

## Competing with wild prediction rules

Vladimir Vovk

Computer Learning Research Centre  
Department of Computer Science  
Royal Holloway, University of London  
Egham, Surrey, England

vovk@cs.rhul.ac.uk

Pittsburgh, June 24, 2006

## Prediction protocol

The square-loss regression:

FOR  $n = 1, 2, \dots$ :

Reality announces  $x_n \in \mathbf{X}$ .

Predictor announces  $\mu_n \in \mathbb{R}$ .

Reality announces  $y_n \in [-Y, Y]$ .

END FOR.

$x_n$ : **object** (the data relevant for predicting  $y_n$ , perhaps including  $n$  and some of the previous  $y_{n-1}, y_{n-2}, \dots$ ;

$y_n$ : its **label**;  $Y$  is fixed throughout.

Predictor's goal:

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

for all  $N = 1, 2, \dots$  and all  $D \in \mathcal{F}$ , with the function class  $\mathcal{F}$  as large as possible.

Predictor's strategies in the prediction protocol: [prediction algorithms](#).

## Competing with RKHS

This is a typical known result: if  $\mathcal{F}$  is a “continuous reproducing kernel Hilbert space with finite imbedding constant”, Predictor can guarantee

$$\frac{1}{N} \sum_{n=1}^N \lambda(x_n, \mu_n, y_n) \leq \frac{1}{N} \sum_{n=1}^N \lambda(x_n, D(x_n), y_n) + O\left(\frac{1}{\sqrt{N}}\right).$$

Unfortunately, Hilbert spaces are rather restrictive.

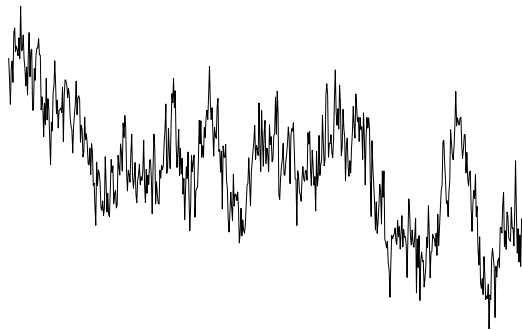
## Hölder exponent

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  $|dx|$ .

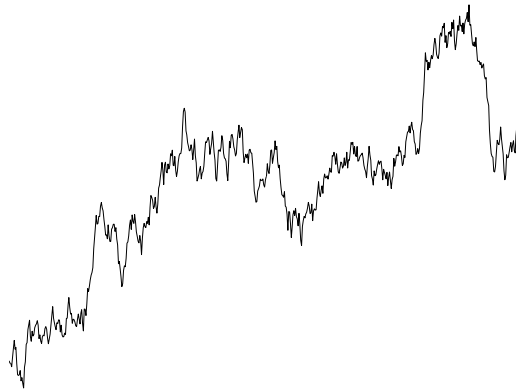
Functions of classical analysis (such as  $\sin x$ ): the Hölder exponent is 1. Typical trajectories of the Brownian motion (more generally, of non-degenerate diffusion processes): the Hölder exponent is  $1/2$ .

## Some examples

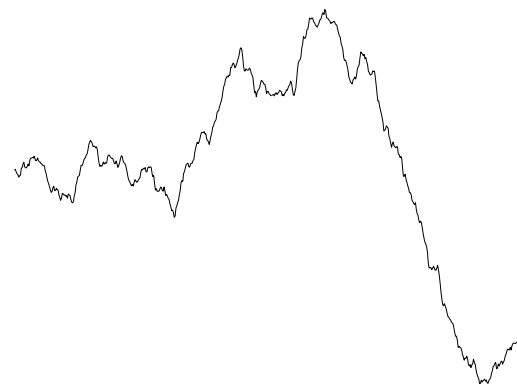
$h = 0.2$



$h = 0.5$



$h = 0.8$



## Sobolev–Slobodetsky norm

Roughly, the Sobolev spaces  $W^{s,p}([0, 1])$ , where  $p \in (1, \infty]$ ,  $s \in (0, 1)$ , and  $s > 1/p$ , are different ways of formalizing the notion of a function on  $[0, 1]$  with Hölder exponent  $h$  above the threshold  $s$ .

Let, for simplicity,  $\mathbf{X} = [0, 1]$ ,  $s \in (0, 1)$  and  $p > 1/s$ . For  $f \in L^p(\mathbf{X})$  define

$$\|f\|_{s,p} := \left( \int_{\mathbf{X}} |f(x)|^p \, dx + \int_{\mathbf{X}} \int_{\mathbf{X}} \left| \frac{f(x) - f(y)}{|x - y|^s} \right|^p \frac{dx \, dy}{|x - y|} \right)^{1/p}.$$

## Sobolev spaces

$W^{s,p}(\mathbf{X})$  = the set of all  $f$  such that  $\|f\|_{s,p} < \infty$ .

