A Cardinal Comparison of Experts

Itay Kavaler [a,⋆] and Rann Smorodinsky [a]

[a] Technion–Israel Institute of Technology

Abstract. In various situations, decision makers face experts that may provide conflicting advice. This advice may be in the form of probabilistic forecasts over critical future events. We consider a setting where the two forecasters provide their advice repeatedly and ask whether the decision maker can learn to compare and rank the two forecasters based on past performance. We take an axiomatic approach and propose three natural axioms that a comparison test should comply with. We propose a test that complies with our axioms. Perhaps, not surprisingly, this test is closely related to the likelihood ratio of the two forecasts over the realized sequence of events. More surprisingly, this test is essentially unique. Furthermore, using results on the rate of convergence of supermartingales, we show that whenever the two experts' advice are sufficiently distinct, the proposed test will detect the informed expert in any desired degree of precision in some fixed finite time.

JEL classification: C11, C70, C73, D83.

Keywords: forecasting; probability; testing.

1 Introduction

Consider an individual who repeatedly consults two weather forecasting websites. It is reasonable to ask what should the individual do when the two forecasts repeatedly contradict. In what way can the individual rank the two? Should the individual trust one site and (eventually) ignore the other?

The weather example above serves as a metaphor for a plethora of settings where a decision maker faces conflicting expert advice. Take for example an elected official who must rely on professional input from civil servants, a patient who receives prognosis from various doctors or, more abstractly, a learning algorithm mechanism that uses input from various sources.

In this paper, we set the stage for defining the notion of a *cardinal comparison test*. The setting we have in mind is a sequential one. At each stage t two forecasters provide a probability over some future event (e.g., the occurrence of rain) and then the event is either realized or its complement is. Before the next day's forecasts the test must rank the two forecasters. We calibrate these ranks so they add up to one. One way to think of the rank is a recommendation for a coin flip to decide which of the two experts' advice should be taken.

We pursue a test that complies with the following set of properties which we consider natural:

Anonymity - A test is *anonymous* if it does not depend on the identity of the experts but only on their forecasts.

Error-free - A test is *error-free* if from their perspective, each of the experts cannot entertain the thought that the other expert will be overwhelmingly preferred (i.e., he assigns relatively lower probability). Another way to think about a notion of an error-free test is to assume that one of the experts has the correct model. In such a case, the test will probably not point at the second expert as the superior one.

⋆ Corresponding author.

E-mail addresses: itayk@campus.technion.ac.il (I.Kavaler), rann@technion.ac.il (R.Smorodinsky).

Reasonable - Let us consider an event, A, that has positive probability according to the first expert but relatively lower probability according to the second. Conditional on the occurrence of event A, a *reasonable* test must assign positive probability to the first expert being better informed than the second.

One thing to emphasize about the cardinal comparison test we pursue and the related properties is that they are not designed to evaluate whether either of the two forecasters is correct in some objective sense. They are only designed to compare the two. To make this point, assume that Nature follows a fair coin for deciding on rain and one forecaster insists on forecasting rain with probability 60% while the other insists on 10%. While both are wrong, a cardinal comparison test should somehow gravitate towards the former one as being better.

There is a large body of literature on expert testing that studies the question of whether a self-proclaimed expert is a true expert or a charlatan (see Section 1.2 for more details) and many of the results point to the difficulty or impossibility of designing such tests that are immune to strategic forecasters.

A comparison test may often be a more natural question than the one on whether the forecaster is correct. Indeed, when a decision maker must act, then she must choose which of the experts to follow. In the case of a single expert, the dismissal of that expert leaves the decision maker working with her own unsubstantiated beliefs, which may lead to an even worse outcome. In case a decision maker faces two forecasters with conflicting input, she may choose to somehow aggregate the two instead of dismissing one or the other. We discuss this alternative line of research in Section 1.2.

1.1 Results

Given an ordered pair of forecasters, f and g, at any finite time t, we consider the corresponding likelihood ratio of the actual outcome and calibrate it so that it and its inverse add up to one. We call this the finite derivative test at time t. We prove that this test is anonymous, error-free and reasonable. Furthermore, modulo an equivalence relation, it is unique. In fact, for any test that differs from the aforementioned construction and which is anonymous and reasonable, there exist two forecasters which render the test not error-free.

Moreover, our constructed test perfectly identifies the correct forecaster whenever the two measures induced by the forecasters are mutually singular with respect to each other. Requiring the test to identify the correct expert when the measures are not mutually singular is shown to be impossible.

A test could potentially take a long while until it converges to a verdict on the better expert. We show that the proposed comparison test converges fast and uniformly. In fact, when disregarding the stages at which the two experts provide similar forecasts, then with high probability the correct verdict will emerge in finite time that is independent of the underlying probabilities.

One can ask whether *ideal* tests can exist, that is, tests that always rank the correct forecaster higher regardless of what forecasting strategies other experts might submit. Unfortunately, this turns out to be impossible, as we discuss in Appendix A. Since an ideal test does not exist, it is natural to explore the ideality of a test over a limited class of data-generating processes. We provide a full characterization for the existence of ideal tests over sets by showing that an ideal test with respect to a set A exists if and only if, A is pairwise mutually singular.

1.2 Related literature

Single expert testing. A substantial part of the literature on expert testing focuses on the single expert setting. This literature dates back to the seminal paper of Dawid (1982), who proposes the calibration test as a scheme to evaluate the validity of weather forecasters. Dawid asserts that a test must not fail a true expert. Foster and Vohra (1998) show how a charlatan, who has no knowledge of the weather, can produce forecasts which are always calibrated. The basic ingredient that allows the charlatan to fool the test is the use of random forecasts. Lehrer (2001) and Sandroni, Smorodinsky, and Vohra (2003) extend this observation to a broader class of calibration-like tests. Finally, Sandroni (2003) shows that there exists no error-free test that is immune to such random charlatans (see also extensions of Sandroni's result in Shmaya (2008) and Olszewski and Sandroni (2008)).

To circumvent the negative results, various authors suggest limiting the set of models for which the test must be error-free (e.g., Al-Najjar, Sandroni, Smorodinsky, and Weinstein (2010) and Pomatto (2016)), or limiting the computational power associated with the charlatan (e.g., Fortnow and Vohra (2009)) or replacing measure theoretic implausibility with topological implausibility by resorting to the notion of category one sets (e.g., Dekel and Feinberg (2006)).

Multiple expert testing. Comparing performance of two (or more) experts gained very little attention in the literature. Apart from our previous work, Kavaler and Smorodinsky (2019), we are only familiar with Al-Najjar and Weinstein (2008). That paper proposes a test based on the likelihood ratio for comparing two experts. They show that if one expert knows the true process whereas the other is uninformed, then one of the following must occur: either, the test correctly identifies the informed expert, or the forecasts made by the uninformed expert are close to those made by the informed one. It turns out that the test they propose is anonymous and reasonable but is not error-free (please refer to Section 5 for the formal definition).

Another approach was suggested by Feinberg and Stewart (2008), who study an infinite horizon model of testing multiple experts, using a cross-calibration test. In their test, N experts are tested simultaneously; each expert is tested according to a calibration restricted to dates where not only does the expert have a fixed forecast but the other experts also have a fixed forecast, possibly with different values. That is to say, where the calibration test checks the empirical frequency of observed outcomes conditional on each forecast, the cross-calibration test checks the empirical frequency of observed outcomes conditional on each profile of forecasts (please refer to Appendix C for the formal definition).

They showed that if an expert predicts according to the data-generating process, the expert is guaranteed to pass the cross-calibration test with probability 1, no matter what strategies the other experts use. In addition, they prove that in the presence of an informed expert, the subset of data-generating processes under which an ignorant expert (a charlatan) will pass the cross-calibration test with positive probability, is topologically "small".

In a previous paper, Kavaler and Smorodinsky (2019), we construct a comparison test over the infinite horizon. In that paper, the test outputs one verdict at the end of all times which is in one of three forms---it points to either one of the forecasters as advantageous or it is indecisive. The main result in that paper was the identification of an essentially unique infinite-horizon, ordinal test that adheres with some natural properties. The properties studied in the current paper (as well as the associated terminology) are inspired by the ones studied in Kavaler and Smorodinsky (2019). The test we identify is based on the likelihood ratio. Interestingly, the tests identified in Al-Najjar and Weinstein (2008) and that identified by Pomatto (2016) for testable paradigms are also based on the likelihood ratio.

An alternative approach to that of comparing and ranking experts is that of aggregating forecasts by a non-Bayesian aggregator. For aggregation schemes that do well in a single stage setting, see Arieli et al. (2018), as well as Levy and Razin (2018a), and Levy and Razin (2018b); for schemes that work well in a repeated setting and produce small regret, see the rich literature in machine learning surveyed in Cesa-Bianchi and Lugosi (2006).

2 Model

At the beginning of each period $t = 1, 2, \ldots$ an outcome, ω_t, drawn randomly by Nature from the set $\Omega = \{0, 1\}$, is realized.[1] A *realization* is an infinite sequence of outcomes, $\omega := \{\omega_1, \omega_2, \ldots\} \in \Omega^\infty$. We denote by $\omega^t := \{\omega_1, \omega_2, \ldots, \omega_t\}$ to be the prefix of length t of ω (sometimes referred to as the partial history of outcomes up to period t) and use the convention that $\omega^0 := \emptyset$. At the risk of abusing notation, we will also use ω^t to denote the cylinder set $\{\hat{\omega} \in \Omega^\infty : \hat{\omega}^t = \omega^t\}$. In other words, ω^t will also denote the set of realizations which share a common prefix of length t. For any t we denote by $\mathcal{G}_t$ the σ-algebra on Ω^∞ generated by the cylinder sets ω^t and let $\mathcal{G}_\infty := \sigma(\bigcup_{t=0}^{\infty} \mathcal{G}_t)$ denote the smallest σ-algebra which consists of all cylinders (also known as the Borel σ-algebra). Let $\Delta(\Omega^\infty)$ be the set of all probability measures defined over the measurable space $(\Omega^\infty, \mathcal{G}_\infty)$.

Before ω_t is realized, two self-proclaimed experts (sometimes referred to as forecasters) simultaneously announce their forecast in the form of a probability distribution over Ω. Let $(\Omega \times \Delta(\Omega) \times \Delta(\Omega))^t$ be the set of all sequences composed of realizations and pairs of forecasts made up to time t and let $\bigcup_{t \geq 0} (\Omega \times \Delta(\Omega) \times \Delta(\Omega))^t$ be the set of all such infinite sequences.

