Mach Learn (2007) 69: 193–212
DOI 10.1007/s10994-007-5021-y

Competing with wild prediction rules

Vladimir Vovk

Received: 21 September 2006 / Revised: 17 July 2007 / Accepted: 7 August 2007 /
Published online: 26 September 2007

Abstract We consider the problem of on-line prediction competitive with a benchmark class of continuous but highly irregular prediction rules. It is known that if the benchmark class is a reproducing kernel Hilbert space, there exists a prediction algorithm whose average loss over the first N examples does not exceed the average loss of any prediction rule in the class plus a "regret term" of $O(N^{-1/2})$. The elements of some natural benchmark classes, however, are so irregular that these classes are not Hilbert spaces. In this paper we develop Banach-space methods to construct a prediction algorithm with a regret term of $O(N^{-1/p})$, where $p \in [2, \infty)$ and $p - 2$ reflects the degree to which the benchmark class fails to be a Hilbert space. Only the square loss function is considered.

Keywords Competitive on-line prediction · Square loss regression · Banach function space

1 Introduction

For simplicity, in this introductory section we only discuss the problem of predicting real-valued labels y_n of objects $x_n \in [0, 1]$ (this will remain our main example throughout the paper). In this paper we are mainly interested in extending the class of the prediction rules our algorithms are competitive with; in other respects, our assumptions are rather restrictive. For example, we always assume that the labels y_n are bounded in absolute value by a known positive constant Y and only consider the problem of square-loss regression (some ideas for extension to a wider range of loss functions can be found in Vovk 2005).

Standard methods allow one to construct a "universally consistent" on-line prediction algorithm, i.e., an on-line prediction algorithm whose average loss over the first N examples does not exceed the average loss of any continuous prediction rule plus $o(1)$. (Such methods were developed in, e.g., Cesa-Bianchi et al. 1996; Kivinen and Warmuth 1997, and, especially, Auer et al. 2002, Sect. 3.2; for an explicit statement see Vovk 2006a.) More specifi-

Editors: Hans Ulrich Simon, Gabor Lugosi, Avrim Blum.

V. Vovk (✉)
Computer Learning Research Centre, Department of Computer Science, Royal Holloway,
University of London, Egham, Surrey TW20 0EX, England, UK
e-mail: vovk@cs.rhul.ac.uk

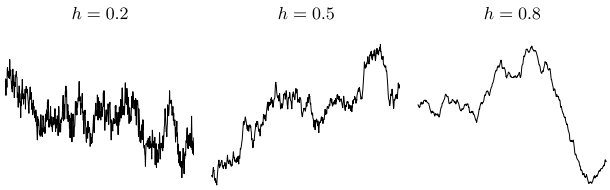

Fig. 1 Functions with Hölder exponent h for three different values of h

cally, for any reproducing kernel Hilbert space (RKHS) on $[0, 1]$ one can construct an on-line prediction algorithm whose average loss does not exceed that of any prediction rule in the RKHS plus $O(N^{-1/2})$; choosing a universal RKHS (Steinwart 2001, Definition 4) gives universal consistency. In this paper we are interested in extending the latter result, which is much more specific than the $o(1)$ provided by universal consistency, to wider benchmark classes of prediction rules. First we discuss limitations of RKHS as benchmark classes.

The regularity of a prediction rule D can be measured by its "Hölder exponent" h, which is informally defined by the condition that $|D(x + dx) - D(x)|$ scale as $|dx|^h$ for small $|dx|$. The most regular continuous functions are those of classical analysis: say, piecewise differentiable with bounded derivatives. For such functions the Hölder exponent is 1. Familiar examples are $x \mapsto \sin x$ and $x \mapsto |x - 1/2|$. Functions much less regular than those of classical analysis are ubiquitous in probability theory: for example, typical trajectories of the Brownian motion (more generally, of non-degenerate diffusion processes) have Hölder exponent $1/2$. Functions with other Hölder exponents $h \in (0, 1)$ can be obtained as typical trajectories of the fractional Brownian motion. Three examples with different values of h are shown in Fig. 1.

Fix a threshold $s \in (0, 1)$. The simplest and most intuitive formalization of the functions with Hölder exponent $h \geq s$ is provided by the function class $\mathcal{C}^s([0, 1])$ consisting of the functions f satisfying $|f(x) - f(y)| = O(|x - y|^s)$. The classes $\mathcal{C}^s([0, 1])$ are called *Hölder spaces* and the elements of $\mathcal{C}^s([0, 1])$ are called *Hölder continuous functions of order s*. The Hölder spaces are nested, $\mathcal{C}^s([0, 1]) \subset \mathcal{C}^{s'}([0, 1])$ when $s' < s$; they are very different from each other, as can be seen from the fact that typical trajectories of the fractional Brownian motion $B^{(h)}$ are in $\mathcal{C}^s([0, 1])$ for $s < h$ and outside $\mathcal{C}^s([0, 1])$ for $s > h$. As we will see in a moment, the standard Hilbert-space methods only work for $\mathcal{C}^s([0, 1])$ with $s > 1/2$ as benchmark classes; our goal is to develop methods that would work for smaller s as well.

It might be argued that the spaces $\mathcal{C}^s([0, 1])$ poorly reflect the intuitive notion of Hölder exponent: they are defined in terms of $\sup_{x,y} |f(x) - f(y)|/|x - y|^s$, and f's behavior in the neighborhood of a single point might too easily disqualify it from being a member of $\mathcal{C}^s([0, 1])$. Replacing sup with a mean (in the sense of L^p) gives the Slobodetsky spaces $B^s_p([0, 1])$ for $p \in [1, \infty]$ (see, e.g., Triebel 1992, 1.2.4, for the formal definition; in the next section we will be discussing much more general spaces). When $p = \infty$, the Slobodetsky spaces reduce to the Hölder spaces, $\mathcal{C}^s([0, 1]) = B^s_\infty([0, 1])$. Results for the case $p < \infty$ immediately carry over to $p = \infty$ since, as we will see in the next section, $\mathcal{C}^s([0, 1]) \subseteq B^{s'}_p([0, 1])$ whenever $s' < s$; s' can be arbitrarily close to s.

All Slobodetsky spaces (including the Hölder spaces) are Banach spaces, but $B^s_2([0, 1])$ are also Hilbert spaces and, for $s > 1/2$, even RKHS. Therefore, they are amenable to the standard methods (see the papers mentioned above; the exposition of Vovk (2006a) is especially close to that of this paper).

The condition $s > 1/p$ appears indispensable in the development of the theory (cf. the reference to the Sobolev embedding theorem in the next section). Since this paper concentrates on the irregular end of the Hölder spectrum, $s < 1/2$, instead of Hilbert spaces, such

as $B_2^s([0,1])$, we now have to deal with Banach spaces, such as $B_p^s([0,1])$ for $p \in (2,\infty)$, which are not Hilbert spaces. The necessary tools are developed in Sects. 3 and 4.

The methods used in Vovk (2006a) relied on the perfect shape of the unit ball in a Hilbert space. If p is not very far from 2, the unit ball in $B_p^s([0,1])$ is not longer perfectly round but still convex enough to allow us to obtain similar results by similar methods. In principle, the condition $s > 1/p$ is not longer an obstacle to coping with any $s > 0$: by taking a large enough p we can reach arbitrarily small s. However, the quality of prediction (at least as judged by our bound) will deteriorate: as we will see (Theorem 1 in the next section), the average loss of our prediction algorithm does not exceed that of any prediction rule in $B_p^s([0,1])$ plus $O(N^{-1/p})$. (This gives a regret term of $O(N^{-s+\epsilon})$ for the prediction rules in $\mathcal{C}^s([0,1])$, where $s \le 1/2$ and $\epsilon > 0$.)

This paper is the journal version of Vovk (2006b). The main difference from the conference version is that Theorem 1 is now applied to a much wider range of standard function spaces, including very smooth spaces (although for such spaces methods based on metric entropy may give better results—cf. Vovk 2006c). Section 3 of the conference version has been removed (to a large degree, it was an aside; besides, extending its results to smoother function spaces would lead to awkward statements).

2 Main result

We consider the following perfect-information prediction protocol:

FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Predictor announces $\mu_n \in \mathbb{R}$.
  Reality announces $y_n \in [-Y, Y]$.
END FOR.

At the beginning of each round n Predictor is given an object x_n whose label is to be predicted. The set of *a priori* possible objects, the *object space*, is denoted $\mathbf{X}$; we always assume $\mathbf{X} \neq \emptyset$. After Predictor announces his prediction μ_n for the object's label he is shown the actual label $y_n \in [-Y, Y]$. We consider the problem of regression, $y_n \in \mathbb{R}$, assuming an upper bound $Y > 0$ on $|y_n|$. The pairs (x_n, y_n) are called *examples*.

