Competitive on-line learning with a convex loss function

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

February 1, 2008

Abstract

We consider the problem of sequential decision making under uncertainty in which the loss caused by a decision depends on the following binary observation. In competitive on-line learning, the goal is to design decision algorithms that are almost as good as the best decision rules in a wide benchmark class, without making any assumptions about the way the observations are generated. However, standard algorithms in this area can only deal with finite-dimensional (often countable) benchmark classes. In this paper we give similar results for decision rules ranging over an arbitrary reproducing kernel Hilbert space. For example, it is shown that for a wide class of loss functions (including the standard square, absolute, and log loss functions) the average loss of the master algorithm, over the first N observations, does not exceed the average loss of the best decision rule with a bounded norm plus $O(N^{-1/2})$. Our proof technique is very different from the standard ones and is based on recent results about defensive forecasting. Given the probabilities produced by a defensive forecasting algorithm, which are known to be well calibrated and to have good resolution in the long run, we use the expected loss minimization principle to find a suitable decision.

1 Introduction

In the simple problem of sequential decision making that we consider in this paper, the loss $\lambda(y_n, \gamma_n)$ (maybe negative) caused by a decision γ_n depends only on the following binary observation y_n. All relevant information available to the decision maker by the time he makes his decision is collected in what we call the datum, x_n. For example, in time series applications the datum may contain all or the most recent observations; in pattern recognition, where the observations are the true classes of patterns, the datum may be the vector of a pattern's attributes.

The traditional approach to this problem assumes a statistical model for the sequence of pairs (x_n, y_n); e.g., statistical learning theory ([21]) assumes that the (x_n, y_n) are generated independently from the same probability distribution. A more recent approach, known in learning theory as "prediction with expert advice" (e.g., [4]) and in information theory as "universal prediction" (e.g., [6, 14]), avoids making assumptions about the way the observations and data are generated. Instead, the goal of the decision maker is to compete with a more or less general benchmark class of decision rules, mapping the xs to the ys (the framework of prediction with expert advice is usually even more general). We will use the phrase "competitive on-line" to refer to this area (as in [23], emphasizing similarities to competitive on-line algorithms in computation theory).

First papers on competitive on-line learning with general loss functions (e.g., [4, 22]) dealt with countable (often finite) benchmark classes. The next step was to consider finite-dimensional benchmark classes (e.g., [8, 12, 23]). This paper continues with infinite-dimensional classes. (Such classes were considered earlier by Kimber and Long [11, 13], who, however, assumed that the benchmark class contains a perfect decision rule.) To get an idea of our central results, the reader is advised to start from Corollaries 1–3.

Our implicit assumption, common with other work in competitive on-line learning, is that the decision maker is "small": his decisions do not affect the future observations. This is not a mathematical assumption: as already mentioned, we do not make any assumptions at all about the way observations are generated; however, interpretation of our results becomes problematic if the decision maker is not small. "Big" decision makers can still use our algorithms for prediction (cf. Remarks 1 and 3 below).

In conclusion of this section we will briefly describe the content of the paper. Our main result is stated in §2, and several examples are given in §3. It is proved in §6; in §4 we describe the main ideas behind the proof and in §5 we prove some preparatory results for §6. Our decision algorithm is explicitly described in §7. We conclude with a short list of directions of further research (§8). A preliminary version of this paper is to appear as [24]. In this new version we made the title of the paper more specific (the old title was even somewhat misleading: in prediction with expert advice, the experts are usually completely free in making their decisions).

2 Main result

Our decision protocol is:

```
FOR $n = 1, 2, \dots$:
    Reality announces $x_n \in \mathbf{X}$.
    Decision Maker announces $\gamma_n \in \Gamma$.
    Reality announces $y_n \in \{0, 1\}$.
END FOR.
```

At each step (or *round*) n Decision Maker makes a decision γ_n whose consequences depend on the *observation* $y_n \in \{0, 1\}$ chosen by Reality. All relevant information available to Decision Maker by the time he makes his decision is collected in x_n, called the *datum.* We assume that the data x_n are elements of a *data space* $\mathbf{X}$ and that the decisions are elements of a *decision space* Γ (both sets assumed non-empty).

Remark 1 In this paper we are interested, first of all, in prediction of future observations. However, our framework allows a fairly wide class of loss functions, not all of which can be interpreted in terms of predictions (such as, e.g., Cover's and long-short games, in the terminology of [23], §2). This is the main reason why we prefer to talk about decision making in general; another reason is that in §5 we will deal with a very different kind of prediction (for which we reserve the term "forecasting").

A *decision strategy* is a strategy for Decision Maker in this protocol (explicitly defined specific strategies will also be called "decision algorithms"). Its performance is measured with a *loss function* $\lambda : \{0,1\} \times \Gamma \to \mathbb{R}$, and so its cumulative loss over the first N rounds is

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

The pair (Γ, λ) is the *game* being played. Decision Maker will compete against a class $\mathcal{F}$, called the *benchmark class*, of functions $D : \mathbf{X} \to \Gamma$ considered as decision rules; the cumulative loss suffered by such a decision rule is

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

Before stating our main result we define some useful notions connected with the two main components of our decision framework, the game (Γ, λ) and the benchmark class $\mathcal{F}$. The reader might want in parallel to read the next section, which describes some important examples of games and benchmark classes.

Games

The *exposure* $\operatorname{Exp}_\lambda(\gamma) \in \mathbb{R}$ of a decision $\gamma \in \Gamma$ is

$$\operatorname{Exp}_\lambda(\gamma) := \lambda(1,\gamma) - \lambda(0,\gamma)$$

and the *exposure* $\operatorname{Exp}_{\lambda,D} : \mathbf{X} \to \mathbb{R}$ of a decision rule D at a point $x \in \mathbf{X}$ is

$$\operatorname{Exp}_{\lambda,D}(x) := \lambda(1, D(x)) - \lambda(0, D(x)).$$

Let $\lambda(p, \gamma)$ be the expected loss caused by taking a decision γ when the probability of 1 is p:

$$\lambda(p,\gamma) := p\lambda(1,\gamma) + (1-p)\lambda(0,\gamma). \tag{1}$$

We only consider games (Γ, λ) such that

$$C_0 := \inf_{\gamma \in \Gamma} \lambda(0, \gamma), \quad C_1 := \inf_{\gamma \in \Gamma} \lambda(1, \gamma) \tag{2}$$

are finite. It is convenient (see, e.g., [9]) to summarize a game by its *superdecision set*

$$\Sigma := \left\{(x, y) \in \mathbb{R}^2 \,|\, \exists \gamma \in \Gamma : x \geq \lambda(0, \gamma) \text{ and } y \geq \lambda(1, \gamma)\right\}; \tag{3}$$

elements of this set will be called *superdecisions*. Superdecisions of the form $(\lambda(0, \gamma), \lambda(1, \gamma))$ will sometimes be called *decisions*. We will assume, additionally, that the set $\Sigma \subseteq \mathbb{R}^2$ is convex and closed. The *Eastern tail* of the game is the function

$$\begin{aligned} f : [C_0, \infty) &\to \mathbb{R} \cup \{\infty\} \\ x &\mapsto \inf\{y \,|(x, y) \in \Sigma\} - C_1 \end{aligned} \tag{4}$$

and its *Northern tail* is

$$\begin{aligned} g : [C_1, \infty) &\to \mathbb{R} \cup \{\infty\} \\ y &\mapsto \inf\{x \,|(x, y) \in \Sigma\} - C_0, \end{aligned} \tag{5}$$

where, as usual, $\inf \emptyset := \infty$; it is clear that f and g are nonnegative everywhere and finite on (C_0, ∞) and (C_1, ∞), respectively.

