Leading strategies
in competitive on-line prediction

Vladimir Vovk
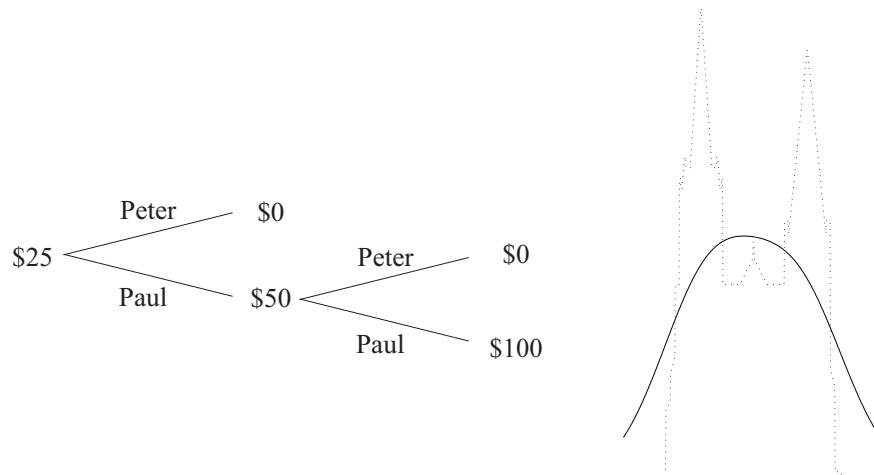
Peter — $0

$25

Paul — $50

Peter — $0

Paul — $100

Abstract

We start from a simple asymptotic result for the problem of on-line regression with the quadratic loss function: the class of continuous limited-memory prediction strategies admits a "leading prediction strategy", which not only asymptotically performs at least as well as any continuous limited-memory strategy but also satisfies the property that the excess loss of any continuous limited-memory strategy is determined by how closely it imitates the leading strategy. More specifically, for any class of prediction strategies constituting a reproducing kernel Hilbert space we construct a leading strategy, in the sense that the loss of any prediction strategy whose norm is not too large is determined by how closely it imitates the leading strategy. This result is extended to the loss functions given by Bregman divergences and by strictly proper scoring rules.

Contents

1   Introduction

Suppose $\mathcal{F}$ is a normed function class of prediction strategies (the "benchmark
class"). It is well known that, under some restrictions on $\mathcal{F}$, there exists a
"master prediction strategy" (sometimes also called a "universal strategy") that
performs almost as well as the best strategies in $\mathcal{F}$ whose norm is not too large
(see, e.g., [9, 5]). The "leading prediction strategies" constructed in this paper
satisfy a stronger property: the loss of any prediction strategy in $\mathcal{F}$ whose norm
is not too large exceeds the loss of a leading strategy by the divergence between
the predictions output by the two prediction strategies. Therefore, the leading
strategy implicitly serves as a standard for prediction strategies F in $\mathcal{F}$ whose
norm is not too large: such a prediction strategy F suffers a small loss to the
degree that its predictions resemble the leading strategy's predictions, and the
only way to compete with the leading strategy is to imitate it.

We start the formal exposition with a simple asymptotic result (Proposition
1 in §2) asserting the existence of leading strategies in the problem of on-line
regression with the quadratic loss function for the class of continuous limited-
memory prediction strategies. To state a non-asymptotic version of this result
(Proposition 2) we introduce several general definitions that are used throughout
the paper. In the following two sections Proposition 2 is generalized in two di-
rections, to the loss functions given by Bregman divergences (§3) and by strictly
proper scoring rules (§4). Competitive on-line prediction typically avoids mak-
ing any stochastic assumptions about the way the observations are generated,
but in §5 we consider, mostly for comparison purposes, the case where observa-
tions are generated stochastically. That section contains most of the references
to the related literature, although there are bibliographical remarks scattered
throughout the paper. The proofs are gathered in §6. The final section, §7,
discusses possible directions of further research.

There are many techniques for constructing master strategies, such as gra-
dient descent, strong and weak aggregating algorithms, following the perturbed
leader, defensive forecasting, to mention just a few. In this paper we will use
the technique of defensive forecasting (proposed in [37] and based on [43, 38]
and much earlier work by Levin [25] and Foster and Vohra [19]). The mas-
ter strategies constructed using defensive forecasting automatically satisfy the
stronger properties required of leading strategies; on the other hand, it is not
clear whether leading strategies can be constructed using other techniques (this
is an interesting open question).

2 On-line quadratic-loss regression

Our general prediction protocol is:

ON-LINE PREDICTION PROTOCOL
FOR $n = 1, 2, \ldots$:
    Reality announces $x_n \in \mathbf{X}$.
    Predictor announces $\mu_n \in \mathbf{P}$.
    Reality announces $y_n \in \mathbf{Y}$.
END FOR.

At the beginning of each round n Predictor is given some side information x_n relevant to predicting the following observation y_n, after which he announces his prediction μ_n. The side information is taken from the *information space* $\mathbf{X}$, the observations from the *observation space* $\mathbf{Y}$, and the predictions from the *prediction space* $\mathbf{P}$; all three sets are assumed non-empty. The error of prediction is measured by a *loss function* $\lambda : \mathbf{Y} \times \mathbf{P} \to \mathbb{R}$, so that $\lambda(y_n, \mu_n)$ is the loss suffered by Predictor on round n.

A *prediction strategy* is a strategy for Predictor in this protocol. More explicitly, each prediction strategy F maps each sequence

$$s = (x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \in \mathbf{S} := \bigcup_{n=1}^{\infty} \left((\mathbf{X} \times \mathbf{Y})^{n-1} \times \mathbf{X} \right) \qquad (1)$$

to a prediction $F(s) \in \mathbf{P}$. We will call $\mathbf{S}$ the *situation space* and its elements *situations*; the situation space is always equipped with the standard sum topology, with each addend equipped with the standard product topology, when $\mathbf{X}$ and $\mathbf{Y}$ are topological spaces (see, e.g., [18], §§2.2–2.3). We will sometimes use the notation

$$s_n := (x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \in \mathbf{S}, \qquad (2)$$

where x_i and y_i are Reality's moves in the on-line prediction protocol.

In this section we will always assume that $\mathbf{Y} = [-Y, Y]$ for some $Y > 0$, $[-Y, Y] \subseteq \mathbf{P} \subseteq \mathbb{R}$, and $\lambda(y, \mu) = (y - \mu)^2$; in other words, we will consider the problem of on-line quadratic-loss regression (with the observations bounded in absolute value by a known constant Y).

Asymptotic result

Let k be a positive integer. We say that a prediction strategy F is *order k Markov* if $F(s_n)$ depends on (2) only via $x_{\max(1, n-k)}, y_{\max(1, n-k)}, \ldots, x_{n-1}, y_{n-1}, x_n$. More explicitly, F is order k Markov if and only if there exists a function

$$f : (\mathbf{X} \times \mathbf{Y})^k \times \mathbf{X} \to \mathbf{P}$$

such that, for all $n > k$ and all (2),

$$F(s_n) = f(x_{n-k}, y_{n-k}, \ldots, x_{n-1}, y_{n-1}, x_n).$$

A *limited-memory* prediction strategy is a prediction strategy which is order k Markov for some k. (The expression "Markov strategy" being reserved for "order 0 Markov strategy".)

