

Leading strategies in competitive on-line prediction

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

Barcelona, 9 October 2006

Plan for this talk:

- Master strategies (traditional) and leading strategies (new) in a simple case
- Briefly: idea of proof of existence of leading strategies
- Non-asymptotic version
- Generalizations

On-line prediction protocol

The game of prediction:

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

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

Predictor announces $\mu_n \in \mathbf{P}$.

Reality announces $y_n \in \mathbf{Y}$.

END FOR.

x_n : side information (the data relevant for predicting y_n , perhaps including some of the previous y_{n-1}, y_{n-2}, \dots)

μ_n : prediction

y_n : observation

Predictor's usual goal in competitive on-line prediction

$\lambda(y_n, \mu_n)$: loss suffered when predicting y_n with μ_n

We want Predictor to be competitive with a “benchmark class” \mathcal{F} of prediction rules $F : \mathbf{X} \rightarrow \mathbf{P}$.

I.e.: we want Predictor to achieve

$$\frac{1}{N} \sum_{n=1}^N \lambda(y_n, \mu_n) \lesssim \frac{1}{N} \sum_{n=1}^N \lambda(y_n, F(x_n))$$

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

Name of the field

Competitive on-line prediction: emphasizes the fact that it is essentially a special case of competitive analysis of on-line algorithms.

Other people prefer: **universal prediction of individual sequences**.

Part of **prediction with expert advice** (the original setting: “free” experts).

Master strategies

Suppose $\mathbf{Y} = [-Y, Y]$, $\mathbf{P} = \mathbb{R}$ and $\lambda(y, \mu) = (y - \mu)^2$ (the quadratic loss function).

If \mathcal{F} is not too big, Predictor can guarantee

$$\frac{1}{N} \sum_{n=1}^N (y_n - \mu_n)^2 \lesssim \frac{1}{N} \sum_{n=1}^N (y_n - F(x_n))^2$$

for all $F \in \mathcal{F}$ and large N . Such a prediction strategy: **master strategy**.

Many ways to prove it, e.g.: Gradient Descent, Aggregating Algorithm, Defensive Forecasting, **The latter: gives more (the difference between the two sides).**

Leading strategies

If \mathcal{F} is not too big, Predictor can guarantee

$$\frac{1}{N} \sum_{n=1}^N (y_n - \mu_n)^2 \approx \frac{1}{N} \sum_{n=1}^N (y_n - F(x_n))^2 - \frac{1}{N} \sum_{n=1}^N (\mu_n - F(x_n))^2$$

for all $F \in \mathcal{F}$ and large N . Such a prediction strategy: [leading strategy](#).

The loss of any $F \in \mathcal{F}$ is determined by how closely it manages to imitate the leading strategy.

Simple asymptotic result

There is a strategy for Predictor that asymptotically dominates every continuous prediction rule:

Theorem Let \mathbf{X} be a metric compact, $\mathbf{Y} = [-Y, Y]$ and $\mathbf{P} = \mathbb{R}$. There exists a strategy for Predictor that guarantees

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^N (y_n - \mu_n)^2 \\ = \frac{1}{N} \sum_{n=1}^N (y_n - F(x_n))^2 - \frac{1}{N} \sum_{n=1}^N (\mu_n - F(x_n))^2 + o(1) \end{aligned}$$

as $N \rightarrow \infty$, for each continuous prediction rule F .

Idea of proof

$$\left| \sum_{n=1}^N (y_n - \mu_n)^2 - \sum_{n=1}^N (y_n - F(x_n))^2 + \sum_{n=1}^N (\mu_n - F(x_n))^2 \right|$$
$$= 2 \left| \sum_{n=1}^N (F(x_n) - \mu_n) (y_n - \mu_n) \right|$$

Defensive forecasting can produce predictions satisfying any “game-theoretic law of probability”.

The smallness of the last sum: a kind of the law of large numbers.

Hilbert function spaces

$C(\mathbf{X})$: the continuous functions on \mathbf{X} with the norm

$$\|F\|_{C(\mathbf{X})} := \sup_{x \in \mathbf{X}} |F(x)|.$$

This is a Banach space (\approx normed function space).

