

A dual approach to universal prediction

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

Dagstuhl, January 31, 2006

Solomonoff–Levin theory of universal prediction

- Take a very large class of prediction strategies (such as all computable strategies).
- Merge them into one strategy (summing with positive weights).

Can be generalized to a wide class of loss functions (Yura Kalnishkan's talk).

Decision protocol:

Loss₀ := 0.

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

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

 Predictor 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; y_n : label; (x_n, y_n) : example; λ : the loss function.

Fundamental loss function:

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

(**log-loss**: how well we can compress the sequence of observations).

Solomonoff: we can merge all computable strategies into one “weakly computable” strategy.

Levin: we can merge all “semicomputable” “semistrategies” into one object of the same kind (the universal semistrategy). This can also be done for many other loss functions.

This talk: an **indirect** approach.

Will be applied to large (but not huge) classes of prediction strategies.

Two very different protocols: **decision** vs. **gambling**.

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

We want to compete against decision rules that are not too wild with no assumptions about Reality.

Suppose $\mathbf{X} = [0, 1]$.

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

$$\|f\|^2 := \int_0^1 f(t)^2 dt + \int_0^1 (f'(t))^2 dt.$$

If $\mathbf{X} = [0, 1]^K$ for $K > 1$: use tensor product.

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

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

for all N and all Sobolev D .

Similar results are true for a wide range of function classes and loss functions.

No upper bound on $\|f\|$, so we have **universal consistency**: for any continuous decision rule D ,

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

This is a minimal property.

How to prove such results?

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

Gambling protocol:

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

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

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

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.

The difference between the two protocols

- In the gambling protocol, our goal to produce true probabilities (probabilities one cannot gamble against).
- In the decision 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

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

This is a typical game-theoretic law of probability.

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 this proof technique possible!)

Recent 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.

Proposition 2 (Takemura) Forecaster has a strategy that ensures $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \cdots$.

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

Research programme

- Choose a law (or laws) of probability you want Forecaster's moves to satisfy. (Several laws have to be merged into one law.)
- Prove the corresponding (continuous) game-theoretic result.
- Apply Proposition 2.
- Streamline and simplify the resulting strategy.

Once you have good probabilities, you can make good decisions minimizing the expected loss.

For example: LLN can be enforced.

Not very interesting: Forecaster 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).

A more useful result

Let \mathcal{F} be a RKHS (“reproducing kernel Hilbert space”) on $[0, 1] \times \mathbf{X}$ (such as: the tensor power of the Sobolev space for $\mathbf{X} = [0, 1]^K$).

Theorem The “K29 algorithm” ensures

$$\left| \frac{1}{N} \sum_{n=1}^N (y_n - p_n) f(p_n, x_n) \right| \leq \frac{c_{\mathcal{F}} \|f\|_{\mathcal{F}}}{2 \sqrt{N}}$$

for all N and all $f \in \mathcal{F}$.

$c_{\mathcal{F}}$: “size” of the RKHS; can be taken as $(1.15)^{K+1}$ for the tensor power of the Sobolev space.

If f is a “soft neighbourhood” of some p^* : **calibration**;
of some x^* : **resolution**.

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, 11, 13, 14, 16.

[Other work](#) in “defensive forecasting” (our name): Foster and Vohra (1998); Fudenberg, Levine, Lehrer, Sandroni, Smorodinsky, Kakade, . . .