Predictions as statements and decisions (draft: comments welcome)

Vladimir Vovk
vovk@cs.rhul.ac.uk
http://vovk.net

October 21, 2013

Abstract

Prediction is a complex notion, and different predictors (such as people, computer programs, and probabilistic theories) can pursue very different goals. In this paper I will review some popular kinds of prediction and argue that the theory of competitive on-line learning can benefit from the kinds of prediction that are now foreign to it.

The standard goal for predictor in learning theory is to incur a small loss for a given loss function measuring the discrepancy between the predictions and the actual outcomes. Competitive on-line learning concentrates on a "relative" version of this goal: the predictor is to perform almost as well as the best strategies in a given benchmark class of prediction strategies. Such predictions can be interpreted as decisions made by a "small" decision maker (i.e., one whose decisions do not affect the future outcomes).

Predictions, or *probability forecasts*, considered in the foundations of probability are statements rather than decisions; the loss function is replaced by a procedure for testing the forecasts. The two main approaches to the foundations of probability are measure-theoretic (as formulated by Kolmogorov) and game-theoretic (as developed by von Mises and Ville); the former is now dominant in mathematical probability theory, but the latter appears to be better adapted for uses in learning theory discussed in this paper.

An important achievement of Kolmogorov's school of the foundations of probability was construction of a universal testing procedure and realization (Levin, 1976) that there exists a forecasting strategy that produces ideal forecasts. Levin's ideal forecasting strategy, however, is not computable. Its more practical versions can be obtained from the results of game-theoretic probability theory. For a wide class of forecasting protocols, it can be shown that for any computable game-theoretic law of probability there exists a computable forecasting strategy that produces ideal forecasts, as far as this law of probability is concerned. Choosing suitable laws of probability we can ensure that the forecasts agree with reality in requisite ways.

Probability forecasts that are known to agree with reality can be used for making good decisions: the most straightforward procedure is to select decisions that are optimal under the forecasts (the principle of minimum expected loss). This gives, *inter alia*, a powerful tool for competitive on-line learning; I will describe its use for designing prediction algorithms that satisfy the property of universal consistency and its more practical versions.

In conclusion of the paper I will discuss some limitations of competitive on-line learning and possible directions of further research.

Contents

1 Introduction

This paper is based on my invited talk at the 19th Annual Conference on Learning Theory (Pittsburgh, PA, June 24, 2006). In recent years COLT invited talks have tended to aim at establishing connections between the traditional concerns of the learning community and the work done by other communities (such as game theory, statistics, information theory, and optimization). Following this tradition, I will argue that some ideas from the foundations of probability can be fruitfully applied in competitive on-line learning.

In this paper I will use the following informal taxonomy of predictions (reminiscent of Shafer's [36], Figure 2, taxonomy of probabilities):

D-predictions are mere Decisions. They can never be true or false but can be good or bad. Their quality is typically evaluated with a loss function.

S-predictions are Statements about reality. They can be tested and, if found inadequate, rejected as false.

F-predictions (or Frequentist predictions) are intermediate between D-predictions and S-predictions. They are successful if they match the frequencies of various observed events.

Traditionally, learning theory in general and competitive on-line learning in particular consider D-predictions. I will start, in Section 2, from a simple asymptotic result about D-predictions: there exists a universally consistent on-line prediction algorithm (randomized if the loss function is not required to be convex in the prediction). Section 3 is devoted to S-prediction and Section 4 to F-prediction. We will see that S-prediction is more fundamental than, and can serve as a tool for, F-prediction. Section 6 explains how F-prediction (and so, indirectly, S-prediction) is relevant for D-prediction. In Section 7 I will prove the result of Section 2 about universal consistency, as well as its non-asymptotic version.

2 Universal consistency

In all prediction protocols in this paper every player can see the other players' moves made so far (they are *perfect-information* protocols). The most basic one is:

Prediction protocol

```
FOR $n = 1, 2, \dots$:
    Reality announces $x_n \in \mathbf{X}$.
    Predictor announces $\gamma_n \in \Gamma$.
    Reality announces $y_n \in \mathbf{Y}$.
END FOR.
```

At the beginning of each round n Predictor is given some data x_n relevant to predicting the following observation y_n; x_n may contain information about n and the previous observations $y_{n-1}, y_{n-2}, \dots$. The data is taken from the *data space* $\mathbf{X}$ and the observations from the *observation space* $\mathbf{Y}$. The predictions γ_n are taken from the *prediction space* Γ, and a prediction's quality in light of the actual observation is measured by a *loss function* $\lambda : \mathbf{X} \times \Gamma \times \mathbf{Y} \to \mathbb{R}$. This is how we formalize D-predictions. The prediction protocol will sometimes be referred to as the "prediction game" (in general, "protocol" and "game" will be used as synonyms, with a tendency to use "protocol" when the players' goals are not clearly stated; for example, a prediction game is a prediction protocol complemented by a loss function).

We will always assume that the data space $\mathbf{X}$, the prediction space Γ, and the observation space $\mathbf{Y}$ are non-empty topological spaces and that the loss function λ is continuous. Moreover, we are mainly interested in the case where $\mathbf{X}$, Γ, and $\mathbf{Y}$ are locally compact metric spaces, the prime examples being Euclidean spaces and their open and closed subsets. Traditionally only loss functions $\lambda(x, \gamma, y) = \lambda(\gamma, y)$ that do not depend on x are considered in learning theory, and this case appears to be most useful and interesting. The reader might prefer to concentrate on this case.

Predictor's total loss over the first N rounds is $\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n)$. As usual in competitive on-line prediction (see [9] for a recent book-length review of the field), Predictor competes with a wide range of *prediction rules* $D : \mathbf{X} \to \Gamma$. The total loss of such a prediction rule is $\sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n)$, and so Predictor's goal is to achieve

$$\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) \lessapprox \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) \tag{1}$$

for all $N = 1, 2, \ldots$ and as many prediction rules D as possible.

Predictor's strategies in the prediction protocol will be called *on-line prediction algorithms* (or *strategies*).

Remark 1 Some common prediction games are not about prediction at all, as this word is usually understood. For example, in Cover's game of sequential investment ([9], Chapter 10) with K stocks,

$$\mathbf{Y} := [0, \infty)^K, \quad \Gamma := \bigl\{(g_1, \ldots, g_K) \in [0, \infty)^K \mid g_1 + \cdots + g_K = 1\bigr\},$$
$$\lambda\bigl((g_1, \ldots, g_K), (y_1, \ldots, y_K)\bigr) := -\ln \sum_{k=1}^{K} g_k y_k.$$

(there is no $\mathbf{X}$; or, more formally, $\mathbf{X}$ consists of one element which is omitted from our notation). The observation y is interpreted as the ratios of the closing to opening price of the K stocks and the "prediction" γ is the proportions of the investor's capital invested in different stocks at the beginning of the round. The loss function is the minus logarithmic increase in the investor's capital. In this example γ can hardly be called a prediction: in fact it is a decision made by a small decision maker, i.e., decision maker whose actions do not affect Reality's future behavior (see Section 8 for a further discussion of this aspect of competitive on-line prediction). For other games of this kind, see [52].

Universal consistency for deterministic prediction algorithms

Let us say that a set in a topological space is *precompact* if its closure is compact. In Euclidean spaces, precompactness means boundedness. An on-line prediction algorithm is *universally consistent* for a loss function λ if its predictions γ_n always satisfy

$$\bigl(\{x_1, x_2, \ldots\} \text{ and } \{y_1, y_2, \ldots\} \text{ are precompact}\bigr)$$
$$\Longrightarrow \limsup_{N \to \infty} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) \right) \le 0 \tag{2}$$

for any continuous prediction rule $D : \mathbf{X} \to \Gamma$. The intuition behind the antecedent of (2), in the Euclidean case, is that the prediction algorithm knows that $\|x_n\|$ and $\|y_n\|$ are bounded but does not know an upper bound in advance.

Of course, universal consistency is only a minimal requirement for successful prediction; we will also be interested in bounds on the predictive performance of our algorithms.

Let us say that the loss function λ is *compact-type* if for each pair of compact sets $A \subseteq \mathbf{X}$ and $B \subseteq \mathbf{Y}$ and each constant M there exists a compact set $C \subseteq \Gamma$ such that

$$\forall x \in A, \gamma \notin C, y \in B : \quad \lambda(x, \gamma, y) > M.$$

More intuitively, we require that $\lambda(x, \gamma, y) \to \infty$ as $\gamma \to \infty$ uniformly in (x, y) ranging over a compact set.

Theorem 1 *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are locally compact metric spaces,* Γ *is a convex subset of a Fréchet space, and the loss function* $\lambda(x, \gamma, y)$ *is continuous, compact-type, and convex in the variable* $\gamma \in \Gamma$. *There exists a universally consistent on-line prediction algorithm.*

To have a specific example in mind, the reader might check that $\mathbf{X} = \mathbb{R}^K$, $\Gamma = \mathbf{Y} = \mathbb{R}^L$, and $\lambda(x, \gamma, y) := \|y - \gamma\|$ satisfy the conditions of the theorem.

Universal consistency for randomized prediction algorithms

When the loss function $\lambda(x, \gamma, y)$ is not convex in γ, two difficulties appear:

- the conclusion of Theorem 1 becomes false if the convexity requirement is removed ([19], Theorem 2);
- in some cases the notion of a continuous prediction rule becomes vacuous: e.g., there are no non-constant continuous prediction rules when $\Gamma = \{0, 1\}$ and $\mathbf{X}$ is connected.

To overcome these difficulties, we consider randomized prediction rules and randomized on-line prediction algorithms (with independent randomizations). It will follow from the proof of Theorem 1 that one can still guarantee that (2) holds, although with probability one; on the other hand, there will be a vast supply of continuous prediction rules.

Remark 2 In fact, the second difficulty is more apparent than real: for example, in the binary case ($\mathbf{Y} = \{0, 1\}$) with the loss function $\lambda(\gamma, y)$ independent of x, there are many non-trivial continuous prediction rules in the canonical form of the prediction game [45] with the prediction set redefined as the boundary of the set of superpredictions [19].

A *randomized prediction rule* is a function $D : \mathbf{X} \to \mathcal{P}(\Gamma)$ mapping the data space into the probability measures on the prediction space; $\mathcal{P}(\Gamma)$ is always equipped with the topology of weak convergence [6]. A *randomized on-line prediction algorithm* is an on-line prediction algorithm in the extended prediction game with the prediction space $\mathcal{P}(\Gamma)$. Let us say that a randomized on-line prediction algorithm is *universally consistent* if, for any continuous randomized prediction rule $D : \mathbf{X} \to \mathcal{P}(\Gamma)$,

$$\bigl(\{x_1, x_2, \ldots\} \text{ and } \{y_1, y_2, \ldots\} \text{ are precompact}\bigr)$$
$$\Longrightarrow \left(\limsup_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, g_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, d_n, y_n) \right) \le 0 \text{ a.s.} \right) \quad (3)$$

where $g_1, g_2, \ldots, d_1, d_2, \ldots$ are independent random variables with g_n distributed as γ_n and d_n distributed as $D(x_n)$, $n = 1, 2, \ldots$. Intuitively, the "a.s." in (3) refers to the algorithm's and prediction rule's internal randomization.

Theorem 2 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be locally compact metric spaces,* Γ *be a metric space, and* λ *be a continuous and compact-type loss function. There exists a universally consistent randomized on-line prediction algorithm.*

Let $\mathbf{X}$ be a metric space. For any discrete (e.g., finite) subset $\{x_1, x_2, \ldots\}$ of $\mathbf{X}$ and any sequence $\gamma_n \in \mathcal{P}(\Gamma)$ of probability measures on Γ there exists a continuous randomized prediction rule D such that $D(x_n) = \gamma_n$ for all n (indeed, it suffices to set $D(x) := \sum_n \phi_n(x)\gamma_n$, where $\phi_n : \mathbf{X} \to [0,1]$, $n = 1, 2, \ldots$, are continuous functions with disjoint supports such that $\phi_n(x_n) = 1$ for all n). Therefore, there is no shortage of randomized prediction rules.

Continuity, compactness, and the statistical notion of universal consistency

In the statistical setting, where (x_n, y_n) are assumed to be generated independently from the same probability measure, the definition of universal consistency was given by Stone [41] in 1977. One difference of Stone's definition from ours is the lack of the requirement that D should be continuous in his definition.

If the requirement of continuity of D is dropped from our definition, universal consistency becomes impossible to achieve: Reality can easily choose $x_n \to c$, where c is a point of discontinuity of D, and y_n in such a way that Predictor's loss will inevitably be much larger than D's. To be more specific, suppose $\mathbf{X} = \Gamma = \mathbf{Y} = [-1, 1]$ and $\lambda(x, \gamma, y) = |y - \gamma|$ (more generally, the loss is zero when $y = \gamma$ and positive when $y \ne \gamma$). No matter how Predictor chooses his predictions γ_n, Reality can choose

$$x_n := \sum_{i=1}^{n-1} \frac{\operatorname{sign} \gamma_i}{3^i}, \quad y_n := -\operatorname{sign} \gamma_n,$$

where the function sign is defined as

$$\operatorname{sign} \gamma := \begin{cases} 1 & \text{if } \gamma \ge 0 \\ -1 & \text{otherwise,} \end{cases}$$

and thus foil (2) for the prediction rule

$$D(x) := \begin{cases} -1 & \text{if } x < \sum_{i=1}^{\infty} (\operatorname{sign} \gamma_i)/3^i \\ 1 & \text{otherwise.} \end{cases}$$

(Indeed, these definitions imply $D(x_n) = -\operatorname{sign}\gamma_n = y_n$ for all n.)

A positive argument in favor of the requirement of continuity of D is that it is natural for Predictor to compete only with computable prediction rules, and continuity is often regarded as a necessary condition for computability (Brouwer's "continuity principle").

Another difference of Stone's definition is that compactness does not play any special role in it (cf. the antecedent of (2)). It is easy to see that the condition that $\{x_1, x_2, \ldots\}$ and $\{y_1, y_2, \ldots\}$ are precompact is essential in our framework. Indeed, let us suppose, e.g., that $\{x_1, x_2, \ldots\}$ is allowed not to be precompact, continuing to assume that $\mathbf{X}$ is a metric space and also assuming that $\mathbf{Y}$ is a convex subset of a topological vector space. Reality can then choose x_n, $n = 1, 2, \ldots$, as a discrete set in $\mathbf{X}$ ([15], 4.1.17). Let $\phi_n : \mathbf{X} \to [0, 1]$, $n = 1, 2, \ldots$, be continuous functions with disjoint supports such that $\phi_n(x_n) = 1$ for all n. For any sequence of observations $y_1, y_2, \ldots$, the function $D(x) := \sum_n \phi_n(x) y_n$ is a continuous prediction rule such that $D(x_n) = y_n$ for all n. Under such circumstances it is impossible to compete with all continuous prediction rules unless the loss function satisfies some very special properties.

As compared to competitive on-line prediction, the statistical setting is rather restrictive. Compactness and continuity may be said to be satisfied automatically: under mild conditions, every measurable prediction rule can be arbitrarily well approximated by a continuous one (according to Luzin's theorem, [14], 7.5.2, combined with the Tietze–Uryson theorem, [15], 2.1.8), and every probability measure is almost concentrated on a compact set (according to Ulam's theorem, [14], 7.1.4).

3 Defensive forecasting

In this and next sections we will discuss S-prediction and F-prediction, which will prepare way for proving Theorems 1 and 2.

Remark 3 In this paper, S-predictions and F-predictions will always be probability measures, whereas typical D-predictions are not measures. This difference is, however, accidental: e.g., in the problem of on-line regression (as in [48], Section 5) different kinds of predictions are objects of the same nature.

Testing predictions in measure-theoretic probability and neutral measures

S-predictions are empirical statements about the future; they may turn out true or false as the time passes. For such statements to be non-vacuous, we need to have a clear idea of when they become falsified by future observations [33]. In principle, the issuer of S-predictions should agree in advance to a protocol of testing his predictions. It can be said that such a protocol provides an empirical meaning to the predictions.

Testing is, of course, a well-developed area of statistics (see, e.g., [10], Chapter 3). A typical problem is: given a probability measure (the "null hypothesis")

P on a set Ω, which observations $\omega \in \Omega$ falsify P? In the context of this paper, P is an S-prediction, or, as we will often say, a probability forecast for $\omega \in \Omega$. Developing Kolmogorov's ideas (see, e.g., [22], Section 4, [23], and [24]), Martin-Löf (1966, [27]) defines a (in some sense, "the") universal statistical test for a computable P. Levin (1976, [26]) modifies Martin-Löf's definition of statistical test (which was, in essence, the standard statistical definition) and extends it to noncomputable P; Levin's 1976 definition is "uniform", in an important sense.

