Efficient testing of forecasts

Ching-Lueh Chang¹ and Yuh-Dauh Lyuu²

¹ Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.
d95007@csie.ntu.edu.tw

² Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.
lyuu@csie.ntu.edu.tw

Abstract. Each day a weather forecaster predicts a probability of each type of weather for the next day. After n days, all the predicted probabilities and the real weather data are sent to a test which decides whether to accept the forecaster as possessing predicting power. Consider tests such that forecasters who know the distribution of nature are passed with high probability. Sandroni shows that any such test can be passed by a forecaster who has no prior knowledge of nature [13], provided that the duration n is known to the forecaster in advance. On the other hand, Fortnow and Vohra [7] show that Sandroni's result may require forecasters with high computational complexity and is thus of little practical relevance in some cases. We consider forecasters who select a deterministic Turing-machine forecaster according to an arbitrary distribution and then use that machine for all future forecasts. We show that forecasters even more powerful than the above ones are required for Sandroni's result. Also, we show that Sandroni's result does not apply when the duration n is not known to the forecaster in advance.

1 Introduction

The weather metaphor motivates the investigations in this paper. A weather forecaster predicts the probability of rain on each day. How do we measure the forecaster's predicting power? Dawid [3] proposes testing whether the announced forecasts are *well calibrated* in the following sense:

Suppose that, in a long (conceptually infinite) sequence of weather forecasts, we look at all those days for which the forecast probability of precipitation was, say, close to some given value ω and (assuming these form an infinite sequence) determine the long run proportion p of such days on which the forecast event (rain) in fact occurred. The plot of p against ω is termed the forecaster's empirical calibration curve. If the curve is the diagonal $p = \omega$, the forecaster may be termed (empirically) *well calibrated*.

A forecaster who knows the distribution of nature can produce well calibrated forecasts [3].

Foster and Vohra [8] show that a forecaster can produce forecasts that will be well calibrated on any distribution adopted by nature. Fudenberg and Levine [9], Lehrer [10], Sandroni, Smorodinsky and Vohra [14] consider stronger forms of calibration tests. They show that a forecaster can pass these stronger tests on any distribution of nature. Thus, to pass these calibration tests, a forecaster

needs no prior knowledge about the distribution of nature in that he does not need to assume anything about the distribution of nature [13]. In this sense, if a forecaster passes a test on any distribution of nature, it means that the forecaster needs no prior knowledge about the distribution of nature to pass that test.

Ideally, we want a test to reject forecasters who have no prior knowledge about the distribution of nature. As a corollary, it is desirable that a test rejects each forecaster on some distribution of nature, for otherwise some forecaster passes the test on all distributions of nature, meaning it needs no prior knowledge about nature to pass the test. A second desirable property is that a test passes forecasters who know the distribution of nature. These two desirable properties are not simultaneously satisfied by calibration tests (see [8, 9, 10, 14]). Sandroni [13] shows that no test satisfies both desirable properties as long as the duration is made known beforehand to the forecaster. In this known-duration case, if a test is such that forecasters who know the distribution of nature are accepted with high probability, then there exists a forecaster that can pass that test on any distribution of nature.

Fortnow and Vohra [7] consider computational bounds on the forecaster and the test. They construct efficient tests that pass forecasters who know the distribution of nature, whereas only computationally powerful forecasters have a chance to pass these tests on all possible distributions of nature. This result implies that Sandroni’s result requires the absence of computational constraints on the forecaster.

There is also closely related literature on theory testing [4, 12, 11]. In these models, an expert proposes a distribution of a stochastic process, which is tested for truthfulness. Some papers [2, 6] discuss identifying which of several experts knows the true distribution of the stochastic process.

In this paper, we present two results. The first proceeds along the computational perspective of Fortnow and Vohra [7], who show that Sandroni’s [13] result may require forecasters with high computational complexity. We consider the family of forecasters who select a deterministic Turing-machine forecaster of any time complexity according to an arbitrary distribution and then use that machine for all future forecasts. We show that Sandroni’s result requires forecasters even more powerful than the above. We achieve it by exhibiting a test such that forecasters who know the distribution of nature are passed with high probability, whereas each of the above-mentioned forecasters is rejected with high probability on some n -day weather sequence for all sufficiently large n . Unlike the results in [7] where Turing-machine forecasters with a certain computational complexity are rejected on some distributions of nature, the above-mentioned forecasters may adopt uncomputable distributions over Turing-machines of arbitrary time complexity.