Predictor's loss on round n is measured by $(y_n - \mu_n)^2$, and so his average loss after N rounds of the game is $\frac{1}{N}\sum_{n=1}^N (y_n - \mu_n)^2$. His goal is to have

$$\frac{1}{N}\sum_{n=1}^N (y_n - \mu_n)^2 \lessapprox \frac{1}{N}\sum_{n=1}^N (y_n - D(x_n))^2$$

($\lessapprox$ meaning "is less than or approximately equal to") for each prediction rule $D : \mathbf{X} \to \mathbb{R}$ that is not "too wild".

2.1 Main theorem

Our main theorem will be fairly general and applicable to a wide range of Banach function spaces. Its implications for some of the standard function spaces will be explained after its statement.

Let U be a Banach space and $S_U := \{u \in U \mid \|u\|_U = 1\}$ be the unit sphere in U. Our methods are applicable only to Banach spaces whose unit spheres do not have very flat areas; a convenient measure of rotundity of S_U is Clarkson's (1936) modulus of convexity

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

(we will be mostly interested in the small values of ϵ).

Let us say that a Banach space $\mathcal{F}$ of real-valued functions f on $\mathbf{X}$ (with the standard pointwise operations of addition and of multiplication by scalar) is a *proper Banach functional space* (PBFS) on $\mathbf{X}$ if, for each $x \in \mathbf{X}$, the evaluation functional $\mathbf{k}_x : f \in \mathcal{F} \mapsto f(x)$ is continuous. We will assume that

$$\mathbf{c}_{\mathcal{F}} := \sup_{x \in \mathbf{X}} \|\mathbf{k}_x\|_{\mathcal{F}^*} < \infty, \tag{2}$$

where $\mathcal{F}^*$ is the dual Banach space (see, e.g., Rudin 1991, Chap. 4).

The following theorem will be proved in Sects. 3 and 4.

Theorem 1 *Let $\mathcal{F}$ be a proper Banach functional space such that*

$$\forall \epsilon \in (0,2] : \delta_{\mathcal{F}}(\epsilon) \geq (\epsilon/2)^p/p \tag{3}$$

for some $p \in [2,\infty)$. There exists a prediction algorithm producing $\mu_n \in [-Y,Y]$ that are guaranteed to satisfy

$$\frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_n)^2 \leq \frac{1}{N}\sum_{n=1}^{N}(y_n - D(x_n))^2 + 40Y\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1}(\|D\|_{\mathcal{F}} + Y)N^{-1/p} \tag{4}$$

for all $N = 1, 2, \ldots$ and all $D \in \mathcal{F}$.

Conditions (2) and (3) are satisfied for the Slobodetsky spaces $B_p^s(\mathbf{X})$, which we will now introduce.

2.2 Besov and Triebel–Lizorkin spaces

Suppose $\mathbf{X}$ is a bounded Lipschitz domain in $\mathbb{R}^m$ (for a definition see, e.g., Triebel 2005, Definition 3). Two standard scales of function spaces are the Besov spaces $B_{p,q}^s(\mathbf{X})$ and the Triebel–Lizorkin spaces $F_{p,q}^s(\mathbf{X})$; in this paper we do not define them (see, e.g., Triebel 1992, especially Chap. I, for the definition) but describe all their properties that we need. In principle, the allowed values of the parameters are $s \in \mathbb{R}$ and $p, q \in (0,\infty]$ (with $p = \infty$ sometimes excluded from the Triebel–Lizorkin scale); however, they are Banach spaces only when $p, q \in [1,\infty]$ (otherwise they are only guaranteed to be quasi-Banach spaces). We will be interested in the case $s \geq 0$ and $p, q \in [1,\infty]$.

These are some of the important special cases of the two scales (for other special cases, see, e.g., Triebel 1992, Chap. 1, and Edmunds and Triebel 1996, 2.2.2):

- the Slobodetsky spaces $B_p^s(\mathbf{X}) := B_{p,p}^s(\mathbf{X}) = F_{p,p}^s(\mathbf{X})$ (for the equality $B_{p,p}^s(\mathbf{X}) = F_{p,p}^s(\mathbf{X})$ see, e.g., Adams and Fournier 2003, 7.67);
- the Hölder–Zygmund spaces $\mathcal{C}^s(\mathbf{X}) := B_\infty^s(\mathbf{X})$ (also called Hölder spaces when s is not an integer number);

– the Bessel potential spaces $H^s_p(\mathbf{X}) := F^s_{p,2}(\mathbf{X})$ (also called Liouville or fractional Sobolev spaces).

In the conference version (Vovk 2006b) of this paper, it was the spaces $B^s_p(\mathbf{X})$ rather than $H^s_p(\mathbf{X})$ that were denoted $W^{s,p}(\mathbf{X})$ and called Sobolev spaces (in the spirit of older literature, such as Nikolsky 1961 or Triebel 1983).

It should be said that in the theory of function spaces one usually does not distinguish between equivalent norms of a given Banach space; in this paper we also adopt this a little sloppy convention (as Hans Triebel describes it in Triebel 1992, 1.2.5).

Let $C(\overline{\mathbf{X}})$ be the Banach space of continuous functions $f : \mathbf{X} \to \mathbb{R}$ with finite norm $\|f\|_{C(\overline{\mathbf{X}})} := \sup_{x\in\mathbf{X}} |f(x)|$ that can be continuously extended to the closure $\overline{\mathbf{X}}$ of $\mathbf{X}$. The Sobolev embedding theorem shows that for $s > m/p$ the function spaces $B^s_{p,q}(\mathbf{X})$ and $F^s_{p,q}(\mathbf{X})$ are continuously embedded in the space $C(\overline{\mathbf{X}})$ (the relevant part of the Sobolev embedding theorem is stated in, e.g., Triebel 2005, Proposition 7(ii); there are other parts of the Sobolev embedding theorem, dealing with the case where the condition $s > m/p$ is not satisfied). In essence, this means that for $s > m/p$ the elements of $B^s_{p,q}(\mathbf{X})$ and $F^s_{p,q}(\mathbf{X})$ are continuous functions that can be extended to $\overline{\mathbf{X}}$ and that the identity mapping from those spaces to $C(\overline{\mathbf{X}})$ is bounded; the latter can be equivalently expressed by the formulas

$$\mathbf{c}_{B^s_{p,q}(\mathbf{X})} < \infty, \qquad \mathbf{c}_{F^s_{p,q}(\mathbf{X})} < \infty$$

as $\mathbf{c}_{\mathcal{F}}$ is just the norm of the embedding $\mathcal{F} \hookrightarrow C(\mathbf{X})$ for any PBFS $\mathcal{F}$ on $\mathbf{X}$.

We are only interested in the case $s > m/p$, and so in view of the Sobolev embedding theorem we sometimes write the argument of $B^s_{p,q}$, $F^s_{p,q}$, and their subclasses as $\overline{\mathbf{X}}$ rather than $\mathbf{X}$ (as we did in Sect. 1).

We can now deduce the following corollary from Theorem 1. It is shown by Cobos and Edmunds (1988, Theorem 3) that (3) is satisfied for the Besov and Triebel–Lizorkin spaces $B^s_{p,q}(\mathbf{X})$, $F^s_{p,q}(\mathbf{X})$ provided $p \in [2,\infty)$ and $q \in [p', p]$, where p' is the conjugate index, defined by the condition $1/p + 1/p' = 1$. In particular, the Slobodetsky spaces $B^s_p(\mathbf{X})$ satisfy (3) for $p \geq 2$. So let $p \in [2,\infty)$ and $s \in (m/p, \infty)$. There exists a constant $C_{s,p} > 0$ and a prediction algorithm producing $\mu_n \in [-Y, Y]$ that are guaranteed to satisfy

$$\frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_n)^2 \leq \frac{1}{N}\sum_{n=1}^{N}(y_n - D(x_n))^2 + YC_{s,p}(\|D\|_{B^s_p(\mathbf{X})} + Y)N^{-1/p} \tag{5}$$

for all $N = 1, 2, \ldots$ and all $D \in B^s_p(\mathbf{X})$.

