $k \in \mathbb{N}$  denote by  $\mathcal{F}_k^\infty$  the  $\sigma$ -algebra generated by the coordinate random variables  $(Z_k, Z_{k+1}, \dots)$ , where, for every  $m \geq 1$ ,  $Z_m(\omega) = \omega_m$  is the outcome in period  $m$ . So,  $\mathcal{F}_k^\infty$  is the collection of all events that do not depend on the first  $k - 1$  realizations of the process. A law  $P$  is *mixing* if for every history  $\omega^n$  it satisfies  $P(\omega^n) > 0$  and

$$\sup_{A \in \mathcal{F}_k^\infty} |P(A|\omega^n) - P(A)| \rightarrow 0 \text{ as } k \rightarrow \infty. \quad (7)$$

Under a mixing measure  $P$ , the information  $\omega^n$  known at time  $n$  has a negligible effect on the probability  $P(A)$  of an event  $A$ , if  $A$  depends on realizations of the process which will occur only in the far enough future.

The fact that a paradigm consists of laws that are mixing does not, without further assumptions, imply that the same paradigm is testable. For instance, if the laws in  $\Lambda$

---

<sup>14</sup>See, for instance, Davidson (1994) and Nze and Doukhan (2004), and the reference therein, for the role of mixing and its generalizations in the analysis of time series, and Giacomini and White (2006, Theorem 1) for an example of applications of mixing in the forecasting literature. In this section, we study mixing in the context of strategic forecasting.

disagree only about the odds of the first realization  $\omega_1$  of the process, testability is not achieved. The next result provides an elementary richness condition that, when added to the mixing assumption, ensures that the paradigm is testable.

For the next result, recall that the measures  $P_1, \dots, P_n$  are *orthogonal* if they satisfy  $\|P_i - P_j\| = 1$  for all  $i \neq j$ . We say  $P_1, \dots, P_n$  are *strongly orthogonal* if for every pair  $P_i$  and  $P_j$ , and every  $k \in \mathbb{N}$ , it holds that  $\sup_{E \in \mathcal{F}_k^\infty} |P_i(E) - P_j(E)| = 1$ .

That is, for any  $k$ , any two distinct measures  $P_i$  and  $P_j$  fully disagree about the probability of some event that does not depend on the first  $k - 1$  realizations. Thus, two laws are strongly orthogonal if they disagree about the probability of events that are arbitrarily far in the future.

**Proposition 2** *Let  $\Lambda$  be a paradigm such that each  $P \in \Lambda$  is mixing and for every  $n \in \mathbb{N}$  there are laws  $P_1, \dots, P_n$  in  $\Lambda$  that are strongly orthogonal. Then  $\Lambda$  is testable.*

In the proof, given  $n$ , we consider a prior  $\mu_n$  that is uniform over  $n$  orthogonal laws  $P_1, \dots, P_n$  in  $\Lambda$ , and show that the induced distribution  $Q_{\mu_n}$  has total-variation distance of at least  $1/n$  from every law in  $\Lambda$ . Hence, by Theorem 2, a non-manipulable test can be obtained by a likelihood-ratio test with respect to  $Q_{\mu_n}$ , for  $n$  suitably large.<sup>15</sup>

The result implies, in particular, that the paradigm of all mixing processes is testable. This is because i.i.d. laws are mixing and the collection of all i.i.d. distributions satisfies the richness condition. Indeed, by the strong law of large numbers, any two distinct i.i.d. laws assign probability 1 to different limiting frequencies. They are therefore strongly orthogonal.

However, the main contribution of Proposition 2 is showing that *any* set of mixing distributions that satisfies the above richness condition is testable. This is an important difference, since, in applications, mixing is usually coupled with additional conditions which further restrict the paradigm under consideration (e.g. assumptions on the rate of convergence in (7), or parametric assumptions on the functional form of the process. See, for instance, Davidson, 1994), and a subset of a testable paradigm is not necessarily testable.<sup>16</sup>

---

<sup>15</sup>Notice that the mixing assumption is crucial for the result. The paradigm  $\Lambda = \Delta(\Omega)$  trivially satisfies the richness assumption but is not testable.

<sup>16</sup>The class of irreducible and aperiodic Markov processes consists of distributions that are mixing, and contains the class of all i.i.d. distributions. It is therefore testable by Proposition 2. Because it is a strict subset of the paradigm of Markov distributions, the result does not follow directly from Proposition 1.

The same observation also explains the relation with Al-Najjar, Sandroni, Smorodinsky, and Weinstein (2010) who study the paradigm of *asymptotically reverse mixing*. This paradigm contains deterministic, i.i.d. and Markov laws. In fact, it is even larger than the class of mixing distributions. Because a subset of a testable paradigm is not necessarily testable, the results in Al-Najjar, Sandroni, Smorodinsky, and Weinstein (2010) do not directly imply (nor are implied) by Propositions 1 or 2.

### 4.3 Uncertainty Sets

A classic topic in the economics of uncertainty is the study of decision making in the presence of ambiguity about the correct law generating the data. We focus here on the approach taken by Hansen and Sargent (2001). In their work, they study non-Bayesian decision makers who are incapable, or unwilling, to formulate a single belief, and who instead consider a set of possible probabilistic models obtained by perturbing a baseline measure  $P$ . In this section, we ask whether such decision makers can, in principle, learn the correct odds of the process by testing the predictions of an expert. We will see that the answer is negative: a paradigm defined according to the methodology of Hansen and Sargent (2001) is, in general, not testable.

Given two laws  $P, Q \in \Delta(\Omega)$  we denote by  $D(P\|Q)$  the corresponding Kullback-Leibler divergence. It is defined as  $D(P\|Q) = \int_{\Omega} \log(dP/dQ) dP$  if  $P$  is absolutely continuous with respect to  $Q$ , and  $D(P\|Q) = +\infty$  otherwise. The Kullback-Leibler divergence is a well known index that measures how difficult it is to distinguish the two distributions.

Hansen and Sargent (2001) study decision makers who consider a set of candidate laws for the process—a set of distributions that are close in Kullback-Leibler divergence to a reference measure  $P$ . The measure  $P$  can be seen as first guess for the correct distribution of the process. In the next proposition we study paradigms  $\Lambda$  that satisfy a similar property.

**Proposition 3** *Let  $P \in \Delta(\Omega)$  and  $\alpha > 0$ . Any paradigm  $\Lambda$  that satisfies  $P \in \Lambda$  and*

$$\Lambda \subseteq \{\tilde{P} \in \Delta(\Omega) : D(\tilde{P}\|P) \leq \alpha\},$$

*is not testable.*

Underlying the result is the following intuition. A key property of the Kullback-Leibler divergence is its convexity. This implies, in particular, that any measure  $Q_{\mu}$

obtained by placing a prior over the paradigm will also satisfy  $D(Q_\mu \| P) \leq \alpha$ . In turn, this upper bound makes the total variation distance  $\|Q_\mu - P\|$  bounded away from 1. Because  $P$  is a law in the paradigm, it follows from Theorem 1 that  $\Lambda$  cannot be testable.

In the previous examples, a property essential for testability was the existence of a sufficiently rich set of laws in the paradigm that are far from each other in total variation. This property does not hold for a set  $\Lambda$  that satisfies the condition of Proposition 3, resulting in a paradigm that is not testable.

#### 4.4 Maximal Paradigms

We have taken as a datum that the paradigm  $\Lambda$  is correctly specified. A paradigm that is incorrectly specified exposes the tester to the risk of rejecting, out of hand, forecasters who are informed but whose predictions lie outside  $\Lambda$ . Adopting a larger paradigm mitigates such a risk. Olszewski (2015) posed the question of which testable paradigms are *maximal*, in the sense of not being included in any other testable paradigm. The next result provides an answer to this open question.

**Proposition 4** *Let  $\epsilon \in (0, 1)$  and fix a law  $P \in \Delta(\Omega)$ . The paradigm*

$$\Lambda_P^\epsilon = \{ \tilde{P} \in \Delta(\Omega) : \|P - \tilde{P}\| > 1 - \epsilon \}$$

*is  $\epsilon$ -testable and is not included in any testable paradigm.*

The paradigm is constructed by simply fixing a distribution  $P$  and considering all laws which are sufficiently far from it. The resulting set  $\Lambda_P^\epsilon$  is not included in any testable paradigm.<sup>17</sup>

As shown in the proof of Proposition 4,  $P$  equals the law  $Q_\mu$  induced by some prior  $\mu$  that assigns probability 1 to the closure of  $\Lambda_P^\epsilon$ . Therefore, by Theorem 2 and the definition of  $\Lambda_P^\epsilon$ , the law  $P$  can be used to construct a non-manipulable likelihood-ratio test where it plays the role of a benchmark against which forecasters' predictions are compared.

---

<sup>17</sup>As shown in the proof,  $\Lambda_P^\epsilon$  is not included in any  $\delta$ -testable paradigm, for all  $\delta > 0$  sufficiently small. However, we do not know if the class of paradigms that are testable, rather than  $\epsilon$ -testable, and have the property of not being strictly included in any testable paradigm, admits a simple characterization. For example, given a non-degenerate law  $P$ , it can be shown that  $\Lambda = \{ \tilde{P} \in \Delta(\Omega) : \|P - \tilde{P}\| = 1 \}$  is testable.  $\Lambda$  is however strictly included in the testable paradigm  $\Lambda' = \{ \tilde{P} \in \Delta(\Omega) : \|P(\cdot|E) - \tilde{P}\| = 1 \}$ , where  $E$  is any event such that  $P(E) \in (0, 1)$ , since  $\Lambda'$  contains the measure  $P(\cdot|E^c)$  but  $\Lambda$  does not.

