

# On-line learning competitive with reproducing kernel Hilbert spaces

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

EURANDOM, 4 October 2005

**Prediction with expert advice:** we are given a pool of decision strategies (more generally, of experts) and our goal is to perform almost as well as the best strategy in the pool. No assumptions about the environment.

**Defensive forecasting**  $\mapsto$  a new proof technique in prediction with expert advice.

This talk: **prediction**  $\mapsto$  **forecasting** or **decision making**.

Decision-making protocol:

Loss<sub>0</sub> := 0.

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

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

    Decision Maker announces  $\gamma_n \in [0, 1]$ .

    Reality II announces  $y_n \in \{0, 1\}$ .

    Loss <sub>$n$</sub>  := Loss <sub>$n-1$</sub>  +  $\lambda(y_n, \gamma_n)$ .

END FOR.

$x_n$ : datum (all relevant information);  $y_n$ : observation;  $\lambda$ : the loss function.

Decision rule  $D : \mathbf{X} \rightarrow [0, 1]$ .

We want to compete against decision rules that are not too weird with no assumptions about Reality. Let  $\mathbf{X} = [0, 1]$  at first.

A “Sobolev norm”  $\|f\|_{\mathcal{S}}$  of  $f : [0, 1] \rightarrow \mathbb{R}$  is defined by

$$\|f\|_{\mathcal{S}}^2 := \left( \int_0^1 f(t) dt \right)^2 + \int_0^1 (f'(t))^2 dt$$

( $\infty$  if  $f$  is not absolutely continuous etc.).

**Proposition** Suppose  $\mathbf{X} = [0, 1]$  and  $\lambda(y, \gamma) = |y - \gamma|$ . Decision Maker has a strategy that guarantees

$$\frac{1}{N} \sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \frac{1}{N} \sum_{n=1}^N \lambda(y_n, D(x_n)) + \frac{\|2D - 1\|_{\mathcal{S}} + 1}{\sqrt{N}}$$

for all  $N$  and  $D$ .

Intuitively: this is about a “small” decision maker.

When is Decision Maker **competitive with  $D$** ? Let

$$f := 2D - 1 \in [-1, 1]$$

(“symmetrized”  $D$ ).

We have

$$\|f\|_{\mathcal{S}} \leq \left| \int_0^1 f(t) dt \right| + \sqrt{\int_0^1 (f'(t))^2 dt} \leq 1 + \text{“mean slope of } f\text{”}.$$

OK if the mean slope  $\ll \sqrt{N}$ . Especially simple case:  
continuous piece-wise linear functions.

Later: other  $\mathbf{X}$ , norms, and loss functions.

## Similar results

Cesa-Bianchi, Conconi, Gentile (2003): specific (and different) loss functions for Decision Maker (counting mistakes) and the decision rules (hinge loss).

Long and Kinber (1994, 1997): there is a perfect decision rule.

Our approach inspired by: Foster and Vohra (1998); Fudenberg, Levine, Lehrer, Sandroni, Smorodinsky, Kakade, . . . ; no formal connections.

The rest of this talk: how to prove this and many other results.

- Game-theoretic vs. measure-theoretic probability (main example: game-theoretic vs. measure-theoretic SLLN)
- Defensive forecasting: laws of probability  $\mapsto$  forecasting algorithms
- Implementation: WLLN  $\mapsto$  ALN
- ALN in RKHS
- Theoretical properties of ALN: calibration and resolution
- Use for decision making

There are 2 main ways to formalize probability: **measure** (Borel / ... / Kolmogorov) vs. **gambling** (von Mises / Ville / Kolmogorov).

I will demonstrate the difference on the simplest **martingale SLLN**. Let  $y_1, y_2, \dots$  be random variables s.t.  $y_n \in [0, 1]$  for all  $n$ ; let  $p_n := \mathbb{E}(y_n \mid y_1, \dots, y_{n-1})$ . Then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N (y_n - p_n) = 0$$

with probability 1.