Remark In fact, Cobos and Edmunds (1988) do not state their results in terms of Clarkson's modulus of convexity; however, it is very easy to deduce (3) for $\mathcal{F} := A^s_{p,q}(\mathbf{X})$, with $A \in \{B, F\}$ and suitable p and q, from their Theorem 3. One of the inequalities in that theorem (combined with the remark following it) is

$$\left(\frac{1}{2}\|f-g\|^p_{A^s_{p,q}(\mathbf{X})} + \frac{1}{2}\|f+g\|^p_{A^s_{p,q}(\mathbf{X})}\right)^{1/p} \leq (\|f\|^{p'}_{A^s_{p,q}(\mathbf{X})} + \|g\|^{p'}_{A^s_{p,q}(\mathbf{X})})^{1/p'}$$

where $f, g \in A^s_{p,q}(\mathbf{X})$ for $A \in \{B, F\}$, $p \in [2,\infty)$, and $q \in [p', p]$. Taking f and g on the unit sphere at a distance of ϵ from each other and setting $h := (f+g)/2$, we obtain

$$\left(\frac{1}{2}\epsilon^p + \frac{1}{2}\|2h\|^p_{A^s_{p,q}(\mathbf{X})}\right)^{1/p} \leq 2^{1/p'},$$

which is equivalent to

$$1 - \|h\|_{A^s_{p,q}(\mathbf{X})} \geq 1 - (1 - (\epsilon/2)^p)^{1/p} \geq (\epsilon/2)^p / p \tag{6}$$

(the last inequality is a special case of $(1-t)^{1/p} \leq 1 - t/p$ valid for $t \in [0,1]$ and $p \geq 1$; to check it, notice that the left-hand side is a concave function of t, and the values and derivatives of the two sides match when $t = 0$). In the case of the spaces L^p this result was obtained by Clarkson (1936, Sect. 3), and Clarkson's bound (before applying the final inequality in (6), of course) was shown to be optimal by Hanner (1956). Cobos and Edmunds's result was further generalized by Takahashi and Kato (1997).

In informal discussions below we will continue to call terms such as the second addend on the right-hand side of (5) the "regret term", and say that the corresponding prediction algorithm is "R-competitive", where R is the regret term.

According to (4), we can take

$$C_{s,p} = 40\sqrt{\mathbf{c}^2_{B^s_p(\mathbf{X})} + 1}$$

in (5), but in fact

$$C_{s,p} = 4 \times 8.68^{1-1/p}\sqrt{\mathbf{c}^2_{B^s_p(\mathbf{X})} + 1} \tag{7}$$

will suffice (see (41) below). In the special case $p = 2$ one can use Hilbert-space methods to improve (7), which now becomes, approximately,

$$11.78\sqrt{\mathbf{c}^2_{B^s_2(\mathbf{X})} + 1}, \tag{8}$$

to

$$2\sqrt{\mathbf{c}^2_{B^s_2(\mathbf{X})} + 1} \tag{9}$$

(Vovk 2006a, Theorem 1); using Banach-space methods we have lost a factor of 5.89.

2.3 Application to the Hölder–Zygmund functions

Let us apply (5) to the Hölder–Zygmund classes $\mathcal{C}^s(\mathbf{X}) := B^s_\infty(\mathbf{X})$. In the case $\mathbf{X} = [0,1]$ and $s \in (0,1)$, the norm in $\mathcal{C}^s(\mathbf{X})$ is equivalent to

$$\|f\|_{\mathcal{C}^s(\mathbf{X})} = \max\left(\sup_{x \in \mathbf{X}} |f(x)|, \sup_{x,y \in \mathbf{X}: x \neq y} \left|\frac{f(x) - f(y)}{|x - y|^s}\right|\right)$$

(and the space $\mathcal{C}^s(\mathbf{X})$ consists of the functions f with finite norm); for a general definition see, e.g., Triebel (1992), 1.2.2.

More generally, let us again assume that $\mathbf{X} \subseteq \mathbb{R}^m$ is a bounded Lipschitz domain. The Sobolev embedding theorem (see Edmunds and Triebel 1996, 2.5.1) implies that there is a continuous embedding

$$\mathcal{C}^s(\mathbf{X}) \hookrightarrow B^{s'}_p(\mathbf{X}) \tag{10}$$

for any $s' < s$ and any p.

Suppose $s \le m/2$ and fix an arbitrarily small $\epsilon > 0$. Applying (5) to $B_p^{s'}(\mathbf{X})$ with $p > m/s$ sufficiently close to $m/s \ge 2$ and to $s' \in (m/p, s)$, we can see from (10) that there exists a constant $C_{s,\epsilon} > 0$ such that

$$\frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_n)^2 \le \frac{1}{N}\sum_{n=1}^{N}(y_n - D(x_n))^2 + YC_{s,\epsilon}(\|D\|_{\mathcal{C}^s(\mathbf{X})} + Y)N^{-s/m+\epsilon} \tag{11}$$

holds for all $N = 1, 2, \ldots$ and all $D \in \mathcal{C}^s(\mathbf{X})$.

3 More geometry of Banach spaces

In the proof of Theorem 1 we will need not only Clarkson's modulus of convexity (1) but a whole range of different moduli of convexity and smoothness. In our description we will often follow Lindenstrauss and Tzafriri (1979); for information about other moduli and further references, see Fuster (2006). We will only consider Banach spaces of dimension at least 2.

3.1 Moduli of convexity and smoothness

A natural modification of Clarkson's modulus of convexity was proposed by Gurary (1967):

$$\delta_U^\dagger(\epsilon) := \inf_{\substack{u,v\in S_U \\ \|u-v\|_U=\epsilon}} \left(1 - \inf_{t\in[0,1]} \|tu + (1-t)v\|_U\right). \tag{12}$$

It is clear that

$$\delta_U(\epsilon) \le \delta_U^\dagger(\epsilon) \le 2\delta_U(\epsilon)$$

(cf. the proof of Lemma 2 below), and it was shown recently (Bárcenas et al. 2004) that this relation cannot be improved.

The standard modulus of smoothness was proposed by Lindenstrauss (1963):

$$\rho_U(\tau) := \sup_{u,v\in S_U} \left(\frac{\|u + \tau v\|_U + \|u - \tau v\|_U}{2} - 1\right), \quad \tau > 0. \tag{13}$$

Lindenstrauss also established a simple but very useful relation of conjugacy (cf. Rockafellar 1970, Sect. 12, although δ is not always convex, as shown by Liokumovich 1973) between δ and ρ:

$$\rho_{U^*}(\tau) = \sup_{\epsilon\in(0,2]} \left(\frac{\epsilon\tau}{2} - \delta_U(\epsilon)\right); \tag{14}$$

we can see that $2\rho_{U^*}$ is the Fenchel transform of $2\delta_U$.

The following inequality will be the basis of the proof of Theorem 1 in the next section. Suppose a PBFS $\mathcal{F}$ satisfies the condition (3) of Theorem 1. By (14) we obtain for the dual space $\mathcal{F}^*$ to $\mathcal{F}$, assuming $\tau \in (0, 1]$:

$$\rho_{\mathcal{F}^*}(\tau) \le \sup_{\epsilon\in(0,2]} \left(\frac{\epsilon\tau}{2} - (\epsilon/2)^p/p\right) = \tau^q/q, \tag{15}$$

where $q := p' = p/(p-1)$ is the conjugate index (the supremum in (15) is attained at $\epsilon = 2\tau^{1/(p-1)}$).

The Banach space U is called *uniformly convex* if $\delta_U(\epsilon) > 0$ for all $\epsilon \in (0, 2]$, and it is called *uniformly smooth* if $\rho_U(\tau)/\tau \to 0$ as $\tau \to 0$. All uniformly convex and all uniformly smooth Banach spaces U are reflexive (i.e., $U^{**} = U$; see, e.g., Lindenstrauss and Tzafriri 1979, Proposition 1.e.3 on p. 61).

If V is a Hilbert space, the "parallelogram identity"

$$\|u+v\|_V^2 + \|u-v\|_V^2 = 2\|u\|_V^2 + 2\|v\|_V^2 \tag{16}$$

immediately gives

$$\delta_V(\epsilon) = 1 - \sqrt{1-(\epsilon/2)^2} \geq \epsilon^2/8$$

and

$$\rho_V(\tau) = \sqrt{1+\tau^2} - 1 \leq \tau^2/2. \tag{17}$$

Nördlander (1960) proved that the unit balls in Hilbert spaces are most convex and smooth: if U is a Banach space and V is a Hilbert space,

$$\begin{aligned} \delta_U(\epsilon) &\leq \delta_V(\epsilon) = 1 - \sqrt{1-(\epsilon/2)^2}, \\ \rho_U(\tau) &\geq \rho_V(\tau) = \sqrt{1+\tau^2} - 1. \end{aligned} \tag{18}$$

The original definitions (1) and (13) of the moduli of convexity and smoothness look very different, and Banaś (1986) proposed a definition of modulus of smoothness similar to (1):

$$\rho_U^\dagger(\tau) := \sup_{\substack{u,v \in S_U \\ \|u-v\|_U = \tau}} \left(1 - \left\|\frac{u+v}{2}\right\|_U\right), \quad \tau \in (0,2). \tag{19}$$

The difference $\rho_U^\dagger(\epsilon) - \delta_U(\epsilon)$ measures the degree to which (the unit ball in) U is deformed (Banaś and Frączek 1993; it is always zero for Hilbert spaces). What we will need in this paper is the modification of (19) in the direction of (12):

$$\rho_U^\ddagger(\tau) := \sup_{\substack{u,v \in S_U \\ \|u-v\|_U = \tau}} \sup_{t \in [0,1]} (1 - \|tu + (1-t)v\|_U), \quad \tau \in (0,2). \tag{20}$$

Since the standard results about moduli of convexity and smoothness are about the definitions (1) and (13), we first need to establish connections between (13) and (20). The first of these results appears in Banaś 1986 (but we still prove it since Banaś (1986) is less easily accessible than most other papers in our bibliography).