Our second result shows that Sandroni’s [13] result does not apply when the duration n of forecasting is unknown beforehand to the forecaster. We do so by exhibiting a test which passes forecasters who know the distribution of nature, whereas any forecaster is rejected on some n -day weather sequence for infinitely many durations n .

Our paper is organized as follows. Section 2 gives formal definitions. Section 3–4 present our results. Section 5 concludes the paper.

2 Definitions

The following setup extends that proposed by Sandroni [13]. A finite state space $S \equiv \{1, \dots, K\}$ categorizes the outcome on each period into one of $1, \dots, K$ where $K \geq 2$. For $\ell \in \mathbb{N}$, denote by S^ℓ the ℓ -dimensional Cartesian product of S and $S^* \equiv \bigcup_{\ell \geq 0} S^\ell$. Let $n \in \mathbb{N}$ be the duration of forecasting (in periods), which may or may not be known to the forecaster and the test. Given a finite outcome sequence $s \in S^*$, the data generating process assigns a probability distribution over the outcomes for the next period. A null outcome sequence is denoted λ .

A distribution over the outcomes of S is called a forecast. The set of distributions over S is denoted ΔS [13]. A (possibly randomized) forecaster F announces a forecast for the $(t+1)$ th period given the outcomes of the first t periods and the previous forecasts by F for these t periods. Formally, the input to a forecaster F is an outcome sequence s and a forecast sequence f where $(s, f) \in \bigcup_{t \geq 0} S^t \times (\Delta S)^t$. F 's output is a forecast for the following period's outcome. Let $s = (s_1, \dots, s_{n-1}) \in S^{n-1}$. If the data generating process adopts s as the outcome sequence for the first $n-1$ periods, the n forecasts announced by a forecaster F are denoted

$$F(\lambda; 1), F(s_1; 2), \dots, F(s_1, \dots, s_{n-1}; n).$$

Here $F(s_1, \dots, s_{i-1}; i)$ is F 's forecast for the i th period. It depends only on the first $i-1$ outcomes in s and the forecasts

$$F(\lambda; 1), \dots, F(s_1, \dots, s_{i-2}; i-1)$$

for them, for $i \geq 1$. For convenience, we write

$$F(s) \equiv (F(\lambda; 1), F(s_1; 2), \dots, F(s_1, \dots, s_{n-1}; n)) \quad (1)$$

for the full forecast sequence (we may also write $F(s_1, \dots, s_{n-1})$ as $s = (s_1, \dots, s_{n-1})$). When the forecaster is a Turing machine, we assume a reasonable encoding for its output format. For example, we may let it output a forecast as a vector consisting of its predicted probabilities of the outcomes in S .

Let $\mathcal{P}$ be the distribution adopted by the data generating process. Given any outcome sequence, $\mathcal{P}$ determines the probability distribution over S for the next period. Given $\mathcal{P}$ and an outcome sequence $s = (s_1, \dots, s_n)$, the correct forecast sequence $\hat{f}(s) = (\hat{f}_1, \dots, \hat{f}_n)$ is one such that $\hat{f}_j$ equals the probability distribution that the data generating process assigns for the j th period conditioned on $(s_1, \dots, s_{j-1})$ being the outcome sequence for the first $j-1$ periods, where $1 \leq j \leq n$. Sensibly, $\hat{f}_j$ depends only on $\mathcal{P}$ and $(s_1, \dots, s_{j-1})$, not on future outcomes $(s_j, \dots, s_n)$. We may write $\hat{f}_j$ as $\hat{f}_j(s_1, \dots, s_{j-1})$ to make the dependency explicit. According to this notation,

$$\hat{f}(s) \equiv (\hat{f}_1(\lambda), \hat{f}_2(s_1), \dots, \hat{f}_n(s_1, \dots, s_{n-1})). \quad (2)$$

