Defensive forecasting and competitive on-line prediction

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

Tokyo, 17 March 2006
My plan:

- Game-theoretic vs. measure-theoretic probability (the difference demonstrated on SLLN)
- **Defensive forecasting**: game-theoretic laws of probability \(\rightarrow\) forecasting algorithms
- Implementation (result only): WLLN \(\rightarrow\) K29
- K29 in function spaces
- Properties of K29: calibration and resolution
- Use for decision making
Glenn’s talk: there are 2 main ways to formalize probability, measure (Borel / ··· / Kolmogorov) vs. gambling (von Mises / Ville / Kolmogorov).

To see the difference (important in defensive forecasting), consider the simplest martingale SLLN. Let $y_1, y_2, \ldots$ be random variables s.t. $y_n \in \{0, 1\}$ for all $n$; let $p_n$ be the conditional probability that $y_n = 1$. Then

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

with probability 1.
Game-theoretic SLLN for binary observations

Forecasting protocol:

$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots$:
   Reality announces $x_n \in X$.
   Forecaster announces $p_n \in [0, 1]$.
   Skeptic 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.

$\mathcal{K}_n$: Skeptic’s capital.

$x_n$: datum (all relevant information, may include some of the previous $y_i$); $y_n$: observation.
Proposition (game-theoretic SLLN) Skeptic has a strategy which guarantees that

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

- either

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

($p_n$ are unbiased) or

$$\lim_{n \to \infty} \mathcal{K}_n = \infty.$$
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 $\mathcal{K}_n \to \infty$ with probability 0.

Game-theoretic SLLN:

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

Caveat: I assumed that Skeptic’s strategy was measurable. Empirical fact: for all kinds of limit theorems, Skeptic’s strategy we construct is measurable; moreover, it is continuous.
Recent (2004) observation: this approach can be used for designing forecasting algorithms.

For any continuous strategy for Skeptic there exists a strategy for Forecaster that does not allow Skeptic's capital to grow.
The difficulty with forecasting

There is no forecasting algorithm that “works” for every sequence. Dawid’s (1985) example:

\[ y_n := \begin{cases} 
1 & \text{if } p_n < 1/2 \\
0 & \text{otherwise.} 
\end{cases} \]

This sequence looks computable and so can be predicted perfectly. But the algorithm producing \( p_n \) is always wrong!
Two very natural “cheats”:

**Continuity**: consider only continuous strategies for Skeptic. Goes back to Kolmogorov’s school of the foundations of probability (Levin 1976).

**Randomness**: allow Forecaster to use randomization (Foster & Vohra 1998 and many followers).

We are using the first cheat. **Brouwer’s principle**: computable functions are always continuous.
Modified protocol:

\[ K_0 := 1. \]

FOR \( n = 1, 2, \ldots \):

- Reality announces \( x_n \in X \).
- Skeptic announces continuous \( S_n : [0, 1] \to \mathbb{R} \).
- Forecaster announces \( p_n \in [0, 1] \).
- Reality announces \( y_n \in \{0, 1\} \).

\[ K_n := K_{n-1} + S_n(p_n)(y_n - p_n). \]

END FOR.

Theorem 1 (Takemura)  Forecaster has a strategy that ensures \( K_0 \geq K_1 \geq K_2 \cdots \).
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

Can be easily generalized; Intermediate Value Theorem \( \mapsto \) numerous fixed point and minimax theorems in topological vector spaces.
Research program I (forecasting)

• 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.

• If necessary, streamline the resulting forecasting algorithm.
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 \( \Phi : [0, 1] \times \mathbf{X} \to H \) (feature mapping to an inner product space) and

\[
c_{\Phi} := \sup_{p, x} \| \Phi(p, x) \| < \infty.
\]

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

\[
\left\| \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n) \Phi(p_n, x_n) \right\| \leq \frac{c_{\Phi}}{\sqrt{N\delta}}
\]

with probability at least 1 – \( \delta \). An easy modification of the standard statement (\( \Phi \equiv 1 \), Kolmogorov 1929). True both measure-theoretically (with \( \Phi \) measurable) and game-theoretically.
Let
\[ k((p, x), (p', x')) = \langle \Phi(p, x), \Phi(p', x') \rangle \]
(the kernel). Suppose \( k \) is continuous in \( p \). Applying Theorem 1 to Kolmogorov's proof: there exists a forecasting strategy (the K29 algorithm with parameter \( k \)) that guarantees

\[
\forall N : \left\| \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n) \Phi(p_n, x_n) \right\| \leq \frac{c \Phi}{\sqrt{N}}
\]

