arXiv:cs/0512059v2 [cs.LG] 25 Jan 2006

Competing with wild prediction rules

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

February 1, 2008

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.

1 Introduction

For simplicity, in this introductory section we only discuss the problem of predicting 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 [36]).

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., [9], [20], and, especially, [4], §3.2; for an explicit statement see [37].) More specifically, 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 ([35], 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 Figure 1.

The intuition behind the informal notion of a function with Hölder exponent h will be captured using function spaces known as Sobolev spaces. Roughly, the Sobolev spaces $W^{s,p}([0,1])$ (defined formally in the next section), where $p \in (1,\infty]$, $s \in (0,1)$, and $s > 1/p$, can be regarded as different ways of formalizing the notion of a function on $[0,1]$ with Hölder exponent $h > s$.

The most familiar Sobolev spaces are the Hölder spaces $W^{s,\infty}([0,1])$, consisting of the functions f satisfying $|f(x) - f(y)| = O(|x-y|^s)$. The Hölder spaces are nested, $W^{s,\infty}([0,1]) \subset W^{s',\infty}([0,1])$ when $s' < s$. (That all Hölder spaces are very different can be seen from the fact that typical trajectories of the fractional Brownian motion $B^{(h)}$, defined in §3, are in $W^{s,\infty}([0,1])$ for $s < h$ and outside $W^{s,\infty}([0,1])$ for $s > h$.) As we will see in a moment, the standard Hilbert-space methods only work for $W^{s,\infty}([0,1])$ with $s > 1/2$ as benchmark classes; our goal is to develop methods that would work for smaller s as well.

The spaces $W^{s,\infty}([0,1])$ are rather awkward analytically and even 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 so f's behavior in the neighborhood of a single point can disqualify it from being a member of $W^{s,\infty}([0,1])$. Replacing sup with the mean (in the sense of L^p) w.r. to a natural "almost finite" measure gives the Sobolev spaces $W^{s,p}([0,1])$ for $p < \infty$. Results for the case $p < \infty$ immediately carry over to $p = \infty$ since, as we will see in the next section, $W^{s,\infty}([0,1]) \subseteq W^{s',p}([0,1])$ whenever $s' < s$; s' can be arbitrarily close to s.

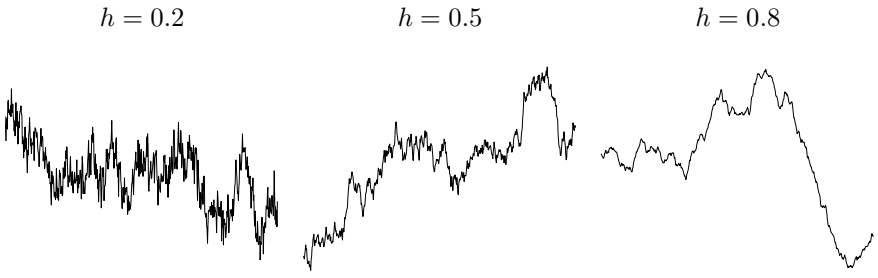

Figure 1: Functions with Hölder exponent h for three different values of h.

All Sobolev spaces (including the Hölder spaces) are Banach spaces, but $W^{s,2}([0,1])$ 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 [37] is especially close to that of this paper, although we wrote H^s instead of $W^{s,2}$ in [37]).

The condition $s > 1/p$ appears indispensable in the development of the theory (cf. the reference to the Sobolev imbedding theorem in the next section). Since this paper concentrates on the irregular end of the Sobolev spectrum, $s < 1/2$, instead of the Hilbert spaces $W^{s,2}([0,1])$ we now have to deal with the Banach spaces $W^{s,p}([0,1])$ with $p \in (2,\infty)$, which are not Hilbert spaces. The necessary tools are developed in §§4–5.

The methods of [37] 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 $W^{s,p}$ 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 $W^{s,p}([0,1])$ plus $O(N^{-1/p})$. (This gives a regret term of $O(N^{-s+\epsilon})$ for the prediction rules in $W^{s,\infty}([0,1])$, where $s < 1/2$ and $\epsilon > 0$.)

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 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".

Main theorem

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

Let U be a Banach space and $S_U := \{u \in U \,|\, \|u\|_U = 1\}$ 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 [10] 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 scalar multiplication) 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., [31], Chapter 4).

The following theorem will be proved in §§4–5.

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 Sobolev spaces $W^{s,p}(\mathbf{X})$, which we will now define.

Sobolev spaces

Suppose $\mathbf{X}$ is an open or closed set in $\mathbb{R}^m$. (The standard theory assumes that $\mathbf{X}$ is open, but the results we need easily extend to closed $\mathbf{X}$.) We only define the Sobolev spaces $W^{s,p}(\mathbf{X})$ for the cases $s \in (0,1)$ and $p > m/s$; for a more general definition see, e.g., [27] (pp. 57, 61) or [1] (Theorem 7.48 and Remark 7.49).

Let $s \in (0,1)$ and $p > m/s$. For a function $f \in L^p(\mathbf{X})$ define

$$\|f\|_{s,p} := \left(\int_{\mathbf{X}} |f(x)|^p \, \mathrm{d}x + \int_{\mathbf{X}} \int_{\mathbf{X}} \left| \frac{f(x) - f(y)}{|x-y|^s} \right|^p \frac{\mathrm{d}x \, \mathrm{d}y}{|x-y|^m} \right)^{1/p} \tag{5}$$

(we use $|\cdot|$ to denote the Euclidean norm in $\mathbb{R}^m$). The Sobolev space $W^{s,p}(\mathbf{X})$ is defined to be the set of all f such that $\|f\|_{s,p} < \infty$. The Sobolev imbedding theorem says that, for a wide range of $\mathbf{X}$ (definitely including our main example $\mathbf{X} = [0,1] \subseteq \mathbb{R}$), the functions in $W^{s,p}(\mathbf{X})$ can be made continuous by a change on a set of measure zero; we will always assume that this is true for our object space $\mathbf{X}$ and consider the elements of $W^{s,p}(\mathbf{X})$ to be continuous functions. Let $C(\mathbf{X})$ be the Banach space of continuous functions $f : \mathbf{X} \to \mathbb{R}$ with finite norm $\|f\|_{C(\mathbf{X})} := \sup_{x \in \mathbf{X}} |f(x)|$. The Sobolev imbedding theorem also says that the imbedding $W^{s,p}(\mathbf{X}) \hookrightarrow C(\mathbf{X})$ (i.e., the function that maps each $f \in W^{s,p}(\mathbf{X})$ to the same function but considered as an element of $C(\mathbf{X})$) is continuous, i.e., that

$$\mathbf{c}_{s,p} := \mathbf{c}_{W^{s,p}(\mathbf{X})} < \infty :$$

notice that $\mathbf{c}_{s,p}$ is just the norm of the imbedding $W^{s,p}(\mathbf{X}) \hookrightarrow C(\mathbf{X})$. These conclusions depend on the condition $p > m/s$ (there are other parts of the Sobolev imbedding theorem, dealing with the case where this condition is not satisfied). For a proof in the case $\mathbf{X} = \mathbb{R}^m$, see, e.g., [2], Theorems 7.34(c) and 7.47(a,c); this implies the analogous statement for $\mathbf{X}$ with smooth boundary since for such $\mathbf{X}$ every $f \in W^{s,p}(\mathbf{X})$ can be extended to an element of $W^{s,p}(\mathbb{R}^m)$ without increasing the norm more than a constant times (see, e.g., [27], p. 81). We will say "domain" to mean a subset of $\mathbb{R}^n$ which satisfies the conditions of regularity mentioned in this paragraph.

