Some uses of metric entropy in on-line learning

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

Edinburgh, September 12, 2006
Plan for this talk (maybe overoptimistic):

- Competitive on-line prediction as strand of learning theory
- Universal prediction strategies
- We need more than universal strategies: competing with dense benchmark classes
- The method of metric entropy
- Its limitations and other methods
- Functional modelling
On-line prediction protocol

The square-loss regression:

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

\(x_n\): signal (the data relevant for predicting \(y_n\), perhaps including some of the previous \(y_{n-1}, y_{n-2}, \ldots\))

\(y_n\): observation

\(Y := 1\) (general case: scaling)
Example

$y_n$: (high) temperature in Edinburgh on day $n$

$x_n$: the data available when the prediction is made

- Our prediction protocol: on-line.
- It is perfect-information: like chess.

How is it related to Prof Temlyakov’s talk?
The other strand of learning theory represented at this workshop: statistical learning theory.

Its basic set-up is **batch**: you are given a training set, and the goal is to come up with a good prediction rule \( F : \mathbf{X} \rightarrow \mathbb{R} \).

It makes the **i.i.d. assumption**:

\((x_n, y_n)\) are generated independently from the same distribution
Goals of learning

Statistical learning theory: come up with a prediction rule $F$ with a small expected loss.

“Expected”: w.r. to the true probability measure generating the signals and observations.

Competitive on-line learning (universal prediction of individual sequences):

- no stochastic assumptions at all
- the goal is a good actual (not expected) performance (no measure $\therefore$ no expectation)
Predictor’s goal in competitive on-line prediction

We want Predictor to achieve

\[
\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 \approx \frac{1}{N} \sum_{n=1}^{N} (y_n - F(x_n))^2
\]

for all \( N = 1, 2, \ldots \) and all \( F \in \mathcal{F} \), for a large function class \( \mathcal{F} \).
Universal prediction strategies

There is a strategy for Predictor that asymptotically dominates every continuous prediction rule:

**Theorem 1** Let $X$ be a metric compact. There exists a strategy for Predictor that guarantees

$$\limsup_{N \to \infty} \left( \frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 - \frac{1}{N} \sum_{n=1}^{N} (y_n - F(x_n))^2 \right) \leq 0$$

for each continuous prediction rule $F$. 
Aggregation of prediction strategies

Lemma Let $F_1, F_2, \ldots$ be a sequence of prediction rules assigned positive weights $w_1, w_2, \ldots$ summing to 1. There is a strategy for Predictor producing $\mu_n \in [-1, 1]$ that are guaranteed to satisfy, for all $N = 1, 2, \ldots$ and all $i = 1, 2, \ldots$,

$$
\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F_i(x_n))^2 + 8 \ln \frac{1}{w_i}.
$$

- you can aggregate any strategies, not just prediction rules
- this is true for a wide class of loss functions
- 8 can be replaced by 2, but the algorithm would be slightly more complicated
Proof sketch

The algorithm maintains the weights $p_{i,n}$ for the prediction rules $F_i$; $w_i = p_{i,0}$ are the initial weights.

At each step the weights are updated

$$p_{i,n} \propto p_{i,n-1} e^{-\eta(y_n-F(x_n))^2}$$

(always sum to 1) and Predictor’s prediction is computed as the weighted average

$$\mu_n := \sum_{i=1}^{\infty} p_{i,n} F_i(x_n).$$

I will call this strategy the mixture of $F_i$ (more generally, of a sequence of prediction strategies).
Proof sketch of Theorem 1

Since $C(X)$ is separable, we can mix a dense sequence of $F_i \in C(X)$. 
Inequalities instead of asymptotics

If \( \mathcal{F} \) is a suitable benchmark class (Banach space, not as big as \( C(\mathbf{X}) \)), Predictor can guarantee

\[
\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + g(\|F\|_{\mathcal{F}}, N)
\]

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

The regret term \( g(\|F\|_{\mathcal{F}}, N) \) must be \( o(N) \) (and not grow too fast with \( \|F\|_{\mathcal{F}} \)).
Metric entropy

Let $A$ be a compact metric space. The metric entropy $\mathcal{H}_\varepsilon(A)$, $\varepsilon > 0$, is the binary logarithm $\log K$ of the minimum number of elements $F_1, \ldots, F_K \in A$ that form an $\varepsilon$-net for $A$.