Lemma 1 (Banaś 1986) *For all* $\tau \in (0,2)$,

$$\frac{\rho_U^\dagger(\tau)}{1-\rho_U^\dagger(\tau)} \leq \rho_U\left(\frac{\tau}{2(1-\rho_U^\dagger(\tau))}\right). \tag{21}$$

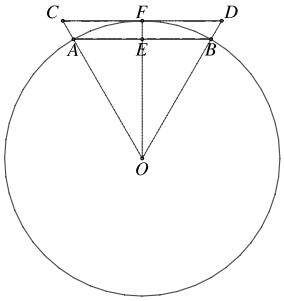

Fig. 2 Relation between ρ and $\rho^\dagger$

Proof Let $c < \rho_U^\dagger(\tau)$ be such that, for some $u, v \in S_U$ satisfying $\|u - v\|_U = \tau$,

$$\left\| \frac{u+v}{2} \right\|_U = 1 - c$$

(it is clear that c can be chosen as close to $\rho_U^\dagger(\tau)$ as we wish). Set

$$u' := \frac{1}{1-c}\frac{u+v}{2}, \qquad v' := \frac{v-u}{\|u-v\|_U}, \qquad \tau' := \frac{1}{1-c}\frac{\tau}{2}$$

(cf. Fig. 2, where $\overrightarrow{OA} = u$, $\overrightarrow{OB} = v$, $\overrightarrow{OE} = (u+v)/2$, $\overrightarrow{OF} = u'$, and $\overrightarrow{FD} = \tau' v'$). Since $u', v' \in S_U$, we have

$$\rho_U(\tau') \geq \frac{\|u' + \tau' v'\|_U + \|u' - \tau' v'\|_U}{2} - 1 = \frac{1}{1-c} - 1,$$

which can be rewritten as

$$\rho_U\left(\frac{\tau}{2(1-c)}\right) \geq \frac{c}{1-c}.$$

Letting $c \to \rho_U^\dagger(\tau)$ completes the proof (the modulus of smoothness is continuous by, e.g., Lindenstrauss and Tzafriri 1979, Proposition 1.e.5 on p. 64). □

Corollary 1 *For all* $\tau \in (0, 1]$,

$$\rho_U^\dagger(\tau) \leq \rho_U(\tau). \tag{22}$$

Proof Let $\tau \in (0, 1]$. Following Banaś (1986), proof of Lemma 1, we obtain

$$\begin{aligned} \rho_U^\dagger(\tau) &= \sup_{\substack{u,v \in S_U \\ \|u-v\|_U = \tau}} \frac{2\|u\|_U - \|u+v\|_U}{2} \\ &\leq \sup_{\substack{u,v \in S_U \\ \|u-v\|_U = \tau}} \frac{\|u+v\|_U + \|u-v\|_U - \|u+v\|_U}{2} = \frac{\tau}{2} \leq \frac{1}{2} \end{aligned}$$

(the first inequality following from the triangle inequality). We can now easily deduce (22) from (21) and the fact that ρ_U is a non-decreasing function (Lindenstrauss and Tzafriri 1979,

Proposition 1.e.5):

$$\rho_U^\dagger(\tau) \le \frac{\rho_U^\dagger(\tau)}{1-\rho_U^\dagger(\tau)} \le \rho_U\left(\frac{\tau}{2(1-\rho_U^\dagger(\tau))}\right) \le \rho_U(\tau). \qquad \square$$

Lemma 2 *For all* $\tau \in (0,2)$,

$$\rho_U^\ddagger(\tau) \le 2\rho_U^\dagger(\tau).$$

Proof Suppose $\rho_U^\ddagger(\tau) > c$. Let $u, v \in S_U$ and $t \in [0,1]$ be such that $\|u-v\|_U = \tau$ and

$$\|tu + (1-t)v\|_U < 1-c.$$

Without loss of generality we assume $t \le 1/2$. Since

$$\begin{aligned}\left\|\frac{u+v}{2}\right\|_U &= \left\|\frac{1-2t}{2-2t}u + \frac{1}{2-2t}(tu+(1-t)v)\right\|_U \\ &\le \frac{1-2t}{2-2t}\|u\|_U + \frac{1}{2-2t}\|tu+(1-t)v\|_U < \frac{1-2t}{2-2t} + \frac{1}{2-2t}(1-c) \\ &= \frac{2-2t-c}{2-2t} \le \frac{2-c}{2} = 1-\frac{c}{2},\end{aligned}$$

we have $\rho_U^\dagger(\tau) > c/2$. $\square$

3.2 Direct sums of uniformly smooth spaces

If U_1 and U_2 are two Banach spaces, their *weighted direct sum* $U_1 \oplus U_2$ is defined to be the Cartesian product $U_1 \times U_2$ with the operations of addition and multiplication by scalar defined by

$$(u_1,u_2)+(u_1',u_2') := (u_1+u_1', u_2+u_2'), \quad c(u_1,u_2) := (cu_1, cu_2);$$

we will equip it with the norm

$$\|(u_1,u_2)\|_{U_1\oplus U_2} := \sqrt{a_1\|u_1\|_{U_1}^2 + a_2\|u_2\|_{U_2}^2}, \tag{23}$$

where a_1 and a_2 are positive constants (to simplify formulas, we do not mention them explicitly in our notation for $U_1 \oplus U_2$). The operation of weighted direct sum provides a means of merging different Banach spaces, which plays an important role in our proof technique (cf. Vovk 2006a, Corollary 4). The "Euclidean" definition (23) of the norm in the direct sum suggests that the sum will be as smooth as the components; this intuition is formalized in the following lemma (essentially a special case of Proposition 17 in Figiel 1976, p. 132).

Lemma 3 *If* U_1 *and* U_2 *are Banach spaces and* $f : (0,1] \to \mathbb{R}$,

$$\begin{aligned}&(\forall \tau \in (0,1]: \quad \rho_{U_1}(\tau) \le f(\tau) \quad \& \quad \rho_{U_2}(\tau) \le f(\tau)) \\ &\implies \quad (\forall \tau \in (0,1]: \quad \rho_{U_1\oplus U_2}(\tau) \le 4.34 f(\tau)).\end{aligned}$$

Proof We will follow the proof of Proposition 17 in Figiel (1976), which is based on the following weak form of the parallelogram identity (16), valid for all Banach spaces:

$$\begin{aligned}&\|u+v\|_U^2+\|u-v\|_U^2-2\|u\|_U^2-2\|v\|_U^2\\&\quad\le 2\|u\|_U(\|u+v\|_U+\|u-v\|_U-2\|u\|_U)\end{aligned}\tag{24}$$

(see Figiel 1976, Lemma 16 on p. 132); it is clear that (24) implies

$$\|u+v\|_U^2+\|u-v\|_U^2-2\|u\|_U^2-2\|v\|_U^2\le 4\|u\|_U^2\rho_U(\|v\|_U/\|u\|_U).\tag{25}$$

Let $u^\dagger=(u_1,u_2)$ and $v^\dagger=(v_1,v_2)$ be arbitrary norm one vectors in $U_1\oplus U_2$. Applying (25) to $(u,v):=(u_1,\tau v_1)$ and $(u,v):=(u_2,\tau v_2)$, we obtain

$$\begin{aligned}&\|u_1+\tau v_1\|_{U_1}^2+\|u_1-\tau v_1\|_{U_1}^2-2\|u_1\|_{U_1}^2-2\tau^2\|v_1\|_{U_1}^2\\&\quad\le 4\|u_1\|_{U_1}^2\rho_{U_1}(\tau\|v_1\|_{U_1}/\|u_1\|_{U_1})\end{aligned}\tag{26}$$

and

$$\begin{aligned}&\|u_2+\tau v_2\|_{U_2}^2+\|u_2-\tau v_2\|_{U_2}^2-2\|u_2\|_{U_2}^2-2\tau^2\|v_2\|_{U_2}^2\\&\quad\le 4\|u_2\|_{U_2}^2\rho_{U_2}(\tau\|v_2\|_{U_2}/\|u_2\|_{U_2}).\end{aligned}\tag{27}$$

