Int J Game Theory (2015) 44:153–163
DOI 10.1007/s00182-014-0423-0

Mostly calibrated

Yossi Feinberg · Nicolas S. Lambert

Accepted: 27 March 2014 / Published online: 16 April 2014

Abstract Prequential testing of a forecaster is known to be manipulable if the test must pass an informed forecaster for all possible true distributions. Stewart (J Econ Theory 146(5):2029–2041, 2011) provides a non-manipulable prequential likelihood test that only fails an informed forecaster on a small, category I, set of distributions. We present a prequential test based on calibration that also fails the informed forecaster on at most a category I set of true distributions and is non-manipulable. Our construction sheds light on the relationship between likelihood and calibration with respect to the distributions they reject.

Keywords Testing forecasters · Calibration test

1 Introduction

We consider the problem of testing a forecaster in a stochastic environment. The forecaster may or may not be informed of the probabilities governing an unfolding sequence of realizations. Our objective is to find a test that can determine whether the forecaster is informed based on her predictions and the realization of the process. The

We are grateful to Colin Stewart for helpful comments and suggestions. Lambert thanks Google Research and the National Science Foundation under grant No. CCF-1101209. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Y. Feinberg (✉) · N. S. Lambert
Graduate School of Business, Stanford University, Stanford, CA, USA
e-mail: yossi@stanford.edu

N. S. Lambert
e-mail: nlambert@stanford.edu

classic example concerns the testing of a weather forecaster. The forecaster makes a daily prediction of the probability of rain the following day. A natural test would be to check whether the predictions are calibrated. This calibration test considers the empirical distribution of rain conditional on the predictions made by the forecaster. For example, it considers the realization of rain following all the days on which she predicted rain with probability 40 % the previous day. The forecaster is said to be calibrated if for every prediction made long enough, the empirical distribution converges to the forecast. In other words, it rained close to 40 % of the time in the days following her prediction of 40 % chance of rain. Foster and Vohra (1998) have shown that an uninformed forecaster could use a randomized prediction strategy that guarantees she will be calibrated no matter what the realization is—hence, no matter what the true distribution is, the uninformed forecaster can manipulate the calibration test.

The manipulability of the calibration test has been extensively generalized, most notably by Sandroni (2003) and Olszewski and Sandroni (2008) for any finite-time rejection test as well as by Shmaya (2008) for all prequential tests—tests that sequentially ask for predictions along the unfolding realization. In contrast, positive results by Dekel and Feinberg (2006) and Olszewski and Sandroni (2009b) showed that there are tests that cannot be manipulated, however they require that the forecaster provide the entire distribution upfront. Other positive results include the testing of multiple forecasters where a prequential test can distinguish *which* forecaster is the informed one. Distinguishing an informed forecaster from an uninformed one was achieved by Al-Najjar and Weinstein (2008) using a likelihood test and by Feinberg and Stewart (2008) using a cross-calibration test. A likelihood test compares the ratio of the products of two forecasters' prediction sequences to determine the informed one under the assumption that one of the forecasters must be informed. A cross-calibration test asks that a forecaster be calibrated conditional on the joint predictions of the forecasters, for example, considering all the periods where one forecaster predicted 20 % chance of rain for the following day and the other predicted 40 % at the same time. Cross-calibration measures the empirical distribution conditional on both predictions. Since the distribution of rain cannot converge both to 40 and 20 % at the same time the test distinguishes the informed forecaster.

An alternative strand of the literature considered tests that restrict the set of possible distributions governing the process. Al-Najjar et al. (2010) considered learnable distributions. These distributions converge to a distinct characteristic that allows the forecaster to provide separating predictions in finite time. Olszewski and Sandroni (2009a) suggested a prequential non-manipulable test that fails the informed forecaster only when the conditional probabilities are within a small interval, e.g., forecasts that get close to 50 %. For general finite sets of outcomes, Babaioff et al (2010) proposed a non-manipulable test that passes the informed forecaster only when the conditional distributions are outside a closed convex set of non-zero measure. Stewart (2011) put forth a likelihood-based test that cannot be manipulated and may fail the informed forecaster on at most a small, category I, set of distributions, i.e., a countable union of nowhere dense sets—sets whose closure has empty interior. Our paper adds to this strand by offering a calibration-based non-manipulable test that may fail the informed forecaster on at most a small, category I, set of distributions. Our results advance the

analogy between likelihood-type and calibration-type tests as viable practical tools for testing experts.