## 5 Discussion and Extensions

### 5.1 Non-asymptotic and Prequential Tests

We now consider the case of non-asymptotic tests where at most  $n$  observations are available to the tester. A paradigm is  $\epsilon$ -testable in  $n$  periods if it admits a test  $T$  such that  $T(\cdot, P)$  is  $\mathcal{F}_n$ -measurable for every  $P$ , accepts the truth with probability at least  $1 - \epsilon$ , and is  $\epsilon$ -nonmanipulable. The next result shows how Theorems 1 and 2 can be adapted to non-asymptotic tests. Given  $n$ , we define the semi-metric  $\rho_n(Q, P) = \max_{E \in \mathcal{F}_n} |Q(E) - P(E)|$ .

**Proposition 5** *Let  $\Lambda$  be a paradigm. If  $\Lambda$  is  $\epsilon$ -testable in  $n$  periods then there exists a prior  $\mu \in \Delta(\bar{\Lambda})$  such that  $\rho_n(Q_\mu, P) > 1 - 2\epsilon$  for every  $P \in \Lambda$ . Conversely, if there exists a prior  $\mu \in \Delta(\bar{\Lambda})$  with the property that  $\rho_n(Q_\mu, P) > 1 - \epsilon$  for every  $P \in \Lambda$ , then the test*

$$T(\omega, P) = \begin{cases} 1 & \text{if } P \in \Lambda \text{ and } P(\omega^n) > Q_\mu(\omega^n) \\ 0 & \text{otherwise} \end{cases}$$

*accepts the truth with probability at least  $1 - \epsilon$  and is  $\epsilon$ -nonmanipulable.*

Hence, similarly to Theorem 1, testability in  $n$  periods is equivalent to a high distance between the law  $Q_\mu$  induced by some prior  $\mu$  and any law in the the paradigm. Conversely, if such a prior exists, then restricting the attention to likelihood-ratio tests is without loss of generality.

An important distinction in the literature on empirical tests is between *prequential* and *non-prequential* tests. Consider, for simplicity, the case where in every period only two outcomes, 0 or 1, can occur. Recall that for every path  $\omega$ , we denote by  $Z_k(\omega) \in \{0, 1\}$  the corresponding outcome in period  $k$ . A test  $T$  is prequential if for every path  $\omega$  and every announced law  $P$ , the result  $T(\omega, P)$  is a function of the one-step ahead predictions generated by  $P$  along the realized sequence  $\omega$ . This is the sequence of probabilities

$$P(Z_1 = 1), P(Z_2 = 1|\omega^1), P(Z_3 = 1|\omega^2), \dots \quad (8)$$

obtained by conditioning, in every period  $t$ , on the realized history  $\omega^t$ . For instance, (8) might correspond to the sequence of rain probabilities reported each day by a weather forecaster.

Many tests used in practice, such as calibration, are prequential. A notable feature of the test described in Proposition 5 is that it shares this property. Because the prior  $\mu$  and the number of observations  $n$  depend on the paradigm but not on the prediction  $P$ , the result of the test is a function only of the probability  $P(\omega^n)$  assigned by the forecaster to the realized history. By the law of total probability,  $P(\omega^n)$  can be computed from the first  $n$  terms in the sequence (8) (provided the history  $\omega^n$  has positive probability under  $P$ ). Hence, the test can be implemented by asking the forecaster to simply report one-step ahead probabilities, rather than to announce a fully specified measure at time 0.

The existing literature has also studied prequential tests. Consider the unrestricted paradigm  $\Lambda = \Delta(\Omega)$ . In this case, as shown by Shmaya (2008), every prequential test is manipulable, even if the test is not required to accept or to reject in finite time. However, there exist tests that are not prequential, and not finite, but cannot be manipulated (see Olszewski (2015)). There are, in addition, specific examples of paradigms that can be tested by means of a nonmanipulable prequential test, as shown by Sandroni and Shmaya (2014).<sup>18</sup>

In contrast to these results, Proposition 5 shows that for non-asymptotic tests the requirement of a prequential test is without loss of generality. A paradigm that is  $\epsilon$ -testable in  $n$  periods remains so even when restricting the attention to prequential tests.

## 5.2 Maxmin and Strategic Forecasters

As discussed in Section 4, a strategic but uninformed forecaster evaluates a strategy  $\zeta$  as

$$\inf_{P \in C} \mathbb{E}_{P \otimes \zeta} [wT + l(1 - T)]$$

where  $C \subseteq \Delta(\Omega)$  is a set of laws. So far, we have considered the case where  $C$  is equal to the paradigm. However, an uninformed forecaster may adopt a less conservative decision making criterion.

To this end, let  $d$  be a distance that metrizes the weak\*-topology on  $\Delta(\Omega)$ , and for every law  $P \in \Delta(\Omega)$  denote by  $B_\delta(P)$  the open ball of radius  $\delta$  around  $P$ . We

---

<sup>18</sup>Sandroni and Shmaya (2014) study the paradigm of exchangeable distributions. This is convex and compact, hence not testable according to the definition of this paper. Their result shows that surprisingly, such a paradigm becomes testable when allowing for non-finite and prequential tests.

consider the specification

$$C = B_\delta(P_o) \cap \Lambda \text{ for some } P_o \in \Lambda. \quad (9)$$

So, under (9), an uninformed forecaster evaluates a strategy by considering the worst-case expected payoff with respect to laws that are within distance  $\delta$  from a reference measure  $P_o$ . Similar definitions appear in robust statistics (Huber, 1981) and economics (Bergemann and Schlag, 2011, and Babaioff, Blumrosen, Lambert and Reingold, 2010). We will not assume that  $P_o$  coincides with the correct law generating the data nor that  $P_o$  is known to the tester.

The definition of testable paradigm can now be strengthened as follows:

**Definition 6** A paradigm  $\Lambda$  is *uniformly testable with precision  $\delta$*  if for every  $\epsilon > 0$  there exists a finite test  $T$  such that:

1.  $T$  accepts the truth with probability at least  $1 - \epsilon$ ;
2. For every strategy  $\zeta$  and every  $P_o \in \Lambda$  there exists a law  $P_\zeta \in \Lambda \cap B_\delta(P_o)$  such that  $\mathbb{E}_{P_\zeta \otimes \zeta}[T] \leq \epsilon$ .

Thus, the test passes a true expert with high probability. In addition, for every strategy  $\zeta$ , there is a law  $P_\zeta$  in the paradigm under which rejection is likely. Given a reference law  $P_o$ , the measure  $P_\zeta$  can be chosen to belong to  $\Lambda \cap B_\delta(P_o)$ . Hence, the test guarantees that the value (9) an uninformed forecaster can expect from participating in the test is negative whenever  $\epsilon$  is sufficiently small. So, the test can screen between the two types of forecasters.

While a complete characterization of paradigms that fulfill the requirements of Definition 6 is beyond the scope of this paper, the next proposition provides a basic sufficient condition for a paradigm to be uniformly testable.

**Proposition 6** *Let  $\Lambda$  be a paradigm. If there exists a prior  $\mu \in \Delta(\bar{\Lambda})$  with support  $\bar{\Lambda}$  and such that  $Q_\mu$  satisfies  $\|Q_\mu - P\| = 1$  for every  $P$  in  $\Lambda$ , then  $\Lambda$  is uniformly testable with precision  $\delta$  for every  $\delta > 0$ .*

The most significant difference with respect to Theorem 1 is the assumption that the prior  $\mu$  has full support over the paradigm. Following the interpretation presented

in Section 2, a Bayesian outside observer endowed with such a prior  $\mu$  is “cautious,” in the sense of assigning positive probability to any open set of possible laws.

For a simple example of prior  $\mu$  that satisfies this property, consider the case where the set of outcomes is binary, i.e.  $X = \{0, 1\}$ , and  $\Lambda$  is the paradigm of all i.i.d. distributions. In this case, the set  $\Lambda$  is closed, and a prior  $\mu$  can be described by a probability measure  $m$  over  $[0, 1]$ . If  $m$  is uniform over the interval  $[0, 1]$ , then  $\mu$  satisfies  $\|Q_\mu - P\| = 1$  for every  $P \in \Lambda$ , by the Strong Law of Large Numbers. Moreover,  $\mu$  has support equal to  $\Lambda$ , and hence satisfies the conditions of Proposition 6. These claims are formally proved in Section A.6 in the Appendix.



# A Appendix

## A.1 Preliminaries

The space of paths  $\Omega$  is endowed with the product topology. Hence, a function that is  $\mathcal{F}_n$ -measurable for some  $n$  is also continuous. This implies that for every finite test  $T$  and any law  $P \in \Delta(\Omega)$  the function  $Q \mapsto \mathbb{E}_Q [T(\cdot, P)]$ ,  $Q \in \Delta(\Omega)$ , is continuous. We will denote by  $\mathcal{H}_n$  the set of histories  $\omega^n$  of length  $n$ .

Recall that the space  $\Delta(\Delta(\Omega))$  is endowed with the weak\* topology. As proved in Phelps (2001) (Proposition 1.1), the function  $\mu \mapsto Q_\mu$  assigning to each prior  $\mu \in \Delta(\Delta(\Omega))$  its barycenter  $Q_\mu$  is continuous. In particular, given a continuous function  $\psi : \Omega \rightarrow \mathbb{R}$ , the map  $\mu \mapsto \int_\Omega \psi(\omega) dQ_\mu(\omega)$ ,  $\mu \in \Delta(\Delta(\Omega))$ , is continuous. In addition,  $Q_\mu$  satisfies  $\int_\Omega \psi(\omega) dQ_\mu(\omega) = \int_{\Delta(\Omega)} (\int_\Omega \psi(\omega) dQ(\omega)) d\mu(Q)$  for every bounded measurable function  $\psi$ . Given a measurable subset  $\Gamma$  of  $\Delta(\Omega)$ , denote by  $\Delta(\Gamma)$  the set of probability measures  $\mu \in \Delta(\Delta(\Omega))$  assigning probability 1 to  $\Gamma$ . The space  $\Delta(\bar{\Gamma})$  is compact (see Aliprantis and Border (2006, Chapter 16)).