The norm (5) (sometimes called the Sobolev–Slobodetsky norm) is only one of the standard norms giving rise to the same topological vector space, and the term "Sobolev space" is usually used to refer to the topology rather than a specific norm; in this paper we will not consider any other norms. The restriction $s \in (0,1)$ is not essential for the results in this paper, but the definition of $\|\cdot\|_{s,p}$ becomes slightly more complicated when $s \geq 1$ (cf. [27]); [2] gives a different but equivalent norm.

For comparison purposes we will also define the spaces $W^{1,p}([0,1])$, $p \in (1,\infty)$: set

$$\|f\|_{1,p} := \left(\int_0^1 |f(x)|^p \, \mathrm{d}x + \int_0^1 |f'(x)|^p \, \mathrm{d}x \right)^{1/p}$$

and include in $W^{1,p}([0,1])$ all absolutely continuous functions $f : [0,1] \to \mathbb{R}$ with $\|f\|_{1,p} < \infty$. We will always assume $\mathbf{X} = [0,1]$ in the case $s = 1$.

We can now deduce the following corollary from Theorem 1. It is known that (3) is satisfied for the Sobolev spaces $W^{s,p}(\mathbf{X})$ (see (44)). Let $p \in [2,\infty)$ and $s \in (m/p, 1)$. There exists a constant $C_{s,p} > 0$ and a prediction algorithm producing $\mu_n \in [-Y, Y]$ that are guaranteed to satisfy

$$\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \frac{1}{N} \sum_{n=1}^{N} (y_n - D(x_n))^2 + Y C_{s,p} \left(\|D\|_{s,p} + Y \right) N^{-1/p} \tag{6}$$

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

In informal discussions below we will continue to call terms such as the second addend on the right-hand side of (6) 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

$$C_{s,p} = 40\sqrt{\mathbf{c}_{s,p}^2 + 1},$$

but in fact

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

will suffice (see (53) below). In the special case $p = 2$ one can use Hilbert-space methods to improve (7), which now becomes, approximately,

$$11.78\sqrt{\mathbf{c}_{s,2}^2 + 1}, \tag{8}$$

to

$$2\sqrt{\mathbf{c}_{s,2}^2 + 1} \tag{9}$$

([37], Theorem 1); using Banach-space methods we have lost a factor of 5.89. For example, in the case $s = 1$, (8) gives $C_{s,p} \approx 17.92$ and (9) gives $C_{s,p} \approx 3.04$ (the value $\mathbf{c}_{1,2}^2 = \coth 1$ was found in [26]; for further details of the case $s = 1, p = 2$, see [37], §4).

Application to the Hölder-continuous functions

An important limiting case of the norm (5) is

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

where $f : \mathbf{X} \to \mathbb{R}$ is, as usual, assumed continuous. The space $W^{s,\infty}(\mathbf{X})$ consists of the functions f with $\|f\|_{s,\infty} < \infty$, and its elements are called *Hölder continuous of order s*.

The Hölder-continuous functions of order s are perhaps the most intuitive formalization of the functions with Hölder exponent $h \geq s$. Let us see what Theorem 1 gives for them.

Suppose that $\mathbf{X}$ is a bounded domain in $\mathbb{R}^m$, $p \in (1, \infty)$, and $s, s' \in (0, 1)$

are such that $s' < s$. If $f \in W^{s,\infty}(\mathbf{X})$,

$$\begin{aligned}
\|f\|_{s',p} &= \left(\int_{\mathbf{X}} |f(x)|^p \, \mathrm{d}x + \int_{\mathbf{X}} \int_{\mathbf{X}} \left| \frac{f(x) - f(y)}{|x-y|^{s'}} \right|^p \frac{\mathrm{d}x \, \mathrm{d}y}{|x-y|^m} \right)^{1/p} \\
&\le \left(C^p + \int_{\mathbf{X}} \int_{\mathbf{X}} \left| \frac{C \, |x-y|^s}{|x-y|^{s'}} \right|^p \frac{\mathrm{d}x \, \mathrm{d}y}{|x-y|^m} \right)^{1/p} \\
&= \left(C^p + C^p \int_{\mathbf{X}} \int_{\mathbf{X}} |x-y|^{-m+sp-s'p} \, \mathrm{d}x \, \mathrm{d}y \right)^{1/p} \\
&\le \left(C^p + C^p \, |\mathbf{X}| \int_0^{\operatorname{diam} \mathbf{X}} t^{-m+sp-s'p} \frac{d}{dt} \left(\frac{\pi^{m/2}}{\Gamma(m/2+1)} t^m \right) \, \mathrm{d}t \right)^{1/p} \\
&= C \left(1 + m \frac{\pi^{m/2}}{\Gamma(m/2+1)} \, |\mathbf{X}| \, \frac{(\operatorname{diam} \mathbf{X})^{(s-s')p}}{(s-s')p} \right)^{1/p}, \quad (10)
\end{aligned}$$

where $C := \|f\|_{s,\infty}$, $|\mathbf{X}|$ stands for the volume (Lebesgue measure) of $\mathbf{X}$, and $\operatorname{diam} \mathbf{X}$ stands for the diameter of $\mathbf{X}$; remember that $\pi^{m/2}/\Gamma(m/2+1)$ is the volume of the unit ball in $\mathbb{R}^m$. Therefore, (10) gives an explicit bound for the norm of the continuous imbedding $W^{s,\infty}(\mathbf{X}) \hookrightarrow W^{s',p}(\mathbf{X})$.

Fix an arbitrarily small $\epsilon > 0$. Applying (6) to $W^{s',p}(\mathbf{X})$ with $p > m/s$ sufficiently close to m/s and to $s' \in (m/p, s)$, we can see from (10) that there exists a constant $C_{s,\epsilon} > 0$ such that

$$\frac{1}{N} \sum_{n=1}^N (y_n - \mu_n)^2 \le \frac{1}{N} \sum_{n=1}^N (y_n - D(x_n))^2 + Y C_{s,\epsilon} \left(\|D\|_{s,\infty} + Y \right) N^{-s/m+\epsilon} \quad (11)$$

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

3 Implications for a stochastic Reality

In this section we discuss implications of Theorem 1 for statistical learning theory and filtering of random processes. Surprisingly, even when Reality follows a specific stochastic strategy, competitive on-line results do not trivialize but provide new meaningful information.