## Game-theoretic SLLN for uniformly bounded observations

Forecasting protocol:

$\mathcal{K}_0 := 1.$

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

Forecaster announces  $p_n \in [0, 1].$

Sceptic announces  $s_n \in \mathbb{R}.$

Reality announces  $y_n \in [0, 1].$

$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - p_n).$

END FOR.

$\mathcal{K}_n$ : Sceptic's capital.

I will be mainly interested in the case:  $y_n \in \{0, 1\}$  (binary probability forecasting); Reality's move  $x_n$  can also be added at the beginning of each round.

## The difference between the two protocols

- In the forecasting protocol, our goal to produce true probabilities (probabilities one cannot gamble against).
- In the decision-making protocol, we are merely minimizing our loss.

Proposition (game-theoretic SLLN) Sceptic has a strategy which guarantees that

- $\mathcal{K}_n$  is never negative
- either

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N (y_n - p_n) = 0$$

( $p_n$  are unbiased) or

$$\lim_{n \rightarrow \infty} \mathcal{K}_n = \infty.$$

The **measure-theoretic SLLN** follows easily: if Reality is **oblivious** (does not pay attention to what her opponents do) and uses a randomized strategy (probability measure  $P$  on the sequences of Reality's moves) and Forecaster computes his moves as conditional expectations w.r. to  $P$ :  $\mathcal{K}_n$  is a non-negative martingale, and so  $\mathcal{K}_n \rightarrow \infty$  with probability 0.

Game-theoretic SLLN:

- Reality need not be oblivious (or even follow a strategy)
- Forecaster need not ignore Sceptic (this is what makes defensive forecasting possible)

Caveat: I assumed that Sceptic's strategy was measurable.  
Fact of life: for all kinds of limit theorems, Sceptic's strategy we construct is measurable; moreover, it is continuous.

Recent (2004) observation: this approach can be used for designing learning algorithms.

For any continuous strategy for Sceptic there exists a strategy for Forecaster that does not allow Sceptic's capital to grow.

Modified protocol:

$\mathcal{K}_0 := 1.$

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

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

Sceptic announces continuous  $S_n : [0, 1] \rightarrow \mathbb{R}.$

Forecaster announces  $p_n \in [0, 1].$

Reality announces  $y_n \in \{0, 1\}.$

$\mathcal{K}_n := \mathcal{K}_{n-1} + S_n(p_n)(y_n - p_n).$

END FOR.

**Theorem 1** (Takemura) Forecaster has a strategy that ensures  $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \dots$ .

## Proof

- choose  $p_n$  so that  $S_n(p_n) = 0$
- if the equation  $S_n(p) = 0$  has no roots (in which case  $S_n$  never changes sign),

$$p_n := \begin{cases} 1 & \text{if } S_n > 0 \\ 0 & \text{if } S_n < 0 \end{cases}$$

QED

Has been greatly extended; Intermediate Value Theorem  $\mapsto$  Ky Fan's fixed point theorem.

## Research programme I (forecasting)

- Open a probability textbook and decide which property (such as LLN, CLT, LIL, Hoeffding's inequality, . . . ) you want Forecaster's moves to satisfy.
- Prove the corresponding game-theoretic result.
- Apply Theorem 1.

[Problem: works too well.]

What does it give in the case of LLN?

In fact, nothing interesting: Forecaster performs his task **too well**. E.g., he can choose

$$p_n := \begin{cases} 1/2 & \text{if } n = 1 \\ y_{n-1} & \text{otherwise,} \end{cases}$$

ensuring

$$\left| \sum_{i=1}^n (y_i - p_i) \right| \leq 1/2$$

for all  $n$  (much better than using the true probabilities).

We need a “convoluted” LLN. Suppose  $\Phi : [0, 1] \times \mathbf{X} \rightarrow \mathcal{H}$  (feature mapping) and

$$\sup_{p,x} \|\Phi(p, x)\|_{\mathcal{H}} \leq 1.$$

The **convoluted LLN**: for any  $\delta \in (0, 1)$ ,

$$\left\| \frac{1}{N} \sum_{n=1}^N (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}} \leq \frac{1}{\sqrt{N\delta}}$$

with probability at least  $1 - \delta$ . An easy modification of the standard statement ( $\Phi \equiv 1$ ). True both measure-theoretically (with  $\Phi$  measurable) and game-theoretically.