**Lemma 1** *Let  $T$  be a finite test. For every strategy  $\zeta$  the function  $P \mapsto \mathbb{E}_{P \otimes \zeta} [T]$ ,  $P \in \Delta(\Omega)$ , is continuous.*

*Proof.* Let  $(\omega_k)$  be a sequence in  $\Omega$  converging to a path  $\omega$ . Given a law  $P$ , the function  $T(\cdot, P)$  is continuous. So,  $T(\omega_k, P) \rightarrow T(\omega, P)$  as  $k \rightarrow \infty$ . Given a strategy  $\zeta$ , Lebesgue's convergence theorem implies  $\mathbb{E}_\zeta [T(\omega_k, \cdot)] \rightarrow \mathbb{E}_\zeta [T(\omega, \cdot)]$  as  $k \rightarrow \infty$ . Hence, for every strategy  $\zeta$  the map  $\omega \mapsto \mathbb{E}_\zeta [T(\omega, \cdot)]$ ,  $\omega \in \Omega$ , is continuous. Fubini's Theorem implies  $\mathbb{E}_{P \otimes \zeta} [T] = \int_\Omega \mathbb{E}_\zeta [T(\omega, \cdot)] dP(\omega)$ . Therefore, for each  $P$ ,  $\int_\Omega \mathbb{E}_\zeta [T(\omega, \cdot)] dP(\omega)$  is the expectation with respect to  $P$  of a continuous function. It follows from the definition of weak\* topology that the map  $P \mapsto \mathbb{E}_{P \otimes \zeta} [T]$ ,  $P \in \Delta(\Omega)$ , is continuous. ■

## A.2 Proofs of Theorems 1 and 2

*Proof of Theorems 1 and 2.* The first half of the proof shows the necessity part of Theorem 1. The second half establishes Theorem 2 and the sufficiency part of Theorem 1.

Assume  $\Lambda$  is testable. Fix  $\epsilon > 0$  and let  $T$  be a test that satisfies the conditions of Definition 4. Given a measure  $P \in \Delta(\Omega)$  and a strategy  $\zeta$ , let  $V(P, \zeta) = \mathbb{E}_{P \otimes \zeta} [T]$ .

The map  $V$  is affine in each argument and for each strategy  $\zeta$  the map  $V(\cdot, \zeta)$  is continuous by Lemma 1. Since  $T$  is  $\epsilon$ -nonmanipulable then

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} V(P, \zeta) \leq \epsilon. \quad (10)$$

Let  $\Delta_o(\Lambda) \subseteq \Delta(\Lambda)$  be the subset of priors on  $\Lambda$  with finite support. We have

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} V(P, \zeta) = \sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{\mu \in \Delta_o(\Lambda)} V(Q_\mu, \zeta) = \sup_{\zeta \in \Delta(\Delta(\Omega))} \min_{\mu \in \Delta(\bar{\Lambda})} V(Q_\mu, \zeta). \quad (11)$$

The first equality follows immediately from the definition of  $Q_\mu$  and the affinity of  $V(\cdot, \zeta)$ . The second equality follows from the continuity of the map  $\mu \mapsto V(Q_\mu, \zeta)$ ,  $\mu \in \Delta(\Delta(\Omega))$ , together with the fact that  $\Delta_o(\Lambda)$  is dense in  $\Delta(\bar{\Lambda})$  (as implied by Aliprantis and Border (2006, Theorem 15.10)) and that  $\Delta(\bar{\Lambda})$  is compact.

The space  $\Delta(\bar{\Lambda})$  is compact and convex and for every  $\zeta$  the map  $\mu \mapsto V(Q_\mu, \zeta)$ ,  $\mu \in \Delta(\Delta(\Omega))$ , is continuous (by Lemma 1) and affine. In addition,  $\Delta(\Delta(\Omega))$  is convex and for every  $\mu$  the map  $V(Q_\mu, \cdot)$  is affine. We can therefore apply Fan's Minmax Theorem (Fan, 1953) to obtain the equality

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \min_{\mu \in \Delta(\bar{\Lambda})} V(Q_\mu, \zeta) = \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{\zeta \in \Delta(\Delta(\Omega))} V(Q_\mu, \zeta). \quad (12)$$

For every  $\mu$ , the function  $V$  satisfies  $V(Q_\mu, \zeta) = \int_{\Delta(\Omega)} \mathbb{E}_{Q_\mu} [T(\cdot, P)] d\zeta(P)$  by Fubini's theorem. So,  $\sup_{\zeta \in \Delta(\Delta(\Omega))} V(Q_\mu, \zeta) = \sup_{P \in \Delta(\Omega)} V(Q_\mu, \delta_P)$ . Hence the right-hand side of (12) can be written as

$$\min_{\mu \in \Delta(\bar{\Lambda})} \sup_{\zeta \in \Delta(\Delta(\Omega))} V(Q_\mu, \zeta) = \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Delta(\Omega)} V(Q_\mu, \delta_P) = \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Delta(\Omega)} \mathbb{E}_{Q_\mu} [T(\cdot, P)]. \quad (13)$$

Taken together, (10), (11) (12) and (13) prove the existence of a prior  $\mu \in \Delta(\bar{\Lambda})$  such that

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} V(P, \zeta) = \sup_{P \in \Delta(\Omega)} \mathbb{E}_{Q_\mu} [T(\cdot, P)] \leq \epsilon.$$

Because the test accepts the truth with probability at least  $1 - \epsilon$ , it follows that

$$\mathbb{E}_P [T(\cdot, P)] - \mathbb{E}_{Q_\mu} [T(\cdot, P)] \geq 1 - 2\epsilon \text{ for all } P \in \Lambda. \quad (14)$$

It follows from standard arguments that the (normalized) total variation distance

$\|Q_\mu - P\|$  satisfies

$$\|Q_\mu - P\| = \sup_{\phi} \left| \int_{\Omega} \phi dQ_\mu - \int_{\Omega} \phi dP \right|$$

where the supremum is taken over all measurable functions  $\phi : \Omega \rightarrow [0, 1]$ .<sup>19</sup>

By letting  $\phi = T(\cdot, P)$ , it follows from (14) that  $\|Q_\mu - P\| \geq 1 - 2\epsilon$  for every  $P \in \Lambda$ . Since  $\epsilon$  is arbitrary, the first part of the proof is concluded.

Now consider a prior  $\mu \in \Delta(\bar{\Lambda})$  such that  $\|Q_\mu - P\| > 1 - \epsilon$  for all  $P \in \Lambda$ . Fix a measure  $P \in \Lambda$ . For any  $n$ ,

$$\max_{E \in \mathcal{F}_n} Q_\mu(E) - P(E) = \max_{E \in \mathcal{F}_n} |Q_\mu(E) - P(E)|.$$

since  $P(E) - Q_\mu(E) = Q_\mu(E^c) - P(E^c)$ . As shown in Halmos (1950, 13D),

$$\max_{E \in \mathcal{F}_n} |Q_\mu(E) - P(E)| \uparrow \|Q_\mu - P\|$$

as  $n \rightarrow \infty$ .<sup>20</sup> Therefore, we can conclude that for each  $P \in \Lambda$  the number

$$n_P = \min \left\{ n : \max_{E \in \mathcal{F}_n} Q_\mu(E) - P(E) > 1 - \epsilon \right\} \quad (15)$$

---

<sup>19</sup>To see this, let  $R = \frac{1}{2}Q_\mu + \frac{1}{2}P$ , and define  $E$  as the event where  $\frac{dQ_\mu}{dR} \geq \frac{dP}{dR}$ , where  $\frac{dQ_\mu}{dR}$  and  $\frac{dP}{dR}$  denote the corresponding Radon-Nikodym derivatives. Given a measurable  $\phi : \Omega \rightarrow [0, 1]$  we have

$$\int_{\Omega} \phi dQ_\mu - \int_{\Omega} \phi dP = \int_{\Omega} \phi \cdot \left( \frac{dQ_\mu}{dR} - \frac{dP}{dR} \right) dR \leq \int_E \left( \frac{dQ_\mu}{dR} - \frac{dP}{dR} \right) dR \leq \|Q_\mu - P\|.$$

and

$$\int_{\Omega} \phi dP - \int_{\Omega} \phi dQ_\mu = \int_{\Omega} \phi \cdot \left( \frac{dP}{dR} - \frac{dQ_\mu}{dR} \right) dR \leq \int_{E^c} \left( \frac{dP}{dR} - \frac{dQ_\mu}{dR} \right) dR \leq \|Q_\mu - P\|.$$

It follows that  $\sup_{\phi} \left| \int_{\Omega} \phi dQ_\mu - \int_{\Omega} \phi dP \right| \leq \|Q_\mu - P\|$ . The converse inequality is obvious.