To define our test consider choosing a probability distribution arbitrarily. This will be our *reference* distribution. The reference distribution will determine the set of probability distributions for which a forecaster who knows the true distribution would fail the test—a set we want to keep small, a category I set. The reference distribution will also determine how the forecaster is being calibrated. We assume the test and the reference distribution are known to the forecaster. The test proceeds by considering sequentially the realization of the process and the sequence of predictions the forecaster provided as this realization materialized, making it a prequential test. We now distinguish two cases. If the forecaster did not make predictions that infinitely often were close to the conditional probabilities of the reference distribution then we fail the forecaster. If the forecasts were infinitely often close to the reference distribution we ask whether the forecasts were calibrated when they came close to the reference distribution. For example, consider the reference distribution that predicts 30 % chance of rain (i.i.d.). The set of distributions that do not infinitely often have conditional predictions close to 30 % is a category I set. For predictions not in that set, i.e., where the expert infinitely often makes predictions close to 30 %, we test if the expert is calibrated with respect to the reference distributions. In this case, it simply means that on these periods, the empirical distribution is indeed 30 %—i.e., we check if the forecaster was calibrated on these periods, much like standard weather forecast calibration tests do. We show that this generalizes to *any* reference distribution. The test will now require the expert to be calibrated with respect to the reference distribution on the periods where he makes predictions close to the conditional probabilities of the reference distribution.

As an example, let the reference distribution predict a probability of 20 % chance of rain if there was no rain the day before, and 80 % chance of rain otherwise. The set of distributions that do not infinitely often have conditional probabilities close to this reference distribution is once again a category I set of distributions. Now we require that the forecaster's predictions are also close to 20 % after a dry day and 80 % the day after it rained infinitely often. Conditional on making these predictions infinitely often we check whether the empirical distribution was close to 20 and 80 % respectively. If the true distribution does infinitely often get close to the reference distribution, the true forecaster will pass the test by reporting the true distribution. However, we show that there is no strategy of the false expert that guarantees passing this test, hence it cannot be manipulated. While standard calibration can be manipulated, it is the additional requirement that predictions coincide infinitely often with a reference distribution that removes manipulation and allows for testing even when there is no restriction on when the true distribution may be close to the reference distribution.

Our results are presented as follows. We first observe that for every reference distribution if the distribution governing the process is indeed close to the reference distribution infinitely often then an informed forecaster will pass the test. We then show that for any given reference distribution the set of distributions that with positive probability never come close to the reference distribution is a category I set of distributions. Hence, for all but a category I set of distributions the informed forecaster will pass the test with probability one (with respect to the true distribution). Finally,

we show that there is no randomized forecasting strategy that would allow an uninformed forecaster to always pass the test. As in Stewart (2011), our results carry over to the Bayesian setting, in which the tester and the forecaster hold a common prior over the possible distributions of the governing process. Under some conditions on the prior, our test passes the informed forecaster with probability one, while it fails the uninformed forecaster with probability one.

2 Definitions

Denote by $\Omega = \{(\omega_t)_{t=1,2,\ldots} \mid \omega_t \in \{0,1\}\}$ the set of possible realizations endowed with the product topology. We consider the Borel σ-algebra on Ω. Let ω^t refer to the first t elements of the sequence ω. Let $\mathcal{H}_t = \{0,1\}^t$ be the set of possible histories of length t.

Let $\Delta(\Omega)$ be the set of Borel probability distributions on Ω endowed with the weak* topology. Given a distribution λ, denote by $\lambda_t(\omega)$ the probability that $\omega_t = 1$ conditional on the history ω^{t-1} when the history has positive λ-probability. Assume $\lambda_t(\omega)$ takes an arbitrary value otherwise. We will occasionally write $\lambda_t(\omega^{t-1})$ to emphasize the independence of λ_t on future realizations.

We are interested in prequential tests of a forecaster. Assume the true distribution generating the realization $\omega \in \Omega$ is the probability distribution μ. At every period t the forecaster is asked to make a prediction as to the probability of $\omega_t = 1$ conditional on observing ω^{t-1}. An informed forecaster knows μ and is assumed to forecast according to μ. An uninformed forecaster does not know μ. At every period t she can choose any forecast in $[0,1]$, as well as randomize her forecast, as a function of the realization ω^{t-1} and her own past realized forecasts.