Let

$$\mathbf{k}((p, x), (p', x')) = \langle \Phi(p, x), \Phi(p', x') \rangle_{\mathcal{H}}.$$

(Remember the “kernel trick”.) Suppose  $\mathbf{k}$  is continuous in  $p$  and  $p'$ . Applying Theorem 1 to a standard proof of LLN:

**Theorem 2** There exists a forecasting strategy (ALN with parameter  $\mathbf{k}$ ) that guarantees

$$\forall N : \left\| \frac{1}{N} \sum_{n=1}^N (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}} \leq \frac{1}{\sqrt{N}}.$$

(Somewhat better than when using the true probabilities, esp. in view of the LIL.)

If the “surface”  $\Phi$  is sufficiently convoluted: “calibration” and “resolution” .

More direct interpretation: RKHS.

## Optimality result

Forecaster can ensure

$$\forall N : \left\| \sum_{n=1}^N (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}} \leq \sqrt{\sum_{n=1}^N p_n (1 - p_n) \|\Phi(p_n, x_n)\|_{\mathcal{H}}^2}.$$

Reality II can ensure

$$\forall N : \left\| \sum_{n=1}^N (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}}^2 \geq \sum_{n=1}^N p_n (1 - p_n) \|\Phi(p_n, x_n)\|_{\mathcal{H}}^2.$$

A **reproducing kernel Hilbert space** (RKHS) on  $\mathbf{X}$  is a Hilbert space  $\mathcal{F}$  of real-valued functions on  $\mathbf{X}$  such that the evaluation functional  $f \in \mathcal{F} \mapsto f(x)$  is continuous for each  $x \in \mathbf{X}$ . By the Riesz–Fischer theorem, for each  $x \in \mathbf{X}$  there exists a function  $\mathbf{k}_x \in \mathcal{F}$  such that

$$f(x) = \langle \mathbf{k}_x, f \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}.$$

Let

$$\mathbf{c}_{\mathcal{F}} := \sup_{x \in \mathbf{X}} \|\mathbf{k}_x\|_{\mathcal{F}};$$

we will be interested in the case  $\mathbf{c}_{\mathcal{F}} < \infty$ .

The corresponding kernel:

$$\mathbf{k}(x, x') := \langle \mathbf{k}_x, \mathbf{k}_{x'} \rangle_{\mathcal{F}}.$$

Theorem 2 implies:

**Theorem 3** ALN with this kernel ensures

$$\left| \frac{1}{N} \sum_{n=1}^N (y_n - p_n) f(p_n, x_n) \right| \leq \frac{\mathbf{c}_{\mathcal{F}} \|f\|_{\mathcal{F}}}{\sqrt{N}}$$

for all  $N$  and  $f$ .

## Optimality result

Forecaster can ensure that for each  $N$ :

$$\left| \sum_{n=1}^N (y_n - p_n) f(p_n, x_n) \right| \leq \|f\|_{\mathcal{F}} \sqrt{\sum_{n=1}^N p_n(1 - p_n) \mathbf{k}((p_n, x_n), (p_n, x_n))}.$$

Reality II can ensure that for each  $N$  there exists a non-zero  $f \in \mathcal{F}$ :

$$\left| \sum_{n=1}^N (y_n - p_n) f(p_n, x_n) \right| \geq \|f\|_{\mathcal{F}} \sqrt{\sum_{n=1}^N p_n(1 - p_n) \mathbf{k}((p_n, x_n), (p_n, x_n))}.$$

## Examples

On  $[0, 1]$ :

$$\|f\|_{\mathcal{S}}^2 := \left( \int_0^1 f(t) dt \right)^2 + \int_0^1 (f'(t))^2 dt$$

with kernel

$$\mathbf{k}(x, x') = \frac{1}{2} \min(x, x') + \frac{1}{2} \min(1 - x, 1 - x') + \frac{5}{6}$$

(Craven and Wahba 1979, used above).