Levin's test is a function $t : \Omega \times \mathcal{P}(\Omega) \to [0, \infty]$, where $\mathcal{P}(\Omega)$ is the set of all Borel probability measures on Ω, assumed to be a topological space. Levin [26] considers the case $\Omega = \{0, 1\}^\infty$ but notes that his argument works for any other "good" compact space with a countable base. We will assume that Ω is a metric compact (which is equivalent to Levin's assumption that Ω is a compact space with a countable base, [15], 4.2.8), endowing $\mathcal{P}(\Omega)$ with the topology of weak convergence (see below for references). Let us say that a function $t : \Omega \times \mathcal{P}(\Omega) \to [0, \infty]$ is a *test of randomness* if it is lower semicontinuous and, for all $P \in \mathcal{P}(\Omega)$,

$$\int_\Omega t(\omega, P) P(\mathrm{d}\omega) \le 1.$$

The intuition behind this definition is that if we first choose a test t, then observe ω, and then find that $t(\omega, P)$ is very large for the observed ω, we are entitled to reject the hypothesis that ω was generated from P (notice that the P-probability that $t(\omega, P) \ge C$ cannot exceed $1/C$, for any $C > 0$).

The following fundamental result is due to Levin ([26], footnote $^{(1)}$), although our proof is slightly different (for details of Levin's proof, see [17], Section 5).

Lemma 1 (Levin) *Let Ω be a metric compact. For any test of randomness t there exists a probability measure P such that*

$$\forall \omega \in \Omega : \quad t(\omega, P) \le 1. \tag{4}$$

Before proving this result, let us recall some useful facts about the probability measures on the metric compact Ω. The Banach space of all continuous functions on Ω with the usual pointwise addition and scalar action and the sup norm will be denoted $C(\Omega)$. By one of the Riesz representation theorems ([14], 7.4.1; see also 7.1.1), the mapping $\mu \mapsto I_\mu$, where $I_\mu(f) := \int_\Omega f \, \mathrm{d}\mu$, is a linear isometry between the set of all finite Borel measures μ on Ω with the total variation norm and the dual space $C'(\Omega)$ to $C(\Omega)$ with the standard dual norm ([34], Chapter 4). We will identify the finite Borel measures μ on Ω with the corresponding $I_\mu \in C'(\Omega)$. This makes $\mathcal{P}(\Omega)$ a convex closed subset of $C'(\Omega)$.

We will be interested, however, in a different topology on $C'(\Omega)$, the weakest topology for which all evaluation functionals $\mu \in C'(\Omega) \mapsto \mu(f)$, $f \in C(\Omega)$, are continuous. This topology is known as the *weak* topology* ([34], 3.14), and the topology inherited by $\mathcal{P}(\Omega)$ is known as the *topology of weak convergence* ([6], Appendix III). The point mass δ_ω, $\omega \in \Omega$, is defined to be the probability measure concentrated at ω, $\delta_\omega(\{\omega\}) = 1$. The simple example of a sequence of point masses δ_{ω_n} such that $\omega_n \to \omega$ as $n \to \infty$ and $\omega_n \ne \omega$ for all n shows that the topology of weak convergence is different from the dual norm topology: $\delta_{\omega_n} \to \delta_\omega$ holds in one but does not hold in the other.

It is not difficult to check that $\mathcal{P}(\Omega)$ remains a closed subset of $C'(\Omega)$ in the weak* topology ([7], III.2.7, Proposition 7). By the Banach–Alaoglu theorem ([34], 3.15) $\mathcal{P}(\Omega)$ is compact in the topology of weak convergence (this is a special case of Prokhorov's theorem, [6], Appendix III, Theorem 6). In the rest of this paper, $\mathcal{P}(\Omega)$ (and all other spaces of probability measures) are always equipped with the topology of weak convergence.

Since Ω is a metric compact, $\mathcal{P}(\Omega)$ is also metrizable (by the well-known Prokhorov metric: [6], Appendix III, Theorem 6).

Proof of Lemma 1: If t takes value ∞, redefine it as $t := \min(t, 2)$. For all $P, Q \in \mathcal{P}(\Omega)$ set

$$\phi(Q, P) := \int_\Omega t(\omega, P) Q(\mathrm{d}\omega).$$

The function $\phi(Q, P)$ is linear in its first argument, Q, and lower semicontinuous (see Lemma 2 below) in its second argument, P. Ky Fan's minimax theorem (see, e.g., [2], Theorem 11.4; remember that $\mathcal{P}(\Omega)$ is a compact convex subset of $C'(\Omega)$ equipped with the weak* topology) shows that there exists $P^* \in \mathcal{P}(\Omega)$ such that

$$\forall Q \in \mathcal{P}(\Omega): \quad \phi(Q, P^*) \le \sup_{P \in \mathcal{P}(\Omega)} \phi(P, P).$$

Therefore,

$$\forall Q \in \mathcal{P}(\Omega): \quad \int_\Omega t(\omega, P^*) Q(\mathrm{d}\omega) \le 1,$$

and we can see that $t(\omega, P^*)$ never exceeds 1. ∎

This proof used the following topological lemma.

Lemma 2 *Suppose $F : X \times Y \to \mathbb{R}$ is a non-negative lower semicontinuous function defined on the product of two metric compacts, X and Y. If Q is a probability measure on Y, the function $x \in X \mapsto \int_Y F(x, y) Q(\mathrm{d}y)$ is also lower semicontinuous.*

Proof The product $X \times Y$ is also a metric compact ([15], 3.2.4 and 4.2.2). According to Hahn's theorem ([15], Problem 1.7.15(c)), there exists a non-decreasing sequence of (non-negative) continuous functions $F_n(x, y)$ such that $F_n(x, y) \to F(x, y)$ as $n \to \infty$ for all $(x, y) \in X \times Y$. Since each F_n is uniformly continuous ([15], 4.3.32), the functions $\int_Y F_n(x, y) Q(\mathrm{d}y)$ are continuous, and by the monotone convergence theorem ([14], 4.3.2) they converge to $\int_Y F(x, y) Q(\mathrm{d}y)$. Therefore, again by Hahn's theorem, $\int_Y F(x, y) Q(\mathrm{d}y)$ is lower semicontinuous. ∎

Lemma 1 says that for any test of randomness t there is a probability forecast P such that t never detects any disagreement between P and the outcome ω, whatever ω might be.

Gacs ([17], Section 3) defines a uniform test of randomness as a test of randomness that is lower semicomputable (lower semicomputability is an "effective" version of the requirement of lower semicontinuity; this requirement is very natural in the context of randomness: cf. [51], Section 3.1). He proves ([17], Theorem 1) that there exists a *universal* (i.e., largest to within a constant factor) uniform test of randomness. If $t(\omega, P) < \infty$ for a fixed universal test t, ω is said to be *random* with respect to P. Applied to the universal test, Lemma 1 says that there exists a "neutral" probability measure P, such that every ω is random with respect to P.

Gacs ([17], Theorem 7) shows that under his definition there are no neutral measures that are computable even in the weak sense of upper or lower semicomputability even for Ω the compactified set of natural numbers. Levin's original definition of a uniform test of randomness involved some extra conditions, which somewhat mitigate (but not solve completely) the problem of non-computability.

Testing predictions in game-theoretic probability

There is an obvious mismatch between the dynamic prediction protocol of Section 2 and the one-step probability forecasting setting of the previous subsection. If we still want to fit the former into the latter, perhaps we will have to take the infinite sequence of data and observations, $x_1, y_1, x_2, y_2, \ldots$, as ω, and so take $\Omega := (\mathbf{X} \times \mathbf{Y})^\infty$. To find a probability measure satisfying a useful property, such as (4) for an interesting t, might be computationally expensive. Besides, this would force us to assume that the x_ns are also generated from P, and it would be preferable to keep them free of any probabilities (we cannot assume that x_n are given constants since they, e.g., may depend on the previous observations).

A more convenient framework is provided by the game-theoretic foundations of probability. This framework was first thoroughly explored by von Mises [29, 30] (see [37], Chapter 2, for von Mises's precursors), and a serious shortcoming of von Mises's theory was corrected by Ville [44]. After Ville, game-theoretic probability was dormant before being taken up by Kolmogorov [23, 24]. The independence of game-theoretic probability from the standard measure-theoretic probability [21] was emphasized by Dawid (cf. his prequential principle in [11, 13]); see [37] for a review.

There is a special player in the game-theoretic protocols who is responsible for testing the forecasts; following [37], this player will be called Skeptic. This is the protocol that we will be using in this paper:

Testing protocol

```
FOR $n = 1, 2, \ldots$:
    Reality announces $x_n \in \mathbf{X}$.
    Forecaster announces $P_n \in \mathcal{P}(\mathbf{Y})$.
    Skeptic announces $f_n : \mathbf{Y} \to \mathbb{R}$ such that $\int_{\mathbf{Y}} f_n \, \mathrm{d}P_n \le 0$.
    Reality announces $y_n \in \mathbf{Y}$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + f_n(y_n)$.
END FOR.
```

Skeptic's move f_n can be interpreted as taking a long position in a security that pays $f_n(y_n)$ after y_n becomes known; according to Forecaster's beliefs encapsulated in P_n, Skeptic does not have to pay anything for this. We write $\int_{\mathbf{Y}} f_n \, \mathrm{d}P_n \le 0$ to mean that $\int_{\mathbf{Y}} f_n \, \mathrm{d}P_n$ exists and is non-positive. Skeptic starts from some initial capital $\mathcal{K}_0$, which is not specified in the protocol; the evolution of $\mathcal{K}_n$, however, is described.

A game-theoretic procedure of testing Forecaster's performance is a strategy for Skeptic in the testing protocol. If Skeptic starts from $\mathcal{K}_0 := 1$, plays so that he never risks bankruptcy (we say that he risks bankruptcy if his move f_n makes it possible for Reality to choose y_n making $\mathcal{K}_n$ negative), and ends up with a very large value $\mathcal{K}_N$ of his capital, we are entitled to reject the forecasts as false. Informally, the role of Skeptic is to detect disagreement between the forecasts and the actual observations, and the current size of his capital tells us how successful he is at achieving this goal.

Defensive forecasting

Levin's Lemma 1 can be applied to any testing procedure t (test of randomness) to produce forecasts that are ideal as far as that testing procedure is concerned. Such ideal forecasts will be called "defensive forecasts"; in this subsection we will be discussing a similar procedure of defensive forecasting in game-theoretic probability.

Let us now slightly change the testing protocol: suppose that right after Reality's first move in each round Skeptic announces his strategy for the rest of that round.

Defensive forecasting protocol

```
FOR $n = 1, 2, \dots$:
    Reality announces $x_n \in \mathbf{X}$.
    Skeptic announces a lower semicontinuous $F_n : \mathbf{Y} \times \mathcal{P}(\mathbf{Y}) \to \mathbb{R}$
        such that $\int_{\mathbf{Y}} F_n(y, P) P(\mathrm{d}y) \le 0$ for all $P \in \mathcal{P}(\mathbf{Y})$.
    Forecaster announces $P_n \in \mathcal{P}(\mathbf{Y})$.
    Reality announces $y_n \in \mathbf{Y}$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + F_n(y_n, P_n)$.
END FOR.
```

This protocol will be used in the situation where Skeptic has chosen in advance, and told Forecaster about, his testing strategy. However, the game-theoretic analogue of Levin's lemma holds even when Skeptic's strategy is disclosed in a piecemeal manner, as in our protocol.

The following lemma can be proven in the same way as (and is a simple corollary of) Levin's Lemma 1. Its version was first obtained by Akimichi Takemura in 2004 [42].

Lemma 3 (Takemura) *Let $\mathbf{Y}$ be a metric compact. In the defensive forecasting protocol, Forecaster can play in such a way that Skeptic's capital never increases, no matter how he and Reality play.*

Proof For all $P, Q \in \mathcal{P}(\mathbf{Y})$ set

$$\phi(Q, P) := \int_{\mathbf{Y}} F_n(y, P) Q(\mathrm{d}y),$$

where F_n is Skeptic's move in round n. The function $\phi(Q, P)$ is linear in Q and lower semicontinuous in P (the latter also follows from Lemma 2 if we notice that the assumption that F is non-negative can be removed: every lower semicontinuous function on a compact set is bounded below). Ky Fan's minimax theorem shows that there exists P^* such that

$$\phi(Q, P^*) \leq \sup_{P \in \mathcal{P}(\mathbf{Y})} \phi(P, P) \leq 0,$$

and we can see that $F_n(y, P^*)$ is always non-positive. Since the increment $\mathcal{K}_n - \mathcal{K}_{n-1}$ equals $F_n(y_n, P_n)$, it suffices to set $P_n := P^*$. ∎

Testing and laws of probability

There are many interesting ways of testing probability forecasts. In fact, every law of probability provides a way of testing probability forecasts (and vice versa, any way of testing probability forecasts can be regarded as a law of probability). As a simple example, consider the strong law of large numbers in the binary case ($\mathbf{Y} = \{0, 1\}$):

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n) = 0 \tag{5}$$

with probability one, where $p_n := P_n(\{1\})$ is the predicted probability that $y_n = 1$. If (5) is violated, we are justified in rejecting the forecasts p_n; in this sense the strong law of large numbers can serve as a test.

In game-theoretic probability theory, the binary strong law of large numbers is stated as follows: Skeptic has a strategy that, when started with $\mathcal{K}_0 := 1$, never risks bankruptcy and makes Skeptic infinitely rich when (5) is violated. We prove many such game-theoretic laws of probability in [37]; all of them exhibit strategies (continuous or easily made continuous) for Skeptic that make him rich when some property of agreement (such as, apart from various laws of large numbers, the law of the iterated logarithm and the central limit theorem) between the forecasts and the actual observations is violated. When Forecaster plays the strategy of defensive forecasting against such a strategy for Skeptic, the property of agreement is guaranteed to be satisfied, no matter how Reality plays.

In the next section we will apply the procedure of defensive forecasting to a law of large numbers found by Kolmogorov in 1929 ([20]; its simple game-theoretic version can be found in [37], Lemma 6.1 and Proposition 6.1).

4 Calibration and resolution

In this section we will see how the idea of defensive forecasting can be used for producing F-predictions. It is interesting that the pioneering work in this direction by Foster and Vohra [16] was completely independent of Levin's idea. The following is our basic probability forecasting protocol (more basic than the protocols of the previous section).

Probability forecasting protocol

```
FOR $n = 1, 2, \dots$:
    Reality announces $x_n \in \mathbf{X}$.
    Forecaster announces $P_n \in \mathcal{P}(\mathbf{Y})$.
    Reality announces $y_n \in \mathbf{Y}$.
END FOR.
```

Forecaster's prediction P_n is a probability measure on $\mathbf{Y}$ that, intuitively, describes his beliefs about the likely values of y_n. Forecaster's strategy in this protocol will be called a *probability forecasting strategy* (or *algorithm*).

Asymptotic theory of calibration and resolution

The following is a simple asymptotic result about the possibility to ensure "calibration" and "resolution".

Theorem 3 *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are locally compact metric spaces. There is a probability forecasting strategy that guarantees*

$$\bigl(\{x_1, x_2, \dots\} \textit{ and } \{y_1, y_2, \dots\} \textit{ are precompact}\bigr) \\ \Longrightarrow \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) = 0 \quad (6)$$

for all continuous functions $f : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathbb{R}$.

This theorem will be proven at the end of this section, and in the rest of this subsection I will explain the intuition behind (6). The discussion here is an extension of that in [47], Section 6. Let us assume, for simplicity, that $\mathbf{X}$ and $\mathbf{Y}$ are compact metric spaces; as before, δ_y, where $y \in \mathbf{Y}$, stands for the probability measure in $\mathcal{P}(\mathbf{Y})$ concentrated on $\{y\}$.

We start from the intuitive notion of calibration (for further details, see [13] and [16]). The probability forecasts P_n, $n = 1, \dots, N$, are said to be "well calibrated" (or "unbiased in the small", or "reliable", or "valid") if, for any $P^* \in \mathcal{P}(\mathbf{Y})$,

$$\frac{\sum_{n=1,\dots,N : P_n \approx P^*} \delta_{y_n}}{\sum_{n=1,\dots,N : P_n \approx P^*} 1} \approx P^* \quad (7)$$

provided $\sum_{n=1,\ldots,N:P_n\approx P^*} 1$ is not too small. The interpretation of (7) is that the forecasts should be in agreement with the observed frequencies. We can rewrite (7) as

$$\frac{\sum_{n=1,\ldots,N:P_n\approx P^*}(\delta_{y_n} - P_n)}{\sum_{n=1,\ldots,N:P_n\approx P^*} 1} \approx 0.$$

Assuming that $P_n \approx P^*$ for a significant fraction of the $n = 1, \ldots, N$, we can further restate this as the requirement that

$$\frac{1}{N} \sum_{n=1,\ldots,N:P_n\approx P^*} \left(g(y_n) - \int_{\mathbf{Y}} g(y) P_n(\mathrm{d}y) \right) \approx 0 \tag{8}$$

for a wide range of continuous functions g (cf. the definition of the topology of weak convergence in the previous section).