Statistical learning theory

In this section we apply the method of [8] to derive a corollary of Theorem 1 for the statistical learning framework, where (x_n, y_n) are assumed to be drawn independently from the same probability distribution on $\mathbf{X} \times [-Y, Y]$.

The *risk* of a prediction rule (formally, a measurable function) $D : \mathbf{X} \to \mathbb{R}$

with respect to a probability distribution P on $\mathbf{X} \times [-Y, Y]$ is defined as

$$\operatorname{risk}_P(D) := \int_{\mathbf{X} \times [-Y,Y]} (y - D(x))^2 P(\mathrm{d}x, \mathrm{d}y).$$

Our current goal is to construct, from a given sample, a prediction rule whose risk is competitive with the risk of small-norm prediction rules in $W^{s,p}(\mathbf{X})$.

Fix an on-line prediction algorithm and a sequence $(x_1, y_1), (x_2, y_2), \ldots$ of examples. For each $n = 1, 2, \ldots$ and each $x \in \mathbf{X}$, define $H_n(x)$ to be the prediction $\mu_n \in \mathbb{R}$ output by the algorithm when fed with $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), x$. We will assume that the functions H_n are always measurable (they are for our algorithm, constructed in the following two sections). The prediction rule

$$\overline{H}_N(x) := \frac{1}{N} \sum_{n=1}^{N} H_n(x)$$

will be said to be *obtained by averaging* from the prediction algorithm.

The following result is an easy application of the method of [8] to (6); we refrain from stating the analogous result based on (11).

Corollary 1 *Let $\mathbf{X}$ be a domain in $\mathbb{R}^m$, $p \geq 2$, $s \in (m/p, 1)$, and let $\overline{H}_N$, $N = 1, 2, \ldots$, be the prediction rule obtained by averaging from some prediction algorithm guaranteeing (6). For any $D \in W^{s,p}(\mathbf{X})$, any probability distribution P on $\mathbf{X} \times [-Y, Y]$, any $N = 1, 2, \ldots$, and any $\delta > 0$,*

$$\operatorname{risk}_P(\overline{H}_N) \leq \operatorname{risk}_P(D) + YC_{s,p} \left(\|D\|_{s,p} + Y \right) N^{-1/p} + 4Y^2 \sqrt{2 \ln \frac{2}{\delta}} N^{-1/2} \quad (12)$$

with probability at least $1 - \delta$.

Proof Without loss of generality we assume that $D(x) \in [-Y, Y]$ for all $x \in \mathbf{X}$ and that $H_n(x) \in [-Y, Y]$ for all $x \in \mathbf{X}$ and n. Outside an event of probability

$$\delta := 2 \exp\left(-\frac{\epsilon^2 N}{8Y^4} \right) \quad (13)$$

we have (some steps will be explained later on)

$$\begin{aligned}
\operatorname{risk}_P(\overline{H}_N) &\leq \frac{1}{N} \sum_{n=1}^{N} \operatorname{risk}_P(H_n) && (14)\\
&\leq \frac{1}{N} \sum_{n=1}^{N} (y_n - H_n(x_n))^2 + \epsilon && (15)\\
&\leq \frac{1}{N} \sum_{n=1}^{N} (y_n - D(x_n))^2 + YC_{s,p} \left(\|D\|_{s,p} + Y \right) N^{-1/p} + \epsilon && (16)\\
&\leq \frac{1}{N} \sum_{n=1}^{N} \operatorname{risk}_P(D) + YC_{s,p} \left(\|D\|_{s,p} + Y \right) N^{-1/p} + 2\epsilon && (17)\\
&= \operatorname{risk}_P(D) + YC_{s,p} \left(\|D\|_{s,p} + Y \right) N^{-1/p} + 2\epsilon. && (18)
\end{aligned}$$

The first inequality, (14), follows from the convexity of the function $t \mapsto t^2$. Inequalities (15) and (17) follow from Hoeffding's martingale inequality ([16]; see also [11], Theorem 9.1 on p. 135). Either of (15) and (17) holds with probability at least $1-\delta/2$; therefore, both will hold with probability at least $1-\delta$. Finally, inequality (16) follows from (6).

Our goal, (12), follows from the inequality between the extreme terms of (14)–(18) if we substitute

$$\epsilon = 2Y^2\sqrt{2\ln\frac{2}{\delta}}N^{-1/2} \tag{19}$$

(which is a different way of writing (13)). ∎

For a fixed δ, the regret term (the sum of the second and third addends on the right-hand side) of (12) grows as $N^{-1/p}$. For a discussion of related results in statistical learning theory, see [37] (versions 1 and 2), §5.

Filtering of random processes

Suppose we are interested in the value of a "signal" $\Theta : [0,1] \to \mathbb{R}$ sequentially observed at moments $t_n := n/N$, $n = 1, \dots, N$, where N is a large positive integer; let $\theta_n := \Theta(t_n)$. The problem is that our observations of θ_n are imperfect, and in fact we see $y_n = \theta_n + \xi_n$, where each noise random variable ξ_n has zero expectation given the past. We assume that Θ belongs to $W^{s,p}([0,1])$ (but do not make any assumptions about the mechanism, deterministic, stochastic, or other, that generated it) and that $\theta_n, y_n \in [-Y, Y]$ for a known constant Y. Let us use the μ_n from Theorem 1 as estimates of the true values θ_n. The elementary equality

$$a^2 = (a-b)^2 - b^2 + 2ab \tag{20}$$

implies

$$\sum_{n=1}^{N}(\mu_n-\theta_n)^2 = \sum_{n=1}^{N}(y_n-\mu_n)^2 - \sum_{n=1}^{N}(y_n-\theta_n)^2 + 2\sum_{n=1}^{N}(y_n-\theta_n)(\mu_n-\theta_n). \tag{21}$$

Hoeffding's inequality in the martingale form shows that, for any $C > 0$,

$$\mathbb{P}\left\{2\sum_{n=1}^{N}(y_n-\theta_n)(\mu_n-\theta_n) \ge C\right\} \le \exp\left(-\frac{C^2}{128Y^4N}\right).$$

Substituting this (with C expressed via the right-hand side, denoted δ) and (6) into (21), we obtain the following corollary, which we state somewhat informally.