A *prequential* test is a function $T : \Delta(\Omega) \times \Omega \mapsto \{\text{PASS}, \text{FAIL}\}$ that satisfies $T(\lambda, \omega) = T(\lambda', \omega)$ if $\lambda_t(\omega) = \lambda'_t(\omega)$ for every t and such that $T(\lambda, \omega) = \text{FAIL}$ whenever there is some t such that $\lambda(\omega^t) = 0$. In other words, T depends only on the sequence of conditional forecasts along the realization and fails the forecaster if she predicts with certainty that the actual realization should not have materialized.[1]

A test *passes a forecaster informed of a true distribution* μ if the set of realizations $\{\omega | T(\mu, \omega) = \text{PASS}\}$ has μ-probability 1. A test *passes an informed forecaster on a set of distributions* $\mathcal{D} \subset \Delta(\Omega)$ if she passes for all $\mu \in \mathcal{D}$. A subset of a topological space is called *category I* if it is a countable union of nowhere dense sets—sets such that the interior of their closure is empty. Category I sets are small in a topological sense.

A pure strategy for an uninformed forecaster is a sequence of predictions $\{\lambda_t(\omega^{t-1})\}_{t=1}^{\infty}$ for every ω. Hence it corresponds to a unique probability distribution in $\Delta(\Omega)$. Given a mixed strategy $\eta \in \Delta(\Delta(\Omega))$ an uninformed forecaster *passes the*

[1] Note that any prequential test as defined above is also a prequential test as defined in Shmaya (2008). Shmaya considers tests as functions of sequences of predictions along a realization. A collection of forecast sequences, one for every realization, defines a unique distribution over sequences (Kolmogorov Extension Theorem), however multiple collections may be mapped into the same distribution. These multiple collections differ only in the probabilities assigned conditional on zero probability finite histories. Hence, by failing on an occurrence of such events our test can be defined as a function of forecasts predictions as well.

test T at ω with at least η-probability $1-\epsilon$ if $\eta(\{\lambda | T(\lambda,\omega) = \text{PASS}\}) \geq 1-\epsilon$, i.e., with at least probability $1-\epsilon$ with respect to her randomized strategy the uninformed forecaster passes the test on the realization ω. A test is *manipulable* if for every $\epsilon > 0$ there exists a mixed strategy η such that the uninformed forecaster passes the test at every $\omega \in \Omega$ with at least η-probability $1-\epsilon$.

A prequential test is called a *Borel test* if it is measurable with respect to the Borel σ-algebra defined on $\Delta(\Omega) \times \Omega$ as above.

We consider the following variant of the calibration test that we call *reference calibration test*.

Fix an arbitrary *reference distribution* $\nu \in \Delta(\Omega)$. Let $\mathcal{I}_1, \ldots, \mathcal{I}_m$ denote the intervals $\mathcal{I}_i = [\beta_i, \beta_{i+1})$ where $0 = \beta_1 < \cdots < \beta_m < 1$ and $\mathcal{I}_m = [\beta_m, 1]$, $m \geq 2$. These intervals need not be small in any sense, and may have different widths. Let $\{\epsilon_n\}_{n=1}^{\infty}$ be a positive sequence converging to zero. The test performs calibration in periods determined by a random variable B_t. The random variable B_t takes the value 1 when the forecaster's prediction is close to the reference distribution ν. Specifically, we define recursively a sequence of periods $\{t_n\}_{n=0}^{\infty}$ as a function of the forecaster's prediction λ, the realization ω, and the reference distribution ν as follows: Let $t_0 = 0$ and

$$t_{n+1} = \inf\left(\{t > t_n \mid |\lambda_t(\omega) - \nu_t(\omega)| < \epsilon_{n+1}\} \cup \{\infty\}\right).$$

t_n denotes the first period strictly after t_{n-1} at which the forecast gets within ϵ_n from the reference distribution and ∞ otherwise. Hence $\lim_n t_n = \infty$.

Define $B_t = 1$ if there is some n such that $t = t_n$, and $B_t = 0$ otherwise. Note that B_t is a function of ω^{t-1} and $\lambda_k(\omega^{k-1})$ for $k \leq t$. Let C_t^i be the random variable such that $C_t^i = 1$ if $\nu_t(\omega) \in \mathcal{I}_i$ and $C_t^i = 0$ otherwise. Note that the periods t_n and the variables B_t depend on ω and λ, and the variables C_t^i also depend on ω.