A (pure) forecasting strategy f is a function that maps finite histories to a probability distribution over Ω. Formally, $f : \bigcup_{t \geq 0} (\Omega \times \Delta(\Omega) \times \Delta(\Omega))^t \longrightarrow \Delta(\Omega)$. Note that each forecast provided by one expert may depend, inter alia, on those provided by the other expert in previous stages. Let F denote the set of all forecasting strategies.

A probability measure $P \in \Delta(\Omega^\infty)$ naturally induces a (set of) corresponding forecasting strategy, denoted f_P, that satisfies for any $\omega \in \Omega^\infty$ and any t such that $P(\omega^t) > 0$

$$f_P(\omega^t, \cdot, \cdot)(\omega_{t+1}) = P(\omega_{t+1} | \omega^t).$$

Thus, the forecasting strategy f_P derives its forecasts from the original measure, P, via Bayes rule. Note that this does not restrict the forecast of f_P over cylinders, ω^t, for which $P(\omega^t) = 0$. [2]

In the other direction, a realization ω, and an ordered pair of forecasting strategies, $\vec{f} := (f, g)$, induce a unique play path, $(\omega, \vec{f}) \in (\Omega \times \Delta(\Omega) \times \Delta(\Omega))^\infty$, where the corresponding t - history is denoted by $(\omega, \vec{f})^t \in (\Omega \times \Delta(\Omega) \times \Delta(\Omega))^t$ started at the Null history, $(\omega, \vec{f})^0 := \emptyset$, which in turn induce a pair of probability measures, denoted for simplicity by (f, g), over Ω^∞, as follows:

$$f(\omega^t) = \prod_{n=1}^{t} f((\omega, \vec{f})^{n-1})[\omega_n], \; g(\omega^t) = \prod_{n=1}^{t} g((\omega, \vec{f})^{n-1})[\omega_n].$$

[1] For expository reasons, we restrict attention to a binary set $\Omega = \{0, 1\}$. The results extend to any finite set.

[2] An expert who uses f_P to derive the correct predictions is referred to as informed, whereas an expert who concocts predictions strategically to pass the test without any knowledge on P is referred to as uninformed.

By Kolomogorov's extension theorem, the above is sufficient in order to derive the whole measure. Observe that a pair of forecasting strategies induces a pair of probability measures, whereas each single forecasting strategy does not induce a single measure due to the dependency between the two forecasters.

2.1 A cardinal comparison test

At each stage t a third party (the 'tester') who observes the forecasts and outcomes compares the performance of both forecasters and decides who she thinks is better. Formally,

Definition 1. *A* cardinal comparison test *is a sequence* $T := (T_t)_{t>0}$*, where* $T_t : (\Omega \times \Delta(\Omega) \times \Delta(\Omega))^\infty \longrightarrow [0,1]$ *is* $\mathcal{G}_t$*—measurable for all* $t > 0$.

In other words, for any t and any realization ω and any sequence of forecasts $\vec{f}$, the tester, conditional on a t - history, announces her level of confidence that the first forecaster (the one using f) is better than the second one (we will interchangeably refer to this as his propensity that f is superior to g).[3] Note that announcing 0.5 means that both are equally capable (this should not be confused with the statement that they are both capable or both incapable). Whenever $T_t(\omega, \vec{f}) = 1$ (respectively, 0) the tester is confident that f outperforms g (respectively, g outperforms f).

Definition 2. T *is called* anonymous *if for all* $\omega \in \Omega^\infty, t > 0$ *and for all* $f, g \in F$,

$$T_t(\omega, f, g) = 1 - T_t(\omega, g, f).$$

In other words, the test's propensity at each period should not depend on the expert's identity. Note that whenever $f = g$ an anonymous test T must output a propensity of 0.5 for all $\omega \in \Omega^\infty$, $t > 0$.

For a given test T, an ordered pair of forecasting strategies $\vec{f} = (f, g)$, and a realization ω, we denote by $T(\omega, \vec{f}) = lim T_t(\omega, \vec{f})$ whenever the limit exists. For $\epsilon \in (0,1)$, let $L^{\vec{f}}_{T,\epsilon} := \{\omega : \ T(\omega, \vec{f}) > \epsilon\}$ be the set of realizations for which the limit of T exists and from some time on assigns a propensity larger than ϵ to f (similarly we denote $R^{\vec{f}}_{T,\epsilon} := \{\omega : \ T(\omega, \vec{f}) < \epsilon\}$). Notice that the following is a straightforward observation derived from Definition 2. If T is an anonymous test, then $\omega \in R^{(f,g)}_{T,\epsilon}$ if and only if $\omega \in L^{(g,f)}_{T,1-\epsilon}$; we use the last for some of our proofs.

When ω is in $L^{\vec{f}}_{T,\epsilon}$ and $\epsilon > 0.5$, the test eventually assigns a higher propensity to f than to g. On the other hand, for $\epsilon < 0.5$, the test assigns a higher propensity to g whenever ω is in $R^{\vec{f}}_{T,\epsilon}$. Thus, we will typically focus on the sets $L^{\vec{f}}_{T,\epsilon}$ with $\epsilon > 0.5$ and on the sets $R^{\vec{f}}_{T,\epsilon}$ for $\epsilon < 0.5$.

2.2 Desirable Properties

In this section, we introduce a set of axioms we deem desirable for a cardinal comparison test. Our first property asserts that any set that is contained in $R^{\vec{f}}_{T,\epsilon}$ must not be assigned a high probability according to f in comparison with the probability assigned by g. In particular, the ratio of these probabilities must be bounded by $\frac{\epsilon}{1-\epsilon}$.

[3] It should be emphasized that the results in this paper hold even for the general case for which definition 1 is extended such that a tester may condition his one step ahead decisions on his own past decisions. Formally, whenever T has the form $T_t : \ (\Omega \times \Delta(\Omega) \times \Delta(\Omega) \times \{f, g\})^\infty \longrightarrow [0,1]$.

Definition 3. *T is* error-free *if for all $\vec{f} := (f,g) \in F \times F$, for all $\epsilon \in (0, \frac{1}{2})$ and for all measurable set A*

$$f(A \cap R^{\vec{f}}_{T,\epsilon}) \leq (\frac{\epsilon}{1-\epsilon}) g(A \cap R^{\vec{f}}_{T,\epsilon}) \tag{1}$$

(Similarly, $g(A \cap L^{\vec{f}}_{T,\epsilon}) \leq (\frac{1-\epsilon}{\epsilon}) f(A \cap L^{\vec{f}}_{T,\epsilon})$ for $\epsilon \in (\frac{1}{2}, 1)$).

Note, in particular, as ϵ approaches 0, the set $R^{\vec{f}}_{T,\epsilon}$ captures the paths where g is clearly deemed better than f and so the property of error-freeness implies that although g may assign a subset of $R^{\vec{f}}_{T,\epsilon}$ a positive probability, it must be the case that f assigns it near-zero probability. On the other hand, whenever ϵ approaches 0.5, the corresponding ratio approaches 1 and so error-freeness requires that f assigns that event a probability no greater than g.

In particular, each forecaster must believe that a test cannot point out the other forecaster as correct. From his perspective, he is either preferred or the test is indecisive.

Consider a set of realizations assigned positive probability by one forecaster whereas his colleague assigns it a relatively lower probability. We shall call a test 'reasonable' if the former forecaster assigns a positive probability to the event that the test will eventually provide a high propensity to her. Formally:

Definition 4. *T is* reasonable *if for all $\vec{f} \in F \times F$, for all $\epsilon \in (0, \frac{1}{2})$ and for all measurable set A,*

$$g(A) > 0 \textit{ and } f(A) < (\frac{\epsilon}{1-\epsilon}) g(A) \Longrightarrow \; g(A \cap R^{\vec{f}}_{T,\epsilon}) > 0. \tag{2}$$

(Similarly, $f(A) > 0$ and $g(A) < (\frac{1-\epsilon}{\epsilon}) f(A) \Longrightarrow f(A \cap L^{\vec{f}}_{T,\epsilon}) > 0$ for $\epsilon \in (\frac{1}{2}, 1)$).

It should be emphasized that reasonableness and error-freeness are not related notions; examples that these properties are independent will be discussed in Section 5.

Remark 1. One could propose to replace error-freeness with a stronger and more appealing property in which a test points out the better informed expert with probability one. Informally, we would like to consider tests that have the following property $f(T(\omega, \vec{f}) = 1) = 1$ whenever $f \neq g$. However, there could be pairs of forecasters that are not equal but induce the same probability distribution. In appendix A, we formalize this and refer to tests that satisfy this stronger requirement as an *ideal*. We, furthermore show, as the name suggests, that such tests essentially do not exist.

3 An error-free and reasonable test

We now turn to propose an anonymous cardinal comparison test that is error-free and reasonable. For any pair of forecasters, $\vec{f} := (f,g) \in F \times F$, $\omega \in \Omega^{\infty}$, $t \geq 0$, the *finite derivative test*, $\mathscr{D}$, is defined as follows:

$$\mathscr{D}_{t+1}(\omega, \vec{f}) = \begin{cases} \frac{f(\omega^t)}{f(\omega^t)+g(\omega^t)}, & g(\omega^t) > 0 \; or \; f(\omega^t) > 0 \\ \frac{1}{2}, & other. \end{cases}$$

It should be noted that the ratio between $\mathscr{D}_{t+1}(\omega, \vec{f})$, the rank associated with the forecast f and $1 - \mathscr{D}_{t+1}(\omega, \vec{f})$, the rank associated with the forecast g, equals the likelihood ratio between the two forecasters. Clearly, $\mathscr{D}$ is anonymous. We turn to show that it is reasonable and error-free. Before doing so, some preliminaries are required.[4]

Lemma 1. Let $\vec{f} := (f, g)$. Then the limit of $\mathscr{D}_t(\cdot, \vec{f})$ exists and is finite $f - a.s.$