Corollary 2 *Let $p \ge 2$, $s \in (1/p, 1)$, and $\delta > 0$. Suppose that $\Theta \in W^{s,p}([0,1])$ and $y_n = \theta_n + \xi_n \in [-Y, Y]$, where $\theta_n := \Theta(n/N) \in [-Y, Y]$ and ξ_n are random variables whose expectation given the past (including θ_n) is zero. With*

probability at least $1-\delta$ *the* μ_n *of (6) satisfy*

$$\frac{1}{N}\sum_{n=1}^{N}(\mu_n-\theta_n)^2 \le YC_{s,p}\left(\|\Theta\|_{s,p}+Y\right)N^{-1/p}+8Y^2\sqrt{2\ln\frac{1}{\delta}}N^{-1/2}. \quad (22)$$

The constant $C_{s,p}$ in (22) is the one in (7). From (11), we can also see that, if we assume $\Theta \in W^{s,\infty}([0,1])$,

$$\frac{1}{N}\sum_{n=1}^{N}(\mu_n-\theta_n)^2 \le YC_{s,\epsilon}\left(\|\Theta\|_{s,\infty}+Y\right)N^{-s+\epsilon}+8Y^2\sqrt{2\ln\frac{1}{\delta}}N^{-1/2} \quad (23)$$

will hold with probability at least $1-\delta$.

It is important that the function Θ in (22) and (23) does not have to be chosen in advance: it can be constructed "step-wise", with $\Theta(t)$ for $t \in (n/N,(n+1)/N]$ chosen at will after observing ξ_n and taking into account all other information that becomes available before and including time n/N. A clean formalization of this intuitive picture seems to require the game-theoretic probability of [32] (although we can get the picture "almost right" using the standard measure-theoretic probability).

In the case where Θ is generated from a diffusion process, it will almost surely belong to $W^{(1-\epsilon)/2,\infty}([0,1])$ (this follows from standard results about the Brownian motion, such as Lévy's modulus theorem: see, e.g., [19], Theorem 9.25), and so the regret term in (22) and (23) can be made $O(N^{-1/2+\epsilon})$, for an arbitrarily small $\epsilon > 0$. The Kalman filter, which is stochastically optimal, gives a somewhat better regret, $O(N^{-1/2})$. Corollary 2, however, does not depend on the very specific assumptions of the Kalman filter: we do not require the linearity, Gaussianity, or even stochasticity of the model; the assumption about the noise ξ_n is minimal (zero expectation given the past). Instead, we have the assumption that all θ_n and y_n are chosen from $[-Y,Y]$. It appears that in practice the interval to which the θ_n and y_n are assumed to belong should change slowly as new data are processed. This is analogous to the situation with the Kalman filter, which, despite assuming linear systems, has found its greatest application to non-linear systems [34]; what is usually used in practice is the "extended Kalman filter", which relies on a slowly changing linearization of the non-linear system.

Until the end of this section we will discuss in more detail the standard stochastic approach to the problem of filtering ([17]; see also [34], [33], §VI.7, and, for a continuous-time version, [18], [25], §10.1). The signal is now modeled as a random process Θ_t, $t \in [0,1]$, governed by the stochastic differential equation

$$d\Theta_t = (a_0(t)+a_1(t)\Theta_t)\,dt + b(t)dB_t, \quad (24)$$

where B_t is the standard Brownian motion (a zero-mean Gaussian continuous stochastic process on $[0,1]$ such that $B_0=0$ and the variance of each increment $B_{t_1}-B_{t_2}$ is $|t_1-t_2|$) and $a_0,a_1,b:[0,1]\to\mathbb{R}$ are bounded Borel functions. The process starts from a random value Θ_0 (modeled as a Gaussian random

variable independent of B_t) and, as before, is observed at points $t_n := n/N$; $\theta_n := \Theta(t_n)$. The observed sequence is $y_n = \theta_n + \sigma\xi_n$ (neither θ_n nor y_n are assumed to be bounded by a known constant), where σ is a positive constant and ξ_n are standard Gaussian random variables independent between themselves and of the initial position Θ_0 and the Brownian motion B_t. In some important respects this is a simplification of the usual filtering problems; e.g., we consider scalar rather than vector Θ_t and y_n.

Earlier we discussed the possibility of positive contributions of competitive on-line results, such as Theorem 1, to the problem of filtering, and now we will briefly explore the connection in the opposite direction: limitations on competitive on-line prediction following from the known optimality properties of the Kalman filter. According to (11), there is a prediction algorithm $O(N^{-s+\epsilon})$-competitive with $W^{s,\infty}([0,1])$, for any $\epsilon > 0$. It remains an open problem to show that the rate $N^{-s+\epsilon}$ (we will disregard plus or minus ϵ in the rest of this section) cannot be improved, but the following considerations make it likely in the case $s \approx 1/2$. (For an alternative argument, see, e.g., Theorem 4 in [37].)

Suppose the prediction rule $D : [0,1] \to \mathbb{R}$ is generated randomly as the trajectory of the stochastic process (24) with $\Theta_0 = 0$, $a_0(t) \equiv 0$, $a_1(t) \equiv 0$, and $b(t) \equiv c > 0$ (i.e., $D(t) = cB_t$, where B is the standard Brownian motion). The positive constant c is chosen small as compared to Y, so that $D(t)$ is unlikely to take values approaching $-Y$ or Y. It is clear that the observations y_n are generated independently (given B) from the normal distribution $N(D(t_n), \sigma^2)$ with mean $D(t_n)$ and variance σ^2; if y_n falls outside $[-Y, Y]$, it is truncated to $Y \operatorname{sign} y_n$. The variance $\sigma^2 > 0$ is assumed to be small enough for the probability of $|y_n| < Y$ to be close to 1 for each n (or we can even take c and σ slightly, say logarithmically, dependent on N so that $\max_n |y_n| < Y$ with a probability tending to 1). According to the standard properties of the Kalman filter (see, e.g., [25], Theorem 13.4, or [33], Theorem VI.7.1), the variance γ_n of the best estimate of θ_n (which is also the best estimate of y_n), $n > 1$, given $y_1, \dots, y_{n-1}$ satisfies the recurrent equation

$$\gamma_{n+1} = \gamma_n + \frac{c^2}{N} - \frac{\gamma_n^2}{\sigma^2 + \gamma_n}.$$

It is clear that γ_n is an increasing sequence tending, as $n \to \infty$, to a limit equal to

$$\frac{c^2 + \sqrt{c^4 + 4c^2\sigma^2 N}}{2N} > \frac{c\sigma}{\sqrt{N}},$$

and that it will move significantly towards this limit already during the first $\sqrt{N}$ rounds (cf. Figure 2). By Hoeffding's inequality, the excess of the total loss of the stochastically best algorithm (the Kalman filter) over the total loss of D will be of order $N^{1/2}$, and so the excess of its average loss will be of order $N^{-1/2}$ (with probability very close to 1).