Nowadays: entropy numbers appear more popular.

Kolmogorov and Tikhomirov 1959 (KT59): 4 main variations on the notion of metric entropy,

$$\mathcal{E}_{2\varepsilon}(A) \leq \mathcal{H}_\varepsilon^{\text{abs}}(A) \leq \mathcal{H}_\varepsilon^{R}(A) \leq \mathcal{H}_\varepsilon(A) \leq \mathcal{E}_\varepsilon(A).$$
Four types of metric compacts

$U_\mathcal{F}$: unit ball in $\mathcal{F}$

KT59 classification and the corresponding regret terms:

(I) finite dimensional function classes $\mathcal{F}$:

$$\mathcal{H}_\epsilon(U_\mathcal{F}) = O \left( \log \frac{1}{\epsilon} \right);$$

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + O(\log N);$$
(II) typical classes $\mathcal{F}$ of analytic functions of $m$ variables:

$$\mathcal{H}_\epsilon(U_\mathcal{F}) = O\left(\log^{m+1} \frac{1}{\epsilon}\right);$$

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + O(\log^{m+1} N);$$

(III) typical classes $\mathcal{F}$ of functions of $m$ real variables with “smoothness indicator” $s$:

$$\mathcal{H}_\epsilon(U_\mathcal{F}) = O\left(\left(\frac{1}{\epsilon}\right)^{m/s}\right);$$

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + O\left(N^{\frac{m}{m+s}}\right);$$
(IV) for classes $\mathcal{F}$ of Lipschitzian functionals on classes of type III (such $\mathcal{F}$ are representative of type IV):

$$\mathcal{H}_\epsilon(U_{\mathcal{F}}) = O\left(C(1/\epsilon)^{m/s}\right);$$

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + O\left(N/\log^{s/m} N\right).$$
State of the art

Regret terms known in competitive on-line prediction (to my knowledge): only types I and III.

Namely:
- $O(N^{1/2})$: Cesa-Bianchi, Long, Warmuth, . . . , starting from 1996, for Hilbert spaces (not always explicit);
- $O(\log N)$: V., Azoury, Warmuth, . . . , starting from 1998, for linear regression (precursor: Foster, 1991);
- $O(N^{1-1/p})$, $p \geq 2$: V., COLT’2006 (June), for Banach spaces that are as convex as $L_p$ (such as $B^s_{p,q}$, $\frac{p}{p-1} \leq q \leq p$: Cobos & Edmunds, 1988).

Now we have the whole spectrum.
Compact benchmark classes

**Theorem 2** Suppose $\mathcal{F}$ is a compact set in $C(X)$. There exists a strategy for Predictor that produces $\mu_n$ with $|\mu_n| \leq 1$ and guarantees, for all $N = 1, 2, \ldots$ and all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2$$

$$+ C \inf_{\epsilon \in (0,1/2]} \left( \mathcal{H}_\epsilon(\mathcal{F}) + \log \log \frac{1}{\epsilon} + \epsilon N + 1 \right),$$

where $C$ is a universal constant.
Proof sketch

- Consider only $\epsilon$ of the form $2^{-i}$, $i = 1, 2, \ldots$.
- Fix, for each $i$, a $2^{-i}$-net $\mathcal{F}_i$ for $\mathcal{F}$ of size $2^{\mathcal{H}_{2^{-i}}(\mathcal{F})}$.
- To each element of $\mathcal{F}_i$ assign weight $\propto i^{-2}2^{-\mathcal{H}_{2^{-i}}(\mathcal{F})}$.
- Mix all these prediction rules.
Banach function spaces as benchmark classes

A Banach space $\mathcal{F}$ is compactly embedded into $C(X)$ if $U_{\mathcal{F}}$ is a compact subset of $C(X)$.