On  $\mathbb{R}$ :

$$\|f\|_{\mathcal{S}'}^2 := \int_{-\infty}^{\infty} f^2(t) dt + \int_{-\infty}^{\infty} (f'(t))^2 dt$$

with kernel

$$\mathbf{k}(x, x') = \frac{1}{2} \exp(-|x - x'|)$$

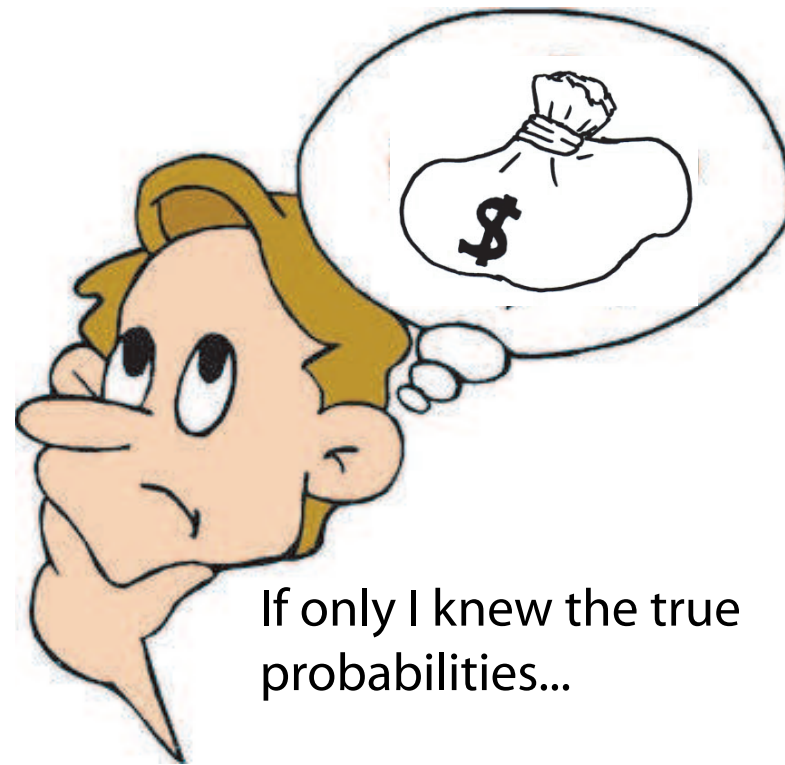
(Thomas-Agnan 1996).

In  $[0, 1]^K$  or  $\mathbb{R}^K$ : tensor products (also popular: thin-plate splines).

Moving between kernels and norms ( $\approx$  inner products): non-trivial. Kernels: used in algorithms; norms: in stating their properties.

## Research programme II (decision making)

- Choose a goal that could be achieved if you knew the true probabilities generating the observations.
- Construct a decision strategy provably achieving your goal.
- Isolate a continuous law of probability on which the proof depends.
- Use defensive forecasting to get rid of the true probabilities.



If only I knew the true probabilities...

The goal: (1) in terms of observables; (2) achievable regardless of what the true probabilities are.

The goal has to be **relative**.

Fix a choice function  $G : [0, 1] \rightarrow \Gamma$ :

$$G(p) \in \arg \min_{\gamma \in \Gamma} \lambda(p, \gamma),$$

where

$$\lambda(p, \gamma) := p\lambda(1, \gamma) + (1 - p)\lambda(0, \gamma).$$

For the square and log loss functions one can take  $G(p) := p$ .

The exposure of  $G$ :

$$\text{Exp}_G(p) := \lambda(1, G(p)) - \lambda(0, G(p))$$

(assumed continuous; a modification of this definition also works for the absolute loss function).