The reference calibration test passes the forecaster if $\sum_t B_t = \infty$ and

$$\lim_{n\to\infty} \frac{\sum_{t=1}^{n} B_t \cdot C_t^i \cdot (\lambda_t(\omega) - \omega_t)}{\sum_{t=1}^{n} B_t \cdot C_t^i} = 0$$

for all i such that $\sum_t B_t \cdot C_t^i = \infty$. Note that, if $\sum_t B_t = \infty$ for a particular λ and ω, there always exists some i such that $\sum_t B_t \cdot C_t^i = \infty$.

The reference calibration test simply adds conditioning on the random variables B_t to a standard calibration test. We perform calibration using the checking rule C_t^i whenever the predictions come close to the reference distribution, and ask that they do so infinitely often. Note that the reference calibration test is prequential as well as Borel since it is a limit of Borel functions.

3 Results

The calibration test was shown to be manipulable by Foster and Vohra (1998). This result was extended, expanded and generalized by Fudenberg and Levine (1999), Lehrer (2001), Vovk and Shafer (2005), Olszewski and Sandroni (2008, 2009b) and

Shmaya (2008), with the latter showing that every Borel prequential test that passes the informed forecaster for every true distribution μ can be manipulated, i.e., there exists a mixed strategy that passes the uninformed forecaster on every realization. We show that our reference calibration test fails the informed forecaster on a small set of true distributions and is non-manipulable.

Fix a reference distribution $\nu \in \Delta(\Omega)$, and define

$$\mathcal{D}_\nu = \left\{ \mu \in \Delta(\Omega) \mid \forall \omega \ \liminf_{t\to\infty} |\mu_t(\omega) - \nu_t(\omega)| = 0 \right\} .$$

The set $\mathcal{D}_\nu$ includes all the distributions with conditionals that get arbitrarily close to ν infinitely often, for every realization. Here, the conditionals of ν are fixed arbitrarily for 0-probability histories.

First we show that for every distribution ν, the set $\mathcal{D}_\nu$ is large in a topological sense.

Proposition 1 *For every reference distribution ν, the set of the distributions that do not belong to $\mathcal{D}_\nu$ is a category I set.*

Proof Observe that

$$\begin{aligned}\Delta(\Omega)\backslash\mathcal{D}_\nu &= \cup_{\omega\in\Omega} \cup_{k\geq 1} \cup_{n\geq 1} \cap_{t\geq n} \{\mu \in \Delta(\Omega) \mid |\mu_t(\omega) - \nu_t(\omega)| \geq 1/k\} \\ &\subseteq \cup_{k\geq 1} \cup_{n\geq 1} \cap_{t\geq n} \cup_{h_{t-1}\in\mathcal{H}_{t-1}} \\ &\qquad \{\mu \in \Delta(\Omega) \mid |\mu_t(h_{t-1}) - \nu_t(h_{t-1})| \geq 1/k\} .\end{aligned}$$

Also note that the set $\{\mu \in \Delta(\Omega) \mid |\mu_t(h_{t-1}) - \nu_t(h_{t-1})| \geq 1/k\}$ is weak* closed. Besides the union $\cup_{h_{t-1}} \{\mu \in \Delta(\Omega) \mid |\mu_t(h_{t-1}) - \nu_t(h_{t-1})| \geq 1/k\}$ is a finite union of closed sets, and so is closed. The set $\mathcal{S}_{n,k}$ defined by

$$\mathcal{S}_{n,k} = \cap_{t\geq n} \cup_{h_{t-1}} \{\mu \in \Delta(\Omega) \mid |\mu_t(h_{t-1}) - \nu_t(h_{t-1})| \geq 1/k\}$$

is an intersection of closed sets, and so is a closed set. Take any $\mu \in \mathcal{S}_{n,k}$. Define $\mu^m \in \Delta(\Omega)$ by

$$\mu^m_t(h_{t-1}) = \begin{cases} \mu_t(h_{t-1}) & \text{for all } t \leq m \text{ and all } h_{t-1} \in \mathcal{H}_{t-1} , \\ \nu_t(h_{t-1}) & \text{for all } t > m \text{ and all } h_{t-1} \in \mathcal{H}_{t-1} . \end{cases}$$

