Competitive on-line statistics

Volodya Vovk

Computer Learning Research Centre
and Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England
vovk@dcs.rhbnc.ac.uk

Technical Report

CLRC-TR-00-03

March 2000

Department of Computer Science
Egham, Surrey TW20 0EX, England

Abstract

A radically new approach to statistical modelling, which combines mathematical techniques of Bayesian statistics with the philosophy of the theory of competitive on-line algorithms, has arisen over the last decade in computer science (to a large degree, under the influence of Dawid's prequential statistics). In this approach, which we call "competitive on-line statistics", it is not assumed that data are generated by some stochastic mechanism; the bounds derived for the performance of competitive on-line statistical procedures are *guaranteed* to hold (and not just hold with high probability or on the average). This paper reviews some results in this area; the new material in it includes the proofs for the performance of the Aggregating Algorithm in the problem of linear regression with square loss.

1 Introduction

1.1 Problem

Making rational decisions is a central problem in science and everyday life. (Polynomials of which degree should I use to fit my data sets? Should I take my umbrella today, tomorrow, etc? Which stocks should I buy and sell this year?) Only rarely we can readily choose the best course of action; more often we will have a more or less extensive (maybe infinite) family of potentially successful decision strategies. (Whether a decision strategy is successful will depend not only on the merits of this strategy but also on the future events which we do not know yet.) However, at the end of the day we must choose one specific decision strategy, so we naturally arrive at this problem: **given a family of decision strategies, find a new decision strategy which will perform, under any circumstances, almost as well as the best (under those circumstances) decision strategy in the family**.

At first, this task might appear hopeless for even moderately interesting families of decision strategies. (Recall that we want the constructed decision strategy to perform almost as well as the best strategy in the family *always*; we do not make any stochastic assumptions about the generation of the future events.) However, for one specific loss function a good merging algorithm has been known for a long time (the Bayesian mixture) and for many more loss functions good merging algorithms have been found in recent years.

The usual *statistical model* is a family of probability distributions reflecting our knowledge or our assumptions about some piece of the world. The Bayesian framework involves another element, the prior distribution on the parameter set, which allows the Bayesian to replace the statistical model by a single probability distribution: the statistical model $\{Q_\theta \mid \theta \in \Theta\}$ with the prior distribution $P(d\theta)$ in Θ is replaced by the Bayesian mixture $Q = \int_\Theta Q_\theta P(d\theta)$. This formula, reinterpreted and generalized, lies at the centre of competitive on-line statistics.

In competitive on-line statistics the statistical model is replaced by the *decision pool*, which is a family of decision strategies; the statistician's goal is to replace this family with a single decision strategy. Recall that the competitive on-line statistician is agnostic: she does not assume anything about the stochastic mechanism generating the data; she does not even assume the existence of such a mechanism.

In the bulk of this paper (Sections 1–4) we describe in some detail only one of the existing merging algorithms (the Aggregating Algorithm [52, 55, 56]) and apply it to only one problem (linear regression with square loss function); a more balanced review of the field will be given in Section 5.

1.2 Aggregating Algorithm

Let Ω be some *sample space*, Γ be a *decision space* and Θ be a *parameter space* (the decision strategies in our decision pool will be indexed by $\theta \in \Theta$). We consider the following perfect-information game between three players, Statistician (usually called Learner in the machine learning literature), Decision Pool and Nature: at each trial $t = 1, 2, \ldots$,

- ⋄ Decision Pool makes a prediction $\xi_t : \Theta \to \Gamma$; $\xi_t(\theta)$ is interpreted as the decision recommended by the decision strategy $\theta \in \Theta$;
- ⋄ Statistician makes her own decision $\gamma_t \in \Gamma$;
- ⋄ Nature chooses some outcome $\omega_t \in \Omega$.

There is some fixed *loss function* $\lambda : \Omega \times \Gamma \to [0, \infty]$; Statistician's goal is to ensure that her cumulative loss

$$L_T(\text{Statistician}) = \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t)$$

is almost as good as the loss

$$L_T(\theta) = \sum_{t=1}^{T} \lambda(\omega_t, \xi_t(\theta))$$

of all or most of the decision strategies $\theta \in \Theta$. Assuming that the number $n = |\Theta|$ of decision strategies in the pool is finite, it is possible to prove, for a wide class of *games* $(\Omega, \Gamma, \lambda)$, that Statistician can ensure that, for all T and θ,

$$L_T(\text{Statistician}) \le c L_T(\theta) + a \ln n, \tag{1}$$

where c and a are some constants. For a wide class of games the *Aggregating Algorithm* (described in the next section and, in more detail, in [55, 56]) ensures that (1) holds with optimal c and a; that algorithm does not require, however, that the decision pool should be finite. Since it is possible to improve c at the expense of deteriorating a and vice versa, the algorithm involves a *learning rate* $\eta \in (0, \infty)$, and the optimal constants $c = c(\eta)$ and $a = a(\eta)$ in (1) depend on η.

The constants $c(\eta)$ and $a(\eta)$ have been found for many games; especially important are the *perfectly mixable* games, for which $c(\eta) = 1$ for some η. The most important for statistics games are perhaps the *log-loss* games; assuming for simplicity that Ω is finite, in the log-loss game with the sample space Ω the decision space Γ is the set of all probability distribution in Ω and the loss function is $\lambda(\omega, \gamma) = -\ln \gamma\{\omega\}$. For this game the Aggregating Algorithm with learning rate $\eta = 1$ coincides with the Bayesian mixture (assuming the uniform prior); the constants are $c(1) = a(1) = 1$. Also important, especially in the problems of regression, is the following *square-loss game*: $\Omega = \Gamma = [-1, 1]$ and

$\lambda(\omega, \gamma) = (\omega - \gamma)^2$. (We assume that the outcomes never exceed some known bound Y; without loss of generality we take $Y = 1$.) In the case of the square-loss game, $c(\eta) = 1$ and $a(\eta) = 2$ for some η. In general, a game is perfectly mixable if its loss function is "strictly convex" in some sense.

1.3 Linear regression

Even if the decision pool is infinite, in a surprisingly wide class of problems it is possible to derive good bounds for competitive on-line procedures; see Section 5 below. In this paper the methods of on-line competitive statistics will be illustrated on the problem of linear regression with the square loss. There are several competitive on-line algorithms for this problem; some of them will be discussed below (Subsection 5.1); we will consider in detail just one of them [54].

We have to extend slightly the protocol of the previous section: we will assume that at the beginning of every trial t Nature outputs a "signal" x_t to be used by Decision Pool and Statistician in making their decisions. We assume that the signals are taken from the L_∞-ball[1] $\{x \in \mathbb{R}^n \mid \|x\|_\infty \le X\}$ of radius X; the decision pool is indexed by the L_1-ball $\Theta = \{\theta \in \mathbb{R}^n \mid \|\theta\|_1 \le C\}$ of radius C; the decision strategy θ recommends prediction $\theta \cdot x_t$ at trial t. Applying the Aggregating Algorithm to this decision pool and a Gaussian prior, the standard bounds for that algorithm imply

$$L_T(\text{Statistician}) \le L_T(\theta) + C^2X^2 + nC^2X^2 \ln(T+1), \qquad (2)$$

for all T and θ. (For the proof of this inequality and its elaborations, see Section 4 below; the assumptions that the signal and parameter spaces should be bounded were made only for simplicity; the essential assumption is that the responses ω_t should be bounded by a known constant.) To see that bound (2) is tight, assume that $n = 1$, $\theta \in [-1, 1]$, $x_t = 1$ for all t, and ω_t are generated by the iid process with the probability $\frac{1+\theta}{2}$ of $\omega_t = 1$ and the probability $\frac{1-\theta}{2}$ of $\omega_t = -1$. It is easy to check (for details, see Subsection 3.3 below) that, when $\theta = 0$ and Statistician uses the Maximum Likelihood estimator for computing γ_t, the **expected value** of the difference between the right-hand and left-hand sides of (2) does not exceed the minute quantity of $\frac{1}{T}$.

As well known from traditional statistics (see, eg, [23], Chap 5), linear regression algorithms and their properties (such as (2)) can be immediately extended to apparently more general regression problems. For example, our results presented below will immediately imply the following (for simplicity we assume $n = 1$ and $x_t, \omega_t \in [-1, 1]$, $\forall t$):

- ◇ if the decision pool consists of all polynomials of degree d, Statistician has

[1] In this paper we will often make use of the norms L_p, $p \ge 1$. Recall that, for any $p > 0$, the L_p norm $\|x\|_p$ of a vector $x \in \mathbb{R}^n$ with coordinates x_i is $\left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$; by definition, $\|x\|_\infty = \max_i |x_i|$. It is well known that $\|x\|_p \le \|x\|_q$ when $p > q > 0$ (see, eg, Beckenbach and Bellman [5], Section 1.16).

a strategy guaranteeing

$$L_T(\text{Statistician}) \le \inf_\theta \left(L_T(\theta) + \|\theta\|_2^2\right) + (d+1)\ln(T+1)$$

($\|\theta\|_2^2$ being the squared L_2 norm of the polynomial's coefficients);

◇ if the decision pool consists of all splines of degree d with k nodes (chosen *a priori*), Statistician has a strategy guaranteeing

$$L_T(\text{Statistician}) \le \inf_\theta \left(L_T(\theta) + \|\theta\|_2^2\right) + (d+k+1)\ln(T+1).$$

1.4 This paper

This introductory section is essentially a revision of the conference version [59] of this paper. In the next section we will discuss in more detail the Aggregating Algorithm and its applications to some particular games. In Sections 3 and 4 we state and prove some elaborations of the results mentioned above about linear regression. In Section 5 we give a brief review of competitive on-line statistics; inevitably, many important results are not even mentioned in it. In the concluding Section 6 we briefly discuss limitations of competitive on-line statistics and put it in a more general context.

2 Aggregating Algorithm

2.1 Generic algorithm

In this subsection we describe the Aggregating Algorithm (AA), first when specialized to the case of perfectly mixable games and then in general (following [55] and [56]). Sometimes (following the tradition of the computational learning literature) we will say "expert" instead of "decision strategy"; the reader can imagine that each decision strategy in the pool is advocated by some expert; "pool of experts" is the same as "decision pool". We will assume that the set Θ of experts is equipped with a σ-algebra; the decision (or *prediction*) space Γ is a topological space equipped with the σ-algebra generated by the open sets, the outcome space Ω is just a set, without any additional structure. The functions $\xi_t : \Theta \to \Gamma$ chosen by Decision Pool are required to be measurable.

We fix a *learning rate* $\eta > 0$, put $\beta = e^{-\eta}$, and fix a probability distribution P_0 in the pool Θ; the *prior distribution* P_0 specifies the initial weights assigned to the experts.

First we describe an algorithm (the *Aggregating Pseudo-Algorithm*, or APA) that is allowed to make not permitted predictions $\gamma \in \Gamma$ but "mixtures", in some sense, of permitted predictions. A *generalized prediction* is defined to be any function of the type $\Omega \to [0,\infty]$; a permitted prediction $\gamma \in \Gamma$ is identified with the generalized prediction g defined by $g(\omega) = \lambda(\omega, \gamma)$. The APA suffers a loss of $g_t(\omega_t)$ after choosing generalized prediction g_t when the actual outcome is ω_t.

The APA works as follows. At every trial $t = 1, 2, \ldots$ Statistician updates the experts' weights,

$$P_t(d\theta) = \beta^{\lambda(\omega_t, \xi_t(\theta))} P_{t-1}(d\theta), \quad \theta \in \Theta, \tag{3}$$

where P_0 is the prior distribution. (So the weight of an expert θ whose prediction $\xi_t(\theta)$ leads to a large loss $\lambda(\omega_t, \xi_t(\theta))$ gets slashed. Recall that (3) is equivalent, by definition, to

$$P_t(E) = \int_E \beta^{\lambda(\omega_t, \xi_t(\theta))} P_{t-1}(d\theta),$$

for all measurable $E \subseteq \Theta$.) The generalized prediction chosen by the APA at trial t is the weighted average of the experts' predictions:

$$g_t(\omega) = \log_\beta \int_\Theta \beta^{\lambda(\omega, \xi_t(\theta))} P^*_{t-1}(d\theta), \tag{4}$$

where P^*_{t-1} are the normalized weights,

$$P^*_{t-1}(d\theta) = \frac{P_{t-1}(d\theta)}{P_{t-1}(\Theta)} \tag{5}$$

(assuming that the denominator is positive; if it is 0, P_0-almost all experts have suffered infinite loss and, therefore, the AA is allowed to choose any prediction). Later on (see the proof of Lemma 1 and Subsection 2.2) we will see that rules (3) and (4) are very natural and generalize Bayes's rule.