The exposure of a decision rule  $D : \mathbf{X} \rightarrow \Gamma$ :

$$\text{Exp}_D(x) := \lambda(1, D(x)) - \lambda(0, D(x)).$$

**Informal corollary** The decisions  $\gamma_n := G(p_n)$  (“ELM principle”), with  $p_n$  output by ALN with a Sobolev kernel, satisfy

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

for all  $N$  and all decision rules  $D$  with a small  $\|\text{Exp}_D\|_{\mathcal{S}}$ .

Proof Subtracting

$$\lambda(p, \gamma) = p\lambda(1, \gamma) + (1 - p)\lambda(0, \gamma)$$

from

$$\lambda(y, \gamma) = y\lambda(1, \gamma) + (1 - y)\lambda(0, \gamma)$$

gives

$$\lambda(y, \gamma) - \lambda(p, \gamma) = (y - p)(\lambda(1, \gamma) - \lambda(0, \gamma)).$$

In conjunction with Theorem 3:

$$\begin{aligned}\sum_{n=1}^N \lambda(y_n, \gamma_n) &= \sum_{n=1}^N \lambda(y_n, G(p_n)) \\ &= \sum_{n=1}^N \lambda(p_n, G(p_n)) + \sum_{n=1}^N \left( \lambda(y_n, G(p_n)) - \lambda(p_n, G(p_n)) \right) \\ &= \sum_{n=1}^N \lambda(p_n, G(p_n)) + \sum_{n=1}^N (y_n - p_n) \left( \lambda(1, G(p_n)) - \lambda(0, G(p_n)) \right) \\ &\approx \sum_{n=1}^N \lambda(p_n, G(p_n))\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{n=1}^N \lambda(p_n, D(x_n)) \\
&= \sum_{n=1}^N \lambda(y_n, D(x_n)) - \sum_{n=1}^N (\lambda(y_n, D(x_n)) - \lambda(p_n, D(x_n))) \\
&= \sum_{n=1}^N \lambda(y_n, D(x_n)) - \sum_{n=1}^N (y_n - p_n) (\lambda(1, D(x_n)) - \lambda(0, D(x_n))) \\
&\qquad\qquad\qquad \approx \sum_{n=1}^N \lambda(y_n, D(x_n)).
\end{aligned}$$

**Summary of the proof technique:** to show that the actual loss of our decision strategy does not exceed the actual loss of a decision rule  $D$  by much, we notice that

- the actual loss  $\sum_{n=1}^N \lambda(y_n, G(p_n))$  of our decision strategy is approximately equal, by Theorem 3, to the (one-step-ahead conditional) expected loss  $\sum_{n=1}^N \lambda(p_n, G(p_n))$  of our strategy;
- since we used the Empirical Loss Minimization principle, the expected loss of our strategy does not exceed the expected loss of  $D$ ;
- the expected loss of  $D$  is approximately equal to its actual loss (again by Theorem 3).

**Theorem 4** (special cases) Suppose  $\lambda(y, \gamma) = (y - \gamma)^2$ .

Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + \frac{3}{8} (\|2D - 1\|_{\mathcal{S}'} + 1) \sqrt{N}$$

for all  $N$  and  $D$ .

Suppose  $\lambda(y, \gamma) = |y - \gamma|$ . Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + \frac{\sqrt{6}}{4} (\|2D - 1\|_{\mathcal{S}'} + 1) \sqrt{N}$$

for all  $N$  and  $D$ .

Suppose

$$\lambda(y, \gamma) = -y \ln \gamma - (1 - y) \ln(1 - \gamma).$$

Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + 0.7 \left( \left\| \ln \frac{D}{1-D} \right\|_{\mathcal{S}'} + 1 \right) \sqrt{N}$$

for all  $N$  and  $D$ .

General theorem: functions with a “convex loss” (if unbounded, the tails must decay faster than  $1/t$ ; in the log loss game, they decay exponentially fast).

Can be extended to non-convex loss functions and loss functions depending on several future observations (work in progress).

## Further details

### Game-theoretic probability:

Glenn Shafer and Vladimir Vovk, [Probability and finance: it's only a game](#), New York: Wiley, 2001

### Defensive forecasting:

<http://www.probabilityandfinance.com>, Working Papers 10, 13, 14.