The theorem

A *reproducing kernel Hilbert space* (RKHS) on $\mathbf{X}$ is a Hilbert space $\mathcal{F}$ of real-valued functions on $\mathbf{X}$ such that the evaluation functional $f \in \mathcal{F} \mapsto f(x)$ is continuous for each $x \in \mathbf{X}$. By the Riesz–Fischer theorem, for each $x \in \mathbf{X}$ there exists a function $\mathbf{K}_x \in \mathcal{F}$ such that

$$f(x) = \langle \mathbf{K}_x, f \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}.$$

Let

$$\mathbf{c}_{\mathcal{F}} := \sup_{x \in \mathbf{X}} \|\mathbf{K}_x\|_{\mathcal{F}}; \tag{6}$$

we will be interested in the case $\mathbf{c}_{\mathcal{F}} < \infty$. With each game (Γ, λ) and each RKHS $\mathcal{F}$ we associate the non-negative (but maybe infinite) constant $\mathbf{c}_{\lambda,\mathcal{F}}$ defined by

$$\begin{aligned} \mathbf{c}_{\lambda,\mathcal{F}}^2 &:= \sup_{p \in (0,1)} \sup_{\gamma \in \Gamma_p} \sup_{x \in \mathbf{X}} p(1-p) \left(\mathrm{Exp}_\lambda^2(\gamma) + \|\mathbf{K}_x\|_{\mathcal{F}}^2\right) \\ &= \sup_{p \in (0,1)} \sup_{\gamma \in \Gamma_p} p(1-p) \left(\mathrm{Exp}_\lambda^2(\gamma) + \mathbf{c}_{\mathcal{F}}^2\right), \end{aligned} \tag{7}$$

where $\Gamma_p := \arg\min_{\gamma \in \Gamma} \lambda(p, \gamma)$ (and $\lambda(p, \gamma)$ is defined by (1)).

The following is our main result.

Theorem 1 *Let the game (Γ, λ) be such that (2) are finite, the superdecision set Σ is convex and closed, and the tails f and g satisfy*

$$f'_+(t) = O(t^{-2}), \quad g'_+(t) = O(t^{-2}) \tag{8}$$

as $t \to \infty$, where f'_+ and g'_+ stand for the right derivatives (see, e.g., [15], §23) of f and g. Let $\mathcal{F}$ be an RKHS on $\mathbf{X}$ and $\mathbf{c}_{\mathcal{F}}$, $\mathbf{c}_{\lambda,\mathcal{F}}$ be defined by (6) and (7). Suppose $\mathbf{c}_{\mathcal{F}} < \infty$. Then $\mathbf{c}_{\lambda,\mathcal{F}} < \infty$ and there is a decision strategy which guarantees that

$$\sum_{n=1}^{N} \lambda(y_n, \gamma_n) \le \sum_{n=1}^{N} \lambda(y_n, D(x_n)) + \mathbf{c}_{\lambda,\mathcal{F}} \left(\left\| \mathrm{Exp}_{\lambda,D} \right\|_{\mathcal{F}} + 1 \right) \sqrt{N} \tag{9}$$

for all $N = 1, 2, \dots$ and all $D : \mathbf{X} \to \Gamma$ with $\mathrm{Exp}_{\lambda,D} \in \mathcal{F}$.

Remark 2 If the loss function λ is bounded, (8) holds trivially. The right derivatives in (8) can be replaced by the corresponding left derivatives, since $\left|f'_+\right| \le \left|f'_-\right|$ and $\left|g'_+\right| \le \left|g'_-\right|$ (see, e.g., [15], Theorem 24.1). Condition (8) can be interpreted as saying that the tails should shrink fast enough. The case $f(t) = g(t) = t^{-1}$ can be considered borderline; Theorem 1 is still applicable in this case, but it ceases to be applicable for tails that shrink less fast.

3 Examples

In this section we first define a specific RKHS and then describe three important games.

Kernels as source of RKHS

We start by describing an equivalent language for talking about RKHS. The *kernel* of an RKHS $\mathcal{F}$ on $\mathbf{X}$ is

$$\mathbf{K}(x, x') := \langle \mathbf{K}_x, \mathbf{K}_{x'} \rangle_{\mathcal{F}}$$

(equivalently, we could define $\mathbf{K}(x, x')$ as $\mathbf{K}_x(x')$ or as $\mathbf{K}_{x'}(x)$). There is a simple internal characterization of the kernels $\mathbf{K}$ of RKHS.

It is easy to check that the function $\mathbf{K}(x, x')$, as we defined it, is symmetric ($\mathbf{K}(x, x') = \mathbf{K}(x', x)$ for all $x, x' \in \mathbf{X}$) and positive definite ($\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \mathbf{K}(x_i, x_j) \ge 0$ for all $m = 1, 2, \dots$, all $(\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$, and all $(x_1, \dots, x_m) \in \mathbf{X}^m$). On the other hand, for every symmetric and positive definite $\mathbf{K} : \mathbf{X}^2 \to \mathbb{R}$ there exists a unique RKHS $\mathcal{F}$ such that $\mathbf{K}$ is the kernel of $\mathcal{F}$ ([1], Théorème 2).

We can see that the notions of a kernel of RKHS and of a symmetric positive definite function on $\mathbf{X}^2$ have the same content, and we will sometimes say "kernel on $\mathbf{X}$" to mean a symmetric positive definite function on $\mathbf{X}^2$. Kernels in this sense are the main source of RKHS in learning theory; see, e.g., [21], [16], and

[18] for numerous examples. Every kernel on $\mathbf{X}$ is a valid parameter for our decision algorithm; to apply Theorem 1 we can use the equivalent definition of $\mathbf{c}_{\mathcal{F}}$,

$$\mathbf{c}_{\mathcal{F}} := \sup_{x\in\mathbf{X}} \sqrt{\mathbf{K}(x,x)}.$$

A long list of RKHS together with their kernels is given in [3], §7.4. For concreteness, in this section we will use the Sobolev space $\mathcal{S}$ of absolutely continuous functions f on $\mathbb{R}$ with finite norm

$$\|f\|_{\mathcal{S}} := \sqrt{\int_{-\infty}^{\infty} f^2(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} (f'(x))^2\,\mathrm{d}x}; \tag{10}$$

its kernel is

$$\mathbf{K}(x,x') = \frac{1}{2}\exp\left(-\left|x-x'\right|\right)$$

(see [20] or [3], §7.4, Example 24). From the last equation we can see that $\mathbf{c}_{\mathcal{S}} = 1/\sqrt{2}$.

The square loss game

For the square loss game, $\Gamma = [0,1]$ and $\lambda(y,\gamma) = (y-\gamma)^2$, and so we have

$$\mathrm{Exp}_\lambda(\gamma) = \lambda(1,\gamma) - \lambda(0,\gamma) = (1-\gamma)^2 - \gamma^2 = 1 - 2\gamma, \tag{11}$$

$$\lambda(p,\gamma) = p(1-\gamma)^2 + (1-p)\gamma^2 = p(1-p) + (\gamma-p)^2,$$

and

$$\Gamma_p = \{p\}. \tag{12}$$

Therefore,

$$\mathbf{c}_{\lambda,\mathcal{F}} = \begin{cases} \mathbf{c}_{\mathcal{F}}/2 & \text{if } \mathbf{c}_{\mathcal{F}} \ge 1 \\ (1+\mathbf{c}_{\mathcal{F}}^2)/4 & \text{if } \mathbf{c}_{\mathcal{F}} < 1; \end{cases}$$

in particular, $\mathbf{c}_{\lambda,\mathcal{S}} = 3/8$ for the Sobolev space (10), and Theorem 1 implies

Corollary 1 *Suppose the decision space is $\mathbf{X} = \mathbb{R}$. There is a decision strategy that guarantees that, for all N and all decision rules $D \in \mathcal{S}$,*

$$\sum_{n=1}^{N} (y_n - \gamma_n)^2 \le \sum_{n=1}^{N} (y_n - D(x_n))^2 + \frac{3}{8}\left(\|2D-1\|_{\mathcal{S}} + 1\right)\sqrt{N}$$

($2D-1$ is the decision rule "normalized" to take values in $[-1,1]$).

Remark 3 The games of this section illustrate Remark 1: here the decisions γ_n are best interpreted as predictions of y_n. Loss functions λ satisfying (12) are called *proper scoring rules*. Such loss functions "encourage honesty": it is optimal to predict with the true probability (provided it is known). We will later see another loss function of this type (the log loss function).

To illustrate Corollary 1, suppose there are constants $c > 1$ and $d > 1$ and a good absolutely continuous decision rule $D : \mathbb{R} \to [0,1]$ such that $|x_n| \le c$, $n = 1, 2, \ldots$, and $|D'(x)| \le d$ for all $x \in \mathbf{X}$. At rounds $N \gg cd^2$ the average loss of our decision algorithm will be almost as good as (or better than) the loss of D. We refrain from giving similar illustrations for the other corollaries in this section.