**Theorem 3** Let $\mathcal{F}$ be a Banach space compactly embedded in $C(X)$. There exists a strategy for Predictor that produces $\mu_n$ with $|\mu_n| \leq 1$ and guarantees, for all $N = 1, 2, \ldots$ and all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \inf_{\epsilon \in (0,1/2]} \left( \mathcal{H}_{\epsilon/\phi}(U_{\mathcal{F}}) + \log \log \frac{1}{\epsilon} + \log \log \phi + \epsilon N + 1 \right),$$

where $C'$ is a universal constant and $\phi := 2 \max(1, \|F\|_{\mathcal{F}})$.  

20
Proof sketch

• Notice that $\mathcal{H}_\epsilon(2^i U_F) = \mathcal{H}_{2^{-i}\epsilon}(U_F)$, $i = 1, 2, \ldots$.

• Apply Theorem 2 to $\mathcal{F} := 2^i U_F$, assigning weight $\propto i^{-2}$ to the corresponding prediction strategy.

• Mix these strategies.
Competing with the continuous prediction rules

Let \( \mathcal{F} \subseteq C(X) \) be a Banach function space dense in \( C(X) \) (densely embedded in \( C(X) \)). The approachability of \( F \in C(X) \) by \( \mathcal{F} \) is

\[
\mathcal{A}_\epsilon^\mathcal{F}(F) := \inf \left\{ \| F^* \|_{\mathcal{F}} \mid \| F - F^* \|_{C(X)} \leq \epsilon \right\}, \quad \epsilon > 0
\]

(finite under our assumption of density).

[equivalent ways of talking about \( \mathcal{A} \): Gagliardo diagram, K norm, ...]
Theorem 4 Let $\mathcal{F}$ be a Banach function space compactly and densely embedded in $C(X)$. Theorem 3's strategy guarantees, for all $N = 1, 2, \ldots$ and $F \in C(X)$,

$$
\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2
$$

$$
+ C \inf_{\epsilon \in (0, 1/2]} \left( \mathcal{H}_{\epsilon/A(\epsilon)}(U_F) + \log \log \frac{1}{\epsilon} + \log \log A(\epsilon) + \epsilon N + 1 \right),
$$

where $C$ is a universal constant and $A(\epsilon) := 2 \max(1, A_{\epsilon}(F))$.

Proof: immediate from Theorem 3.
Theorem 4: source of many universal prediction strategies.

Many Banach spaces of types II and III are compactly and densely embedded in $C(X)$.

Given any Banach space compactly and densely embedded in $C(X)$ Theorem 4 produces a universal prediction strategy.
Example 1 of type II class

Let $K$ be a simply connected continuum in $\mathbb{C}$ containing more than one point and $G$ be a connected open set such that $K \subseteq G \subseteq \mathbb{C}$.

$A^K_G$: the set of all complex-valued functions on $K$ that admit a bounded analytic continuation to $G$.

The norm:

$$\|f|_K\|_{A^K_G} := \sup_{z \in G} |f(z)|,$$

where $f : G \rightarrow \mathbb{C}$ ranges over the bounded analytic functions.
Example 1 cont.

\[ \mathcal{H}_\varepsilon \left( U_{A^K_G} \right) \sim \tau(G, K) \log^2 \frac{1}{\varepsilon} \]

(KT59; hypothesised by Kolmogorov and proved independently by Babenko and Erokhin).

Theorem 3 gives:

\[
\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C\tau(G, K) \log^2 N
\]

for all real-valued \( F \in A^K_G \) and from some \( N \) on, where \( C \) is a universal constant.

Vitushkin: \( \tau(G, K) = 1/(2 \log \lambda) \) if \( K = [-1, 1] \) and \( G \) is the ellipse \( E_\lambda \) with the sum of semi-axes equal to \( \lambda > 1 \) and with foci at the points \( \pm 1 \).
Example 2 of type II class

Let $h > 0$.