<sup>20</sup>We provide here a sketch of the proof. Let  $\mathcal{F} = \cup_n \mathcal{F}_n$ . It can be verified that for every  $Q \in \Delta(\Omega)$ , the collection of events  $E$  for which there exists a sequence  $(F_m)$  in  $\mathcal{F}$  such that  $\lim_n Q(E \Delta F_n) = 0$  is a  $\sigma$ -algebra containing  $\mathcal{F}$ . Hence it equals  $\mathcal{B}$ . Fix  $B \in \mathcal{B}$  and let  $(E_m)$  be a sequence in  $\mathcal{F}$  such that  $\lim_m (P + Q_\mu)(B \Delta E_m) = 0$ . Hence  $\lim_m |P(B) - P(E_m)| \leq \lim_m P(B \Delta E_m) = 0$ . Similarly,  $\lim_m |Q_\mu(B) - Q_\mu(E_m)| = 0$ .

Because, for every  $m$ ,

$$|P(B) - Q_\mu(B)| \leq |P(B) - P(E_m)| + |P(E_m) - Q_\mu(E_m)| + |Q_\mu(B) - Q_\mu(E_m)|$$

then letting  $m \rightarrow \infty$ , it follows that  $|P(B) - Q_\mu(B)| \leq \sup_{F \in \mathcal{F}} |P(F) - Q_\mu(F)|$ . Because  $B$  is arbitrary, then  $\|P - Q_\mu\| \leq \sup_{\mathcal{F}} |P(F) - Q_\mu(F)| \leq \|P - Q_\mu\|$ . Since  $\sup_{F \in \mathcal{F}_n} |P(F) - Q_\mu(F)| \uparrow \sup_{\mathcal{F}} |P(F) - Q_\mu(F)|$  as  $n \rightarrow \infty$ , the result is established.

is well defined. Consider now the test

$$T(\omega, P) = \begin{cases} 1 & \text{if } P \in \Lambda \text{ and } P(\omega^{n_P}) > Q_\mu(\omega^{n_P}) \\ 0 & \text{otherwise} \end{cases}$$

We now prove that  $T$  is measurable. First we show that for every  $k \in \mathbb{N}$  the set  $\{P \in \Lambda : n_P = k\}$  is measurable. For every  $n$  and every  $E \in \mathcal{F}_n$  the function  $P \mapsto P(E)$ ,  $P \in \Delta(\Omega)$ , is continuous. Because  $\mathcal{F}_n$  is finite, it follows that  $\varphi_n : P \mapsto \max_{E \in \mathcal{F}_n} Q_\mu(E) - P(E)$ ,  $P \in \Delta(\Omega)$ , is measurable. Since  $\Lambda$  is measurable the restriction of  $\varphi_n$  on  $\Lambda$  is also measurable. The set  $\{P \in \Lambda : n_P = k\}$  can be written as  $\{P \in \Lambda : \varphi_k(P) > 1 - \epsilon\}$  if  $k = 1$ , or as the intersection

$$\bigcap_{1 \leq n < k} \{P \in \Lambda : \varphi_n(P) \leq 1 - \epsilon\} \cap \{P \in \Lambda : \varphi_k(P) > 1 - \epsilon\}$$

if  $k > 1$ . Hence  $\{P \in \Lambda : n_P = k\}$  is measurable. For each path  $\omega$ , the function  $T(\omega, \cdot)$  is measurable: For each  $n$ , the set  $\{P \in \Delta(\Omega) : T(\omega, P) = 1\}$  is given by the union over  $k > 1$  of all sets of the form

$$\{P \in \Delta(\Omega) : P(\omega^k) - Q_\mu(\omega^k) > 0\} \cap \{P \in \Lambda : n_P = k\}.$$

It follows that  $T(\omega, \cdot)$  is measurable. For each  $\omega \in \Omega$  and  $P \in \Delta(\Omega)$ , the function  $T(\cdot, P)$  is continuous and  $T(\omega, \cdot)$  is measurable. That is,  $T$  is a Carathéodory function. It follows then from Lemma 4.51 in Aliprantis and Border (2016) that  $T$  is measurable.

We now show that  $P(\{\omega : T(\omega, P) = 1\}) > 1 - \epsilon$  and  $Q_\mu(\{\omega : T(\omega, P) = 1\}) < \epsilon$  for each  $P \in \Lambda$ . The proof follows Lehmann and Romano (2006, Chapter 16). Let  $P \in \Lambda$ , and denote by  $A^P$  the set  $\{\omega : P(\omega^{n_P}) > Q_\mu(\omega^{n_P})\}$ . Recall  $\mathcal{H}_{n_P}$  is the set of all histories of length  $n_P$ . For every  $E \in \mathcal{F}_{n_P}$  we have

$$\begin{aligned} P(E) - Q_\mu(E) &= \sum_{\omega^{n_P} \in \mathcal{H}_{n_P} : \omega^{n_P} \subseteq E} P(\omega^{n_P}) - Q_\mu(\omega^{n_P}) \\ &\leq \sum_{\omega^{n_P} \in \mathcal{H}_{n_P} : \omega^{n_P} \subseteq E \cap A^P} P(\omega^{n_P}) - Q_\mu(\omega^{n_P}) \\ &\leq \sum_{\omega^{n_P} \in \mathcal{H}_{n_P} : \omega^{n_P} \subseteq A^P} P(\omega^{n_P}) - Q_\mu(\omega^{n_P}). \end{aligned}$$

Therefore  $P(A^P) - Q_\mu(A^P) = \max_{E \in \mathcal{F}_{n_P}} P(E) - Q_\mu(E) > 1 - \epsilon$ . So  $P(A^P) > 1 - \epsilon$  (in

particular, the test  $T$  accepts the truth with probability at least  $1 - \epsilon$  and  $Q_\mu(A^P) < \epsilon$ . We can now show that  $T$  is  $\epsilon$ -nonmanipulable. For every strategy  $\zeta$ , we have

$$V(Q_\mu, \zeta) = \int_{\Delta(\Omega)} Q_\mu(A^P) d\zeta(P) < \epsilon. \quad (16)$$

Using again the fact that  $\mu \mapsto V(Q_\mu, \zeta)$ ,  $\mu \in \Delta(\Delta(\Omega))$ , is continuous and  $\Delta_o(\Lambda)$  is dense in  $\Delta(\bar{\Lambda})$ , we can find a prior  $\mu_\zeta \in \Delta_o(\Lambda)$  such that

$$V(Q_{\mu_\zeta}, \zeta) = \sum_{P \in \Lambda} \mu_\zeta(P) V(P, \zeta) < \epsilon$$

Hence, there must exist some law  $P_\zeta \in \Lambda$  in the support of  $\mu_\zeta$  such that  $V(P_\zeta, \zeta) < \epsilon$ . Because  $\epsilon$  is arbitrary, we conclude that  $\Lambda$  is testable. ■

*Proof of Corollary 1.* As shown in Phelps (2001, Proposition 1.2) a law  $P$  belongs to the weak\*-closed convex hull of  $\Lambda$  if and only if there exists a prior  $\mu \in \Delta(\bar{\Lambda})$  such that  $P = Q_\mu$ . The result now follows immediately from Theorem 1 and the definition of  $I$ . ■

### A.3 Proof of Theorem 3

The next result is a version of the Neyman-Pearson lemma. The standard proof parallels the proof of Theorem 3.2.1 in Lehmann and Romano (2006) and is therefore omitted.

**Theorem 4 (Neyman-Pearson Lemma)** *Let  $P_0, P_1 \in \Delta(\Omega)$ . Given  $n \in \mathbb{N}$  and  $\alpha \in [0, 1]$ , let  $\Phi$  be the set of  $\mathcal{F}_n$ -measurable functions  $\phi : \Omega \rightarrow [0, 1]$  that satisfy  $E_{P_0}[\phi] \geq \alpha$ . Let*

$$\lambda = \sup \{k \in \mathbb{R} : P_0(\{\omega : P_0(\omega^n) \geq kP_1(\omega^n)\}) \geq \alpha\}$$

and, letting  $0 \cdot \infty = 0$ , define

$$\begin{aligned} \delta &= P_0(\{\omega : P_0(\omega^n) > \lambda P_1(\omega^n)\}) \\ \gamma &= P_0(\{\omega : P_0(\omega^n) = \lambda P_1(\omega^n)\}) \end{aligned}$$

The function

$$\phi^*(\omega) = \begin{cases} 1 & \text{if } P_0(\omega^n) > \lambda P_1(\omega^n) \\ \frac{\alpha - \delta}{\gamma} & \text{if } P_0(\omega^n) = \lambda P_1(\omega^n) \text{ and } \gamma > 0 \\ 0 & \text{otherwise} \end{cases}$$

is a solution to  $\min_{\phi \in \Phi} \mathbb{E}_{P_1}[\phi]$ .

*Proof of Theorem 3.* Fix a paradigm  $\Lambda$ , testing times  $(n_P)$  and a probability  $\alpha \in [0, 1]$ . Denote by  $\mathcal{T}$  the class of finite tests that are bounded by  $(n_P)$  and accept the truth with probability at least  $\alpha$ .

For every  $P \in \Lambda$ , let  $\Phi_P$  be the set of  $\mathcal{F}_{n_P}$ -measurable functions  $\phi : \Omega \rightarrow [0, 1]$  that satisfy  $E_P[\phi] \geq \alpha$ . Define the function  $f : \Delta(\bar{\Lambda}) \rightarrow \mathbb{R}$  as

$$f(\mu) = \sup_{P \in \Lambda} \min_{\phi \in \Phi_P} \mathbb{E}_{Q_\mu}[\phi].$$

The function  $f$  is lower-semicontinuous: Fix  $P \in \Lambda$ . The set  $\Phi_P$  can be identified with a subset of  $[0, 1]^m$ , where  $m$  is the cardinality of the set of histories of length  $n_P$ . It is then immediate to verify that  $\Phi_P$  is compact. It follows from the theorem of the maximum that the map  $Q \mapsto \min_{\phi \in \Phi_P} \mathbb{E}_Q[\phi]$ ,  $Q \in \Delta(\Omega)$ , is continuous. Thus, the continuity of the map  $\mu \mapsto Q_\mu$ ,  $\mu \in \Delta(\Delta(\Omega))$ , implies that the map  $\mu \mapsto \min_{\phi \in \Phi_P} \mathbb{E}_{Q_\mu}[\phi]$ ,  $\mu \in \Delta(\Delta(\Omega))$  is a composition of continuous functions. Thus,  $f$  is a supremum of continuous functions. Hence  $f$  is lower-semicontinuous and so attains a minimum on  $\Delta(\bar{\Lambda})$ . Let  $\mu^*$  be a prior which minimizes  $f$ .