The fact that good calibration is only a necessary condition for good forecasting performance can be seen from the following standard example [13, 16]: if $\mathbf{Y} = \{0, 1\}$ and

$$(y_1, y_2, y_3, y_4, \ldots) = (1, 0, 1, 0, \ldots),$$

the forecasts $P_n(\{0\}) = P_n(\{1\}) = 1/2$, $n = 1, 2, \ldots$, are well calibrated but rather poor; it would be better to forecast with

$$(P_1, P_2, P_3, P_4, \ldots) = (\delta_1, \delta_0, \delta_1, \delta_0, \ldots).$$

Assuming that each datum x_n contains the information about the parity of n (which can always be added to x_n), we can see that the problem with the former forecasting strategy is its lack of resolution: it does not distinguish between the data with odd and even n. In general, we would like each forecast P_n to be as specific as possible to the current datum x_n; the resolution of a probability forecasting algorithm is the degree to which it achieves this goal (taking it for granted that x_n contains all relevant information).

Analogously to (8), the forecasts P_n, $n = 1, \ldots, N$, may be said to have good resolution if, for any $x^* \in \mathbf{X}$,

$$\frac{1}{N} \sum_{n=1,\ldots,N:x_n\approx x^*} \left(g(y_n) - \int_{\mathbf{Y}} g(y) P_n(\mathrm{d}y) \right) \approx 0 \tag{9}$$

for a wide range of continuous g. We can also require that the forecasts P_n, $n = 1, \ldots, N$, should have good "calibration-cum-resolution": for any $(x^*, P^*) \in \mathbf{X} \times \mathcal{P}(\mathbf{Y})$,

$$\frac{1}{N} \sum_{n=1,\ldots,N:(x_n,P_n)\approx(x^*,P^*)} \left(g(y_n) - \int_{\mathbf{Y}} g(y) P_n(\mathrm{d}y) \right) \approx 0 \tag{10}$$

for a wide range of continuous g. Notice that even if forecasts have both good calibration and good resolution, they can still have poor calibration-cum-resolution.

To make sense of the $\approx$ in, say, (8), we can replace each "crisp" point $P^* \in \mathcal{P}(\mathbf{Y})$ by a "fuzzy point" $I_{P^*} : \mathcal{P}(\mathbf{Y}) \to [0, 1]$; I_{P^*} is required to be continuous, and we might also want to have $I_{P^*}(P^*)=1$ and $I_{P^*}(P)=0$ for all P outside a small neighborhood of P^*. (The alternative of choosing $I_{P^*}:=\mathbb{I}_A$, where A is a small neighborhood of P^* and $\mathbb{I}_A$ is its indicator function, does not work because of Oakes's and Dawid's examples [32, 12]; I_{P^*} can, however, be arbitrarily close to $\mathbb{I}_A$.) This transforms (8) into

$$\frac{1}{N}\sum_{n=1}^{N} I_{P^*}(P_n)\left(g(y_n)-\int_{\mathbf{Y}} g(y)P_n(\mathrm{d}y)\right)\approx 0,$$

which is equivalent to

$$\frac{1}{N}\sum_{n=1}^{N}\left(f(P_n,y_n)-\int_{\mathbf{Y}} f(P_n,y)P_n(\mathrm{d}y)\right)\approx 0, \tag{11}$$

where $f(P,y):=I_{P^*}(P)g(y)$. It is natural to require that (11) should hold for a wide range of continuous functions $f(P,y)$, not necessarily of the form $I_{P^*}(P)g(y)$.

In the same way we can transform (9) into

$$\frac{1}{N}\sum_{n=1}^{N}\left(f(x_n,y_n)-\int_{\mathbf{Y}} f(x_n,y)P_n(\mathrm{d}y)\right)\approx 0$$

and (10) into

$$\frac{1}{N}\sum_{n=1}^{N}\left(f(x_n,P_n,y_n)-\int_{\mathbf{Y}} f(x_n,P_n,y)P_n(\mathrm{d}y)\right)\approx 0.$$

We can see that the consequent of (6) can be interpreted as the forecasts having good calibration-cum-resolution; the case where $f(x,P,y)$ depends only on P and y corresponds to good calibration, and the case where $f(x,P,y)$ depends only on x and y corresponds to good resolution.

Calibration-cum-resolution bounds

A more explicit result about calibration and resolution is given in terms of "reproducing kernel Hilbert spaces". Let $\mathcal{F}$ be a Hilbert space of functions on a set Ω (with the pointwise operations of addition and scalar action). Its *imbedding constant* $\mathbf{c}_{\mathcal{F}}$ is defined by

$$\mathbf{c}_{\mathcal{F}}:=\sup_{\omega\in\Omega}\sup_{f\in\mathcal{F}:\|f\|_{\mathcal{F}}\le 1} f(\omega). \tag{12}$$

We will be interested in the case $\mathbf{c}_{\mathcal{F}}<\infty$ and will refer to $\mathcal{F}$ satisfying this condition as *reproducing kernel Hilbert spaces (RKHS) with finite imbedding constant.*

The Hilbert space $\mathcal{F}$ is called a *reproducing kernel Hilbert space* (RKHS) if all evaluation functionals $f\in\mathcal{F}\mapsto f(\omega)$, $\omega\in\Omega$, are bounded; the class of

RKHS with finite imbedding constant is a subclass of the class of RKHS. Let $\mathcal{F}$ be an RKHS on Ω. By the Riesz–Fischer theorem, for each $\omega \in \Omega$ there exists a function $\mathbf{k}_\omega \in \mathcal{F}$ (the *representer* of ω in $\mathcal{F}$) such that

$$f(\omega) = \langle \mathbf{k}_\omega, f \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}. \tag{13}$$

If Ω is a topological space and the mapping $\omega \mapsto \mathbf{k}_\omega$ is continuous, $\mathcal{F}$ is called a *continuous RKHS*. If $\Omega = \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$ and $\mathbf{k}_\omega = \mathbf{k}_{x,P,y}$ is a continuous function of $(P, y) \in \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$ for each $x \in \mathbf{X}$, we will say that $\mathcal{F}$ is *forecast-continuous.*

Theorem 4 *Let $\mathbf{Y}$ be a metric compact and $\mathcal{F}$ be a forecast-continuous RKHS on $\mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$ with finite imbedding constant $\mathbf{c}_{\mathcal{F}}$. There is a probability forecasting strategy that guarantees*

$$\left| \sum_{n=1}^{N} \left(f(x_n, P_n, y_n) - \int_{\mathbf{Y}} f(x_n, P_n, y)\, P_n(\mathrm{d}y) \right) \right| \le 2\mathbf{c}_{\mathcal{F}} \left\| f \right\|_{\mathcal{F}} \sqrt{N}$$

for all N and all $f \in \mathcal{F}$.

Before proving Theorem 4 we will give an example of a convenient RKHS $\mathcal{F}$ that can be used in its applications. Let us consider a finite $\mathbf{Y}$, represent $\mathcal{P}(\mathbf{Y})$ as a simplex in a Euclidean space, and suppose that $\mathbf{X}$ is a bounded open subset of a Euclidean space. The interior $\operatorname{Int}\mathcal{P}(\mathbf{Y})$ of $\mathcal{P}(\mathbf{Y})$ can be regarded as a bounded open subset of a Euclidean space, and so the product $\mathbf{X} \times \operatorname{Int}\mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$ can also be regarded as a bounded open set Ω in a Euclidean space of dimension $K := \dim \mathbf{X} + |\mathbf{Y}| - 1$: namely, as a disjoint union of $|\mathbf{Y}|$ copies of the bounded open set $\mathbf{X} \times \operatorname{Int}\mathcal{P}(\mathbf{Y})$.

For a smooth function $u : \Omega \to \mathbb{R}$ and $m \in \{0, 1, \ldots\}$ define

$$\left\| u \right\|_m := \sqrt{\sum_{0 \le |\alpha| \le m} \int_{\Omega} (D^\alpha u)^2}, \tag{14}$$

where $\int_\Omega$ stands for the integral with respect to the Lebesgue measure on Ω, α runs over the multi-indices $\alpha = (\alpha_1, \ldots, \alpha_K) \in \{0, 1, \ldots\}^K$, and

$$|\alpha| := \alpha_1 + \cdots + \alpha_K, \quad D^\alpha u := \frac{\partial^{|\alpha|} u}{\partial_{t_1}^{\alpha_1} \cdots \partial_{t_K}^{\alpha_K}}$$

($(t_1, \ldots, t_K)$ is a typical point of the Euclidean space containing Ω). Let $H^m(\Omega)$ be the completion of the set of smooth function on Ω with respect to the norm (14). According to the Sobolev imbedding theorem ([1], Theorem 4.12), $H^m(\Omega)$ can be identified with an RKHS of continuous functions on the closure $\overline{\Omega}$ of Ω with a finite imbedding constant. This conclusion depends on the assumption $m > K/2$, which we will always be making.

It is clear that every continuous function f on $\overline{\Omega}$ can be approximated, arbitrarily closely, by a function from $H^m(\Omega)$: even the functions in $C^\infty(\mathbb{R}^K)$, all of which belong to all Sobolev spaces on Ω, are dense in $C(\overline{\Omega})$ ([1], 2.29).

There is little doubt that Sobolev spaces $H^m(\Omega)$ are continuous under our assumption $m > K/2$ and for "nice" Ω, although I am not aware of any general results in this direction.

Proof of Theorem 4

If $f : \Omega \to \mathcal{H}$ is a function taking values in a topological vector space $\mathcal{H}$ and P is a finite measure on its domain Ω, the integral $\int_\Omega f \,\mathrm{d}P$ will be understood in Pettis's ([34], Definition 3.26) sense. Namely, the integral $\int_\Omega f \,\mathrm{d}P$ is defined to be $h \in \mathcal{H}$ such that

$$\Lambda h = \int_\Omega (\Lambda f) \,\mathrm{d}P \tag{15}$$

for all $\Lambda \in \mathcal{H}^*$. The existence and uniqueness of the Pettis integral is assured if Ω is a compact topological space (with P defined on its Borel σ-algebra), $\mathcal{H}$ is a Banach space, and f is continuous ([34], Theorems 3.27, 3.20, and 3.3).

Remark 4 Another popular notion of the integral for vector-valued functions is Bochner's (see, e.g., [53]), which is more restrictive than Pettis's (in particular, the Bochner integral always satisfies (15)). Interestingly, the Bochner integral $\int_\Omega f \,\mathrm{d}P$ exists for all measurable functions $f : \Omega \to \mathcal{H}$ (with Ω a measurable space) provided $\mathcal{H}$ is a separable Banach space and $\int_\Omega \|f\|_{\mathcal{H}} \,\mathrm{d}P < \infty$ (this follows from Bochner's theorem, [53], Theorem 1 in Section V.5, and Pettis's measurability theorem, [53], the theorem in Section V.4). No topological conditions are imposed on Ω or f, but there is the requirement of separability (which is essential, again by Bochner's theorem and Pettis's measurability theorem). This requirement, however, may be said to be satisfied automatically under the given sufficient conditions for the existence of the Pettis integral: since $f(\Omega)$ is a compact metric space, it is separable ([15], 4.1.18), and we can redefine $\mathcal{H}$ as the smallest closed linear subspace containing $f(\Omega)$. Therefore, we can use all properties of the Bochner integral under those conditions.

We start from a corollary (a version of Kolmogorov's 1929 result) of Lemma 3.

Lemma 4 *Suppose* $\mathbf{Y}$ *is a metric compact. Let* $\Phi_n : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathcal{H}$, $n = 1, 2, \ldots$, *be functions taking values in a Hilbert space* $\mathcal{H}$ *such that, for all* n *and* x, $\Phi_n(x, P, y)$ *is a continuous function of* $(P, y) \in \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$. *There is a probability forecasting strategy that guarantees*

$$\left\| \sum_{n=1}^{N} \Psi_n (x_n, P_n, y_n) \right\|_{\mathcal{H}}^2 \le \sum_{n=1}^{N} \| \Psi_n (x_n, P_n, y_n) \|_{\mathcal{H}}^2 \tag{16}$$

for all N, *where*

$$\Psi_n (x, P, y) := \Phi_n (x, P, y) - \int_{\mathbf{Y}} \Phi_n (x, P, y) \, P(\mathrm{d}y).$$

Proof According to Lemma 3, it suffices to check that

$$S_N := \left\| \sum_{n=1}^{N} \Psi_n \left(x_n, P_n, y_n\right) \right\|_{\mathcal{H}}^2 - \sum_{n=1}^{N} \left\| \Psi_n \left(x_n, P_n, y_n\right) \right\|_{\mathcal{H}}^2 \tag{17}$$

is the capital process of some strategy for Skeptic in the defensive forecasting protocol. Since

$$\begin{aligned} S_N - S_{N-1} &= \left\| \sum_{n=1}^{N-1} \Psi_n \left(x_n, P_n, y_n\right) + \Psi_N \left(x_N, P_N, y_N\right) \right\|_{\mathcal{H}}^2 \\ &\quad - \left\| \sum_{n=1}^{N-1} \Psi_n \left(x_n, P_n, y_n\right) \right\|_{\mathcal{H}}^2 - \left\| \Psi_N \left(x_N, P_N, y_N\right) \right\|_{\mathcal{H}}^2 \\ &= \left\langle 2 \sum_{n=1}^{N-1} \Psi_n \left(x_n, P_n, y_n\right), \Psi_N \left(x_N, P_N, y_N\right) \right\rangle_{\mathcal{H}} \\ &= \left\langle A, \Psi_N \left(x_N, P_N, y_N\right) \right\rangle_{\mathcal{H}}, \end{aligned}$$

where we have introduced the notation A for the element $2 \sum_{n=1}^{N-1} \Psi_n \left(x_n, P_n, y_n\right)$ of $\mathcal{H}$ known at the beginning of the Nth round, and, by the definition of the Pettis integral,

$$\int_{\mathbf{Y}} \left\langle A, \Psi_N \left(x_N, P_N, y\right) \right\rangle_{\mathcal{H}} P_N(\mathrm{d}y) = \left\langle A, \int_{\mathbf{Y}} \Psi_N \left(x_N, P_N, y\right) P_N(\mathrm{d}y) \right\rangle_{\mathcal{H}} = 0, \tag{18}$$

the difference $S_N - S_{N-1}$ coincides with Skeptic's gain in the Nth round of the testing protocol when he makes the valid move $f_N(y) := \left\langle A, \Psi_N \left(x_N, P_N, y\right) \right\rangle_{\mathcal{H}}$. It remains to check that $F_N(y, P) := \left\langle A, \Psi_N \left(x_N, P, y\right) \right\rangle_{\mathcal{H}}$ will be a valid move in the defensive forecasting protocol, i.e., that the function F_N is lower semicontinuous; we will see that it is in fact continuous. By Lemma 5 below, the function $\int_{\mathbf{Y}} \Phi_N(x, P, y) P(\mathrm{d}y)$ is continuous in P; therefore, the function Ψ_N is continuous in (P, y). This implies that $\left\langle A, \Psi_N \left(x_N, P, y\right) \right\rangle_{\mathcal{H}}$ is a continuous function of (P, y). ▮

The proof of Lemma 4 used the following lemma.

Lemma 5 *Suppose* $\mathbf{Y}$ *is a metric compact and* $\Phi : \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathcal{H}$ *is a continuous mapping into a Hilbert space* $\mathcal{H}$. *The mapping* $P \in \mathcal{P}(\mathbf{Y}) \mapsto \int_{\mathbf{Y}} \Phi(P, y) P(\mathrm{d}y)$ *is also continuous.*

Proof Let $P_n \to P$ as $n \to \infty$; our goal is to prove that $\int_{\mathbf{Y}} \Phi(P_n, y) P_n(\mathrm{d}y) \to \int_{\mathbf{Y}} \Phi(P, y) P(\mathrm{d}y)$. We have:

$$\left\| \int_{\mathbf{Y}} \Phi(P_n, y) P_n(\mathrm{d}y) - \int_{\mathbf{Y}} \Phi(P, y) P(\mathrm{d}y) \right\|_{\mathcal{H}}$$

$$\begin{aligned}\leq \left\| \int_{\mathbf{Y}} \Phi(P_n, y) P_n(\mathrm{d}y) - \int_{\mathbf{Y}} \Phi(P, y) P_n(\mathrm{d}y) \right\|_{\mathcal{H}} \\ + \left\| \int_{\mathbf{Y}} \Phi(P, y) P_n(\mathrm{d}y) - \int_{\mathbf{Y}} \Phi(P, y) P(\mathrm{d}y) \right\|_{\mathcal{H}}. \quad (19)\end{aligned}$$

The first addend on the right-hand side can be bounded above by

$$\int_{\mathbf{Y}} \left\| \Phi(P_n, y) - \Phi(P, y) \right\|_{\mathcal{H}} P_n(\mathrm{d}y)$$

([34], 3.29), and the last expression tends to zero since Φ is uniformly continuous ([15], 4.3.32). The second addend on the right-hand side of (19) tends to zero by the continuity of the mapping $Q \in \mathcal{P}(\mathbf{Y}) \mapsto \int_{\mathbf{Y}} f(y) Q(\mathrm{d}y)$ for a continuous f ([7], III.4.2, Proposition 6). ∎

The following variation on Lemma 5 will be needed later.

