

## A third way in probability forecasting

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

Joint work with [Akimichi Takemura](#) (University of Tokyo) and  
[Glenn Shafer](#) (Rutgers University)

EURANDOM, 8 October 2004

## Plan for this talk:

- Game-theoretic probability vs. Kolmogorov's axioms (main example: game-theoretic vs. measure-theoretic SLLN)
- **Defensive forecasting**: theorems about probability  $\mapsto$  forecasting algorithms
- Implementation: LLN  $\mapsto$  **K29**
- Theoretical properties of K29
- Experimental results

There are 2 ways (at least) to do probability, both very old:  
measure (Kolmogorov) vs. gambling (von Mises / Ville).

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

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - p_i) = 0$$

with probability 1.

## Game-theoretic SLLN for uniformly bounded observations

Protocol:

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

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

Forecaster announces  $p_n \in \mathbb{R}$ .

Skeptic announces  $S_n \in \mathbb{R}$ .

Reality announces  $y_n \in [A, B]$ .

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

END FOR.

(Similar to games in de Finetti's foundations of Bayesian statistics.)  $\mathcal{K}_n$ : Skeptic's capital.

I will be mainly interested in the case:  $y_n \in \{0, 1\}$  (binary probability forecasting).

Theorem (game-theoretic SLLN) Skeptic has a strategy which guarantees that

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

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - p_i) = 0$$

( $p_i$  are unbiased) or

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

Sometimes it is difficult to decide which term of the disjunction will be true! E.g., the prices  $p_n$  in Iowa Electronic Markets (<http://www.biz.uiowa.edu/iem>) either allow us to become infinitely rich or are unbiased. Which?

General definition: an event  $E$  is **almost certain** if Skeptic has a strategy that does not risk bankruptcy and makes him infinitely rich if  $E$  fails to happen.

Or: Skeptic **can force**  $E$ .

I will almost prove this simple SLLN.

Usual tricks:

- if  $E_1, E_2, \dots$  are almost certain,  $\cap E_i$  is also almost certain [combine the corresponding strategies]
- we can replace  $\mathcal{K}_n \rightarrow \infty$  with  $\sup_n \mathcal{K}_n = \infty$  [occasionally set aside half of your capital  $\mathcal{K}_n$ ]: “weakly forcing”
- suppose  $p_n = 0$  for all  $n$

**Lemma** Suppose  $\epsilon > 0$ . Then Skeptic can “weakly force”

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n y_i \leq \epsilon.$$

The same argument, with  $-\epsilon$  in place of  $\epsilon$ :

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n y_i \geq -\epsilon \quad \text{a.s.}$$

Combine these for all  $\epsilon$ .

**Proof of the lemma** Skeptic always buys  $\epsilon\mathcal{K}_{n-1}$  at trial  $n$ ; then

$$\mathcal{K}_n = \prod_{i=1}^n (1 + \epsilon y_i).$$

On the paths where  $\mathcal{K}_n$  is bounded:

$$\prod_{i=1}^n (1 + \epsilon y_i) \leq D$$

$$\sum_{i=1}^n \ln(1 + \epsilon y_i) \leq \ln D;$$

since  $\ln(1 + t) \geq t - t^2$  whenever  $t \geq -\frac{1}{2}$ ,

$$\epsilon \sum_{i=1}^n y_i - \epsilon^2 \sum_{i=1}^n y_i^2 \leq \ln D$$

$$\epsilon \sum_{i=1}^n y_i - \epsilon^2 n \leq \ln D$$

$$\epsilon \sum_{i=1}^n y_i \leq \epsilon^2 n + \ln D$$

$$\frac{1}{n} \sum_{i=1}^n y_i \leq \epsilon + \frac{\ln D}{\epsilon n}.$$

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 (and 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 and even efficiently computable.

## Game-theoretic philosophy of probability

“The probability of  $E$  is  $p$ .” What does this mean? How do we falsify (or verify) this statement? Popular answers: frequentist; Bayesian.