Multiplying (26) by a_1 and (27) by a_2 and summing the resulting inequalities now gives

$$\begin{aligned}&\|u^\dagger+\tau v^\dagger\|_{U_1\oplus U_2}^2+\|u^\dagger-\tau v^\dagger\|_{U_1\oplus U_2}^2-2-2\tau^2\\&\quad\le 4\sum_{j=1}^2 a_j\|u_j\|_{U_j}^2\rho_{U_j}(\tau\|v_j\|_{U_j}/\|u_j\|_{U_j}).\end{aligned}\tag{28}$$

To estimate the sum over $j=1,2$, notice that:

– when $\|v_j\|_{U_j}\le\|u_j\|_{U_j}$,

$$\rho_{U_j}(\tau\|v_j\|_{U_j}/\|u_j\|_{U_j})\le\rho_{U_j}(\tau)\|v_j\|_{U_j}/\|u_j\|_{U_j}$$

(by the convexity of ρ, following from the convexity of the Fenchel transform, (14), and the reflexivity of all uniformly convex and all uniformly smooth spaces);

– when $\|v_j\|_{U_j}>\|u_j\|_{U_j}$,

$$\rho_{U_j}(\tau\|v_j\|_{U_j}/\|u_j\|_{U_j})\le L\rho_{U_j}(\tau)(\|v_j\|_{U_j}/\|u_j\|_{U_j})^2$$

(where $L<3.18$ is a constant satisfying $\rho(\sigma)/\sigma^2\le L\rho(\tau)/\tau^2$ for all positive $\tau\le\sigma$; see Figiel 1976, Proposition 10 on p. 128 and the remark after its proof).

Using the Cauchy–Schwarz inequality, the sum can be bounded above as follows:

$$\begin{aligned}&\sum_{j=1}^2 a_j\|u_j\|_{U_j}^2\rho_{U_j}(\tau\|v_j\|_{U_j}/\|u_j\|_{U_j})\\&\quad\le\sum_{j=1}^2 a_j\|v_j\|_{U_j}\rho_{U_j}(\tau)\max(\|u_j\|_{U_j},L\|v_j\|_{U_j})\end{aligned}$$

$$\leq \left(\sum_{j=1}^{2} a_j \|v_j\|_{U_j}^2\right)^{1/2} \left(\sum_{j=1}^{2} a_j (\rho_{U_j}(\tau))^2 (\|u_j\|_{U_j}^2 + L^2 \|v_j\|_{U_j}^2)\right)^{1/2}$$

$$\leq \left(\sum_{j=1}^{2} f^2(\tau) a_j (\|u_j\|_{U_j}^2 + L^2 \|v_j\|_{U_j}^2)\right)^{1/2} = \sqrt{L^2+1}\, f(\tau) \tag{29}$$

(the last line assuming $\tau \in (0, 1]$). Now we have all we need to deduce the conclusion of the lemma (some steps will be explained after the equation): when $\tau \in (0, 1]$,

$$\frac{1}{2}(\|u^\dagger + \tau v^\dagger\|_{U_1 \oplus U_2} + \|u^\dagger - \tau v^\dagger\|_{U_1 \oplus U_2})$$

$$\leq \left(\frac{1}{2}(\|u^\dagger + \tau v^\dagger\|_{U_1 \oplus U_2}^2 + \|u^\dagger - \tau v^\dagger\|_{U_1 \oplus U_2}^2)\right)^{1/2}$$

$$\leq (1 + \tau^2 + 2\sqrt{L^2+1}\, f(\tau))^{1/2} \leq (1+\tau^2)^{1/2} + \sqrt{L^2+1}\, f(\tau)$$

$$\leq 1 + f(\tau) + \sqrt{L^2+1}\, f(\tau) = 1 + (1 + \sqrt{L^2+1}) f(\tau)$$

(the first inequality follows from the convexity of the function $t \mapsto t^2$, the second from (28) and (29), the third from the mean-value theorem, and the fourth from Nördlander's bound (18)). It remains to compare the resulting inequality with the definition of the modulus of convexity and remember that $L < 3.18$. □

4 Proof of Theorem 1

In this section we partly follow the proof of Theorem 1 in Vovk (2006a) (Sect. 6).

4.1 The BBK29 algorithm

Let U be a Banach space. We say that a function $\Phi : [-Y, Y] \times \mathbf{X} \to U$ is *forecast-continuous* if $\Phi(\mu, x)$ is continuous in $\mu \in [-Y, Y]$ for every fixed $x \in \mathbf{X}$. For such a Φ the function

$$f_n(y, \mu) := \left\| \sum_{i=1}^{n-1} (y_i - \mu_i) \Phi(\mu_i, x_i) + (y - \mu) \Phi(\mu, x_n) \right\|_U - \left\| \sum_{i=1}^{n-1} (y_i - \mu_i) \Phi(\mu_i, x_i) \right\|_U \tag{30}$$

is continuous in $\mu \in [-Y, Y]$.

BANACH-SPACE BALANCED K29 ALGORITHM (BBK29)
Parameter: forecast-continuous $\Phi : [-Y, Y] \times \mathbf{X} \to U$, with U a Banach space
FOR $n = 1, 2, \ldots$:
Read $x_n \in \mathbf{X}$.
Define $f_n : [-Y, Y]^2 \to \mathbb{R}$ by (30).
Output any root $\mu \in [-Y, Y]$ of $f_n(-Y, \mu) = f_n(Y, \mu)$ as μ_n;
if there are no such roots, output $\mu_n \in \{-Y, Y\}$

such that $\sup_{y\in[-Y,Y]} f_n(y,\mu_n) \le 0$.
Read $y_n \in [-Y,Y]$.
END FOR.

The validity of this description depends on the existence of $\mu \in \{-Y,Y\}$ satisfying $\sup_{y\in[-Y,Y]} f_n(y,\mu) \le 0$ when the equation $f_n(-Y,\mu) = f_n(Y,\mu)$ does not have roots $\mu \in [-Y,Y]$. The existence of such a μ is easy to check: if $f_n(-Y,\mu) < f_n(Y,\mu)$ for all $\mu \in [-Y,Y]$, take $\mu := Y$ to obtain

$$f_n(-Y,\mu) < f_n(Y,\mu) = 0$$

and, hence, $\sup_{y\in[-Y,Y]} f_n(y,\mu) \le 0$ by the convexity of (30) in y; if $f_n(-Y,\mu) > f_n(Y,\mu)$ for all $\mu \in [-Y,Y]$, setting $\mu := -Y$ leads to

$$f_n(Y,\mu) < f_n(-Y,\mu) = 0$$

and, hence, $\sup_{y\in[-Y,Y]} f_n(y,\mu) \le 0$. The parameter Φ of the BBK29 algorithm will sometimes be called the *feature mapping*.

Theorem 2 *Let Φ be a forecast-continuous mapping from $[-Y,Y]\times\mathbf{X}$ to a Banach space U and set $\mathbf{c}_\Phi := \sup_{\mu\in[-Y,Y],x\in\mathbf{X}} \|\Phi(\mu,x)\|_U$. Suppose $\rho_U(\tau) \le a\tau^q, \forall \tau \in (0,1]$, for some constants $q \ge 1$ and $a \ge 1/q$. The BBK29 algorithm with parameter Φ outputs $\mu_n \in [-Y,Y]$ such that*

$$\left\| \sum_{n=1}^{N} (y_n - \mu_n)\Phi(\mu_n, x_n) \right\|_U \le 2Y\mathbf{c}_\Phi (2aqN)^{1/q} \tag{31}$$

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

Proof Set

$$S_N := \left\| \sum_{n=1}^{N} (y_n - \mu_n)\Phi(\mu_n, x_n) \right\|_U;$$

our goal is to prove

$$S_N \le 2Y\mathbf{c}_\Phi (2aqN)^{1/q}.$$

For $N = 1$, this follows from

$$2Y\mathbf{c}_\Phi \le 2Y\mathbf{c}_\Phi (2aqN)^{1/q},$$

which in turn follows from $2aq \ge 1$, which in turn follows from the condition $a \ge 1/q$. It remains to prove that

$$S_{N-1} \le 2Y\mathbf{c}_\Phi (2aq(N-1))^{1/q}$$

implies

$$S_N \le 2Y\mathbf{c}_\Phi (2aqN)^{1/q} \tag{32}$$

for $N \ge 2$. Without loss of generality we assume that $f_N(-Y,\mu_N) = f_N(Y,\mu_N)$ and replace S_N in (32) by $F_N := S_{N-1} + f_N(Y,\mu_N)$ (using the convexity of $f_N(y,\mu_N)$ in y).