The AA is obtained from the APA by replacing each generalized prediction g_t by a permitted prediction $\gamma_t = \Sigma(g_t)$, where the *substitution function* Σ maps every generalized prediction $g : \Omega \to [0, \infty]$ into a permitted prediction $\Sigma(g) \in \Gamma$. The AA as described in [55] requires that a "minimax" substitution function should be chosen, but it is often convenient to relax this requirement (in order to make the algorithm more efficient in some important cases, such as linear regression). Notice that the generalized predictions output by the APA are always of the form

$$\omega \mapsto \log_\beta \int_\Theta \beta^{\lambda(\omega, \gamma)} Q(d\gamma),$$

Q ranging over the probability distributions in Γ; we let $\mathcal{P}(\lambda, \eta)$ stand for the set of all generalized predictions of this form. We say that a substitution function Σ is *perfect* (for given λ and η) if, for every generalized prediction $g \in \mathcal{P}(\lambda, \eta)$,

$$\lambda(\omega, \Sigma(g)) \leq g(\omega).$$

First we will assume that such a substitution function exists (in this case we will say that our game $(\Omega, \Gamma, \lambda)$ is η-*mixable*).

Now we can describe how the AA works: its only difference from the APA is that it outputs $\Sigma(g_t)$ instead of the APA's generalized prediction g_t, where Σ is a fixed perfect substitution function. When Statistician follows $\mathrm{AA}(\eta, P_0)$ (ie, the AA with learning rate η and prior P_0) we will write $L_T(\mathrm{AA}(\eta, P_0))$ in place of $L_T(\text{Statistician})$; we will also use analogous notation for the APA and other algorithms.

All our proofs are based on the following property of the APA (see [55]).

Lemma 1 *For any learning rate $\eta > 0$, prior P_0, and $T = 1, 2, \ldots$,*

$$L_T(\mathrm{APA}(\eta, P_0)) = \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \tag{6}$$

Proof We are required to deduce (6) from the weight update rule (3) and the formula (4) for computing pseudopredictions. It is more natural, however, to move in another direction, from (6) and (3) to (4). If we want to achieve goal (6) (this goal is the generalization of the formula $Q = \int_\Theta Q_\theta P(d\theta)$ for Bayesian mixture; see Subsection 2.2 below), we have little choice but to compute pseudopredictions using (4) and (3). Indeed, noticing that

$$P_{T-1}(d\theta) = \beta^{L_{T-1}(\theta)} P_0(d\theta)$$

(this immediately follows from (3)) and assuming (6), we obtain:

$$g_T(\omega_T) = L_T(\mathrm{APA}(\eta, P_0)) - L_{T-1}(\mathrm{APA}(\eta, P_0)) = \log_\beta \frac{\int_\Theta \beta^{L_T(\theta)} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)}$$

$$= \log_\beta \frac{\int_\Theta \beta^{L_{T-1}(\theta)+\lambda(\omega_T,\xi_T(\theta))} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} = \log_\beta \frac{\int_\Theta \beta^{\lambda(\omega_T,\xi_T(\theta))} P_{T-1}(d\theta)}{P_{T-1}(\Theta)},$$

which coincides with (4). Reversing this argument, we can see that (4) is not only necessary but also sufficient for (6). ∎

For an η-mixable game, Lemma 1 implies

$$L_T(\mathrm{AA}(\eta, P_0)) \le \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \tag{7}$$

In particular, if there are only finitely many experts and every expert is assigned the same weight $\frac{1}{n}$ (n is the number of experts), we have, for any $\theta \in \Theta$,

$$L_T(\mathrm{AA}(\eta, P_0)) \le \log_\beta \left(\frac{1}{n} \sum_{\theta\in\Theta} \beta^{L_T(\theta)} \right) \le \log_\beta \left(\frac{1}{n} \beta^{L_T(\theta)} \right) = L_T(\theta) + \frac{\ln n}{\eta},$$

which coincides with (1) with $c = 1$ and $a = \frac{1}{\eta}$.

Recall that a game is called perfectly mixable if it is η-mixable for some learning rate $\eta > 0$. Now we will discuss the case where our game is not perfectly mixable (or the game is perfectly mixable but we are interested in η for which it is not η-mixable). In this case no perfect substitution function exists, so we need a different criterion for choosing the substitution function Σ; our choice of Σ will depend on η. First we define the important notion of the *mixability curve* $c(\eta)$. For any $\eta \in (0, \infty)$ we put

$$c(\eta) = \inf \{c \mid \forall g \in \mathcal{P}(\lambda, \eta)\, \exists \delta \in \Gamma\, \forall \omega \colon \lambda(\omega, \delta) \le c g(\omega)\}, \tag{8}$$

where $\inf \emptyset$ is set to ∞. A related function is $a(\eta) = \frac{c(\eta)}{\eta}$ (cf (14)). As shown in [55], $c(\eta)$ and $a(\eta)$ are continuous and monotonic functions ($c(\eta)$ nondecreasing and $a(\eta)$ nonincreasing).

We will always assume that our substitution function $\Sigma = \Sigma_\eta$ satisfies

$$\forall \eta\, \forall \omega \colon \lambda(\omega, \Sigma_\eta(g)) \le c(\eta) g(\omega) \tag{9}$$

for any pseudoprediction $g \in \mathcal{P}(\lambda, \eta)$. We can satisfy this requirement provided that the infimum in (8) is attained; it is under mild assumptions (for details, see [55] and [56]) about the game $(\Omega, \Gamma, \lambda)$.

A natural way to ensure assumption (9) is to require that, for every pseudoprediction g,

$$\Sigma_\eta(g) \in \arg\min_{\gamma \in \Gamma} \sup_{\omega \in \Omega} \frac{\lambda(\omega, \gamma)}{g(\omega)} \tag{10}$$

(where $\frac{0}{0}$ is set to 0); a pleasant feature of such a definition would be the independence of Σ_η from η.

The following approach, however, seems to be more computationally efficient in many situations: we require that

$$\Sigma_\eta(g) \in \arg\min_{\gamma \in \Gamma} \sup_{\omega \in \Omega} \left(\lambda(\omega, \gamma) - c(\eta) g(\omega)\right) \tag{11}$$

(again min is attained under mild assumptions about the game $(\Omega, \Gamma, \lambda)$) and

$$(g_1(\omega) - g_2(\omega) \text{ does not depend on } \omega) \Longrightarrow (\Sigma_\eta(g_1) = \Sigma_\eta(g_2)). \tag{12}$$

(Assumption (12) is always compatible with (11) but is typically incompatible with (10).) A crucial advantage of assumption (12) is that when running the AA we do not need to normalize the weights $P_t(d\theta)$, since the pseudoprediction

$$\omega \mapsto \log_\beta \int_\Theta \beta^{\lambda(\omega, \xi_t(\theta))} P_{t-1}(d\theta)$$

calculated from the unnormalized weights will differ from the pseudoprediction (4) calculated from the normalized weights by only an additive constant. Besides avoiding normalization of the weights at every trial, which is often computationally difficult, this way of defining the substitution function can lead to significant simplifications of the AA in particular applications; see, eg, Subsection 2.4 below.

We usually drop the index η of Σ_η.

For a game which is not necessarily η-mixable, we have, instead of (7),

$$L_T(\mathrm{AA}(\eta, P_0)) \le c(\eta) \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta); \tag{13}$$

the other inequalities given above will also remain true if their right-hand side is multiplied by $c(\eta)$. In particular, in the case of $n < \infty$ experts we have

$$L_T(\mathrm{AA}) \le c(\eta) L_T(\theta) + a(\eta) \ln n, \tag{14}$$

where $a(\eta) = \frac{c(\eta)}{\eta}$; cf (1).

2.2 Log-loss games

Let us assume, for simplicity, that the sets Ω and Θ are finite. In the *log-loss game*, Γ is defined to be the set of all probability distributions in Ω and the loss function is defined as $\lambda(\omega, \gamma) = -\ln \gamma(\omega)$ (to simplify the notation, we let $\gamma(\omega)$ stand for the more formal $\gamma(\{\omega\})$). Let P_0 be a prior probability distribution in Θ; set $\eta = 1$ (and, therefore, $\beta = 1/e$). With every decision strategy $\theta \in \Theta$ we associate the unique probability distribution Q_θ in Ω^∞ such that the expert θ's prediction $\xi_t(\theta) = \xi_t^\theta$ (the notation with θ as an upper index will be more convenient in this subsection) is (a variant of) the conditional Q_θ-probability for different values of ω_t given $\omega_1, \ldots, \omega_{t-1}$. In other words, expert θ is a Bayesian whose subjective probability distribution is Q_θ; at the beginning of every trial t this expert quotes his subjective probabilities for ω_t given the past $\omega_1 \ldots \omega_{t-1}$.

The weight update rule (3),

$$P_t(\theta) = \beta^{\lambda(\omega_t, \xi_t^\theta)} P_{t-1}(\theta),$$

becomes

$$P_t(\theta) = \xi_t^\theta(\omega) P_{t-1}(\theta);$$

therefore, the normalized version (cf (5))

$$P_t^*(\theta) = \frac{Q_\theta(\omega_1 \ldots \omega_t) P_0(\theta)}{\sum_\theta Q_\theta(\omega_1 \ldots \omega_t) P_0(\theta)}$$

of P_t is identical to the posterior probability of θ after observing $\omega_1 \ldots \omega_t$.

It is easy to see that the integral in (4) is the predictive distribution of the Bayesian mixture $\int_\Theta Q_\theta P_0(d\theta)$; therefore, generalized prediction (4) corresponds to a permitted prediction (viz the predictive distribution). It is clear that the AA (identical, in this case, to the APA) corresponds to the Bayesian mixture in the same sense as every decision strategy θ corresponds to Q_θ (ie, at every trial t the AA outputs the predictive distribution, according to the Bayesian mixture, of ω_t given the past).

To summarize, in the case of the log-loss game:

- ◇ the weight update rule (3) is identical to Bayes's theorem;
- ◇ the AA is identical to the predictive version of the Bayesian mixture;
- ◇ the APA is identical to the predictive version of the Bayesian mixture; this fact is also expressed by Lemma 1.

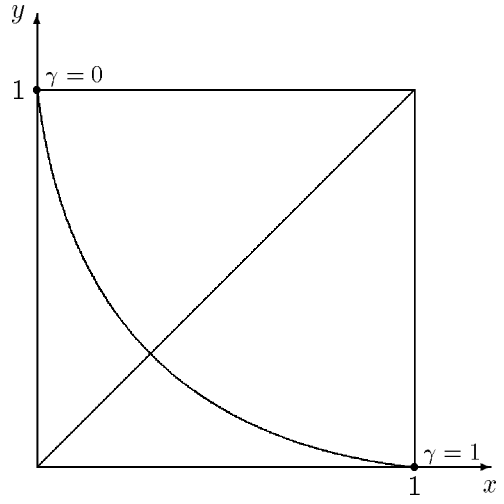

Figure 1 The substitution function for the simple prediction game

2.3 Simple prediction game

Let us consider the following game $(\Omega, \Gamma, \lambda)$, which we call the *simple prediction game*:

$$\Omega = \Gamma = \{0, 1\}, \quad \lambda(\omega, \gamma) = \begin{cases} 0, & \text{if } \omega = \gamma, \\ 1, & \text{otherwise.} \end{cases}$$

(Therefore, in this game Statistician is trying to predict a binary classification, 0 or 1; at every trial she suffers a loss of 1 if she makes a mistake.) Suppose the pool contains a finite number n of decision strategies. To clarify the notions of pseudoprediction, substitution function, etc, in the rest of this subsection we will apply the AA to predicting in the simple prediction game using advice of n experts. (In this case the AA becomes the Weighted Majority Algorithm; see [43] and [53].)