Lemma 6 *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are metric compacts and* $\Phi : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathcal{H}$ *is a continuous mapping into a Hilbert space* $\mathcal{H}$. *The mapping* $(x, P) \in \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \mapsto \int_{\mathbf{Y}} \Phi(x, P, y) P(\mathrm{d}y)$ *is also continuous.*

Proof Let $x_n \to x$ and $P_n \to P$ as $n \to \infty$. To prove $\int_{\mathbf{Y}} \Phi(x_n, P_n, y) P_n(\mathrm{d}y) \to \int_{\mathbf{Y}} \Phi(x, P, y) P(\mathrm{d}y)$ we can use a similar argument to that in the previous lemma applied to

$$\begin{aligned}&\left\| \int_{\mathbf{Y}} \Phi(x_n, P_n, y) P_n(\mathrm{d}y) - \int_{\mathbf{Y}} \Phi(x, P, y) P(\mathrm{d}y) \right\|_{\mathcal{H}} \\ &\quad \leq \left\| \int_{\mathbf{Y}} \Phi(x_n, P_n, y) P_n(\mathrm{d}y) - \int_{\mathbf{Y}} \Phi(x, P, y) P_n(\mathrm{d}y) \right\|_{\mathcal{H}} \\ &\quad\quad + \left\| \int_{\mathbf{Y}} \Phi(x, P, y) P_n(\mathrm{d}y) - \int_{\mathbf{Y}} \Phi(x, P, y) P(\mathrm{d}y) \right\|_{\mathcal{H}}. \quad ∎\end{aligned}$$

Now we can begin the actual proof of Theorem 4. Take as $\Phi(x, P, y)$ the representer $\mathbf{k}_{x,P,y}$ of the evaluation functional $f \in \mathcal{F} \mapsto f(x, P, y)$:

$$\langle f, \mathbf{k}_{x,P,y} \rangle_{\mathcal{F}} = f(x, P, y), \quad \forall (x, P, y) \in \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}, f \in \mathcal{F}.$$

Set

$$\mathbf{k}_{x,P} := \int_{\mathbf{Y}} \mathbf{k}_{x,P,y} P(\mathrm{d}y);$$

the function $\mathbf{k}_{x,P}$ is continuous in P by Lemma 5.

Theorem 4 will easily follow from the following lemma, which itself is an easy implication of Lemma 4.

Lemma 7 *Let* $\mathbf{Y}$ *be a metric compact and* $\mathcal{F}$ *be a forecast-continuous RKHS on* $\mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$. *There is a probability forecasting strategy that guarantees*

$$\left|\sum_{n=1}^{N}\left(f\left(x_n,P_n,y_n\right)-\int_{\mathbf{Y}}f\left(x_n,P_n,y\right)P_n(\mathrm{d}y)\right)\right| \\ \le \|f\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}\left\|\mathbf{k}_{x_n,P_n,y_n}-\mathbf{k}_{x_n,P_n}\right\|_{\mathcal{F}}^{2}}$$

for all N and all $f \in \mathcal{F}$.

Proof Using Lemma 4 (with all Ψ_n equal, $\Psi_n(x,P,y) := \mathbf{k}_{x,P,y}-\mathbf{k}_{x,P}$), we obtain:

$$\left|\sum_{n=1}^{N}\left(f\left(x_n,P_n,y_n\right)-\int_{\mathbf{Y}}f\left(x_n,P_n,y\right)P_n(\mathrm{d}y)\right)\right| \\ =\left|\sum_{n=1}^{N}\left(\langle f,\mathbf{k}_{x_n,P_n,y_n}\rangle_{\mathcal{F}}-\int_{\mathbf{Y}}\langle f,\mathbf{k}_{x_n,P_n,y}\rangle_{\mathcal{F}}P_n(\mathrm{d}y)\right)\right| \\ =\left|\left\langle f,\sum_{n=1}^{N}\left(\mathbf{k}_{x_n,P_n,y_n}-\mathbf{k}_{x_n,P_n}\right)\right\rangle_{\mathcal{F}}\right| \le \|f\|_{\mathcal{F}}\left\|\sum_{n=1}^{N}\left(\mathbf{k}_{x_n,P_n,y_n}-\mathbf{k}_{x_n,P_n}\right)\right\|_{\mathcal{F}} \\ \le \|f\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}\left\|\mathbf{k}_{x_n,P_n,y_n}-\mathbf{k}_{x_n,P_n}\right\|_{\mathcal{F}}^{2}}. \quad \blacksquare$$

Remark 5 The algorithm of Lemma 7 is a generalization of the K29 algorithm of [50]. It would be interesting also to analyze the K29* algorithm (called the algorithm of large numbers in [47] and [46]).

To deduce Theorem 4 from Lemma 7, notice that $\|\mathbf{k}_{x,P,y}\|_{\mathcal{F}} \le \mathbf{c}_{\mathcal{F}}$ (by Lemma 8 below), $\|\mathbf{k}_{x,P}\|_{\mathcal{F}} \le \int_{\mathbf{Y}}\|\mathbf{k}_{x,P,y}\|_{\mathcal{F}}P(\mathrm{d}y) \le \mathbf{c}_{\mathcal{F}}$, and, therefore,

$$\sum_{n=1}^{N}\left\|\mathbf{k}_{x_n,P_n,y_n}-\mathbf{k}_{x_n,P_n}\right\|_{\mathcal{F}}^{2} \le 4\mathbf{c}_{\mathcal{F}}^{2}N.$$

This completes the proof apart from Lemma 8.

Let $\mathcal{F}$ be an RKHS on Ω. The norm of the evaluation functional $f \in \mathcal{F} \mapsto f(\omega)$ will be denoted by $\mathbf{c}_{\mathcal{F}}(\omega)$. It is clear that $\mathcal{F}$ is an RKHS with finite imbedding constant if and only if

$$\mathbf{c}_{\mathcal{F}} := \sup_{\omega\in\Omega}\mathbf{c}_{\mathcal{F}}(\omega) \tag{20}$$

is finite; the constants in (20) and (12) coincide. The next lemma, concluding the proof of Theorem 4, asserts that the norm $\|\mathbf{k}_{\omega}\|_{\mathcal{F}}$ of the representer of ω in $\mathcal{F}$ coincides with the norm $\mathbf{c}_{\mathcal{F}}(\omega)$ of the evaluation functional $f \mapsto f(\omega)$.

Lemma 8 *Let $\mathcal{F}$ be an RKHS on Ω. For each $\omega \in \Omega$,*

$$\|\mathbf{k}_{\omega}\|_{\mathcal{F}} = \mathbf{c}_{\mathcal{F}}(\omega). \tag{21}$$

Proof Fix $\omega \in \Omega$. We are required to prove

$$\sup_{f:\|f\|_{\mathcal{F}} \le 1} |f(\omega)| = \|\mathbf{k}_\omega\|_{\mathcal{F}} .$$

The inequality $\le$ follows from

$$|f(\omega)| = |\langle f, \mathbf{k}_\omega \rangle_{\mathcal{F}}| \le \|f\|_{\mathcal{F}} \|\mathbf{k}_\omega\|_{\mathcal{F}} \le \|\mathbf{k}_\omega\|_{\mathcal{F}} ,$$

where $\|f\|_{\mathcal{F}} \le 1$. The inequality $\ge$ follows from

$$|f(\omega)| = \frac{\mathbf{k}_\omega(\omega)}{\|\mathbf{k}_\omega\|_{\mathcal{F}}} = \frac{\langle \mathbf{k}_\omega, \mathbf{k}_\omega \rangle_{\mathcal{F}}}{\|\mathbf{k}_\omega\|_{\mathcal{F}}} = \|\mathbf{k}_\omega\|_{\mathcal{F}} ,$$

where $f := \mathbf{k}_\omega / \|\mathbf{k}_\omega\|_{\mathcal{F}}$ and $\|\mathbf{k}_\omega\|_{\mathcal{F}}$ is assumed to be non-zero (if it is zero, $\mathbf{k}_\omega = 0$, which implies $\mathbf{c}_{\mathcal{F}}(\omega) = 0$, and (21) still holds). ∎

Reproducing kernels

In this subsection we start preparations for proving Theorem 3. But first we need to delve slightly deeper into the theory of RKHS. An equivalent language for talking about RKHS is provided by the notion of a reproducing kernel, and this subsection defines reproducing kernels and summarizes some of their properties. For a detailed discussion, see, e.g., [3, 4] or [28].

The *reproducing kernel* of an RKHS $\mathcal{F}$ on Ω is the function $\mathbf{k} : \Omega^2 \to \mathbb{R}$ defined by

$$\mathbf{k}(\omega, \omega') := \langle \mathbf{k}_\omega, \mathbf{k}_{\omega'} \rangle_{\mathcal{F}}$$

(equivalently, we could define $\mathbf{k}(\omega, \omega')$ as $\mathbf{k}_\omega(\omega')$ or as $\mathbf{k}_{\omega'}(\omega)$). The origin of this name is the "reproducing property" (13).

There is a simple internal characterization of reproducing kernels of RKHS. First, it is easy to check that the function $\mathbf{k}(\omega, \omega')$, as we defined it, is symmetric,

$$\mathbf{k}(\omega, \omega') = \mathbf{k}(\omega', \omega), \quad \forall (\omega, \omega') \in \Omega^2,$$

and positive definite,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} t_i t_j \mathbf{k}(\omega_i, \omega_j) \ge 0, \qquad \forall m = 1, 2, \dots, (t_1, \dots, t_m) \in \mathbb{R}^m, (\omega_1, \dots, \omega_m) \in \Omega^m.$$

On the other hand, for every symmetric and positive definite $\mathbf{k} : \Omega^2 \to \mathbb{R}$ there exists a unique RKHS $\mathcal{F}$ on Ω such that $\mathbf{k}$ is the reproducing kernel of $\mathcal{F}$ ([3], Theorem 2 on p. 143).

We can see that the notions of a reproducing kernel of RKHS and of a symmetric positive definite function on Ω^2 have the same content, and we will sometimes say "kernel on Ω" to mean a symmetric positive definite function on Ω^2. Kernels in this sense are the main source of RKHS in learning theory:

cf. [43, 35, 38]. Every kernel on $\mathbf{X}$ is a valid parameter for our prediction algorithms. In general, it is convenient to use RKHS in stating mathematical properties of prediction algorithms, but the algorithms themselves typically use the more constructive representation of RKHS via their reproducing kernels.

It is easy to see that $\mathcal{F}$ is a continuous RKHS if and only if its reproducing kernel is continuous (see [40] or [47], Appendix B of the arXiv technical report). A convenient equivalent definition of $\mathbf{c}_{\mathcal{F}}$ is

$$\mathbf{c}_{\mathcal{F}} = \mathbf{c}_{\mathbf{k}} := \sup_{\omega \in \Omega} \sqrt{\mathbf{k}(\omega, \omega)} = \sup_{\omega, \omega' \in \Omega} \sqrt{|\mathbf{k}(\omega, \omega')|}, \tag{22}$$

$\mathbf{k}$ being the reproducing kernel of an RKHS $\mathcal{F}$ on Ω.

Let us say that a family $\mathcal{F}$ of functions $f : \Omega \to \mathbb{R}$ is *universal* if Ω is a topological space and for every compact subset A of Ω every continuous function on A can be arbitrarily well approximated in the metric $C(A)$ by functions in $\mathcal{F}$ (in the case of compact Ω this coincides with the definition given in [40] as Definition 4).

We have already noticed the obvious fact that the Sobolev spaces $H^m(\Omega)$ on bounded open $\Omega \subseteq \mathbb{R}^K$, $K < 2m$, are universal. There is a price to pay for the obviousness of this fact: the reproducing kernels of the Sobolev spaces are known only in some special cases (see, e.g., [5], Section 7.4). This complicates checking their continuity.

On the other hand, some very simple continuous reproducing kernels, such as the Gaussian kernel

$$\mathbf{k}(\omega, \omega') := \exp\left(-\frac{\|\omega - \omega'\|^2}{\sigma^2}\right)$$

($\|\cdot\|$ being the Euclidean norm and σ being an arbitrary positive constant) on the Euclidean space $\mathbb{R}^K$ and the infinite polynomial kernel

$$\mathbf{k}(\omega, \omega') := \frac{1}{1 - \langle \omega, \omega' \rangle}$$

($\langle \cdot, \cdot \rangle$ being the Euclidean inner product) on the Euclidean ball $\{\omega \in \mathbb{R}^K \mid \|\omega\| < 1\}$, are universal ([40], Examples 1 and 2). Their universality is not difficult to prove but not obvious (and even somewhat counterintuitive in the case of the Gaussian kernel: *a priori* one might expect that only smooth functions that are almost linear at scales smaller than σ can belong to the corresponding RKHS). On the other hand, their continuity is obvious.

Universal function space on the Hilbert cube

Remember that the *Hilbert cube* is the topological space $[0,1]^\infty$ ([15], 2.3.22), i.e., the topological product of a countable number of closed intervals $[0,1]$. As the next step in the proof of Theorem 3, in this subsection we construct a universal RKHS on the Hilbert cube with finite imbedding constant; the idea of the construction is to "mix" Sobolev spaces on $[0,1]^K$ for $K = 1, 2, \ldots$ (or the spaces mentioned at the end of the previous subsection, for which both continuity and universality are proven).

Let $\mathcal{F}_K$, $K = 1, 2, \ldots$, be the set of all functions f on the Hilbert cube such that $f(t_1, t_2, \ldots)$ depends only on $t_1, \ldots, t_K$ and whose norm (14) (with $\Omega := [0,1]^K$) is finite for $m := K$. Equipping $\mathcal{F}_K$ with this norm we obtain an RKHS with finite imbedding constant. Let c_K be the imbedding constant of $\mathcal{F}_K$. It will be convenient to modify each $\mathcal{F}_K$ by scaling the inner product:

$$\langle \cdot, \cdot \rangle_{\mathcal{F}'_K} := c_K^2 2^K \langle \cdot, \cdot \rangle_{\mathcal{F}_K} ;$$

the scaled $\mathcal{F}_K$ will be denoted $\mathcal{F}'_K$. By (13), the representer $\mathbf{k}'_\omega$ of ω in $\mathcal{F}'_K$ can be expressed as $\mathbf{k}'_\omega = c_K^{-2} 2^{-K} \mathbf{k}_\omega$ via the representer $\mathbf{k}_\omega$ of ω in $\mathcal{F}_K$. Therefore, the imbedding constant of $\mathcal{F}'_K$ is $2^{-K/2}$, and it is obvious that $\mathcal{F}'_K$ inherits from $\mathcal{F}_K$ the property of being a universal RKHS for functions that only depend on $t_1, \ldots, t_K$.