Proof. For $\omega \in \Omega^\infty$ where $f(\omega^t) > 0$ define the likelihood ratio between the two forecasters at time t as

$$D^t_f g(\omega) = \prod_{n=1}^{t} \frac{g((\omega, \vec{f})^{n-1})[\omega_n]}{f((\omega, \vec{f})^{n-1})[\omega_n]},$$

and observe that $\mathscr{D}_{t+1}(\omega, \vec{f}) = \frac{1}{1 + D^t_f g(\omega)}$.[5] Applying Lemma 2 from Kavaler and Smorodinsky (2019), we know that the limit of $D^t_f g$, denoted $D_f g$, exists and is finite $f - a.s.$ It readily follows that $\mathscr{D}(\omega, \vec{f}) := \frac{1}{1+D_f g(\omega)} := lim \mathscr{D}_t(\omega, \vec{f})$ exists and is finite $f - a.s.$ □

Now that we have established the existence and the finiteness of the test $\mathscr{D}$, let us prove that it complies with the two central properties for cardinal comparison tests:

Proposition 1. $\mathscr{D}$ is *error-free.*

Proof. Let $\vec{f} := (f, g) \in F \times F$, $\epsilon \in (0, \frac{1}{2})$, and a measurable set A. From Lemma 1, the limit of $\mathscr{D}_t(\cdot, \vec{f})$ exists and is finite $f - a.s.$ Hence,

$$A \cap R^{\vec{f}}_{\mathscr{D},\epsilon} = \{\omega \in A\colon \ \frac{1}{1 + D_f g(\omega)} \in R^{\vec{f}}_{\mathscr{D},\epsilon}\} \subset \{\omega\colon \ D_f g(\omega) \geq \frac{1-\epsilon}{\epsilon}\}.$$

Thus, applying Kavaler and Smorodinsky (2019), Lemma 2, part b, we obtain

$$f(A \cap R^{\vec{f}}_{\mathscr{D},\epsilon}) \leq (\frac{\epsilon}{1-\epsilon}) g(A \cap R^{\vec{f}}_{\mathscr{D},\epsilon}).$$

Similarly, by applying Kavaler and Smorodinsky (2019), Lemma 2, part a, we show that $g(A \cap L^{\vec{f}}_{\mathscr{D},1-\epsilon}) \leq (\frac{\epsilon}{1-\epsilon}) f(A \cap L^{\vec{f}}_{\mathscr{D},1-\epsilon})$, and $\mathscr{D}$ is error-free. □

Proposition 2. $\mathscr{D}$ is *reasonable.*

Proof. Let $\vec{f} \in F \times F$, $\epsilon \in (0, \frac{1}{2})$, and a measurable set A, and suppose (w.l.o.g.) that

$$g(A) > 0 \ and \ f(A) < (\frac{\epsilon}{1-\epsilon}) g(A). \tag{3}$$

Denote $A_1 := (A \cap L^{\vec{f}}_{\mathscr{D},\epsilon}) \cup (A \cap \{\omega\colon \ \mathscr{D}(\omega, \vec{f}) = \epsilon\})$, $A_2 := (A \cap R^{\vec{f}}_{\mathscr{D},\epsilon})$ and observe that $A = A_1 \cup A_2$. Assume by contradiction that $f(A_2) = 0$ and notice that by the construction,

$$A_1 \subset \{\omega\colon \ lim \mathscr{D}_{\sqcup}(\omega, \vec{\{}) = \frac{\infty}{\infty + \mathscr{D}_{\{}\}(\omega)} \geq \epsilon\}.$$

Thus, applying Kavaler and Smorodinsky (2019), Lemma 2, part a, together with $f(A) = f(A_1)$, we obtain that $f(A) \geq (\frac{\epsilon}{1-\epsilon}) g(A)$ which contradicts (3) and hence $f(A_2) > 0$. By similar consideration

[4] Notice that $\mathscr{D}$ is unaffected by the so-called "counterfactual" predictions. These predictions are referred to events which may not occur. On the contrary, the outcome of $\mathscr{D}$ depends only on predictions which were made along the realized play path.

[5] If $f((\omega, \vec{f})^{n-1})[\omega_n] = 0$ for some n, we set $D^t_f g(\omega) = \infty$ for all $t \geq n$.

we show that $f(A \cap L^{\vec{f}}_{\mathscr{D},\epsilon}) > 0$ whenever $f(A) > 0$ *and* $g(A) < (\frac{1-\epsilon}{\epsilon}) f(A)$ for $\epsilon \in (\frac{1}{2}, 1)$, and therefore $\mathscr{D}$ is reasonable. [6] □

Propositions 1 and 2 jointly prove our first main theorem:

Theorem 1. $\mathscr{D}$ is an anonymous, reasonable and error-free test.

We now turn to show that the finite derivative test is essentially the unique anonymous cardinal comparison test that is reasonable and error-free.

4 Uniqueness

Although there may be other error-free and reasonable cardinal comparison tests, they are essentially equivalent to the finite derivative test. To motivate this idea, consider the following example.

Example 1. Consider the realization $\tilde{\omega} := (1,1,1,,,)$, and two forecasters $\tilde{f}$ and $\tilde{g}$, both using a coin to make predictions. $\tilde{f}$ uses a fair coin whereas $\tilde{g}$ uses a biased coin with probability one for the outcome to be 1. Let $\overrightarrow{h^t_1}$ be the history of length t induced by $(\tilde{\omega}, \tilde{f}, \tilde{g})$ and let $\overleftarrow{h^t_1}$ be the one induced by $(\tilde{\omega}, \tilde{g}, \tilde{f})$. Let $c > 1$ and consider the following test:

$$T_t(\omega, \vec{f}) = \begin{cases} \mathscr{D}_t(\omega, \vec{f}), & other \\ \frac{1}{1+c \cdot D^t_f g(\omega)}, & (\omega, \vec{f})^t = \overrightarrow{h^t_1} \\ 1 - \frac{1}{1+c \cdot D^t_f g(\omega)}, & (\omega, \vec{f})^t = \overleftarrow{h^t_1}. \end{cases}$$

Hence, the propensities of T differ from those provided by $\mathscr{D}$ only along the play paths $\overrightarrow{h_1}, \overleftarrow{h_1}$, in which case the limit of T converges slower to $1, 0$, respectively, than $\mathscr{D}$.

Proposition 3. *T is an anonymous error-free and a reasonable test.*

Proof. Let $\vec{f} := (f, g) \in F \times F$, $\epsilon \in (0, \frac{1}{2})$ and a measurable set A. Recall that $\vec{f}$ and $\tilde{\omega}$ induce a unique play path, $(\tilde{\omega}, \vec{f})$. Thus, if $\tilde{\omega} \notin A \cap R^{\vec{f}}_{T,\epsilon}$ or $\tilde{\omega} \in A \cap R^{\vec{f}}_{T,\epsilon}$ and $(\tilde{\omega}, \vec{f}) \neq \overleftarrow{h_1}$, then the construction yields $A \cap R^{\vec{f}}_{T,\epsilon} = A \cap R^{\vec{f}}_{\mathscr{D},\epsilon}$. In addition, note that $(\tilde{\omega}, \vec{f}) = \overrightarrow{h_1}$ implies that $T_t(\omega, \vec{f}) \longrightarrow 1$, in which case $\tilde{\omega} \notin R^{\vec{f}}_{T,\epsilon}$. If, on the other hand, $\tilde{\omega} \in A \cap R^{\vec{f}}_{T,\epsilon}$ and $(\tilde{\omega}, \vec{f}) = \overleftarrow{h_1}$ then $T_t(\omega, \vec{f}) \longrightarrow 0$ as $cD^t_g f(\tilde{\omega}) \longrightarrow 0$. In which case $f(\tilde{\omega}) = \tilde{f}(\tilde{\omega}) = 0$. Since by Propositions 1 $\mathscr{D}$ is error-free the following is obtain

$$f(A \cap R^{\vec{f}}_{T,\epsilon}) = f(A \setminus \tilde{\omega} \cap R^{\vec{f}}_{T,\epsilon}) = f(A \setminus \tilde{\omega} \cap R^{\vec{f}}_{\mathscr{D},\epsilon}) \leq (\frac{\epsilon}{1-\epsilon}) g(A \setminus \tilde{\omega} \cap R^{\vec{f}}_{\mathscr{D},\epsilon}) \leq (\frac{\epsilon}{1-\epsilon}) g(A \cap R^{\vec{f}}_{T,\epsilon}).$$

The case for which $\epsilon \in (\frac{1}{2}, 1)$ is analogous and hence omitted. We therefore conclude that T is error-free.

To see why T is reasonable assume that $g(A) > 0$ and $f(A) \leq (\frac{\epsilon}{1-\epsilon}) g(A)$. Similar consideration shows that either $\tilde{\omega} \in A \cap R^{\vec{f}}_{T,\epsilon}$ and $(\tilde{\omega}, \vec{f}) = \overleftarrow{h_1}$, in which case $g(A \cap R^{\vec{f}}_{T,\epsilon}) = \tilde{g}(\tilde{\omega}) = 1$, or $g(A \cap R^{\vec{f}}_{T,\epsilon}) =$

[6] In fact we show a stronger result: since f is monotone and $R^{\vec{f}}_{\mathscr{D},\epsilon} = \bigcup_{\bar{\epsilon} \in \mathbb{Q} \cap (0,\epsilon]} R^{\vec{f}}_{\mathscr{D},\bar{\epsilon}}$ it follows that, conditional on $f(A_2) > 0$ there exists $\bar{\epsilon} < \epsilon$ such that $g(A \cap R^{\vec{f}}_{\mathscr{D},\bar{\epsilon}}) > 0$.

$g(A \cap R^{\vec{f}}_{\mathscr{D},\epsilon}) > 0$ where the most-right inequality holds since by proposition 2 $\mathscr{D}$ is a reasonable test. The proof for $\epsilon \in (\frac{1}{2}, 1)$ is analogous. Finally, by construction, the anonymity of $\mathscr{D}$ implies the anonymity of T. □

To capture the concept of equivalence, we introduce the following equivalence relation over tests;

Definition 5. *Let* $\vec{f} := (f, g) \in F \times F$. *We say that* $\mathrm{T} \sim_{\vec{f}} \hat{T}$ *if*

$$f(\{\omega : T(\omega, \vec{f}) \neq \hat{T}(\omega, \vec{f})\}) = g(\{\omega : T(\omega, \vec{f}) \neq \hat{T}(\omega, \vec{f})\}) = 0.$$

We say that $T \sim \hat{T}$ *if and only if* $T \sim_{\vec{f}} \hat{T}$ *for all* $\vec{f}$.

That is, two tests are equivalent if and only if, given an ordered pair of forecasting strategies, there is zero probability according to each forecaster that the tests will converge to different propensities.