The absolute loss game

In this game, $\lambda(y, \gamma) = |y - \gamma|$ with $\Gamma = [0,1]$. We find:

$$\mathrm{Exp}_\lambda(\gamma) = \lambda(1, \gamma) - \lambda(0, \gamma) = (1 - \gamma) - \gamma = 1 - 2\gamma$$

(the same as in the square loss case, (11)),

$$\lambda(p, \gamma) = p(1 - \gamma) + (1 - p)\gamma = p + (1 - 2p)\gamma,$$

and

$$\Gamma_p = \begin{cases} \{0\} & \text{if } p < 1/2 \\ \{1\} & \text{if } p > 1/2 \\ [0,1] & \text{if } p = 1/2. \end{cases}$$

Therefore,

$$\mathbf{c}_{\lambda,\mathcal{F}} = \frac{1}{2}\sqrt{1 + \mathbf{c}_{\mathcal{F}}^2}$$

(in particular, $\mathbf{c}_{\lambda,\mathcal{S}} = \sqrt{6}/4$), and we have the following corollary of Theorem 1.

Corollary 2 *Let $\mathbf{X} = \mathbb{R}$. There is a decision strategy that produces decisions γ_n such that, for all N and all $D \in \mathcal{S}$,*

$$\sum_{n=1}^{N} |y_n - \gamma_n| \le \sum_{n=1}^{N} |y_n - D(x_n)| + \frac{\sqrt{6}}{4}\left(\|2D - 1\|_{\mathcal{S}} + 1\right)\sqrt{N}. \qquad (13)$$

The log loss game

For the log loss game, $\Gamma = (0, 1)$ and

$$\lambda(y, \gamma) = -y \ln \gamma - (1 - y)\ln(1 - \gamma).$$

For this game, $\mathbf{c}_{\lambda,\mathcal{F}} < \infty$ (assuming $\mathbf{c}_{\mathcal{F}} < \infty$) since its tails satisfy

$$f'(t) = g'(t) = -\frac{1}{e^t - 1} \sim -e^{-t} = O(t^{-2});$$

this will be also clear from the following direct calculation. Since

$$\mathrm{Exp}_\lambda(\gamma) = \lambda(1, \gamma) - \lambda(0, \gamma) = -\ln \gamma + \ln(1 - \gamma) = \ln \frac{1 - \gamma}{\gamma},$$

$$\lambda(p,\gamma) = -p\ln\gamma - (1-p)\ln(1-\gamma) = \lambda(p,p) + D(p,\gamma)$$

(where $D(p,\gamma) := p\ln\frac{p}{\gamma} + (1-p)\ln\frac{1-p}{1-\gamma}$ is the Kullback distance between p and γ, known to take its minimal value in γ at $\gamma = p$), and $\Gamma_p = \{p\}$, we can bound $\mathbf{c}_{\lambda,\mathcal{F}}$ from above as follows:

$$\begin{aligned}\mathbf{c}^2_{\lambda,\mathcal{F}} = \sup_{p\in(0,1)} p(1-p)\left(\left(\ln\frac{1-p}{p}\right)^2 + \mathbf{c}^2_{\mathcal{F}}\right)\\ \le \mathbf{c}^2_{\mathcal{F}}/4 + \sup_{p\in(0,1)} p(1-p)\left(\ln\frac{1-p}{p}\right)^2\\ \approx \mathbf{c}^2_{\mathcal{F}}/4 + 0.439 \le \mathbf{c}^2_{\mathcal{F}}/4 + 0.44.\end{aligned}$$

Of course, for specific values of $\mathbf{c}_{\mathcal{F}}$ it is better to find the $\sup_{p\in(0,1)}$ directly, without using this bound. Such a direct calculation shows that $\mathbf{c}_{\lambda,\mathcal{S}} \approx 0.693 \le 0.7$, and Theorem 1 now implies the following.

Corollary 3 *Some decision strategy in the log loss game with $\mathbf{X} = \mathbb{R}$ produces decisions γ_n such that, for all N and all $D : \mathbf{X} \to (0,1)$ with the log-likelihood ratio $\ln\frac{D}{1-D}$ in $\mathcal{S}$,*

$$\sum_{n=1}^N \lambda(y_n,\gamma_n) \le \sum_{n=1}^N \lambda(y_n, D(x_n)) + 0.7\left(\left\|\ln\frac{D}{1-D}\right\|_{\mathcal{S}} + 1\right)\sqrt{N}.$$

4 Idea of the proof of Theorem 1

This section describes the intuition behind the proof. The following sections, which carry out the proof, are formally independent of this section. We will also describe a general research program that may lead, it can be hoped, to many other results.

Game-theoretic probability

Our proof technique is based on a game-theoretic alternative to the standard measure-theoretic axioms of probability ([17]). Many of the standard laws of probability, including the weak and strong laws of large numbers, the central limit theorem, and the law of the iterated logarithm, can be restated in terms of perfect information games involving three key players: Reality, Forecaster, and Skeptic. A typical game-theoretic law of probability states that Skeptic has a strategy which, without risking bankruptcy, greatly enriches him if the law is violated. All such strategies for Skeptic were explicitly constructed continuous functions; game-theoretic laws of probability with a continuous strategy for Skeptic will be called "continuous laws of probability".

Game-theoretic probability as developed in [17] was to a large degree parallel to measure-theoretic probability. Following [7] and the literature that this paper spawned, paper [27] pointed out a surprising feature of game-theoretic probability: for any continuous law of probability, Forecaster has a strategy that prevents Skeptic's capital from growing (cf. Lemma 1 below). In other words, for any continuous law of probability there is a forecasting strategy that is perfect as far as this law is concerned (we will say "perfect relative to" this law). This result was obtained in [27] for binary forecasting, and in [26] it was extended to more general protocols. Forecasting strategies obtained in this way from various laws of probability were called "defensive forecasting" strategies.

General procedure

Now we are ready to describe a general procedure whose implementation leads, in the most straightforward case, to Theorem 1.

Choose a goal which could be achieved if you knew the true probabilities generating the observations. It is important that this goal should be "practical", in the sense of being stated in terms of observable quantities, such as data, decisions, and observations. The goal is not allowed to contain theoretical quantities, such as the true probabilities themselves, and it should be achievable no matter what the true probabilities are. Construct a decision strategy which, using the true probabilities, leads to the goal.

Realistically, however, we do not know the true probabilities. To get rid of them, isolate the law of probability on which the proof that your decision strategy achieves the goal depends; typically, this law can be stated as a continuous game-theoretic law of probability. (If the proof depends on several laws, they should first be merged into a single law.) There is a forecasting strategy whose forecasts are at least as good as (and often better than) the true probabilities, as far as the law you have just isolated is concerned. It remains to feed your decision strategy with those forecasts.

Implementing this procedure for various interesting goals appears to be a promising research program.

Introduction to the proof

In this paper our goal is to achieve (9), which we roughly rewrite as

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

where the informal notation $\lessapprox$ is used to mean that the left-hand side does not exceed the right-hand side plus a quantity small as compared to N. The goal is stated in terms of the observables.

Let us see how our goal could be achieved if we knew the true probabilities p_n that $y_n = 1$ (slightly more formally, p_n is the conditional probability that $y_n = 1$ given the available information). By the law of large numbers (see, e.g., [19], Theorem VII.5.4, for a suitable measure-theoretic statement and [17], Theorem 4.1, for its game-theoretic counterpart), we expect

$$\left|\sum_{n=1}^{N} f(p_n, x_n)(y_n - p_n)\right| \ll N \tag{14}$$

if f is a bounded function (assumed measurable in the measure-theoretic case). If f is allowed to range over a function class $\mathcal{F}$ that is not excessively wide, (14) will still continue to hold uniformly in f.

Suppose, for simplicity, that Γ_p is a singleton for all $p \in [0,1]$; the only element of Γ_p will be denoted $G(p)$. Our decision strategy will make the decision $G(p_n)$ at round n, i.e., the decision that leads to the smallest expected loss. We will sometimes say that G is our “choice function”.