Proposition 1 *Let* $\mathbf{Y} = \mathbf{P} = [-Y, Y]$ *and* $\mathbf{X}$ *be a compact metric space. There exists a strategy for Predictor that produces* $\mu_n \in [-Y, Y]$ *and guarantees*

$$\frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_n)^2 + \frac{1}{N}\sum_{n=1}^{N}(\mu_n - \phi_n)^2 - \frac{1}{N}\sum_{n=1}^{N}(y_n - \phi_n)^2 \to 0 \qquad (3)$$

as $N \to \infty$ *for the predictions* ϕ_n *output by any continuous limited-memory prediction strategy* F *taking values in* $[-Y, Y]$.

The strategy whose existence is asserted by Proposition 1 is a leading strategy in the sense discussed in §1: the average loss of a continuous limited-memory strategy F is determined by how well it manages to imitate the leading strategy. And once we know the predictions made by F and by the leading strategy, we can find the excess loss of F over the leading strategy without need to know the actual observations y_n.

Leading strategies for reproducing kernel Hilbert spaces

In this subsection we will state a non-asymptotic version of Proposition 1. Let $\mathbf{P} = \mathbb{R}$. Since $\mathbf{P}$ is now a vector space, the sum of two prediction strategies and the product of a scalar (i.e., real number) and a prediction strategy can be defined pointwise:

$$(F_1 + F_2)(s) := F_1(s) + F_2(s), \quad (cF)(s) := cF(s), \qquad s \in \mathbf{S}.$$

Let $\mathcal{F}$ be a Hilbert space of prediction strategies (with the pointwise operations of addition and multiplication by scalar). Its *embedding constant* $\mathbf{c}_{\mathcal{F}}$ is defined by

$$\mathbf{c}_{\mathcal{F}} := \sup_{s \in \mathbf{S}} \sup_{F \in \mathcal{F}: \|F\|_{\mathcal{F}} \leq 1} |F(s)|. \qquad (4)$$

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 embedding constant*. (More generally, $\mathcal{F}$ is said to be an *RKHS* if the internal supremum in (4) is finite for each $s \in \mathbf{S}$.) In our informal discussions we will be assuming that $\mathbf{c}_{\mathcal{F}}$ is a moderately large constant.

Proposition 2 *Let* $\mathbf{Y} = [-Y, Y]$, $\mathbf{P} = \mathbb{R}$, *and* $\mathcal{F}$ *be an RKHS of prediction strategies with finite embedding constant* $\mathbf{c}_{\mathcal{F}}$. *There exists a strategy for Predictor that produces* $\mu_n \in [-Y, Y]$ *and guarantees*

$$\left| \sum_{n=1}^{N}(y_n - \mu_n)^2 + \sum_{n=1}^{N}(\mu_n - \phi_n)^2 - \sum_{n=1}^{N}(y_n - \phi_n)^2 \right|$$

$$\leq 2Y\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1}\,(\|F\|_{\mathcal{F}} + Y)\sqrt{N}, \qquad \forall N \in \{1, 2, \ldots\}\ \forall F \in \mathcal{F}, \quad (5)$$

where ϕ_n *are* F*'s predictions,* $\phi_n := F(s_n)$.

For an F whose norm is not too large (i.e., F satisfying $\|F\|_{\mathcal{F}} \ll N^{1/2}$), (5) shows that

$$\frac{1}{N}\sum_{n=1}^{N}(y_n - \phi_n)^2 \approx \frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_n)^2 + \frac{1}{N}\sum_{n=1}^{N}(\mu_n - \phi_n)^2\,.$$

Proposition 1 is obtained by applying Proposition 2 to a large ("universal") RKHS. The details will be given in §6, and here we will only demonstrate this idea with a simple but non-trivial example.

Let k and m be positive integer constants such that $m > k/2$. The *Sobolev space* $W^{m,2}([-Y,Y]^k)$ is the completion of the vector space of all smooth functions $f : [-Y,Y]^k \to \mathbb{R}$ (with the pointwise operations of addition and multiplication by scalar) equipped with the norm

$$\|f\| := \left(\sum_{0 \le |\alpha| \le m} \int_{[-Y,Y]^k} (D^\alpha f)^2\right)^{1/2},$$

$\alpha = (\alpha_1, \ldots, \alpha_k) \in \{0, 1, \ldots\}^k$ ranging over the multi-indices of order $|\alpha| := \alpha_1 + \cdots + \alpha_k \le m$ and D^α being the operator of differentiating α_1 times over the first variable of f, α_2 times over the second variable of f,..., α_k times over the kth variable of f. (See, e.g., [2] for further details and properties of Sobolev spaces.) A prediction strategy F will be included in $\mathcal{F}$ if its predictions ϕ_n satisfy

$$\phi_n = \begin{cases} 0 & \text{if } n \le k \\ f(y_{n-k}, \ldots, y_{n-1}) & \text{otherwise,} \end{cases}$$

where f is a function from the Sobolev space $W^{m,2}([-Y,Y]^k)$; $\|F\|_{\mathcal{F}}$ is defined to be the Sobolev norm of f. Every continuous function of $(y_{n-k}, \ldots, y_{n-1})$ can be arbitrarily well approximated by functions in $W^{m,2}([-Y,Y]^k)$, and so $\mathcal{F}$ is a suitable class of prediction strategies if we believe that neither $x_1, \ldots, x_n$ nor $y_1, \ldots, y_{n-k-1}$ are useful in predicting y_n.

In this paper we do not describe the prediction algorithm achieving (5); such an algorithm is described in [35] under the name "K29 algorithm" (cf. the proof of Proposition 2 in §6). The algorithm can be implemented efficiently using, e.g., the simple bisection method (see the journal version of [36], the end of §5).

Very large benchmark classes

Some interesting benchmark classes of prediction strategies are too large to equip with the structure of RKHS [36]. However, an analogue of Proposition 2 can also be proved for some Banach spaces $\mathcal{F}$ of prediction strategies (with the pointwise operations of addition and multiplication by scalar) for which the constant $\mathbf{c}_{\mathcal{F}}$ defined by (4) is finite. The *modulus of convexity* of a Banach space U is defined as the function

$$\delta_U(\epsilon) := \inf_{\substack{u,v \in S_U \\ \|u-v\|_U = \epsilon}} \left(1 - \left\|\frac{u+v}{2}\right\|_U\right), \quad \epsilon \in (0, 2],$$

4

where $S_U := \{u \in U \mid \|u\|_U = 1\}$ is the unit sphere in U.

The existence of leading strategies (in a somewhat weaker sense than in Proposition 2) is asserted in the following result.