A Banach space is a **Hilbert space** if it satisfies the parallelogram identity

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

(equivalently: if the norm can be defined in terms of an inner product).

Embedding constant and a simple example

Suppose $\mathcal{F} \subseteq C(\mathbf{X})$ is a Hilbert space (with a different norm from that of $C(\mathbf{X})$). Define

$$\mathbf{c}_{\mathcal{F}} := \sup_{F: \|F\|_{\mathcal{F}} \leq 1} \|F\|_{C(\mathbf{X})}$$

(the **embedding constant**) and suppose $\mathbf{c}_{\mathcal{F}} < \infty$.

Let $\mathbf{X} = [0, 1]$. A **Sobolev space** \mathcal{F} : all absolutely continuous functions on $[0, 1]$ with finite

$$\|F\|_{\mathcal{F}} := \sqrt{\int_0^1 (F(x))^2 dx + \int_0^1 (F'(x))^2 dx}.$$

In this case (Marti, 1983):

$$\mathbf{c}_{\mathcal{F}} = \sqrt{\coth 1} \approx 1.15.$$

Non-asymptotic result for the quadratic loss function

Theorem Let $\mathbf{Y} = [-Y, Y]$, $\mathbf{P} = \mathbb{R}$, and $\mathcal{F} \subseteq C(\mathbf{X})$ be a Hilbert space with finite embedding constant $c_{\mathcal{F}}$. There exists a strategy for Predictor that guarantees

$$\left| \sum_{n=1}^N (y_n - \mu_n)^2 - \sum_{n=1}^N (y_n - F(x_n))^2 + \sum_{n=1}^N (\mu_n - F(x_n))^2 \right| \leq 2Y \sqrt{c_{\mathcal{F}}^2 + 1} (\|F\|_{\mathcal{F}} + Y) \sqrt{N}, \quad \forall N \in \{1, 2, \dots\} \quad \forall F \in \mathcal{F}.$$

Generalizations

The result about existence of leading strategies for the quadratic loss function $\lambda(y, \mu) = (y - \mu)^2$ can be generalized to:

- loss functions given by Bregman divergences
- loss functions given by strictly proper scoring rules

Details: see the paper.

Result for the logarithmic loss function

A popular strictly proper scoring rule is the **log loss function**

$$\mathbf{Y} = \{0, 1\}, \quad \mathbf{P} = (0, 1), \quad \lambda(y, \mu) := \begin{cases} -\ln \mu & \text{if } y = 1 \\ -\ln(1 - \mu) & \text{if } y = 0. \end{cases}$$

Define the **Kullback–Leibler divergence** between two predictions (a.k.a. relative entropy):

$$D(\mu \parallel \mu') := \mu \ln \frac{\mu}{\mu'} + (1 - \mu) \ln \frac{1 - \mu}{1 - \mu'}.$$

Theorem Let $\mathbf{Y} = \{0, 1\}$, $\mathbf{P} = (0, 1)$, λ be the log loss function, and $\mathcal{F} \subseteq C(\mathbf{X})$ be a Hilbert space with finite $c_{\mathcal{F}}$. There exists a strategy for Predictor that guarantees, for all prediction rules F ,

$$\left| \sum_{n=1}^N \lambda(y_n, \mu_n) - \sum_{n=1}^N \lambda(y_n, F(x_n)) + \sum_{n=1}^N D(\mu_n \| F(x_n)) \right| \leq \frac{\sqrt{c_{\mathcal{F}}^2 + 1.8}}{2} \left(\left\| \ln \frac{F}{1-F} \right\|_{\mathcal{F}} + 1 \right) \sqrt{N}, \quad \forall N \in \{1, 2, \dots\}.$$

(The norm is set to infinity for functions not belonging to \mathcal{F} .)

Open problems

- Can leading strategies be obtained using other methods, such as Gradient Descent or Aggregating Algorithm? (This might lead to much better bounds for some benchmark classes.)
- What is the widest class of loss functions for which leading strategies exist?