Denote by  $\phi_P^*$  the test obtained by applying the Neyman-Pearson lemma when setting  $P_0 = P$ ,  $P_1 = Q_{\mu^*}$  and  $n = n_P$  in the statement of Theorem 4. Denote also by  $\lambda_P$ ,  $\delta_P$  and  $\gamma_P$  the corresponding quantities. Let  $T^*$  be the test defined as

$$T^*(\omega, P) = \begin{cases} \phi_P^*(\omega) & \text{if } P \in \Lambda \\ 0 & \text{if } P \notin \Lambda. \end{cases}$$

We now show that  $T^*$  is a well-defined test belonging to  $\mathcal{T}$ . By definition, the test is finite and accepts the truth with probability at least  $\alpha$ . It remains to show it is measurable. By Lemma 4.51 in Aliprantis and Border (2016), it is enough to prove that  $T(\omega, \cdot)$  is measurable for every  $\omega$ . We first show that the map  $P \mapsto \lambda_P$ ,

$P \in \Lambda$ , mapping each measure to the corresponding threshold  $\lambda_P \in [0, \infty]$  in the likelihood-ratio test, is measurable. For every  $k \in \mathbb{R}$  let

$$\Gamma_k = \{P \in \Lambda : P(\{\omega : P(\omega^{n_P}) \geq kQ_{\mu^*}(\omega^{n_P})\}) \geq \alpha\}.$$

Notice that  $\Gamma_k$  can be written as

$$\bigcup_{m \in \mathbb{N}} (\{P \in \Lambda : n_P = m\} \cap \{P \in \Lambda : P(\{\omega : P(\omega^m) \geq kQ_{\mu^*}(\omega^m)\}) \geq \alpha\})$$

Each set  $\{P \in \Lambda : n_P = m\}$  is measurable. For each  $\omega^m$  the function  $P \mapsto P(\omega^m)$ ,  $P \in \Delta(\Omega)$ , is continuous. So, for each history  $\omega^m$  the set

$$\Upsilon_{\omega^m} = \{P \in \Lambda : P(\omega^m) \geq kQ_{\mu^*}(\omega^m)\}$$

is measurable. Let  $1_{\Upsilon_{\omega^m}}$  be the indicator function of  $\Upsilon_{\omega^m}$  and notice that

$$P(\{\omega : P(\omega^m) \geq kQ_{\mu^*}(\omega^m)\}) = \sum_{\omega^m \in \mathcal{H}_m} P(\omega^m) 1_{\Upsilon_{\omega^m}}(P),$$

where the latter is a measurable function of  $P$ . It then follows that each set of the form

$$\{P \in \Lambda : P(\{\omega : P(\omega^m) \geq kQ_{\mu^*}(\omega^m)\}) \geq \alpha\}$$

is measurable. Thus,  $\Gamma_k$  is measurable. This in turn yields that for each  $k$  the function  $P \mapsto k1_{\Gamma_k}(P)$  is measurable. Notice that  $\lambda_P = \sup_{k \in \mathbb{Q}} k1_{\Gamma_k}(P)$  for every  $P$ . Thus, we can conclude that the function  $P \mapsto \lambda_P$  (mapping  $\Delta(\Omega)$  to  $\mathbb{R} \cup \{\infty\}$ ) is measurable. Now fix a path  $\omega$ . An argument analogous to that one used to prove the measurability of the set  $\Gamma_k$  shows that  $\{P \in \Lambda : P(\omega^{n_P}) > \lambda_P Q_{\mu^*}(\omega^{n_P})\}$  and  $\{P \in \Lambda : P(\omega^{n_P}) = \lambda_P Q_{\mu^*}(\omega^{n_P})\}$  are measurable and that  $\delta_P$  and  $\gamma_P$  are measurable functions of  $P$ . It is then routine to verify that  $T(\omega, \cdot)$  is measurable. We can therefore conclude that  $T$  is a well defined test belonging to  $\mathcal{T}$ .

We now show that  $T^*$  is a least manipulable test in the class  $\mathcal{T}$ . Let  $T \in \mathcal{T}$ . As in the proof of Theorems 1 and 2, given any test  $T \in \mathcal{T}$  we can apply Fan's minmax theorem to conclude

$$\sup_{\zeta \in \Delta(\Delta(\Omega))} \inf_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta} [T] = \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Delta(\Omega)} \mathbb{E}_{Q_\mu} [T(\cdot, P)].$$

It is without loss of generality to assume that  $T(\omega, P) = 0$  for every  $\omega$  and  $P \notin \Lambda$ . So, the expression can be simplified to

$$\sup_{\zeta \in \Delta(\Lambda)} \inf_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta} [T] = \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Lambda} \mathbb{E}_{Q_\mu} [T(\cdot, P)]. \quad (17)$$

The test  $T$  is finite and accepts the truth with probability at least  $\alpha$ . So, it satisfies  $T(\cdot, P) \in \Phi_P$  for every  $P \in \Lambda$ . Thus,

$$\begin{aligned} \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Lambda} \mathbb{E}_{Q_\mu} [T(\cdot, P)] &\geq \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Lambda} \min_{\phi \in \Phi_P} \mathbb{E}_{Q_\mu} [\phi] \\ &= \min_{\mu \in \Delta(\bar{\Lambda})} f(\mu) \\ &= \sup_{P \in \Lambda} \min_{\phi \in \Phi_P} \mathbb{E}_{Q_{\mu^*}} [\phi]. \end{aligned}$$

The essential idea is that the test  $T^*$  has been defined to satisfy, for every  $P \in \Lambda$ ,

$$\mathbb{E}_{Q_{\mu^*}} [T^*(\cdot, P)] = \min_{\phi \in \Phi_P} \mathbb{E}_{Q_{\mu^*}} [\phi]$$

This means that

$$\begin{aligned} \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Lambda} \mathbb{E}_{Q_\mu} [T(\cdot, P)] &\geq \sup_{P \in \Lambda} \min_{\phi \in \Phi_P} \mathbb{E}_{Q_{\mu^*}} [\phi] \\ &= \sup_{P \in \Lambda} \mathbb{E}_{Q_{\mu^*}} [T^*(\cdot, P)] \\ &\geq \min_{\mu \in \Delta(\bar{\Lambda})} \sup_{P \in \Lambda} \mathbb{E}_{Q_\mu} [T^*(\cdot, P)]. \end{aligned}$$

By applying (17) to both  $T$  and  $T^*$  we now obtain

$$\sup_{\zeta \in \Delta(\Lambda)} \inf_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta} [T] \geq \sup_{\zeta \in \Delta(\Lambda)} \inf_{P \in \Lambda} \mathbb{E}_{P \otimes \zeta} [T^*].$$

Hence,  $T^*$  is less manipulable than  $T$ . ■

#### A.4 Proof of Propositions 1-4

*Proof of Proposition 1.* Fix two outcomes  $x, y \in X$ . Let  $N_n(\omega)$  be the number of periods where outcome  $x$  occurs along the path  $\omega$  up to time  $n$ , and let  $N_\infty(\omega) = \sup_n N_n(\omega)$ . In addition, define  $N_n[x \rightarrow y](\omega)$  to be the number of periods, up to time



$n$ , where the outcome  $x$  is followed in the next period by  $y$ .

For every transition  $\pi$ , let  $E_\pi$  be the set of paths  $\omega$  such that  $N_\infty(\omega) = \infty$  and

$$\lim_{n \rightarrow \infty} \frac{N_n[x \rightarrow y](\omega)}{N_n(\omega)} = \pi(x)(y) \quad (18)$$

It is a standard result that every Markov  $P_{\rho,\pi}$  satisfies  $P_{\rho,\pi}(\{N_\infty < \infty\} \cup E_\pi) = 1$ . We include here a proof for completeness. Let  $A = \{N_\infty < \infty\} \cup E_\pi$  and notice that  $P_{\rho,\pi}(A) = \sum_{z \in X} P_{\rho,\pi}(\{\omega_1 = z\})P_{z,\pi}(A)$ , where  $P_{z,\pi}$  denotes the Markov law with transition  $\pi$  and initial probability putting mass 1 on  $z$ . Write  $X = S \cup R_1 \cup \dots \cup R_n$ , where  $S$  is the set of transient states and  $(R_i)$  are disjoint maximal irreducible sets of states (see Theorem 6.2.13 in Dembo, 2015). If  $x$  is transient then  $P_{\rho,\pi}(\{N_\infty < \infty\}) = 1$ , hence  $P_{\rho,\pi}(A) = 1$  as desired. Assume  $x$  is not transient, and  $x \in R_1$  without loss of generality. If  $z \in R_i$  and  $i > 1$  then  $P_{z,\pi}(\{N_\infty = 0\}) = 1$ . So  $P_{z,\pi}(A) = 1$ . If  $z \in R_1$  then it is well known that  $P_{z,\pi}$ , being irreducible, satisfies  $P_{z,\pi}(E_\pi) = 1$ <sup>21</sup>. Thus  $P_{z,\pi}(A) = 1$ . Hence  $P_{z,\pi}(A) = 1$  for every recurrent state. Now let  $z \in S$  and consider the stopping time  $\tau(\omega) = \inf\{n : \omega_n \notin S\}$ . Because  $z$  is transient and  $X$  is finite then  $P_{z,\pi}(\{\tau < \infty\}) = 1$ . Because  $P_{w,\pi}(A) = 1$  for every  $w \notin S$  it follows from the strong Markov property (Proposition 6.1.16 in Dembo, 2015) that  $P_{z,\pi}(A) = 1$ .