In the case where there are just two possible outcomes, say 0 and 1 (as in the simple prediction game), it is convenient to represent every prediction $\gamma \in \Gamma$ as the point $(\lambda(0, \gamma), \lambda(1, \gamma))$ of the (x, y)-plane. There are two permitted predictions in the simple prediction game, 0 and 1, which are depicted as small filled circles in Figure 1. It is also convenient to represent every pseudoprediction $g : \Omega \to [0, \infty]$ as the point $(g(0), g(1))$ of the (x, y)-plane. Possible mixtures

$$\Big(\log_\beta \left(\beta p + (1 - p)\right), \log_\beta \left(p + \beta(1 - p)\right)\Big) \tag{15}$$

(see (4); this expression only depends on the total weight $p = P^*_{t-1}(\Theta_t(1))$ of the experts $\Theta_t(1) = \{\theta \in \Theta \mid \xi_t(\theta) = 1\}$ who predict 1 at trial t; notice that $1 - p$ is the total weight $P^*_{t-1}(\Theta_t(0))$ of the experts $\Theta_t(0) = \{\theta \in \Theta \mid \xi_t(\theta) = 0\}$ who predict 0 at trial t) of the permitted predictions are shown in Figure 1 by the curve (which we will call the *pseudoprediction curve*) connecting the two permitted predictions; (15) is a parametric equation of this curve (the parameter p ranges between 0 and 1).

trial no	Pool's weights	Pool's predictions	Statistician's pseudoprediction	Her prediction	outcome	Pool's losses
1	$(1.00, 1.00, 1.00)$	$(1, 1, 0)$	$(-0.55, -0.86)$	1	0	$(1, 1, 0)$
2	$(0.37, 0.37, 1.00)$	$(0, 1, 1)$	$(0.14, -0.41)$	1	1	$(1, 0, 0)$
3	$(0.14, 0.37, 1.00)$	$(0, 1, 0)$	$(-0.24, 0.24)$	0	1	$(1, 0, 1)$

Table 1 Example of execution of the AA for the simple prediction game

It is clear from Figure 1 that in the simple prediction game $1/c(\eta)$ (see (8)) equals the abscissa (equivalently, the ordinate) of the intersection of the pseudoprediction curve and the straight line $x = y$; this intersection corresponds to $p = 1/2$ in (15), which gives

$$\frac{1}{c(\eta)} = \log_\beta \left(\frac{1+\beta}{2}\right);$$

equivalently,

$$c(\eta) = \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}} = \frac{\eta}{\ln \frac{2}{1+\exp(-\eta)}}.$$

Now it is clear that for every η there is essentially only one substitution function Σ_η satisfying (9): if g is above the line $x = y$, Σ_η should map g to the prediction $\gamma = 0$; if g is below the line $x = y$, Σ_η should map g to the prediction $\gamma = 1$; and only if g is exactly on the line $x = y$, $\Sigma_\eta(g)$ can be defined arbitrarily. Notice that (15) being above the line $x = y$ means that $x < y$, i.e.,

$$\log_\beta (\beta p + (1-p)) < \log_\beta (p + \beta(1-p)),$$

which is equivalent to $p < 1/2$; analogously, (15) being below the line $x = y$ is equivalent to $p > 1/2$. This means that the AA predicts according to the weighted majority of the experts, which explains the name "Weighted Majority Algorithm".

We conclude this subsection with an example of execution of the AA for the simple prediction game. Table 1 describes the AA's behaviour in the first 3 trials in the situation where: there are three experts with equal initial weights who give predictions $(1, 1, 0)$, $(0, 1, 1)$ and $(0, 1, 0)$; the actual outcomes (Nature's moves) are 0, 1 and 1; the learning rate is 1 (and so $\beta = 1/e$). The weights given in Table 1 are *unnormalized*, and instead of (15) we use its unnormalized version

$$\Big(\log_\beta (\beta p + q), \log_\beta (p + \beta q)\Big), \tag{16}$$

where $p = P_{t-1}(\Theta_t(1))$ is the total unnormalized weight of the experts $\Theta_t(1) = \{\theta \in \Theta \mid \xi_t(\theta) = 1\}$ who predict 1 at trial t and $q = P_{t-1}(\Theta_t(0))$ is the total unnormalized weight of the experts $\Theta_t(0) = \{\theta \in \Theta \mid \xi_t(\theta) = 0\}$ who predict 0 at trial t. We simplify (16) to

$$\left(\ln \frac{1}{q + p/e}, \ln \frac{1}{p + q/e}\right).$$

2.4 Some results about the square-loss game

In this section we discuss the square-loss game; recall that in this game

$$\Omega = [-1, 1],\ \Gamma = \mathbb{R},\ \lambda(\omega, \gamma) = (\omega - \gamma)^2.$$

It is shown in [52] that this game is perfectly mixable if and only if $\eta \le \frac{1}{2}$, under the restriction $\omega_t \in \{-1, 1\}$. Haussler et al [33] proved that this restriction can be removed. The goal of this section is to prove these facts and to find an explicit expression for a substitution function in the square-loss game.

First we consider the restricted square-loss game in which it is required that $\omega_t \in \{-1, 1\}$. The following lemma is proven in [52] and Haussler et al [33], but we will give a simple independent proof.

Lemma 2 *The restricted square-loss game is η-mixable if and only if $\eta \le \frac{1}{2}$.*

Proof Let us represent (analogously to the previous subsection) a pseudoprediction $(g(-1), g(1))$ in the restricted square-loss game by the point

$$(x, y) = (e^{-\eta g(-1)}, e^{-\eta g(1)}) \in [0, 1]^2$$

of the (x, y)-plane. Therefore, the set of permitted predictions will be represented by the parametrical curve

$$(x(\gamma), y(\gamma)) = \left(e^{-\eta(-1-\gamma)^2}, e^{-\eta(1-\gamma)^2}\right),$$

where γ ranges over $[-1, 1]$. The game is η-mixable if and only if this curve veers to the left (when moving in the direction in which γ increases), ie if and only if

$$\det \begin{pmatrix} x'(\gamma) & y'(\gamma) \\ x''(\gamma) & y''(\gamma) \end{pmatrix} \ge 0, \quad \forall \gamma. \tag{17}$$

Computing the derivatives we find

$$(x'(\gamma), x''(\gamma)) \propto (-(1+\gamma), -1 + 2\eta(1+\gamma)^2),$$
$$(y'(\gamma), y''(\gamma)) \propto ((1-\gamma), -1 + 2\eta(1-\gamma)^2),$$

with positive proportionality constants; therefore, condition (17) can be equivalently transformed as

$$-(1+\gamma)(-1 + 2\eta(1-\gamma)^2) - (1-\gamma)(-1 + 2\eta(1+\gamma)^2) \ge 0, \quad \forall \gamma;$$
$$\eta(1-\gamma^2) \le \frac{1}{2}, \quad \forall \gamma;$$
$$\eta \le \frac{1}{2}.$$

▮

The following lemma is an elaboration of Haussler et al's [33] result that the restriction $\omega_t \in \{-1, 1\}$ (assumed in [52]) can be removed. It asserts that any substitution function for the restricted game is also a substitution function in the full game.

Lemma 3 *Fix Y_1 and Y_2 such that $Y_1 < Y_2$. Let y and p range over $[Y_1, Y_2]$ and $\mathbb{R}$, respectively, and $\lambda(y,p) = (y-p)^2$ be the square loss function. Let P be a probability distribution in $\mathbb{R}$; put*

$$f(y) = \log_\beta \int \beta^{(y-p)^2} P(dp).$$

(So f is the pseudoprediction corresponding to P.) For every $\gamma \in \mathbb{R}$,

$$(\lambda(Y_1,\gamma) \le f(Y_1) \ \& \ \lambda(Y_2,\gamma) \le f(Y_2)) \Longrightarrow (\lambda(y,\gamma) \le f(y), \forall y \in [Y_1, Y_2]).$$

Proof It is sufficient to prove that for any fixed γ the function $\lambda(y,\gamma) - f(y)$ is convex in $y \in \mathbb{R}$. Since

$$\lambda(y,\gamma) - f(y) = (y-\gamma)^2 - \log_\beta \int \beta^{(y-p)^2} P(dp)$$
$$= (y-\gamma)^2 + \frac{1}{\eta} \ln \int e^{-\eta(y-p)^2} P(dp),$$

it suffices to notice that

$$\frac{\partial^2}{\partial y^2} \ln \int e^{-\eta(y-p)^2} P(dp) = \frac{1}{\left(\int e^{-\eta(y-p)^2} P(dp)\right)^2}$$
$$\times \Bigg(\int e^{-\eta(y-p)^2} (-2\eta(y-p))^2 P(dp) \int e^{-\eta(y-p)^2} P(dp)$$
$$+ \int e^{-\eta(y-p)^2} (-2\eta) P(dp) \int e^{-\eta(y-p)^2} P(dp)$$
$$- \left(\int e^{-\eta(y-p)^2} (-2\eta(y-p)) P(dp) \right)^2 \Bigg)$$
$$\ge \frac{1}{\left(\int e^{-\eta(y-p)^2} P(dp)\right)^2} \left(\int e^{-\eta(y-p)^2} (-2\eta) P(dp) \int e^{-\eta(y-p)^2} P(dp) \right) = -2\eta$$

(the inequality follows from the Schwarz inequality; see, eg, [5], Section 1.18). ∎

In the case of the square-loss game with $y_t \in [-Y, Y]$, the minimax in the sense of (10) substitution function is

$$\Sigma(g) = Y \frac{\sqrt{g(-Y)} - \sqrt{g(Y)}}{\sqrt{g(-Y)} + \sqrt{g(Y)}}.$$

As explained in Subsection 2.1, it is likely that we obtain more efficient algorithms if we use (11) and (12) instead. It is clear that $\gamma = \Sigma_\eta(g)$ will satisfy

$$(-Y-\gamma)^2 - g(-Y) = (Y-\gamma)^2 - g(Y),$$

which gives

$$\gamma = \frac{g(-Y) - g(Y)}{4Y}. \tag{18}$$

(This expression is one of the bounds in [33].)

In the case $y_t \in [-Y, Y]$, $\eta = \frac{1}{2}$ of Lemma 2 should be replaced by $\eta = \frac{1}{2Y^2}$.

Remark 1 Kivinen and Warmuth [41] ask the question when the *weighted average* will be a perfect substitution function, ie, when it is true that

$$\forall\omega\forall P\colon \lambda\left(\omega, \int_\Gamma \gamma P(d\gamma)\right) \le \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} P(d\gamma),$$

with P running over the probability distributions in Γ (cf (9)). We can see that the weighted average will be a perfect substitution function if and only if the function $\beta^{\lambda(\omega,\gamma)}$ in concave in γ. For the square-loss game this is equivalent to

$$\frac{\partial^2}{\partial\gamma^2} e^{-\eta(\omega-\gamma)^2} \le 0, \ \forall\omega,\gamma;$$

for $\omega,\gamma \in [-Y,Y]$, simple calculations give that this is true if and only if $\eta \le \frac{1}{8Y^2}$ (which is 4 times worse than the constant in Lemma 2; we will meet this factor of 4 again in Section 3 below).

2.5 Other games

As we already mentioned, the log-loss and square-loss games are perfectly mixable. Some important games are not perfectly mixable, such as the simple prediction game, where

$$c(\eta) = \frac{\eta}{\ln\frac{2}{1+\exp(-\eta)}},\ a(\eta) = \frac{1}{\ln\frac{2}{1+\exp(-\eta)}}.$$

When we take $\Gamma = [0,1]$ instead (the *absolute loss* game), $c(\eta)$ and $a(\eta)$ are halved (as evident from Figure 1) and the game becomes "almost perfectly mixable", in the sense that $c(\eta) \to 1$ as $\eta \to 0$ (this is also true if $\Omega = [0,1]$).

Unfortunately, some important perfectly mixable games, such as the square-loss game, cease to be perfectly mixable when the sets of possible predictions and outcomes are widened; in the case of the square-loss game $c(\eta)$ will jump to ∞ (for any $\eta \in (0,\infty)$) when the predictions and outcomes are allowed to take values in the whole real line $\mathbb{R}$. It is interesting to characterize the loss functions $\lambda(\omega,\gamma)$ giving perfectly mixable games when ω and γ are allowed to run over the whole of $\mathbb{R}$. An obvious example of such a game is

$$\Omega = \Gamma = \mathbb{R},\ \lambda(\omega,\gamma) = (\tanh\omega - \tanh\gamma)^2,$$

but perhaps there are less trivial examples.