For the reproducing kernel $\mathbf{k}'_K(\omega, \omega')$ of $\mathcal{F}'_K$ we have

$$|\mathbf{k}'_K(\omega, \omega')| = \left| \langle \mathbf{k}'_\omega, \mathbf{k}'_{\omega'} \rangle_{\mathcal{F}'_K} \right| \le \|\mathbf{k}'_\omega\|_{\mathcal{F}'_K} \|\mathbf{k}'_{\omega'}\|_{\mathcal{F}'_K} \le 2^{-K/2} 2^{-K/2} = 2^{-K},$$

where $\mathbf{k}'_\omega$ and $\mathbf{k}'_{\omega'}$ stand for the representers in $\mathcal{F}'_K$. Define an RKHS $\mathcal{G}_K$ as the set of all functions $f : [0,1]^\infty \to \mathbb{R}$ that can be decomposed into a sum $f = f_1 + \cdots + f_K$, where $f_k \in \mathcal{F}'_k$, $k = 1, \ldots, K$. The norm of f is defined as the infimum

$$\|f\|_{\mathcal{G}_K} := \inf \sqrt{\sum_{k=1}^K \|f_k\|^2_{\mathcal{F}'_k}}$$

over all such decompositions. According to the theorem on p. 353 of [4], $\mathcal{G}_K$ is an RKHS whose reproducing kernel $\mathbf{k}_K$ satisfies

$$\mathbf{k}_K(\omega, \omega') = \sum_{k=1}^K \mathbf{k}'_k(\omega, \omega') \in \left[-1 + 2^{-K}, 1 - 2^{-K}\right].$$

The limiting RKHS of $\mathcal{G}_K$, $K \to \infty$, is defined in [4], Section I.9 (Case B), in two steps. Let $\mathcal{F}_0$ consist of the functions in $\mathcal{G}_K$, $K = 1, 2, \ldots$; the $\mathcal{F}_0$-norm of a function $g \in \mathcal{G}_K$ is defined as

$$\|g\|_{\mathcal{F}_0} := \inf_{k \ge K} \|g\|_{\mathcal{G}_k}.$$

In general, the space $\mathcal{F}_0$ is not complete. Therefore, a larger space $\mathcal{F}_0^*$ is defined: $f \in \mathcal{F}_0^*$ if there is a Cauchy sequence f_n in $\mathcal{F}_0$ such that

$$\forall \omega \in [0,1]^\infty : f(\omega) = \lim_{n \to \infty} f_n(\omega); \tag{23}$$

the norm of such an f is defined as

$$\|f\|_{\mathcal{F}_0^*} := \inf \lim_{n \to \infty} \|f_n\|_{\mathcal{F}_0},$$

where the infimum is taken over all Cauchy sequences satisfying (23). By Theorem II on p. 367 of [4], $\mathcal{F}_0^*$ is an RKHS with reproducing kernel

$$\mathbf{k}^*(\omega,\omega') = \sum_{k=1}^{\infty} \mathbf{k}'_k(\omega,\omega') \in [-1,1]; \tag{24}$$

therefore, its imbedding constant is finite (at most 1: see (22)).

Lemma 9 *The RKHS $\mathcal{F}_0^*$ on the Hilbert cube is universal and continuous.*

Proof The Hilbert cube is a topological space that is both compact (by Tikhonov's theorem, [15], 3.2.4) and metrizable; for concreteness, let us fix the metric

$$\rho\left((t_1,t_2,\ldots),(t'_1,t'_2,\ldots)\right) := \sum_{k=1}^{\infty} 2^{-k}\left|t_k - t'_k\right|.$$

Let f be a continuous function on the Hilbert cube. Since every continuous function on a compact metric space is uniformly continuous ([15], 4.3.32), the function

$$g\left(t_1,t_2,\ldots\right) := f\left(t_1,\ldots,t_K,0,0,\ldots\right)$$

can be made arbitrarily close to f, in metric $C([0,1]^\infty)$, by making K sufficiently large. It remains to notice that g can be arbitrarily closely approximated by a function in $\mathcal{F}_K$ and that every function in $\mathcal{F}_K$ belongs to $\mathcal{F}_0^*$.

The continuity of $\mathcal{F}_0^*$ follows from the Weierstrass M-test and the expression (24) of its reproducing kernel via the reproducing kernels of the spaces $\mathcal{F}'_K$, $K = 1,2,\ldots$, with imbedding constant 2^{-K}. ∎

Corollary 1 *For any compact metric space Ω there is a continuous universal RKHS $\mathcal{F}$ on Ω with finite imbedding constant.*

Proof It is known ([15], 4.2.10) that every compact metric space can be homeomorphically imbedded into the Hilbert cube; let $F : \Omega \to [0,1]^\infty$ be such an imbedding. The image $F(\Omega)$ is a compact subset of the Hilbert cube ([15], 3.1.10). Let $\mathcal{F}$ be the class of all functions $f : \Omega \to \mathbb{R}$ such that $f(F^{-1}) : F(\Omega) \to \mathbb{R}$ is the restriction of a function in $\mathcal{F}_0^*$ to $F(\Omega)$; the norm of f is defined as the infimum of the norms of the extensions of $f(F^{-1})$ to the whole of the Hilbert cube. According to the theorem on p. 351 of [4], this function space is an RKHS whose reproducing kernel is $\mathbf{k}(\omega,\omega') := \mathbf{k}^*(F(\omega),F(\omega'))$, where $\mathbf{k}^*$ is the reproducing kernel of $\mathcal{F}_0^*$; we can see that $\mathcal{F}$ is a continuous RKHS with finite imbedding constant.

Let us see that the RKHS $\mathcal{F}$ is universal. Take any continuous function $g : \Omega \to \mathbb{R}$. By the Tietze–Uryson theorem ([15], 2.1.8), $g(F^{-1}) : F(\Omega) \to \mathbb{R}$ can be extended to a continuous function g_1 on $[0,1]^\infty$. Let $g_2 \in \mathcal{F}_0^*$ be a function that is close to g_1 in the $C([0,1]^\infty)$ norm. Then $g_2(F) : \Omega \to \mathbb{R}$ will belong to $\mathcal{F}$ and will be close to g in the $C(\Omega)$ norm. ∎

Proof of Theorem 3

We start by proving the theorem under the assumption that $\mathbf{X}$ and $\mathbf{Y}$ are compact metric spaces. As explained above, in this case $\mathcal{P}(\mathbf{Y})$ is also compact and metrizable; therefore, $\Omega := \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$ is also compact and metrizable. Let f be a continuous real-valued function on Ω; our goal is to establish the consequent of (6).

Let $\mathcal{F}$ be a universal and continuous RKHS on Ω with finite imbedding constant (cf. Corollary 1). If $g \in \mathcal{F}$ is at a distance at most ϵ from f in the $C(\Omega)$ metric, we obtain from Theorem 4:

$$\begin{aligned} &\limsup_{N\to\infty} \left| \frac{1}{N} \sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) \right| \\ &\le \limsup_{N\to\infty} \left| \frac{1}{N} \sum_{n=1}^{N} \left(g\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} g\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) \right| + 2\epsilon = 2\epsilon. \quad (25) \end{aligned}$$

Since this can be done for any $\epsilon > 0$, the proof for the case of compact $\mathbf{X}$ and $\mathbf{Y}$ is complete.

The rest of the proof is based on the following game (an abstract version of the "doubling trick", [9]) played in a topological space X:

Game of removal $G(X)$

```
FOR $n = 1, 2, \dots$:
    Remover announces compact $K_n \subseteq X$.
    Evader announces $p_n \notin K_n$.
END FOR.
```

Winner: Evader if the set $\{p_1, p_2, \dots\}$ is precompact; Remover otherwise.

Intuitively, the goal of Evader is to avoid being removed to the infinity. Without loss of generality we will assume that Remover always announces a non-decreasing sequence of compact sets: $K_1 \subseteq K_2 \subseteq \cdots$.

Lemma 10 (Gruenhage) *Remover has a winning strategy in $G(X)$ if X is a locally compact and paracompact space.*

Proof We will follow the proof of Theorem 4.1 in [18] (the easy direction). If X is locally compact and σ-compact, there exists a non-decreasing sequence $K_1 \subseteq K_2 \subseteq \cdots$ of compact sets covering X, and each K_n can be extended to compact K_n^* so that $\operatorname{Int} K_n^* \supseteq K_n$ ([15], 3.3.2). Remover will obviously win $G(X)$ choosing $K_1^*, K_2^*, \dots$ as his moves.

If X is the sum of locally compact σ-compact spaces X_s, $s \in S$, Remover plays, for each $s \in S$, the strategy described in the previous paragraph on the subsequence of Evader's moves belonging to X_s. If Evader chooses $p_n \in X_s$ for infinitely many X_s, those X_s will form an open cover of the closure of

$\{p_1, p_2, \ldots\}$ without a finite subcover. If x_n are chosen from only finitely many X_s, there will be infinitely many x_n chosen from some X_s, and the result of the previous paragraph can be applied. It remains to remember that each locally compact paracompact can be represented as the sum of locally compact σ-compact subsets ([15], 5.1.27). ∎

Now it is easy to prove the general theorem. Forecaster's strategy ensuring (6) will be constructed from his strategies $\mathcal{S}(A, B)$ ensuring the consequent of (6) under the condition $\forall n : (x_n, y_n) \in A \times B$ for given compact sets $A \subseteq \mathbf{X}$ and $B \subseteq \mathbf{Y}$ and from Remover's winning strategy in $G(\mathbf{X} \times \mathbf{Y})$ (remember that, by Stone's theorem, [15], 5.1.3, all metric space are paracompact and that the product of two locally compact spaces is locally compact, [15], 3.3.13; therefore, Lemma 10 is applicable to $G(\mathbf{X} \times \mathbf{Y})$). Without loss of generality we assume that Remover's moves are always of the form $A \times B$ for $A \subseteq \mathbf{X}$ and $B \subseteq \mathbf{Y}$. Forecaster will be playing two games in parallel: the probability forecasting game and the auxiliary game of removal $G(\mathbf{X} \times \mathbf{Y})$ (in the role of Evader).

Forecaster asks Remover to make his first move $A_1 \times B_1$ in the game of removal. He then plays the probability forecasting game using the strategy $\mathcal{S}(A_1, B_1)$ until Reality chooses $(x_n, y_n) \notin A_1 \times B_1$ (forever if Reality never chooses such (x_n, y_n)). As soon as such (x_n, y_n) is chosen, Forecaster, in his Evader hat, announces (x_n, y_n) and notes Remover's move (A_2, B_2). He then plays the probability forecasting game using the strategy $\mathcal{S}(A_2, B_2)$ until Reality chooses $(x_n, y_n) \notin A_2 \times B_2$, etc.

Let us check that this strategy for Forecaster will always ensure (6). If Reality chooses (x_n, y_n) outside Forecaster's current $A_k \times B_k$ finitely often, the consequent of (6) will be satisfied. If Reality chooses (x_n, y_n) outside Forecaster's current $A_k \times B_k$ infinitely often, the set $\{(x_n, y_n) \,|\, n = 1, 2, \ldots\}$ will not be precompact, and so the antecedent of (6) will be violated.

5 Implications for probability theory

This section is an aside; its results are not used in the rest of the paper.

As we discussed at the end of Section 3, the procedure of defensive forecasting can be applied to virtually any law of probability (stated game-theoretically) to obtain a probability forecasting strategy whose forecasts are guaranteed to satisfy this law. Unfortunately, the standard laws of probability theory are often not strong enough to produce interesting probability forecasting strategies ([50], Section 4.1). In particular, for the purpose of this paper it would be easiest to apply the procedure of defensive forecasting to a law of probability asserting that (6) holds for all continuous functions f simultaneously with probability one. I am not aware of such results, but in the derivation of Theorem 5 we essentially proved one. In this section this result will be stated formally (as Theorem 5).

In general, it can be hoped that probability theory and competitive on-line prediction have a potential to enrich each other; not only laws of probability can be translated into probability forecasting strategies via defensive forecasting, but also the needs of competitive on-line prediction can help identify and fill gaps in the existing probability theory.

Game-theoretic result

Let us say that Skeptic *can force* some property E of the players' moves x_n, P_n, y_n, $n = 1, 2, \ldots$, in the testing protocol if he has a strategy guaranteeing that (1) his capital $\mathcal{K}_n$ is always non-negative, and (2) either E is satisfied or $\lim_{n\to\infty} \mathcal{K}_n = \infty$. The properties that can be forced by Skeptic are the game-theoretic analogue of the properties that hold with probability one in measure-theoretic probability theory ([37], Section 8.1).

The following is a corollary from the proof (rather than the statement, which is why we also call it a theorem) of Theorem 3. Its interpretation is that the true probabilities have good calibration-cum-resolution.

Theorem 5 *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are locally compact metric spaces. Skeptic can force*

$$\left(\{x_1, x_2, \ldots\} \text{ and } \{y_1, y_2, \ldots\} \text{ are precompact}\right) \Longrightarrow \left(\forall f : \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) = 0 \right) \qquad (26)$$

in the testing protocol, where f *ranges over all continuous functions* $f : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathbb{R}$.

Proof of Theorem 5

We will follow the proof of Theorem 3, starting from an analogue of Lemma 4.

Lemma 11 *Suppose* $\mathbf{Y}$ *is a metric compact. Let* $\Phi : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathcal{H}$ *be a function taking values in a Hilbert space* $\mathcal{H}$ *such that, for each* x, $\Phi(x, P, y)$ *is a continuous function of* $(P, y) \in \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$. *Suppose* $\sup_{x,P,y} \|\Phi(x, P, y)\|_{\mathcal{H}} < \infty$ *and set*

$$\Psi\left(x, P, y\right) := \Phi\left(x, P, y\right) - \int_{\mathbf{Y}} \Phi\left(x, P, y\right) P(\mathrm{d}y).$$

Skeptic can force

$$\left\| \sum_{n=1}^{N} \Psi\left(x_n, P_n, y_n\right) \right\|_{\mathcal{H}} = O\left(\sqrt{N} \log N\right) \qquad (27)$$

as $N \to \infty$.

Proof Let

$$c := \sup_{x,P,y} \|\Psi(x, P, y)\|_{\mathcal{H}} < \infty.$$

For $k, N = 1, 2, \ldots$, define

$$S_N^k := \begin{cases} 2^k + S_N & \text{if } c^2 N \le 2^k \\ S_{N-1}^k & \text{otherwise,} \end{cases}$$

where S_N is defined as in (17) (with Ψ in place of Ψ_n in all references to the proof of Lemma 4). Let us check that

$$S_N^* := \sum_{k=1}^{\infty} k^{-2} 2^{-k} S_N^k \tag{28}$$

is a capital process (obviously non-negative) of a strategy for Skeptic started with a finite initial capital. Since $S_0^k = 2^k$, the initial capital $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$ is indeed finite. It is also easy to see that the series (28) is convergent and that (18) still holds, where

$$A = \sum_{k=K}^{\infty} k^{-2} 2^{-k} 2 \sum_{n=1}^{N-1} \Psi(x_n, P_n, y_n)$$

for some K.

Skeptic can force $S_N^* \le C$, where C can depend on the path

$$x_1, P_1, y_1, x_2, P_2, y_2, \ldots$$

chosen by the players (see Lemma 3.1 in [37] or, for a simpler argument, the end of the proof of Theorem 3 in [49]). Therefore, he can force $k^{-2} 2^{-k} S_N^k \le C$ for all k. Setting $k := \lceil \log(c^2 N) \rceil$ (with log standing for the binary logarithm), we can rewrite the inequality $S_N^k \le C k^2 2^k$ as

$$2^k + S_N \le C k^2 2^k,$$

which implies

$$\begin{aligned} \left\| \sum_{n=1}^{N} \Psi\left(x_n, P_n, y_n\right) \right\|_{\mathcal{H}}^2 &\le C k^2 2^k \\ &\le C \left(\log(c^2 N) + 1\right)^2 2^{\log(c^2 N) + 1} = O\left(N \log^2 N\right). \quad ▮ \end{aligned}$$

The following analogue of Theorem 4 immediately follows from Lemma 11 and the proof of Lemma 7.

Lemma 12 *Let* $\mathbf{Y}$ *be a metric compact and* $\mathcal{F}$ *be a forecast-continuous RKHS on* $\mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$ *with finite imbedding constant. Skeptic can force*

$$\sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) = O\left(\|f\|_{\mathcal{F}} \sqrt{N} \log N \right)$$

as $N \to \infty$*, where the* O *is uniform in* $f \in \mathcal{F}$*.*

In its turn Lemma 12 immediately implies the statement of Theorem 5 in the case of compact $\mathbf{X}$ and $\mathbf{Y}$ (where the antecedent of (26) is automatically true): we can use the same argument based on (25).

Now let $\mathbf{X}$ and $\mathbf{Y}$ be any locally compact metric spaces. Skeptic can use the same method based on Remover's winning strategy in the game of removal as that used by Forecaster in the proof of Theorem 3 (see p. 26). This completes the proof of Theorem 5.

Measure-theoretic result

In this subsection we will use some notions of measure-theoretic probability theory, such as regular conditional distributions; all needed background information can be found in, e.g., [39].

Corollary 2 *Suppose $\mathcal{F}_n$, $n = 0, 1, \ldots$, is a filtration (increasing sequence of σ-algebras), $\mathbf{X}$ and $\mathbf{Y}$ are compact metric spaces, x_n, $n = 1, 2, \ldots$, are $\mathcal{F}_{n-1}$-measurable random elements taking values in $\mathbf{X}$, y_n, $n = 1, 2, \ldots$, are $\mathcal{F}_n$-measurable random elements taking values in $\mathbf{Y}$, and $P_n \in \mathcal{P}(\mathbf{Y})$ are regular conditional distributions of y_n given $\mathcal{F}_{n-1}$. Then*

$$\forall f: \quad \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(f(x_n, P_n, y_n) - \int_{\mathbf{Y}} f(x_n, P_n, y)\, P_n(\mathrm{d}y) \right) = 0 \qquad (29)$$

holds with probability one, where f ranges over all continuous functions $f : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathbb{R}$.

