

Defensive forecasting for decision making

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

Santander, 6 July 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: a new proof technique in prediction with expert advice.

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

Decision-making protocol:

Loss₀ := 0.

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

 Reality announces $x_n \in [0, 1]^K$.

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

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

 Loss _{n} := Loss _{$n-1$} + $\lambda(y_n, \gamma_n)$.

END FOR.

x_n : object; K : the number of attributes; y_n : label; λ : the loss function.

Decision rule $D : [0, 1]^K \rightarrow [0, 1]$.

We want to compete against decision rules that are not too weird with no assumptions about Reality. Let $K = 1$ at first.

The **Fermi–Sobolev norm** $\|f\|$ of $f : [0, 1] \rightarrow \mathbb{R}$ is defined by

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

Assume, for simplicity, that f is **Lipschitzian**:

$$\sup_{t_1 \neq t_2} \frac{|f(t_1) - f(t_2)|}{|t_1 - t_2|} < \infty.$$

Proposition Suppose $K = 1$ and $\lambda(y, \gamma) = |y - \gamma|$. Decision Maker has a strategy that guarantees

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

for all N and all Lipschitzian D .

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

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

(“symmetrized” D).

We have

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

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

Later: $K > 1$, other loss functions.

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 **K29**
- Theoretical properties of K29: calibration and resolution; Fourier kernel
- Use for decision making

There are 2 main ways to formalize probability: **measure** (Borel / Radon / Fréchet / 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 $p_n \in [0, 1]$.

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.

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) Skeptic 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 Skeptic (this is what makes defensive forecasting possible!)

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

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

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

Modified protocol:

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

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

Reality announces $x_n \in [0, 1]^K.$

Skeptic 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

Research program 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.

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 $\Phi : [0, 1] \times [0, 1]^K \rightarrow H$ (feature mapping) and

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

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

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

with probability at least $1 - \delta$. An easy modification of the standard statement ($\Phi \equiv 1$, Kolmogorov 1929). True both measure-theoretically (with Φ measurable) and game-theoretically.

Let

$$\mathbf{k}((p, x), (p', x')) = \Phi(p, x) \cdot \Phi(p', x').$$

(Remember the “kernel trick”.) Suppose \mathbf{k} is continuous in p . Applying Theorem 1 to Kolmogorov’s proof: there exists a forecasting strategy (the [K29 algorithm with parameter \$\mathbf{k}\$](#)) that guarantees

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

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

If the “surface” Φ is sufficiently convoluted: “calibration” and “resolution”; e.g., think of the Gaussian kernel: distant points are virtually orthogonal.

We can take $H = \mathbb{R}^\infty$ with the dot product

$$u \cdot v := \sum_{m \in M} a_m u_m v_m$$

for fixed $a_m > 0$, $\sum_m a_m = 1$, and

$$\Phi(p, x) := (f_m)_{m \in M},$$

where

$$f_{m_0, m_1, \dots, m_K}(p, x_1, \dots, x_K) := \cos \pi m_0 p \cos \pi m_1 x_1 \cdots \cos \pi m_K x_K$$

is the Fourier basis.

For suitable a_m and the corresponding kernel:

Theorem 2 K29 with this kernel ensures

$$\left| \frac{1}{N} \sum_{n=1}^N (y_n - p_n) f(p_n, x_n) \right| \leq \left(\frac{2}{\sqrt{3}} \right)^{K+1} \frac{\|f\|_{\text{FS}}}{\sqrt{N}}$$

for all N and all Lipschitzian f .

Definition: The FS norm of $f(t_0, t_1, \dots, t_K)$ is defined by

$$\|f\|_{\text{FS}}^2 := \sum_{\{i_1, \dots, i_k\} \subseteq \{0, 1, \dots, K\}} S^2 \left(\int_0^1 \cdots \int_0^1 f(t_0, \dots, t_{K+1}) dt_{i_1} \cdots dt_{i_k} \right),$$

where the seminorm $S(g)$ of $g(s_1, \dots, s_k)$ is defined by

$$S^2(g) := \int_0^1 \cdots \int_0^1 \left(\frac{\partial^k g(s_1, \dots, s_k)}{\partial s_1 \cdots \partial s_k} \right)^2 ds_1 \cdots ds_k.$$

Research program II (decision making)

- Supposing you know the true probabilities generating the observations, construct a decision strategy with desirable properties.
- Isolate a continuous law of probability implying those desirable properties.
- Use defensive forecasting to get rid of the true probabilities; this can be done as defensive forecasts are as good as the true probabilities, as far as the given law of probability is concerned.

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 **Lipschitzian**; a modification of this definition also works for the absolute loss function).

The **exposure** of a decision rule $D : [0, 1]^K \rightarrow \Gamma$:

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

(assumed Lipschitzian).

Informal corollary The decisions $\gamma_n := G(p_n)$ (“ELM principle”), with p_n output by K29 with the Fourier 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 $\|\text{Exp}_D\|_{FS}$.

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:

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

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

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 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 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 2).

Theorem 3 Suppose $\lambda(y, \gamma) = |y - \gamma|$ or $\lambda(y, \gamma) = (y - \gamma)^2$.

Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + \frac{1}{\sqrt{2}} \left(\left(\frac{2}{\sqrt{3}} \right)^K \|2D - 1\|_{\text{FS}} + 1 \right) \sqrt{N}$$

for all N and all Lipschitzian D .

Suppose

$$\lambda(y, \gamma) = \begin{cases} -\ln \gamma & \text{if } y = 1 \\ -\ln(1 - \gamma) & \text{if } y = 0. \end{cases}$$

Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + 0.73 \left(\left(\frac{2}{\sqrt{3}} \right)^K \left\| \ln \frac{1 - D}{D} \right\|_{\text{FS}} + 1 \right) \sqrt{N}$$

for all N and all Lipschitzian D .

Can be extended to essentially [any](#) loss function.

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 8, 10, 13, 14.

[Previous work](#): Foster and Vohra (1998); Fudenberg, Levine, Lehrer, Sandroni, Smorodinsky, Kakade, . . .