Proposition 4. *The relation* $\sim$ *is an equivalence relation on* $\top := \{T : T\text{—}cardinal\ comparison\ test\}$.

The proof of Proposition 4 is relegated to Appendix B. The next theorem asserts that, up to an equivalence class representative, there exists a unique anonymous reasonable and error-free cardinal comparison test. That is, any anonymous test $T \nsim T_{\mathscr{D}}$ which is reasonable, admits an error. To this end, we will show that any $T \nsim T_{\mathscr{D}}$ can be associated with a pair of forecasting strategies for which the error-free condition fails. More importantly, the power of the theorem stems from the premise that T admits an error at any pair $\vec{f}$ whenever $T \nsim_{\vec{f}} \mathscr{D}$.

Before proceeding, we make the observation that Definition 5 can be stated equivalently by the next lemma which is invoked in our adjacent uniqueness theorem proof.

Lemma 2. Let $\vec{f} := (f, g) \in F \times F$. Then $T \sim_{\vec{f}} \hat{T}$ if and only if for all $\epsilon \in (0, 1) \cap \mathbb{Q}$

$$f((L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{\hat{T},\epsilon}) \cup (L^{\vec{f}}_{\hat{T},\epsilon} \cap R^{\vec{f}}_{T,\epsilon})) = g((L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{\hat{T},\epsilon}) \cup (L^{\vec{f}}_{\hat{T},\epsilon} \cap R^{\vec{f}}_{T,\epsilon})) = 0.$$

The proof of Lemma 2 is supplemented to Appendix B.

Theorem 2. Let T be an anonymous and reasonable cardinal comparison test. If $T \nsim \mathscr{D}$ then T is not error-free.

Proof. Assume by contradiction that T is error-free. Let $\vec{f} := (f, g)$ be such that $T \nsim_{\vec{f}} \mathscr{D}$; then from Lemma 2 there exits $\epsilon \in (0, 1)$ such that (w.l.o.g. for f)

$$f((L^{\vec{f}}_{\mathscr{D},\epsilon} \cap R^{\vec{f}}_{T,\epsilon}) \cup (L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{\mathscr{D},\epsilon})) > 0.$$

We shall consider the following cases which result in a contradiction.

Case 1: $f(L^{\vec{f}}_{\mathscr{D},\epsilon} \cap R^{\vec{f}}_{T,\epsilon}) > 0$. Assume that $\epsilon \in (\frac{1}{2}, 1)$ and observe that since $L^{\vec{f}}_{\mathscr{D},\epsilon} = \bigcup_{n \in \mathbb{N}: n > \lceil \frac{1}{1-\epsilon} \rceil}^{\infty} L^{\vec{f}}_{\mathscr{D},\epsilon+\frac{1}{n}}$ and f is monotone with respect to inclusion, there exist $\epsilon < \epsilon_1$ such that $f(\hat{A_1} := L^{\vec{f}}_{\mathscr{D},\epsilon_1} \cap R^{\vec{f}}_{T,\epsilon}) > 0$. By Proposition 1, $\mathscr{D}$ is error-free where $\hat{A_1} \subset L^{\vec{f}}_{\mathscr{D},\epsilon_1}$; hence

$$g(\hat{A_1}) \leq (\frac{1-\epsilon_1}{\epsilon_1}) f(\hat{A_1}) < (\frac{1-\epsilon}{\epsilon}) f(\hat{A_1}).$$

In addition, by the assumption T is reasonable hence

$$f(\hat{A_1} \cap L^{\vec{f}}_{T,\epsilon}) > 0,$$

which yields a contradiction since $R^{\vec{f}}_{T,\epsilon}, L^{\vec{f}}_{T,\epsilon}$ are disjoint sets.

For $\epsilon \in (0,\frac{1}{2})$ note that $R^{\vec{f}}_{T,\epsilon} = \bigcup_{n\in\mathbb{N}:n>\lceil\frac{1}{\epsilon}\rceil}^{\infty} R^{\vec{f}}_{T,\epsilon-\frac{1}{n}}$, and hence there exists $\epsilon_2 < \epsilon$ such that $f(A_2 := \{L^{\vec{f}}_{\mathscr{D},\epsilon} \cap R^{\vec{f}}_{T,\epsilon_2}\}) > 0$. By the assumption T is an error-free test, hence

$$f(\hat{A_2}) \leq (\frac{\epsilon_2}{1-\epsilon_2})g(\hat{A_1}) < (\frac{\epsilon}{1-\epsilon})g(\hat{A_2}).$$

In addition, by Proposition 1, $\mathscr{D}$ is reasonable hence

$$g(\hat{A_2} \cap R^{\vec{f}}_{\mathscr{D},\epsilon}) > 0,$$

which yields a contradiction since $R^{\vec{f}}_{\mathscr{D},\epsilon}, L^{\vec{f}}_{\mathscr{D},\epsilon}$ are disjoint sets.

Case 2: $f(\hat{A_3} := L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{\mathscr{D},\epsilon}) > 0$. Assume (w.l.o.g) that $\epsilon \in (\frac{1}{2},1)$. By the assumption T is an error-free test where, by Proposition 1, $\mathscr{D}$ is reasonable, therefore, the contradiction

$$g(\hat{A_3} \cap L^{\vec{f}}_{T,\epsilon}) > 0,$$

follows analogously from Case 1 and hence omitted. □

5 Independence of axioms

The notions of error-freeness and reasonableness which were introduced in Subsection 2.2 are not related; obviously, as the sets: $\{T = \frac{1}{2}\}$, $R^{\vec{f}}_{T,\epsilon}$, $L^{\vec{f}}_{T,1-\epsilon}$ are disjoint, inequality (1) is satisfied trivially and hence the constant fair test, $T_t(\omega,\vec{f}) \equiv 1/2$, is error-free and is not reasonable. Using the result of Theorem 1, the next example illustrates that reasonableness does not imply error-freeness.

Example 2. Let $\overrightarrow{h_2}, \overleftarrow{h_2}$ be play paths composed of the realization $\tilde{\omega} := (1,1,1,,,)$, and pairs of forecasts along $\tilde{\omega}$ which, from day two onward, are shown to have similar forecasts according to an iid distribution with parameter 1, where on day one, one forecast assigns 1 to the outcome 1 whereas the other assigns half. Let $\overrightarrow{h^t_2}, \overleftarrow{h^t_2}$ be the corresponding uniquely induced t - history and consider the following test:

$$T_t(\omega,\vec{f}) = \begin{cases} \mathscr{D}_t(\omega,\vec{f}), & other \\ 0, & (\omega,\vec{f})^t = \overrightarrow{h^t_2} \\ 1, & (\omega,\vec{f})^t = \overleftarrow{h^t_2}. \end{cases}$$

Proposition 5. *T is anonymous and reasonable but is not error-free.*

Proof. Since $\mathscr{D}$ is anonymous and $T_t(\omega,f,g) = 1 - T_t(\omega,g,f)$ whenever $(\omega,\vec{f})^t$ equals $\overrightarrow{h^t_2}$ or $\overleftarrow{h^t_2}$, it follows that T is anonymous. Further, let $\vec{\tilde{f}} = (\tilde{f},\tilde{g})$ be such that $(\tilde{\omega},\vec{\tilde{f}}) = \overrightarrow{h_2}$. Since $\{\tilde{\omega}\} = R^{\vec{\tilde{f}}}_{T,\epsilon}$ for all $\epsilon \in (0,\frac{1}{2})$ one has

$$\tilde{f}(\{\tilde{\omega}\} \cap R^{\vec{\tilde{f}}}_{T,\frac{1}{3}}) = 1 > \frac{1}{2}\tilde{g}(\{\tilde{\omega}\} \cap R^{\vec{\tilde{f}}}_{T,\frac{1}{3}})$$

and hence T is not error-free. To verify that T is a reasonable test let $\vec{f} := (f,g)$, a measurable set A, and $\epsilon \in (0,\frac{1}{2})$. If $(\tilde{\omega},\vec{f}) \neq \overrightarrow{h_2}$ and $(\tilde{\omega},\vec{f}) \neq \overleftarrow{h_2}$, then by the construction $T_t(\cdot,\vec{f}) \equiv \mathscr{D}_t(\cdot,\vec{f})$ and since by Proposition 2 $\mathscr{D}$ is a reasonable test, condition (2) is satisfied. Now, observe that if $(\tilde{\omega},\vec{f}) = \overrightarrow{h_2}$ and $\tilde{\omega} \in A$ then $f(A) = 1 > \frac{\epsilon}{1-\epsilon}g(A)$ which rules out the left hand-side of condition (2). If, on the other hand, $(\tilde{\omega},\vec{f}) = \overrightarrow{h_2}$ and $\tilde{\omega} \notin A$, then $g(A) > 0$ implies that $\hat{\omega} := (0,1,1,,,) \in A$ where $g(A \setminus \hat{\omega}) = 0$, in which case $f(A) = 0$ and $T_t(\hat{\omega},\vec{f}) \equiv \mathscr{D}_t(\hat{\omega},\vec{f}) = 0$. Hence, since by Proposition 2 $\mathscr{D}$ is reasonable we have $g(\hat{\omega} \cap R^{\vec{f}}_{T,\epsilon}) = g(A \cap R^{\vec{f}}_{T,\epsilon}) = g(A \cap R^{\vec{f}}_{\mathscr{D},\epsilon}) > 0$. For the remaining case, note that if $(\tilde{\omega},\vec{f}) = \overleftarrow{h_2}$ then either $\tilde{\omega} \in A$, in which case $f(A) \geq \frac{1}{2} > \frac{\epsilon}{1-\epsilon}g(A)$, or $\tilde{\omega} \notin A$ yielding that $g(A) = 0$. Since the case for which $\epsilon \in (\frac{1}{2},1)$ is proven analogously, the result follows. □

Al-Najjar and Weinstein (2008) introduce an alternative cardinal comparison test:

$$L_t(\omega,f,g) = \begin{cases} 0, & \frac{g(\omega^t)}{f(\omega^t)} > 1 \\ 0.5, & other \\ 1, & \frac{g(\omega^t)}{f(\omega^t)} < 1. \end{cases}$$

Note that this test differs from $\mathscr{D}$ whenever the likelihood ratio is high but finite. In our case, the test does not prefer any expert but provides a relative ranking, whereas the likelihood ratio test, L, does.[7]