(somewhat better than when using the true probabilities, esp. in view of the LIL).
Problem with Research Program I in the binary case: works too well. Already in response to WLLN, Theorem 1 produces predictions that satisfy most other laws. Might be interesting for unbounded $y_n$ (connections with empirical processes).
The K29 algorithm with parameter \( k \)

FOR \( n = 1, 2, \ldots : \)
  Read \( x_n \in X \).
  Set \( S_n(p) := \sum_{i=1}^{n-1} k((p, x_n), (p_i, x_i))(y_i - p_i) \) for \( p \in [0, 1] \).
  Output any root \( p \) of \( S_n(p) = 0 \) as \( p_n \);
    if there are no roots, \( p_n := (1 + \text{sign } S_n)/2 \).
  Read \( y_n \in \{0, 1\} \).
END FOR.

Since \( S_n \) is continuous, \( \text{sign } S_n \) is well defined in this context.

Intuition: \( p_n \) is chosen so that \( p_i \) are unbiased forecasts for \( y_i \) on the rounds \( i = 1, \ldots, n - 1 \) for which \((p_i, x_i)\) is similar to \((p_n, x_n)\).
A reproducing kernel Hilbert space (RKHS) on $Z$ (such as $X$ or $[0, 1] \times X$) is a Hilbert space $\mathcal{F}$ of real-valued functions on $Z$ such that the evaluation functional $f \in \mathcal{F} \mapsto f(z)$ is continuous for each $z \in Z$. By the Riesz–Fischer theorem, for each $z \in Z$ there exists a function $k_z \in \mathcal{F}$ such that

$$f(z) = \langle k_z, f \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}.$$ 

Let

$$c_\mathcal{F} := \sup_{z \in Z} \|k_z\|_{\mathcal{F}};$$

we will be interested in the case $c_\mathcal{F} < \infty$. 
The corresponding kernel:

\[ k(z, z') := \langle k_z, k_{z'} \rangle_F; \]

c can be equivalently defined as \( \sup_z k(z, z) \). The K29 property stated earlier implies (when applied to \( \Phi(p, x) := k_{p,x} \)):

**Theorem 2** Let \( F \) be a RKHS on \([0, 1] \times X\). K29 with the kernel \( k \) ensures

\[
\left| \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n) f(p_n, x_n) \right| \leq \frac{c_F \|f\|_F}{\sqrt{N}}
\]

for all \( N \) and \( f \).
Examples

A “Sobolev norm” $\|f\|_S$ of $f : [0, 1] \to \mathbb{R}$ is defined by

$$\|f\|_S^2 := \left( \int_0^1 f(t) \, dt \right)^2 + \int_0^1 (f'(t))^2 \, dt$$

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

Its kernel is

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

(Craven and Wahba 1979); so $c_S = 4/3$. 
For functions on $\mathbb{R}$:

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

with kernel

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

(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.
Calibration and resolution (informal discussion)

The forecasts $p_n$, $n = 1, \ldots, N$, are well calibrated if, for any $p^* \in [0, 1]$, 

$$\frac{\sum_{n=1,\ldots,N} p_n \approx p^* y_n}{\sum_{n=1,\ldots,N} p_n \approx p^* 1} \approx p^*$$

provided $\sum_{n=1,\ldots,N} p_n \approx p^* 1$ is not too small.

Can be rewritten as 

$$\frac{\sum_{n=1,\ldots,N} p_n \approx p^* (y_n - p_n)}{\sum_{n=1,\ldots,N} p_n \approx p^* 1} \approx 0.$$
The forecasts $p_n$, $n = 1, \ldots, N$, have good resolution if, for any $x^* \in X$,

$$\frac{\sum_{n=1, \ldots, N: x_n \approx x^*} (y_n - p_n)}{\sum_{n=1, \ldots, N: x_n \approx x^*} 1} \approx 0$$

provided the denominator is not too small.