We have $\mu^m \notin \mathcal{S}_{n,k}$ and the sequence $\{\mu^m\}_{m=1}^\infty$ converges to μ in the weak* topology. Hence $\mathcal{S}_{n,k}$ has empty interior. So $\mathcal{S}_{n,k}$ is a closed nowhere dense set. Hence, $\cup_{k\geq 1}\cup_{n\geq 1}$ $\mathcal{S}_{n,k}$ is the countable union of nowhere dense sets—a category I set. □

Second we observe that for every reference distribution ν, an informed forecaster passes the reference calibration test on every distribution of $\mathcal{D}_\nu$.

Proposition 2 *Fix the reference distribution ν and the corresponding reference calibration test. The test passes the informed forecaster on the set $\mathcal{D}_\nu$.*

Proof Noting that for every $\mu \in \mathcal{D}_\nu$ we have $\sum_t B_t = \infty$, the proposition follows from the standard calibration theorems (cf. Dawid 1982; Kalai et al. 1999). □

Finally we show that the reference calibration test is not manipulable.

Proposition 3 *Fix the reference distribution ν and the corresponding reference calibration test. For every mixed strategy $\eta \in \Delta(\Delta(\Omega))$ of an uninformed forecaster, there exist uncountably many distributions $\mu \in \mathcal{D}_\nu$ such that the test fails the forecaster with (η, μ)-probability 1. That is, for a probability one set of realizations of the randomized forecast strategy and a probability one set of realizations of the process governed by μ, the forecaster fails the test.*

Proof As in previous work (cf.Olszewski and Sandroni 2009a; Babaioff et al 2010; Stewart 2011) to prove non-manipulability we will allow Nature to randomize over the "feasible" set of distributions $\mathcal{D}_\nu$ and observe that, for an appropriate choice of randomization, the compound distribution falls outside the feasible set. By doing so the uninformed forecaster fails the test on the compound distribution and also, by the section theorems, on a probability one set of distributions in $\mathcal{D}_\nu$.

Let $|\mathcal{I}_k|$ denote the length of interval $\mathcal{I}_k$, and let $\delta = \min\{|\mathcal{I}_1|, \ldots, |\mathcal{I}_m|\}/2$. We introduce the random variables Z^{h_t} taking values in [0, 1], for every $t \geq 0$ and every history $h_t \in \mathcal{H}_t$. There are countably many such random variables, and only finitely many for each fixed time t.

To define the Z^{h_t}'s, we also introduce the random variables $\tilde{Z}_t$, $t \geq 1$, taking values in [0, 1]. These variables are independently and identically distributed according to $P(\tilde{Z}_t \leq x) = (1 - \delta) + \delta x$. We let $Z^{h_{t-1}} = \tilde{Z}_t$ if $\nu_t(h_{t-1}) \in \mathcal{I}_m$, and $Z^{h_{t-1}} = 1 - \tilde{Z}_t$ otherwise. Note that the Z^{h_t}'s are perfectly correlated for every given t, but are independent across t's. Besides every Z^{h_t} is distributed following a mixture of a uniform distribution with weight δ and a $1 - \delta$ atom either at 0 or at 1.

By the Kolmogorov Extension Theorem, every realization of the family of variables Z^{h_t} yields a unique distribution over Ω, defined by $\mu_t(\omega) = Z^{\omega^{t-1}}$. Hence P yields a distribution ξ over $\Delta(\Omega)$. Denote by $\bar{\mu}_t(\omega)$ the expected conditional probability at time t, defined by $\bar{\mu}_t(\omega) = \int \mu_t(\omega) \mathrm{d}\xi(\mu)$. Note that $\bar{\mu}_t(\omega) = \delta/2$ and $\bar{\mu}_t(\omega) = 1 - \delta/2$ if $\nu_t(\omega) \in \mathcal{I}_m$ and $\nu_t(\omega) \notin \mathcal{I}_m$, respectively.

The distribution $\xi \in \Delta(\Delta(\Omega))$ has the following properties:

1. Whenever $\nu_t(\omega) \in \mathcal{I}_i$ for some $i \leq m-1$ we have $\bar{\mu}_t(\omega) - \nu_t(\omega) \geq \delta/2$. Similarly if $\nu_t(\omega) \in \mathcal{I}_m$ we have $\bar{\mu}_t(\omega) - \nu_t(\omega) \leq -\delta/2$.
2. The distribution ξ has no atoms. In particular every set of distributions with ξ-probability 1 is uncountable.
3. The distribution ξ assigns probability 1 to the set of distributions $\mathcal{D}_\nu$.