$A_h$: the vector space of all periodic period $2\pi$ complex-valued functions on the real line $\mathbb{R}$ that admit a bounded analytic continuation to the strip $\{z \in \mathbb{C} | |\Im z| < h\}$

The norm:

$$\|f\|_{A_h} := \sup_{z:|\Im z|<h} |f(z)|,$$

where $f$ ranges over the bounded analytic functions on $\{z | |\Im z| < h\}$. 
Example 2 cont.

\[ \mathcal{H}_\epsilon(U_{A_h}) \sim \frac{2}{h \log e} \log^2 \frac{1}{\epsilon} \]

(KT59, Vitushkin).

Theorem 3 now gives

\[
\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + \frac{C}{h} \log^2 N
\]

for all real-valued \( F \in A_h \) and from some \( N \) on, where \( C \) is a universal constant.

Both \( A_h \) and \( A_{E,\lambda}^{[-1,1]} \) are dense, and so give rise to universal prediction strategies.
Example 1 of type III spaces

Suppose $X$ is a subset of Euclidean space, $X \subseteq \mathbb{R}^m$, which is a minimally regular domain (bounded and coincides with the interior of its closure).

Every $B_{p,q}^s(X)$ with $s > m/p$ is compactly embedded in $C(X)$. Edmunds and Triebel’s (1996) general result implies

$$\mathcal{H}_\epsilon \left( U_{B_{p,q}^s(X)} \right) \asymp (1/\epsilon)^{m/s}$$

(where $U_{B_{p,q}^s(X)}$ is considered a subset of $C(X)$).

[The same as in Prof Triebel’s talk!]
Example 1 cont.

Theorem 3 then shows that

\[ \sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 \]

\[ + C_{X,s,p,q} \left( \|F\|_{B_{p,q}^s(X)} + 1 \right)^{\frac{m}{m+s}} N^{\frac{m}{m+s}} \]

for all \( F \in B_{p,q}^s(X) \) from some \( N \) on.

Cucker & Smale 2002 obtain the rate \( N^{\frac{m}{m+s}} \) for \( H^s(X) \) under the i.i.d. assumption ((5) with \( \delta := 1 \)).
Example 2 (type 2.5?): smooth RKHS

Cucker and Smale (2001): if $\mathcal{F}$ is an RKHS with a $C^\infty$ reproducing kernel on $\mathbb{R}^2$ for a compact set $\mathbb{R}^m$,

$$\mathcal{H}_\epsilon(U_\mathcal{F}) = O\left((1/\epsilon)^{2m/h}\right)$$

for an arbitrary $h > m$.

From Theorem 3: for an arbitrarily small $\delta > 0$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + N\delta$$

for all $F \in \mathcal{F}$ from some $N$ on.

The regret term is worse than poly-log: the class of analytic functions is much narrower than that of infinitely differentiable functions.
Two limitations of the metric entropy method

- It gives prediction strategies that cannot be written in a closed form (and are not computationally efficient).
- It does not give optimal regret terms (at least for type III classes whose members are not very smooth): even exponents, not only constants.
Alternatives:

• apply the aggregating algorithm to $F$ without “discretization”: weighted summation $\rightarrow$ integration (continuous mixing)

• defensive forecasting (a new method originating in the game-theoretic foundations for probability)

• Gradient Descent and its versions, following the perturbed leader, etc.

The first tends to give the best constants; the second is almost as good. [Attention to constants in learning theory: perhaps impetus is coming from experimental machine learning.] Other methods: often computationally very efficient.
Comparisons with the method of “defensive forecasting”

Many of the Besov spaces $B_{p,q}^s(X)$ are “uniformly convex”.

Clarkson’s modulus of convexity:

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

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

If a Banach space $\mathcal{F}$ is continuously embedded in $C(X)$, the embedding constant is

$$c_\mathcal{F} := \sup_{F \in U_\mathcal{F}} \|F\|_{C(X)} < \infty.$$
Proposition (my COLT’2006 paper) Let $\mathcal{F}$ be a Banach space continuously embedded in $C(\mathbf{X})$ and 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 strategy for Predictor producing $\mu_n$ that are guaranteed to satisfy

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

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

$\mathcal{F}$ is not required to be compactly embedded in $C(\mathbf{X})$.