Proof Since $\mathbf{X}$ and $\mathbf{Y}$ are automatically complete and separable, regular conditional distributions exist by the corollary of Theorem II.7.5 in [39]. Our derivation of Corollary 2 from Theorem 5 will follow the standard recipe ([37], Section 8.1).

Skeptic's strategy forcing (29) (i.e., the consequent of (26)) can be chosen measurable (in the sense that $f_n(y)$ is a measurable function of y and the previous moves $x_1, P_1, y_1, \ldots, x_n, P_n$). This makes his capital process $\mathcal{K}_n$, $n = 0, 1, \ldots$, a martingale (in the usual measure-theoretic sense) with respect to the filtration $(\mathcal{F}_n)$. This martingale is non-negative and tends to infinity where (29) fails; standard results of probability theory (such as Doob's inequality, [39], Theorem VII.3.1.III, or Doob's convergence theorem, [39], Theorem VII.4.1) imply that (29) holds with probability one. ∎

6 Defensive forecasting for decision making: asymptotic theory

Our D-prediction algorithms are built on top of probability forecasting algorithms: D-predictions are found by minimizing the expected loss, with the expectation taken with respect to the probability forecast. The first problem that we have to deal with is the possibility that the minimizer of the expected loss will be a discontinuous function, whereas continuity is essential for the method of defensive forecasting (cf. Theorem 3, where f has to be a continuous function).

Continuity of choice functions

It will be convenient to use the notation

$$\lambda(x,\gamma,P) := \int_{\mathbf{Y}} \lambda(x,\gamma,y) P(\mathrm{d}y),$$

where P is a probability measure on $\mathbf{Y}$. Let us say that $G : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \to \Gamma$ is a (precise) *choice function* if it satisfies

$$\lambda(x, G(x,P), P) = \inf_{\gamma\in\Gamma} \lambda(x,\gamma,P), \quad \forall x \in \mathbf{X}, P \in \mathcal{P}(\mathbf{Y}).$$

As we said, a serious problem in implementing the expected loss minimization principle is that there might not exist a continuous choice function G; this is true even if $\mathbf{X}$, Γ, and $\mathbf{Y}$ are metric compacts and the loss function is continuous. If, however, the loss function $\lambda(x,\gamma,y)$ is convex in $\gamma \in \Gamma$, there exists an approximate choice function (although a precise choice function may still not exist).

The simplest example of a prediction game is perhaps the *simple prediction game*, in which there are no data, $\Gamma = \mathbf{Y} = \{0,1\}$ and $\lambda(\gamma,y) := |y-\gamma|$ (omitting the xs from our notation). There are no continuous approximate choice functions in this case, since there are no non-trivial (taking more than one value) continuous functions from the connected space $\mathcal{P}(\mathbf{Y})$ to Γ. If we allow randomized predictions, the simple prediction game effectively transforms into the following *absolute loss game*: $\Gamma = [0,1]$, $\mathbf{Y} = \{0,1\}$, $\lambda(\gamma,y) := |y-\gamma|$. Intuitively, the prediction γ in this game is the bias of the coin tossed to choose the prediction in the simple prediction game, and $|y-\gamma|$ is the expected loss in the latter.

Unfortunately, there is still no continuous choice function in the absolute loss game. It is easy to check that any choice function G must satisfy

$$G(P) := \begin{cases} 1 & \text{if } P(\{1\}) > 1/2 \\ 0 & \text{if } P(\{1\}) < 1/2, \end{cases} \tag{30}$$

but the case $P(\{1\}) = 1/2$ is a point of bifurcation: both predictions $\gamma = 1$ and $\gamma = 0$ are optimal, as indeed is every prediction in between. If $P(\{1\}) = 1/2$, the predictor finds himself in a position of Buridan's ass: he has several equally attractive decisions to choose from. It is clear that G defined by (30) cannot be continuously extended to the whole of $\mathcal{P}(\{0,1\})$.

We have to look for approximate choice functions. Under natural compactness and convexity conditions, they exist by the following lemma.

Lemma 13 *Let X be a paracompact, Y be a non-empty compact convex subset of a topological vector space, and $f : X \times Y \to \mathbb{R}$ be a continuous function such* *that $f(x,y)$ is convex in $y \in Y$ for each $x \in X$. For any $\epsilon > 0$ there exists a continuous "approximate choice function" $g : X \to Y$ such that*

$$\forall x \in X: \quad f(x, g(x)) \leq \inf_{y \in Y} f(x,y) + \epsilon. \tag{31}$$

Proof Each $(x,y) \in X \times Y$ has a neighborhood $A_{x,y} \times B_{x,y}$ such that $A_{x,y}$ and $B_{x,y}$ are open sets in X and Y, respectively, and

$$\sup_{A_{x,y} \times B_{x,y}} f - \inf_{A_{x,y} \times B_{x,y}} f < \frac{\epsilon}{2}.$$

For each $x \in X$ choose a finite subcover of the cover $\{A_{x,y} \times B_{x,y} \,|\, x \in A_{x,y}, y \in Y\}$ of $\{x\} \times Y$ and let A_x be the intersection of all $A_{x,y}$ in this subcover. The sets A_x constitute an open cover of X such that

$$(x_1 \in A_x, x_2 \in A_x) \Longrightarrow |f(x_1, y) - f(x_2, y)| < \frac{\epsilon}{2} \tag{32}$$

for all $x \in X$ and $y \in Y$. Since X is paracompact, there exists ([15], Theorem 5.1.9) a locally finite partition $\{\phi_i \mid i \in I\}$ of unity subordinated to the open cover of X formed by all A_x, $x \in X$. For each $i \in I$ choose $x_i \in X$ such that $\phi_i(x_i) > 0$ (without loss of generality we can assume that such x_i exists for each $i \in I$) and choose $y_i \in \arg\min_y f(x_i, y)$. Now we can set

$$g(x) := \sum_{i \in I} \phi_i(x) y_i.$$

Inequality (31) follows, by (32) and the convexity of $f(x,y)$ in y, from

$$\begin{aligned}
\forall y \in Y: \quad f(x, g(x)) &= f\left(x, \sum_i \phi_i(x) y_i\right) \leq \sum_i \phi_i(x) f(x, y_i) \\
&\leq \sum_i \phi_i(x) f(x_i, y_i) + \frac{\epsilon}{2} \leq \sum_i \phi_i(x) f(x_i, y) + \frac{\epsilon}{2} \\
&\leq \sum_i \phi_i(x) f(x, y) + \epsilon = f(x,y) + \epsilon,
\end{aligned}$$

where i ranges over the finite number of $i \in I$ for which $\phi_i(x)$ is non-zero. ∎

Suppose that $\mathbf{X}$ and $\mathbf{Y}$ are compact metric spaces, Γ is a compact convex subset of a topological vector space, and $\lambda(x,\gamma,y)$ is continuous in (x,γ,y) and convex in $\gamma \in \Gamma$ (therefore, by Lemma 6, $\lambda(x,\gamma,P)$ is continuous in $(x,\gamma,P) \in \mathbf{X} \times \Gamma \times \mathcal{P}(\mathbf{Y})$, and it is convex in γ). Taking $\mathbf{X} \times \mathcal{P}(\mathbf{Y})$ as X and Γ as Y, we can see that for each $\epsilon > 0$ there exists an approximate choice function G satisfying

$$\lambda(x, G(x,P), P) \leq \inf_{\gamma \in \Gamma} \lambda(x,\gamma,P) + \epsilon, \quad \forall x \in \mathbf{X}, P \in \mathcal{P}(\mathbf{Y}). \tag{33}$$

Proof of a weak form of Theorem 1

Suppose $\mathbf{X}$ and $\mathbf{Y}$ are compact metric spaces and Γ is a compact convex subset of a topological vector space. In this subsection we will prove the existence of a prediction algorithm guaranteeing (2) (whose antecedent can now be ignored) with ≤ 0 replaced by $\le \epsilon$ for all continuous prediction rules D for an arbitrarily small constant $\epsilon > 0$. Let G satisfy (33). If Predictor chooses his predictions by applying the approximate choice function G to x_n and probability forecasts P_n for y_n satisfying (6) of Theorem 3, we will have

$$\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) = \sum_{n=1}^{N} \lambda(x_n, G(x_n, P_n), y_n)$$

$$= \sum_{n=1}^{N} \lambda(x_n, G(x_n, P_n), P_n) + \sum_{n=1}^{N} \Bigl(\lambda(x_n, G(x_n, P_n), y_n) - \lambda(x_n, G(x_n, P_n), P_n)\Bigr)$$

$$= \sum_{n=1}^{N} \lambda(x_n, G(x_n, P_n), P_n) + o(N) \le \sum_{n=1}^{N} \lambda(x_n, D(x_n), P_n) + \epsilon N + o(N)$$

$$= \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) - \sum_{n=1}^{N} \Bigl(\lambda(x_n, D(x_n), y_n) - \lambda(x_n, D(x_n), P_n)\Bigr) + \epsilon N + o(N)$$

$$= \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) + \epsilon N + o(N). \quad (34)$$

7 Defensive forecasting for decision making: loss bounds

The goal of this section is to finish the proof of Theorem 1 and to establish its non-asymptotic version. We will start with the latter.

Results

Let $\mathcal{F}$ be an RKHS on $\mathbf{X}\times\mathbf{Y}$ with finite imbedding constant. For each prediction rule $D : \mathbf{X} \to \Gamma$, define a function $\lambda_D : \mathbf{X} \times \mathbf{Y} \to \mathbb{R}$ by

$$\lambda_D(x, y) := \lambda(x, D(x), y).$$

The notation $\|f\|_{\mathcal{F}}$ will be used for all functions $f : \mathbf{X} \times \mathbf{Y} \to \mathbb{R}$: we just set $\|f\|_{\mathcal{F}} := \infty$ for $f \notin \mathcal{F}$. We will continue to use the notation $\mathbf{c}_{\mathcal{F}}$ for the imbedding constant (defined by (12), where $\Omega := \mathbf{X} \times \mathbf{Y}$). Set

$$\mathbf{c}_\lambda := \sup_{x\in\mathbf{X},\gamma\in\Gamma,y\in\mathbf{Y}} \lambda(x, \gamma, y) - \inf_{x\in\mathbf{X},\gamma\in\Gamma,y\in\mathbf{Y}} \lambda(x, \gamma, y);$$

this is finite if λ is continuous and $\mathbf{X}, \Gamma, \mathbf{Y}$ are compact.

Theorem 6 *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are compact metric spaces,* Γ *is a convex compact subset of a topological vector space and the loss function* $\lambda(x, \gamma, y)$ *is continuous in* (x, γ, y) *and convex in* $\gamma \in \Gamma$*. Let* $\mathcal{F}$ *be a forecast-continuous RKHS on* $\mathbf{X} \times \mathbf{Y}$ *with finite imbedding constant* $\mathbf{c}_{\mathcal{F}}$*. There is an on-line prediction algorithm that guarantees*

$$\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) \le \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2} \left(\|\lambda_D\|_{\mathcal{F}} + 1\right) \sqrt{N} + 1 \tag{35}$$

for all prediction rules D *and all* $N = 1, 2, \ldots$*.*

An application of Hoeffding's inequality immediately gives the following corollary (we postpone the details of the simple proof until p. 37).

Corollary 3 *Suppose* $\mathbf{X}, \Gamma, \mathbf{Y}$ *are compact metric spaces and the loss function* λ *is continuous. Let* $N \in \{1, 2, \ldots\}$ *and* $\delta \in (0, 1)$*. There is a randomized on-line prediction algorithm achieving*

$$\sum_{n=1}^{N} \lambda(x_n, g_n, y_n) \le \sum_{n=1}^{N} \lambda(x_n, d_n, y_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2} \left(\|\lambda_D\|_{\mathcal{F}} + 1\right) \sqrt{N} + \mathbf{c}_\lambda \sqrt{2 \ln \frac{1}{\delta}} \sqrt{N} + 1$$

with probability at least $1 - \delta$ *for any randomized prediction rule* $D : \mathbf{X} \to \mathcal{P}(\Gamma)$*;* g_n *and* d_n *are independent random variables distributed as* γ_n *and* $D(x_n)$*, respectively.*

The above results are non-vacuous only when λ_D is an element of the function space $\mathcal{F}$. If $\mathcal{F}$ is a Sobolev space, this condition follows from D being in the Sobolev space and the smoothness of λ. For example, Moser proved in 1966 the following result concerning composition in Sobolev spaces. Let Ω be a smooth bounded domain in $\mathbb{R}^K$ and m be an integer number satisfying $2m > K$. If $u \in H^m(\Omega)$ and $\Phi \in C^m(\mathbb{R})$, then $\Phi \circ u \in H^m(\Omega)$ (see [31]; for further results, see [8]).

Two special cases of calibration-cum-resolution

In the chain (34) we applied the law of large numbers (the property of good calibration-cum-resolution) twice: in the third and fifth equalities. It is easy to see, however, that in fact the fifth equality depends only on resolution and the third equality, although it depends on calibration-cum-resolution, involves a known function f (in the notation of (6)). We will say that the fifth equality depends on "general resolution" whereas the third equality depends on "specific calibration-cum-resolution". This limited character of the required calibration-cum-resolution becomes important for obtaining good bounds on the predictive performance: in the following subsections we will construct prediction algorithms that satisfy the properties of specific calibration-cum-resolution and general resolution and merge them into one algorithm; we will start from the last step.

Synthesis of prediction algorithms

The following corollary of Lemma 4 will allow us to construct prediction algorithms that achieve two goals simultaneously (specific calibration-cum-resolution and general resolution).

Corollary 4 *Let* $\mathbf{Y}$ *be a metric compact and* $\Phi_{n,j} : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathcal{H}_j$*,* $n = 1, 2, \ldots$*,* $j = 0, 1$*, be functions taking values in Hilbert spaces* $\mathcal{H}_j$ *and such that* $\Phi_{n,j}(x, P, y)$ *is continuous in* (P, y) *for all* n *and both* j*. Let* a_0 *and* a_1 *be two positive constants. There is a probability forecasting strategy that guarantees*

$$\left\| \sum_{n=1}^{N} \Psi_{n,j}(x_n, P_n, y_n) \right\|_{\mathcal{H}_j}^2 \le \frac{1}{a_j} \sum_{n=1}^{N} \left(a_0 \left\| \Psi_{n,0}(x_n, P_n, y_n) \right\|_{\mathcal{H}_0}^2 + a_1 \left\| \Psi_{n,1}(x_n, P_n, y_n) \right\|_{\mathcal{H}_1}^2 \right)$$

for all N *and for both* $j = 0$ *and* $j = 1$*, where*

$$\Psi_{n,j}(x, P, y) := \Phi_{n,j}(x, P, y) - \int_{\mathbf{Y}} \Phi_{n,j}(x, P, y) \, P(\mathrm{d}y).$$

Proof Define the "weighted direct sum" $\mathcal{H}$ of $\mathcal{H}_0$ and $\mathcal{H}_1$ as the Cartesian product $\mathcal{H}_0 \times \mathcal{H}_1$ equipped with the inner product

$$\langle g, g' \rangle_{\mathcal{H}} = \langle (g_0, g_1), (g'_0, g'_1) \rangle_{\mathcal{H}} := \sum_{j=0}^{1} a_j \langle g_j, g'_j \rangle_{\mathcal{H}_j}.$$

Now we can define $\Phi : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathcal{H}$ by

$$\Phi_n(x, P, y) := \left(\Phi_{n,0}(x, P, y), \Phi_{n,1}(x, P, y) \right).$$

It is clear that $\Phi_n(x, P, y)$ is continuous in (P, y) for all n. Applying the strategy of Lemma 4 to it and using (16), we obtain

$$\begin{aligned} a_j \left\| \sum_{n=1}^{N} \Psi_{n,j}(x_n, P_n, y_n) \right\|_{\mathcal{H}_j}^2 &\le \left\| \left(\sum_{n=1}^{N} \Psi_{n,0}(x_n, P_n, y_n), \sum_{n=1}^{N} \Psi_{n,1}(x_n, P_n, y_n) \right) \right\|_{\mathcal{H}}^2 \\ &= \left\| \sum_{n=1}^{N} \Psi_n(x_n, P_n, y_n) \right\|_{\mathcal{H}}^2 \le \sum_{n=1}^{N} \left\| \Psi_n(x_n, P_n, y_n) \right\|_{\mathcal{H}}^2 \end{aligned}$$

$$= \sum_{n=1}^{N} \sum_{j=0}^{1} a_j \left\| \Psi_{n,j}(x_n, P_n, y_n) \right\|_{\mathcal{H}_j}^2 . \quad \blacksquare$$

Suppose $\mathbf{X}, \Gamma, \mathbf{Y}$ are metric compacts and $\mathcal{F}$ is a forecast-continuous RKHS on $\mathbf{X} \times \mathbf{Y}$. Let $G_n : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \to \Gamma$ be a sequence of approximate choice functions satisfying

$$\lambda(x, G_n(x,P), P) < \inf_{\gamma \in \Gamma} \lambda(x, \gamma, P) + 2^{-n}, \quad \forall x \in \mathbf{X}, P \in \mathcal{P}(\mathbf{Y})$$

(they exist by (33)). Corollary 4 will be applied to $a_0 = a_1 = 1$ and to the mappings

$$\Psi_{n,0}(x,P,y) := \lambda(x, G_n(x,P), y) - \lambda(x, G_n(x,P), P), \tag{36}$$

$$\Psi_{n,1}(x,P,y) := \mathbf{k}_{x,y} - \mathbf{k}_{x,P}, \tag{37}$$

where $\mathbf{k}_{x,y}$ is the evaluation functional at (x,y) for $\mathcal{F}$ and $\mathbf{k}_{x,P}$ is the mean of $\mathbf{k}_{x,y}$ with respect to $P(\mathrm{d}y)$. It is easy to see that

$$\left\| \Psi_{n,0}(x,P,y) \right\|_{\mathbb{R}} = \left| \Psi_{n,0}(x,P,y) \right| \le \mathbf{c}_\lambda, \quad \left\| \Psi_{n,1}(x,P,y) \right\|_{\mathcal{F}} \le 2\mathbf{c}_{\mathcal{F}}. \tag{38}$$

Specific calibration-cum-resolution

Corollary 4 immediately implies:

Lemma 14 *The probability forecasting strategy of Corollary 4 based on (36) and (37) guarantees*

$$\left| \sum_{n=1}^{N} \Big(\lambda(x_n, G_n(x_n, P_n), y_n) - \lambda(x_n, G_n(x_n, P_n), P_n) \Big) \right| \le \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2} \sqrt{N}.$$

Proof This follows from

$$\left| \sum_{n=1}^{N} \Big(\lambda(x_n, G_n(x_n, P_n), y_n) - \lambda(x_n, G_n(x_n, P_n), P_n) \Big) \right|^2 \le \sum_{n=1}^{N} \left(\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2 \right)$$

(see (38)). $\blacksquare$

General resolution I

The following lemma is proven similarly to Lemma 7.