Fix $N \geq 2$. We will assume that

$$S_{N-1} \leq 2Y\mathbf{c}_\Phi(2aq(N-1))^{1/q} \quad \& \quad F_N > 2Y\mathbf{c}_\Phi(2aqN)^{1/q} \tag{33}$$

and arrive at a contradiction. Using $F_N > 2Y\|\Phi(\mu_N, x_N)\|$ (which follows from (33)), Corollary 1, Lemma 2, and the definition of $\rho^{\ddagger}$, we obtain:

$$\begin{aligned}
&2a\left(\frac{2Y\|\Phi(\mu_N, x_N)\|}{F_N}\right)^q \\
&\quad \geq 2\rho_U\left(\frac{2Y\|\Phi(\mu_N, x_N)\|}{F_N}\right) \\
&\quad \geq 2\rho_U^{\dagger}\left(\frac{2Y\|\Phi(\mu_N, x_N)\|}{F_N}\right) \geq \rho_U^{\ddagger}\left(\frac{2Y\|\Phi(\mu_N, x_N)\|}{F_N}\right) \\
&\quad \geq 1 - \left\| t\frac{\sum_{n=1}^{N-1}(y_n - \mu_n)\Phi(\mu_n, x_n) + (-Y - \mu_N)\Phi(\mu_N, x_N)}{F_N} \right. \\
&\qquad \left. + (1-t)\frac{\sum_{n=1}^{N-1}(y_n - \mu_n)\Phi(\mu_n, x_n) + (Y - \mu_N)\Phi(\mu_N, x_N)}{F_N} \right\| \\
&\quad = 1 - \frac{S_{N-1}}{F_N},
\end{aligned} \tag{34}$$

where the moduli of smoothness are understood to be zero at $\tau = 0$, and $t \in [0, 1]$ is chosen such that

$$t(-Y - \mu_N) + (1-t)(Y - \mu_N) = 0$$

(i.e., $t := \frac{1}{2} - \frac{\mu_N}{2Y}$). The inequality between the extreme terms of (34) can be rewritten as

$$S_{N-1} \geq F_N\left(1 - 2a\left(\frac{2Y\|\Phi(\mu_N, x_N)\|}{F_N}\right)^q\right).$$

As the right-hand side is a monotonically increasing function of F_N (which can be checked by differentiation), in combination with (33) the last inequality gives

$$2Y\mathbf{c}_\Phi(2aq(N-1))^{1/q} > 2Y\mathbf{c}_\Phi(2aqN)^{1/q}(1 - 2a((2aqN)^{-1/q})^q),$$

i.e.,

$$(N-1)^{1/q} > N^{1/q}\left(1 - \frac{1}{qN}\right).$$

It remains to rewrite the last inequality as

$$N^{1/q} - (N-1)^{1/q} < \frac{1}{q}N^{1/q-1} \tag{35}$$

and notice that, by the mean-value theorem, the left-hand side of (35) equals

$$\frac{1}{q}(N - \theta)^{1/q-1}$$

for some $\theta \in (0, 1)$: as $1/q - 1 \leq 0$, we have the required contradiction. □

4.2 The feature mapping for the proof of Theorem 1

In the proof of Theorem 1 we need two feature mappings from $[-Y, Y] \times \mathbf{X}$ to different Banach spaces: first, $\Phi_1(\mu, x) := \mu$ (mapping to the Banach space $\mathbb{R}$), and second, $\Phi_2 : [-Y, Y] \times \mathbf{X} \to \mathcal{F}^*$ such that $\Phi_2(\mu, x)$ is the evaluation functional $\mathbf{k}_x : f \mapsto f(x)$, $f \in \mathcal{F}$. We combine them into one feature mapping

$$\Phi(\mu, x) := \big(\Phi_1(\mu, x), \Phi_2(\mu, x)\big) \tag{36}$$

to the weighted direct sum $U := \mathbb{R} \oplus \mathcal{F}^*$, with the weights a_1 and a_2 to be chosen later. By Lemma 3, (15), and (17), $\rho_U(\tau) \le a\tau^q$, where $a := 4.34/q$. With the help of Theorem 2, we obtain for the BBK29 algorithm with parameter Φ:

$$\begin{aligned}\left|\sum_{n=1}^{N}(y_n - \mu_n)\mu_n\right| &= \left\|\sum_{n=1}^{N}(y_n - \mu_n)\Phi_1(\mu_n, x_n)\right\|_{\mathbb{R}} \\ &\le \frac{1}{\sqrt{a_1}}\left\|\sum_{n=1}^{N}(y_n - \mu_n)\Phi(\mu_n, x_n)\right\|_U \\ &\le \frac{1}{\sqrt{a_1}}2Y\mathbf{c}_\Phi(2aqN)^{1/q}\end{aligned} \tag{37}$$

and

$$\begin{aligned}&\left|\sum_{n=1}^{N}(y_n - \mu_n)D(x_n)\right| \\ &\quad= \left|\sum_{n=1}^{N}(y_n - \mu_n)\mathbf{k}_{x_n}(D)\right| \\ &\quad= \left|\left(\sum_{n=1}^{N}(y_n - \mu_n)\mathbf{k}_{x_n}\right)(D)\right| \le \left\|\sum_{n=1}^{N}(y_n - \mu_n)\mathbf{k}_{x_n}\right\|_{\mathcal{F}^*}\|D\|_{\mathcal{F}} \\ &\quad= \left\|\sum_{n=1}^{N}(y_n - \mu_n)\Phi_2(\mu_n, x_n)\right\|_{\mathcal{F}^*}\|D\|_{\mathcal{F}} \\ &\quad\le \frac{1}{\sqrt{a_2}}\left\|\sum_{n=1}^{N}(y_n - \mu_n)\Phi(\mu_n, x_n)\right\|_U\|D\|_{\mathcal{F}} \le \frac{1}{\sqrt{a_2}}2Y\mathbf{c}_\Phi(2aqN)^{1/q}\|D\|_{\mathcal{F}}\end{aligned} \tag{38}$$

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

4.3 Proof proper

The proof is based on the inequality

$$\begin{aligned}\sum_{n=1}^{N}(y_n - \mu_n)^2 &= \sum_{n=1}^{N}(y_n - D(x_n))^2 + 2\sum_{n=1}^{N}(D(x_n) - \mu_n)(y_n - \mu_n) - \sum_{n=1}^{N}(D(x_n) - \mu_n)^2 \\ &\le \sum_{n=1}^{N}(y_n - D(x_n))^2 + 2\sum_{n=1}^{N}(D(x_n) - \mu_n)(y_n - \mu_n).\end{aligned} \tag{39}$$

Using this inequality and (37–38) with $a_1 := Y^{-2}$ and $a_2 := 1$, we obtain for the $\mu_n \in [-Y, Y]$ output by the BBK29 algorithm with Φ as parameter:

$$\sum_{n=1}^{N}(y_n - \mu_n)^2 \le \sum_{n=1}^{N}(y_n - D(x_n))^2 + 2\left|\sum_{n=1}^{N}\mu_n(y_n - \mu_n)\right| + 2\left|\sum_{n=1}^{N}D(x_n)(y_n - \mu_n)\right|$$
$$\le \sum_{n=1}^{N}(y_n - D(x_n))^2 + 4Y\mathbf{c}_\Phi(2aqN)^{1/q}(\|D\|_{\mathcal{F}} + Y). \tag{40}$$

Since

$$\mathbf{c}_\Phi \le \sqrt{a_1Y^2 + a_2\mathbf{c}_{\mathcal{F}}^2} = \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1},$$

we can see that (4) holds with

$$4(2aq)^{1/q} = 4 \times 8.68^{1/p'} \tag{41}$$

in place of 40.

5 Banach kernels

An RKHS can be defined as a PBFS in which the norm is expressed via an inner product as $\|f\| = \sqrt{\langle f, f\rangle}$. It is well known that all information about an RKHS $\mathcal{F}$ on a set Z is contained in its "reproducing kernel", which is a symmetric positive definite function on Z^2 (Aronszajn 1950, Sects. I.1–I.2). The reproducing kernel can be regarded as the constructive representation of its RKHS, and it is the reproducing kernel rather than the RKHS itself that serves as a parameter of various machine-learning algorithms. In this section we will introduce a similar constructive representation for PBFS.