Proposition 6. *L is anonymous and reasonable and is not error-free.*

Proof. Let $\epsilon \in (\frac{1}{2},1)$. Let g be a forecasting strategy which deterministically predicts $\tilde{\omega}$, and let f be such that it predicts $(1-\epsilon)$ at day one and meets g from day two onward regardless of any past history. Note that whenever f is assumed to be the true measure, then $L_t(\tilde{\omega},f,g) = \frac{1}{1-\epsilon} > 1$ for all $t > 0$ and so expert g is determinstically ranked by 1 along $(\tilde{\omega},\vec{f})$ yielding $\tilde{\omega} \in L^{\vec{f}}_{L,\epsilon}$. A simple calculation shows that

$$g(\tilde{\omega} \cap L^{\vec{f}}_{L,\epsilon}) > \frac{1-\epsilon}{\epsilon} f(\tilde{\omega} \cap L^{\vec{f}}_{L,\epsilon})$$

as $g(\tilde{\omega}) = 1$. Since ϵ is taken arbitrarily, the following important conclusion can be drawn: not only is L not error-free but it admits an arbitrarily large error. The fact that L is reasonable follows directly from Proposition 2. □

6 Decisiveness in finite time

In this section we provide a natural sufficient condition for which a tester achieves a higher level of confidence in favor of the informed forecaster with any desired degree of precision in some fixed finite time. To this end, we show the existence of a uniform bound on the rate at which a cardinal comparison test converges. Consider expert $f's$ point of view. Not only should he maintain that,

[7] A different existing test, which was introduced in Feinberg and Stewart (2008), is the cross-calibration test which is discussed in the introduction. However, it turns out that this test does not naturally induce a cardinal comparison test; rather than ranking the experts, this test outputs a binary verdict (pass/fail) for each of the two experts separately and hence may rule out anonymity. Moreover, it can be shown that any cardinal comparison test which naturally ranks an expert according to its empirical frequency would fail to be reasonable if an expert, who had calibrated only along one profile and had failed along all others, is tested against an informed expert .

whenever expert g's forecasts are different from his, then he should eventually be ranked higher than him, but if expert g's forecasts are relatively far, then this should essentially happen uniformly fast. Indeed, as we show in this section, this holds for our finite derivative test. This observation tightly builds on a theory of active supermartingales due to Fudenberg and Levine (1992).

To determine whether a test is 'almost' certain about a forecaster requires the two forecasters to provide significantly different forecasts as captured by the following definition:

Definition 6. *A pair of forecasting strategies* $\vec{f} := (f,g)$ *is* $\epsilon - close$ *along* ω *at period* $t > 0$, *if*

$$|f((\omega,\vec{f})^{t-1})[\omega_t] - g((\omega,\vec{f})^{t-1})[\omega_t]| < \epsilon$$

The next theorem asserts that, given an arbitrarily small $\epsilon > 0$, there exists a finite uniform bound, K, which is independent of any pair of forecasting strategies, such that if the forecasts of the uninformed expert are sufficiently different from those of the informed one in more than K periods, then the finite derivative test, $\mathscr{D}$, will eventually settle on the informed expert with a high level of confidence. In the latter scenario, it furthermore surprisingly asserts that, given any sufficiently large time n, $\mathscr{D}_n$ ranks the informed expert higher than $(1-\epsilon)$ and up to ϵ - amount of accuracy as it would have ranked had it continued to rank the expert following his test to infinity.

Theorem 3. For all $0 < \epsilon < 1$ there exists $K = K(\epsilon)$ such that for all $\vec{f} := (f,g)$, and for all $n > 0$, there is a set of which the probability according to f is at least $(1-\epsilon)$ such that for any ω in that set:

1. Either $\vec{f}$ is $\epsilon - close$ along ω in all but K periods in $\{1...n\}$ or
2. $\omega \in L^{f}_{\mathscr{D},1-\epsilon}$. Furthermore, $|\mathscr{D}_t(\omega,\vec{f}) - \mathscr{D}_n(\omega,\vec{f})| < \epsilon$ for all $t \geq n$.

In words, with high probability, given any sufficiently large n and any sufficiently small ϵ, the only reason that the tester is not 'almost' settled on the correct forecaster at time n (and onward) is because the uninformed expert made excellent predictions along the play path. Moreover, Theorem 3 is universal in the following manner: The bound on the number of periods in which the two experts' forecasts must be different, K, for the finite derivative test to rank the informed one higher, depends on the required level of accuracy, but is independent of any pair of forecasting strategies, f or g.

The proof of Theorem 3 is relegated to Appendix B. Nevertheless let us briefly provide some technical intuition. At the heart of the proof of Theorem 3 lies a theorem due to regarding the rate of decrease of *active* supermartingales. Consider an abstract setting with a probability measure P in $\Delta(\Omega^{\infty})$ and a filtration $\{\mathscr{G}_t\}_{t=1}^{\infty}$.

Definition 7. *A* $(\mathscr{G}_t)$ *- adapted, real-valued process* $\tilde{\mathscr{D}} := \{\tilde{\mathscr{D}}_t\}_{t=0}^{\infty}$ *is called a* supermartingale *under* P *if*

1. $E|\tilde{\mathscr{D}}_t| < \infty$ for all $t > 0$;
2. $E[\tilde{\mathscr{D}}_t|\mathscr{G}_s] \leq \tilde{\mathscr{D}}_t$ for all $s \leq t$, $P - a.s.$

Intuitively, a supermartingale is a process that decreases on average. The proof of Theorem 3 implies that the finite derivative test is associated with a supermartingale property with respect to the natural filtration which is defined in Section 2. Let us further consider the following class of supermartingales called active supermartingales. This notion was first introduced in Fudenberg and Levine (1992) who studied reputations in infinitely repeated games:

Definition 8. *A non-negative supermartingale $\tilde{\mathscr{D}}$ is* active *with activity $\psi \in (0,1)$ under P if*

$$P(\{\omega : |\frac{\tilde{\mathscr{D}}_t(\omega)}{\tilde{\mathscr{D}}_{t-1}(\omega)} - 1| > \psi\}|\tilde{\omega}^{k-1}) > \psi$$

for P - almost all histories $\tilde{\omega}^{t-1}$ such that $\tilde{\mathscr{D}}_{t-1}(\tilde{\omega}) > 0$.

In other words, a supermartingale has activity ψ if the probability of a jump of size ψ at time t exceeds ψ for almost all histories. Note that $\tilde{\mathscr{D}}$ being a supermartingale, is weakly decreasing in expectations. Showing that it is active implies that $\tilde{\mathscr{D}}_t$ substantially goes up or down relative to $\tilde{\mathscr{D}}_{t-1}$ with probability bounded away from zero in each period. Fudenberg and Levine (1992), Theorem A.1, showed the following remarkable result

Theorem 4 (Fudenberg and Levine (1992)). For every $\epsilon > 0$, $\psi \in (0,1)$, and $0 < \underline{D} < 1$ there is a time $K < \infty$ such that

$$P(\{\omega : \underset{t>K}{sup}\tilde{\mathscr{D}}_t(\omega) \leq \underline{D}\}) \geq 1-\epsilon$$

for every active supermartingale $\{\tilde{\mathscr{D}}_t\}$ with $\tilde{\mathscr{D}}_0 \equiv 1$ and activity ψ.

Theorem 4 asserts that if $\tilde{\mathscr{D}}$ is an active supermartingale with activity ψ, then there is a fixed time K by which, with high probability, $\tilde{\mathscr{D}}_t$ drops below $\underline{D}$ and remains below $\underline{D}$ for all future periods. It should be noted that the power of the theorem stems from the fact that the bound, K, depends solely on the parameters $\epsilon > 0$, ψ and $\underline{D}$, and is otherwise independent of the underlying stochastic process P.

We exploit the active supermartingale property in a different way. In the context of cardinal comparison testing, we consider two strategies, one for each expert, which are updated using Bayes rule. Given sufficiently small $\epsilon > 0$, our comparative test ranks an expert depending on whether the posterior odds ratio is above or below ϵ. The active supermartingale result implies that there is a uniform bound (independent of neither the length of the game nor the true distribution) on the number of periods where the uninformed expert can be substantially wrong, without being detected, such that if this bound is exceeded, the probability that the tester ranks high the uninformed expert is small.

7 Concluding remarks

The paper proposes a normative approach to the challenge of comparing between two forecasters who repeatedly provide probabilistic forecasts. The paper postulates three basic norms: anonymity, error-freeness and reasonableness and provides a cardinal comparison test, the finite derivative test, that complies with them. It also shows that this test is essentially unique. Finally, it shows that the test converges fast and hence is meaningful in finite time. In the future we hope to extend our results to settings with more than two forecasters and study alternative sets of norms.

7.1 Implications

The approach taken in this paper can be considered as a contribution to the hypothesis testing literature in statistics where a forecaster is associated with a hypothesis. In this context we propose a hypothesis test that complies with a set of fundamental properties which we refer to as axioms. In contrast, a central thrust for the hypothesis testing literature (for two hypotheses) is the pair of notions of significance level and power of a test. In that literature one hypothesis is considered as the null hypothesis while the other serves as an alternative. A test is designed to either reject the null hypothesis, in which case it accepts the alternative, or fail to reject it (a binary outcome). The significance level of a test is the probability of rejecting the null hypothesis whenever it is correct (type-1 error) while the power of the test is the probability of rejecting the null hypothesis assuming the alternative one is correct (the complement of a type-2 error).

In contrast with the aforementioned binary outcome that is prevalent in the hypothesis testing literature we allow, in addition, for an inconclusive outcome. Recall the celebrated Neyman-Pearson lemma which characterizes a test with the maximal power subject to an upper bound on the significance level. The possibility of an inconclusive (ranking) outcome, in our framework, allows us to design a test where both type-1 and type-2 errors have relatively low probability.[8]

Interestingly, the test proposed in the Neyman-Pearson lemma, similar to ours, also hinges on the likelihood ratio.[9] In our approach we, a priori, treat both hypotheses symmetrically. In the statistics literature, however, this is not the case and the null hypothesis is, in some sense, the status quo hypothesis. This asymmetry is manifested, for example, in the Neyman-Pearson lemma.

Note that in order to design a test that complies with a given significance level and a given power one must know the full specification of the two hypotheses. This is in contrast with our test which is universal, in the sense that it does not rely on the specifications of the two forecasts. Finally, let us comment that whereas hypothesis testing is primarily discussed in the context of a finite sample, typically from some iid distribution, our framework allows for sequences of forecasts that are dependent on past outcomes as well as past forecasts of the other expert.