Game-theoretic answer: this does not mean much and cannot be falsified (unless  $p$  is extreme). Probabilities have meaning *en masse* (e.g.: probabilistic theory; probability forecasts made during the last year). When somebody announces 1000 probabilities, our interpretation of what he means is: “I do not expect you to become rich gambling at the odds given by my probabilities”. His claim is falsified if we do get rich this way.

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  $x_n \in \mathbf{X}.$

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.

**Main Theorem** Forecaster has a strategy that ensures  
 $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \dots$

Notice: we do not use the requirement  $\mathcal{K}_n \geq 0, \forall n.$

## 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),

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

## Three approaches to probability forecasting

The **traditional** approach: Reality-centered (statistics, PAC theory, “conformal prediction”). Some assumptions about Reality’s randomized strategy are made (such as i.i.d.; often more). The other two approaches: no assumptions about Reality whatsoever.

“**Prediction with expert advice**”, in a wide sense: Forecaster-centered (Peter Grünwald’s talk). The starting point is a pool of forecasting strategies; they can be merged into one strategy. May have originated with Solomonoff in 1960 (merging all computable strategies).

Defensive forecasting: Skeptic-centered.

- 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 Main Theorem.

What does it give in the case of SLLN?

In fact, nothing interesting: Forecaster performs his task **too well**.

In response to Skeptic's strategy forcing

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - p_i) \leq \epsilon$$

he will always choose  $p_n := 1$ .

In response to Skeptic's strategy forcing

$$-\epsilon \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - p_i) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - p_i) \leq \epsilon$$

he will choose

$$p_n := \begin{cases} 1 & \text{if } \sum_{i=1}^{n-1} (y_i - p_i) > 0 \\ 0 & \text{if } \sum_{i=1}^{n-1} (y_i - p_i) < 0 \end{cases}$$

(limit as  $\epsilon \rightarrow 0$ ). Therefore he ensures:

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

for all  $n$  (better than in the case of randomness).

We need a better LLN (“in the small”):  $p_n$  should be unbiased in the neighborhood of each  $p^* \in [0, 1]$ .

Take the strategy

$$S_n(p) := \epsilon I_{p^*}(p) \mathcal{K}_{n-1},$$

where  $I_{p^*}$  is a “smooth neighborhood” of  $p^*$ . The above calculation now gives

$$\frac{\sum_{i=1}^n I_{p^*}(p_i)(y_i - p_i)}{\sum_{i=1}^n I_{p^*}(p_i)} \leq \epsilon + \frac{\ln D}{\epsilon \sum_{i=1}^n I_{p^*}(p_i)}.$$

Mix these strategies for different  $p^*$ , and you will get Skeptic’s strategy demonstrating the LLN “in the small”.

## Simplification

Take:  $I_{p^*} =$  Gaussian bells  $N(p^*, \sigma)$ ; mix using the Lebesgue measure over  $p^*$ ;  $\epsilon \rightarrow 0$ . Then

$$S_n(p) = \sum_{i=1}^{n-1} K(p, p_i)(y_i - p_i),$$

where

$$K(p, p') := \exp\left(-\frac{(p - p')^2}{4\sigma^2}\right)$$

(Gaussian kernel with an unusual parameterization).

## K29 algorithm

We can take any kernels on  $[0, 1]^2$  and we can allow dependence on the objects.

Let  $K : ([0, 1] \times \mathbf{X})^2 \rightarrow \mathbb{R}$  be a Mercer kernel. After seeing the object  $x_n$  on round  $n$  Forecaster outputs an arbitrary root  $p = p_n$  of the equation  $S_n(p) = 0$ , where

$$S_n(p) = \sum_{i=1}^{n-1} K((p, x_n), (p_i, x_i))(y_i - p_i);$$

if this equation has no roots,

$$p_n := (\text{sign}(S_n) + 1)/2.$$

Problem: because of  $\epsilon \rightarrow 0$ , it is not clear what we proved.  
Easier to start from scratch.