A *Banach kernel* B on a set Z is a function that maps each finite non-empty sequence $z_1, \dots, z_n$ of distinct elements of Z to a seminorm $(t_1, \dots, t_n) \mapsto \|(t_1, \dots, t_n)\|_{B(z_1,\dots,z_n)}$ on $\mathbb{R}^n$ and satisfies the following conditions (familiar from Kolmogorov's existence theorem, Kolmogorov 1933, Sect. III.4):

- for each $n = 1, 2, \dots$, each sequence $z_1, \dots, z_n$ of distinct elements of Z, each sequence $(t_1, \dots, t_n) \in \mathbb{R}^n$, and each permutation $\begin{pmatrix} 1 & 2 & \dots & n \\ i_1 & i_2 & \dots & i_n \end{pmatrix}$,

$$\|(t_{i_1}, \dots, t_{i_n})\|_{B(z_{i_1},\dots,z_{i_n})} = \|(t_1, \dots, t_n)\|_{B(z_1,\dots,z_n)}$$

  (in words, the seminorm of $(t_1, \dots, t_n) \in \mathbb{R}^n$ corresponding to $(z_1, \dots, z_n) \in Z^n$ does not change if $(t_1, \dots, t_n)$ and $(z_1, \dots, z_n)$ are permuted in the same way);
- for each $n = 1, 2, \dots$, each $k = 1, \dots, n$, each sequence $z_1, \dots, z_n$ of distinct elements of Z, and each sequence $(t_1, \dots, t_k) \in \mathbb{R}^k$,

$$\|(t_1, \dots, t_k)\|_{B(z_1,\dots,z_k)} = \|(t_1, \dots, t_k, 0, \dots, 0)\|_{B(z_1,\dots,z_n)}$$

  (in words, the seminorm of $(t_1, \dots, t_k) \in \mathbb{R}^k$ corresponding to $(z_1, \dots, z_k) \in Z^k$ does not change if $(t_1, \dots, t_k)$ is extended by 0s and $(z_1, \dots, z_k)$ is extended arbitrarily).

The *Banach kernel of a mapping* $\Phi : Z \to U$ to a Banach space U is the Banach kernel B defined by

$$\|(t_1, \ldots, t_n)\|_{B(z_1,\ldots,z_n)} := \|t_1\Phi(z_1) + \cdots + t_n\Phi(z_n)\|_U .$$

Proposition 1 *For each Banach kernel B on Z there exists a Banach space U and a mapping $\Phi : Z \to U$ such that B is the Banach kernel of Φ.*

Proposition 1 is a special case of the following Proposition 2, but we still need to prove it as the proof of Proposition 2 depends on it.

Proof of Proposition 1 Let U_1 be the set of all formal linear combinations $t_1z_1 + \cdots + t_nz_n$, where $n \in \{0, 1, 2, \ldots\}$, $(t_1, \ldots, t_n) \in (\mathbb{R} \setminus \{0\})^n$, and $z_1, \ldots, z_n$ are distinct elements of Z. (There is only one linear combination, denoted 0, corresponding to $n = 0$.) We do not distinguish linear combinations if they have the same addends (perhaps listed in different orders). The set U_1 is a linear space with the obvious operations of addition and multiplication by scalar: in the sum the addends that are multiples of the same $z \in Z$ should be grouped together (and removed if the resulting coefficient is zero) and multiplication by 0 gives 0.

For each linear combination $t_1z_1 + \cdots + t_nz_n \in U_1$, $n > 0$, its seminorm is defined to be $\|(t_1, \ldots, t_n)\|_{B(z_1,\ldots,z_n)}$, and the seminorm of $0 \in U_1$ is defined to be 0; it is easy to check that this is indeed a seminorm (it is well defined because of the first condition in the definition of Banach kernel, and the triangle inequality follows from the second condition). Two linear combinations are said to be *equivalent* if their difference has zero seminorm (this is indeed an equivalence relation because of the second condition). Let U_2 be the set of all equivalence classes.

The norm of $u \in U_2$ can be defined as the seminorm of any element of the equivalence class u. It remains to take the completion of U_2 as U and to define $\Phi : Z \to U$ so that $\Phi(z)$ is the equivalence class containing $1z \in U_1$. □

The *Banach kernel of a PBFS* $\mathcal{F}$ on Z is the Banach kernel B defined by

$$\|(t_1, \ldots, t_n)\|_{B(z_1,\ldots,z_n)} := \|t_1\mathbf{k}_{z_1} + \cdots + t_n\mathbf{k}_{z_n}\|_{\mathcal{F}^*},$$

where $\mathbf{k}_z : \mathcal{F} \to \mathbb{R}$, $z \in Z$, is the evaluation functional $f \in \mathcal{F} \mapsto f(z)$.

Proposition 2 *For each Banach kernel B on Z there exists a proper Banach functional space $\mathcal{F}$ on Z such that B is the Banach kernel of $\mathcal{F}$.*

Proof Let $\Phi : Z \to U$ be a mapping to a Banach space U such that B is the Banach kernel of Φ (such a Φ exists by Proposition 1). Without loss of generality we will assume that $\Phi(Z)$ spans U. Define $\mathcal{F}$ to be the set of all functions $f : Z \to \mathbb{R}$ of the form

$$f(z) := \phi(\Phi(z)), \tag{42}$$

where ϕ is a continuous linear functional on U, $\phi \in U^*$. The norm of the function (42) is $\|f\|_{\mathcal{F}} := \|\phi\|_{U^*}$. We will prove that $\mathcal{F}$ is a PBFS and that B is the Banach kernel of $\mathcal{F}$.

It is obvious that $\mathcal{F}$ is a linear space (under the usual pointwise operations of addition and multiplication by scalar) and that $\|f\|_{\mathcal{F}}$ is well-defined (i.e., does not depend on the choice of ϕ satisfying (42): there is only one such ϕ). All defining properties of a norm are

clearly satisfied for $\|\cdot\|_{\mathcal{F}}$; in particular, $\|f\|_{\mathcal{F}} = 0$ implies $f = 0$. The completeness of $\mathcal{F}$ follows from the completeness of U^*. The boundedness of the evaluation functionals for $\mathcal{F}$ means that, for each fixed $z \in Z$,

$$\sup_{\phi:\|\phi\|_{U^*}\le 1} |\phi(\Phi(z))| < \infty;$$

this immediately follows from the definition of $\|\cdot\|_{U^*}$. This completes the proof that $\mathcal{F}$ is a PBFS.

It remains to check that B is the Banach kernel of $\mathcal{F}$, i.e., that

$$\|(t_1,\dots,t_n)\|_{B(z_1,\dots,z_n)} = \|\phi \mapsto t_1\phi(\Phi(z_1)) + \cdots + t_n\phi(\Phi(z_n))\|_{U^{**}} \tag{43}$$

for all $n = 1, 2, \dots$, all $(t_1,\dots,t_n) \in (\mathbb{R}\setminus\{0\})^n$, and all distinct $z_1,\dots,z_n \in Z$. We can rewrite (43) as

$$\|(t_1,\dots,t_n)\|_{B(z_1,\dots,z_n)} = \|\phi \mapsto \phi(t_1\Phi(z_1) + \cdots + t_n\Phi(z_n))\|_{U^{**}};$$

since B is the Banach kernel of Φ, this is equivalent to

$$\|t_1\Phi(z_1) + \cdots + t_n\Phi(z_n)\|_U = \|\phi \mapsto \phi(t_1\Phi(z_1) + \cdots + t_n\Phi(z_n))\|_{U^{**}}.$$

The last equality follows from the fact that the canonical embedding of U into U^{**} is an isometry (Rudin 1991, Sect. 4.5). □

Remark A Banach kernel B on a set Z can be visualized as a family $b(z_1,\dots,z_n) \subseteq \mathbb{R}^n$, n ranging over $\{1, 2, \dots\}$ and $z_1,\dots,z_n$ over sequences of distinct elements of Z, of balanced convex sets containing a neighborhood of zero. Such a family can be obtained from B by replacing each seminorm $\|\cdot\|_{B(z_1,\dots,z_n)}$ with the unit ball in that seminorm; it is well known that the seminorm and the corresponding unit ball carry the same information (see, e.g., Rudin 1991, Theorems 1.34 and 1.35). Of course, the sets $b(z_1,\dots,z_n)$ should satisfy the two conditions of consistency analogous to those in the definition of a Banach kernel; e.g., the second condition becomes: for all $n = 1, 2, \dots$, all $k = 1,\dots,n$, and all $(z_1,\dots,z_n) \in Z^n$ whose elements are all different, the set $b(z_1,\dots,z_k)$ is the intersection of $b(z_1,\dots,z_n)$ and the subspace $z_{k+1} = \cdots = z_n = 0$.