Proposition 3 *Let $\mathbf{Y} = [-Y, Y]$, $\mathbf{P} = \mathbb{R}$, and $\mathcal{F}$ be a Banach space of prediction strategies having a finite embedding constant $\mathbf{c}_{\mathcal{F}}$ (see (4)) and satisfying*

$$\forall \epsilon \in (0, 2] : \delta_{\mathcal{F}}(\epsilon) \geq (\epsilon/2)^p / p$$

for some $p \in [2, \infty)$. There exists a strategy for Predictor that guarantees

$$\left| \sum_{n=1}^{N} (y_n - \mu_n)^2 + \sum_{n=1}^{N} (\mu_n - \phi_n)^2 - \sum_{n=1}^{N} (y_n - \phi_n)^2 \right|$$
$$\leq 40Y \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \left(\|F\|_{\mathcal{F}} + Y \right) N^{1-1/p}, \qquad \forall N \in \{1, 2, \ldots\} \ \forall F \in \mathcal{F}, \quad (6)$$

where ϕ_n are F's predictions.

The example of a benchmark class of prediction strategies given after Proposition 2 but with f ranging over the Sobolev space $W^{s,p}([-Y, Y]^k)$, $s > k/p$, is covered by this proposition: see, e.g., [36], §2. The parameter s (in general, not necessarily integer) describes the "degree of regularity" of the elements of $W^{s,p}$, and taking sufficiently large p we can reach arbitrarily irregular functions in the Sobolev hierarchy.

3 Predictions evaluated by Bregman divergences

A *predictable process* is a function F mapping the situation space $\mathbf{S}$ to $\mathbb{R}$, $F : \mathbf{S} \to \mathbb{R}$. Notice that for any function $\psi : \mathbf{P} \to \mathbb{R}$ and any prediction strategy F the composition $\psi(F)$ (mapping each situation s to $\psi(F(s))$) is a predictable process; such compositions will be used in Theorems 1–3 below. A Hilbert space $\mathcal{F}$ of predictable processes (with the usual pointwise operations) is called an *RKHS with finite embedding constant* if (4) is finite.

The notion of Bregman divergence was introduced in [8], and is now widely used in competitive on-line prediction (see, e.g., [20, 6, 21, 23, 10]). Suppose $\mathbf{Y} = \mathbf{P}$ is a closed interval of the real line $\mathbb{R}$ (although it would be interesting to extend Theorem 1 to the case where $\mathbb{R}$ is replaced by any Euclidean, or even Hilbert, space). Let Ψ and Ψ' be two real-valued functions defined on $\mathbf{Y}$. The expression

$$d_{\Psi, \Psi'}(y, z) := \Psi(y) - \Psi(z) - \Psi'(z)(y - z), \quad y, z \in \mathbf{Y}, \quad (7)$$

is said to be the corresponding *Bregman divergence* if $d_{\Psi, \Psi'}(y, z) > 0$ whenever $y \neq z$. (Bregman divergence is usually defined for y and z ranging over a Euclidean space.) In all our examples Ψ will be a strictly convex continuously

differentiable function and Ψ' its derivative, in which case we abbreviate $d_{\Psi, \Psi'}$ to d_Ψ.

We will be using the standard notation

$$\|f\|_{C(A)} := \sup_{y \in A} |f(y)|, \tag{8}$$

where A is a subset of the domain of f. The *diameter* of a set $A \subseteq \mathbb{R}$ is defined as

$$\mathrm{diam}(A) := \sup_{a, a' \in A} |a - a'|.$$

Theorem 1 *Suppose* $\mathbf{Y} = \mathbf{P}$ *is a closed interval of the real line* $\mathbb{R}$. *Let* $\mathcal{F}$ *be an RKHS of predictable processes with finite embedding constant* $\mathbf{c}_{\mathcal{F}}$ *and* Ψ, Ψ' *be continuous real-valued functions on* $\mathbf{Y} = \mathbf{P}$. *There exists a strategy for Predictor that guarantees, for all prediction strategies* F *and* $N = 1, 2, \ldots$,

$$\left| \sum_{n=1}^{N} d_{\Psi, \Psi'} (y_n, \mu_n) + \sum_{n=1}^{N} d_{\Psi, \Psi'} (\mu_n, \phi_n) - \sum_{n=1}^{N} d_{\Psi, \Psi'} (y_n, \phi_n) \right|$$
$$\leq \mathrm{diam}(\mathbf{Y}) \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \left(\|\Psi'(F)\|_{\mathcal{F}} + \|\Psi'\|_{C(\mathbf{Y})} \right) \sqrt{N}, \quad (9)$$

where ϕ_n *are* F's *predictions.*

The expression $\|\Psi'(F)\|_{\mathcal{F}}$ in (9) is interpreted as ∞ when $\Psi'(F) \notin \mathcal{F}$; in this case (9) holds vacuously. Similar conventions will be implicit in all following statements.

Two of the most important Bregman divergences are obtained from the convex functions $\Psi(y) := y^2$ and $\Psi(y) := y \ln y + (1 - y) \ln(1 - y)$ (negative entropy, defined for $y \in (0, 1)$); they are the quadratic loss function

$$d_\Psi(y, z) = (y - z)^2 \tag{10}$$

and the relative entropy (also known as the Kullback–Leibler divergence)

$$d_\Psi(y, z) = D(y \parallel z) := y \ln \frac{y}{z} + (1 - y) \ln \frac{1 - y}{1 - z}, \tag{11}$$

respectively. If we apply Theorem 1 to them, (10) leads (assuming $\mathbf{Y} = [-Y, Y]$) to a weaker version of Proposition 2, with the right-hand side of (9) twice as large as that of (5), and (11) leads to the following corollary.

Corollary 1 *Let* $\epsilon \in (0, 1/2)$, $\mathbf{Y} = \mathbf{P} = [\epsilon, 1 - \epsilon]$, *and the loss function be*

$$\lambda(y, \mu) = D(y \parallel \mu)$$

(defined in (11)). Let $\mathcal{F}$ *be an RKHS of predictable processes with finite embedding constant* $\mathbf{c}_{\mathcal{F}}$. *There exists a strategy for Predictor that guarantees, for all*

prediction strategies F,

$$\left| \sum_{n=1}^{N} \lambda\left(y_n, \mu_n\right) + \sum_{n=1}^{N} \lambda\left(\mu_n, \phi_n\right) - \sum_{n=1}^{N} \lambda\left(y_n, \phi_n\right) \right|$$

$$\leq \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \left(\left\| \ln \frac{F}{1-F} \right\|_{\mathcal{F}} + \ln \frac{1-\epsilon}{\epsilon} \right) \sqrt{N}, \qquad \forall N \in \{1, 2, \ldots\},$$

where ϕ_n are F's predictions.

The log likelihood ratio $\ln \frac{F}{1-F}$ appears because $\Psi'(y) = \ln \frac{y}{1-y}$ in this case.

Analogously to Proposition 2, Theorem 1 (as well as Theorems 2–3 in the next section) can be easily generalized to Banach spaces of predictable processes. One can also state asymptotic versions of Theorems 1–3 similar to Proposition 1; and the continuous limited-memory strategies of Proposition 1 could be replaced by the equally interesting classes of continuous stationary strategies (as in [40]) or Markov strategies (possibly discontinuous, as in [39]). We will have to refrain from pursuing these developments in this paper.