By Mercer's theorem:

$$K((p, x), (p', x')) = \Phi(p, x) \cdot \Phi(p', x'),$$

where  $\Phi : \mathbf{X} \rightarrow H$  and  $H$  is a Hilbert space.

Let

$$C := \sup_{p, x} \|\Phi(p, x)\| < \infty$$

(often  $C = 1$ ).

**Theorem** The K29 algorithm guarantees

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

for each  $N$ .

If  $\Phi$  is sufficiently twisted: unbiasedness in the small.

Consider the “soft neighborhood”

$$I_{(p^*, x^*)}(p, x) := K((p^*, x^*), (p, x))$$

of  $(p^*, x^*) \in [0, 1] \times \mathbf{X}$ . Using the Schwarz inequality:

**Corollary** The K29 algorithm with parameter  $K \geq 0$  ensures

$$\left| \frac{\sum_{n=1}^N (y_n - p_n) I_{(p^*, x^*)}(p_n, x_n)}{\sum_{n=1}^N I_{(p^*, x^*)}(p_n, x_n)} \right| \leq \frac{C^2 \sqrt{N}}{\sum_{n=1}^N I_{(p^*, x^*)}(p_n, x_n)}$$

for each  $N$  and each  $(p^*, x^*) \in ([0, 1] \times \mathbf{X})$ .

Therefore: we can expect unbiasedness in the “soft neighborhood” of  $(p^*, x^*)$  when

$$\sum_{n=1}^N I_{(p^*, x^*)}(p_n, x_n) \gg \sqrt{N}.$$

What is the precise statement of the LLN in the small?

**Theorem (WLLN in the small)** In the game with finite horizon  $N$ ,

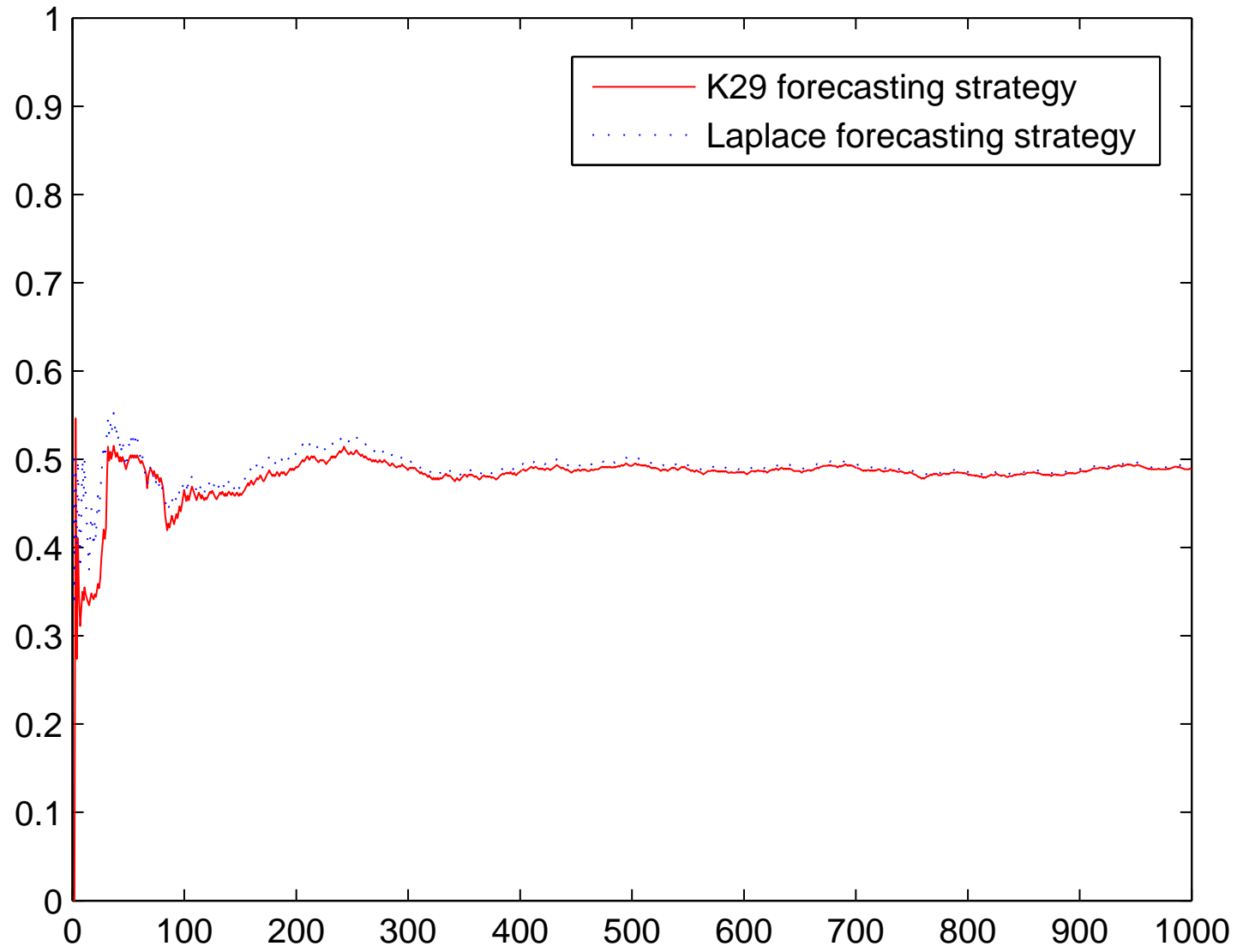
$$\underline{\mathbb{P}} \left\{ \left\| \frac{1}{N} \sum_{n=1}^N (y_n - p_n) \Phi(p_n, x_n) \right\| < \frac{C}{\sqrt{N\delta}} \right\} \geq 1 - \delta$$