When $p = 2$: $40 \mapsto 2$; continuous mixing: $\sqrt{c_\mathcal{F}^2 + 1} \mapsto c_\mathcal{F}$ (optimal).
Convexity of Besov spaces

Clarkson (1936): for \( p \in [2, \infty) \),
\[
\delta_{L^p}(\epsilon) \geq 1 - (1 - (\epsilon/2)^p)^{1/p} \geq (\epsilon/2)^p/p.
\]

Extended to some other Besov spaces by Cobos and Edmunds (1988): the modulus of convexity of each \( B^s_{p,q}(\mathbb{R}^m) \), \( s \in \mathbb{R} \), \( p \in [2, \infty) \) and \( q \in [p/(p-1), p] \), also satisfies
\[
\delta_{B^s_{p,q}(\mathbb{R}^m)}(\epsilon) \geq 1 - (1 - (\epsilon/2)^p)^{1/p} ;
\]
easily extends to \( B^s_{p,q}(X) \).
Defensive forecasting for Besov spaces

Let $p \in [2, \infty)$, $q \in [p/(p - 1), p]$ and $s \in (m/p, \infty)$. There exist a constant $C_{X,s,p,q} > 0$ and a strategy for Predictor producing $\mu_n$ that are guaranteed to satisfy

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_{X,s,p,q} \left( \|F\|_{B_{p,q}^s} + 1 \right) N^{1 - 1/p}$$

for all $N = 1, 2, \ldots$ and all $F \in B_{p,q}^s(X)$. 
Comparison for the Hölder–Zygmund spaces $C^s(X) := B_{\infty, \infty}^s(X)$

For $s = k + \alpha$, where $k$ is integer and $\alpha \in (0, 1)$, $C^s(X)$ consists of the functions whose $k$th partial derivatives exist and are all Hölder continuous of order $\alpha$.

Defensive forecasting works better than metric entropy at the “rough” end of the scale $C^s(X)$:
Suppose $s \in (0, m/2]$. The DF exponent $1 - 1/p$ of $N$ can be taken arbitrarily close to $1 - s/m$, and we can see that it is then better than the ME exponent of $N$:

$$1 - \frac{s}{m} < \frac{m}{m + s}.$$  

For example, if $m = 1$, $s \approx 1/2$ (typical trajectories of the Brownian motion are of this type) defensive forecasting gives approximately $N^{1/2}$ whereas metric entropy gives approximately $N^{2/3}$. 
• Suppose $s \in (m/2, m)$. The DF exponent of $N$ can always be taken as $1/2$, and it is still better than the ME exponent of $N$:

$$\frac{1}{2} < \frac{m}{m + s}.$$ 

• Suppose $s \in [m, \infty)$. A weakness of the method of defensive forecasting (in its current state) is that it cannot give regret terms better than $O(N^{1/2})$. Therefore, the method of metric entropy beats defensive forecasting for smooth $C^s(X)$, $s > m$. 
Functional modelling

Competitive on-line prediction: statistical models $\mapsto$ functional models (=benchmark classes)

What if we choose a “wrong” model?

It appears that: choosing a meagre (but dense in $C(X)$) class is safer than choosing a rich class.
Choosing a wrong class of type II

Lemma Let $0 < h < H < \infty$ and let $F \in A_h$. For small enough $\epsilon > 0$,

$$\log A_{\epsilon}^{A_H}(F) \leq C \frac{H}{h} \log \frac{1}{\epsilon},$$

where $C$ is a universal constant.

Proof idea: Functions in $A_h$ can be very well approximated in $C(X)$ by low-degree trigonometric polynomials (Akhiezer's theorem), whose $A_H$ norm is not too large.
In combination with Theorem 4 this lemma gives:

**Corollary** Let $0 < h < H < \infty$. The strategy for Predictor constructed earlier for the benchmark class $A_H$ guarantees

$$
\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \frac{H^2}{h^3} \log^2 N
$$

for each $F \in A_h$ from some $N$ on, where $C$ is a universal constant.
Cost of using a wrong class

This is what happens with the regret term:

- if we use $A_h$ instead of $A_H$ (err on the side of caution):
  \[
  \frac{1}{H} \log^2 N \rightarrow \frac{1}{h} \log^2 N
  \]
  (lose a factor of $H/h$);

- if we use $A_H$ instead of $A_h$ (being too optimistic):
  \[
  \frac{1}{h} \log^2 N \rightarrow \frac{H^2}{h^3} \log^2 N
  \]
  (lose a factor of $(H/h)^2$).