A test is deterministic. It receives an outcome sequence s and a corresponding forecast sequence f where $(s, f) \in \bigcup_{t \geq 0} S^t \times (\Delta S)^t$ and decides whether to accept the forecaster as possessing predicting power. A natural criterion is that, whatever the distribution $\mathcal{P}$ adopted by the data generating process, an outcome sequence $s = (s_1, \dots, s_n)$ together with the correct forecast sequence $\hat{f}(s)$ must be accepted with high probability. Here the probability is taken over the random variables $s_1, \dots, s_n$ being the outcome sequence whose distribution is determined by the data generating process. A test satisfying this criterion is said to pass the truth with high probability [7].

A few comparisons with Sandroni's [13] definitions can be made. In Sandroni's definition, the duration n is known to the forecaster and the test; the data after the n th period are simply ignored. The forecaster receives its input from $\bigcup_{t=0}^{n-1} S^t \times (\Delta S)^t$, and the test receives its input from $S^n \times (\Delta S)^n$. This paper, however, considers the consequences on the same forecaster (and test) over arbitrarily large duration n . Thus, in our definition the forecaster receives its input from $\bigcup_{t \geq 0} S^t \times (\Delta S)^t$, and the test receives its input from $\bigcup_{n \geq 0} S^n \times (\Delta S)^n$.

3 Forecasters with arbitrary time complexity

We first review previous theorems. The following theorem is due to Sandroni [13] (for the interpretations of this theorem please refer to their paper).

Theorem 1. ([13]) *Let $S \equiv \{1, \dots, K\}$ be the finite state space, $0 < \epsilon < 1$, and $n \in \mathbb{N}$ be the duration of forecasting. Consider forecasters receiving input from $\bigcup_{t=0}^{n-1} S^t \times (\Delta S)^t$ and tests receiving input from $S^n \times (\Delta S)^n$. If a test T_n passes the truth with probability at least $1 - \epsilon$, there is a forecaster F_n (which may be randomized) that is accepted by T_n with probability at least $1 - \epsilon$ on any data generating process.*

The existence of the forecaster implied in Theorem 1 is shown via Fan's minimax theorem [5]. The proof is nonconstructive, and there seem no reasons for the implied forecaster to be efficient. Indeed, Fortnow and Vohra [7] proceed along this computational perspective and show the following.

Theorem 2. ([7]) *Let $S \equiv \{1, \dots, K\}$ be the finite state space. For any time-constructible $t(n)$, there is a test T of time complexity at most $\text{poly}(n)t(n)$ and an infinite sequence $s^* = (s_1^*, s_2^*, \dots)$ over S with the following properties:*

1. *For each duration n , the test T passes the truth with high probability.*
2. *For any deterministic Turing-machine forecaster F with time complexity $t(n)$, T rejects $((s_1^*, \dots, s_n^*), F(s_1^*, \dots, s_{n-1}^*))$ for sufficiently large n .*

Theorem 2 complements Theorem 1 by giving a test which passes the truth with high probability, whereas any deterministic forecaster of time complexity $t(n)$ is rejected on some outcome sequence if the duration is sufficiently long. Fortnow and Vohra [7] also show that Theorem 2 carries over to randomized forecasters of time complexity $t(n)$.

Theorem 3. ([7]) *Let $S \equiv \{1, \dots, K\}$ be the finite state space. For any time-constructible $t(n)$, there is a test T of time complexity at most $\text{poly}(n)t(n)$ and an infinite sequence $s^* = (s_1^*, s_2^*, \dots)$ over S with the following properties:*

1. *For each duration n , the test T passes the truth with high probability.*
2. *For any randomized Turing-machine forecaster F with time complexity $t(n)$, T rejects $((s_1^*, \dots, s_n^*), F(s_1^*, \dots, s_{n-1}^*))$ with high probability for sufficiently large n .*

Clearly, for the test in Theorem 3, the forecaster implied in Theorem 1 must be highly complicated—it cannot simply be the same randomized $t(n)$ -time Turing machine for all durations n . Fortnow and Vohra [7] also show tests such that the forecasters implied in Theorem 1 can be used to do factorization or even solve **PSPACE**-complete problems, a strong indication of the high complexity of the forecasters in Theorem 1. Based on these results, Fortnow and Vohra [7] conclude that Theorem 1 may be of little practical relevance in some cases.