There are many interesting loss functions for which the mixability curves have not been found, such as the "robust" loss functions mentioned in Press et al [45], Section 15.7: the *Lorentzian* loss function

$$\lambda(\omega,\gamma) = \ln\left(1 + \frac{(\omega-\gamma)^2}{2}\right),$$

Andrew's loss function

$$\lambda(\omega,\gamma) = \begin{cases} 1-\cos\frac{\omega-\gamma}{c}, \text{ if } |\omega-\gamma| < c\pi, \\ 2, \text{ otherwise,} \end{cases}$$

and *Tukey's biweight* function

$$\lambda(\omega,\gamma) = \begin{cases} 1-\left(1-\frac{(\omega-\gamma)^2}{c^2}\right)^3, \text{ if } |\omega-\gamma| < c, \\ 1, \text{ otherwise} \end{cases}$$

An important class of games of prediction are the *probability games*, in which the predictions γ and the outcomes ω are probability distributions in some set, which we, for simplicity, will assume to be finite and of the form $\{1,\ldots,K\}$. The most important probability games are perhaps the Kullback–Leibler game, the Hellinger game and the χ^2-game with

$$\Omega = \Gamma = \{1,\ldots,K\}$$

and the loss functions

$$\lambda(\omega,\gamma) = \mathbf{E}_\omega \ln\frac{\omega}{\gamma} = \sum_{k=1}^{K} \omega_k \ln\frac{\omega_k}{\gamma_k},$$

$$\lambda(\omega,\gamma) = \sum_{k=1}^{K} \left(\sqrt{\omega_k}-\sqrt{\gamma_k}\right)^2,$$

$$\lambda(\omega,\gamma) = \sum_{k=1}^{K} \frac{(\omega_k-\gamma_k)^2}{\gamma_k},$$

respectively (some applications of these loss functions are described in [8, 51]). Notice that the Kullback–Leibler game includes the log-loss game as a special case (take the degenerate ωs). The following result was proven in Haussler et al [33] (Example 4.3) under the assumption that $K=2$.

Lemma 4 *The Kullback–Leibler game is* 1*-mixable. The AA for the Kullback–Leibler game with learning rate* 1 *coincides with the Bayesian mixture.*

Proof Fix learning rate $\eta=1$. From considering the log-loss game it is clear that if the Kullback–Leibler game is 1-mixable, the weighted average (see Remark 1 above) must be the only substitution function. Let us prove that it will indeed be a substitution function. Suppose $p+q=1$ and $\gamma,\delta\in\Gamma$; we are required to prove

$$\lambda(\omega, p\gamma+q\delta) \le \log_\beta\left(p\beta^{\lambda(\omega,\gamma)} + q\beta^{\lambda(\omega,\delta)}\right),$$

where $\beta = 1/e$; or, after equivalent transformations:

$$\beta^{\lambda(\omega,p\gamma+q\delta)} \ge p\beta^{\lambda(\omega,\gamma)} + q\beta^{\lambda(\omega,\delta)},$$

$$e^{-\sum_k \omega_k \ln \frac{\omega_k}{p\gamma_k+q\delta_k}} \geq p e^{-\sum_k \omega_k \ln \frac{\omega_k}{\gamma_k}} + q e^{-\sum_k \omega_k \ln \frac{\omega_k}{\delta_k}},$$
$$e^{\sum_k \omega_k \ln(p\gamma_k+q\delta_k)} \geq p e^{\sum_k \omega_k \ln \gamma_k} + q e^{\sum_k \omega_k \ln \delta_k},$$
$$\prod_k (p\gamma_k + q\delta_k)^{\omega_k} \geq p \prod_k \gamma_k^{\omega_k} + q \prod_k \delta_k^{\omega_k}.$$

The last inequality follows from the concavity of the function

$$f : (\gamma_1, \ldots, \gamma_K) \mapsto \prod_k \gamma_k^{\omega_k}.$$

To check that this function is indeed concave, notice that the following quadratic form is negative definite:

$$\sum_{k,j} \frac{\partial^2 f}{\partial\gamma_k \partial\gamma_j} x_k x_j \approx \sum_{k\neq j} \frac{\omega_k\omega_j}{\gamma_k\gamma_j} x_k x_j + \sum_k \frac{\omega_k(\omega_k - 1)}{\gamma_k^2} x_k^2$$

$$\approx \sum_{k\neq j} \omega_k\omega_j y_k y_j + \sum_k \omega_k(\omega_k - 1) y_k^2 = \sum_{k,j} \omega_k\omega_j y_k y_j - \sum_k \omega_k y_k^2$$

$$= \left(\sum_k \omega_k y_k\right)^2 - \sum_k \omega_k y_k^2 = (\mathbf{E}_\omega y)^2 - \mathbf{E}_\omega(y^2)$$

(where $\approx$ means that the quadratic forms on both sides of this sign are either simultaneously negative definite or not); the negative definiteness of the last expression follows from Lyapunov's inequality (see, eg, [5], Section 1.16). ∎

Another interesting class of games are connected with finance; the simplest game of this kind is *Cover's game*:

$$\Omega = [0,\infty)^K,\ \Gamma = \{(p_1, \ldots, p_K) \in [0,\infty)^K \mid p_1 + \cdots + p_K = 1\},$$

$$\lambda((y_1, \ldots, y_K), (p_1, \ldots, p_K)) = -\ln \sum_{k=1}^K y_k p_k;$$

the financial meaning of this game is described in, eg, [18, 58]. Like the Kullback–Leibler game, Cover's game is also a generalization of the log-loss game (with a finite sample space): it suffices to consider "degenerate" $\omega \in \Omega$ of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$. (Such "degenerate" ω can be regarded as "crisp events"; they can be generalized to "fuzzy events" $\omega = (y_1, \ldots, y_K) \in [0,1]^K$, where y_k is interpreted as the weight of evidence in favour of k occurring.) Cover's "universal portfolio" algorithm [17, 18] is the AA applied to this game and a particular decision pool, the constant rebalanced portfolios (cf [58]); however, the AA and Cover's algorithm were found independently of each other (the AA was obtained by bridging the Weighted Majority Algorithm and the Bayesian mixing rule).

Cover's game ignores transaction costs and the possibility of "short selling" in financial markets. The *long-short game*, introduced by Watkins [58], takes the possibility of "short selling" into account. Both Cover's game and the long-short game are perfectly mixable. Taking into account transaction costs leads to a plethora of new games (cf Blum and Kalai [7] and Vovk and Watkins [58]).

The constants $c(\eta)$ and $a(\eta)$ for the games mentioned above were found in [22, 43, 33, 18, 52, 55, 58]. We have already noticed that a game is perfectly mixable if its loss function is "strictly convex" in some sense. The exact statement in the binary case (ie, where the outcome space Ω consists of only 2 elements) can be found in [52] (Lemma 2) and [33]; it is an open problem to find a simple general criterion.

3 Regression: results

In this section we will show in detail how to apply the AA to the problem of linear regression with the square loss; our main assumption is that the response variable is bounded. It turns out that for this particular problem the AA (when applied to the decision pool of linear functions with a Gaussian prior) resembles, but is different from, the well-known Ridge Regression (RR) procedure. From general results about the AA we deduce a guaranteed bound on the difference between the AA's performance and the best, in some sense, linear regression function's performance. We show that the AA attains the optimal constant in our bound, whereas the constant attained by the RR procedure in general can be 4 times worse.

3.1 Explicit algorithm

In the problem of regression, we consider the following protocol of interaction between Statistician and Nature (a modification of the protocol of the previous section):

FOR $t = 1, 2, \ldots$
  Nature chooses $x_t \in \mathbb{R}^n$
  Statistician chooses prediction $\gamma_t \in \mathbb{R}$
  Nature chooses $y_t \in [-Y, Y]$
END FOR.

(Notice that we now use the notation y_t, rather than ω_t, for the outcomes chosen by Nature; this (x, y) notation is conventional in the regression literature.) This is a "perfect-information" protocol: either player can see the other player's moves. The parameters of our protocol are: a fixed positive number n (the dimensionality of our regression problem) and an upper bound $Y > 0$ on the value y_t returned by Nature. It is important, however, that our algorithm for playing this game (on the part of Statistician) *does not need to know* Y.

In this subsection we only give a description of our regression algorithm; its rigorous derivation from the general AA will be given in the next section. (It is usually a non-trivial task to represent the AA in a computationally efficient form, and the case of on-line linear regression is not an exception.) Here we will only briefly describe the main idea of how the AA is applied to our regression problem. Let Σ be a perfect substitution function for the square-loss game and

$\eta = \frac{1}{2Y^2}$. Our experts are indexed by $\theta \in \mathbb{R}^n$. At trial t, expert θ outputs the prediction $\xi_t(\theta) = \theta \cdot x_t$; all such predictions are averaged in accordance with the experts' weights and Σ is applied to the resulting generalized prediction. (For details, see the next section.)

Fix n and $a > 0$. The algorithm is as follows:

> $A = aI$; $b = 0$
> FOR TRIAL $t = 1, 2, \ldots$:
>   read new $x_t \in \mathbb{R}^n$
> $A = A + x_t x_t'$
>   output prediction $\gamma_t = b' A^{-1} x_t$
>   read new $y_t \in \mathbb{R}$
> $b = b + y_t x_t$
> END FOR.

In this description, A is an $n \times n$ matrix (which is always symmetrical and positive definite), $b \in \mathbb{R}^n$, I is the unit $n \times n$ matrix, and 0 is the all-0 vector. As usual, vectors are identified with one-column matrices; B' stands for the transpose of matrix B.

This algorithm will be denoted AAR (the AA for Regression), or AAR(n, a). It should be remembered, however, that in principle the AA can be applied to the problem of regression in many different ways, and the AAR was derived from the AA under some strong extra assumptions (including a Gaussian prior on the decision pool).

3.2 Upper bounds

In this subsection we state results describing the predictive performance of the AAR; they (as well as the results stated in the rest of this section) will be proven in the next section. Recall that our decision pool consists of the linear functions $x_t \mapsto \theta \cdot x_t$, where $\theta \in \mathbb{R}^n$. At every trial t expert θ and Statistician suffer loss $(y_t - \theta \cdot x_t)^2$ and $(y_t - \gamma_t)^2$, respectively. A typical bound for the AAR that we would like to prove is: assuming that the signals x_t and the parameter θ are confined to the unit balls in the metrics L_∞ and L_1, respectively,

$$L_T(\mathrm{AAR}) \le L_T(\theta) + n \ln(T+1) + 1, \qquad (19)$$

for all T and θ (this is (2) with $C = X = 1$).

Remark 2 In the case $n = 1$ and $x_t = 1$, $\forall t$, inequality (19) differs from Freund's [27] Theorem 4 only in the additive constant. In that paper Freund noticed that for the problems that he considered the adversarial bounds of competitive on-line statistics are only a tiny amount worse than the average-case bounds for some stochastic strategies for Nature; this paper shows other manifestations of this phenomenon.

For compact pools of experts (which, in our setting, corresponds to the set of possible weights θ being bounded and closed) it is usually possible to derive bounds (such as (19)) where Statistician's loss is compared to the best expert's loss. In the case of non-compact pool, however, we need to give Statistician a start on remote experts. Specifically, instead of comparing Statistician's performance to $\inf_\theta L_T(\theta)$, we compare it to $\inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2\right)$ (thus giving Statistician a start of $a\|\theta\|_2^2$ on expert θ), where $a > 0$ is a constant reflecting our prior expectations about the "complexity" $\|\theta\|_2 = \sqrt{\sum_{i=1}^n \theta_i^2}$ of successful experts.

This idea of giving a start to Statistician allows us to prove stronger results; eg, the following elaboration of (19) holds:

$$L_T(\mathrm{AAR}) \le \inf_\theta \left(L_T(\theta) + \|\theta\|_2^2\right) + n \ln(T+1) \tag{20}$$

(this inequality still assumes that $\|x_t\|_\infty \le 1$ and $|y_t| \le 1$ for all t, but θ is unbounded).