Acknowledgments

Smorodinsky gratefully acknowledges the United States-Israel Binational Science Foundation and the National Science Foundation (grant 2016734), the German-Israel Foundation (grant I-1419-118.4/2017), the Ministry of Science and Technology (grant 19400214), the Technion VPR grants, and the Bernard M. Gordon Center for Systems Engineering at the Technion.

[8] Note that we abuse the statistical terminology. In statistics the notion of rejection is always used in the context of the null hypothesis. In our model, we assume symmetry between the alternatives and so we discuss rejection also in the context of the alternative hypothesis. As a consequence, an error of type-1 is defined as the probability of accepting the alternative hypothesis whenever the null hypothesis is correct, and symmetrically, an error of type-2 is the probability of accepting the null hypothesis whenever the alternative one is correct.

[9] The test proposed in the Neyman-Pearson lemma rejects the null hypothesis whenever the likelihood ratio falls below some positive threshold.

Bibliography

N.I. Al-Najjar and J. Weinstein. Comparative testing of experts. *Econometrica*, 76(3):541–559, 2008.

N.I. Al-Najjar, A. Sandroni, R. Smorodinsky, and J. Weinstein. Testing theories with learnable and predictive representations. *Journal of Economic Theory*, 145(6):2203–2217, 2010.

I. Arieli, Y. Babichenko, and R. Smorodinsky. Robust forecast aggregation. *Proceedings of the National Academy of Sciences*, 115(52):E12135–E12143, 2018.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77:605–613, 1982.

E. Dekel and Y. Feinberg. Non-bayesian testing of a stochastic prediction. *Review of Economic Studies*, 73:893–936, 2006.

Y. Feinberg and C. Stewart. Testing multiple forecasters. *Econometrica*, 76:561–582, 2008.

L. Fortnow and R. Vohra. The complexity of forecast testing. *Econometrica*, 77:93–105, 2009.

D.P. Foster and R.V. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

D. Fudenberg and D. Levine. Maintaining a reputaion when strategies are imperfectly observed. *Review of Economic Studies*, 59:561–579, 1992.

I. Kavaler and R. Smorodinsky. On comparison of experts. *Games and Economic Behavior*, 118: 94–109, 2019.

E. Lehrer. Any inspection is manipulable. *Econometrica*, 69:1333–1347, 2001.

G. Levy and R. Razin. Combining forecasts in the presence of ambiguity over correlation structures. *Unpublished results*, 2018a.

G. Levy and R. Razin. An explanation-based approach to combining forecasts. *Unpublished results*, 2018b.

W. Olszewski and A. Sandroni. Manipulability of future-independent tests. *Econometrica*, 76:1437–1466, 2008.

L. Pomatto. Testable forecasts. *Caltech. Unpublished results*, 2016.

A. Sandroni. The reproducible properties of correct forecasts. *International Journal of Game Theory*, 32:151–159, 2003.

A. Sandroni, R. Smorodinsky, and R. Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28:141–153, 2003.

E. Shmaya. Many inspections are manipulable. *Theoretical Economics*, 3:367–382, 2008.

APPENDIX

A On ideal tests

Recall that an error-free test eliminates the necessity of pointing out the less informed expert. A stronger and more appealing property is to point out the better informed expert, in which case the tester eventually settles on one forecaster as being better than the other. We consider tests that exhibit such a property as ideal. Formally,

Definition 9. *T is* decisive *on f at $(\omega,\vec{f})$ (respectively, g) if $T_t(\omega,\vec{f}) \longrightarrow 1$ (respectively, $(1-T_t(\omega,\vec{f}))\longrightarrow 1$).*

For a given $T,\vec{f}$, we denote by

$$A^{\vec{f}}_{T,f} := \{\omega :\ T \textit{ is decisive on } f \textit{ at } (\omega,\vec{f})\},$$

to be the measurable set of realizations (in $L^{\vec{f}}_{T,N}$) for which T is decisive on f at $(\omega,\vec{f})$.

Definition 10. *A test T is* ideal with respect to $A \subseteq F$ *if for all $\vec{f} := (f_, g \neq f) \in A \times A$*

$$f(A^{\vec{f}}_{T,f}) = g(A^{\vec{f}}_{T,g}) = 1.$$

It is called ideal *if it is ideal with respect to F.*

In other words, whenever the left expert knows the actual data-generating process and the right expert does not, an ideal test will surely identify the informed expert.

Trivially, any ideal test with respect to a subset of forecasts A is also error-free with respect to the same set. The following is a straightforward corollary of Theorem 1.

Corollary 1. There exists no ideal test with respect to a set of forecasts A whenever it contains two forecasts which induce measures, one of which is absolutely continuous with respect to the other.

This immediately entails:

Corollary 2. There exists no ideal test.

However, whenever A contains no such pair of forecasts, then an ideal test does exist. To prove this we must first accurately define the notion of mutually singular forecasts.

Definition 11. *Two forecasting strategies, $\vec{f} = (f, g \neq f) \in F$, are said to be mutually singular with respect to each other, if there exist two disjoint sets*

$$C^{\vec{f}}_f, C^{\vec{f}}_g \subset (\Omega \times \Delta(\Omega) \times \Delta(\Omega))^\infty$$

such that[10]

$$f(\{\omega :\ (\omega,\vec{f}) \in C^{\vec{f}}_f\}) = g(\{\omega :\ (\omega,\vec{f}) \in C^{\vec{f}}_g\}) = 1.$$

A set $A \subseteq F$ is pairwise mutually singular if for all $\vec{f} = (f, g) \neq f) \in A$, f, g are mutually singular with respect to each other.

[10] Recall that $\vec{f}$ induces a unique play path $(\omega,\vec{f})$.

The next lemma asserts that a reasonable test is able to perfectly distinguish between far measures which are induced from forecasting strategies which are mutually singular with respect to each other.

Lemma 3. Let $f, g \neq f \in F$ which are mutually singular with respect to each other. If T is reasonable then

$$f(A^{\vec{f}}_{T,f}) = g(A^{\vec{f}}_{T,g}) = 1.$$

The proof of Lemma 3 is relegated to Appendix B. It should be noted that Lemma 3 holds even for T which is not error-free.

The next theorem provides a necessary and sufficient condition for the existence of an ideal test over sets.

Theorem 5. There exists an anonymous ideal test with respect to A if and only if A is pairwise mutually singular.

Proof. $\Longleftarrow$ Directly follows from Lemma 3 and Proposition 2.

$\Longrightarrow$ Let T be an ideal anonymous test with respect to a set A. Let $\vec{f} := (f, g \neq f) \in A \times A$ and denote

$$C^T_f := \{(\omega, \vec{f}) : \omega \in A^{\vec{f}}_{T,f}\},\ C^T_g := \{(\omega, \vec{f}) : \omega \in A^{\vec{f}}_{T,g}\}.$$

Since $A^{\vec{f}}_{T,f}$, $A^{\vec{f}}_{T,g}$ are disjoint, it follows that C^T_f, C^T_g are disjoint where T ideal yields

$$f(\{\omega : (\omega, \vec{f}) \in C^T_f\}) = f(A^{\vec{f}}_{T,f}) = g(A^{\vec{f}}_{T,g}) = g(\{\omega : (\omega, \vec{f}) \in C^{\vec{f}}_g\}) = 1.$$

□

We conclude the paper with an example of an ideal test over a domain of mutually singular forecasts:

Example 3. Let

$$A_{iid} \times A_{iid} := \{\vec{f} := (f, g) : \text{there exist } a_f, a_g \in [0,1] \text{ s.t } f(\omega^t)[1] \equiv a_f,\ g(\omega^t)[1] \equiv a_g \text{ for all } \omega \in \Omega^\infty\}$$

and for $\omega \in \Omega^\infty$ denote the average realization by

$$a_\omega := \lim_{t\to\infty} \left(\frac{\sum_{n=1}^{t} 1_{\{\omega_n = 1\}}}{t} \right)$$

(whenever the limit exists). Let $\vec{f} \in A_{iid} \times A_{iid}$ such that $a_f \neq a_g$ and observe that for any ω

$$a_\omega = a_f \iff \lim_{t\to\infty} D^t_f g(\omega) = 0 \iff \lim_{t\to\infty} \mathscr{D}_t(\omega, \vec{f})) = 1.$$

Since the induced measures f, g are iid with different parameters, a mere application of the law of large numbers yields

$$f(A^{\vec{f}}_{\mathscr{D},f}) = 1 \ and \ g(A^{\vec{f}}_{\mathscr{D},f}) = 0,$$

showing that $\mathscr{D}$ is ideal with respect to A_{iid}.

B Missing proofs

Proof of Lemma 2. Observe that for all $\omega \in \{\mathrm{T} \neq \hat{T}\}$ there exists $\epsilon \in (0,1) \cap \mathbb{Q}$ such that either $\hat{T}(\omega, \vec{f}) < \epsilon < T(\omega, \vec{f})$ or $T(\omega, \vec{f}) < \epsilon < \hat{T}(\omega, \vec{f})$. We thus have

$$\{\hat{T} < T\} = \bigcup_{\epsilon \in (0,1) \cap \mathbb{Q}} \{\hat{T} < \epsilon < T\} = \bigcup_{\epsilon \in (0,1) \cap \mathbb{Q}} (L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{\hat{T},\epsilon}) \tag{B.1}$$

as well as

$$\{\hat{T} > T\} = \bigcup_{\epsilon \in (0,1) \cap \mathbb{Q}} \{T < \epsilon < \hat{T}\} = \bigcup_{\epsilon \in (0,1) \cap \mathbb{Q}} (L^{\vec{f}}_{\hat{T},\epsilon} \cap R^{\vec{f}}_{T,\epsilon}). \tag{B.2}$$

$\Longleftarrow$Assume by contradiction that $T \nsim_{\vec{f}} \hat{T}$ and observe that since $\{\mathrm{T} \neq \hat{T}\} = \{\hat{T} < T\} \cup \{\hat{T} > T\}$ it follows that (w.l.o.g. for f) either $f(\{\hat{T} < T\}) > 0$ or $f(\{\hat{T} > T\})$. If $f(\{\hat{T} < T\}) > 0$ then the most right equality of (B.1) implies that there exists $\epsilon' \in (0,1) \cap \mathbb{Q}$ such that $f(L^{\vec{f}}_{T,\epsilon'} \cap R^{\vec{f}}_{\hat{T},\epsilon'})$ >0 yielding a contradiction (similarly whenever $f(\{\hat{T} < T\}) > 0$).