We are now ready to describe our results. In Theorem 2, the test T has a running time of $\text{poly}(n)t(n)$ and is guaranteed to reject any $t(n)$ -time forecasters when the data generating process adopts s^* as the outcome sequence. The test, therefore, is more complex in terms of running time than the forecasters it is to reject. The $\text{poly}(n)t(n)$ running time of T is inherent in the proof of [7] in that T simulates deterministic Turing machines of time complexity $t(n)$. We improve upon Theorem 2 by allowing T to run in $\text{poly}(n)$ time and reject every Turing-computable deterministic forecaster of any time complexity—not just $t(n)$. For our test T , the forecasters implied in Theorem 1 can not be the same deterministic Turing machine (of any time complexity) for all durations n .

Lemma 4. *Let $S \equiv \{1, \dots, K\}$ be the finite state space. There is a polynomial-time test T and an infinite sequence $s^* = (s_1^*, s_2^*, \dots)$ over S with the following properties:*

1. *For each duration n , the test T passes the truth with high probability.*
2. *For any Turing-computable deterministic forecaster F , T rejects*

$$((s_1^*, \dots, s_n^*), F(s_1^*, \dots, s_{n-1}^*))$$

for sufficiently large n .

Proof. Consider the doubly fractal sequence [1]

$$1, 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 5, \dots$$

Let $h(j)$ be the j th element of the sequence. Note that $h(j) \leq j$. We begin by describing the test T . The input to T is some outcome sequence $(s_1, \dots, s_n)$ and a forecast sequence $(f_1, \dots, f_n)$ where f_i is the forecast probability distribution over S for the i th period. For each $1 \leq j \leq n$ the test T marks j if s_j equals the smallest outcome among those with the smallest forecast probability in f_j . Now T rejects if for some $1 \leq k \leq \log n$, every $j \leq n$ with $h(j) = k$ gets marked; otherwise, T accepts. It is clear that T runs in polynomial time.

We proceed to construct s^* . Let $F_1, F_2, \dots$ be an enumeration of deterministic Turing machines that compute recursive functions. Such machines must halt on all inputs. Pick s_1^* arbitrarily from S . Inductively, once $(s_1^*, \dots, s_{j-1}^*)$ is determined, set $s_j^* \in S$ to be the outcome with the smallest forecast probability in $F_{h(j)}(s_1^*, \dots, s_{j-1}^*; j)$. If there are ties, s_j^* should be the one with the smallest value. In this way s^* is constructed. The intuition is that s_j^* is set to foil the forecast of $F_{h(j)}$ as much as possible.

We now argue item 1. Let the data generating process adopt an arbitrary distribution $\mathcal{P}$. Let $s = (s_1, \dots, s_n)$ be the outcome sequence. Here s and thus $\hat{f}(s)$ (defined in Eq. (2)) consist of random variables whose distributions are determined by the data generating process. On input s and $\hat{f}(s)$, what is the probability that T marks every $1 \leq j \leq n$ with $h(j) = k$, for a particular $k \leq \log n$? Conditioned on any particular realization $(w_1, \dots, w_{j-1})$ of the outcome sequence for the first $j-1$ periods where $h(j) = k$, let $u \in S$ be the smallest outcome among those with the smallest probability assigned by $\hat{f}_j(w_1, \dots, w_{j-1})$. Then j gets marked if and only if s_j turns out to be u , which happens with probability at most $1/K$ since u is least probable among all K possible outcomes. Since the conditional realization $(w_1, \dots, w_{j-1})$ is arbitrary, the probability that T marks every $1 \leq j \leq n$ with $h(j) = k$ is at most

$$(1/K)^{|\{1 \leq j \leq n \mid h(j)=k\}|}.$$

For $k \leq \log n$, it is not hard to see that

$$|\{1 \leq j \leq n \mid h(j) = k\}| > \sqrt{n}.$$

Therefore, the probability that T marks every $1 \leq j \leq n$ with $h(j) = k$ for some $k \leq \log n$ is at most

$$\sum_{k=1}^{\log n} (1/K)^{\sqrt{n}} = o(1).$$

That is, the correct forecast sequence is rejected only with $o(1)$ probability. This completes item 1.