The forecasts $p_n$, $n = 1, \ldots, N$, have good calibration-cum-resolution if, for any $(p^*, x^*) \in [0, 1] \times X$,

$$\frac{\sum_{n=1, \ldots, N: (p_n, x_n) \approx (p^*, x^*)} (y_n - p_n)}{\sum_{n=1, \ldots, N: (p_n, x_n) \approx (p^*, x^*)} 1} \approx 0$$

provided the denominator is not too small.
For concreteness: \textit{calibration}.

To make sense of the $\approx$, consider a “soft neighborhood” $f \in S$ of $p^*$: $f(p^*) = 1$ and $f(p) = 0$ unless $p$ is close to $p^*$.

The K29 forecasts will be well calibrated,
\[
\frac{\sum_{n=1,\ldots,N} f(p_n)(y_n - p_n)}{\sum_{n=1,\ldots,N} f(p_n)} \approx 0,
\]
if $\|f\|_S$ is not large and
\[
\sum_{n=1}^{N} f(p_n) \gg \sqrt{N}.
\]
Competitive on-line prediction: we are given a pool of decision strategies and our goal is to perform almost as well as the best strategy in the pool. No assumptions about the reality.

Defensive forecasting → a new proof technique in competitive on-line prediction.

This talk: prediction ← forecasting or decision making.
Decision-making protocol:

\[
\text{Loss}_0 := 0.
\]
\[
\text{FOR } n = 1, 2, \ldots :
\]
\[
\text{Reality announces } x_n \in X.
\]
\[
\text{Decision Maker announces } \gamma_n \in \Gamma.
\]
\[
\text{Reality announces } y_n \in \{0, 1\}.
\]
\[
\text{Loss}_n := \text{Loss}_{n-1} + \lambda(y_n, \gamma_n).
\]
\[
\text{END FOR.}
\]

\(\lambda\): the loss function.
The difference between the two protocols

- In the forecasting protocol, our goal to produce probabilistic statements (in principle, they can be falsified: turn out to be false).

- In the decision-making protocol, we are merely minimizing our loss.
Decision rule $D : X \to \Gamma$.

We want to compete against decision rules that are not too irregular with no assumptions about Reality. Let $X = [0, 1]$ at first. Irregularity is measured with the Sobolev norm.
Proposition  Suppose $X = \Gamma = [0, 1]$ and $\lambda(y, \gamma) = |y - \gamma|$. Decision Maker has a strategy that guarantees

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

for all $N$ and $D$. 
When is Decision Maker competitive with $D$? Let

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

(“symmetrized” $D$).

We have

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

OK if the mean slope $\ll \sqrt{N}$. Especially simple case: continuous piece-wise linear functions (dense in $C([0, 1])$).
No upper bound on $\|f\|_S$, so we have universal consistency: for any continuous prediction rule $D$, 

$$\limsup_{N \to \infty} \left( \frac{1}{N} \sum_{n=1}^{N} \lambda(y_n, \gamma_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(y_n, D(x_n)) \right) \leq 0.$$ 

This is a minimal property.
Research program 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.
The goal should be:

1. in terms of observables;

2. achievable regardless of what the true probabilities are.

The goal has to be relative.
Fix a choice function $G : [0, 1] \to \Gamma$:

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

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 $D : X \to \Gamma$:

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

Informal statement  Suppose $\|\text{Exp}_G\|_S$ is not large. The decisions $\gamma_n := G(p_n)$ ("ELM principle"), with $p_n$ output by ALN with a Sobolev kernel, satisfy

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

for all $N$ and all decision rules $D$ with $\|\text{Exp}_D\|_S$ not too large.
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 2:

\[
\sum_{n=1}^{N} \lambda(y_n, \gamma_n) = \sum_{n=1}^{N} \lambda(y_n, G(p_n))
\]

\[
= \sum_{n=1}^{N} \lambda(p_n, G(p_n)) + \sum_{n=1}^{N} \left( \lambda(y_n, G(p_n)) - \lambda(p_n, G(p_n)) \right)
\]

\[
= \sum_{n=1}^{N} \lambda(p_n, G(p_n)) + \sum_{n=1}^{N} (y_n - p_n) \left( \lambda(1, G(p_n)) - \lambda(0, G(p_n)) \right)
\]