It might be slightly better to be a pessimist (but not much difference).
Caveat (for the previous and following slides): I am talking about the available performance guarantees, which might not be optimal.
Choosing a wrong type

Conclusion: if you optimistically choose type II instead of type III, you might lose half of the smoothness \((s \mapsto s/2)\).

**Lemma** Let \(h > 0\) and let \(F : \mathbb{R} \rightarrow \mathbb{R}\) be a non-zero periodic function with period \(2\pi\) whose \(k\)th derivative \((k \in \{0, 1, \ldots\})\) exists and is Hölder continuous of order \(\alpha \in (0, 1]\) with coefficient \(c\). Set \(s := k + \alpha\). For small enough \(\epsilon > 0\),

\[
\log A_{\epsilon}^{A_h}(F) \leq C h \left(\frac{12c}{\epsilon}\right)^{1/s},
\]

where \(C\) is a universal constant.

**Proof idea:** use Jackson’s theorem instead of Akhiezer’s.
Combining with Theorem 4:

**Corollary** Let $F : \mathbb{R} \to \mathbb{R}$ be a periodic period $2\pi$ function whose $k$th derivative ($k \geq 0$) is Hölder continuous of order $\alpha$ with coefficient $c$. The strategy for Predictor constructed for the class $A_h$ guarantees

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + Ch^{s/2}c_s^{1/2}N^{s/2}$$

from some $N$ on, where $s := k + \alpha$ and $C$ is a universal constant.

The growth rate $N^{2/(s+2)} = N^{1/(s/2+1)}$ of the regret term is worse than the rate $N^{1/(s+1)}$ obtained (using ME) for a prediction strategy designed specifically for functions with Hölder continuous derivatives.
Choosing a wrong class of type III

I will state two simple corollaries of

\[ s_0 \neq s_1 \implies (B_{p,q_0}^{s_0}, B_{p,q_1}^{s_1})_{\theta,r} = B_{p,r}^{(1-\theta)s_0 + \theta s_1}, \]

for the Hölder–Zygmund spaces \( \mathcal{C}^s(X) := B_{\infty,\infty}^{s}(X) \).
Defensive forecasting bound

The regret term is of order, approximately,

$$\|F\|_{\mathcal{C}^s(X)} N^{1-s/m}$$

(1)

for the benchmark class $\mathcal{C}^s(X)$, $0 < s \leq m/2$, and of order

$$\|F\|_{\mathcal{C}^S(X)} N^{1-S/m}$$

(2)

for the benchmark class $\mathcal{C}^S(X)$, $0 < S \leq m/2$.

Let $s < S$. Achieving (2) automatically achieves (1) (ignoring constant factors).
Metric entropy bound

The regret terms are of order, approximately,

$$\|F\|_{\mathcal{C}^s(X)}^m N^{m+s}$$  \hspace{1cm} (1)

for the benchmark class $\mathcal{C}^s(X)$ and

$$\|F\|_{\mathcal{C}^S(X)}^m N^{m+S}$$  \hspace{1cm} (2)

for $\mathcal{C}^S(X)$, where $0 < s < S$.

Achieving (2) again automatically achieves (1) (ignoring constant factors).
Possible directions of further research

- Find computationally efficient prediction strategies for benchmark classes such as $A^K_G$ and $A_h$ (type II) and Besov spaces $B_{p,q}^s$ with $m/(m + s) < 1/2$.

- Extend the metric entropy method to discontinuous prediction rules.

- Complement the available performance guarantees with lower bounds.

- Study the “relation of domination” between various a priori plausible benchmark classes: e.g., some of them may turn out to be useless or nearly useless on purely theoretical grounds.
Full proofs for this talk

http://www.vovk.net (the front page)

Recent review of the field

Nicolò Cesa-Bianchi and Gábor Lugosi
Prediction, learning, and games
New York: Cambridge University Press, 2006