We now move on to item 2. Fix a k and consider F_k . From the construction of s^* , the forecast sequence $F_k(s_1^*, \dots, s_{n-1}^*)$ made by F_k is such that for every j with $h(j) = k$, s_j^* is the smallest among the outcomes to which $F_k(s_1^*, \dots, s_{j-1}^*; j)$ assigns the smallest probability. Therefore, on input $(s_1^*, \dots, s_n^*)$ and the forecast sequence $F_k(s_1^*, \dots, s_{n-1}^*)$ announced by F_k , each j with $h(j) = k$ gets marked by T and thus T rejects if n satisfies $k \leq \log n$. $\square$

The next theorem generalizes Lemma 4 by including a class of randomized forecasters.

Theorem 5. *Let $S \equiv \{1, \dots, K\}$ be the finite state space. There is a polynomial-time test T and an infinite sequence $s^* = (s_1^*, s_2^*, \dots)$ over S with the following properties:*

1. *For each duration n , the test T passes the truth with high probability.*

2. Consider any forecaster F that adopts an arbitrary distribution over deterministic Turing-machine forecasters to select one machine for use, and then uses that same machine for all future forecasts. The test T rejects

$$((s_1^*, \dots, s_n^*), F(s_1^*, \dots, s_{n-1}^*))$$

with high probability for sufficiently large n .

Proof. T and s^* are as in Lemma 4, which already shows that T passes the truth with high probability for each duration n . Let F adopt a distribution Q over deterministic Turing-machine forecasters to select one machine for use, and then uses that same machine for all future forecasts. Let $F_1, F_2, \dots$ be an enumeration of deterministic Turing-machine forecasters. Clearly,

$$\sum_{i=1}^{\infty} \Pr[F \text{ selects } F_i] = 1.$$

For any $\epsilon > 0$, there is an $m \in \mathbb{N}$ such that

$$\sum_{i=1}^m \Pr[F \text{ selects } F_i] > 1 - \epsilon. \quad (3)$$

Assume the data generating process adopts s^* as the outcome sequence. Lemma 4 guarantees that T rejects all $F_1, \dots, F_m$ for sufficiently long duration n . Once F selects a machine for forecasting, it stays with that machine thereafter. Hence inequality (3) shows that, with probability more than $1 - \epsilon$, T rejects F . $\square$

4 Impossibility of working for all durations

A natural generalization of Theorem 1 to the case where the duration n is unknown to the forecaster is to have the forecaster implied in Theorem 1 pass the test with high probability for all durations n and data generating processes. This is because a forecaster who does not know the duration n beforehand is guaranteed to pass a test with high probability if and only if it is so for every duration n (note that $n \in \mathbb{N}$ may be arbitrary). Formally, given an arbitrary test T that passes the truth with high probability, is it possible to find a forecaster F such that for each duration n and each data generating process, F is accepted by T with high probability? However, this is impossible as the following theorem shows.

Theorem 6. *Let $S \equiv \{1, \dots, K\}$ be the finite state space and $0 < \epsilon < 1$. There is a polynomial-time test T with the following properties:*

1. *For each duration n , the test T passes the truth with probability greater than $1 - \epsilon$.*
2. *For every randomized forecaster F and $\delta > 0$, there are infinitely many $n \in \mathbb{N}$ and outcome sequences $s^* = (s_1^*, \dots, s_n^*)$ such that T accepts $(s^*, F(s_1^*, \dots, s_{n-1}^*))$ with probability at most $1/2 + \delta$.*

Proof. We first describe the test T . Consider an arbitrary outcome sequence $s = (s_1, \dots, s_n)$ and a forecast sequence $f = (f_1, \dots, f_n)$. Denote by $f_i^{[1]}$ the probability f_i assigns to the outcome $1 \in S$. On input (s, f) , the test T rejects only in the following two cases:

case 1. $s_1 = \dots = s_n = 1$ and $\prod_{i=1}^n f_i^{[1]} < \epsilon/2$.

case 2. $s_1 = \dots = s_{n-1} = 1, s_n \neq 1$, and $f_n^{[1]} > 1 - \epsilon/2$.