The measure  $m$  assigns probability 1 to the set  $\Pi_+ \subseteq \Pi$  of transition probabilities  $\pi$  that satisfy  $\pi(y)(z) \in (0, 1)$  for all  $y, z \in X$ . Let  $\pi \in \Pi_+$ . Then  $P_\pi$  is irreducible and satisfies  $P_\pi(N_\infty = \infty) = 1$ . Hence  $P_\pi(E_\pi) = 1$ . Therefore, given a Markov law  $P_{\rho,\sigma}$  with transition  $\sigma \in \Pi$ ,

$$\begin{aligned} & P_{\rho,\sigma}(\{N_\infty < \infty\} \cup E_\sigma) - Q_\mu(\{N_\infty < \infty\} \cup E_\sigma) \\ &= 1 - \int_{\Pi_+} P_\pi(E_\sigma) dm(\pi) = 1 \end{aligned}$$

where the last equality follows from the fact that  $\pi(x)(y) \neq \sigma(x)(y)$  implies  $E_\sigma \cap E_\pi = \emptyset$  (hence  $P_\pi(E_\sigma) = 0$ ) and  $\{\pi \in \Pi : \pi(x)(y) = \sigma(x)(y)\}$  has probability 0 under  $m$ . Therefore  $\|Q_\mu - P_{\rho,\pi}\| = 1$ . ■

*Proof of Proposition 2.* Let  $P_1, \dots, P_n$  in  $\Lambda$  be strongly orthogonal. Denote by  $\mathcal{F}^\infty = \bigcap_{k=1}^\infty \mathcal{F}_k^\infty$  the tail  $\sigma$ -algebra. We first show that for any  $i \neq j$  we can find an event  $A_{ij}$  that belongs to  $\mathcal{F}^\infty$  and satisfies  $P_i(A_{ij}) = 1$  and  $P_j(A_{ij}) = 0$ . As shown by

<sup>21</sup>See, for example, <http://www.statslab.cam.ac.uk/~james/Markov/s110.pdf>

Goldstein (1979, Proposition 4.1) it holds that

$$\lim_{k \rightarrow \infty} \sup_{E \in \mathcal{F}_k^\infty} |P_i(E) - P_j(E)| = \sup_{E \in \mathcal{F}^\infty} |P_i(E) - P_j(E)|.$$

Hence  $\sup_{E \in \mathcal{F}^\infty} |P_i(E) - P_j(E)| = 1$ . Now let  $R_i$  and  $R_j$  be the restriction on  $\mathcal{F}^\infty$  of  $P_i$  and  $P_j$ , respectively. Let  $R = \frac{1}{2}R_i + \frac{1}{2}R_j$  and define  $A_{ij} \in \mathcal{F}^\infty$  as the event where the two Radon-Nikodym derivatives satisfy  $\frac{dR_i}{dR} \geq \frac{dR_j}{dR}$ . For every  $E \in \mathcal{F}^\infty$ , the event  $A_{ij}$  satisfies

$$P_i(E) - P_j(E) = \int_E \left( \frac{dR_i}{dR} - \frac{dR_j}{dR} \right) dR \leq R_i(A_{ij}) - R_j(A_{ij}) = P_i(A_{ij}) - P_j(A_{ij}).$$

Hence  $P_i(A_{ij}) = 1$  and  $P_j(A_{ij}) = 0$ . Let  $E_i = \bigcap_{j \neq i} A_{i,j}$ . Then each  $E_i$  is tail-measurable and satisfies  $P_i(E_i) = 1$  and  $P_j(E_i) = 0$  for every  $j \neq i$ . If  $i \neq j$  then  $A_{i,j}$  and  $A_{j,i}$  are disjoint, hence  $E_i$  and  $E_j$  are disjoint as well. By enlarging  $E_n$  to be equal to the complement of  $\bigcup_{i=1}^{n-1} E_i$ , we can assume that  $E_1, \dots, E_n$  form a partition of  $\Omega$ . Each of its element are tail-measurable.

Let  $\mu$  be a uniform prior over  $P_1, \dots, P_n$ . Then  $Q_\mu(E_i) = 1/n$  for every  $i = 1, \dots, n$ . Fix  $P \in \Lambda$ . Because  $P$  is mixing, it satisfies  $P(F) \in \{0, 1\}$  for every event  $F$  that is tail-measurable (see Theorem 13.18 in Davidson, 1994). Because  $E_1, \dots, E_n$  is a partition of  $\Omega$  that consists of tail-measurable events, then  $P \in \Lambda$  satisfies  $P(E_{i_P}) = 1$  for some  $i_P \in \{1, \dots, n\}$ . Hence

$$\|Q_\mu - P\| \geq P(E_{i_P}) - Q_\mu(E_{i_P}) = P(E_{i_P}) - \mu(P_{i_P}) = 1 - 1/n.$$

Since  $P$  and  $n$  are arbitrary, it follows from Theorem 1 that  $\Lambda$  is testable. ■

*Proof of Proposition 3.* The map  $\tilde{P} \mapsto D(\tilde{P} \| P)$ ,  $\tilde{P} \in \Delta(\Omega)$ , is convex and lower semicontinuous (see, for instance, Lemma 1.4.3 in Dupuis and Ellis (2011)). Convexity implies that for every  $\mu \in \Delta(\Lambda)$  with finite support, i.e. every  $\mu$  such that  $Q_\mu \in \text{co}(\Lambda)$ ,

$$D(Q_\mu \| P) \leq \sum_{\tilde{P} \in \Lambda} \mu(\tilde{P}) D(\tilde{P} \| P) \leq \alpha.$$

Lower semincontinuity implies the set  $\{\tilde{P} \in \Delta(\Omega) : D(\tilde{P} \| P) \leq \alpha\}$  is closed. Hence  $D(Q_\mu \| P) \leq \alpha$  for every  $Q_\mu \in \overline{\text{co}}(\Lambda)$ , i.e. for every  $\mu \in \Delta(\overline{\Lambda})$  as shown in the proof of Corollary 1. The normalized total-variation distance between  $Q_\mu$  and  $P$  and the

Kullback-Leibler divergence  $D(Q_\mu \| P)$  are related by the inequality

$$\|Q_\mu - P\|^2 \leq 1 - e^{-D(Q_\mu \| P)}$$

(see equation (4) in Sason and Verdú (2016)). Thus, every  $\mu \in \Delta(\bar{\Lambda})$  satisfies  $\|Q_\mu - P\| \leq \sqrt{1 - e^{-\alpha}} < 1$ . Since  $P \in \Lambda$ , it follows from Theorem 1 that  $\Lambda$  is not testable. ■

*Proof of Proposition 4.* As shown by Theorem 2, in order to prove that  $\Lambda_P^\epsilon$  is  $\epsilon$ -testable it is enough to find a prior  $\mu \in \Delta(\bar{\Lambda}_P^\epsilon)$  such that  $P = Q_\mu$ . Consider the set  $N = \{\omega : P(\{\omega\}) = 0\}$ . Each  $\omega \in N$  satisfies  $\delta_\omega \in \Lambda_P^\epsilon$ . Notice that  $P$  can have at most countably many atoms, so  $N$  is dense. The function  $\omega \mapsto \delta_\omega$ ,  $\omega \in \Omega$ , is continuous, and so  $\{\delta_\omega : \omega \in N\}$  is dense in  $\{\delta_\omega : \omega \in \Omega\}$ . We can therefore conclude that  $\{\delta_\omega : \omega \in \Omega\} \subseteq \bar{\Lambda}_P^\epsilon$ . Consider now the prior defined as  $\mu(\Gamma) = P(\{\omega : \delta_\omega \in \Gamma\})$  for every measurable set  $\Gamma \subseteq \Delta(\Omega)$ . Standard arguments shows that  $\mu$  is well defined and satisfies  $Q_\mu = P$ . Because  $\mu(\{\delta_\omega : \omega \in \Omega\}) = 1$ , then  $\mu \in \Delta(\bar{\Lambda}_P^\epsilon)$ . Therefore,  $\Lambda_P^\epsilon$  is  $\epsilon$ -testable.