4 Predictions evaluated by strictly proper scoring rules

In this section we consider the case where $\mathbf{Y} = \{0, 1\}$ and $\mathbf{P} \subseteq [0, 1]$. Every loss function $\lambda : \mathbf{Y} \times \mathbf{P} \to \mathbb{R}$ will be extended to the domain $[0, 1] \times \mathbf{P}$ by the formula

$$\lambda(p, \mu) := p\lambda(1, \mu) + (1-p)\lambda(0, \mu);$$

intuitively, $\lambda(p, \mu)$ is the expected loss of the prediction μ when the probability of $y = 1$ is p. Let us say that a loss function λ is a *strictly proper scoring rule* if

$$\forall p, \mu \in \mathbf{P} : p \neq \mu \implies \lambda(p, p) < \lambda(p, \mu)$$

(it is optimal to give the prediction equal to the true probability of $y = 1$ when the latter is known and belongs to $\mathbf{P}$). In this case the function

$$d_\lambda(\mu, \phi) := \lambda(\mu, \phi) - \lambda(\mu, \mu)$$

can serve as a measure of difference between predictions μ and ϕ: it is non-negative and is zero only when $\mu = \phi$. (Cf. [14], §4.)

The *exposure* of a loss function λ is defined as

$$\mathrm{Exp}_\lambda(\mu) := \lambda(1, \mu) - \lambda(0, \mu), \quad \mu \in \mathbf{P}.$$

Theorem 2 *Let $\mathbf{Y} = \{0, 1\}$, $\mathbf{P} = [0, 1]$, λ be a continuous strictly proper scoring rule, and $\mathcal{F}$ be an RKHS of predictable processes with finite embedding constant $\mathbf{c}_{\mathcal{F}}$. There exists a strategy for Predictor that guarantees, for all prediction*

strategies F and all $N = 1, 2, \ldots,$

$$\left| \sum_{n=1}^{N} \lambda\left(y_n, \mu_n\right) + \sum_{n=1}^{N} d_\lambda\left(\mu_n, \phi_n\right) - \sum_{n=1}^{N} \lambda\left(y_n, \phi_n\right) \right|$$
$$\leq \frac{\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1}}{2} \left(\|\mathrm{Exp}_\lambda(F)\|_{\mathcal{F}} + \|\mathrm{Exp}_\lambda\|_{C(\mathbf{P})} \right) \sqrt{N}, \quad (12)$$

where ϕ_n are F's predictions.

Two popular strictly proper scoring rules are the quadratic loss function $\lambda(y, \mu) := (y - \mu)^2$ and the log loss function

$$\lambda(y, \mu) := \begin{cases} -\ln \mu & \text{if } y = 1 \\ -\ln(1 - \mu) & \text{if } y = 0. \end{cases}$$

Applied to the quadratic loss function, Theorem 2 becomes essentially a special case of Proposition 2 (with $Y = 1/2$). For the log loss function we have $d_\lambda(\mu, \phi) = D(\mu \,\|\, \phi)$, and one might hope to obtain a version of Corollary 1 for $\mathbf{Y} = \{0, 1\}$ from Theorem 2. Unfortunately, Theorem 2 is not sufficient since the exposure of the log loss function is unbounded (and from the formal point of view, the log loss function $\lambda(y, \mu)$ is not even defined for some $\mu \in [0, 1]$, namely for $\mu \in \{0, 1\}$). This can be done using methods of [37], and it is even possible to get rid of the restriction $\mathbf{P} = [\epsilon, 1 - \epsilon]$ in Corollary 1. Since the log loss function plays a fundamental role in information theory (the cumulative log loss corresponds to the code length), we state this result as our next theorem.

Theorem 3 *Let $\mathbf{Y} = \{0, 1\}$, $\mathbf{P} = (0, 1)$, λ be the log loss function, and $\mathcal{F}$ be an RKHS of predictable processes with finite embedding constant $\mathbf{c}_{\mathcal{F}}$. There exists a strategy for Predictor that guarantees, for all prediction strategies F,*

$$\left| \sum_{n=1}^{N} \lambda\left(y_n, \mu_n\right) + \sum_{n=1}^{N} D\left(\mu_n \,\|\, \phi_n\right) - \sum_{n=1}^{N} \lambda\left(y_n, \phi_n\right) \right|$$
$$\leq \frac{\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1.8}}{2} \left(\left\| \ln \frac{F}{1 - F} \right\|_{\mathcal{F}} + 1 \right) \sqrt{N}, \qquad \forall N \in \{1, 2, \ldots\},$$

where ϕ_n are F's predictions.

It is clear that the restriction to strictly proper scoring rules is essential. Consider, e.g., the absolute loss function $\lambda(y, \mu) := |y - \mu|$, the sequence of observations $(y_1, y_2, \ldots) := (0, 1, 0, 1, \ldots)$, and $\mathcal{F}$ consisting of the constant prediction strategies. All strategies in $\mathcal{F}$ will suffer essentially the same loss, and it is clear that leading strategies do not exist; Theorem 2 is not applicable as the absolute loss function is not a proper scoring rule.

5 Stochastic Reality and Jeffreys's law

In this section we revert to the quadratic regression framework of §2 and assume $\mathbf{Y} = \mathbf{P} = [-Y, Y]$, $\lambda(y, \mu) = (y - \mu)^2$. (It will be clear that similar results hold for Bregman divergences and strictly proper scoring rules, but we stick to the simplest case since our main goal in this section is to discuss the related literature.)

Proposition 4 *Suppose $\mathbf{Y} = \mathbf{P} = [-Y, Y]$. Let F be a prediction strategy and $y_n \in [-Y, Y]$ be generated as $y_n := F(s_n) + \xi_n$ (remember that s_n are defined by (2)), where the noise random variables ξ_n have expected value zero given s_n. For any other prediction strategy G, any $N \in \{1, 2, \ldots\}$, and any $\delta \in (0, 1)$,*

$$\left| \sum_{n=1}^{N} (y_n - \phi_n)^2 + \sum_{n=1}^{N} (\phi_n - \mu_n)^2 - \sum_{n=1}^{N} (y_n - \mu_n)^2 \right| \leq 8Y^2 \sqrt{2 \ln \frac{2}{\delta}} \sqrt{N} \quad (13)$$

with probability at least $1 - \delta$, where ϕ_n are F's predictions and μ_n are G's predictions.

Combining Proposition 4 with Proposition 2 we obtain the following corollary.

Corollary 2 *Suppose $\mathbf{Y} = [-Y, Y]$ and $\mathbf{P} = \mathbb{R}$. Let $\mathcal{F}$ be an RKHS of prediction strategies with finite embedding constant $\mathbf{c}_{\mathcal{F}}$, G be a prediction strategy whose predictions μ_n always belong to $[-Y, Y]$ and are guaranteed to satisfy (5) (a "leading prediction strategy"), F be a prediction strategy in $\mathcal{F}$ taking values in $[-Y, Y]$, and $y_n \in [-Y, Y]$ be generated as $y_n := F(s_n) + \xi_n$, where the noise random variables ξ_n have expected value zero given s_n. For any $N \in \{1, 2, \ldots\}$ and any $\delta \in (0, 1)$, the conjunction of*

$$\left| \sum_{n=1}^{N} (y_n - \mu_n)^2 - \sum_{n=1}^{N} (y_n - \phi_n)^2 \right|$$
$$\leq Y \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \left(\|F\|_{\mathcal{F}} + Y \right) \sqrt{N} + 4Y^2 \sqrt{2 \ln \frac{2}{\delta}} \sqrt{N} \quad (14)$$

and

$$\sum_{n=1}^{N} (\phi_n - \mu_n)^2 \leq Y \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \left(\|F\|_{\mathcal{F}} + Y \right) \sqrt{N} + 4Y^2 \sqrt{2 \ln \frac{2}{\delta}} \sqrt{N} \quad (15)$$

holds with probability at least $1 - \delta$, where ϕ_n are F's predictions and μ_n are G's predictions.