Notice that

$$\lambda(y,\gamma) - \lambda(p,\gamma) = (y-p)\bigl(\lambda(1,\gamma) - \lambda(0,\gamma)\bigr)$$

always holds (this can be checked by subtracting (1) from $\lambda(y,\gamma) := y\lambda(1,\gamma) + (1-y)\lambda(0,\gamma)$). In conjunction with the law of large numbers (14) this implies

$$\begin{aligned}
\sum_{n=1}^{N} \lambda(y_n,\gamma_n) &= \sum_{n=1}^{N} \lambda(y_n, G(p_n))\\
&= \sum_{n=1}^{N} \lambda(p_n, G(p_n)) + \sum_{n=1}^{N} \bigl(\lambda(y_n, G(p_n)) - \lambda(p_n, G(p_n))\bigr)\\
&= \sum_{n=1}^{N} \lambda(p_n, G(p_n)) + \sum_{n=1}^{N} (y_n - p_n)\bigl(\lambda(1, G(p_n)) - \lambda(0, G(p_n))\bigr) \lessapprox \sum_{n=1}^{N} \lambda(p_n, G(p_n))\\
&\le \sum_{n=1}^{N} \lambda(p_n, D(x_n)) = \sum_{n=1}^{N} \lambda(y_n, D(x_n)) - \sum_{n=1}^{N} \bigl(\lambda(y_n, D(x_n)) - \lambda(p_n, D(x_n))\bigr)\\
&= \sum_{n=1}^{N} \lambda(y_n, D(x_n)) - \sum_{n=1}^{N} (y_n - p_n)\bigl(\lambda(1, D(x_n)) - \lambda(0, D(x_n))\bigr)\\
&\lessapprox \sum_{n=1}^{N} \lambda(y_n, D(x_n)).
\end{aligned} \tag{15}$$

This shows that we can achieve our goal if we know the true probabilities, and it remains to replace the true probabilities with the forecasts that are perfect relative to the law of large numbers.

For clarity, let us summarize the idea of the proof expressed by (15). To show that the actual loss of our decision strategy does not exceed the actual loss of a decision rule D by much, we notice that:

- the actual loss $\sum_{n=1}^{N} \lambda(y_n, G(p_n))$ of our decision strategy is approximately equal, by the law of large numbers, to the (one-step-ahead conditional) expected loss $\sum_{n=1}^{N} \lambda(p_n, G(p_n))$ of our strategy;

- since we used the expected loss minimization principle, the expected loss of our strategy does not exceed the expected loss of D;
- the expected loss $\sum_{n=1}^{N} \lambda(p_n, D(x_n))$ of D is approximately equal to its actual loss $\sum_{n=1}^{N} \lambda(y_n, D(x_n))$ (by the law of large numbers).

To get the strongest possible result, we will have to use more specific laws of probability than the general law of large numbers. It will be convenient to use the following informal terminology introduced in [25]. Let p_n be the forecasts output by some forecasting strategy (rather than the true probabilities). We say that the forecasting strategy has good *calibration-cum-resolution* if the left-hand side of (14) is much less than N for a relatively wide class of functions $f : [0,1] \times \mathbf{X} \to \mathbb{R}$ and large N. We say that the strategy has good *calibration* if

$$\left| \sum_{n=1}^{N} (y_n - p_n) f(p_n) \right| \ll N$$

for a wide class of functions $f : [0,1] \to \mathbb{R}$ and large N. Finally, we say that the strategy has good *resolution* if

$$\left| \sum_{n=1}^{N} (y_n - p_n) f(x_n) \right| \ll N$$

for a wide class of $f : \mathbf{X} \to \mathbb{R}$ and for large N. For a detailed discussion and examples, see [25].

Notice that in applying the law of large numbers to establishing the two approximate inequalities in (15) we need not general $f = f(p,x)$ but only $f = f(p)$ (known in advance) and $f = f(x)$. In particular, we only need calibration and resolution separately, not calibration-cum-resolution. These are the two specific probability laws we will be concerned with.

The requirement that Γ_p should always be a singleton (in fact, we will even need the function $G(p)$ to be continuous) is restrictive: for example, it is not satisfied for the absolute loss function. To deal with this problem, we will have to consider forecasting strategies that output extended forecasts $(p_n, q_n) \in [0,1]^2$, where p_n is the forecast of y_n and the extra component q_n will play a more technical role.

The next section is devoted to constructing a perfect forecasting strategy relative to the law of large numbers. In the following section we will be able to prove Theorem 1.

5 The algorithm of large numbers

This section is the core of our proof of Theorem 1. First we describe a forecasting protocol in which Forecaster tries to predict the observations chosen by Reality. Following [17], we introduce another player, Skeptic, who is allowed to bet at the odds implied by Forecaster's moves.

Binary Forecasting Game I
Players: Reality, Forecaster, Skeptic
Protocol:

```
FOR $n = 1, 2, \dots$:
    Reality announces $x_n \in \mathbf{X}$.
    Forecaster announces $(p_n, q_n) \in [0,1]^2$.
    Skeptic announces $s_n \in \mathbb{R}$.
    Reality announces $y_n \in \{0,1\}$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - p_n)$.
END FOR.
```

The real forecast is p_n (the "probability" that $y_n = 1$), which is interpreted as the price Forecaster charges for a ticket paying y_n; s_n is the number of tickets Skeptic decides to buy. The protocol describes not only the players' moves but also the changes in Skeptic's capital $\mathcal{K}_n$; its initial value $\mathcal{K}_0$ can be an arbitrary real number. Skeptic demonstrates that the forecasts are poor if he manages to multiply his initial capital (assumed positive) manyfold without risking bankruptcy (i.e., $\mathcal{K}_n$ becoming negative). Forecaster also provides an additional number $q_n \in [0,1]$ which does not affect Skeptic's capital; intuitively, the role of q_n is to help those of Forecaster's customers who find themselves in a position of Buridan's ass (find two or more actions equally attractive in view of the forecast p_n) to break the tie.

The main difference between our decision protocol (stated at the beginning of §2) and the protocols of this section is that in the latter Forecaster implicitly claims (by pricing the tickets) that he has the fullest possible knowledge of the way Reality chooses the observations, and Skeptic tries to prove him wrong by gambling against him. In the decision protocol, Decision Maker does no make any such claims and simply tries to minimize his losses.

It will be convenient to make the set $[0,1]^2$ from which the forecasts (p_n, q_n) are chosen into a topological space. The *lexicographic square* $\pounds$ is defined to be the set $[0,1]^2$ equipped with the following linear order: if (x_1, y_1) and (x_2, y_2) are two points in $\pounds$, $(x_1, y_1) < (x_2, y_2)$ means that either $x_1 < x_2$ or $x_1 = x_2, y_1 < y_2$. (Cf. [5], Problem 3.12.3(d).) The topology on the lexicographic square is, as usual, generated by the open intervals

$$(a, b) := \{u \in \pounds \mid a < u < b\},$$

a and b ranging over $\pounds$. As a topological space, the lexicographic square is normal ([5], Problem 1.7.4(d)), compact ([5], Problem 3.12.3(a), [10], Problem 5.C), and connected ([5], Problem 6.3.2(a), [10], Problem 1.I(d)).

As in [27], we will see that for any continuous strategy for Skeptic there exists a strategy for Forecaster that does not allow Skeptic's capital to grow, regardless of what Reality is doing. To state this observation in its strongest form, we make Skeptic announce his strategy for each round before Forecaster's move on that round rather than announce his full strategy at the beginning of the game. Therefore, we consider the following perfect-information game:

Binary Forecasting Game II
Players: Reality, Forecaster, Skeptic
Protocol:

```
FOR $n = 1, 2, \dots$:
    Reality announces $x_n \in \mathbf{X}$.
    Skeptic announces continuous $S_n : \pounds \to \mathbb{R}$.
    Forecaster announces $(p_n, q_n) \in \pounds$.
    Reality announces $y_n \in \{0, 1\}$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + S_n(p_n, q_n)(y_n - p_n)$.
END FOR.
```

Lemma 1 *Forecaster has a strategy in Binary Forecasting Game II that ensures* $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \geq \cdots$.

Before proving this lemma, we will need another lemma, which will play the role of the Intermediate Value Theorem, used in [25].

Lemma 2 *If a continuous function $f : \pounds \to \mathbb{R}$ takes both positive and negative values, there exists $x \in \pounds$ such that $f(x) = 0$.*

Proof A continuous image of a connected compact set is connected ([5], Theorem 6.1.4) and compact ([5], Theorem 3.1.10). Therefore, $f(\pounds)$ is a closed interval. ∎

Proof of Lemma 1 Forecaster can now use the following strategy to ensure $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \cdots$:

- if the function $S_n(p, q)$ takes value 0, choose (p_n, q_n) such that $S_n(p_n, q_n) = 0$;
- if S_n is always positive, take $p_n := 1$ and choose $q_n \in [0, 1]$ arbitrarily;
- if S_n is always negative, take $p_n := 0$ and choose $q_n \in [0, 1]$ arbitrarily. ∎

A kernel $\mathbf{K}$ on $\pounds \times \mathbf{X}$ is *forecast-continuous* if the function $\mathbf{K}((p, q, x), (p', q', x'))$ is continuous in $(p, q, p', q') \in \pounds^2$, for each fixed $(x, x') \in \mathbf{X}^2$. (Kernels on $\pounds \times \mathbf{X}$ are defined analogously to kernels on $\mathbf{X}$.) For such a kernel the function

$$S_n(p, q) := \sum_{i=1}^{n-1} \mathbf{K}((p, q, x_n), (p_i, q_i, x_i))(y_i - p_i) + \frac{1}{2}\mathbf{K}((p, q, x_n), (p, q, x_n))(1 - 2p) \tag{16}$$

is continuous in $(p, q) \in \pounds$.