Since the sample paths of diffusion processes almost surely belong to $W^{s,\infty}([0,1])$ for all $s \in (0, 1/2)$, we can see that no prediction algorithm can be $O(N^{-1/2-\epsilon})$-competitive with $W^{1/2-\epsilon,\infty}([0,1])$. Therefore, if we disregard the

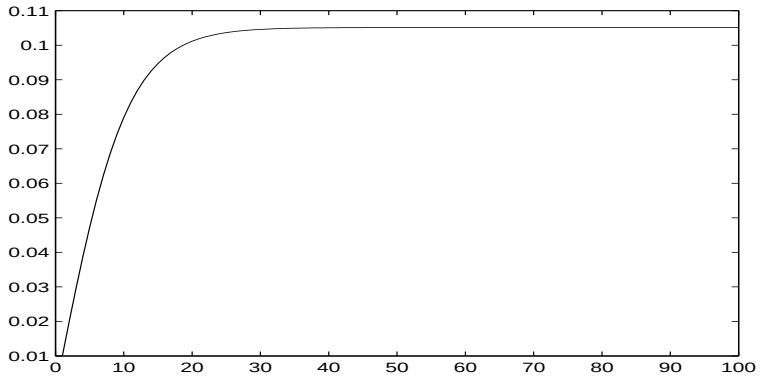

Figure 2: The growth of the Kalman filter's error γ_n, $n = 1, \dots, N$, for $c = \sigma^2 = 1$ and $N = 100$; the final value γ_N is approximately $N^{-1/2}$.

epsilons, our algorithm achieves the optimal rate of decay in N of the regret term for $s \approx 1/2$.

A similar argument might have also worked in the case $s < 1/2$ had we known an analogue of the Kalman filter result for the fractional Brownian motion, where B is replaced with a stochastic process $B^{(h)}$, $h \in (0, 1/2)$, defined in the same way except that the variance of each increment $B^{(h)}_{t_1} - B^{(h)}_{t_2}$ is $|t_1 - t_2|^{2h}$ (notice that $B = B^{(1/2)}$). Unfortunately, we know of no such result, although a step in this direction is made in [29].

4 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 [23]; for information about other moduli and further references, see [13]. We will only consider Banach spaces of dimension at least 2.

Moduli of convexity and smoothness

A natural modification of Clarkson's modulus of convexity was proposed by Gurary [14]:

$$\delta^\dagger_U(\epsilon) := \inf_{\substack{u,v \in S_U \\ \|u-v\|_U = \epsilon}} \left(1 - \inf_{t \in [0,1]} \|tu + (1-t)v\|_U\right). \tag{25}$$

It is clear that

$$\delta_U(\epsilon) \le \delta^\dagger_U(\epsilon) \le 2\delta_U(\epsilon)$$

(cf. the proof of Lemma 2 below), and it was shown recently [7] that this relation cannot be improved.

The standard modulus of smoothness was proposed by Lindenstrauss [22]:

$$\rho_U(\tau) := \sup_{u,v\in S_U} \left(\frac{\|u+\tau v\|_U + \|u-\tau v\|_U}{2} - 1 \right), \quad \tau > 0. \tag{26}$$

Lindenstrauss also established a simple but very useful relation of conjugacy (cf. [30], §12, although δ is not always convex [24]) between δ and ρ:

$$\rho_{U^*}(\tau) = \sup_{\epsilon\in(0,2]} \left(\frac{\epsilon\tau}{2} - \delta_U(\epsilon) \right); \tag{27}$$

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 (27) we obtain for the dual space $\mathcal{F}^*$ to $\mathcal{F}$, assuming $\tau \in (0,1]$:

$$\rho_{\mathcal{F}^*}(\tau) \le \sup_{\epsilon\in(0,2]} \left(\frac{\epsilon\tau}{2} - (\epsilon/2)^p/p \right) = \tau^q/q, \tag{28}$$

where $q := p/(p-1)$ (the supremum in (28) 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 $\rho_U(\tau) \to 0$ as $\tau \to 0$. All uniformly convex and all uniformly smooth Banach spaces U are reflexive (i.e., $U^{**} = U$; see, e.g., [23], Proposition 1.e.3 on p. 61).

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

$$\|u+v\|_V^2 + \|u-v\|_V^2 = 2\|u\|_V^2 + 2\|v\|_V^2 \tag{29}$$

immediately gives

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

and

$$\rho_V(\tau) = \sqrt{1+\tau^2} - 1 \le \tau^2/2. \tag{30}$$

Nördlander [28] 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,

$$\begin{aligned} \delta_U(\epsilon) &\le \delta_V(\epsilon) = 1 - \sqrt{1-(\epsilon/2)^2}, \\ \rho_U(\tau) &\ge \rho_V(\tau) = \sqrt{1+\tau^2} - 1. \end{aligned} \tag{31}$$

The original definitions (1) and (26) of the moduli of convexity and smoothness look very different, and Banaś [5] proposed a definition of modulus of smoothness similar to (1):

$$\rho_U^\dagger(\tau) := \sup_{\substack{u,v\in S_U \\ \|u-v\|_U=\tau}} \left(1 - \left\| \frac{u+v}{2} \right\|_U \right), \quad \tau \in (0,2). \tag{32}$$

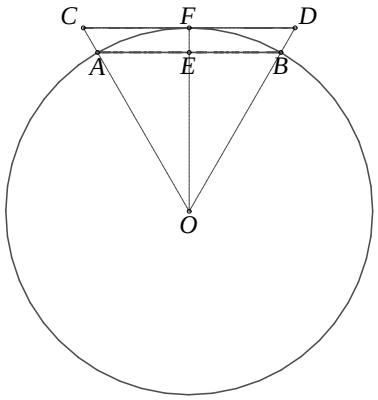

Figure 3: Relation between ρ and $\rho^\dagger$.

The difference $\rho_U^\dagger(\epsilon) - \delta_U(\epsilon)$ measures the degree to which (the unit ball in) U is deformed [6] (it is always zero for Hilbert spaces). What we will need in this paper is the modification of (32) in the direction of (25):

$$\rho_U^\ddagger(\tau) := \sup_{\substack{u,v\in S_U \\ \|u-v\|_U=\tau}} \sup_{t\in[0,1]} \left(1 - \|tu + (1-t)v\|_U\right), \quad \tau \in (0,2). \tag{33}$$

Since the standard results about moduli of convexity and smoothness are about the definitions (1) and (26), we first need to establish connections between (26) and (33). The first of these results appears in [5] (but we still prove it since [5] is less easily accessible than most other papers in our bibliography).

Lemma 1 ([5]) *For all $\tau \in (0,2)$,*

$$\frac{\rho_U^\dagger(\tau)}{1-\rho_U^\dagger(\tau)} \le \rho_U\left(\frac{\tau}{2\left(1-\rho_U^\dagger(\tau)\right)}\right). \tag{34}$$

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

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

(cf. Figure 3, 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

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

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., [23], Proposition 1.e.5 on p. 64). ∎

Corollary 3 *For all $\tau \in (0,1]$,*

$$\rho_U^\dagger(\tau) \leq \rho_U(\tau). \tag{35}$$

Proof Let $\tau \in (0,1]$. Following [5], proof of Lemma 1, we obtain

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

We can now easily deduce (35) from (34) and the fact that ρ_U is a non-decreasing function ([23], Proposition 1.e.5):

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

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

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

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 $t \leq 1/2$. Since

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

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

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

$$\|(u_1, u_2)\|_{U_1 \oplus U_2} := \sqrt{a_1 \|u_1\|_{U_1}^2 + a_2 \|u_2\|_{U_2}^2}, \tag{36}$$

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. [37], Corollary 4). The "Euclidean" definition (36) 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 [12], p. 132).