Suppose, as a way of contradiction, that  $\Lambda_P^\epsilon \subseteq \Lambda$ , where  $\Lambda$  is a paradigm that is  $\epsilon'$ -testable and  $\epsilon' < \frac{\epsilon}{2}$ . As shown in the proof of Theorem 1, there exists a prior  $\nu \in \Delta(\bar{\Lambda})$  such that  $\|Q_\nu - Q\| \geq 1 - 2\epsilon'$  for every  $Q \in \Lambda$ . Equivalently,

$$\{Q \in \Delta(\Omega) : \|Q - Q_\nu\| < 1 - 2\epsilon'\} \subseteq \Lambda^c$$

By assumption,  $\Lambda^c \subseteq (\Lambda_P^\epsilon)^c = \{Q \in \Delta(\Omega) : \|Q - P\| \leq 1 - \epsilon\}$ , so

$$\{Q \in \Delta(\Omega) : \|Q - Q_\nu\| < 1 - 2\epsilon'\} \subseteq \{Q \in \Delta(\Omega) : \|Q - P\| \leq 1 - \epsilon\}. \quad (19)$$

To show that this leads to a contradiction, let  $R \in \Delta(\Omega)$  be a measure such that  $\|R - Q_\nu\| = \|R - P\| = 1$ . For instance, let  $R = \delta_\omega$  for some path  $\omega$  that is not an atom of either  $Q_\nu$  or  $P$ . Fix  $t \in (2\epsilon', \epsilon)$  and consider the measure  $tQ_\nu + (1 - t)R$ . We have

$$\|tQ_\nu + (1 - t)R - Q_\nu\| = (1 - t)\|R - Q_\nu\| = (1 - t) < 1 - 2\epsilon'.$$

Hence, it follows from (19) that  $\|tQ_\nu + (1 - t)R - P\| \leq 1 - \epsilon$ . Now let  $E$  be an event

such that  $R(E) = 1$  and  $Q_\nu(E) = P(E) = 0$ . Then

$$1 - \epsilon \geq \|tQ_\nu + (1 - t)R - P\| \geq tQ_\nu(E) + (1 - t)R(E) - P(E) = 1 - t.$$

By construction,  $1 - t > 1 - \epsilon$ . So we obtain a contradiction. Therefore,  $\Lambda_P^\epsilon$  is not included in any testable paradigm. ■

## A.5 Other Proofs

*Proof of Proposition 5.* Let  $\Lambda$  be  $\epsilon$ -testable in  $n$  periods. Then, by substituting the total-variation distance with the semi-distance  $\rho_n$  and following the same arguments used in the proof of Theorem 1, it follows that there exists a prior  $\mu \in \Delta(\bar{\Lambda})$  such that  $\rho_n(Q_\mu, P) > 1 - 2\epsilon$  for all  $P \in \Lambda$ .

Only one change is necessary: the same argument applied in the proof of Theorem 1 shows that  $\rho_n(P, Q) = \max_\phi |\int_\Omega \phi dP - \int_\Omega \phi dQ|$  where the maximum is taken over all functions  $\phi : \Omega \rightarrow [0, 1]$  that are  $\mathcal{F}_n$ -measurable.

Conversely, let  $\mu \in \Delta(\bar{\Lambda})$  be a prior such that  $\|Q_\mu - P\|_n > 1 - \epsilon$  for all  $P \in \Lambda$ . This part of the proof follows, verbatim, the proof of Theorem 2 (notice that by assumption  $n_P \leq n$  for every  $P \in \Lambda$ ). ■

The next result will be used in the proof of Proposition 6. In what follows,  $B_\delta(P)$  denote the open ball of radius  $\delta$  around  $P$  with respect to the same metric  $d$  fixed in the main text. Recall that the *support* of a measure  $\mu \in \Delta(\Delta(\Omega))$  is the unique closed set  $\Gamma \subseteq \Delta(\Omega)$  with the property that  $\mu(\Gamma^c) = 0$  and for every open set  $V \subseteq \Delta(\Omega)$ , if  $V \cap \Gamma \neq \emptyset$  then  $\mu(V \cap \Gamma) > 0$  (see Aliprantis and Border, 2016, 12.3).

**Lemma 2** *Let  $\mu \in \Delta(\Delta(\Omega))$  be a prior and let  $\Gamma \subseteq \Delta(\Omega)$  be its support. For every  $\delta > 0$  there exists a constant  $\lambda > 0$  such that*

$$\mu(B_\delta(P)) \geq \lambda \text{ for all } P \in \Gamma.$$

*Proof of Lemma 2.* Suppose not. Then there must exist  $\delta > 0$  and a sequence  $(P_n)$  in  $\Gamma$  such that  $\mu(B_\delta(P_n)) \rightarrow 0$  as  $n \rightarrow \infty$ . The space  $\Delta(\Omega)$  is compact and  $\Gamma \subseteq \Delta(\Omega)$  is closed. Hence, it is compact. So, we can assume (taking a subsequence if necessary) that  $P_n$  converges to a law  $P \in \Gamma$ . Fix a law  $Q$ . Assume  $Q \in B_{\delta/2}(P)$ .

Then  $d(P_n, Q) < \delta$  for all  $n$  large enough. Thus  $Q \in B_\delta(P_n)$  for all  $n$  large enough. Thus,

$$1_{B_{\delta/2}(P)}(Q) \leq \liminf_{n \rightarrow \infty} 1_{B_\delta(P_n)}(Q) \text{ for every } Q \in \Gamma$$

where  $1_{B_{\delta/2}(P)}$  denotes the indicator function of  $B_{\delta/2}(P)$ . By applying Fatou's lemma, we can then conclude that

$$\mu(B_{\delta/2}(P)) \leq \int_{\Delta(\Omega)} \liminf_{n \rightarrow \infty} 1_{B_\delta(P_n)} d\mu \leq \liminf_n \mu(B_\delta(P_n)) = 0$$

Hence  $\mu(B_{\delta/2}(P)) = 0$ . Since  $P \in \Gamma$ , then  $\mu$  must assign positive probability to every neighborhood of  $P$ , so we reach a contradiction, and the proof is finished. ■

*Proof of Proposition 6.* By Lemma 2, there exists a  $\lambda > 0$  such that  $\mu(B_\delta(P)) \geq \lambda$  for every  $P \in \Lambda$ . Fix a sequence  $(\epsilon_n)$  such that  $\epsilon_n \downarrow 0$ . Because  $\|Q_\mu - P\| = 1$  for every  $P \in \Lambda$ , then, as shown in the proof of Theorem 2, we can find for every  $n$  a finite test  $T_n$  with the properties that  $T_n$  accepts the truth with probability at least  $1 - \epsilon_n$  and for every strategy  $\zeta$ , by equation (16),

$$\mathbb{E}_{Q_\mu \otimes \zeta}[T_n] = \int_{\Lambda} \mathbb{E}_{P \otimes \zeta}[T_n] d\mu(P) \leq \epsilon_n.$$

By applying Markov's inequality, for every  $k > 0$  and  $\zeta$ , we have

$$\mu\left(\left\{P \in \bar{\Lambda} : \mathbb{E}_{P \otimes \zeta}[T_n] \leq k\epsilon_n\right\}\right) \geq 1 - \frac{\mathbb{E}_{Q_\mu \otimes \zeta}[T_n]}{k\epsilon_n} \geq 1 - \frac{1}{k}$$

Fix  $\epsilon > 0$  and choose  $k$  large enough such that  $1 - \frac{1}{k} + \lambda > 1$ . In addition, given  $k$  choose  $N$  large enough such that  $k\epsilon_n \leq \epsilon$  for all  $n > N$ . Now fix a particular  $n > N$ . Given  $P_o \in \Lambda$  and a strategy  $\zeta$ , we have

$$\begin{aligned} & \mu\left(\left\{P \in \bar{\Lambda} \cap B_\delta(P_o) : \mathbb{E}_{P \otimes \zeta}[T_n] \leq \epsilon\right\}\right) \\ & \geq \mu\left(\left\{P \in \bar{\Lambda} : \mathbb{E}_{P \otimes \zeta}[T_n] \leq k\epsilon_n\right\} \cap B_\delta(P_o)\right) \\ & = \mu\left(\left\{P \in \bar{\Lambda} : \mathbb{E}_{P \otimes \zeta}[T_n] \leq k\epsilon_n\right\}\right) \\ & \quad + \mu(B_\delta(P_o)) - \mu\left(\left\{P \in \bar{\Lambda} : \mathbb{E}_{P \otimes \zeta}[T_n] \leq k\epsilon_n\right\} \cup B_\delta(P_o)\right) \\ & \geq 1 - \frac{1}{k} + \lambda - 1 > 0. \end{aligned}$$

This implies we can select a measure  $P_\zeta \in \bar{\Lambda} \cap B_\delta(P_o)$  such that  $\mathbb{E}_{P_\zeta \otimes \zeta}[T_n] \leq \epsilon$ . By

continuity of the map  $P \mapsto E_{P \otimes \zeta} [T_n]$  we can then select a measure  $P'_\zeta \in \Lambda \cap B_\delta(P_0)$  such that  $\mathbb{E}_{P'_\zeta \otimes \zeta} [T_n] \leq 2\epsilon$ . Because  $P_0$  is arbitrary, then it follows that the test  $T_n$  satisfies the conditions of Definition 6 for  $2\epsilon$ . Because  $\epsilon$  is arbitrary, it follows that  $\Lambda$  is uniformly testable with precision  $\delta$ . ■

## A.6 The Paradigm of i.i.d. Distributions is Uniformly Testable

Let  $X = \{0, 1\}$  and for every  $\theta \in [0, 1]$  let  $P^\theta$  be the i.i.d. distribution where the probability of outcome 1 is equal to  $\theta$  in every period. Consider the paradigm  $\Lambda = \{P^\theta : \theta \in [0, 1]\}$ . It is immediate to verify that the function  $\theta \mapsto P^\theta$  is continuous.

Let  $m$  be the uniform distribution on  $[0, 1]$ , and define the prior  $\mu \in \Delta(\Lambda)$  as  $\mu(\Gamma) = m(\{\theta : P^\theta \in \Gamma\})$  for every measurable  $\Gamma \subseteq \Delta(\Omega)$ . Then  $\mu$  has support equal to  $\Lambda$ . This follows from the fact that for every open set  $V \subseteq \Delta(\Omega)$ , the set  $\{\theta : P^\theta \in V\}$  is open. Therefore, it has positive probability under  $m$ .