The test T clearly runs in polynomial time.

To argue item 1, fix $n \in \mathbb{N}$ and a distribution $\mathcal{P}$ adopted by the data generating process. Let $s = (s_1, \dots, s_n)$ be the outcome sequence for the first n periods and $\hat{f}(s)$ be the corresponding correct forecast sequence. Here s and thus $\hat{f}(s)$ are random variables whose distributions are determined by the data generating process. By the definition of the correct forecast sequence $\hat{f}(s)$, the probability that $s_1 = \dots = s_n = 1$ is

$$\prod_{i=1}^n \Pr[s_i = 1 \mid s_1 = \dots = s_{i-1} = 1] = \prod_{i=1}^n \hat{f}_i^{[1]}(1^{i-1}),$$

where $\hat{f}_i^{[1]}(1^{i-1})$ is the probability assigned to the outcome $1 \in S$ in $\hat{f}_i(1^{i-1})$. Note that $\hat{f}_i^{[1]}(1^{i-1})$ is not a random variable but a fixed number determined by $\mathcal{P}$ for $1 \leq i \leq n$. Now, feed $(s, \hat{f}(s))$ to T . The test T rejects due to case 1 only if $s_1, \dots, s_n$ all turn out to be 1 and $\prod_{i=1}^n \hat{f}_i^{[1]}(1^{i-1}) < \epsilon/2$. But

$$\prod_{i=1}^n \hat{f}_i^{[1]}(1^{i-1})$$

is precisely the probability that $s_1 = \dots = s_n = 1$. Hence, if $\prod_{i=1}^n \hat{f}_i^{[1]}(1^{i-1}) < \epsilon/2$ holds, T rejects due to case 1 with probability less than $\epsilon/2$. Similarly, T rejects due to case 2 only if $s_1, \dots, s_{n-1}$ turn out to be 1 but s_n turns out otherwise, and $\hat{f}_n^{[1]}(1^{n-1}) > 1 - \epsilon/2$. Assume $\hat{f}_n^{[1]}(1^{n-1}) > 1 - \epsilon/2$. When conditioned on $s_1 = \dots = s_{n-1} = 1$, the probability that s_n turns out 1 is $\hat{f}_n^{[1]}(1^{n-1}) > 1 - \epsilon/2$, so that s_n turns out otherwise with probability less than $\epsilon/2$. Therefore, the probability that T rejects due to case 2 is less than $\epsilon/2$. To sum up, the probability that T rejects $(s, \hat{f}(s))$ on either case is less than $\epsilon/2 + \epsilon/2 = \epsilon$. This completes item 1.

We now move on to item 2. Let F be an arbitrary (possibly randomized) forecaster and $\delta > 0$. Let $m > d_{\epsilon, \delta}$ for some $d_{\epsilon, \delta}$ to be determined later. Assume the data generating process picks the outcome sequence $1^m \in S^m$ for the first m periods. T rejects $(1^m, F(1^{m-1}))$ because of case 1 if $\prod_{i=1}^m F^{[1]}(1^{i-1}; i) < \epsilon/2$. Thus, either

$$\Pr[T \text{ accepts } (1^m, F(1^{m-1}))] \leq \frac{1}{2} + \delta, \quad (4)$$

or

$$\Pr\left[\prod_{i=1}^m F^{[1]}(1^{i-1}; i) \geq \epsilon/2\right] > \frac{1}{2} + \delta, \quad (5)$$