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

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

Proof We will follow the proof of Proposition 17 in [12], which is based on the following weak form of the parallelogram identity (29), valid for all Banach spaces:

$$\|u+v\|_U^2 + \|u-v\|_U^2 - 2\|u\|_U^2 - 2\|v\|_U^2 \\ \le 2\|u\|_U \left(\|u+v\|_U + \|u-v\|_U - 2\|u\|_U\right) \tag{37}$$

(see [12], Lemma 16 on p. 132); it is clear that (37) implies

$$\|u+v\|_U^2 + \|u-v\|_U^2 - 2\|u\|_U^2 - 2\|v\|_U^2 \le 4\|u\|_U^2 \, \rho_U\left(\|v\|_U / \|u\|_U\right). \tag{38}$$

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 (38) to $(u, v) := (u_1, \tau v_1)$ and $(u, v) := (u_2, \tau v_2)$, we obtain

$$\|u_1 + \tau v_1\|_{U_1}^2 + \|u_1 - \tau v_1\|_{U_1}^2 - 2\|u_1\|_{U_1}^2 - 2\tau^2 \|v_1\|_{U_1}^2 \\ \le 4\|u_1\|_{U_1}^2 \, \rho_{U_1}\left(\tau \|v_1\|_{U_1} / \|u_1\|_{U_1}\right) \tag{39}$$

and

$$\|u_2 + \tau v_2\|_{U_2}^2 + \|u_2 - \tau v_2\|_{U_2}^2 - 2\|u_2\|_{U_2}^2 - 2\tau^2 \|v_2\|_{U_2}^2 \\ \le 4\|u_2\|_{U_2}^2 \, \rho_{U_2}\left(\tau \|v_2\|_{U_2} / \|u_2\|_{U_2}\right). \tag{40}$$

Multiplying (39) by a_1 and (40) by a_2 and summing now gives

$$\left\|u^\dagger + \tau v^\dagger\right\|_{U_1 \oplus U_2}^2 + \left\|u^\dagger - \tau v^\dagger\right\|_{U_1 \oplus U_2}^2 - 2 - 2\tau^2 \\ \leq 4 \sum_{j=1}^{2} a_j \left\|u_j\right\|_{U_j}^2 \rho_{U_j}\left(\tau \left\|v_j\right\|_{U_j} / \left\|u_j\right\|_{U_j}\right). \quad (41)$$

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

- when $\left\|v_j\right\|_{U_j} \leq \left\|u_j\right\|_{U_j}$,

  $$\rho_{U_j}\left(\tau \left\|v_j\right\|_{U_j} / \left\|u_j\right\|_{U_j}\right) \leq \rho_{U_j}(\tau) \left\|v_j\right\|_{U_j} / \left\|u_j\right\|_{U_j}$$

  (by the convexity of ρ, following from the convexity of the Fenchel transform, (27), and the reflexivity of all uniformly convex and all uniformly smooth spaces);

- when $\left\|v_j\right\|_{U_j} > \left\|u_j\right\|_{U_j}$,

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

  (where $L < 3.18$ is a constant satisfying $\rho(\sigma)/\sigma^2 \leq L\rho(\tau)/\tau^2$ for all positive $\tau \leq \sigma$; see [12], Proposition 10 on p. 128 and the remark after its proof).

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

$$\sum_{j=1}^{2} a_j \left\|u_j\right\|_{U_j}^2 \rho_{U_j}\left(\tau \left\|v_j\right\|_{U_j} / \left\|u_j\right\|_{U_j}\right) \\ \leq \sum_{j=1}^{2} a_j \left\|v_j\right\|_{U_j} \rho_{U_j}(\tau) \max\left(\left\|u_j\right\|_{U_j}, L \left\|v_j\right\|_{U_j}\right) \\ \leq \left(\sum_{j=1}^{2} a_j \left\|v_j\right\|_{U_j}^2\right)^{1/2} \left(\sum_{j=1}^{2} a_j \left(\rho_{U_j}(\tau)\right)^2 \left(\left\|u_j\right\|_{U_j}^2 + L^2 \left\|v_j\right\|_{U_j}^2\right)\right)^{1/2} \\ \leq \left(\sum_{j=1}^{2} f^2(\tau) a_j \left(\left\|u_j\right\|_{U_j}^2 + L^2 \left\|v_j\right\|_{U_j}^2\right)\right)^{1/2} = \sqrt{L^2+1} f(\tau) \quad (42)$$

(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

$\tau \in (0,1]$,

$$\begin{aligned}
&\frac{1}{2}\left(\left\|u^\dagger + \tau v^\dagger\right\|_{U_1\oplus U_2} + \left\|u^\dagger - \tau v^\dagger\right\|_{U_1\oplus U_2}\right) \\
&\qquad\qquad \le \left(\frac{1}{2}\left(\left\|u^\dagger + \tau v^\dagger\right\|^2_{U_1\oplus U_2} + \left\|u^\dagger - \tau v^\dagger\right\|^2_{U_1\oplus U_2}\right)\right)^{1/2} \\
&\qquad \le \left(1 + \tau^2 + 2\sqrt{L^2+1}f(\tau)\right)^{1/2} \le \left(1+\tau^2\right)^{1/2} + \sqrt{L^2+1}f(\tau) \\
&\qquad\qquad \le 1 + f(\tau) + \sqrt{L^2+1}f(\tau) = 1 + \left(1+\sqrt{L^2+1}\right)f(\tau)
\end{aligned}$$

(the first inequality follows from the convexity of the function $t \mapsto t^2$, the second from (41) and (42), the third from the mean-value theorem, and the fourth from Nördlander's bound (31)). It remains to compare the resulting inequality with the definition of the modulus of convexity and remember that $L < 3.18$. ∎

Convexity and smoothness for Sobolev spaces

It was shown by Clarkson [10] (§3) that, for $p \in [2,\infty)$,

$$\delta_{L^p}(\epsilon) \ge 1 - \left(1 - (\epsilon/2)^p\right)^{1/p}.$$

(And this bound was shown to be optimal in [15].) A quick inspection of the standard proofs (see, e.g., [2], 2.34–2.40) shows that the underlying measurable space Ω and measure μ of $L^p = L^p(\Omega,\mu)$ can be essentially arbitrary (only the degenerate case where $\dim L^p < 2$ should be excluded), although this generality is usually not emphasized.