$\Longrightarrow$ Assume by contradiction that (w.l.o.g. for f) $f((L^{\vec{f}}_{T,\epsilon'} \cap R^{\vec{f}}_{\hat{T},\epsilon'}) \cup (L^{\vec{f}}_{\hat{T},\epsilon'} \cap R^{\vec{f}}_{T,\epsilon'})) > 0$ for some $\epsilon' \in (0,1) \cap \mathbb{Q}$. Therefore, either $f(\bigcup_{\epsilon \in (0,1) \cap \mathbb{Q}} (L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{\hat{T},\epsilon})) > 0$ or $f(\bigcup_{\epsilon \in (0,1) \cap \mathbb{Q}} (L^{\vec{f}}_{\hat{T},\epsilon} \cap R^{\vec{f}}_{T,\epsilon})) > 0$. In which case, using (B.1) and (B.2), we conclude $f(\{\hat{T} < T\} \cup \{\hat{T} > T\}) > 0$ which contradicts that $\mathrm{T} \sim_{\vec{f}} \hat{T}$. □

Proof of proposition 4. Let $T, T_1, T_2 \in \top$, $\vec{f} \in F \times F$.
Reflexivity: Applying Lemma 2 it is readily seen that $f(L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{T,\epsilon}) = g(L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{T,\epsilon}) = 0$ as $L^{\vec{f}}_{T,\epsilon}, R^{\vec{f}}_{T,\epsilon}$ are disjoint sets for all $\epsilon \in (0,1)$. Thus, $T \sim_{\vec{f}} T$.
Anonymity: From Lemma 2 we obtain that for all $\epsilon \in (0,1) \cap \mathbb{Q}$,

$$\begin{aligned} T_1 \sim_{\vec{f}} T_2 \iff & f((L^{\vec{f}}_{T_1,\epsilon} \cap R^{\vec{f}}_{T_2,\epsilon}) \cup (L^{\vec{f}}_{T_2,\epsilon} \cap R^{\vec{f}}_{T_1,\epsilon})) = g((L^{\vec{f}}_{T_1,\epsilon} \cap R^{\vec{f}}_{T_2,\epsilon}) \cup (L^{\vec{f}}_{T_2,\epsilon} \cap R^{\vec{f}}_{T_1,\epsilon})) \\ &= f((L^{\vec{f}}_{T_2,\epsilon} \cap R^{\vec{f}}_{T_1,\epsilon}) \cup (L^{\vec{f}}_{T_1,\epsilon} \cap R^{\vec{f}}_{T_2,\epsilon})) = g((L^{\vec{f}}_{T_2,\epsilon} \cap R^{\vec{f}}_{T_1,\epsilon}) \cup (L^{\vec{f}}_{T_1,\epsilon} \cap R^{\vec{f}}_{T_2,\epsilon})) \\ &\iff T_2 \sim_{\vec{f}} T_1. \end{aligned}$$

Transitivity: Suppose by contradiction that $T_1 \sim_{\vec{f}} T$, and $T \sim_{\vec{f}} T_2$ where $T_1 \nsim_{\vec{f}} T_2$. Then from Lemma 2 (wl.o.g. for f) we are provided with $\bar{\epsilon} \in (0,1)$ such that

$$f((L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}}) \cup (L^{\vec{f}}_{T_2,\bar{\epsilon}} \cap R^{\vec{f}}_{T_1,\bar{\epsilon}})) > 0,$$

where for all $\epsilon \in (0,1)$,

$$\begin{aligned} T_1 \sim_{\vec{f}} T &\Longrightarrow f((L^{\vec{f}}_{T_1,\epsilon} \cap R^{\vec{f}}_{T,\epsilon}) \cup (L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{T_1,\epsilon})) = 0, \\ T \sim_{\vec{f}} T_2 &\Longrightarrow f((L^{\vec{f}}_{T,\epsilon} \cap R^{\vec{f}}_{T_2,\epsilon}) \cup (L^{\vec{f}}_{T_2,\epsilon} \cap R^{\vec{f}}_{T,\epsilon})) = 0. \end{aligned} \tag{B.3}$$

Case 1: $f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}}) > 0$. Note that,

$$\begin{aligned} & f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}}) = \\ &\qquad = f((L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap L^{\vec{f}}_{T,\bar{\epsilon}}) \cup (L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap (L^{\vec{f}}_{T,\bar{\epsilon}})^c)) \\ &\qquad = f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap L^{\vec{f}}_{T,\bar{\epsilon}}) + f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap R^{\vec{f}}_{T,\bar{\epsilon}}) + f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap \{T = \bar{\epsilon}\}). \end{aligned}$$

Thus, if $f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap L^{\vec{f}}_{T,\bar{\epsilon}}) > 0$, then $f(R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap L^{\vec{f}}_{T,\bar{\epsilon}}) > 0$, which contradicts the second condition of (B.3); otherwise if $f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap R^{\vec{f}}_{T,\bar{\epsilon}}) > 0$, then $f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T,\bar{\epsilon}}) > 0$, which contradicts the first condition of (B.3). Otherwise, $f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap R^{\vec{f}}_{T_2,\bar{\epsilon}} \cap \{T = \bar{\epsilon}\}) > 0$ implies that $f(L^{\vec{f}}_{T_1,\bar{\epsilon}} \cap \{T = \bar{\epsilon}\}) > 0$. In addition, since $\{L^{\vec{f}}_{T_1,\bar{\epsilon}+\frac{1}{n}}\}_{n > \lceil \frac{1}{1-\bar{\epsilon}} \rceil}$ is increasing to $L^{\vec{f}}_{T_1,\bar{\epsilon}}$ and $R^{\vec{f}}_{T,\bar{\epsilon}} \subset \{R^{\vec{f}}_{T,\bar{\epsilon}+\frac{1}{n}}\}$ for all n it follows that there exists a sufficiently large n' and $\epsilon' = \bar{\epsilon} + \frac{1}{n'}$ such that $f(L^{\vec{f}}_{T_1,\epsilon'} \cap R^{\vec{f}}_{T,\epsilon'}) > 0$, which again contradicts the first condition of (B.3).

Case 2: $f(L^{\vec{f}}_{T_2,\bar{\epsilon}} \cap R^{\vec{f}}_{T_1,\bar{\epsilon}}) > 0$. The contradiction follows analogously from Case 1 and hence omitted. □

Proposition 7. *If T is reasonable then for all $\vec{f}$ and $\epsilon \in (\frac{1}{2}, 1)$ and for all measurable set A*

$$f(A \cap R^{\vec{f}}_{T,\epsilon}) > 0 \implies g(A \cap R^{\vec{f}}_{T,\epsilon}) > 0.$$

(similarly for g where $\epsilon \in (0, \frac{1}{2})$).

Proof. Let $\vec{f}$ and $\epsilon \in (\frac{1}{2}, 1)$ and a measurable set A, and (w.l.o.g.) assume by contradiction that

$$f(A \cap R^{\vec{f}}_{T,\epsilon}) > 0 \implies g(A \cap R^{\vec{f}}_{T,\epsilon}) = 0.$$

Since $0 = g(A \cap R^{\vec{f}}_{T,\epsilon}) < (\frac{1-\epsilon}{\epsilon}) f(A \cap R^{\vec{f}}_{T,\epsilon})$ and T is reasonable (2) yields that $f(A \cap R^{\vec{f}}_{T,\epsilon} \cap L^{\vec{f}}_{T,\epsilon}) > 0$, which contradicts the fact as $R^{\vec{f}}_{T,\epsilon}, L^{\vec{f}}_{T,\epsilon}$ are disjoint sets. □

Proof of Lemma 3. W.l.o.g. let A be such that: $f(A) = 1,\ g(A) = 0$, and let $\epsilon \in (\frac{1}{2}, 1)$. Assume that $f(A \cap R^{\vec{f}}_{T,\epsilon}) > 0$, T is reasonable, therefore applying Proposition 7 with the set A yields

$$g(A \cap R^{\vec{f}}_{T,\epsilon}) > 0$$

which contradicts the assumption that $g(A) = 0$. Hence, $f(A \cap R^{\vec{f}}_{T,\epsilon}) = 0$. On the other hand, since the left-hand side of condition (2) is satisfied trivially for A, we are provided with $f(A \cap L^{\vec{f}}_{T,\epsilon}) \geq f(A \cap L^{\vec{f}}_{T,\epsilon}) > 0$. As a result,

$$1 = f(A) = f(A \cap R^{\vec{f}}_{T,\epsilon}) + f(A \cap L^{\vec{f}}_{T,\epsilon}) + f(A \cap \{T = \epsilon\})$$

and therefore $f(A \cap L^{\vec{f}}_{T,\epsilon^{1.01}}) = 1$. Similarly, assuming that $f(B) = 0,\ g(B) = 1$ we obtain that $g(R^{\vec{f}}_{T,\epsilon^{0.99}}) = 1$ for all $\epsilon \in (0, \frac{1}{2})$. Since $A^{\vec{f}}_{T,f} \subset L^{\vec{f}}_{T,\epsilon}$ for all $\epsilon \in (\frac{1}{2}, 1)$ and $L^{\vec{f}}_{T,\epsilon}$ is decreasing as $\epsilon \to 1$ (as the partition is refined) it follows that $f(A^{\vec{f}}_{T,f}) = f(\bigcap_{\epsilon} L^{\vec{f}}_{T,\epsilon}) = 1$ and the result follows. □

B.1 Decisiveness in finite time

The proof of Theorem 3 is generalized to the case where the number of elements, $|\Omega|$, is arbitrary and it is relied on achieving a uniform bound on the up-crossing probability of any non-negative supermartingale which admits sufficiently (finite) many fluctuations.