$\underline{\mathbb{P}}$ : “game-theoretic lower probability”; but the theorem remains true (as usual) for measure-theoretic probability.

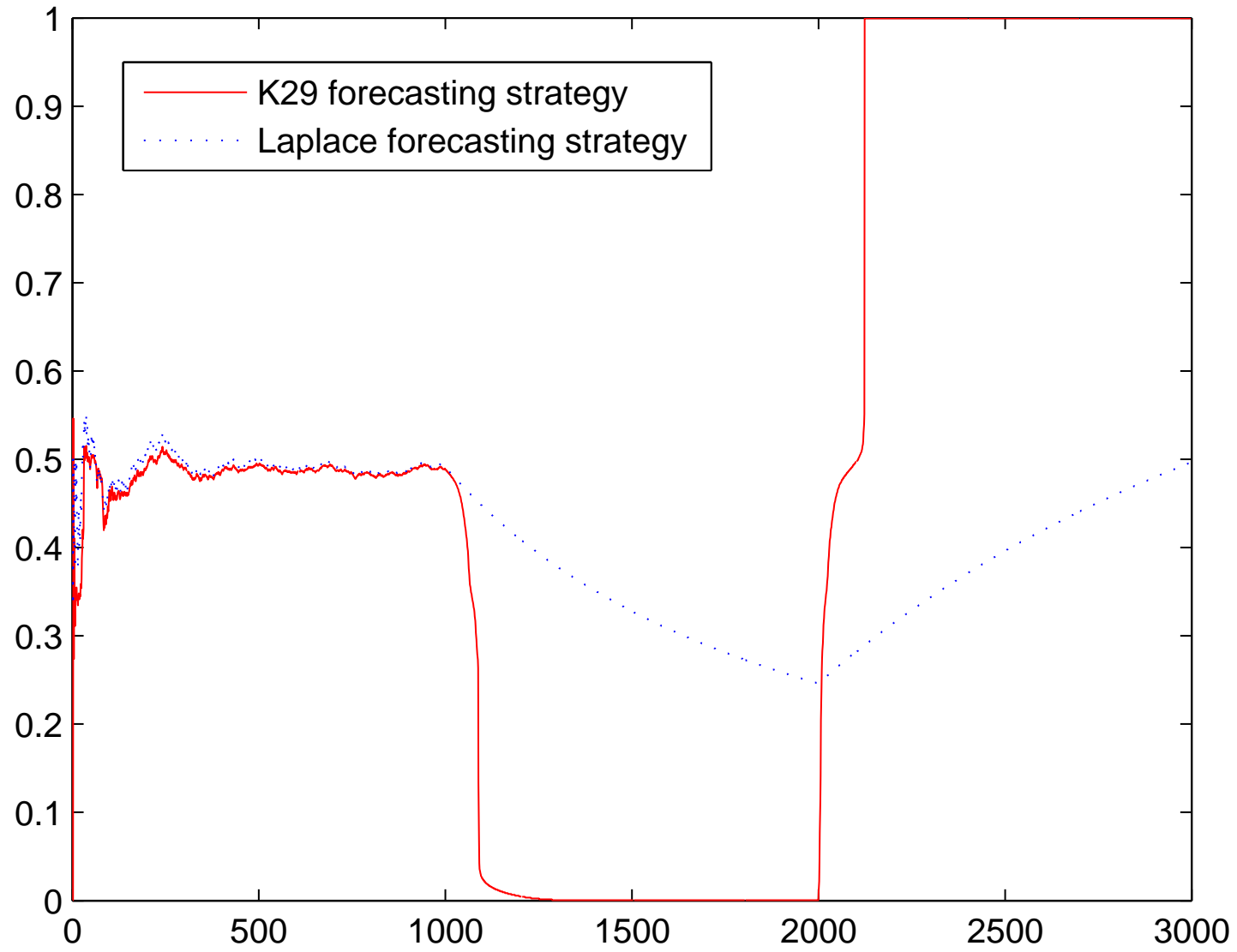
When  $\Phi \equiv 1$ , this was proven by Kolmogorov in 1929. K29 is Forecaster’s strategy obtained by Main Theorem from Skeptic’s strategy demonstrating WLLN in the small.

## Experimental results I: artificially generated data

1000 Bernoulli bits with parameter  $\theta = 0.5$ ; the K29 ( $\sigma = 0.01$ ) and Laplace ( $p_n := (k + 1)/(n + 1)$ ) forecasting strategies are almost indistinguishable:



These two forecasting strategies can behave very differently; the 1000 bits are complemented with 1000 0s followed by 1000 1s:



## Experimental results II: benchmark data set

TIC (or CoIL 2000) data set: 5822 training and 4000 test examples.

To compare K29 with the results in the literature: [off-line K29](#) (each test example is processed as if its was following the 5822 training examples). Kernel used: the tensor product

$$K((p, x), (p', x')) = \exp\left(-\frac{(p - p')^2}{4\sigma^2}\right) (x \cdot x')^d,$$

where  $d \in \{1, 2, \dots\}$ .

$\sigma = 0.1$  (the accuracy we are aiming for)

Zadrozny and Elkan, KDD'2002:

Method	Training MSE	Test MSE
NB	0.12845	0.13551
Sigmoid NB	0.10536	0.10905
PAV naive Bayes	0.10315	0.10818
SVM	0.11942	0.11889
Sigmoid SVM	0.11080	0.11122
PAV SVM	0.10974	0.11200

JL&BZ'04  
 0.10737  
 (best of 8;  
 "probing")

K29 run in the off-line fashion:

Degree	1	2	3	4	5
Test MSE	0.11033	0.10884	0.10752	0.10688	0.10668
Degree	6	7	8	9	10
Test MSE	0.10680	0.10715	0.10775	0.10862	0.10992

## 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 and 9

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

## Directions of further research

- Develop computationally efficient algorithms for the multi-label case. (Existence: an immediate corollary of the Poincaré–Hopf theorem. Regression: efficient. Mean/variance prediction: kind of efficient.)
- Defensive forecasting against:
  - Hoeffding’s inequality
  - central limit theorem
  - law of the iterated logarithm (and statements intermediate between SLLN and LIL)