The lexicographic algorithm of large numbers ($\pounds$ALN)
Parameter: forecast-continuous kernel $\mathbf{K}$ on $\pounds \times \mathbf{X}$

```
FOR $n = 1, 2, \dots$:
    Read $x_n \in \mathbf{X}$.
    Define $S_n(p,q)$ by (16), $(p,q) \in \pounds$.
    Output any root $(p,q)$ of $S_n(p,q) = 0$ as $(p_n, q_n)$;
        if there are no roots,
        set $p_n := (1 + \operatorname{sign} S_n)/2$ and set $q_n$ to any number in $[0,1]$.
    Read $y_n \in \{0,1\}$.
END FOR.
```

(Notice that $\operatorname{sign} S_n$ is well defined by Lemma 2.) It is well known that there exists a function $\Phi : \pounds \times \mathbf{X} \to \mathcal{H}$ (a *feature mapping* taking values in a Hilbert space $\mathcal{H}$) such that

$$\mathbf{K}(a,b) = \Phi(a) \cdot \Phi(b), \ \ \forall a, b \in \pounds \times \mathbf{X}. \tag{17}$$

(For example, we can take the RKHS on $\pounds \times \mathbf{X}$ with kernel $\mathbf{K}$ as $\mathcal{H}$ and take $a \mapsto \mathbf{K}_a$ as the feature mapping Φ; there are, however, easier and more transparent constructions.) It can be shown that $\Phi(p,q,x)$ is *forecast-continuous*, i.e., continuous in $(p,q) \in \pounds$ for each fixed $x \in \mathbf{X}$, if and only if the kernel $\mathbf{K}$ defined by (17) is forecast-continuous (see, e.g., [25], Appendix B).

Theorem 2 *Let $\mathbf{K}$ be the kernel defined by (17) for a forecast-continuous feature mapping $\Phi : \pounds \times \mathbf{X} \to \mathcal{H}$. The lexicographic algorithm of large numbers with parameter $\mathbf{K}$ outputs (p_n, q_n) such that*

$$\left\| \sum_{n=1}^{N} (y_n - p_n) \Phi(p_n, q_n, x_n) \right\|^2 \le \sum_{n=1}^{N} p_n (1 - p_n) \left\| \Phi(p_n, q_n, x_n) \right\|^2 \tag{18}$$

always holds for all $N = 1, 2, \ldots$.

Proof Following £ALN Forecaster ensures that Skeptic will never increase his capital with the strategy

$$s_n := \sum_{i=1}^{n-1} \mathbf{K}\left((p_n, q_n, x_n), (p_i, q_i, x_i)\right)(y_i - p_i) \\ + \frac{1}{2} \mathbf{K}\left((p_n, q_n, x_n), (p_n, q_n, x_n)\right)(1 - 2p_n). \tag{19}$$

Using the formula

$$(y_n - p_n)^2 = p_n(1 - p_n) + (1 - 2p_n)(y_n - p_n)$$

(which can be checked by setting $y_n := 0$ and $y_n := 1$), we can see that the increase in Skeptic's capital when he follows (19) is

$$
\begin{aligned}
\mathcal{K}_N - \mathcal{K}_0 &= \sum_{n=1}^{N} s_n(y_n - p_n) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{n-1} \mathbf{K}\left((p_n, q_n, x_n), (p_i, q_i, x_i)\right)(y_n - p_n)(y_i - p_i) \\
&\quad + \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, q_n, x_n), (p_n, q_n, x_n)\right)(1 - 2p_n)(y_n - p_n) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} \mathbf{K}\left((p_n, q_n, x_n), (p_i, q_i, x_i)\right)(y_n - p_n)(y_i - p_i) \\
&\quad - \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, q_n, x_n), (p_n, q_n, x_n)\right)(y_n - p_n)^2 \\
&\quad + \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, q_n, x_n), (p_n, q_n, x_n)\right)(1 - 2p_n)(y_n - p_n) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} \mathbf{K}\left((p_n, q_n, x_n), (p_i, q_i, x_i)\right)(y_n - p_n)(y_i - p_i) \\
&\quad - \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, q_n, x_n), (p_n, q_n, x_n)\right) p_n(1 - p_n) \\
&= \frac{1}{2} \left\| \sum_{n=1}^{N} (y_n - p_n)\Phi(p_n, q_n, x_n) \right\|^2 - \frac{1}{2} \sum_{n=1}^{N} p_n(1 - p_n) \left\| \Phi(p_n, q_n, x_n) \right\|^2,
\end{aligned}
$$

which immediately implies (18). ▮

Resolution

This subsection makes the next step in our proof of Theorem 1. Its forecasting protocol is:

```
FOR $n = 1, 2, \dots$:
    Reality announces $x_n \in \mathbf{X}$.
    Forecaster announces $(p_n, q_n) \in [0, 1]$.
    Reality announces $y_n \in \{0, 1\}$.
END FOR.
```

Our goal is to prove the following result (although in §6 we will need a slight modification of this result rather than the result itself).

Theorem 3 *Let $\mathcal{F}$ be an RKHS on $\mathbf{X}$. The forecasts (p_n, q_n) output by £ALN always satisfy*

$$\left|\sum_{n=1}^{N}(y_n - p_n)f(x_n)\right| \le \frac{\mathbf{c}_{\mathcal{F}}}{2}\left\|f\right\|_{\mathcal{F}}\sqrt{N}$$

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

Proof Applying ALN to the feature mapping $x \in \mathbf{X} \mapsto \mathbf{K}_x \in \mathcal{F}$ and using (18), we obtain

$$\begin{aligned}\left|\sum_{n=1}^{N}(y_n - p_n)f(x_n)\right| &= \left|\sum_{n=1}^{N}(y_n - p_n)\left\langle \mathbf{K}_{x_n}, f\right\rangle_{\mathcal{F}}\right| \\ &= \left|\left\langle\sum_{n=1}^{N}(y_n - p_n)\mathbf{K}_{x_n}, f\right\rangle_{\mathcal{F}}\right| \le \left\|\sum_{n=1}^{N}(y_n - p_n)\mathbf{K}_{x_n}\right\|_{\mathcal{F}}\left\|f\right\|_{\mathcal{F}} \\ &\le \left\|f\right\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}p_n(1-p_n)\mathbf{K}(x_n,x_n)} \le \mathbf{c}_{\mathcal{F}}\left\|f\right\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}p_n(1-p_n)} \qquad (20)\end{aligned}$$

for any $f \in \mathcal{F}$. ▮

Remark 4 In the terminology introduced in the previous section, Theorem 3 is about resolution. This is sufficient for the purpose of this paper, but it is easy to see that similar statements hold for calibration-cum-resolution and calibration. For example, let $\mathcal{F}$ be an RKHS on £ × $\mathbf{X}$. The forecasts (p_n, q_n) output by £ALN always satisfy

$$\left|\sum_{n=1}^{N}(y_n - p_n)f(p_n, q_n, x_n)\right| \le \frac{\mathbf{c}_{\mathcal{F}}}{2}\left\|f\right\|_{\mathcal{F}}\sqrt{N}$$

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

6 Proof of Theorem 1

Before starting the proof proper, we need to discuss two topics: choosing a suitable choice function and "mixing" different feature mappings.