Theorem 1 *For any positive integer n and any $a > 0$,*

$$\begin{aligned} L_T(\mathrm{AAR}(n,a)) &\le \inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2\right) + Y^2 \ln\det\left(I + \frac{1}{a}\sum_{t=1}^T x_t x_t'\right) \\ &\le \inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2\right) + Y^2 \sum_{i=1}^n \ln\left(1 + \frac{1}{a}\sum_{t=1}^T x_{t,i}^2\right). \end{aligned}$$

If, in addition, $\|x_t\|_\infty \le X$, $\forall t$,

$$L_T(\mathrm{AAR}(n,a)) \le \inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2\right) + nY^2 \ln\left(\frac{TX^2}{a} + 1\right). \tag{21}$$

The last inequality of this theorem implies inequalities (20) (it suffices to put $X = Y = a = 1$) and (2) (put $a = X^2$).

To interpret the term

$$\ln\det\left(I + \frac{1}{a}\sum_{t=1}^T x_t x_t'\right)$$

in Theorem 1, notice that it can be rewritten as

$$n \ln T + \ln\det\left(\frac{1}{T} I + \frac{1}{a}\mathrm{cov}(X_1, \dots, X_n)\right),$$

where $\mathrm{cov}(X_1, \dots, X_n)$ is the empirical covariance matrix of the predictor variables (in other words, $\mathrm{cov}(X_1, \dots, X_n)$ is the covariance matrix of the random vector which takes the values $x_1, \dots, x_T$ with equal probability $\frac{1}{T}$). We can see that this term is typically close to $n \ln T$.

3.3 Lower bounds

The following simple argument (sketched in Subsection 1.3 above) shows that bound (19) is tight in a certain (rather weak) sense. Let $n = 1$, $\theta \in [-1, 1]$ and, for all t, $x_t = 1$ and $y_t \in \{-1, 1\}$ (ie, we are interested in the problem of classification in the absence of any signals from Nature). Suppose the data are generated stochastically by the iid process with probability $\frac{1+\theta}{2}$ of $y_t = 1$ and probability $\frac{1-\theta}{2}$ of $y_t = -1$. (Therefore, we are essentially assuming the Bernoulli model.) A good estimate of θ after trial T is the Maximum Likelihood Estimate (MLE): $p_T = \frac{1}{T}\sum_{t=1}^{T} y_t$ (with, say, $p_0 = 0$). The expected loss of this estimator at trial $T + 1$ is

$$\mathbf{E}\left((p_T - y_{T+1})^2\right) = \mathbf{E}\left((p_T - \theta)^2\right) + \mathbf{E}\left((\theta - y_{T+1})^2\right) = \frac{1}{T}(1-\theta^2) + (1-\theta^2)$$

(when $T = 0$, $\frac{1}{T}(1-\theta^2)$ should be replaced by θ^2); therefore, its cumulative loss over the first T trials is

$$\theta^2 + (1-\theta^2)\sum_{t=1}^{T-1}\frac{1}{t} + T(1-\theta^2)$$

$$\geq \theta^2 + (1-\theta^2)\int_1^T \frac{1}{t} + T(1-\theta^2) = \theta^2 + (1-\theta^2)\ln T + T(1-\theta^2).$$

On the other hand, the expected loss of the best expert over the first T trials is

$$\mathbf{E}\sum_{t=1}^{T}\left(y_t - \frac{Y_T}{T}\right)^2$$

$$= \sum_{t=1}^{T}\mathbf{E}(y_t-\theta)^2 - 2\sum_{t=1}^{T}\mathbf{E}\,(y_t-\theta)\left(\frac{Y_T}{T}-\theta\right) + \sum_{t=1}^{T}\mathbf{E}\left(\frac{Y_T}{T}-\theta\right)^2$$

$$= T(1-\theta^2) - 2T\frac{1}{T}(1-\theta^2) + T\frac{1}{T^2}T(1-\theta^2) = T(1-\theta^2) - (1-\theta^2),$$

where $Y_T = \sum_{t=1}^{T} y_t$. We can see that

$$\mathbf{E}(L_T(\mathrm{MLE}) - L_T^*) \geq (1-\theta^2) + \theta^2 + (1-\theta^2)\ln T = 1 + (1-\theta^2)\ln T,$$

L_T^* being the loss of the best expert after T trials. For $\theta = 0$ we obtain

$$\mathbf{E}L_T(\mathrm{MLE}) \geq \mathbf{E}L_T^* + \ln T + 1,$$

so the MLE cannot much improve on (19) even on the average; the improvement of at most $\ln(T+1) - \ln T \leq \frac{1}{T}$ is minute. We can see that even in the case where Nature is not adversarial but instead follows a simple stochastic strategy, inequality (19) is tight.[2]

[2]Replacing the inequality

$$\sum_{t=1}^{T-1}\frac{1}{t} \geq \int_1^T \frac{1}{t} = \ln T$$

A weakness of this argument is that, despite the common belief that the MLE is a reasonable estimator in the case of the Bernoulli model, it is desirable to have some lower bounds applicable to *any* strategy for Statistician. The next theorem is a step in this direction.

Theorem 2 *Fix n (the number of attributes), Y (the upper bound on $|y_t|$) and $a > 0$. For any $\epsilon > 0$ there exist a constant C and a stochastic strategy for Nature such that $\|x_t\|_\infty = 1$ and $|y_t| = Y$, for all t, and, for any stochastic strategy for Statistician,*

$$\mathbf{E}\left(L_T(\text{Statistician}) - \inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2\right)\right) \geq (n-\epsilon)Y^2 \ln T - C, \quad \forall T$$

(cf inequality (21)).

In the proof of this theorem (see the next section) we will exhibit a suitable strategy for Nature (it will be just a mixture of the iid processes considered above with a beta prior).

3.4 Comparisons with Ridge Regression

The first method proposed for linear regression (by Gauss) was the Least Squares method; in the square-loss game, this means that we choose the best expert. Of course, the expert who has been the best in the past may perform badly in the future; to overcome this danger of overfitting, the RR procedure was proposed. In its simplest form, this procedure is just another implementation of the idea of giving Statistician a start on remote experts: as an estimate of θ we take the value where the inf in (21) is attained. The explicit formula is

$$\theta = \left(aI + \sum_{t=1}^T x_t x_t'\right)^{-1} \left(\sum_{t=1}^T y_t x_t\right)$$

(see Subsection 4.5 below), where a is a positive *ridge constant*; the Least Squares procedure corresponds to $a = 0$. Using this estimate to predict y_{T+1}, we obtain the following *RR prediction*:

$$\theta' x_{T+1} = \left(\sum_{t=1}^T y_t x_t\right)' \left(aI + \sum_{t=1}^T x_t x_t'\right)^{-1} x_{T+1}; \tag{22}$$

the on-line prediction algorithm using this formula for computing its predictions will be called the *RR procedure*. Notice that the RR procedure is very similar to the AAR; it can be obtained by swapping two lines in the AAR:

---

used above by the more accurate

$$\sum_{t=1}^{T-1} \frac{1}{t} \geq 1 + \int_2^T \frac{1}{t} = \ln T + 1 - \ln 2,$$

we can see that our worst-case bound for the AAR is actually better than the precise average-case bound for the MLE when $T > 2$.

$A = aI$; $b = 0$
FOR TRIAL $t = 1, 2, \ldots$:
  read new $x_t \in \mathbb{R}^n$
  output prediction $\gamma_t = b' A^{-1} x_t$
  $A = A + x_t x_t'$
  read new $y_t \in \mathbb{R}$
  $b = b + y_t x_t$
END FOR.

Correspondingly, the predictions output by the AAR are given by a formula very similar to (22):

$$\gamma_{T+1} = \left(\sum_{t=1}^{T} y_t x_t \right)' \left(aI + \sum_{t=1}^{T+1} x_t x_t' \right)^{-1} x_{T+1}.$$

It is easy to see that the RR procedure amounts to applying the Least Squares procedure (RR with $a = 0$) to an enlarged training set: we add n more observations, specifically $(x_{-n+1}, y_{-n+1}), \ldots, (x_0, y_0)$, where $(x_{-n+i}, y_{-n+i}) = (\sqrt{a} e_i, 0)$, e_i is the ith column of the unit $n \times n$ matrix, and $a > 0$ is the ridge constant. In the AAR, we add one more observation to our training set: namely, the observation $(x_{T+1}, 0)$, where x_{T+1} are the attributes of the example to be classified. Therefore, the AAR makes a further step towards shrinking the predictions to 0; hopefully, this extra shrinking will make the algorithm even more resistant to overfitting.

It is possible to deduce an explicit relation between the AAR's prediction γ_T and the RR's prediction r_T at trial T (this was first done by Kostas Skouras): by the Sherman-Morrison formula (40),

$$\gamma_T = b_{T-1}' A_T^{-1} x_T = b_{T-1}' A_{T-1}^{-1} x_T - b_{T-1}' \frac{\left(A_{T-1}^{-1} x_T \right) \left(A_{T-1}^{-1} x_T \right)'}{1 + x_T' A_{T-1}^{-1} x_T} x_T$$

$$= r_T - \frac{b_{T-1}' A_{T-1}^{-1} x_T x_T' A_{T-1}^{-1} x_T}{1 + x_T' A_{T-1}^{-1} x_T} = r_T \left(1 - \frac{x_T' A_{T-1}^{-1} x_T}{1 + x_T' A_{T-1}^{-1} x_T} \right) = \frac{r_T}{1 + x_T' A_{T-1}^{-1} x_T},$$

where the notations A_t and b_t are used in the sense of the AAR, ie $A_t = aI + \sum_{s=1}^{t} x_s x_s'$, $b_t = \sum_{s=1}^{t} y_s x_s$. Now we can see the nature of the dependence of γ_T on x_T: γ_T is a rational function, namely a linear function divided by a quadratic function. The limiting behaviour of γ_T and r_T is very different as $x_T \to \infty$: typically, $r_T \to \infty$, whereas always $\gamma_T \to 0$. Notice that, according to this formula, in the problem of classification (where the outcome space is $\{-1, 1\}$) the AAR gives the same categorical prediction, $+$ or $-$, as the RR procedure.

It can be seen that the RR sometimes gives results that are not sensible in our framework where $y_t \in [-Y, Y]$ and the goal is to compete against the best linear regression function. For example, suppose that $n = 1$, $Y = 1$, and Nature generates outcomes (x_t, y_t), $t = 1, 2, \ldots$, where

$$a \ll x_1 \ll x_2 \ll \ldots, \quad y_t = \begin{cases} 1, & \text{if } t \text{ odd}, \\ -1, & \text{if } t \text{ even}. \end{cases}$$

At trial $t = 2, 3, \ldots$ the RR (more accurately, its natural modification which truncates its predictions to $[-1, 1]$) will give prediction $\gamma_t = y_{t-1}$ equal to the previous response, and so will suffer a loss of about $4T$ over T trials. On the other hand, the AAR's prediction will be close to 0, and so the cumulative loss of the AAR over the first T trials will be about T, which is close to the best expert's loss. We can see that the RR procedure in this situation is forced to suffer a loss 4 times as big as the AA's loss. (It is interesting that the same constant 4 occurs in Theorem 4 below; cf Theorem 1.)

The lower bound proven in Subsection 3.3 does not imply that our regression algorithm is better than the RR in our adversarial framework. (Moreover, the idea of the proof of Theorem 2 given in the next section is to lower bound the performance of the RR in the situation where the expected loss of the RR is optimal.) Theorem 1 asserts that

$$L_T(\text{Statistician}) \le \inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2 \right) + Y^2 \sum_{i=1}^n \ln \left(1 + \frac{1}{a} \sum_{t=1}^T x_{t,i}^2 \right) \tag{23}$$

when Statistician follows the AAR. The next theorem shows that the RR sometimes violates this inequality.

Theorem 3 *Let $n = 1$ (the number of attributes) and $Y = 1$ (the upper bound on $|y_t|$); fix $a > 0$. Nature has a strategy such that, when Statistician follows* $\mathrm{RR}(n, a)$,

$$L_T(\text{Statistician}) = 4T + o(T), \tag{24}$$

$$\inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2 \right) \le T, \tag{25}$$

$$\ln \left(1 + \frac{1}{a} \sum_{t=1}^T x_t^2 \right) = 2T \ln 2 + O(1) \tag{26}$$

as $T \to \infty$ (and, therefore, (23) is violated).