The first property is direct. The second property owes to the weight δ on the uniform distribution given independently to the conditional probabilities at every period. Finally, to obtain the third property, fix some $K, t_0 \in \mathbb{N}$. Let $\mathcal{S}_{K,t_0}$ be the set of distributions μ such that, for some $\omega \in \Omega$, $|\mu_t(\omega) - \nu_t(\omega)| > 1/K$ for every $t \geq t_0$. Note that for any given t, the probability that there exists some ω for which $|\mu_t(\omega) - \nu_t(\omega)| > 1/K$ cannot be greater than $1 - \delta/K$, because of the weight δ on the uniform distribution of the random conditional probabilities, and because all the random conditional probabilities are perfectly correlated at any given time. Since those conditional probabilities are distributed independently across time and $1 - \delta/K$ is bounded away from 1, the set $\mathcal{S}_{K,t_0}$ has ξ-probability 0. If a distribution μ does not belong to $\mathcal{D}_\nu$, it belongs to

$\mathcal{S}_{K,t_0}$ for some K, t_0. As there are only countably many such sets, the complement of $\mathcal{D}_\nu$ has ξ-probability 0, and so the set $\mathcal{D}_\nu$ has ξ-probability 1.

The remainder of the proof makes use of the following lemma:

Lemma 1 *Let η be a mixed strategy of the forecaster. For every $i = 1, \ldots, m$, there is a ξ-probability 1 set of distributions $\mathcal{S}^i \subset \Delta(\Omega)$ such that for every $\mu \in \mathcal{S}^i$, with (η, μ)-probability 1 on predictions λ and realizations ω, we have*

$$\lim_{n\to\infty} \frac{\sum_{t=1}^n B_t \cdot C_t^i \cdot (\omega_t - \bar{\mu}_t(\omega))}{\sum_{t=1}^n B_t \cdot C_t^i} = 0$$

whenever $\sum_t B_t \cdot C_t^i = \infty$.

Proof (Lemma 1) Consider the probability distribution U on infinite sequences $[0, 1]^\infty$ that distributes uniformly and independently every element of the sequence. Let $\{\alpha_t\}_{t=1}^\infty$ denote a realization of U. We consider the product space of the following three processes: Forecasts λ generated by the given mixed strategy η, sequences $\{\alpha_t\}_{t=1}^\infty$ sampled according to the uniform i.i.d. U, and distributions μ generated by ξ as defined above.

Define recursively the stochastic process $\{Y_t\}_{t=1}^\infty$ by $Y_t = \mathbb{1}\{\alpha_t \le \mu_t(Y_1, \ldots, Y_{t-1})\}$. That is, $Y_t = 1$ if the uniform i.i.d. realization is less or equal to the conditional distribution of μ given $Y_1, \ldots, Y_{t-1}$, and $Y_t = 0$ otherwise. Note that $Pr(Y_t = 1|Y_1, \ldots, Y_{t-1}) = \bar{\mu}_t(Y_1, \ldots, Y_{t-1})$. By the standard calibration theorems we have that with (η, U, ξ)-probability 1,

$$\lim_{n\to\infty} \frac{\sum_{t=1}^n B_t \cdot C_t^i \cdot (Y_t - \bar{\mu}_t(Y_1, \ldots, Y_{t-1}))}{\sum_{t=1}^n B_t \cdot C_t^i} = 0$$

if $\sum_t B_t \cdot C_t^i = \infty$. Hence there is a ξ-probability 1 set of distributions $\mathcal{S}^i$ such that for every $\mu \in \mathcal{S}^i$, with (η, U)-probability 1 we have (for the given μ),

$$\lim_{n\to\infty} \frac{\sum_{t=1}^n B_t \cdot C_t^i \cdot (Y_t - \bar{\mu}_t(Y_1, \ldots, Y_{t-1}))}{\sum_{t=1}^n B_t \cdot C_t^i} = 0$$

if $\sum_t B_t \cdot C_t^i = \infty$ (See, for example, Halmos 1974 Chapter VII). The conclusion follows observing that Y_t is distributed according to $\mu_t(Y_1, \ldots, Y_{t-1})$ as is ω_t. □