\[\approx \sum_{n=1}^{N} \lambda(p_n, G(p_n)) \]
\[ \leq \sum_{n=1}^{N} \lambda(p_n, D(x_n)) \]

\[ = \sum_{n=1}^{N} \lambda(y_n, D(x_n)) - \sum_{n=1}^{N} \left( \lambda(y_n, D(x_n)) - \lambda(p_n, D(x_n)) \right) \]

\[ = \sum_{n=1}^{N} \lambda(y_n, D(x_n)) - \sum_{n=1}^{N} (y_n - p_n) \left( \lambda(1, D(x_n)) - \lambda(0, D(x_n)) \right) \]

\[ \approx \sum_{n=1}^{N} \lambda(y_n, D(x_n)). \]
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 2, 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 Expected 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 2).
Theorem 3 (special cases: specific loss functions and the Sobolev space $S'$ on $\mathbb{R}$) Let $\Gamma = [0, 1]$ and $X = \mathbb{R}$. Suppose $\lambda(y, \gamma) = (y - \gamma)^2$. Decision Maker has a strategy that guarantees

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

for all $N$ and $D$.

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

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

for all $N$ and $D$. 

40
Suppose
\[ \lambda(y, \gamma) = -y \ln \gamma - (1 - y) \ln(1 - \gamma). \]
Decision Maker has a strategy that guarantees
\[ \sum_{n=1}^{N} \lambda(y_n, \gamma_n) \leq \sum_{n=1}^{N} \lambda(y_n, D(x_n)) + 0.7 \left( \left\| \ln \frac{D}{1 - D} \right\|_{S'} + 1 \right) \sqrt{N} \]
for all \( N \) and \( D \).

General theorem: any RKHS in pace of \( S' \); convex loss functions (if unbounded, the tails must decay faster than \( 1/t \); in the log loss game, they decay exponentially fast).

Natural developments: extend to non-convex loss functions (with a little randomization) and loss functions depending on several future observations.
Limitations of defensive forecasting

Competitive on-line prediction: its goal implicitly assumes a small decision maker.

Remember a typical guarantee:

\[
\sum_{n=1}^{N} \lambda(y_n, \gamma_n) \leq \sum_{n=1}^{N} \lambda(y_n, D(x_n)) + (\|2D - 1\|_S + 1) \sqrt{N}.
\]
Ideal probability forecasts (actual) are not enough in big decision making!

Simple example: $\Gamma = \{0, 1\}$, $\lambda$ is given by the matrix

\[
\begin{array}{c|cc}
\text{Reality} & 0 & 1 \\
\hline
0 & 1 & 2 \\
1 & 2 & 0 \\
\end{array}
\]

Reality’s strategy: $y_n := \gamma_n$. Decision Maker’s theory: Reality always chooses $y_n = 0$. 
Decision Maker’s mistake: he was being greedy (concentrated on exploitation and completely neglected exploration). But:

- he acted optimally given his beliefs,
- his beliefs have been verified by what actually happened.

We have to worry about what would have happened if we had acted in a different way.
My hope: game-theoretic probability has an important role to play in big decision making as well. A standard picture in the philosophy of science (Popper, Kuhn, Lakatos, . . . ): science progresses via struggle between (probabilistic) theories. It is possible that something like this happens in individual (human and animal) learning as well. **Testing** of probabilistic theories is crucial. The game-theoretic version of Cournot's principle: more flexible; at each time we know to what degree a theory has been falsified.
Small decision making is important; two popular examples in learning theory: prediction (evaluated with a loss function) and portfolio selection.

Big decision making: might be even more important in practice, but also might be mathematically less elegant (cf. PDE).
Related literature


Randomization approach to calibration: Foster and Vohra (1998); Fudenberg, Levine, Lehrer, Sandroni, Smorodinsky,... (Asymptotic results.)

Continuity approach rediscovered by Kakade and Foster (2004). (Asymptotic results.)

Hannan 1957: the beginning of competitive on-line prediction.

Littlestone, Warmuth, Vovk, Cesa-Bianchi, Freund, Schapire,... (from 1989): “prediction with expert advice”, with numerous applications to competitive on-line prediction.
Further details

Game-theoretic probability:


Defensive forecasting: