

Non-asymptotic calibration and resolution

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

ALT'2005 (Singapore)
10 October 2005, 16:15

My plan for this talk:

- Game-theoretic vs. measure-theoretic probability (main example: game-theoretic vs. measure-theoretic SLLN)
- **Defensive forecasting** in the binary case: laws of probability
↳ forecasting algorithms
- Implementation: WLLN \mapsto **ALN** (=K29*)
- ALN in RKHS
- Theoretical properties of ALN: calibration and resolution

The following talk: 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. Reality I: x_n , Reality II: y_n .

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 observation: this approach can be used for designing forecasting 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 I announces $x_n \in \mathbf{X}.$

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

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

Reality II 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; the talk after next.

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

$$\left| \sum_{n=1}^N (y_n - p_n) \right| \leq 1/2$$

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

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

(the sup is finite and Φ is normalized).

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 Φ 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.)

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

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

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 (main) 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 and each $f \in \mathcal{F}$:

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

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

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

Example

If f is a function 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

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

(Craven and Wahba 1979).

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

Calibration and resolution (informal discussion)

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

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

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

Can be rewritten as

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

The forecasts p_n , $n = 1, \dots, N$, have **good resolution** if, for any $x^* \in \mathbf{X}$,

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

provided the denominator is not too small.

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

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

provided the denominator is not too small.

For concreteness: **calibration**.

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

The forecasts will be well calibrated,

$$\frac{\sum_{n=1, \dots, N} f(p_n)(y_n - p_n)}{\sum_{n=1, \dots, N} f(p_n)} \approx 0,$$

if $\|f\|_{\mathcal{S}}$ is not large and

$$\sum_{n=1}^N f(p_n) \gg \sqrt{N}.$$