We can see that if the "true" (in the sense of outputting the true expectations) strategy F belongs to the RKHS $\mathcal{F}$ and $\|F\|_{\mathcal{F}}$ is not too large, not only the loss of the leading strategy will be close to that of the true strategy, but their predictions will be close as well.

Jeffreys's law

In the rest of this section we will explain the connection of this paper with the phenomenon widely studied in probability theory and the algorithmic theory of randomness and dubbed "Jeffreys's law" by Dawid [12, 15]. The general statement of "Jeffreys's law" is that two successful prediction strategies produce similar predictions (cf. [12], §5.2). To better understand this informal statement, we first discuss two notions of success for prediction strategies.

As argued in [41], there are (at least) two very different kinds of predictions, which we will call "S-predictions" and "D-predictions". Both S-predictions and D-predictions are elements of $[-Y, Y]$ (in our current context), and the prefixes "S-" and "D-" refer to the way in which we want to evaluate their quality. S-predictions are Statements about Reality's behavior, and they are successful if they withstand attempts to falsify them; standard means of falsification are statistical tests (see, e.g., [11], Chapter 3) and gambling strategies ([32]; for a more recent exposition, see [29]). D-predictions do not claim to be falsifiable statements about Reality; they are Decisions deemed successful if they lead to a good cumulative loss.

As an example, let us consider the predictions ϕ_n and μ_n in Proposition 4. The former are S-predictions; e.g., they can be rejected if (13) fails to happen for a small δ and fixed G (the complement of (13) can be used as the critical region of a statistical test). The latter are D-predictions: we are only interested in their cumulative loss. If ϕ_n are successful ((13) holds for a moderately small δ) and μ_n are successful (in the sense of their cumulative loss being close to the cumulative loss of the successful S-predictions ϕ_n; this is the best that can be achieved as, by (13), the latter cannot be much larger than the former), they will be close to each other, in the sense $\sum_{n=1}^{N} (\phi_n - \mu_n)^2 \ll N$. We can see that Proposition 4 implies a "mixed" version of Jeffreys's law, asserting the proximity of S-predictions and D-predictions.

Similarly, Corollary 2 is also a mixed version of Jeffreys's law: it asserts the proximity of the S-predictions ϕ_n (which are part of our falsifiable model $y_n = \phi_n + \xi_n$) and the D-predictions μ_n (successful in the sense of leading to a good cumulative loss; cf. (5)).

Proposition 2 immediately implies two "pure" versions of Jeffreys's laws for D-predictions:

- if a prediction strategy F with $\|F\|_{\mathcal{F}}$ not too large performs well, in the sense that its loss is close to the leading strategy's loss, F's predictions will be similar to the leading strategy's predictions; more precisely,

$$\sum_{n=1}^{N} (\phi_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - \phi_n)^2 - \sum_{n=1}^{N} (y_n - \mu_n)^2$$
$$+ 2Y \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \left(\|F\|_{\mathcal{F}} + Y \right) \sqrt{N};$$

- therefore, if two prediction strategies F_1 and F_2 with $\|F_1\|_{\mathcal{F}}$ and $\|F_2\|_{\mathcal{F}}$

not too large perform well, in the sense that their loss is close to the leading strategy's loss, their predictions will be similar.

It is interesting that the leading strategy can be replaced by a master strategy for the second version: if F_1 and F_2 gave very different predictions and both performed almost as well as the master strategy, the mixed strategy $(F_1 + F_2)/2$ would beat the master strategy; this immediately follows from

$$\left(\frac{\phi_1 + \phi_2}{2} - y\right)^2 = \frac{(\phi_1 - y)^2 + (\phi_2 - y)^2}{2} - \left(\frac{\phi_1 - \phi_2}{2}\right)^2,$$

where ϕ_1 and ϕ_2 are F_1's and F_2's predictions, respectively, and y is the observation.

The usual versions of Jeffreys's law are, however, statements about S-predictions. The quality of S-predictions is often evaluated using universal statistical tests (as formalized by Martin-Löf [26]) or universal gambling strategies (Levin [24], Schnorr [28]). For example, Theorem 7.1 of [13] and Theorem 3 of [33] state that if two computable S-prediction strategies are both successful, their predictions will asymptotically agree. Earlier, somewhat less intuitive, statements of Jeffreys's law were given in terms of absolute continuity of probability measures: see, e.g., [7] and [22]. Solomonoff [30] proved a version of Jeffreys's law that holds "on average" (rather than for individual sequences).

This paper is, to my knowledge, the first to state a version of Jeffreys's law for D-predictions (although a step in this direction was made in Theorem 8 of [34]).

6 Proofs

In this section we prove, or give proof sketches of, Propositions 1–4, Theorems 1–3 and Corollary 2. We start from the proof of Proposition 2, which is most intuitive.

Proof of Propositions 2 and 3

Noticing that

$$\left|\sum_{n=1}^{N} (y_n - \mu_n)^2 + \sum_{n=1}^{N} (\mu_n - \phi_n)^2 - \sum_{n=1}^{N} (y_n - \phi_n)^2\right|$$

$$= 2\left|\sum_{n=1}^{N} (\phi_n - \mu_n)(y_n - \mu_n)\right|$$

$$\leq 2\left|\sum_{n=1}^{N} \mu_n (y_n - \mu_n)\right| + 2\left|\sum_{n=1}^{N} \phi_n (y_n - \mu_n)\right|, \quad (16)$$

we can use the results of [35], §6, asserting the existence of a prediction strategy producing predictions $\mu_n \in [-Y, Y]$ that satisfy

$$\left| \sum_{n=1}^{N} \mu_n (y_n - \mu_n) \right| \leq Y^2 \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \sqrt{N} \qquad (17)$$

(see (24) in [35]) and

$$\left| \sum_{n=1}^{N} \phi_n (y_n - \mu_n) \right| \leq Y \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \, \|F\|_{\mathcal{F}} \, \sqrt{N} \qquad (18)$$

(see (25) in [35]).

The intuition behind (17) is that it is a special case of the requirement of good calibration of the predictions μ_n. More generally and informally, the predictions are said to be well calibrated if, for any $\mu^* \in \mathbf{P}$ and for large enough N, the average of y_n over those of the first N rounds of the prediction protocol that satisfy $\mu_n \approx \mu^*$ is close to μ^*. On the other hand, the intuition behind (18) is that it expresses the property of good resolution; informally, the predictions are said to have good resolution if, for any $x^* \in \mathbf{X}$ and for large enough N, the average of $y_n - \mu_n$ over those of the first N rounds of the prediction protocol that satisfy $x_n \approx x^*$ is close to zero. The method of defensive forecasting allows Predictor to enforce the properties of good calibration and resolution given a proof, expressed in a suitable (game-theoretic) language, that the predictions calculated from the true probability mechanism generating the data satisfy these properties. This idea is implemented in various contexts in [38, 35, 36, 41].