The Sobolev imbedding theorem says: the functions in  $W^{s,p}(\mathbf{X})$  can be made continuous by a change on a set of measure zero (only if  $p > 1/s$ ).

All  $W^{s,p}(\mathbf{X})$  are Banach spaces, but they are **Hilbert spaces only when  $p = 2$** . Because of the condition  $s > 1/p$  only the functions with Hölder exponent  $h > 1/2$  are amenable to the usual Hilbert-space methods of analysis.

The boundary of the domain of application of the standard methods:  $D$  that look like the Brownian motion.

## Intuitive picture I

Standard methods rely 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.

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  $s$ .

## Intuitive picture II

However, the quality of prediction (as judged by our bound) will deteriorate: if  $p > 2$ ,

$$\frac{1}{N} \sum_{n=1}^N \lambda(x_n, \mu_n, y_n) \leq \frac{1}{N} \sum_{n=1}^N \lambda(x_n, D(x_n), y_n) + O(N^{-1/p}).$$

The regret term:  $O(N^{-1/2}) \mapsto O(N^{-1/p})$ ,

## Convexity of Banach spaces

Let  $U$  be a Banach space and  $S_U := \{u \in U \mid \|u\|_U = 1\}$  the unit sphere in  $U$ .

Clarkson's modulus of convexity of  $U$ :

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

(we will be mostly interested in the small values of  $\epsilon$ ).

## Proper Banach functional spaces

A Banach space  $\mathcal{F}$  of real-valued functions  $f$  on  $\mathbf{X}$  (with the standard pointwise operations) 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 the **imbedding constant**

$$\mathbf{c}_{\mathcal{F}} := \sup_{x \in \mathbf{X}} \|\mathbf{k}_x\|_{\mathcal{F}^*} < \infty, \quad (1)$$

where  $\mathcal{F}^*$  is the dual Banach space.

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

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

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

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

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

The condition of the theorem is satisfied for the Sobolev spaces  $W^{s,p}([0, 1])$ ,  $p \in [2, \infty)$  and  $s \in (1/p, 1)$ .

## Implication for the Sobolev spaces $W^{s,p}([0, 1])$

Therefore, there exists a constant  $C_{s,p} > 0$  and a prediction algorithm producing  $\mu_n \in [-Y, Y]$  that are guaranteed to satisfy

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

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

## Constant

General theorem: we can take

$$C_{s,p} = 40\sqrt{c_{s,p}^2 + 1}$$

(where  $c_{s,p} := c_{W^{s,p}(\mathbf{X})}$ ) but the proof shows that in fact

$$C_{s,p} = 4 \times 8.68^{1-1/p} \sqrt{c_{s,p}^2 + 1}$$

will suffice.

In the special case  $p = 2$ : using Banach-space methods I lost a factor of 5.89.

## Implication for Hölder spaces

The most familiar Sobolev spaces are the Hölder spaces  $W^{s,\infty}([0, 1])$ , consisting of the functions  $f$  satisfying  $|f(x) - f(y)| = O(|x - y|^s)$ . Let  $s < 1/2$ .

There is a continuous imbedding  $W^{s,\infty}(\mathbf{X}) \hookrightarrow W^{s',p}(\mathbf{X})$  whenever  $s' < s$ ;  $s'$  can be arbitrarily close to  $s$ . Therefore: for an arbitrarily small  $\epsilon > 0$  that there exists a constant  $C_{s,\epsilon} > 0$  such that

$$\frac{1}{N} \sum_{n=1}^N (y_n - \mu_n)^2 \leq \frac{1}{N} \sum_{n=1}^N (y_n - D(x_n))^2 + Y C_{s,\epsilon} (\|D\|_{s,\infty} + Y) N^{-s+\epsilon}$$

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

## Open problem

If  $\dim \mathcal{F} = K < \infty$ , Predictor can guarantee

$$\frac{1}{N} \sum_{n=1}^N \lambda(x_n, \mu_n, y_n) \leq \frac{1}{N} \sum_{n=1}^N \lambda(x_n, D(x_n), y_n) + O\left(\frac{K \log N}{N}\right)$$

(the “Vovk–Azoury–Warmuth forecaster”).

How to connect the regret term  $O(K \log N/N)$  with  $O(N^{-1/2})$  (Hilbert spaces) and  $O(N^{-1/p})$  (Banach spaces)?