It remains to show that  $\|Q_\mu - P^\theta\| = 1$  for each  $\theta \in [0, 1]$ . Let  $E_\theta \subseteq \Omega$  be the event where the limiting frequency of outcome 1 equals  $\theta$ . By the Strong Law of Large Numbers we have  $P^\theta(E_\theta) = 1$  and  $P^{\theta'}(E_\theta) = 0$  for each  $\theta' \neq \theta$ . Hence  $Q_\mu(E_\theta) = \int_0^1 P^{\theta'}(E_\theta) d\theta' = 0$ .

## A.7 Relation with Stewart (2011)

Stewart (2011) studies strategic forecasting in an environment where the tester is a Bayesian endowed with a prior  $\mu$  over  $\Delta(\Omega)$ . Stewart (2011) considers a (non-finite) likelihood-ratio test which compares the forecaster's predictions to the tester's predictions induced by  $Q_\mu$ . The paper studies priors  $\mu$  for which the quantity

$$\varepsilon = \int P \left\{ \omega : \sum_{t=1}^T (Q_\mu(\omega^t | \omega^{t-1}) - P(\omega^t | \omega^{t-1}))^2 \text{ converges} \right\} d\mu(P).$$

is sufficiently small. Intuitively, this implies that over time a true expert is able to provide more precise predictions than the tester. For every strategy  $\zeta$ , a strategic but ignorant forecaster will fail the test in Stewart (2011) with probability 1 under  $Q_\mu \otimes \zeta$ , while a true expert will almost surely pass the test, for a class of measures  $P$  that has probability  $1 - \varepsilon$  under  $\mu$ .

To see more clearly the connection between the two papers, consider the case where  $\varepsilon$  is zero, and define the paradigm  $\Lambda$  consisting of all distributions  $P$  for which

$\sum_{t=1}^T (Q_\mu(\omega^t|\omega^{t-1}) - P(\omega^t|\omega^{t-1}))^2$  diverges  $P$ -almost surely. The assumption that  $\varepsilon = 0$  implies that  $\mu(\Lambda) = 1$ . The same argument used by Stewart (2011) in the proof of his main result shows that every such  $P \in \Lambda$  has the property that the likelihood-ratio  $P(\omega^t) \setminus Q_\mu(\omega^t)$  diverges,  $P$ -almost surely. In turn, by an application of the Lebesgue-Decomposition Theorem,<sup>22</sup> this implies that each  $P$  in  $\Lambda$  is orthogonal to the predictive distribution  $Q_\mu$ . That is, the two have total variation distance 1. So, while there are many modeling differences between the two papers, given a prior  $\mu$  that satisfies the condition  $\varepsilon = 0$  in Stewart (2011), there is a paradigm that satisfies the conditions of Theorem 1 with respect to  $\mu$ .

At a conceptual level, Stewart (2011) and the present paper provide complementary perspectives on the use of the log-likelihood ratio as a way to screen forecasters. In this paper, we take as a primitive a paradigm, while the prior  $\mu$  and the corresponding Bayesian forecaster are endogenously derived from the test. Stewart (2011) takes as a primitive the prior, and the tester is willing to discard measures that have low probability under her subjective belief.

## References

- Al-Najjar, N., Pomatto, L., and A. Sandroni. (2014). “Claim Validation.” *American Economic Review*, 104(11), 3725-36.
- Aliprantis, C. D., and K. Border. (2006). *Infinite dimensional analysis: a hitchhiker’s guide*. Springer.
- Alkema, L., Raftery, A. E., and S.J. Clark. (2007). “Probabilistic projections of HIV prevalence using Bayesian melding.” *The Annals of Applied Statistics*, 229-248.
- Al-Najjar, N.I., Sandroni, A., Smorodinsky, R. and J. Weinstein (2010). “Testing theories with learnable and predictive representations.” *Journal of Economic Theory*, 145(6), 2203-2217.
- Al-Najjar, N., and J. Weinstein (2008). “Comparative testing of experts.” *Econometrica*, 76(3), 541-559.

---

<sup>22</sup>see Theorem 2, p. 525, in Shiryaev (1996).

- Babaioff, M., L. Blumrosen, N. Lambert and O. Reingold (2011). “Only valuable experts can be valued.” *Proceedings of the 12th ACM conference on electronic commerce*, 221-222.
- Bergemann, D., and K. Schlag (2011). “Robust monopoly pricing.” *Journal of Economic Theory*, 146, 2527–2543.
- Cerreia-Vioglio, S., Maccheroni, F. and M. Marinacci (2013). “Classical subjective expected utility.” *Proceedings of the National Academy of Sciences*, 110(17), 6754-6759.
- Corradi, V., and N. R. Swanson (2006). “Predictive density evaluation.” *Handbook of economic forecasting*, 1, 197-284.
- Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford.
- Dawid, A. P. (1982). “The well-calibrated Bayesian.” *Journal of the American Statistical Association*, 77(379), 605-610.
- Dembo, A. (2015). *Probability Theory*. Lecture notes.
- Diebold, F. X., Tay, A. S., and K. F. Wallis. (1997). “Evaluating density forecasts of inflation: the Survey of Professional Forecasters.” *National bureau of economic research*.
- Dupuis, P. and Ellis, R.S. (2011). *A weak convergence approach to the theory of large deviations*.
- Fan, K. (1953). “Minimax theorems.” *Proceedings of the National Academy of Sciences*, 39(1), 42-47.
- Feinberg, Y., and N. Lambert (2015). “Mostly calibrated.” *International Journal of Game Theory*, 44:153-163.
- Foster, D. P., and R. Vohra (1998). “Asymptotic calibration.” *Biometrika*, 85(2), 379-390.
- Foster, D. P., and R. Vohra (2011). “Calibration: Respice, adspice, prospice.” In *Advances in Economics and Econometrics: Tenth World Congress*, volume 1 (Daron Acemoglu, Manuel Arellano, and Eddie Dekel, eds.), 423-442, Cambridge University Press.



- Giacomini, R., and H. White. (2006). "Tests of conditional predictive ability." *Econometrica*, 74(6), 1545-1578.
- Goldstein, S. (1979). "Maximal coupling." *Probability Theory and Related Fields*, 46, 193-204.
- Gilboa, I. and D. Schmeidler (1989). "Maxmin expected utility with non-unique prior." *Journal of mathematical economics*, 18(2), 141-153.
- Gneiting, T., and A. E. Raftery (2005). "Weather forecasting with ensemble methods." *Science*, 310 (5746), 248-249.
- Hansen, L.P, and T. J. Sargent. (2001) "Robust control and model uncertainty." *American Economic Review*, 9(2), 60-66.
- Huber, P. J. (1981). *Robust statistics*. Wiley.
- Jackson, M., E. Kalai and R. Smorodinsky (1999). "Bayesian representation of stochastic processes under learning: de Finetti revisited." *Econometrica*, 67(4), 875-893.
- Jordan T.H., Chen Y.T., Gasparini P., Madariaga R., Main I., et al. (2011). "Operational earthquake forecasting: state of knowledge and guidelines for utilization." *Annals of Geophysics* 54, 315–91.
- Kavaler, I., and R. Smorodinsky(2017). "On Comparison Of Experts." arXiv preprint arXiv:1710.09461.
- Lehmann, E. L., and J.P. Romano (2006). *Testing statistical hypotheses*. Springer.
- Neyman, J. and E. S. Pearson. (1933). "The testing of statistical hypotheses in relation to probabilities a priori." *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 29, No.4.
- Nze, P. A., and P. Doukhan. (2004). "Weak dependence: models and applications to econometrics." *Econometric Theory*, 20(6), 995-1045.
- Olszewski, W. (2015). "Calibration and expert testing." *Handbook of Game Theory with Economic Applications*, 4, 949-984.
- Olszewski, W., and A. Sandroni. (2008). "Manipulability of future-independent tests." *Econometrica*, 76(6), 1437-1466.

- Olszewski, W., and A. Sandroni. (2009). “Strategic manipulation of empirical tests.” *Mathematics of Operations Research*, 34.1, 57-70.
- Olszewski, W., and A. Sandroni. (2009b). “A nonmanipulable test.” *The Annals of Statistics*. 1013-1039.
- Phelps, R. R. (2001). *Lectures on Choquet’s theorem*. Springer.
- Raftery A.E., Li N., Sevcikova H., Gerland P., Heilig G.K. (2012). “Bayesian probabilistic population projections for all countries.” *Proceedings of the National Academy of Sciences*, 109, 13915–21
- Sandroni, A. (2003). “The reproducible properties of correct forecasts.” *International Journal of Game Theory*, 32(1), 151-159.
- Sandroni, A., and E. Shmaya (2014). “A prequential test for exchangeable theories.” *Journal of Dynamics and Games*, 1(3), 497-505
- Sason, I. and Verdú, S. (2016). “ $f$ -divergence Inequalities.” *IEEE Transactions on Information Theory*, 62(11), 5973-6006.
- Shmaya, E. (2008). “Many inspections are manipulable.” *Theoretical Economics*, 3(3), 367-382.
- Shiryaev, A.N. (1996). *Probability*, Second Edition, Springer.
- Starr, R. M. (1969). “Quasi-equilibria in markets with non-convex preferences.” *Econometrica*, 37(1), 25-38.
- Stewart, C. (2011). “Nonmanipulable bayesian testing.” *Journal of Economic Theory*. 146(5), 2029-2041.
- Tetlock P.E. (2005). *Political Expert Judgement*. Princeton.
- Timmermann, A. (2000). “Density forecasting in economics and finance.” *Journal of Forecasting*, 19(4), 231-234.
- Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.