Nature's strategy that ensures the inequality of Theorem 3 is very simple: she generates, for $t = 1, 2, \ldots$,

$$x_t = 2^{t-1}, \quad y_t = \begin{cases} 1, & \text{if } t \text{ is odd}, \\ -1, & \text{if even}. \end{cases}$$

The RR prediction can be obtained as the mean of the posterior distribution (as proven in Subsection 4.5), whereas the AAR's prediction is obtained by using a more sophisticated substitution function. Let RR^Y be the modification of the RR procedure in which the experts' predictions $\theta \cdot x_t$ at every trial t

are clipped to $[-Y, Y]$ before averaging. The properties of the posterior mean as a substitution function have been studied by Kivinen and Warmuth [41]; their result (reproduced in Remark 1 above) immediately implies the following theorem (for details see the next section).

Theorem 4 *For any fixed n, $Y > 0$ and $a > 0$,*

$$
\begin{aligned}
L_T(\mathrm{RR}^Y(n,a)) &\le \inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2\right) + 4Y^2 \ln\det\left(I + \frac{1}{a}\sum_{t=1}^T x_t x_t'\right) \\
&\le \inf_\theta \left(L_T(\theta) + a\|\theta\|_2^2\right) + 4Y^2 \sum_{i=1}^n \ln\left(1 + \frac{1}{a}\sum_{t=1}^T x_{t,i}^2\right),
\end{aligned}
$$

provided all y_t belong to $[-Y, Y]$.

It is not clear if this remains true when RR^Y is replaced with RR (but this is not a particularly interesting question in view of the better bound of Theorem 1).

4 Regression: proofs

4.1 Derivation of the AAR

Let $a > 0$ be an arbitrary constant. Consider the prior distribution P_0 in the set $\mathbb{R}^n$ of possible weights θ with the Gaussian density

$$(a\eta/\pi)^{n/2} e^{-a\eta\|\theta\|_2^2} d\theta \tag{27}$$

(the value of the normalizing constant is given in, eg, [5], Chap 2, Theorem 3). The loss of expert θ over the first T trials is

$$\sum_{t=1}^T (y_t - x_t'\theta)^2 = \theta'\left(\sum_{t=1}^T x_t x_t'\right)\theta - 2\left(\sum_{t=1}^T y_t x_t'\right)\theta + \sum_{t=1}^T y_t^2. \tag{28}$$

Therefore, the loss of the APA is

$$
\begin{gathered}
\log_\beta \int_{\mathbb{R}^n} d\theta (a\eta/\pi)^{n/2} \\
\times \exp\left(-\eta\theta'\left(aI + \textstyle\sum_{t=1}^T x_t x_t'\right)\theta + 2\eta\left(\sum_{t=1}^T y_t x_t'\right)\theta - \eta\sum_{t=1}^T y_t^2\right).
\end{gathered} \tag{29}
$$

In accordance with (18) and (29), the AA's prediction should be[3]

$$
\begin{aligned}
\gamma_T &= \frac{1}{4Y}\log_\beta \frac{\beta^{g_T(-Y)}}{\beta^{g_T(Y)}} \\
&= \frac{1}{4Y}\log_\beta \frac{\int_{\mathbb{R}^n} e^{-\eta\theta'\left(aI + \sum_{t=1}^T x_t x_t'\right)\theta + 2\eta\left(\sum_{t=1}^{T-1} y_t x_t' - Y x_T'\right)\theta - \eta\left(\sum_{t=1}^{T-1} y_t^2 + Y^2\right)} d\theta}{\int_{\mathbb{R}^n} e^{-\eta\theta'\left(aI + \sum_{t=1}^T x_t x_t'\right)\theta + 2\eta\left(\sum_{t=1}^{T-1} y_t x_t' + Y x_T'\right)\theta - \eta\left(\sum_{t=1}^{T-1} y_t^2 + Y^2\right)} d\theta}
\end{aligned}
$$

[3] See below for the explanation of some steps in this derivation.

$$= \frac{1}{4Y}\log_\beta \frac{\int_{\mathbb{R}^n} e^{-\eta\theta'\left(aI+\sum_{t=1}^T x_t x_t'\right)\theta+2\eta\left(\sum_{t=1}^{T-1} y_t x_t' - Y x_T'\right)\theta} d\theta}{\int_{\mathbb{R}^n} e^{-\eta\theta'\left(aI+\sum_{t=1}^T x_t x_t'\right)\theta+2\eta\left(\sum_{t=1}^{T-1} y_t x_t' + Y x_T'\right)\theta} d\theta} \tag{30}$$

$$= \frac{1}{4Y}\log_\beta e^{-\eta F\left(aI+\sum_{t=1}^T x_t x_t', -2\sum_{t=1}^{T-1} y_t x_t', 2Y x_T'\right)} \tag{31}$$

$$= \frac{1}{4Y} F\left(aI + \sum_{t=1}^T x_t x_t', -2\sum_{t=1}^{T-1} y_t x_t', 2Y x_T'\right)$$

$$= \left(\sum_{t=1}^{T-1} y_t x_t'\right)\left(aI + \sum_{t=1}^T x_t x_t'\right)^{-1} x_T \tag{32}$$

(notice that η and Y disappeared), where we used the notation

$$F(A,b,x) = \inf_{\theta\in\mathbb{R}^n}\left(\theta' A\theta + b'\theta + x'\theta\right) - \inf_{\theta\in\mathbb{R}^n}\left(\theta' A\theta + b'\theta - x'\theta\right)$$

and the fact that the function F can be transformed as follows:

$$F(A,b,x) = \inf_{\theta\in\mathbb{R}^n}\left(\theta' A\theta + b'\theta + x'\theta\right) - \inf_{\theta\in\mathbb{R}^n}\left(\theta' A\theta + b'\theta - x'\theta\right)$$

$$= -\frac{1}{4}(b+x)'A^{-1}(b+x) + \frac{1}{4}(b-x)'A^{-1}(b-x) = -b'A^{-1}x.$$

To see why the transition from (30) to (31) is justified, notice that a *horizontal translation*

$$(\theta+c)'A(\theta+c) = \theta'A\theta + 2c'A\theta + c'Ac$$

of a positive definite quadratic form $\theta'A\theta$ can have an arbitrary vector coefficient $2c'A$ in front of θ; therefore, the quadratic forms in the numerator and denominator of (30) can be obtained from each other by horizontal translation and adding a constant. It is clear that the ratio of the integrals in (30) depends only on this additive constant, which is written in the form $F(\ldots)$ in (31).

4.2 Upper bounds: Proof of Theorem 1

The proof of the first inequality in Theorem 1 (whose derivation from known properties of the general AA does not present any difficulties) will be given in the next paragraph; the second inequality follows from, eg, [5], Chap 2, Theorem 7; the third inequality is trivial.

Set $\eta = \frac{1}{2Y^2}$ (cf Subsection 2.4). Let the maximum of the expression under the exp sign in (29) be attained at a point $\theta = \theta_0$ (this is also the point where the minimum is attained in the statement of the theorem). Subtracting the "regularized" loss

$$\theta_0'\left(aI + \sum_{t=1}^T x_t x_t'\right)\theta_0 - 2\left(\sum_{t=1}^T y_t x_t'\right)\theta_0 + \sum_{t=1}^T y_t^2$$

of expert θ_0 (see (28)) from (29), we obtain

$$\log_\beta \int_{\mathbb{R}^n} d\theta (a\eta/\pi)^{n/2} e^{-\eta B(\theta)}, \tag{33}$$

where $B(\theta)$ is a translation (ie, horizontal translation plus a constant) of the quadratic form

$$\theta' \left(aI + \sum_{t=1}^{T} x_t x_t' \right) \theta; \tag{34}$$

since $\inf_\theta B(\theta) = 0$, $B(\theta)$ is a horizontal translation of (34) (cf the last paragraph of the previous section), and so (33) equals

$$\log_\beta \int_{\mathbb{R}^n} d\theta (a\eta/\pi)^{n/2} \exp\left(-\eta\theta' \left(aI + \sum_{t=1}^{T} x_t x_t' \right) \theta \right).$$

Directly evaluating the integral (see, eg, [5], Chap 2, Theorem 3), we can transform this expression to

$$\log_\beta \left((a\eta/\pi)^{n/2} \frac{\pi^{n/2}}{\sqrt{\det\left(a\eta I + \eta \sum_{t=1}^{T} x_t x_t'\right)}} \right) = -\frac{1}{2} \log_\beta \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right)$$

$$= \frac{1}{2\eta} \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right) = Y^2 \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right).$$

4.3 Lower bounds I: Proof of Theorem 2

Our stochastic strategy for Nature will be such that $x_{t,i} \in \{0,1\}$, for all $t = 1, 2, \ldots$ and $i = 1, \ldots, n$. First we consider the 1D case ($n = 1$), where we will have $x_t = 1$, $\forall t$ (it will be easy to generalize to arbitrary n). Without loss of generality we will only consider the case of $Y = \frac{1}{2}$: the general case can be obtained by a simple rescaling of y_ts and θ. It is clear that instead of assuming $y_t \in \{-\frac{1}{2}, \frac{1}{2}\}$ we can, and will, assume $y_t \in \{0,1\}$ (since $x_t = 1$, we can shift θ and y_t by $\frac{1}{2}$); we will see that the best value of θ will satisfy $\theta \in [0,1]$.

Nature's strategy is as follows: first she generates $p \in [0,1]$ from the beta distribution with parameters (A, A), where A is a large positive constant; at trial t she sets $y_t = 1$ with probability p and $y_t = 0$ with probability $1 - p$, independently of the previous trials.

Since Nature's strategy is known, it is easy to find the best, on the average, strategy for Statistician (the Bayesian strategy). The prior density is proportional to

$$p^{A-1}(1-p)^{A-1};$$

therefore, taking k to denote the number of 1s among the first T trials, we obtain that the posterior density is proportional to

$$p^{k+A-1}(1-p)^{T-k+A-1}$$

after the first T trials. Since Nature generates 1 at trial $T+1$ with probability

$$\int_0^1 \frac{p^{k+A}(1-p)^{T-k+A-1}}{B(k+A, T-k+A)} dp = \frac{B(k+A+1, T-k+A)}{B(k+A, T-k+A)}$$

$$= \frac{\Gamma(k+A+1)\Gamma(T-k+A)/\Gamma(T+2A+1)}{\Gamma(k+A)\Gamma(T-k+A)/\Gamma(T+2A)} = \frac{k+A}{T+2A}$$

(conditional on what has happened in the first T trials) and square loss is a proper scoring rule (see, eg, Dawid [20]), the Bayesian strategy for Statistician will output

$$p_T = \frac{k+A}{T+2A}$$

at trial $T+1$.

For a fixed p, the expected loss of the Bayesian strategy at trial $T+1$ is

$$\mathbf{E}_p\left(p_T - y_{T+1}\right)^2 = \mathbf{E}_p\left(p_T - p\right)^2 + \mathbf{E}_p\left(y_{T+1} - p\right)^2$$

$$= \mathbf{E}_p\left(\frac{k+A}{T+2A} - p\right)^2 + p(1-p)$$

$$= \mathbf{E}_p\left(\frac{k+A}{T+2A} - \frac{pT+A}{T+2A}\right)^2 + \left(p - \frac{pT+A}{T+2A}\right)^2 + p(1-p)$$

$$= \mathbf{E}_p\left(\frac{k-pT}{T+2A}\right)^2 + \left(\frac{2Ap-A}{T+2A}\right)^2 + p(1-p)$$

$$= \frac{Tp(1-p)}{(T+2A)^2} + \left(\frac{2Ap-A}{T+2A}\right)^2 + p(1-p);$$

the first equality in this chain follows from the fact that it holds conditionally on the first T trials and the third equality follows from

$$\mathbf{E}_p\frac{k+A}{T+2A} = \frac{pT+A}{T+2A}.$$

Therefore, the expected value (for p fixed) of the loss of the Bayesian strategy over the first T trials is

$$\mathbf{E}_p L_T(\text{Statistician}) = \sum_{t=0}^{T-1}\left(\frac{tp(1-p)}{(t+2A)^2} + \left(\frac{2Ap-A}{t+2A}\right)^2 + p(1-p)\right)$$

$$= p(1-p)T + p(1-p)\ln T + O(1).$$

Averaging over the prior distribution and using

$$\int p(1-p)\frac{p^{A-1}(1-p)^{A-1}}{B(A,A)}dp = \frac{B(A+1,A+1)}{B(A,A)} = \frac{A^2}{(2A+1)(2A)} = \frac{A}{4A+2}, \tag{35}$$

we find

$$\mathbf{E}L_T(\text{Statistician}) = \frac{A}{4A+2}T + \frac{A}{4A+2}\ln T + O(1). \tag{36}$$

Now we find the expected cumulative loss over the first T trials, for a fixed p, of the best expert $\theta = k/T$:

$$\mathbf{E}_p\sum_{t=1}^{T}\left(y_t - \frac{k}{T}\right)^2 = \mathbf{E}_p\sum_{t=1}^{T}y_t^2 - 2\mathbf{E}_p\sum_{t=1}^{T}y_t\frac{k}{T} + \mathbf{E}_p T(k/T)^2$$

$$= Tp - \frac{2}{T}\sum_{t,s}\mathbf{E}_p(y_t y_s) + \frac{1}{T}\Big((pT)^2 + p(1-p)T\Big)$$
$$= Tp - \frac{2}{T}(T^2 - T)p^2 - \frac{2}{T}Tp + \frac{1}{T}\Big((pT)^2 + p(1-p)T\Big)$$
$$= T\left(p - 2p^2 + p^2\right) + O(1) = p(1-p)T + O(1).$$

Again using (35), we further obtain

$$\mathbf{E}\inf_\theta L_T(\theta) = \mathbf{E}\sum_{t=1}^{T}\left(y_t - \frac{k}{T}\right)^2 = \frac{A}{4A+2}T + O(1). \tag{37}$$

Now the statement of Theorem 2 for $n = 1$ immediately follows from comparison of (36) and (37): for any $\epsilon > 0$ we will be able to take A large enough.