It is easy to see (cf. [2], 3.5–3.6) that the modulus of convexity of each Sobolev space $W^{s,p}(\mathbf{X})$, $s \in (0,1)$ and $p \in [2,\infty)$, also satisfies

$$\delta_{W^{s,p}(\mathbf{X})}(\epsilon) \ge 1 - \left(1 - (\epsilon/2)^p\right)^{1/p}. \tag{43}$$

Indeed, with each $f \in W^{s,p}(\mathbf{X})$ we can associate a function $\overline{f} : \mathbf{X} \cup \mathbf{X}^2 \to \mathbb{R}$ (we regard the sets $\mathbf{X}$ and $\mathbf{X}^2$ as disjoint) such that

$$\begin{aligned}
\overline{f}(x) &= f(x) && \text{for } x \in \mathbf{X}, \\
\overline{f}(x,y) &= \frac{f(x) - f(y)}{|x-y|^s} && \text{for } (x,y) \in \mathbf{X}^2;
\end{aligned}$$

the measure on $\mathbf{X}\cup\mathbf{X}^2$ coincides with the Lebesgue measure on the measurable subsets of $\mathbf{X}$ and with the measure whose density is $(x,y) \in \mathbf{X}^2 \mapsto |x-y|^{-m}$, with respect to the Lebesgue measure, on the measurable subsets of $\mathbf{X}^2$. The

bound (43) can now be deduced from Clarkson's result as follows:

$$\begin{aligned}
\delta_{W^{s,p}(\mathbf{X})}(\epsilon) &:= \inf_{\substack{f,g\in S_{W^{s,p}(\mathbf{X})}\\ \|f-g\|_{W^{s,p}(\mathbf{X})}=\epsilon}} \left(1-\left\|\frac{f+g}{2}\right\|_{W^{s,p}(\mathbf{X})}\right)\\
&= \inf_{\substack{f,g:\mathbf{X}\to\mathbb{R}\\ \overline{f},\overline{g}\in L^p(\mathbf{X}\cup\mathbf{X}^2)\\ \|\overline{f}-\overline{g}\|_{L^p(\mathbf{X}\cup\mathbf{X}^2)}=\epsilon}} \left(1-\left\|\frac{\overline{f}+\overline{g}}{2}\right\|_{L^p(\mathbf{X}\cup\mathbf{X}^2)}\right)\\
&\geq \inf_{\substack{u,v\in L^p(\mathbf{X}\cup\mathbf{X}^2)\\ \|u-v\|_{L^p(\mathbf{X}\cup\mathbf{X}^2)}=\epsilon}} \left(1-\left\|\frac{u+v}{2}\right\|_{L^p(\mathbf{X}\cup\mathbf{X}^2)}\right)\\
&= \delta_{L^p(\mathbf{X}\cup\mathbf{X}^2)}(\epsilon) \geq 1-\left(1-(\epsilon/2)^p\right)^{1/p}.
\end{aligned}$$

Since, for $t\in[0,1]$ and $p\geq 1$, $(1-t)^{1/p}\leq 1-t/p$ (the left-hand side is a concave function of t, and the values and derivatives of the two sides match when $t=0$), we have

$$\delta_{W^{s,p}(\mathbf{X})}(\epsilon) \geq (\epsilon/2)^p/p. \tag{44}$$

Therefore, as we said in §2, the Sobolev spaces indeed satisfy the condition (3) of Theorem 1.

5 Proof of Theorem 1

In this section we partly follow the proof of Theorem 1 in [37] (§6).

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

$$\begin{aligned}
f_n(y,\mu) := &\left\|\sum_{i=1}^{n-1}(y_i-\mu_i)\Phi\big(\mu_i,x_i\big)+(y-\mu)\Phi\big(\mu,x_n\big)\right\|_U\\
&-\left\|\sum_{i=1}^{n-1}(y_i-\mu_i)\Phi\big(\mu_i,x_i\big)\right\|_U
\end{aligned} \tag{45}$$

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 (45).
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 (45) 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*

$$\left\| \sum_{n=1}^{N} (y_n - \mu_n)\Phi(\mu_n, x_n) \right\|_U \le 2Y\mathbf{c}_\Phi \left(2aqN\right)^{1/q} \tag{46}$$

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

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

For $N = 1$, this follows from

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

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

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

implies

$$S_N \le 2Y\mathbf{c}_\Phi \left(2aqN\right)^{1/q} \tag{47}$$

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 (47) by $f_N := f_N(Y, \mu_N)$.

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

$$S_{N-1} \le 2Y\mathbf{c}_\Phi \left(2aq(N-1)\right)^{1/q} \quad \& \quad f_N > 2Y\mathbf{c}_\Phi \left(2aqN\right)^{1/q} \tag{48}$$

and arrive at a contradiction. By the definition of $\rho^\ddagger$,

$$S_{N-1} \ge f_N \left(1 - \rho_U^\ddagger \left(\frac{2Y \left\|\Phi(\mu_N, x_N)\right\|}{f_N}\right)\right)$$

(cf. Figure 3). Since $f_N > 2Y\mathbf{c}_\Phi$ (remember that we are assuming (48)), by Corollary 3 and Lemma 2 this implies

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

As the right-hand side is a monotonically increasing function of f_N (which can be checked by differentiation), in combination with (48) the last inequality gives

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

i.e.,

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

It remains to rewrite the last inequality as

$$N^{1/q} - (N-1)^{1/q} < \frac{1}{q} N^{1/q-1} \tag{49}$$

and notice that, by the mean-value theorem, the left-hand side of (49) equals

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

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

The feature mapping for the proof of Theorem 1

In the proof of Theorem 1 we will 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

$$\Phi(\mu, x) := \left(\Phi_1(\mu, x), \Phi_2(\mu, x)\right) \tag{50}$$

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, (28), and (30), $\rho_U(\tau) \le a\tau^q$, where $a := 4.34/q$. With the help of Theorem 2, we obtain for the BBK29 algorithm with parameter Φ:

$$\left|\sum_{n=1}^{N}(y_n-\mu_n)\mu_n\right| = \left\|\sum_{n=1}^{N}(y_n-\mu_n)\Phi_1(\mu_n,x_n)\right\|_{\mathbb{R}}$$
$$\le \frac{1}{\sqrt{a_1}}\left\|\sum_{n=1}^{N}(y_n-\mu_n)\Phi(\mu_n,x_n)\right\|_U \le \frac{1}{\sqrt{a_1}}2Y\mathbf{c}_\Phi\left(2aqN\right)^{1/q} \quad (51)$$

and

$$\left|\sum_{n=1}^{N}(y_n-\mu_n)D(x_n)\right| = \left|\sum_{n=1}^{N}(y_n-\mu_n)\mathbf{k}_{x_n}(D)\right| = \left|\left(\sum_{n=1}^{N}(y_n-\mu_n)\mathbf{k}_{x_n}\right)(D)\right|$$
$$\le \left\|\sum_{n=1}^{N}(y_n-\mu_n)\mathbf{k}_{x_n}\right\|_{\mathcal{F}^*}\|D\|_{\mathcal{F}} = \left\|\sum_{n=1}^{N}(y_n-\mu_n)\Phi_2(\mu_n,x_n)\right\|_{\mathcal{F}^*}\|D\|_{\mathcal{F}}$$
$$\le \frac{1}{\sqrt{a_2}}\left\|\sum_{n=1}^{N}(y_n-\mu_n)\Phi(\mu_n,x_n)\right\|_U\|D\|_{\mathcal{F}} \le \frac{1}{\sqrt{a_2}}2Y\mathbf{c}_\Phi\left(2aqN\right)^{1/q}\|D\|_{\mathcal{F}} \quad (52)$$

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

Proof proper

The proof is based on the inequality

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

(immediately following from (20)). Using this inequality and (51)–(52) 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:

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

Since

$$\mathbf{c}_\Phi \le \sqrt{a_1 Y^2 + a_2 \mathbf{c}_{\mathcal{F}}^2} = \sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1},$$

we can see that (4) holds with

$$4(2aq)^{1/q} = 4 \times 8.68^{1/q} \tag{53}$$

in place of 40.