where the probability is taken over the random variables in $F(1^{m-1})$. Assume inequality (5) holds. The event $\prod_{i=1}^m F^{[1]}(1^{i-1}; i) \geq \epsilon/2$ implies

$$\left| \{i \mid F^{[1]}(1^{i-1}; i) \leq 1 - \epsilon/2, 1 \leq i \leq m\} \right| \leq c_\epsilon$$

where

$$c_\epsilon \equiv \frac{\log(\epsilon/2)}{\log(1 - \epsilon/2)}.$$

The above and inequality (5) imply

$$\Pr \left[\left| \{i \mid F^{[1]}(1^{i-1}; i) \leq 1 - \epsilon/2, 1 \leq i \leq m\} \right| \leq c_\epsilon \right] > \frac{1}{2} + \delta$$

where the probability is taken over the random variables in $F(1^{m-1})$. Now set $d_{\epsilon, \delta} \equiv c_\epsilon(1 + 2\delta)/\delta$. The above inequality trivially implies

$$\Pr \left[\left| \{i \mid F^{[1]}(1^{i-1}; i) \leq 1 - \epsilon/2, m - d_{\epsilon, \delta} < i \leq m\} \right| \leq c_\epsilon \right] > \frac{1}{2} + \delta,$$

which is equivalent to

$$\Pr \left[\left| \{i \mid F^{[1]}(1^{i-1}; i) > 1 - \epsilon/2, m - d_{\epsilon, \delta} < i \leq m\} \right| \geq d_{\epsilon, \delta} - c_\epsilon \right] > \frac{1}{2} + \delta \quad (6)$$

where the probability is taken over the random variables in $F(1^{m-1})$.

Now add the random variable r uniformly distributed over $\{m - d_{\epsilon, \delta} + 1, m - d_{\epsilon, \delta} + 2, \dots, m\}$ and independent of the random variables in $F(1^{m-1})$. Inequality (6) and the independence of r from the random variables in $F(1^{m-1})$ imply

$$\Pr \left[F^{[1]}(1^{r-1}; r) > 1 - \epsilon/2 \right] > \left(\frac{1}{2} + \delta \right) \frac{d_{\epsilon, \delta} - c_\epsilon}{d_{\epsilon, \delta}} \quad (7)$$

where the probability is taken over r and the random variables in $F(1^{m-1})$. Due to the independence of r from the random variables in $F(1^{m-1})$, we have

$$\Pr \left[F^{[1]}(1^{r-1}; r) > 1 - \epsilon/2 \right] = \sum_{i=m-d_{\epsilon, \delta}+1}^m \frac{\Pr \left[F^{[1]}(1^{i-1}; i) > 1 - \epsilon/2 \right]}{d_{\epsilon, \delta}} \quad (8)$$

where the probability on the left-hand side is taken over r and the random variables in $F(1^{m-1})$, and those within the summation are over the random variables in $F(1^{m-1})$. Inequalities (7) and (8) imply the existence of a number $i(m) \in \{m - d_{\epsilon, \delta} + 1, m - d_{\epsilon, \delta} + 2, \dots, m\}$ with

$$\Pr \left[F^{[1]}(1^{i(m)-1}; i(m)) > 1 - \epsilon/2 \right] > \left(\frac{1}{2} + \delta \right) \cdot \frac{d_{\epsilon, \delta} - c_\epsilon}{d_{\epsilon, \delta}} = \frac{1}{2} + \frac{\delta}{2} \quad (9)$$

where the probability is taken over the random variables in $F(1^{m-1})$.

Consider the sequence $1^{i(m)-1}2 \in S^*$ standing for $\overbrace{1 \cdots 1}^{i(m)-1}2$. Observe that inequality (9) holds also when the probability is taken over the random variables in $F(1^{i(m)-1})$ because 1^m and $1^{i(m)-1}$ share a common prefix, $1^{i(m)-1}$.