In the case of $n > 1$ the following stochastic strategy for Nature will satisfy the requirement of Theorem 2. She starts by independently generating n numbers $p_i \in [0,1]$, $i = 1, \ldots, n$, from the beta distribution with parameters (A, A); at trial t she sets $x_t = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the only 1 is the ith component of the vector x_t and

$$i = \begin{cases} n, & \text{if } t \bmod n = 0, \\ t \bmod n, & \text{otherwise,} \end{cases}$$

$y_t = Y$ with probability p_i and $y_t = -Y$ with probability $1 - p_i$, independently of the previous trials. This construction implies a subtrahend of anY^2 in the right-hand side of the inequality in the statement of the theorem (anY^2 is an upper bound on $a\|\theta\|_2^2$), but we included this subtrahend in the constant C.

4.4 Lower bounds II: Proof of Theorem 3

In this subsection we will prove Theorem 3; Nature will play the strategy described after the statement of the theorem. First let us prove that $\gamma_T - y_{T-1} \to 0$ as $T \to \infty$ (that is, Statistician's prediction is almost the repetition of the last observed response). After trial $T \gg 1$ the observed (x_t, y_t) are

$$(1,1), (c,-1), (c^2,1), \ldots \tag{38}$$

if we take x_T (resp y_T) to be our unit of length for measuring the xs (resp ys); here $c = \frac{1}{2}$. Let us imagine that the data set (38) is infinite; for large T, this will be a good approximation. The best least-squares approximation $y = \theta x$ to this data set is determined from the condition

$$(1-\theta)^2 + (1+c\theta)^2 + (1-c^2\theta)^2 + (1+c^3\theta)^2 + \ldots \to \min.$$

Differentiating this function in θ and summing the geometrical progressions, we obtain

$$\frac{1}{1-c^2}\theta = \frac{1}{1+c},$$

which gives $\theta = 1 - c$. After observing $x_{T+1} = \frac{1}{c}$ at the next trial, Statistician will give prediction

$$\gamma_{T+1} = \frac{\theta}{c} = \frac{1-c}{c} = 1.$$

This completes the proof of $\gamma_T - y_{T-1} \to 0$ $(T \to \infty)$.

Equations (24) and (25) are now obvious; (26) is also easy to check:

$$\ln\left(1 + \frac{1}{a}\sum_{t=1}^{T} x_t^2\right) = \ln\left(\sum_{t=1}^{T} 2^{2(t-1)}\right) + O(1) = 2T\ln 2 + O(1).$$

4.5 Derivation of the RR and proof of Theorem 4

Recall that the RR estimate was defined as the arg min in (21). First we "derive" the RR prediction showing that it is precisely the mean of the posterior distribution. (This is a well-known fact, but its proof is so simple that we give it here.)

Taking the same prior over θ as before, (27), and associating with the expert θ the probability forecasting system whose prediction at trial t is the Gaussian distribution with mean $x_t'\theta$ and variance $\frac{1}{2\eta}$, we obtain that the posterior distribution is proportional to

$$\exp\left(-a\eta\|\theta\|_2^2 - \eta L_T(\theta)\right) \propto \exp\left(-\eta\theta'\left(aI + \sum_{t=1}^{T} x_t x_t'\right)\theta + 2\eta\left(\sum_{t=1}^{T} y_t x_t'\right)\theta\right)$$

(cf (29)). The posterior mean equals the minimum of these expressions; minimizing the first expression, we obtain our original definition of the RR estimate; minimizing the second one,

$$\frac{\partial}{\partial\theta}\left(-\eta\theta'\left(aI + \sum_{t=1}^{T} x_t x_t'\right)\theta + 2\eta\left(\sum_{t=1}^{T} y_t x_t'\right)\theta\right) = 0,$$

we obtain the RR formula (22).

To prove Theorem 4, it suffices to take $\eta = \frac{1}{8Y^2}$ instead of $\eta = \frac{1}{2Y^2}$ in the proof in Subsection 4.2 (cf Remark 1).

5 Review of literature

The first paper on competitive on-line statistics was, probably, DeSantis et al [22], which performed a competitive analysis of the Bayesian mixing scheme for the log-loss prediction game. Later Littlestone and Warmuth [43] and Vovk [53] introduced an on-line algorithm (called the Weighted Majority Algorithm by the former authors) for the simple prediction game. These two algorithms (the Bayesian mixing scheme and the Weighted Majority Algorithm) were generalized by the AA, proposed in Vovk [52]. The AA has been proven to be optimal in some simple cases (see Haussler et al [33], Vovk [55], Watkins's theorem stated in [56]). The AA is a member of a wide family of algorithms called "multiplicative weight" or "exponential weight" algorithms.

We call this area "competitive on-line statistics" because of the closeness of its ideology to that of the theory of competitive on-line algorithms (see, eg, [46]). The adjective "competitive" refers to inequalities such as (1) and (2) above (or analogous inequalities for competitive on-line algorithms); in this approach, the goal of the algorithm to perform as well as the best, retrospectively, decision strategy. The adjective "on-line" means that the data ω_t (or (x_t, y_t)) arrive sequentially; this is more or less a computer-science analogue of "prequential" in "prequential statistics".

In the rest of this section we will give a brief review of several selected subareas of competitive on-line statistics (those which are especially closely connected with the material presented in this paper) and, in conclusion, discuss some connections between competitive on-line statistics and more traditional statistics and information theory. Many exciting developments and applications will not be discussed; the interested reader can consult the literature on sleeping experts and World-wide Web applications (see, eg, Freund et al [31], Cohen and Singer [16]), boosting, game theory and linear programming (Freund and Schapire [30, 28, 29]), pruning decision trees (Helmbold and Schapire [35], Pereira and Singer [44], Takimoto et al [49]), etc.

5.1 Competitive on-line regression

First we review the subarea of competitive on-line statistics most closely connected to the rest of this paper, regression with the square loss function.

The competitive approach to regression started, probably, from Foster [26], who considered a variant of the RR procedure. In particular, his Theorem 1 asserts that, in the situation where the response variable takes values in $\{0, 1\}$, the examples x_t belong to $[0, 1]^n$ and the weights are nonnegative and satisfy $\|\theta\|_1 = 1$,

$$L_T \leq L_T^* + n \ln(n(T + 1)) + 2;$$

this result is very similar to inequality (19). No lower bounds are given in [26].

Cesa-Bianchi et al [12] performed, under the square loss, a competitive analysis of the standard Gradient Descent Algorithm and Kivinen and Warmuth [40] complemented it by a competitive analysis of a modification of the Gradient Descent, which they call the Exponentiated Gradient Algorithm. The bounds obtained by Cesa-Bianchi et al [12] and Kivinen and Warmuth [40] are of the following type: at every trial T,

$$L_T \leq c L_T^* + O(1), \tag{39}$$

where L_T is the loss (over the first T trials) of the on-line algorithm, L_T^* is the loss of the best (by trial T) linear regression function, and c is a constant,

$c > 1$; specifically, $c = 2.25$ for the Gradient Descent and $c = 3$ for the Exponentiated Gradient. (See though Remark 4 below.) These bounds hold under the following assumptions: for the Gradient Descent, it is assumed that the L_2 norm of the weights and of all data items are bounded by constant 1; for the Exponentiated Gradient, that the L_1 norm of the weights and the L_∞ norm of all data items are bounded by 1.

In many interesting cases bound (39) might appear weak. For example, suppose that our comparison class contains a "true" regression function, but its values are corrupted by Gaussian noise (that is, we assume that $y_t = \theta \cdot x_t + \xi_t$, where ξ_t are independent $\mathcal{N}(0, \sigma^2)$ random variables, $\sigma^2 > 0$); we also assume that $\|\theta\|_1 \le 1$ and $\|x_t\|_\infty \le 1$ for all t. The performance of the AAR is given by (19); therefore, the difference $L_T - L_T^*$ is bounded by a logarithmic function of T. As concerns (39), L_T^* will grow linearly in T, and inequality (39) will only bound the difference $L_T - L_T^*$ by a linear function of T. However, the bounds proven in [12, 40] have some important advantages before our bounds (such as (19)); for example:

- $O(1)$ in (39) depends on the number of parameters n logarithmically in the case of the Exponentiated Gradient Algorithm and does not depend on n at all in the case of Gradient Descent, whereas our bound depends on n linearly.
- Our bounds are not as good as, say, the bound

  $$L_T(\mathrm{GD}) \le 2.25 \inf_\theta \left(L_T(\theta) + X\|\theta\|_2^2 \right)$$

  ([12], Theorem IV.1; the only assumption here is that all $\|x_t\|_2 \le X$) in the noise-free case where $L_T(\theta) = 0$ for some θ.
- The most naive implementation of the AAR would require $O(n^3)$ arithmetic operations at every trial. It is easy to see that there are ways to implement it more efficiently; for example, we can use the value A_{t-1}^{-1} of the inverse to A at trial $t-1$ to compute its value A_t^{-1} at trial t: by the Sherman-Morrison formula ([45], Section 2.10),

  $$A_t^{-1} = A_{t-1}^{-1} - \frac{\left(A_{t-1}^{-1} x_t\right)\left(A_{t-1}^{-1} x_t\right)'}{1 + x_t' A_{t-1}^{-1} x_t}. \tag{40}$$

  Therefore, we only need $O(n^2)$ arithmetic operations per trial. This is still not as good as for the Gradient Descent Algorithm and Exponentiated Gradient Algorithm: they require only $O(n)$ operations per trial.

It remains to be seen if it is possible to combine the advantages of the AAR, Gradient Descent and Exponentiated Gradient in one algorithm. A step in this direction is done by Azoury and Warmuth [3], who obtained the bound of our Theorem 1 (see Theorem 3.5 in [3]) for an algorithm motivated by the methodology of [12] and [40]. Another approach might be to try the AA with a different prior.

Remark 3 Let us formally verify the above claim that L_T^* grows linearly in T. Our model is $Y = X\theta + \xi$, where $Y = (y_1, \ldots, y_T)'$ is the observed sequence of responses and $X = (x_1, \ldots, x_T)'$ is the observed sequence of signals. Without loss of generality we assume that the true value of θ is 0; for simplicity, we consider the 1D case, $n = 1$. The least squares estimate of θ (ie, the best expert) is

$$\hat{\theta} = (X'X)^{-1}X'Y$$

(cf (22)), and so the residual sum of squares is

$$Y'Y - Y'X\hat{\theta} = Y'Y - Y'X(X'X)^{-1}X'Y = \xi'\xi - \xi'X(X'X)^{-1}X'\xi$$

$$= \sum_{t=1}^{T} \xi_t^2 - \frac{\left(\sum_{t=1}^{T} \xi_t x_t\right)^2}{\sum_{t=1}^{T} x_t^2}.$$

The minuend of the final expression is the loss suffered by the true regression function, which grows as $\sigma^2 T$ (by the law of large numbers); the subtrahend is much smaller by Chebyshev's inequality: it is non-negative and

$$\mathbf{E}\frac{\left(\sum_{t=1}^{T} \xi_t x_t\right)^2}{\sum_{t=1}^{T} x_t^2} = \sigma^2.$$

Notice that the Gaussian noise model violates the assumption $|y_t| \le Y$, but this does not matter since we know in advance that $\|x_t\|_\infty \le 1$ and $\|\theta\|_1 \le 1$.