Replacing (17) and (18) with the corresponding statements for Banach function spaces ([36], (28) and (29)) we obtain the proof of Proposition 3.

Remark In [35] we considered only prediction strategies F for which $F(s_n)$ depends on s_n (see (2)) via x_n; in the terminology of this paper these are (order 0) Markov strategies. It is easy to see that considering only Markov strategies does not lead to a loss of generality: if we redefine the object x_n as $x_n := s_n$, any prediction strategy will become a Markov prediction strategy.

Proof of Proposition 1

Proposition 1 will follow from the following lemma, proved (without stating it explicitly) in [31] (proof of Theorem 2). Remember that $C(A)$, where A is any topological space, is the set of all continuous real-valued functions on A equipped with the sup norm (8). We will use the notation ℓ_2 for the Hilbert space of infinite sequences $(a_1, a_2, \ldots) \in \mathbb{R}^\infty$ with componentwise addition and multiplication by scalar and inner product

$$\langle (a_1, a_2, \ldots), (a_1', a_2', \ldots) \rangle := \sum_{n=1}^{\infty} a_n a_n'.$$

Lemma 1 ([31]) *Let $\mathcal{G}$ be a separable set in $C(Z)$. There exists an RKHS $\mathcal{F}$ on Z with finite embedding constant such that $\mathcal{F}$ is dense in $\mathcal{G}$ in metric $C(Z)$.*

Proof Let $F_1, F_2, \ldots$ be a dense (in metric $C(Z)$) sequence of elements of $\mathcal{G}$. Set

$$\Phi_n := \begin{cases} 2^{-n} \|F_n\|_{C(Z)}^{-1} F_n & \text{if } F_n \neq 0 \\ 0 & \text{otherwise,} \end{cases}$$

$\Phi(z) := (\Phi_1(z), \Phi_2(z), \ldots) \in \ell_2$ for $z \in Z$,

$$\mathbf{K}(z, z') := \langle \Phi(z), \Phi(z') \rangle_{\ell_2}, \qquad z, z' \in Z,$$

and let $\mathcal{F}$ be the unique RKHS with reproducing kernel $\mathbf{K}$ (see the Moore–Aronszajn theorem in [3], Theorem 2). It is clear that $\mathbf{c}_{\mathcal{F}}^2 = \sup_z \mathbf{K}(z, z)$ is finite. By Lemma 2 below, each F_n belongs to $\mathcal{F}$ since it can be represented as

$$\left\langle 2^n \|F_n\|_{C(Z)} e_n, \Phi(\cdot) \right\rangle_{\ell_2},$$

where $e_n \in \ell_2$ consists of all 0s except a 1 at the nth position. Therefore, $\mathcal{F}$ is dense in $\mathcal{G}$. ∎

The following lemma (which is contained in, e.g., Theorem 1 in §2.1 of [27]) was used in the proof.

Lemma 2 *Let $\Phi : Z \to H$, where H is a Hilbert space. The RKHS corresponding to the reproducing kernel $\mathbf{K}(z, z') := \langle \Phi(z), \Phi(z') \rangle_H$ consists of all functions $\langle v, \Phi(\cdot) \rangle_H$, $v \in H$, with the inner product of $\langle v, \Phi(\cdot) \rangle_H$ and $\langle v', \Phi(\cdot) \rangle_H$ equal to $\langle p(v), p(v') \rangle_H$, p standing for the projection onto the closure of the span of $\Phi(Z)$.*

Proof By the Moore–Aronszajn theorem ([3], Theorem 2) there is a unique RKHS with reproducing kernel $\mathbf{K}$, so we only need to check that the function space $\mathcal{F}$ defined in the statement of the lemma is an RKHS with $\mathbf{K}$ as reproducing kernel.

First we need to check that the inner product is well defined. This follows from the obvious fact that the equality of the functions $\langle v, \Phi(\cdot) \rangle_H$ and $\langle v', \Phi(\cdot) \rangle_H$ for $v, v' \in \overline{\text{span}}(\Phi(Z))$ implies $v = v'$. The continuity of each evaluation functional is also obvious.

The representer of $z \in Z$ is $\mathbf{k}_z(\cdot) := \langle \Phi(z), \Phi(\cdot) \rangle_H$ (in the sense that $\langle \mathbf{k}_z, f \rangle_{\mathcal{F}} = f(z)$ for each $f \in \mathcal{F}$) and so the reproducing kernel $\langle \mathbf{k}_z, \mathbf{k}_{z'} \rangle_{\mathcal{F}}$ of $\mathcal{F}$ indeed coincides with $\mathbf{K}$. ∎

Now we can easily deduce Proposition 1 from Proposition 2. The set of all continuous order k Markov strategies is a separable set in the Banach space $C(\mathbf{S})$ of continuous prediction strategies with the sup norm (by [18], Corollary 4.2.18). Therefore, the set $\mathcal{G}$ of all continuous limited-memory strategies is separable in $C(\mathbf{S})$.

13

Let $\mathcal{F}$ be the RKHS whose existence is asserted by Lemma 1; we will see that any strategy for Predictor satisfying (5) and $\mu_n \in [-Y, Y]$ will satisfy (3) with $\phi_n \in [-Y, Y]$ output by a continuous limited-memory strategy F taking values in $[-Y, Y]$. Indeed, for any $\epsilon > 0$ we can find $F^* \in \mathcal{F}$ that is ϵ-close in $C(\mathbf{S})$ to F. If ϕ_n are F's predictions and ϕ_n^* are F^*'s predictions, (5) implies that

$$
\left| \frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 + \frac{1}{N} \sum_{n=1}^{N} (\mu_n - \phi_n)^2 - \frac{1}{N} \sum_{n=1}^{N} (y_n - \phi_n)^2 \right|
$$

$$
\leq \left| \frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 + \frac{1}{N} \sum_{n=1}^{N} (\mu_n - \phi_n^*)^2 - \frac{1}{N} \sum_{n=1}^{N} (y_n - \phi_n^*)^2 \right| + 8(Y + \epsilon)\epsilon
$$

$$
\leq 2Y \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \left(\|F^*\|_{\mathcal{F}} + Y \right) \frac{1}{\sqrt{N}} + 8(Y + \epsilon)\epsilon \leq 10(Y + \epsilon)\epsilon
$$

from some N on. Since ϵ can be taken arbitrarily small, we have (3).