Lemma 15 *The probability forecasting strategy of Corollary 4 based on (36) and (37) guarantees*

$$\left| \sum_{n=1}^{N} \Big(\lambda(x_n, D(x_n), y_n) - \lambda(x_n, D(x_n), P_n) \Big) \right| \le \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2} \left\| \lambda_D \right\|_{\mathcal{F}} \sqrt{N}.$$

Proof This follows from

$$\begin{aligned}
&\left|\sum_{n=1}^{N}\Bigl(\lambda(x_n,D(x_n),y_n)-\lambda(x_n,D(x_n),P_n)\Bigr)\right| \\
&\qquad\qquad\qquad = \left|\sum_{n=1}^{N}\Bigl(\lambda_D(x_n,y_n)-\lambda_D(x_n,P_n)\Bigr)\right| \\
&\qquad = \left|\sum_{n=1}^{N}\langle\lambda_D,\mathbf{k}_{x_n,y_n}-\mathbf{k}_{x_n,P_n}\rangle_{\mathcal{F}}\right| \le \|\lambda_D\|_{\mathcal{F}}\left\|\sum_{n=1}^{N}(\mathbf{k}_{x_n,y_n}-\mathbf{k}_{x_n,P_n})\right\|_{\mathcal{F}} \\
&\qquad\qquad\qquad \le \|\lambda_D\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}(\mathbf{c}_\lambda^2+4\mathbf{c}_{\mathcal{F}}^2)} = \sqrt{\mathbf{c}_\lambda^2+4\mathbf{c}_{\mathcal{F}}^2}\,\|\lambda_D\|_{\mathcal{F}}\sqrt{N}
\end{aligned}$$

(we have used Corollary 4 and (38)). ∎

Proof of Theorem 6

Let $\gamma_n := G_n(x_n,P_n)$ where P_n are produced by the probability forecasting strategy of Corollary 4 based on (36) and (37). Following (34) and using the previous two lemmas, we obtain:

$$\begin{aligned}
\sum_{n=1}^{N}\lambda(x_n,\gamma_n,y_n) &= \sum_{n=1}^{N}\lambda(x_n,G_n(x_n,P_n),y_n) \\
&= \sum_{n=1}^{N}\lambda(x_n,G_n(x_n,P_n),P_n) \\
&\quad + \sum_{n=1}^{N}\Bigl(\lambda(x_n,G_n(x_n,P_n),y_n)-\lambda(x_n,G_n(x_n,P_n),P_n)\Bigr) \\
&\le \sum_{n=1}^{N}\lambda(x_n,G_n(x_n,P_n),P_n) + \sqrt{\mathbf{c}_\lambda^2+4\mathbf{c}_{\mathcal{F}}^2}\sqrt{N} \\
&\le \sum_{n=1}^{N}\lambda(x_n,D(x_n),P_n) + \sqrt{\mathbf{c}_\lambda^2+4\mathbf{c}_{\mathcal{F}}^2}\sqrt{N}+1 \\
&= \sum_{n=1}^{N}\lambda(x_n,D(x_n),y_n) + \sqrt{\mathbf{c}_\lambda^2+4\mathbf{c}_{\mathcal{F}}^2}\sqrt{N}+1 \\
&\quad - \sum_{n=1}^{N}\Bigl(\lambda(x_n,D(x_n),y_n)-\lambda(x_n,D(x_n),P_n)\Bigr) \\
&\le \sum_{n=1}^{N}\lambda(x_n,D(x_n),y_n) + \sqrt{\mathbf{c}_\lambda^2+4\mathbf{c}_{\mathcal{F}}^2}\,(\|\lambda_D\|_{\mathcal{F}}+1)\sqrt{N}+1.
\end{aligned}$$

Proof of Corollary 3

Since $\lambda(x_n, g_n, y_n) - \lambda(x_n, d_n, y_n)$ never exceeds $\mathbf{c}_\lambda$ in absolute value, Hoeffding's inequality ([9], Corollary A.1) shows that

$$\mathbb{P}\Biggl\{\sum_{n=1}^{N}\Bigl(\lambda(x_n, g_n, y_n) - \lambda(x_n, d_n, y_n)\Bigr) - \sum_{n=1}^{N}\Bigl(\lambda(x_n, \gamma_n, y_n) - \lambda(x_n, D(x_n), y_n)\Bigr) \\ > t\Biggr\} \le \exp\left(-\frac{t^2}{2\mathbf{c}_\lambda^2 N}\right)$$

for every $t > 0$. Choosing t satisfying

$$\exp\left(-\frac{t^2}{2\mathbf{c}_\lambda^2 N}\right) = \delta,$$

i.e.,

$$t := \mathbf{c}_\lambda \sqrt{2 \ln \frac{1}{\delta}} \sqrt{N},$$

we obtain the statement of Corollary 3.

General resolution II

To prove Theorem 1, we will need the following variation on Lemma 15.

Lemma 16 *The probability forecasting strategy of Corollary 4 based on (36) and (37) guarantees*

$$\left|\sum_{n=1}^{N}\left(f(x_n, y_n) - \int_{\mathbf{Y}} f(x_n, y) P(\mathrm{d}y)\right)\right| \le \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_\mathcal{F}^2}\, \|f\|_\mathcal{F} \sqrt{N}$$

for any $f \in \mathcal{F}$.

Proof Following the proof of Lemma 15:

$$\left|\sum_{n=1}^{N}\left(f(x_n, y_n) - \int_{\mathbf{Y}} f(x_n, y) P_n(\mathrm{d}y)\right)\right| \\ = \left|\sum_{n=1}^{N} \langle f, \mathbf{k}_{x_n, y_n} - \mathbf{k}_{x_n, P_n}\rangle_\mathcal{F}\right| \le \|f\|_\mathcal{F} \left\|\sum_{n=1}^{N} (\mathbf{k}_{x_n, y_n} - \mathbf{k}_{x_n, P_n})\right\|_\mathcal{F} \\ \le \|f\|_\mathcal{F} \sqrt{\sum_{n=1}^{N} (\mathbf{c}_\lambda^2 + 4\mathbf{c}_\mathcal{F}^2)} = \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_\mathcal{F}^2}\, \|f\|_\mathcal{F} \sqrt{N}. \quad \blacksquare$$

Proof of Theorem 1

As in the proof of Theorem 3, we first assume that $\mathbf{X}$, Γ, and $\mathbf{Y}$ are compact. Let us first see that the prediction algorithm of Theorem 6 fed with a suitable RKHS guarantees the consequent of (2) for all continuous D. Let $\mathcal{F}$ be a universal and continuous RKHS on $\mathbf{X} \times \mathbf{Y}$ with finite imbedding constant $\mathbf{c}_{\mathcal{F}}$.

Fix a continuous decision rule $D : \mathbf{X} \to \Gamma$. For any $\epsilon > 0$, we can find a function $f \in \mathcal{F}$ that is ϵ-close in $C(\mathbf{X} \times \mathbf{Y})$ to $\lambda(x, D(x), y)$. Following (34) and the similar chain in the proof of Theorem 6, we obtain:

$$\begin{aligned}
\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) &= \sum_{n=1}^{N} \lambda(x_n, G_n(x_n, P_n), y_n) \\
&= \sum_{n=1}^{N} \lambda(x_n, G_n(x_n, P_n), P_n) \\
&\quad + \sum_{n=1}^{N} \Big(\lambda(x_n, G_n(x_n, P_n), y_n) - \lambda(x_n, G_n(x_n, P_n), P_n) \Big) \\
&\le \sum_{n=1}^{N} \lambda(x_n, G_n(x_n, P_n), P_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2}\sqrt{N} \\
&\le \sum_{n=1}^{N} \lambda(x_n, D(x_n), P_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2}\sqrt{N} + 1 \\
&= \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2}\sqrt{N} + 1 \\
&\quad - \sum_{n=1}^{N} \Big(\lambda(x_n, D(x_n), y_n) - \lambda(x_n, D(x_n), P_n) \Big) \\
&\le \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2}\sqrt{N} + 1 \\
&\quad - \sum_{n=1}^{N} \Big(f(x_n, y_n) - \int_{\mathbf{Y}} f(x_n, y) P_n(y) \Big) + 2\epsilon N \\
&\le \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2} \left(\|f\|_{\mathcal{F}} + 1 \right) \sqrt{N} + 1 \\
&\quad + 2\epsilon N.
\end{aligned}$$

We can see that

$$\limsup_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) \right) \le 2\epsilon;$$

since this is true for any $\epsilon > 0$, the consequent of (2) holds.

It remains to get rid of the assumption of compactness of $\mathbf{X}$, Γ, and $\mathbf{Y}$. We will need the following lemma.

Lemma 17 *Under the conditions of Theorem 1, for each pair of compact sets $A \subseteq \mathbf{X}$ and $B \subseteq \mathbf{Y}$ there exists a compact set $C = C(A,B) \subseteq \Gamma$ such that for each continuous prediction rule $D : \mathbf{X} \to \Gamma$ there exists a continuous prediction rule $D' : \mathbf{X} \to C$ that dominates D in the sense*

$$\forall x \in A, y \in B : \quad \lambda(x, D'(x), y) \le \lambda(x, D(x), y). \tag{39}$$

Proof Without loss of generality A and B are assumed non-empty. Fix any $\gamma_0 \in \Gamma$. Let

$$M_1 := \sup_{(x,y)\in A\times B} \lambda(x,\gamma_0,y),$$

let $C_1 \subseteq \Gamma$ be a compact set such that

$$\forall x \in A, \gamma \notin C_1, y \in B : \quad \lambda(x,\gamma,y) > M_1 + 1,$$

let

$$M_2 := \sup_{(x,\gamma,y)\in A\times C_1\times B} \lambda(x,\gamma,y).$$

and let $C_2 \subseteq \Gamma$ be a compact set such that

$$\forall x \in A, \gamma \notin C_2, y \in B : \quad \lambda(x,\gamma,y) > M_2 + 1.$$

It is obvious that $M_1 \le M_2$ and $\gamma_0 \in C_1 \subseteq C_2$.

Let us now check that C_1 lies inside the interior of C_2. Indeed, for any fixed $(x,y) \in A \times B$ and $\gamma \in C_1$, we have $\lambda(x,\gamma,y) \le M_2$; since $\lambda(x,\gamma',y) > M_2 + 1$ for all $\gamma' \notin C_2$, some neighborhood of γ will lie completely in C_2.

Let $D : \mathbf{X} \to \Gamma$ be a continuous prediction rule. We will show that (39) holds for some continuous prediction rule D' taking values in the compact set C_2. Namely, we define

$$D'(x) := \begin{cases} D(x) & \text{if } D(x) \in C_1 \\ \frac{\rho(D(x),\Gamma\setminus C_2)}{\rho(D(x),C_1)+\rho(D(x),\Gamma\setminus C_2)} D(x) + \frac{\rho(D(x),C_1)}{\rho(D(x),C_1)+\rho(D(x),\Gamma\setminus C_2)} \gamma_0 & \text{if } D(x) \in C_2 \setminus C_1 \\ \gamma_0 & \text{if } D(x) \in \Gamma \setminus C_2 \end{cases}$$

where ρ is the metric on Γ; the denominator $\rho(D(x),C_1) + \rho(D(x),\Gamma \setminus C_2)$ is always positive since already $\rho(D(x),C_1)$ is positive. Assuming C_2 convex (which can be done by [34], Theorem 3.20(c)), we can see that D' indeed takes values in C_2. The only points x at which the continuity of D' is not obvious are those for which $D(x)$ lies on the boundary of C_1: one has to use the fact that C_1 is covered by the interior of C_2.

It remains to check (39); the only non-trivial case is $D(x) \in C_2 \setminus C_1$. By the convexity of $\lambda(x,\gamma,y)$ in γ, the inequality in (39) will follow from

$$\frac{\rho(D(x),\Gamma\setminus C_2)}{\rho(D(x),C_1)+\rho(D(x),\Gamma\setminus C_2)}\lambda(x,D(x),y)$$
$$+\frac{\rho(D(x),C_1)}{\rho(D(x),C_1)+\rho(D(x),\Gamma\setminus C_2)}\lambda(x,\gamma_0,y)\le\lambda(x,D(x),y),$$

i.e.,

$$\lambda(x,\gamma_0,y)\le\lambda(x,D(x),y).$$

Since the left-hand side of the last inequality is at most M_1 and its right-hand side exceeds M_1+1, it holds true. ∎

For each pair of compact $A\subseteq\mathbf{X}$ and $B\subseteq\mathbf{Y}$ fix a compact $C(A,B)\subseteq\Gamma$ as in the lemma. Similarly to the proof of Theorem 3, Predictor's strategy ensuring (2) is constructed from Remover's winning strategy in $G(\mathbf{X}\times\mathbf{Y})$ and from Predictor's strategies $\mathcal{S}(A,B)$ outputting predictions $\gamma_n\in C(A,B)$ and ensuring the consequent of (2) for $D:A\to C(A,B)$ under the assumption that $(x_n,y_n)\in A\times B$ for given compact $A\subseteq\mathbf{X}$ and $B\subseteq\mathbf{Y}$. Remover's moves are assumed to be of the form $A\times B$ for compact $A\subseteq\mathbf{X}$ and $B\subseteq\mathbf{Y}$. Predictor is simultaneously playing the game of removal $G(\mathbf{X}\times\mathbf{Y})$ as Evader.

Predictor asks Remover to make his first move $A_1\times B_1$ in the game of removal. Predictor then plays the prediction game using the strategy $\mathcal{S}(A_1,B_1)$ until Reality chooses $(x_n,y_n)\notin A_1\times B_1$ (forever if Reality never chooses such (x_n,y_n)). As soon as such (x_n,y_n) is chosen, Predictor announces (x_n,y_n) in the game of removal and notes Remover's response (A_2,B_2). He then continues playing the prediction game using the strategy $\mathcal{S}(A_2,B_2)$ until Reality chooses $(x_n,y_n)\notin A_2\times B_2$, etc.

Let us check that this strategy for Predictor will always ensure (2). If Reality chooses (x_n,y_n) outside Predictor's current $A_k\times B_k$ finitely often, the consequent of (2) will be satisfied for all continuous $D:\mathbf{X}\to C(A_K,B_K)$ ((A_K,B_K) being Remover's last move) and so, by Lemma 17, for all continuous $D:\mathbf{X}\to\Gamma$. If Reality chooses (x_n,y_n) outside Predictor's current $A_k\times B_k$ infinitely often, the set of (x_n,y_n), $n=1,2,\ldots$, will not be precompact, and so the antecedent of (2) will be violated.