Proof of Theorem 3. Let $\epsilon \in (0,1)$. We will show that there exists a uniform constant $K = K(\epsilon)$ such that on the set of histories ω^t of $f-probability-(1-\epsilon)$, and for all $n > 0$, only two scenarios are possible; if there exists a subsequence of times $(t_i)_{i=1}^{K+1} \subset \{1,...,n\}$, and there exists a subsequence of corresponding outcomes $(\varpi_{t_i})_{i=1}^{K+1} \subset \Omega^{K+1}$ such that $|f((\omega,\vec{f})^{t_i-1})[\varpi_{t_i}] - g((\omega,\vec{f})^{t_i-1})[\varpi_{t_i}]| \geq \epsilon$ for all $1 \leq i \leq K+1$, then, the limit of $\mathscr{D}$ is strictly greater than $(1-\epsilon)$, and more importantly, the value of $\mathscr{D}$ at time n, $\mathscr{D}_n$, is $\epsilon - close$ for all ranks from time n onward. In all other scenarios, $|f((\omega,\vec{f})^{t-1})[\varpi] - g((\omega,\vec{f})^{t-1})[\varpi]| < \epsilon$ for all $\varpi \in \Omega$ in all but K periods t in $\{1,...,n\}$.

Construction of the faster process As in Fudenberg and Levine (1992), define an increasing sequence of stopping times $\{\tau_k\}_{k=0}^{\infty}$ relative to $\{D_f^t g\}$ and ϵ inductively as follows. First set $\tau_0 = 0$ and if $\tau_{k-1}(\omega) = \infty$ set $\tau_k(\omega) = \infty$. If$\tau_{k-1}(\omega) < \infty$ set $\tau_k(\omega)$ to be the smallest integer $t > \tau_{k-1}(\omega)$ such that either

$$f(\omega^{t-1}) > 0 \ and \ f(\{\bar{\omega} \in \Omega^{\infty} : \ |\frac{D_f^t g(\bar{\omega})}{D_f^{t-1} g(\bar{\omega}))}\text{-}1| > \frac{\epsilon}{|\Omega|}\}|\ \omega^{t-1}\}) > \frac{\epsilon}{|\Omega|} \tag{B.1.1}$$

or

$$\frac{D_f^t g(\omega)}{D_f^{\tau_{k-1}} g(\omega)} - 1 \geq \frac{\epsilon}{2|\Omega|}. \tag{B.1.2}$$

If there is no such t, set $\tau_k(\omega) = \infty$. Now define the process $\{\tilde{\mathscr{D}}_k\}_{k=0}^{\infty}$ by $\tilde{\mathscr{D}}_k = D_f^{\tau_k} g$ if $\tau_k < \infty$ and $\tilde{\mathscr{D}}_k = 0$ if $\tau_k = \infty$. Now, From Fudenberg and Levine (1992), Lemma 4.1, $(D_f^t g(\omega) := \frac{g(\omega^t)}{f(\omega^t)})_{t>0}$ is a supermartingale; hence from a standard result, the process $\{\tilde{\mathscr{D}}_k\}_{k=0}^{\infty}$ is a supermartingale. Furthermore, by Fudenberg and Levine (1992), Lemma 4.3, $\{\tilde{\mathscr{D}}_k\}_{k=0}^{\infty}$ is an active supermartingale with activity $\frac{\epsilon}{2|\Omega|}$.

Applying Theorem 4 with ϵ, $|\Omega|$, $acitivity = \frac{\epsilon}{2|\Omega|}$, and $\tilde{\mathscr{D}}_0 \equiv 1$, there exists an integer $K = K(\epsilon) > 0$ (depending only on these variables) such that for any active supermartingale $\{\tilde{\mathscr{D}}_k\}$ with activity $\frac{\epsilon}{2|\Omega|}$, one has

$$f(\sup_{k>K} \tilde{\mathscr{D}}_k < \epsilon) > 1 - \epsilon. \tag{B.1.3}$$

In addition, by Fudenberg and Levine (1992), Lemma 4.2, if $|f((\omega,\vec{f})^t)[\varpi] - g((\omega,\vec{f})^t)[\varpi]| > \epsilon$, for some $\varpi \in \Omega$ then condition (B.1.1) holds. Consequently, the process $\{\tilde{\mathscr{D}}_k\}_{k=0}^{\infty}$ takes into account all observations where $|f((\omega,\vec{f})^t)[\varpi] - g((\omega\vec{f})^t)[\varpi]| > \epsilon$ for some $\varpi \in \Omega$ and omits only observations where $|f((\omega,\vec{f})^t)[\varpi] - g((\omega,\vec{f})^t)[\varpi]| \leq \epsilon$ for all $\varpi \in \Omega$ (although, by condition (B.1.2), not necessarily all of them).

As a result, under the assumption that expert f is truthful (meaning, the realizations are generated via f), there exists a constant $K = K(\epsilon)$, which does not depend on the true process f or the forecasting strategy g, so that on the set of histories, ω^t, of probability $(1-\epsilon)$ under f, in all but K periods either $|f((\omega,\vec{f})^t)[\varpi] - g((\omega,\vec{f})^t)[\varpi]| \leq \epsilon$ for all $\varpi \in \Omega$ or $D_f^t g(\omega) < \epsilon$.

Now assume that there exist $K+1$ periods $(t_i)_{i=1}^{K+1} \subset \{1,...,n\}$ and $(\varpi_{t_i})_{i=1}^{K+1} \subset \Omega^{K+1}$ such that $|f((\omega,\vec{f})^{t_i-1})[\varpi_{t_i}] - g((\omega,\vec{f})^{t_i-1})[\varpi_{t_i}]| \geq \epsilon$ for all $1 \leq i \leq K+1$ with $f(\omega^{t_i-1}) > 0$ and let $n > K+1$. Then equation (B.1.3) ensures us that with $f-probability-(1-\epsilon)$

$$\tilde{\mathscr{D}}_{K+1} = D_f^{\tau_{K+1}} g < \epsilon \tag{B.1.4}$$

where by condition (B.1.2) for any $t \geq n \geq \tau_{K+1}$ we obtain that either $\tilde{\mathscr{D}}_t$ drops below ϵ or

$$D_f^t g(\omega) < D_f^{\tau_{K+1}} g(\omega)(1 + \frac{\epsilon}{2}) < \epsilon(1+\epsilon) \tag{B.1.5}$$

and hence it cannot exceed $\epsilon(1+\epsilon)$.

We conclude that there exists a constant K, which does not depend on the forecasting strategies f, g, such that for any sufficiently large $n > K$, with f - probability - $(1-\epsilon)$; if there exist $K+1$ periods in which f and g are slightly different above ϵ along a play path then the likelihood ratio at any point t after n never exceeds $\epsilon(1+\epsilon)$.

Now from (B.1.4) and (B.1.5) we conclude that with f - probability - $(1-\epsilon)$, either $|f((\omega,\vec{f})^{t-1})[\varpi] - g((\omega,\vec{f})^{t-1})[\varpi]| < \epsilon$ for all $\varpi \in \Omega$ in all but K periods t in $\{1,...,n\}$, or

$$1 - \frac{\epsilon(1+\epsilon)}{1+\epsilon(1+\epsilon)} = \frac{1}{1+\epsilon(1+\epsilon)} < \frac{1}{1+D_f^t g(\omega)} = \mathscr{D}_t(\omega,\vec{f}) \leq 1 \qquad \text{(B.1.6)}$$

for all $t \geq n$, and as a result, the liminf of $\mathscr{D}$ is always greater than $1 - \frac{\epsilon(1+\epsilon)}{1+\epsilon(1+\epsilon)}$. Consequently, for all $t \geq n$ inequality (B.1.6) yields

$$1 - \frac{\epsilon(1+\epsilon)}{1+\epsilon(1+\epsilon)} < |\mathscr{D}_n(\omega,\vec{f}) - \mathscr{D}_t(\omega,\vec{f})| \leq 1,$$

and hence $|\mathscr{D}_n(\omega,\vec{f}) - \mathscr{D}_t(\omega,\vec{f})| < \frac{\epsilon(1+\epsilon)}{1+\epsilon(1+\epsilon)}$, which, together with the first scenario, holds with f - probability - $(1-\epsilon)$. Since $\frac{\epsilon(1+\epsilon)}{1+\epsilon(1+\epsilon)} < \epsilon$, the result follows. □

C The cross-calibration test

We now restate the cross-calibration test as suggested by Feinberg and Stewart (2008). Fix a positive integer $N > 4$ and divide the interval $[0,1]$ into N equal closed subintervals $I_1, ..., I_N$, so that $I_j = [\frac{j-1}{N}, \frac{j}{N}]$, $1 \leq j \leq N$. All results in their paper hold when $[0,1]$ is replaced with the set of distributions over any finite set Ω and the intervals I_j are replaced with a cover of the set of distributions by sufficiently small closed convex subsets. At the beginning of each period $t = 1, 2...$, all forecasters (or experts) $i \in \{0,..,M-1\}$ simultaneously announce predictions $I_t^i \in \{I_1, ..., I_N\}$, which are interpreted as probabilities with which the realization 1 will occur in that period. We assume that forecasters observe both the realized outcome and the predictions of the other forecasters at the end of each period.

The cross-calibration test is defined over outcomes $(\omega_t, I_t^0, ..., I_t^{M-1})_{t=1}^{\infty}$, which specify, for each period t, the realization $\omega_t \in \Omega$, together with the prediction intervals announced by each of the M forecasters. Given any such outcome and any M - tuple $l = (I_{l^0}, ..., I_{l^{M-1}}) \in \{I_1, ..., I_N\}^M$, let $\zeta_t^l = 1_{I_t^i = I_{l^i}, \, \forall i=0,...,M-1}$, and

$$\nu_t^l = \sum_{n=1}^{t} \zeta_n^l,$$

which represents the number of times that the forecast profile l is chosen up to time n. For $\nu_t^l > 0$, the frequency f_n^l of realizations conditional on this forecast profile is given by

$$f_t^l = \frac{1}{\nu_t^l} \sum_{n=1}^{t} \zeta_n^l \omega_n.$$

Forecaster i passes the cross-calibration test at the outcome $(\omega_t, I_t^0, ..., I_t^{M-1})_{t=1}^{\infty}$ if

$$\underset{t\to\infty}{limsup} |f_t^l - \frac{2l^i - 1}{2N}| \leq \frac{1}{2N}$$

for every l satisfying $\lim_{t\to\infty} v_t^l = \infty$.

In the case of a single forecaster, the cross-calibration test reduces to the classic calibration test, which checks the frequency of realizations conditional on each forecast that is made infinitely often. With multiple forecasters, the cross-calibration test checks the empirical frequencies of the realization conditional on each profile of forecasts that occurs infinitely often. Note that if an expert is cross-calibrated, he will also be calibrated.