The canonical choice function

Let us say that a straight line $(1-p)x + py = c$ in the (x,y)-plane, where $p \in [0,1]$ and $c \in \mathbb{R}$, is *southwest of* the superdecision set Σ (defined by (3)) if

$$\forall(x,y) \in \Sigma : (1-p)x + py \ge c.$$

For each $p \in [0,1]$ let $c(p)$ be the largest c (which obviously exists) such that the line $(1-p)x + py = c$ is southwest of Σ. It is clear that, for $p \in (0,1)$, the line $(1-p)x + py = c(p)$ intersects Σ and the intersection, being compact and convex, has the form $[A(p), B(p)]$, where $A(p)$ and $B(p)$ are points (perhaps $A(p) = B(p)$) on the line. For concreteness, let $A(p)$ be northwest of $B(p)$ (i.e., if $A(p) = (A_0, A_1)$ and $B(p) = (B_0, B_1)$, we assume that $A_0 \leq B_0$ and $A_1 \geq B_1$). Now we can define the *canonical choice function* G associated with (Γ, λ) as follows:

- if $0 < p < 1$ and $q \in [0,1]$, $G(p,q)$ is defined to be any $\gamma \in \Gamma$ satisfying
$$(\lambda(0,\gamma), \lambda(1,\gamma)) = (1-q)A(p) + qB(p);$$
the existence of such a γ is obvious;
- if $p = 0$ and $q \in [0,1]$, $G(p,q)$ is defined to be any fixed $\gamma_0 \in \Gamma$ satisfying
$$(\lambda(0,\gamma_0), \lambda(1,\gamma_0)) = (C_0, f(C_0))$$
(C_0 and f are defined in (2) and (4)); if $f(C_0) = \infty$, such a γ_0 does not exist and $G(p,q)$ is undefined;
- if $p = 1$ and $q \in [0,1]$, $G(p,q)$ is defined to be any fixed $\gamma_1 \in \Gamma$ such that
$$(\lambda(0,\gamma_1), \lambda(1,\gamma_1)) = (g(C_1), C_1)$$
(C_1 and g are defined in (2) and (5)); if $g(C_1) = \infty$, such a γ_1 does not exist and $G(p,q)$ is undefined.

It is easy to see that the function $(\lambda(0, G(p,q)), \lambda(1, G(p,q)))$ is continuous in $(p,q) \in \operatorname{dom} G$ and, therefore, $\operatorname{Exp}_{\lambda,G}(p,q) := \operatorname{Exp}_\lambda(G(p,q))$ is continuous in $(p,q) \in \operatorname{dom} G$. We defined G in such a way that it is a "perfect" choice function: $\lambda(p, G(p,q)) = \inf_{\gamma\in\Gamma} \lambda(p,\gamma)$ for virtually all (p,q) (in any case, for all $(p,q) \in \operatorname{dom} G$).

Mixing

In the proof of Theorem 1 we will mix the feature mapping $\Phi_0(p,q,x) := \operatorname{Exp}_{\lambda,G}(p,q)$ (into $\mathcal{H}_0 := \mathbb{R}$) and the feature mapping $\Phi_1(p,q,x) := \mathbf{K}_x$ used in the proof of Theorem 3 (as discussed in §4, we will have to achieve two goals simultaneously, only one of them connected with resolution). This can be done using the following corollary of Theorem 2.

Corollary 4 *Let $\Phi_j : \pounds \times \mathbf{X} \to \mathcal{H}_j$, $j = 0,1$, be forecast-continuous mappings from $\pounds \times \mathbf{X}$ to Hilbert spaces $\mathcal{H}_j$. The forecasts output by $\pounds$ALN with a suitable kernel parameter always satisfy*

$$\left\| \sum_{n=1}^N (y_n - p_n) \Phi_j(p_n, q_n, x_n) \right\|_{\mathcal{H}_j}^2 \\ \leq \sum_{n=1}^N p_n(1-p_n) \left(\|\Phi_0(p_n,q_n,x_n)\|_{\mathcal{H}_0}^2 + \|\Phi_1(p_n,q_n,x_n)\|_{\mathcal{H}_1}^2 \right)$$

for all N and for both $j = 0$ and $j = 1$.

Proof Define the 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

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

Now we can define $\Phi : \pounds \times \mathbf{X} \to \mathcal{H}$ by

$$\Phi(p, q, x) := (\Phi_0(p, q, x), \Phi_1(p, q, x)) ;$$

the corresponding kernel is

$$\begin{aligned}\mathbf{K}((p, q, x), (p', q', x')) &:= \langle \Phi(p, q, x), \Phi(p', q', x') \rangle_{\mathcal{H}} \\ &= \sum_{j=0}^{1} \langle \Phi_j(p, q, x), \Phi_j(p', q', x') \rangle_{\mathcal{H}_j} = \sum_{j=0}^{1} \mathbf{K}_j((p, q, x), (p', q', x')),\end{aligned}$$

where $\mathbf{K}_0$ and $\mathbf{K}_1$ are the kernels corresponding to Φ_0 and Φ_1, respectively. It is clear that this kernel is forecast-continuous. Applying $\pounds$ALN to it and using (18), we obtain

$$\begin{aligned}&\left\| \sum_{n=1}^{N} (y_n - p_n) \Phi_j(p_n, q_n, x_n) \right\|_{\mathcal{H}_j}^2 \\ &\quad \le \left\| \left(\sum_{n=1}^{N} (y_n - p_n) \Phi_0(p_n, q_n, x_n), \sum_{n=1}^{N} (y_n - p_n) \Phi_1(p_n, q_n, x_n) \right) \right\|_{\mathcal{H}}^2 \\ &\quad = \left\| \sum_{n=1}^{N} (y_n - p_n) \Phi(p_n, q_n, x_n) \right\|_{\mathcal{H}}^2 \le \sum_{n=1}^{N} p_n(1 - p_n) \left\| \Phi(p_n, q_n, x_n) \right\|_{\mathcal{H}}^2 \\ &\quad = \sum_{n=1}^{N} p_n(1 - p_n) \sum_{j=0}^{1} \left\| \Phi_j(p_n, q_n, x_n) \right\|_{\mathcal{H}_j}^2 . \quad \blacksquare\end{aligned}$$

Merging Φ_0 and Φ_1 by Corollary 4, we obtain

$$\begin{aligned}\left| \sum_{n=1}^{N} (y_n - p_n) \operatorname{Exp}_{\lambda,G}(p_n, q_n) \right| &= \left\| \sum_{n=1}^{N} (y_n - p_n) \Phi_0(p_n, q_n, x_n) \right\|_{\mathbb{R}} \\ &\le \sqrt{\sum_{n=1}^{N} p_n(1 - p_n) \left(\operatorname{Exp}_{\lambda,G}^2(p_n, q_n) + \mathbf{K}(x_n, x_n) \right)} \qquad (21)\end{aligned}$$

and, using (20),

$$\begin{aligned}\left|\sum_{n=1}^{N}(y_n-p_n)f(x_n)\right| &\le \left\|\sum_{n=1}^{N}(y_n-p_n)\mathbf{K}_{x_n}\right\|_{\mathcal{F}}\|f\|_{\mathcal{F}}\\ &= \left\|\sum_{n=1}^{N}(y_n-p_n)\Phi_1(p_n,q_n,x_n)\right\|_{\mathcal{F}}\|f\|_{\mathcal{F}}\\ &\le \|f\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}p_n(1-p_n)\left(\mathrm{Exp}^2_{\lambda,G}(p_n,q_n)+\mathbf{K}(x_n,x_n)\right)}, \qquad (22)\end{aligned}$$

for each function $f \in \mathcal{F}$.

Proof: Part I

In this subsection we will assume that $\operatorname{dom} G = \pounds$. Subtracting (1) from $\lambda(y,\gamma) = y\lambda(1,\gamma) + (1-y)\lambda(0,\gamma)$, we obtain

$$\lambda(y,\gamma) - \lambda(p,\gamma) = (y-p)\big(\lambda(1,\gamma)-\lambda(0,\gamma)\big) = (y-p)\,\mathrm{Exp}_\lambda(\gamma) \qquad (23)$$

(we already did this in §4, but we promised that the rest of the paper would be formally independent of §4). Using the last equality and (21)–(22), we obtain for the decision strategy $\gamma_n := G(p_n,q_n)$ based on the (p_n,q_n) output by £ALN with the merged kernel as parameter:

$$\begin{aligned}&\sum_{n=1}^{N}\lambda(y_n,\gamma_n) = \sum_{n=1}^{N}\lambda(y_n,G(p_n,q_n))\\ &= \sum_{n=1}^{N}\lambda(p_n,G(p_n,q_n)) + \sum_{n=1}^{N}\big(\lambda(y_n,G(p_n,q_n)) - \lambda(p_n,G(p_n,q_n))\big)\\ &= \sum_{n=1}^{N}\lambda(p_n,G(p_n,q_n)) + \sum_{n=1}^{N}(y_n-p_n)\,\mathrm{Exp}_{\lambda,G}(p_n,q_n)\\ &\le \sum_{n=1}^{N}\lambda(p_n,G(p_n,q_n)) + \sqrt{\sum_{n=1}^{N}p_n(1-p_n)\left(\mathrm{Exp}^2_{\lambda,G}(p_n,q_n)+\mathbf{K}(x_n,x_n)\right)}\\ &\le \sum_{n=1}^{N}\lambda(p_n,D(x_n)) + \sqrt{\sum_{n=1}^{N}p_n(1-p_n)\left(\mathrm{Exp}^2_{\lambda,G}(p_n,q_n)+\mathbf{K}(x_n,x_n)\right)}\\ &= \sum_{n=1}^{N}\lambda(y_n,D(x_n)) - \sum_{n=1}^{N}\big(\lambda(y_n,D(x_n)) - \lambda(p_n,D(x_n))\big)\\ &\quad + \sqrt{\sum_{n=1}^{N}p_n(1-p_n)\left(\mathrm{Exp}^2_{\lambda,G}(p_n,q_n)+\mathbf{K}(x_n,x_n)\right)}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} \lambda(y_n, D(x_n)) - \sum_{n=1}^{N} (y_n - p_n) \operatorname{Exp}_{\lambda,D}(x_n)\big) \\
&\quad + \sqrt{\sum_{n=1}^{N} p_n(1-p_n) \left(\operatorname{Exp}^2_{\lambda,G}(p_n,q_n) + \mathbf{K}(x_n,x_n)\right)} \\
&\le \sum_{n=1}^{N} \lambda(y_n, D(x_n)) + \left\|\operatorname{Exp}_{\lambda,D}\right\|_{\mathcal{F}} \sqrt{\sum_{n=1}^{N} p_n(1-p_n) \left(\operatorname{Exp}^2_{\lambda,G}(p_n,q_n) + \mathbf{K}(x_n,x_n)\right)} \\
&\quad + \sqrt{\sum_{n=1}^{N} p_n(1-p_n) \left(\operatorname{Exp}^2_{\lambda,G}(p_n,q_n) + \mathbf{K}(x_n,x_n)\right)} \\
&= \sum_{n=1}^{N} \lambda(y_n, D(x_n)) + \left(\left\|\operatorname{Exp}_{\lambda,D}\right\|_{\mathcal{F}} + 1\right) \sqrt{\sum_{n=1}^{N} p_n(1-p_n) \left(\operatorname{Exp}^2_{\lambda,G}(p_n,q_n) + \mathbf{K}(x_n,x_n)\right)} \\
&\le \sum_{n=1}^{N} \lambda(y_n, D(x_n)) + \left(\left\|\operatorname{Exp}_{\lambda,D}\right\|_{\mathcal{F}} + 1\right) \mathbf{c}_{\lambda,\mathcal{F}} \sqrt{N}.
\end{aligned}
$$

It remains to show that $\mathbf{c}_{\lambda,\mathcal{F}} < \infty$ (assuming $\mathbf{c}_{\mathcal{F}} < \infty$, here and in the rest of this section). In this case, $\operatorname{dom} G = \pounds$, this is easy: essentially, this is the case of a bounded loss function (the reservation "essentially" is needed since Γ can contain "litter"—decisions dominated by other decisions in Γ). Since $\operatorname{Exp}_{\lambda,G}$ is continuous and $\pounds$ is compact,

$$\sup_{(p,q)\in\pounds} p(1-p) \left(\operatorname{Exp}^2_{\lambda,G}(p,q) + \mathbf{c}^2_{\mathcal{F}}\right) < \infty.$$

Proof: Part II

The *stripped lexicographic square* is the subset

$${}^{\dagger}\pounds^{\dagger} := (0,1) \times [0,1]$$

of $\pounds$. In this subsection we consider the case $\operatorname{dom} G = {}^{\dagger}\pounds^{\dagger}$.

The order and topology on ${}^{\dagger}\pounds^{\dagger}$ are inherited from $\pounds$. The following analogue of Lemma 2 still holds.

Lemma 3 *If a continuous function $f : {}^{\dagger}\pounds^{\dagger} \to \mathbb{R}$ takes both positive and negative values, it also takes the value* 0.

Proof See the proof of Lemma 2; $f({}^{\dagger}\pounds^{\dagger})$ is still a connected set in $\mathbb{R}$. ∎

A kernel $\mathbf{K}$ on ${}^{\dagger}\pounds^{\dagger} \times \mathbf{X}$ is *forecast-continuous* if the function $\mathbf{K}\left((p,q,x),(p',q',x')\right)$ is continuous in $(p,q,p',q') \in ({}^{\dagger}\pounds^{\dagger})^2$. The function (16) is then continuous in $(p,q) \in {}^{\dagger}\pounds^{\dagger}$, and for our current kernel

$$\mathbf{K}\left((p,q,x),(p',q',x')\right) = \operatorname{Exp}_{\lambda,G}(p,q) \operatorname{Exp}_{\lambda,G}(p',q') + \langle \mathbf{K}_x, \mathbf{K}_{x'} \rangle_{\mathcal{F}} \qquad (24)$$

it equals

$$\begin{aligned} S_n(p,q) = \sum_{i=1}^{n-1} \Big(&\operatorname{Exp}_{\lambda,G}(p,q) \operatorname{Exp}_{\lambda,G}(p_i,q_i) + \langle \mathbf{K}_{x_n}, \mathbf{K}_{x_i} \rangle_{\mathcal{F}} \Big)(y_i - p_i) \\ &+ \frac{1}{2}\Big(\operatorname{Exp}^2_{\lambda,G}(p,q) + \|\mathbf{K}_{x_n}\|^2_{\mathcal{F}} \Big)(1-2p) \\ &= A \operatorname{Exp}_{\lambda,G}(p,q) + B + \frac{1}{2} \operatorname{Exp}^2_{\lambda,G}(p,q)(1-2p) + Cp, \end{aligned} \tag{25}$$

where A, B, and C do not depend on (p,q). Since $\operatorname{dom} G = {}^{\dagger}\pounds^{\dagger}$, $\left|\operatorname{Exp}_{\lambda,G}(p,q)\right| \to \infty$ as $p \to 0$ or $p \to 1$, and so

$$\lim_{\substack{(p,q)\to(1,0)\\(p,q)\in{}^{\dagger}\pounds^{\dagger}}} S_n(p,q) = -\infty \tag{26}$$

and

$$\lim_{\substack{(p,q)\to(0,1)\\(p,q)\in{}^{\dagger}\pounds^{\dagger}}} S_n(p,q) = \infty. \tag{27}$$

The *stripped lexicographic ALN* (or, briefly, ${}^{\dagger}\pounds^{\dagger}$ALN) is defined as the lexicographic ALN except that:

- its parameter is a forecast-continuous kernel $\mathbf{K}$ on ${}^{\dagger}\pounds^{\dagger} \times \mathbf{X}$;
- it outputs a root (p,q) (an element of ${}^{\dagger}\pounds^{\dagger} = \operatorname{dom} S_n$) of the equation $S_n(p,q) = 0$ as (p_n,q_n) and crashes if this equation does not have roots (this will never happen for the kernel (24)).

Because of (26) and (27), ${}^{\dagger}\pounds^{\dagger}$ALN applied to the kernel (24) on ${}^{\dagger}\pounds^{\dagger} \times \mathbf{X}$ still ensures that (18) holds for our feature mapping (Φ_0, Φ_1); this algorithm never crashes and, of course, never outputs (p_n,q_n) with $p_n \in \{0,1\}$. We can see that the proof of (9) given in the previous subsection still works.

Let us now prove that $\mathbf{c}_{\lambda,\mathcal{F}} < \infty$ when $\operatorname{dom} G = {}^{\dagger}\pounds^{\dagger}$. It suffices to check that

$$\limsup_{\substack{(p,q)\to(0,1)\\(p,q)\in{}^{\dagger}\pounds^{\dagger}}} p \operatorname{Exp}^2_{\lambda,G}(p,q) < \infty \tag{28}$$

and

$$\limsup_{\substack{(p,q)\to(1,0)\\(p,q)\in{}^{\dagger}\pounds^{\dagger}}} (1-p) \operatorname{Exp}^2_{\lambda,G}(p,q) < \infty. \tag{29}$$

For example, let us demonstrate (29). Without loss of generality, we replace (8) with

$$f'_-(t) = O(t^{-2}), \quad g'_-(t) = O(t^{-2}) \tag{30}$$

(this can be done since $f'_-(t) \le f'_+(t) \le f'_-(t+1)$ and $g'_-(t) \le g'_+(t) \le g'_-(t+1)$). Consider the decision

$$(X,Y) := (\lambda(0, G(p,q)), \lambda(1, G(p,q))) .$$

Since $-\frac{1-p}{p}$ is a subgradient (see, e.g., [15], Section 23) of $f(x)$ at X, (30) implies that $1-p = O(X^{-2})$, i.e., $(1-p)X^2 = O(1)$. Since $|\mathrm{Exp}_{\lambda,G}(p,q)| = X - Y \le X - C_1$ for $(p,q) < (1,0)$ sufficiently close to $(1,0)$, (29) indeed holds.