We now return to the proof of our proposition. Applying Lemma 1 to every $i = 1, \ldots, m$, there is a set of distributions $\mathcal{S} = \cap_i \mathcal{S}^i$ with ξ-probability 1 and such that, for all $\mu \in \mathcal{S}$, with μ-probability 1 on ω,

$$\lim_{n\to\infty} \frac{\sum_{t=1}^n B_t \cdot C_t^i \cdot (\omega_t - \bar{\mu}_t(\omega))}{\sum_{t=1}^n B_t \cdot C_t^i} = 0$$

if $\sum_t B_t \cdot C_t^i = \infty$.

By construction of the process B_t, $B_t \cdot (\lambda_t - \nu_t) \to 0$ for every λ and ω as $t \to \infty$. Also, by property #1 above, for all n large enough,

$$\left| \frac{\sum_{t=1}^{n} B_t \cdot C_t^i \cdot (\bar{\mu}_t(\omega) - \nu_t(\omega))}{\sum_{t=1}^{n} B_t \cdot C_t^i} \right| > \frac{\delta}{3}$$

if $\sum_t B_t \cdot C_t^i = \infty$.

Finally, we have $\omega_t - \lambda_t = (\omega_t - \bar{\mu}_t) + (\bar{\mu}_t - \nu_t) + (\nu_t - \lambda_t)$. The averages of the first and third pairs on the right hand side converge to 0 when multiplied by $B_t \cdot C_t^i$ but the middle pair was shown to be bounded away from 0. Hence for each $\mu \in \mathcal{S}$, with μ-probability 1 on ω,

$$\frac{\sum_{t=1}^{n} B_t \cdot C_t^i \cdot (\omega_t - \lambda_t(\omega))}{\sum_{t=1}^{n} B_t \cdot C_t^i}$$

does not converge to zero if $\sum_t B_t \cdot C_t^i = \infty$.

By property #3, $\xi(\mathcal{D}_\nu) = 1$. Hence the set $\mathcal{S} \cap \mathcal{D}_\nu$ has ξ-probability 1, and the conclusion follows by property #2. □

We also note that, much like Stewart (2011), our reference calibration test can be extended to the Bayesian setting albeit with somewhat stronger restrictions on the prior over the set of true distributions. In particular, assume $P \in \Delta(\Delta(\Omega))$ is a prior that satisfies *(a)* $P(\mathcal{D}_\nu) = 1$, *(b)* the average distribution with respect to the prior denoted $\bar{\mu} \in \Delta(\Omega)$ and defined by $\bar{\mu}_t(\omega) = \int \mu_t(\omega) dP(\mu)$ satisfies that there is some $\epsilon > 0$ such that for every ω there is some t_0 with $|\bar{\mu}_t(\omega) - \nu_t(\omega)| > \epsilon$ for all $t > t_0$ and *(c)* in addition, for every i, $\bar{\mu}_t(\omega) - \nu_t(\omega)$ has the same sign whenever $\nu_t(\omega) \in \mathcal{I}_i$.

Then we have the following corollary of Proposition 3:

Corollary 1 *Let P be a prior over distributions with reference ν satisfying the condition above. Then there exists a reference calibration test with ν that satisfies:*

- *The informed forecaster passes the test with P-probability 1: The set of true distributions μ such that the informed forecaster passes the test has P-probability 1.*
- *An uninformed forecaster fails the test with P-probability 1: For every given mixed strategy $\eta \in \Delta(\Delta(\Omega))$ of the forecaster, the set of true distributions μ such that the forecaster passes the test with positive (η, μ)-probability has P-probability 0.*

4 Concluding remarks

The non-manipulability of the reference calibration test is the result of adding an additional hurdle to the standard calibration test. Now the forecaster must also get close to the reference distribution along all realizations. This hurdle is exactly what a true distribution must possess to pass the test. Fortunately, most distributions do get infinitely often close to any given reference distribution, since those distributions that do not, form a category I set. In Stewart (2011) the added "burden of proof" requires the

forecaster to perform better than a given reference distribution in the likelihood test. In that case, the true distribution must be sufficiently far from the reference distribution to ensure that an informed forecaster can beat the reference distribution. Interestingly, the distributions that do not pass this likelihood test are those that quickly converge to the reference distribution. Both the distributions that do not have a converging subsequence and those that converge quickly[2] are topologically small, category I, sets of distributions.