Proof of Theorem 2

Define

$$\lambda(x,\gamma,y):=\int_\Gamma\lambda(x,g,y)\gamma(\mathrm{d}g),\tag{40}$$

where γ is a probability measure on Γ. This is the loss function in a new game of prediction with the prediction space $\mathcal{P}(\Gamma)$. When γ ranges over $\mathcal{P}(C)$ (identified with the subset of $\mathcal{P}(\Gamma)$ consisting of the measures concentrated on C) for a compact C, the loss function (40) is continuous by Lemma 6. We need the following analogue of Lemma 17.

Lemma 18 *Under the conditions of Theorem 2, for each pair of compact sets $A\subseteq\mathbf{X}$ and $B\subseteq\mathbf{Y}$ there exists a compact set $C=C(A,B)\subseteq\Gamma$ such that* *for each continuous randomized prediction rule $D : \mathbf{X} \to \mathcal{P}(\Gamma)$ there exists a continuous randomized prediction rule $D' : \mathbf{X} \to \mathcal{P}(C)$ such that (39) holds (D' dominates D "on average").*

Proof Define γ_0, C_1, and C_2 as in the proof of Lemma 17. Fix a continuous function $f_1 : \Gamma \to [0,1]$ such that $f_1 = 1$ on C_1 and $f_1 = 0$ on $\Gamma \setminus C_2$ (such an f_1 exists by the Tietze–Uryson theorem, [15], 2.1.8). Set $f_2 := 1 - f_1$. Let $D : \mathbf{X} \to \mathcal{P}(\Gamma)$ be a continuous randomized prediction rule. For each $x \in \mathbf{X}$, split $D(x)$ into two measures on Γ absolutely continuous with respect to $D(x)$: $D_1(x)$ with Radon–Nikodym density f_1 and $D_2(x)$ with Radon–Nikodym density f_2; set

$$D'(x) := D_1(x) + |D_2(x)|\,\delta_{\gamma_0}$$

(letting $|P| := P(\Gamma)$ for $P \in \mathcal{P}(\Gamma)$). It is clear that D' is continuous (in the topology of weak convergence, as usual), takes values in $\mathcal{P}(C_2)$, and

$$\begin{aligned}\lambda(x, D'(x), y) &= \int_\Gamma \lambda(x,\gamma,y) f_1(\gamma) D(x)(\mathrm{d}\gamma) + \lambda(x,\gamma_0,y)\int_\Gamma f_2(\gamma) D(x)(\mathrm{d}\gamma)\\ &\le \int_\Gamma \lambda(x,\gamma,y) f_1(\gamma) D(x)(\mathrm{d}\gamma) + \int_\Gamma M_1 f_2(\gamma) D(x)(\mathrm{d}\gamma)\\ &\le \int_\Gamma \lambda(x,\gamma,y) f_1(\gamma) D(x)(\mathrm{d}\gamma) + \int_\Gamma \lambda(x,\gamma,y) f_2(\gamma) D(x)(\mathrm{d}\gamma) = \lambda(x, D(x), y)\end{aligned}$$

for all $(x,y) \in A \times B$. ∎

Fix one of the mappings $(A,B) \mapsto C(A,B)$ whose existence is asserted by the lemma.

We will prove that the strategy of the previous subsection with $\mathcal{P}(C(A,B))$ in place of $C(A,B)$ applied to the new game is universally consistent. Let $D : \mathbf{X} \to \mathcal{P}(\Gamma)$ be a continuous randomized prediction rule, i.e., a continuous prediction rule in the new game. Let (A_K, B_K) be Remover's last move (if Remover makes infinitely many moves, the antecedent of (3) is false, and there is nothing to prove), and let $D' : \mathbf{X} \to \mathcal{P}(C(A_K, B_K))$ be a continuous randomized prediction rule satisfying (39) with $A := A_K$ and $B := B_K$. From some n on our randomized prediction algorithm produces $\gamma_n \in \mathcal{P}(\Gamma)$ concentrated on $C(A_K, B_K)$, and they will satisfy

$$\begin{aligned}&\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(x_n,\gamma_n,y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda(x_n, D(x_n), y_n)\right)\\ &\qquad\le \limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(x_n,\gamma_n,y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda(x_n, D'(x_n), y_n)\right) \le 0. \qquad (41)\end{aligned}$$

The loss function is bounded in absolute value on the compact set $A_K \times (C(A_K, B_K) \cup D(A_K)) \times B_K$ by a constant c. The law of the iterated logarithm (see, e.g., [37], (5.8)) implies that

$$\limsup_{N\to\infty} \frac{\left|\sum_{n=1}^{N} \left(\lambda(x_n, g_n, y_n) - \lambda(x_n, \gamma_n, y_n)\right)\right|}{\sqrt{2c^2 N \ln\ln N}} \le 1,$$

$$\limsup_{N\to\infty} \frac{\left|\sum_{n=1}^{N} \left(\lambda(x_n, d_n, y_n) - \lambda(x_n, D(x_n), y_n)\right)\right|}{\sqrt{2c^2 N \ln\ln N}} \le 1$$

with probability one. Combining the last two inequalities with (41) gives

$$\limsup_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, g_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, d_n, y_n) \right) \le 0 \text{ a.s.}$$

This immediately implies (3).

8 Conclusion

In this section I will list what I think are interesting directions of further research.

The data space as a bottleneck

It is easy to see that if we set $\mathbf{X} := \sum_{n=0}^{\infty} \mathbf{Y}^n$ and

$$x_n := (y_1, \ldots, y_{n-1}),$$

it becomes impossible to compete even with the simplest prediction rules $D : \mathbf{X} \to \mathbf{Y}$: there needs be no connection between the restrictions of D to $\mathbf{Y}^n$ for different n. The requirement that $y_1, \ldots, y_{n-1}$ should be compressed into an element x_n of a locally compact space $\mathbf{X}$ restricts the set of possible prediction rules so that it becomes manageable. We can consider $\mathbf{X}$ to be the necessary bottleneck in our notion of a prediction rule, and the requirement of local compactness of $\mathbf{X}$ makes it narrow enough for us to be able to compete with all continuous prediction rules. A natural question is: can the requirement of the local compactness of $\mathbf{X}$ be weakened while preserving the existence of on-line prediction algorithms competitive with the continuous prediction rules? (And it should be remembered that our (2) might be a poor formalization of the latter property if sizeable pieces of $\mathbf{X}$ cannot be expected to be compact.)

Randomization

It appears that various aspects of randomization in this paper and competitive on-line prediction in general deserve further study. For example, the bound of Corollary 3 is based on the worst possible outcome of Predictor's randomization and the best possible outcome of the prediction rule's randomization (disregarding an event of probability at most δ). This is unfair to Predictor. Of course, comparing the expected values of Predictor's and the prediction rule's loss would be an even worse solution: this would ignore the magnitude of the likely deviations of the loss from its expected value. It would be too crude to use the variance as the only indicator of the likely deviations, and it appears that the right formalization should involve the overall distribution of the deviations.

A related observation is that, when using a prediction strategy based on defensive forecasting, Predictor needs randomization only when there are several very different predictions with similar expected losses with respect to the current probability forecast P_n. Since P_n are guaranteed to agree with reality, we would not expect that Predictor will often find himself in such a position provided Reality is neutral (rather than an active opponent). Predictor's strategy will be almost deterministic. It would be interesting to formalize this intuition.

Limitations of competitive on-line prediction

In conclusion, I will briefly discuss two serious limitations of this paper.

First, the main results of this paper only concern one-step-ahead prediction. In a more general framework the loss function would depend not only on y_n but on other future outcomes as well. There are simple ways of extending our results in this direction: e.g., if the loss function $\lambda = \lambda(x_n, \gamma_n, y_n, y_{n+1})$ depends on both y_n and y_{n+1}, we could run two on-line prediction algorithms with the observation space $\mathbf{Y}^2$, one responsible for choosing γ_n for odd n and the other for even n. However, cleaner and more principled approaches are needed.

As we noted earlier (see Remark 1), the general interpretation of D-predictions is that they are decisions made by a small decision maker. To see why the decision maker is assumed small, let us consider (1), which the kind of guarantee (such as (35)) provided in competitive on-line prediction (although see [9], Section 7.11, for a recent advance). Predictor's and the prediction rule D's losses are compared on the same sequence $x_1, y_1, x_2, y_2, \ldots$ of data and observations. If Predictor is a big decision maker (i.e., his decisions affect Reality's future behavior) the interpretation of (1) becomes problematic: presumably, $x_1, y_1, x_2, y_2, \ldots$ resulted from Predictor's decisions γ_n, and D's loss should be evaluated on a different sequence: the sequence $x_1^*, y_1^*, x_2^*, y_2^*, \ldots$ resulting from D's decisions $D(x_n)$.

The approach of this paper is based on defensive forecasting: the ability to produce ideal, in important aspects, probability forecasts. It is interesting that ideal probability forecasts are not sufficient in big decision making. As a simple example, consider the game where there is no $\mathbf{X}$, $\Gamma = \mathbf{Y} = \{0, 1\}$, and the loss function λ is given by the matrix

	$y = 0$	$y = 1$
$\gamma = 0$	1	2
$\gamma = 1$	2	0

Reality's strategy is $y_n := \gamma_n$, but Predictor's initial theory is that Reality always chooses $y_n = 0$.

Predictor's "optimal" strategy based on his initial beliefs is to always choose $\gamma_n = 0$ suffering loss 1 at each step. His initial beliefs are reinforced with every move by Reality. Intuitively it is clear that Predictor's mistake in not choosing $\gamma_n \equiv 1$ is that he was being greedy (concentrated on exploitation and completely neglected exploration). However,

- he acted optimally given his beliefs,
- his beliefs have been verified by what actually happened.

In big decision making we have to worry about what would have happened if we had acted in a different way.

My hope is that game-theoretic probability has an important role to play in big decision making as well. A standard picture in the philosophy of science (see, e.g., [33, 25]) is that science progresses via struggle between (probabilistic) theories, and it is conceivable that something like this also happens in individual (human and animal) learning. Based on good theories (the ones that survives serious attempts to overthrow them) we can make good decisions. Testing of probabilistic theories is crucial in this process, and the game-theoretic version of the testing process (gambling against the theory) is much more flexible than the standard approach to testing statistical hypotheses: at each time we know to what degree the theory has been falsified. It is important, however, that the skeptic testing the theory should not only do this playing the imaginary game with the imaginary capital; he should also venture in the real world. Predictor's theory that Reality always chooses $y_n = 0$ would not survive for more than one round had it been tested (by choosing a sub-optimal, from the point of view of the old theory, decision).

Big decision making is a worthy goal but it is very difficult to prove anything about it, and elegant mathematical results might be beyond our reach for some time. Small decision making is also important but much easier; in many cases we can do it almost perfectly.

Acknowledgments

I am grateful to the COLT'2006 co-chairs for inviting me to give the talk on which this paper is based. Theorems 1 and 2 provide a partial answer to a question asked by Nicolò Cesa-Bianchi. This work was partially supported by MRC (grant S505/65).

References

[1] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, Amsterdam, second edition, 2003.

[2] Ravi P. Agarwal, Maria Meehan, and Donal O'Regan. *Fixed Point Theory and Applications*. Cambridge University Press, Cambridge, 2001.

[3] Nachman Aronszajn. La théorie générale des noyaux reproduisants et ses applications, première partie. *Proceedings of the Cambridge Philosophical Society*, 39:133–153 (additional note: p. 205), 1944. The second part of this paper is [4].

[4] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer, Boston, 2004.

[6] Patrick Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

[7] Nicolas Bourbaki. *Eléments de mathématique, Book VI, Intégration, Chapters 1–4.* Hermann, Paris, first edition, 1952.

[8] Haim Brezis and Petru Mironescu. Composition in fractional Sobolev spaces. *Discrete and Continuous Dynamical Systems*, 7:241–246, 2001.

[9] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, Cambridge, 2006.

[10] David R. Cox and David V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[11] A. Philip Dawid. Statistical theory: the prequential approach. *Journal of the Royal Statistical Society* A, 147:278–292, 1984.

[12] A. Philip Dawid. Self-calibrating priors do not exist: Comment. *Journal of the American Statistical Association*, 80:340–341, 1985. This is a contribution to the discussion in [32].

[13] A. Philip Dawid. Probability forecasting. In Samuel Kotz, Norman L. Johnson, and Campbell B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley, New York, 1986.

[14] Richard M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, England, 2002. Originally published in 1989.

[15] Ryszard Engelking. *General Topology*, volume 6 of *Sigma Series in Pure Mathematics.* Heldermann, Berlin, second edition, 1989. First edition: 1977 (Państwowe Wydawnictwo Naukowe, Warsaw).

[16] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

[17] Peter Gács. Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science*, 341:91–137, 2005.

[18] Gary Gruenhage. The story of a topological game. *Rocky Mountain Journal of Mathematics*, 2006. To appear.

[19] Yuri Kalnishkan and Michael V. Vyugin. The Weak Aggregating Algorithm and weak mixability. In Peter Auer and Ron Meir, editors, *Proceedings of the Eighteenth Annual Conference on Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 188–203, Berlin, 2005. Springer.

[20] Andrei N. Kolmogorov. Sur la loi des grands nombres. *Atti della Reale Accademia Nazionale dei Lincei. Classe di scienze fisiche, matematiche, e naturali. Rendiconti Serie VI*, 185:917–919, 1929.

[21] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin, 1933. English translation: *Foundations of the Theory of Probability.* Chelsea, New York, 1950.

[22] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.

[23] Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions of Information Theory*, IT-14:662–664, 1968.

[24] Andrei N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983.

[25] Thomas S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, Chicago, 1962. Third edition: 1996.

[26] Leonid A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17:337–340, 1976.

[27] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[28] Herbert Meschkowski. *Hilbertsche Räume mit Kernfunktion.* Springer, Berlin, 1962.

[29] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.

[30] Richard von Mises. *Wahrscheinlichkeit, Statistik, und Wahrheit.* Springer, Berlin, 1928. English translation (1939): *Probability, Statistics and Truth.* William Hodge, London.

[31] Jürgen K. Moser. A rapidly convergent iteration method and non-linear partial differential equations. *Annali della Scuola Normale Superiore di Pisa III*, 20:265–315, 1966. The second part of this paper, based on Moser's Pisa lectures, appears on pp. 499–535.

[32] David Oakes. Self-calibrating priors do not exist (with discussion). *Journal of the American Statistical Association*, 80:339–342, 1985.

[33] Karl R. Popper. *Logik der Forschung*. Springer, Vienna, 1934. An English translation, *The Logic of Scientific Discovery*, was published by Hutchinson, London, in 1959.

[34] Walter Rudin. *Functional Analysis*. McGraw-Hill, Boston, second edition, 1991.

[35] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[36] Glenn Shafer. The unity and diversity of probability. *Statistical Science*, 5:435–444, 1990.

[37] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[38] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

[39] Albert N. Shiryaev. *Probability*. Springer, New York, second edition, 1996. Third Russian edition published in 2004.

[40] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[41] Charles J. Stone. Consistent nonparametric regression (with discussion). *Annals of Statistics*, 5:595–645, 1977.

[42] Akimichi Takemura. Personal communication to Glenn Shafer. Takemura's observation was prompted by Shafer's talk at the University of Tokyo, May 2004.

[43] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[44] Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

[45] Vladimir Vovk. Aggregating strategies. In Mark Fulk and John Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

[46] Vladimir Vovk. Competitive on-line learning with a convex loss function. Technical Report `arXiv:cs.LG/0506041` (version 3), `arXiv.org` e-Print archive, September 2005.

[47] Vladimir Vovk. Non-asymptotic calibration and resolution. Technical Report `arXiv:cs.LG/0506004` (version 3), `arXiv.org` e-Print archive, August 2005.

[48] Vladimir Vovk. Leading strategies in competitive on-line learning. Submitted for publication, May 2006.

[49] Vladimir Vovk and Glenn Shafer. Good randomized sequential probability forecasting is always possible, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #7, June 2003 (revised September 2004). Also published in the *Journal of the Royal Statistical Society* B **67**, 747–763 (2005).

[50] Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. Technical Report `arXiv:cs.LG/0505083`, `arXiv.org` e-Print archive, May 2005.

[51] Vladimir Vovk and Vladimir V. V'yugin. On the empirical validity of the Bayesian method. *Journal of Royal Statistical Society* B, 55:253–266, 1993.

[52] Vladimir Vovk and Chris J. H. C. Watkins. Universal portfolio selection. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 12–23, New York, 1998. Association for Computing Machinery.

[53] Kôsaku Yosida. *Functional Analysis*, volume 123 of *Die Grundlehren der Mathematischen Wissenschaften*. Springer, New York, 1965. Sixth edition: 1980.