Proof: Part III

In this subsection we consider the remaining possibilities for $\operatorname{dom} G$. Let us define the *left-stripped lexicographic ALN* (${}^\dagger\pounds$ALN for brief) as the lexicographic ALN except that:

- its parameter is a forecast-continuous kernel $\mathbf{K}$ on ${}^\dagger\pounds \times \mathbf{X}$, where the *left-stripped lexicographic square*

  $$ {}^\dagger\pounds := (0,1] \times [0,1] $$

  is equipped with the order and topology inherited from $\pounds$;
- it outputs a root $(p,q) \in {}^\dagger\pounds$ of the equation $S_n(p,q) = 0$ as (p_n, q_n); if this equation does not have roots in ${}^\dagger\pounds$, we set $p_n := 1$ and set $q_n \in [0,1]$ arbitrarily (we will make sure that this happens only when S_n is everywhere positive).

In a similar way we define the *right-stripped lexicographic square* $\pounds^\dagger$ and the *right-stripped lexicographic ALN* ($\pounds^\dagger$ALN), which always outputs $(p_n, q_n) \in \pounds^\dagger$; when $S_n(p,q) = 0$ does not have roots $(p,q) \in \pounds^\dagger$ we now set $p_n := 0$.

We only consider the case $\operatorname{dom} G = {}^\dagger\pounds$ (the case $\operatorname{dom} G = \pounds^\dagger$ is treated analogously); this corresponds to $f(C_0) = \infty$ and $f(C_1) < \infty$. Since S_n is continuous, the absence of roots of $S_n = 0$ in ${}^\dagger\pounds$ in conjunction with (27) means that S_n is positive everywhere on ${}^\dagger\pounds$, and so setting $p_n := 1$ in this case guarantees that ${}^\dagger\pounds$ALN still ensures (18). It remains to notice that (28) still holds.

7 The algorithm

In this short section we extract the decision strategy achieving (9) from our proof of Theorem 1. As we have already noticed (see (25)),

$$ S_n(p,q) = \sum_{i=1}^{n-1} \Bigl(\mathrm{Exp}_{\lambda,G}(p,q)\,\mathrm{Exp}_{\lambda,G}(p_i,q_i) + \mathbf{K}(x_n,x_i) \Bigr)(y_i - p_i) \\ + \frac{1}{2}\Bigl(\mathrm{Exp}^2_{\lambda,G}(p,q) + \mathbf{K}(x_n,x_n) \Bigr)(1-2p); \quad (31) $$

this immediately leads to the following explicit description.

An algorithm achieving (9)

Parameters: game with loss function λ and canonical choice function G;
kernel $\mathbf{K}$ on $\mathbf{X}$

```
FOR $n = 1, 2, \dots$:
    Read $x_n \in \mathbf{X}$.
    Define $S_n(p,q)$ by (31) for all $(p,q) \in \pounds$ for which $G(p,q)$ is defined.
    Define $(p_n, q_n)$ as any root $(p,q)$ of $S_n(p,q) = 0$;
        if there are no roots,
        set $p_n := (1 + \operatorname{sign} S_n)/2$ and set $q_n$ to any number in $[0,1]$.
    Set $\gamma_n := G(p_n, q_n)$.
    Read $y_n \in \{0,1\}$.
END FOR.
```

(We saw in the previous section that $\operatorname{sign} S_n$ is well defined and is -1 or 1 in this context.)

The canonical choice functions for the three examples of games given in §3 are as follows: $G(p,q) = p$ for the square loss and log loss games, and

$$G(p,q) = \begin{cases} 0 & \text{if } p < 1/2 \\ 1 & \text{if } p > 1/2 \\ q & \text{if } p = 1/2 \end{cases} \tag{32}$$

for the absolute loss game.

8 Directions of further research

In this section we discuss informally what we consider to be interesting directions of further research.

Non-convex games

Theorem 1 assumes that the superdecision set is convex. The assumption of convexity is convenient but not indispensable. We will only discuss the simplest non-convex game.

The loss function for the *simple loss game* is the same as for the absolute loss game, $\lambda(y,\gamma) = |y - \gamma|$, but $\Gamma = \{0,1\}$. Now the approach we have used in this paper does not work: since Γ consists of two elements, there is no non-trivial continuous choice function $G : \pounds \to \Gamma$ (every continuous image of $\pounds$ is connected: [5], Theorem 6.1.4).

A natural idea ([4]) is to allow Decision Maker to use randomization. The expected loss of a strategy making decision 1 with probability γ and 0 with probability $1 - \gamma$ is $|y - \gamma|$, where y is the actual observation; therefore, for the simple loss game a randomized decision strategy can guarantee the following analogue of (13):

$$\sum_{n=1}^{N} \mathbb{E}|y_n - \gamma_n| \le \sum_{n=1}^{N} |y_n - D(x_n)| + \frac{\sqrt{6}}{4} \left(\|2D - 1\|_{\mathcal{S}} + 1 \right) \sqrt{N}, \tag{33}$$

where $\mathbb{E}$ refers to the strategy's internal randomization (the decision rules D can be allowed to take values in $[0,1]$).

The disadvantage of (33) is that typically we are interested in the strategy's actual rather than expected loss. Our derivation of (33) shows the role of randomization: with our choice function (32) no randomization is required unless $p = 1/2$. Typically, we rarely find ourselves in a situation of complete uncertainty, $p_n = 1/2$; therefore, only a little bit of randomization is needed, essentially for tie breaking. The actual loss will be very close to the expected loss. It would be interesting to derive formal statements along these lines.

Non-binary observations

It would also be interesting to extend this paper's results to more general observation spaces (first of all, to carry them over to least-squares regression and multi-class classification). The two apparent obstacles to such extensions are that the fundamental equality (23) looks tailored to the binary case $y \in \{0,1\}$ and that Lemma 1 ceases to be obvious outside the binary case. However, (23) only states, in the terminology of [17], that $\lambda(p,\gamma)$ is the game-theoretic expected value of $\lambda(y,\gamma)$ (and that reproducing $\lambda(y,\gamma)$ given $\lambda(p,\gamma)$ can be accomplished by buying $\lambda(1,\gamma) - \lambda(0,\gamma)$ tickets paying y and costing p each). Similar equalities hold for many other forecasting protocols. And an analogue of Lemma 1 for a wide class of forecasting protocols is proved in [26].

Optimality

An important problem is to investigate the optimality of our algorithm, described in §7: is the bound (9) tight? (The tightness of the bounds in Theorem 2 and Equation (20) is established in [25].)

Acknowledgments

This work was partially supported by MRC (grant S505/65) and Royal Society.

References

[1] 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 [2].

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

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

[4] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44:427–485, 1997.

[5] Ryszard Engelking. *General Topology.* Heldermann, Berlin, second edition, 1989.

[6] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

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

[8] Yoav Freund. Predicting a binary sequence almost as well as the optimal biased coin. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 89–98, New York, 1996. Association for Computing Machinery.

[9] 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.

[10] John L. Kelley. *General Topology.* Van Nostrand, Princeton, NJ, 1957.

[11] Don Kimber and Philip M. Long. On-line learning of smooth functions of a single variable. *Theoretical Computer Science*, 148:141–156, 1995.

[12] Jyrki Kivinen and Manfred K. Warmuth. Exponential Gradient versus Gradient Descent for linear predictors. *Information and Computation*, 132:1–63, 1997.

[13] Philip M. Long. Improved bounds about on-line learning of smooth functions of a single variable. *Theoretical Computer Science*, 241:25–35, 2000.

[14] Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[15] R. Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

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

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

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

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

[20] Christine Thomas-Agnan. Computing a family of reproducing kernels for statistical applications. *Numerical Algorithms*, 13:21–32, 1996.

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

[22] 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.

[23] Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[24] Vladimir Vovk. Defensive forecasting with expert advice. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Artificial Intelligence*, pages 444–458, Berlin, 2005. Springer. To appear. Full version: Technical report `arXiv:cs.LG/0506041` (version 2), `arXiv.org` e-Print archive, July 2005.

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

[26] Vladimir Vovk, Ilia Nouretdinov, Akimichi Takemura, and Glenn Shafer. Defensive forecasting for linear protocols. Technical Report `arXiv:cs.LG/0506007`, `arXiv.org` e-Print archive, June 2005.

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