Now we can state more explicitly the prediction algorithm described above and guaranteeing (4). Let B be the Banach kernel of the benchmark class $\mathcal{F}$ in (4). Following (30) (with Φ defined by (36)), define

$$\begin{aligned} f_n(y,\mu) := &\left(\frac{1}{Y^2}\left(\sum_{i=1}^{n-1}(y_i-\mu_i)\mu_i + (y-\mu)\mu\right)^2\right. \\ &\left. + \|(y_1-\mu_1,\dots,y_{n-1}-\mu_{n-1},y-\mu)\|^2_{B(x_1,\dots,x_{n-1},x_n)}\right)^{1/2} \\ &- \left(\frac{1}{Y^2}\left(\sum_{i=1}^{n-1}(y_i-\mu_i)\mu_i\right)^2\right. \\ &\left. + \|(y_1-\mu_1,\dots,y_{n-1}-\mu_{n-1})\|^2_{B(x_1,\dots,x_{n-1})}\right)^{1/2}. \end{aligned} \tag{44}$$

This allows us to give the kernel representation of BBK29 with Φ defined by (36); its parameter is a Banach kernel on the object space $\mathbf{X}$.

ALGORITHM GUARANTEEING (4)
Parameter: Banach kernel B of $\mathcal{F}$
FOR $n = 1, 2, \ldots$:
Read $x_n \in \mathbf{X}$.
Define $f_n : [-Y, Y]^2 \to \mathbb{R}$ by (44).
Output any root $\mu \in [-Y, Y]$ of $f_n(-Y, \mu) = f_n(Y, \mu)$ as μ_n;
if there are no such roots, output $\mu_n \in \{-Y, Y\}$
such that $\sup_{y \in [-Y,Y]} f_n(y, \mu_n) \le 0$.
Read $y_n \in [-Y, Y]$.
END FOR.

This, of course, assumes that the function

$$(n \in \{1, 2, \ldots\}, (t_1, \ldots, t_n) \in \mathbb{R}^n, (x_1, \ldots, x_n) \in \mathbf{X}^n) \mapsto \|(t_1, \ldots, t_n)\|_{B(x_1,\ldots,x_n)}$$

is efficiently computable. Perhaps the easiest way to implement the step "Output any root..." of the algorithm is to use the bisection method (see, e.g., Press et al. 1992, Sect. 9.1). To see how it can be applied, set

$$g(\mu) := f_n(Y, \mu) - f_n(-Y, \mu)$$

and remember the argument for the validity of the BBK29 algorithm given after the algorithm's description. If $g(\mu)$ is positive for $\mu = -Y$ and negative for $\mu = Y$, we can use the bisection method to find a root of $g(\mu) = 0$, as required. If this condition is not satisfied, we have one (or both) of the following cases:

- $f_n(Y, -Y) \le f_n(-Y, -Y) = 0$, which implies $\sup_{y \in [-Y,Y]} f_n(y, -Y) \le 0$ (by the convexity of $f_n(y, \mu)$ in y) and enables us to set $\mu_n := -Y$;
- $f_n(-Y, Y) \le f_n(Y, Y) = 0$, which implies $\sup_{y \in [-Y,Y]} f_n(y, Y) \le 0$ and enables us to set $\mu_n := Y$.

Acknowledgements I am grateful to Glenn Shafer and the participants of the workshop "Metric entropy and applications in analysis, learning theory and probability" (Edinburgh, September 2006), especially Fernando Cobos, for useful discussions. The referees' comments have been very helpful. This work was partially supported by EPSRC (grant EP/F002998/1), MRC (grant S505/65), Cyprus Research Promotion Foundation, and the Royal Society.

References

Adams, R. A., & Fournier, J. J. F. (2003). *Pure and applied mathematics: Vol. 140. Sobolev spaces* (2nd ed.). Amsterdam: Academic Press.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.

Auer, P., Cesa-Bianchi, N., & Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, *64*, 48–75.

Banaś, J. (1986). On moduli of smoothness of Banach spaces. *Bulletin of the Polish Academy of Sciences. Mathematics*, *34*, 287–293.

Banaś, J., & Frączek, K. (1993). Deformation of Banach spaces. *Commentationes Mathematicae Universitatis Carolinae*, *34*, 47–53.

Bárcenas, D., Gurary, V. I., Sánchez, L., & Ullán, A. (2004). On moduli of convexity in Banach spaces. *Quaestiones Mathematicae*, *27*, 137–145.

Cesa-Bianchi, N., Long, P. M., & Warmuth, M. K. (1996). Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, *7*, 604–619.

Clarkson, J. A. (1936). Uniformly convex spaces. *Transactions of the American Mathematical Society*, *40*, 396–414.

Cobos, F., & Edmunds, D. E. (1988). Clarkson's inequalities, Besov spaces and Triebel–Sobolev spaces. *Zeitschrift für Analysis und ihre Anwendungen*, *7*, 229–232.

Edmunds, D. E., & Triebel, H. (1996). *Cambridge tracts in mathematics: Vol. 120. Function spaces, entropy numbers, differential operators*. Cambridge: Cambridge University Press.

Figiel, T. (1976). On the moduli of convexity and smoothness. *Studia Mathematica*, *56*, 121–155.

Fuster, E. L. (2006). Moduli and constants: ... what a show! Available on the Internet (accessed in July 2007).

Gurary, V. I. (1967). On differential properties of the convexity moduli of Banach spaces. *Matematicheskie Issledovaniya*, *2*(1), 141–148 (in Russian).

Hanner, O. (1956). On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, *3*, 239–244.

Kivinen, J., & Warmuth, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, *132*, 1–63.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer. English translation: Foundations of the Theory of Probability. Chelsea, New York (1950).

Lindenstrauss, J. (1963). On the modulus of smoothness and divergent series in Banach spaces. *Michigan Mathematical Journal*, *10*, 241–252.

Lindenstrauss, J., & Tzafriri, L. (1979). *Ergebnisse der Mathematik und ihrer Grenzgebiete: Vol. 97. Classical Banach spaces II: function spaces*. Berlin: Springer.

Liokumovich, V. I. (1973). The existence of B-spaces with non-convex modulus of convexity. *Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika*, *12*, 43–50 (in Russian).

Nikolsky, S. M. (1961). On embedding, continuation and approximation theorems for differentiable functions of several variables. *Russian Mathematical Surveys*, *16*(5), 55–104.

Nördlander, G. (1960). The modulus of convexity in normed linear spaces. *Arkiv för Matematik*, *4*, 15–17.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C* (2nd ed.). Cambridge: Cambridge University Press.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton: Princeton University Press.

Rudin, W. (1991). *Functional analysis. International series in pure and applied mathematics* (2nd ed.). Boston: McGraw–Hill.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.

Takahashi, Y., & Kato, M. (1997). Clarkson and random Clarkson inequalities for $L_r(X)$. *Mathematische Nachrichten*, *188*, 341–348.

Triebel, H. (1983). *Monographs in mathematics: Vol. 78. Theory of function spaces*. Basel: Birkhäuser.

Triebel, H. (1992). *Monographs in mathematics: Vol. 84. Theory of function spaces II*. Basel: Birkhäuser.

Triebel, H. (2005). Sampling numbers and embedding constants. *Proceedings of the Steklov Institute of Mathematics*, *248*, 268–277.

Vovk, V. (2005). Defensive prediction with expert advice. In S. Jain, H. U. Simon & E. Tomita (Eds.), *Lecture notes in artificial intelligence: Vol. 3734. Proceedings of the sixteenth international conference on algorithmic learning theory* (pp. 444–458). Berlin: Springer. Full version: Technical report. arXiv:cs.LG/0506041 "Competitive on-line learning with a convex loss function" (version 3), arXiv.org e-Print archive, September 2005.

Vovk, V. (2006a). On-line regression competitive with reproducing kernel Hilbert spaces. In J.-Y. Cai, S. B. Cooper & A. Li (Eds.), *Lecture notes in computer science: Vol. 3959. Theory and applications of models of computation. Proceedings of the third annual conference on computation and logic* (pp. 452–463). Berlin: Springer. Full version: Technical report. arXiv:cs.LG/0511058, arXiv.org e-Print archive, January 2006.

Vovk, V. (2006b). Competing with wild prediction rules. In G. Lugosi & H. U. Simon (Eds.), *Lecture notes in artificial intelligence: Vol. 4005. Proceedings of the nineteenth annual conference on learning theory* (pp. 559–573). Berlin: Springer. This is the conference version of this paper.

Vovk, V. (2006c). *Metric entropy in competitive on-line prediction*. Technical report. arXiv:cs.LG/0609045, arXiv.org e-Print archive.