We note that our results can be extended by requiring that the predictions are close to the reference distribution on a given infinite sub-sequence of times. In other words, we can require that the expert be cross-calibrated on any fixed sub-sequence of times (with respect to the reference distribution). In this case, we have to exclude all distributions that do not get infinitely often close to the reference distribution *on that particular sub-sequence.* While this set is larger, it is still a category I set, and all of our results continue to hold.

The likelihood test in Stewart (2011) uses a reference measure much in the same way as Al-Najjar and Weinstein (2008) used two forecasters under the assumption that one of them is informed. In Stewart (2011) the reference distribution is assumed to be the "false" distribution which a true informed forecaster should beat. In our paper the reference distribution represents a desired property of the true distribution—getting close to the reference distribution infinitely often. Moreover, it is the forecaster's predictions that determine when she is cross-calibrated with respect to the reference distribution. Here the reference distribution plays a somewhat different role than the one played by either the informed or the uninformed forecaster in Feinberg and Stewart (2008).

We point out that instead of cross-calibrating along a sequence converging to the reference distribution, our proofs would hold if we considered cross-calibrating whenever the forecaster makes predictions within the same intervals $\mathcal{I}_k$ as the reference distribution forecasts. Adding the requirement that the forecaster's predictions fall in these intervals infinitely often, we fail on a category I set of true distributions and retain the non-manipulability of a prequential test.[3]

Finally, while having a category I set of distributions for which the true expert might fail may seem like a small set, we point out that this is a topologically determined set that may not be intuitively small. Our test fails all true distributions that do not come infinitely often close to the reference distribution on every realization. This is a small set because most distributions, in the category sense, are "all over the place" and have conditional probabilities that take values that are dense in the interval [0, 1], quite an extreme non-systemic property for a stochastic process. This is in contrast with rules governing the distribution over time—such as being learnable (see Al-Najjar et al. 2010) where the distributions that are considered tend to have converging properties over time.

[2] Specifically, Stewart (2011) requires that the sum of squared differences on the conditionals be bounded.

[3] We are very grateful to Colin Stewart for bringing this to our attention.

References

Al-Najjar N, Weinstein J (2008) Comparative testing of experts. Econometrica 76(3):541–559
Al-Najjar N, Sandroni A, Smorodinsky R, Weinstein J (2010) Testing theories with learnable and predictive representations. J Econ Theory 145(6):2203–2217
Babaioff M, Blumrosen L, Lambert NS, Reingold O (2010) Testing expert forecasts for parameterized sets of distributions. Microsoft research technical report
Dawid A (1982) The well-calibrated Bayesian. J Am Stat Assoc 77(379):605–610
Dekel E, Feinberg Y (2006) Non-Bayesian testing of a stochastic prediction. Rev Econ Stud 73(4):893–906
Feinberg Y, Stewart C (2008) Testing multiple forecasters. Econometrica 76(3):561–582
Foster D, Vohra R (1998) Asymptotic calibration. Biometrika 85(2):379–390
Fudenberg D, Levine D (1999) An easier way to calibrate. Games Econ Behav 29(1–2):131–137
Halmos P (1974) Measure theory. Springer, New York
Kalai E, Lehrer E, Smorodinsky R (1999) Calibrated forecasting and merging. Games Econ Behav 29(1–2):151–169
Lehrer E (2001) Any inspection is manipulable. Econometrica 69(5):1333–1347
Olszewski W, Sandroni A (2008) Manipulability of future-independent tests. Econometrica 76(6):1437–1466
Olszewski W, Sandroni A (2009a) A non-manipulable test. Ann Stat 37(2):1013–1039
Olszewski W, Sandroni A (2009b) Strategic manipulation of empirical tests. Math Oper Res 34(1):57–70
Sandroni A (2003) The reproducible properties of correct forecasts. Int J Game Theory 32(1):151–159
Shmaya E (2008) Many inspections are manipulable. Theor Econ 3(3):367–382
Stewart C (2011) Non-manipulable Bayesian testing. J Econ Theory 146(5):2029–2041
Vovk V, Shafer G (2005) Good randomized sequential probability forecasting is always possible. J R Stat Soc Ser B (Statistical Methodology) 67(5):747–763