Proof of Theorem 1

The proof is based on the *generalized law of cosines*

$$
d_{\Psi, \Psi'}(y, \phi) = d_{\Psi, \Psi'}(\mu, \phi) + d_{\Psi, \Psi'}(y, \mu) - (\Psi'(\phi) - \Psi'(\mu))(y - \mu) \qquad (19)
$$

(which follows directly from the definition (7)). From (19) we deduce

$$
\left| \sum_{n=1}^{N} d_{\Psi, \Psi'}(y_n, \mu_n) + \sum_{n=1}^{N} d_{\Psi, \Psi'}(\mu_n, \phi_n) - \sum_{n=1}^{N} d_{\Psi, \Psi'}(y_n, \phi_n) \right|
$$

$$
= \left| \sum_{n=1}^{N} (\Psi'(\phi_n) - \Psi'(\mu_n))(y_n - \mu_n) \right|
$$

$$
\leq \left| \sum_{n=1}^{N} \Psi'(\mu_n)(y_n - \mu_n) \right| + \left| \sum_{n=1}^{N} \Psi'(\phi_n)(y_n - \mu_n) \right|. \qquad (20)
$$

From Theorem 3 in [42] we can see that there is a prediction strategy guaranteeing

$$
\left| \sum_{n=1}^{N} \Psi'(\mu_n)(y_n - \mu_n) \right| \leq \mathrm{diam}(\mathbf{Y}) \|\Psi'\|_{C(\mathbf{Y})} \sqrt{N} \qquad (21)
$$

and from Theorem 4 in [42] we can see that there is a prediction strategy guaranteeing

$$
\left| \sum_{n=1}^{N} \Psi'(\phi_n)(y_n - \mu_n) \right| \leq \mathrm{diam}(\mathbf{Y}) \mathbf{c}_{\mathcal{F}} \|\Psi'(F)\|_{\mathcal{F}} \sqrt{N}. \qquad (22)
$$

We need, however, a single strategy guaranteeing some versions of (21) and (22). Such a strategy can be obtained by merging a strategy guaranteeing (21) and a strategy guaranteeing (22) (as in [35], Corollaries 3 and 4).

Setting

$$\Phi(\mu, s) := \left(\frac{\Psi'(\mu)}{\|\Psi'\|_{C(\mathbf{Y})}}, \mathbf{k}_s \right) \in \mathbb{R} \times \mathcal{F}, \quad \mu \in \mathbf{P}, \ s \in \mathbf{S}, \tag{23}$$

so that $\mathbf{c}_\Phi := \sup_{\mu,s} \|\Phi(\mu, s)\| \leq \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1}$ ($\mathbf{k}_s$ is the representer of s and $\mathbb{R} \times \mathcal{F}$ is equipped with the standard inner product of the direct sum of $\mathbb{R}$ and $\mathcal{F}$), and letting μ_n be output by the K29 algorithm (as described in [42]) based on (23), we obtain

$$\left| \sum_{n=1}^{N} \Psi'(\mu_n)(y_n - \mu_n) \right| \leq \|\Psi'\|_{C(\mathbf{Y})} \left\| \sum_{n=1}^{N} (y_n - \mu_n)\Phi(\mu_n, s_n) \right\|_{\mathbb{R} \times \mathcal{F}}$$

$$\leq \|\Psi'\|_{C(\mathbf{Y})} \operatorname{diam}(\mathbf{Y}) \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \sqrt{N} \tag{24}$$

from Theorem 3 of [42], and we obtain

$$\left| \sum_{n=1}^{N} \Psi'(\phi_n)(y_n - \mu_n) \right| = \left| \sum_{n=1}^{N} (y_n - \mu_n) \langle \mathbf{k}_{s_n}, \Psi'(F) \rangle_{\mathcal{F}} \right|$$

$$= \left| \left\langle \sum_{n=1}^{N} (y_n - \mu_n) \mathbf{k}_{s_n}, \Psi'(F) \right\rangle_{\mathcal{F}} \right| \leq \|\Psi'(F)\|_{\mathcal{F}} \left\| \sum_{n=1}^{N} (y_n - \mu_n) \mathbf{k}_{s_n} \right\|_{\mathcal{F}}$$

$$\leq \|\Psi'(F)\|_{\mathcal{F}} \left\| \sum_{n=1}^{N} (y_n - \mu_n)\Phi(\mu_n, s_n) \right\|_{\mathbb{R} \times \mathcal{F}}$$

$$\leq \|\Psi'(F)\|_{\mathcal{F}} \operatorname{diam}(\mathbf{Y}) \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1} \sqrt{N} \tag{25}$$

from the proof of Theorem 4 and from Theorem 3 of [42].

Combining (20) with (24) and (25) we can see that (23) produces a strategy guaranteeing (9).

Remark As we mentioned earlier, the leading constant in the bound of Theorem 1 (and its corollary) is worse than those in other results in this paper, in the intersection of their domains of application. The explanation is that Theorem 1 is based on the K29 algorithm, whereas all other results are based on the more sophisticated "K29* algorithm".

Proof sketch of Theorem 2

The proof is similar to that of Theorem 1, with the role of the generalized law of cosines (19) played by the equation

$$\lambda(y, \phi) = a + \lambda(y, \mu) + b(y - \mu) \tag{26}$$

15

for some $a = a(\mu, \phi)$ and $b = b(\mu, \phi)$. Since y can take only two possible values, suitable a and b are easy to find: it suffices to solve the linear system

$$\begin{cases} \lambda(1, \phi) = a + \lambda(1, \mu) + b(1 - \mu) \\ \lambda(0, \phi) = a + \lambda(0, \mu) + b(-\mu). \end{cases}$$

Subtracting these equations we obtain $b = \mathrm{Exp}(\phi) - \mathrm{Exp}(\mu)$ (abbreviating Exp_λ to Exp), which in turn gives $a = d_\lambda(\mu, \phi)$. Therefore, (26) gives

$$\left| \sum_{n=1}^N \lambda(y_n, \mu_n) + \sum_{n=1}^N d_\lambda(\mu_n, \phi_n) - \sum_{n=1}^N \lambda(y_n, \phi_n) \right|$$

$$= \left| \sum_{n=1}^N (\mathrm{Exp}(\phi_n) - \mathrm{Exp}(\mu_n))(y_n - \mu_n) \right|$$

$$\leq \left| \sum_{n=1}^N \mathrm{Exp}(\mu_n)(y_n - \mu_n) \right| + \left| \sum_{n=1}^N \mathrm{Exp}(\phi_n)(y_n - \mu_n) \right|. \quad (27)$$

There are prediction strategies that guarantee

$$\left| \sum_{n=1}^N \mathrm{Exp}(\mu_n)(y_n - \mu_n) \right| \leq \frac{1}{2} \|\mathrm{Exp}\|_{C(\mathbf{P})} \sqrt{N} \quad (28)$$

(cf. [37], Theorem 2) and there are prediction strategies that guarantee

$$\left| \sum_{n=1}^N \mathrm{Exp}(F(s_n))(y_n - \mu_n) \right| \leq \frac{\mathbf{c}_{\mathcal{F}}}{2} \|\mathrm{Exp}(F)\|_{\mathcal{F}} \sqrt{N} \quad (29)$$

(cf. [37], Theorem 3); merging such strategies as in [35], Corollaries 3 and 4, we can easily obtain (12) from (27), (28), and (29).