6 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 Z is contained in its "reproducing kernel", which is a symmetric positive definite function on Z^2 ([3], §§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 $\|\cdot\|_{B(z_1,\dots,z_n)}$ on $\mathbb{R}^n$ and satisfies the following conditions (familiar from Kolmogorov's existence theorem [21], §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 $\left(\begin{smallmatrix} 1 & 2 & \cdots & n \\ i_1 & i_2 & \cdots & i_n \end{smallmatrix}\right)$,

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

- 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 $(t_1, \dots, t_k) \in \mathbb{R}^k$,

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

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

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

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 $t_1 z_1 + \cdots + t_n z_n$, 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 $t_1 z_1 + \cdots + t_n z_n \in U_1$, $n > 0$, 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

$$f(z) := \phi(\Phi(z)), \tag{54}$$

where ϕ is a continuous linear functional on U, $\phi \in U^*$. The norm of the function (54) 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 (54): 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

$$\|(t_1,\dots,t_n)\|_{B(z_1,\dots,z_n)} = \|\phi \mapsto t_1\phi(\Phi(z_1)) + \dots + t_n\phi(\Phi(z_n))\|_{U^{**}} \tag{55}$$

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 (55) as

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

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

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

The last equality follows from the fact that the canonical imbedding of U into U^{**} is an isometry ([31], §4.5). ∎

Remark A Banach kernel B on 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., [31], 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 hyperplane $z_{k+1} = \dots = z_n = 0$.

Now we can state more explicitly the prediction algorithm described above and guaranteeing (4). Following (45) (with Φ defined by (50)), define

$$\begin{aligned} f_n(y,\mu) := \Bigg(& \frac{1}{Y^2}\left(\sum_{i=1}^{n-1}(y_i-\mu_i)\mu_i + (y-\mu)\mu\right)^2 \\ & + \|(y_1-\mu_1,\dots,y_{n-1}-\mu_{n-1},y-\mu)\|^2_{B(x_1,\dots,x_{n-1},x_n)}\Bigg)^{1/2} \\ & - \Bigg(\frac{1}{Y^2}\left(\sum_{i=1}^{n-1}(y_i-\mu_i)\mu_i\right)^2 \\ & + \|(y_1-\mu_1,\dots,y_{n-1}-\mu_{n-1})\|^2_{B(x_1,\dots,x_{n-1})}\Bigg)^{1/2}. \end{aligned} \tag{56}$$

This allows us to give the kernel representation of BBK29 with Φ defined by (50); 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 (56).
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.

Acknowledgments

I am grateful to Glenn Shafer for a series of useful discussions. This work was partially supported by MRC (grant S505/65) and the Royal Society.

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. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[4] 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.

[5] Józef Banaś. On moduli of smoothness of Banach spaces. *Bulletin of the Polish Academy of Sciences. Mathematics*, 34:287–293, 1986.

[6] Józef Banaś and Krzysztof Frączek. Deformation of Banach spaces. *Commentationes Mathematicae Universitatis Carolinae*, 34:47–53, 1993.

[7] Diómedes Bárcenas, Vladimir I. Gurary, Luisa Sánchez, and Antonio Ullán. On moduli of convexity in Banach spaces. *Quaestiones Mathematicae*, 27:137–145, 2004.

[8] Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50:2050–2057, 2004.

[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] James A. Clarkson. Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40:396–414, 1936.

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

[12] T. Figiel. On the moduli of convexity and smoothness. *Studia Mathematica*, 56:121–155, 1976. Available free of charge at `http://matwbn.icm.edu.pl`.

[13] E. Llorens Fuster. Moduli and constants: ... what a show! Available on the Internet (accessed in November 2005), May 2005.

[14] Vladimir I. Gurary. On differential properties of the complexity moduli of Banach spaces (in Russian). *Matematicheskie Issledovaniya*, 2:141–148, 1967.

[15] Olof Hanner. On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3:239–244, 1956.

[16] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[17] Rudolph E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82D:35–45, 1960.

[18] Rudolph E. Kalman and Richard S. Bucy. New results in linear filtering and prediction theory. *Transactions of the ASME—Journal of Basic Engineering*, 83D:95–108, 1961.

[19] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, New York, second edition, 1991.

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

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

[22] Joram Lindenstrauss. On the modulus of smoothness and divergent series in Banach spaces. *Michigan Mathematical Journal*, 10:241–252, 1963.

[23] Joram Lindenstrauss and Lior Tzafriri. *Classical Banach Spaces II: Function Spaces*, volume 97 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, Berlin, 1979.

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

[25] Robert S. Liptser and Albert N. Shiryaev. *Statistika sluchainykh protsessov*. Nauka, Moscow, 1974. English translation: *Statistics of Random Processes*. Springer, New York. In two volumes: *General Theory* (1977) and *Applications* (1978).

[26] J. T. Marti. Evaluation of the least constant in Sobolev's inequality for $H^1(0, s)$. *SIAM Journal on Numerical Analysis*, 20:1239–1242, 1983.

[27] Sergei M. Nikolsky. On imbedding, continuation and approximation theorems for differentiable functions of several variables. *Russian Mathematical Surveys*, 16(5):55–104, 1961. Russian original in: *Uspekhi matematicheskikh nauk*, 16(5):63–114.

[28] G. Nördlander. The modulus of convexity in normed linear spaces. *Arkiv för Matematik*, 4:15–17, 1960.

[29] Carl J. Nuzman and H. Vincent Poor. Linear estimation of self-similar processes via Lamperti's transformation. *Journal of Applied Probability*, 37:429–452, 2000.

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

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

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

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

[34] H. W. Sorenson. Least-squares estimation: from Gauss to Kalman. *IEEE Spectrum*, 7:63–68, 1970.

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

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

[37] Vladimir Vovk. On-line regression competitive with reproducing kernel Hilbert spaces. Technical Report `arXiv:cs.LG/0511058` (version 2), `arXiv.org` e-Print archive, January 2006.