Inequality (9) therefore says that, if the data generating process adopts the outcome sequence $1^{i(m)-1}2$ for the first $i(m)$ periods, then with probability more than $1/2 + \delta/2$ the forecast that F makes for the $i(m)$ th period attaches more than $1 - \epsilon/2$ to the probability on the outcome 1. But then T rejects $(1^{i(m)-1}2, F(1^{i(m)-1}))$ with probability more than $1/2 + \delta/2$ (see case 2). In summary, for each $m > d_{\epsilon,\delta}$, either inequality (4) holds, or T rejects $(1^{i(m)-1}2, F(1^{i(m)-1}))$ with probability more than $1/2 + \delta/2$. $\square$

Our proof of Theorem 6 actually shows the following stronger result.

Theorem 7. *Let $S \equiv \{1, \dots, K\}$ be the finite state space and $0 < \epsilon < 1$. There is a polynomial-time test T with the following properties:*

1. *For each duration n , the test T passes the truth with probability greater than $1 - \epsilon$.*
2. *Consider an arbitrary, possibly randomized forecaster F and $\delta > 0$. Let $d_{\epsilon,\delta} \equiv (1 + 2\delta) \log(\epsilon/2) / (\delta \log(1 - \epsilon/2))$. For sufficiently large $m \in \mathbb{N}$, there is an $n \in \{m - d_{\epsilon,\delta} + 1, m - d_{\epsilon,\delta} + 2, \dots, m\}$ and an outcome sequence $s^* = (s_1^*, \dots, s_n^*) \in S^n$ such that T accepts $(s^*, F(s_1^*, \dots, s_{n-1}^*))$ with probability at most $1/2 + \delta$.*

Thus, for the test T in Theorem 7, the outcome sequences on which a forecaster (computable or not) cannot perform well are not rare. One difference of Theorem 7 from [11] is that T may accept sequence pairs with prefixes rejected by T .

5 Conclusion

We have built computationally efficient tests with various desirable properties. The tests that we construct pass the truth with high probability, and it is hard in various ways to pass these tests for all data generating processes. Thus, it is hard for a forecaster to pass our tests without prior knowledge on the underlying data generating process.

Unlike most previous works except that by Fortnow and Vohra [7], our results take a computational perspective on forecast testing. Our tests run in polynomial time, and our first result requires the forecaster to have enormous computational power to pass the test on all data generating processes. As suggested by Fortnow and Vohra [7], we believe that taking a computational perspective may shed some light on many other problems previously studied without computational considerations.

Bibliography

- [1] Franklin T. Adams-Watters. <http://www.research.att.com/~njas/sequences/a002260.txt>.
- [2] N. I. Al-Najjar and J. Weinstein. Comparative testing of experts. Levine's Working Paper Archive 321307000000000590, Department of Economics, UCLA, 2006.
- [3] A.P. Dawid. The well calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–613, 1982.
- [4] E. Dekel and Y. Feinberg. Non-Bayesian testing of a stochastic prediction. *Review of Economic Studies*, 73(4):893–906, 2006.
- [5] K. Fan. Minimax theorems. *Proceedings of the National Academy of Science U.S.A.*, 39:42–47, 1953.
- [6] Y. Feinberg and C. Stewart. Testing multiple forecasters. Research Paper Series 1957, Graduate School of Business, Stanford University, 2007.
- [7] L. Fortnow and R.V. Vohra. The complexity of forecast testing. Technical Report TR06-149, Electronic Colloquium on Computational Complexity, 2006.
- [8] D.P. Foster and R.V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.
- [9] D. Fudenberg and D. Levine. Conditional universal consistency. *Games and Economic Behavior*, 29:104–130, 1999.
- [10] E. Lehrer. Any inspection rule is manipulable. *Econometrica*, 69(5):1333–1347, 2001.
- [11] W. Olszewski and A. Sandroni. Counterfactual predictions. Technical report, Northwestern University, Department of Economics, 2006.
- [12] W. Olszewski and A. Sandroni. Strategic manipulation of empirical tests. Technical report, Northwestern University, Department of Economics, 2006.
- [13] A. Sandroni. The reproducible properties of correct forecasts. *International Journal of Game Theory*, 32(1):151–159, 2003.
- [14] A. Sandroni, R. Smorodinsky, and R.V. Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28(1):141–153, 2003.