Proof sketch of Theorem 3

It is shown in [37] that there is a prediction strategy guaranteeing

$$\left| \sum_{n=1}^N \mathrm{Exp}(\mu_n)(y_n - \mu_n) \right| \leq \sqrt{\sum_{n=1}^N \mu_n(1 - \mu_n)\left(\mathrm{Exp}^2(\mu_n) + \mathbf{K}(s_n, s_n)\right)} \quad (30)$$

and

$$\left| \sum_{n=1}^N \mathrm{Exp}(F(s_n))(y_n - \mu_n) \right|$$

$$\leq \|\mathrm{Exp}(F)\|_{\mathcal{F}} \sqrt{\sum_{n=1}^N \mu_n(1 - \mu_n)\left(\mathrm{Exp}^2(\mu_n) + \mathbf{K}(s_n, s_n)\right)} \quad (31)$$

16

(see (21), (22), and the subsection "Proof: Part II" in [37]), where $\mathbf{K}$ is the reproducing kernel of $\mathcal{F}$. Comparing (30) and (31) with (27), we can see that Theorem 3 will follow from

$$\sqrt{\sum_{n=1}^{N} \mu_n(1 - \mu_n)\left(\mathrm{Exp}^2(\mu_n) + \mathbf{K}(s_n, s_n)\right)} \leq \frac{\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1.8}}{2}\sqrt{N},$$

which in turn will follow from

$$\mu(1 - \mu)\left(\ln^2 \frac{\mu}{1 - \mu} + \mathbf{c}_{\mathcal{F}}^2\right) \leq \frac{\mathbf{c}_{\mathcal{F}}^2 + 1.8}{4}.$$

It remains to notice that $\mu(1 - \mu) \leq 1/4$ and to calculate

$$\sup_{\mu}\left(4\mu(1 - \mu)\ln^2 \frac{\mu}{1 - \mu}\right) \approx 1.76 \leq 1.8.$$

Proof of Proposition 4

This proposition immediately follows from the first equality in (16) (with the roles of μ_n and ϕ_n interchanged) and Hoeffding's inequality (the latter is stated in, e.g., [16], p. 135).

Proof of Corollary 2

On an event of probability at least $1 - \delta$ we will have the conjunction of (5) and (13). Inequality (15) now obtains as the average of (5) and (13) with all the straight brackets "|" (signifying absolute value) removed; by the average of inequalities $a \leq b$ and $c \leq d$ we mean the inequality $(a + c)/2 \leq (b + d)/2$. Inequality (14) with the straight brackets removed obtains as the average of (5) with the straight brackets removed and (13) with the $|\cdots|$ replaced by $-(\cdots)$. Inequality (14) with the $|\cdots|$ replaced by $-(\cdots)$ obtains as the average of (5) with the $|\cdots|$ replaced by $-(\cdots)$ and (13) with the straight brackets removed.

7 Conclusion

The existence of master strategies (strategies whose loss is less than or close to the loss of any strategy with not too large a norm) can be shown for a very wide class of loss functions. On the contrary, leading strategies appear to exist for a rather narrow class of loss functions. It would be very interesting to delineate the class of loss functions for which a leading strategy does exist. In particular, does this class contain any natural loss functions except Bregman divergences and strictly proper scoring rules?

Even if a leading strategy does not exist, one might look for a strategy G such that the loss of any strategy F whose norm is not too large lies between the loss of G plus some measure of difference between F's and G's predictions and the loss of G plus another measure of difference between F's and G's predictions.

References

[1] Robert A. Adams. *Sobolev Spaces*, volume 65 of *Pure and Applied Mathematics*. Academic Press, New York, first edition, 1975.

[2] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, Amsterdam, second edition, 2003. This new edition is not a superset of [1]: some less important material is deleted.

[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), 1943. 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] Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.

[6] Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43:211–246, 2001.

[7] David Blackwell and Lester Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–886, 1962.

[8] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, 7:200–217, 1967.

[9] Nicolò Cesa-Bianchi, Philip M. Long, and Manfred K. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.

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

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

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

[13] A. Philip Dawid. Calibration-based empirical probability (with discussion). *Annals of Statistics*, 13:1251–1285, 1985.

[14] A. Philip Dawid. Proper measures of discrepancy, uncertainty and dependence, with applications to predictive experimental design. Technical Report 139, Department of Statistical Science, University College London, November 1994. This technical report was revised (and its title was slightly changed) in August 1998.

[15] A. Philip Dawid. Probability, causality and the empirical world: a Bayes–de Finetti–Popper–Borel synthesis. *Statistical Science*, 19:44–57, 2004.

[16] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer, New York, 1996.

[17] Jacques Ellul. *The Technological Bluff*. Eerdmans, Grand Rapids, MI, 1990. Translated by Geoffrey W. Bromiley. The French original: *Le bluff technologique*, Hachette, Paris, 1988.

[18] Ryszard Engelking. *General Topology*, volume 6 of *Sigma Series in Pure Mathematics*. Heldermann, Berlin, second edition, 1989.

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

[20] David P. Helmbold, Jyrki Kivinen, and Manfred K. Warmuth. Relative loss bounds for single neurons. *IEEE Transactions on Neural Networks*, 10:1291–1304, 1999.

[21] Mark Herbster and Manfred K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

[22] Yury M. Kabanov, Robert Sh. Liptser, and Albert N. Shiryaev. To the question of absolute continuity and singularity of probability measures. *Mathematics of the USSR—Sbornik*, 33:203–221, 1977.

[23] Jyrki Kivinen and Manfred K. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45:301–329, 2001.

[24] Leonid A. Levin. On the notion of a random sequence. *Soviet Mathematics Doklady*, 14:1413–1416, 1973.

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

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

[27] Saburou Saitoh. *Integral Transforms, Reproducing Kernels and their Applications*, volume 369 of *Pitman Research Notes in Mathematics*. Longman, Harlow, England, 1997.

[28] Claus P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit*. Springer, Berlin, 1971.

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

[30] Ray J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24:422–432, 1978.

[31] Ingo Steinwart, Don Hush, and Clint Scovel. Function classes that approximate the Bayes risk. In Gábor Lugosi and Hans Ulrich Simon, editors, *Proceedings of the Nineteenth Annual Conference on Learning Theory*, volume 4005 of *Lecture Notes in Artificial Intelligence*, pages 79–93, Berlin, 2006. Springer.

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

[33] Vladimir Vovk. On a randomness criterion. *Soviet Mathematics Doklady*, 35:656–660, 1987.

[34] Vladimir Vovk. Probability theory for the Brier game. *Theoretical Computer Science*, 261:57–79, 2001.

[35] Vladimir Vovk. On-line regression competitive with reproducing kernel Hilbert spaces, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #11, November 2004.

[36] VladimirVovk. Competing with wild prediction rules, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #16, December 2005.

[37] VladimirVovk. Defensive prediction with expert advice, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #14, June 2005. First posted in May 2005.

[38] VladimirVovk. Non-asymptotic calibration and resolution, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #13, May 2005. First posted in November 2004.

[39] Vladimir Vovk. Competing with Markov prediction strategies. Technical Report `arXiv:cs.LG/0607136`, `arXiv.org` e-Print archive, July 2006.

[40] Vladimir Vovk. Competing with stationary prediction strategies. Technical Report `arXiv:cs.LG/0607067`, `arXiv.org` e-Print archive, July 2006.

[41] VladimirVovk. Predictions as statements and decisions, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #17, June 2006.

[42] Vladimir Vovk, Ilia Nouretdinov, Akimichi Takemura, and Glenn Shafer. Defensive forecasting for linear protocols, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #10, February 2005.

[43] Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #8, September 2004 (revised January 2005). A shorter version is published in the Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics and available electronically at `http://www.gatsby.ucl.ac.uk/aistats/`.