Remark 4 Cesa-Bianchi et al [12] and Kivinen and Warmuth [40] also construct some "semi-on-line" algorithms; such algorithms must be *a priori* given a good upper estimate K on the loss of the best regression function. For such algorithms, they derive bounds of the type $L_T \le L_T^* + O(\sqrt{K})$. Using the usual "doubling trick" (Cesa-Bianchi et al [10]), it is possible to obtain bounds of the type $L_T \le L_T^* + O(\sqrt{L_T^*})$ in the pure on-line setting (see Cesa-Bianchi et al [12], Theorem IV.4). In our example of a true regression function corrupted by Gaussian noise, the difference $L_T - L_T^*$ will be bounded by a linear function of $\sqrt{T}$, which is still worse than (19) (even though the AA is purely on-line and does not require *a priori* estimates of L_T^*).

Yamanishi [64] analyzed the application of the AA to a wide class of decisions pools and obtained loss bounds analogous to (19) in a very general setting. One of the decision pools (Examples 1 and 9 of [64]) is the following: the weights θ and examples x_t are known to belong to the unit L_2-ball in $\mathbb{R}^n$ and all their components are nonnegative; responses y_t are known to belong to $[0, 1]$. Taking the uniform distribution as the prior for the AA, Yamanishi obtained the following analogue of our inequality (19):

$$L_T(\mathrm{AA}) \le \inf_\theta (L_T(\theta)) + \frac{n}{4} \ln \frac{2Tn}{\pi} + \frac{1}{2} \ln \frac{\pi^{n/2}}{\Gamma(1 + n/2)} + o(1). \qquad (41)$$

Comparing this to (21) with $Y = \frac{1}{2}$ (recall that Yamanishi assumes $y_t \in [0, 1]$), we can see that the coefficients before the leading term $\ln T$ match; it should be borne in mind that (41) is a special case of a much more general result (Theorem 2 in [64]; it first appeared in Yamanishi [63]). In the more recent paper [65] Yamanishi proves non-asymptotic results of the same kind and proves lower bounds matching his upper bounds.

Forster [25] has discovered a surprising property of the AAR predictions: according to his Theorem 1, they minimize the maximal extra loss, compared to the best expert, that might be suffered in trial T. He also gives a simplified proof of our Theorem 1.

Feder and Singer [24] obtain results very similar to ours in the case of linear autoregression with square loss. In the case where the order p of autoregression is known, they obtain the bound

$$L_T(\text{Universal Algorithm}) \le \inf_\theta L_T(\theta) + 4A^2 p \ln T + O(1) \tag{42}$$

(this bound follows from Theorem 2 in [24]; Theorem 1 has a similar form), where A is the *a priori* bound on the values taken by y_t (which is to be predicted using the values $y_{t-1}, \dots, y_{t-p}$) and $O(1)$ stays bounded as $T \to \infty$. Notice that (42) involves an extra factor of 4 (familiar from Subsection 3.4) before $\ln T$ as compared to (21). This must have happened because they used the posterior mean as a substitution function (see the discussion in Subsection 3.4) rather than following the more opportunistic approach of the AA.

A distinctive feature of the competitive approach to regression is that one does not make any assumptions about stochastic properties of the data-generating mechanism; for example, in Section 3 our only assumption about the data was that $|y_t| \le Y$, $\forall t$. In some situations (if the data were generated by a partially known stochastic mechanism) this feature is a disadvantage, but often it will be an advantage.

The assumption that y_t are bounded is natural in many applications (the problem of classification, in which always $y_t \in \{-1, 1\}$, predicting a student's score ranging from 0% to 100%, etc), but in some other applications it might look artificial. It would be very interesting to relax or remove this assumption; some ideas on "tuning" the lower and upper bounds a and b can be found in Vovk and Gammerman [57].

5.2 Tracking the best expert

So far we have been interested in competing against the best linear regressor $y = \theta \cdot x$; the components of θ are some fixed weights. Paper [37] by Herbster and Warmuth considers the case where θ is allowed to (slowly) change over time. This is a special case of a more general problem of tracking the best expert, which originated in Littlestone and Warmuth [43] and was further studied by, among others, Herbster and Warmuth [36], Auer and Warmuth [2], Vovk [56]. In the problem of tracking the best expert we start with a pool of "basic" decision strategies. It is known that the strategies in the pool are too inflexible to attain a good performance, but we expect that first one of the strategies will work well, then another, etc, hoping that we will only rarely have to switch between the strategies. Two main approaches to the problem of tracking the best expert are:

- ◇ We can try to modify the known merging algorithms to make them more adaptive; this approach was used in, eg, [36].
- ◇ We can construct the pool of "superstrategies" (an example of a superstrategy: start in the morning following strategy 3, at 1:15 pm switch to strategy 1 and at 5 pm switch to strategy 2) and then apply the AA or another "static" merging algorithm; this approach was used in [56].

Typical bounds for the total on-line loss of a merging algorithm tracking the best expert are the sum of the loss of the best "switching schedule" between the basic strategies plus the total cost for switching between the experts (say, a constant times the number of switches) plus a small overhead (such as the logarithm of the number of experts). The work on tracking the best expert has been applied to predicting disk idle times [34] and load balancing problems [6]; there is no doubt that this work will find further important applications.

5.3 Bandit problems and reinforcement learning

We already mentioned that the philosophy of competitive on-line statistics, as described above, only "works" when Nature is oblivious to Statistician's decisions (ω_T does not depend on $\gamma_1, \ldots, \gamma_{T-1}$): if Nature is not oblivious, it is not longer possible to interpret inequality (1) as saying that Statistician performs not much worse than the best of the decision strategies: if we had followed strategy θ we would perhaps have observed a different sequence of outcomes. The assumption that Nature is oblivious is always justified when γ_t are predictions (say, the atmosphere does not care about our predicting rain), but it can also be justified for decisions different from predictions, such as portfolio selection by a small investor (see Subsection 2.5).

The situation is changing now: eg, papers [1] and [9] present competitive on-line results for bandit problems, in which Nature by no means is oblivious. It turns out that it is even possible to study the exploration—exploitation tradeoff in the competitive on-line framework.

5.4 Predictive complexity

A recent application of the Aggregating Algorithm is to generalize the notion of Kolmogorov complexity (see, eg, Li and Vitanyi [42]). The idea is to apply the AA (or some other merging algorithm) to the "universal decision pool" containing every computable decision strategy (such a pool can be constructed from a universal Turing machine). The loss of the resulting "decision strategy" (actually it will be a decision strategy only in a generalized sense) on a data sequence x is called the *predictive complexity* of x. When applied to the log-loss game, this leads to a variant of Kolmogorov complexity. As well as being a fundamental concept *per se*, the notion of predictive complexity allows us to define the notion of randomness for prediction games different from the log-loss game; for details, see [50]. (Though even the standard notion of log-loss randomness seems to be grossly under-used presently: see [60].) Another application is to generalizing the MDL principle to games different from the log-loss one: see [57]. For further information about predictive complexity the reader can consult Kalnichkan [39, 38] and Vovk and Watkins [58].

5.5 Statistics and information theory

It is clear from Subsection 2.2 that the log loss function plays a fundamental role in probability theory and statistics (esp Bayesian statistics); the role of the log loss function is as prominent in information theory, where it is interpreted as the code length[4]. To make the connections between the usual statistical notions, such as log-likelihood, maximum likelihood estimate (MLE) and observed Fisher information, and competitive on-line statistics clearer, we will explain how they can be generalized to different loss functions.

Let μ be a prior probability density in Θ and $\eta > 0$ be some learning rate; as usual, we set $\beta = e^{-\eta}$. As described earlier, when Statistician can predict with "pseudo-predictions", we will have at each trial

$$L_T(\text{Statistician}) = \log_\beta \int \beta^{L_T(\theta)} \mu(\theta) d\theta.$$

When Statistician has to make permitted predictions $\gamma \in \Gamma$, we will instead obtain

$$L_T(\text{Statistician}) \leq c(\eta) \log_\beta \int \beta^{L_T(\theta)} \mu(\theta) d\theta. \tag{43}$$

The function $L(\theta) = L_T(\theta)$ is a generalization of the minus log-likelihood function in statistics. Let us for simplicity consider only the 1D case where $\Theta \subseteq \mathbb{R}$. The two most important characteristics of the "minus log-likelihood function" $L(\theta)$ are the maximum likelihood estimate

$$\hat\theta = \arg\min_\theta L(\theta)$$

and the observed Fisher information

$$\hat\jmath = L''(\hat\theta)$$

(the usual Fisher information is the expectation of the observed Fisher information). The functions $\hat\theta$ and $\hat\jmath$ have a long history in statistics; eg, it was Fisher's idea (Dawid [21], Section 6) that to make the maximum likelihood estimator $\hat\theta$ sufficient it should be complemented by the "ancillary statistic" $\hat\jmath$.

[4]This is true for lossless compression; the competitive on-line theory of lossy compression would require more sophisticated loss functions.

Let us simplify (43) under "regularity conditions" (without specifying them precisely). Function $\beta^{L(\theta)}$ rapidly vanishes outside the interval centered at $\hat{\theta}$ of length about $1/\sqrt{\hat{\jmath}}$ (because

$$L(\theta) - L(\hat{\theta}) \approx \frac{1}{2}\hat{\jmath}(\theta - \hat{\theta})^2$$

near $\hat{\theta}$), but inside that interval it is close to $\beta^{L(\hat{\theta})}$. So we obtain from (43):

$$L(\text{Statistician}) \leq c(\eta) \log_\beta \left(\beta^{L(\hat{\theta})} \mu(\hat{\theta}) / \sqrt{\hat{\jmath}}\right) + \text{const},$$

ie,

$$L(\text{Statistician}) \leq c(\eta) L(\hat{\theta}) + a(\eta) \ln \frac{\sqrt{\hat{\jmath}}}{\mu(\hat{\theta})} + \text{const}. \tag{44}$$

(Using Laplace's approximation, the constant const can be evaluated to $a(\eta) \ln \sqrt{2\pi}$.)

In many cases $\hat{\jmath}$ will be of the order of magnitude T, the ordinal number of the trial. Therefore, (44) gives an $O(\log T)$ "overhead" for perfectly mixable games; a similar logarithmic term, $n \ln(T+1)$, occurs in (19). These terms are analogous to the term $\frac{n}{2} \ln T$ occurring in the analysis of the log-loss game and its generalizations, in particular in Rissanen's theory of stochastic complexity ([48]; extended to more general loss functions by Yamanishi [64]), Wallace's theory of minimum message length (Wallace and Boulton [61], Wallace and Freeman [62]), minimax regret analysis (Barron and Xie [4]).

We showed above that our upper bound for the problem of regression is in some sense optimal; viz, we have shown that the coefficient n before $n \ln(T+1)$ in (19) cannot be improved. Related results are Rissanen's [48] proof that his term $\frac{n}{2} \ln T$ is optimal.

In traditional statistics it is usually assumed that the outcomes ω_t are generated by some true distribution, and in some papers the log loss is averaged with respect to the true distribution, which naturally leads to considering the Kullback–Leibler distance: see Clarke and Barron [13, 14]. It is natural to assume that the results of those papers can be extended to the Kullback–Leibler game (see Subsection 2.5 above); as proven in Subsection 2.5, the AA for this game is exactly the Bayesian mixing rule, which was considered in [13, 14]. In the recent paper by Clarke and Dawid [15] no assumptions are made about true distributions; it is entirely competitive on-line.

Competitive on-line statistics is concerned with, of course, the on-line performance of statistical procedures. Many procedures, like the RR, can be used in both batch and on-line setting. Dawid's [19] prequential approach to statistics recommends using on-line performance as a measure of quality of batch algorithms; therefore, competitive on-line results also provide a justification for the use of the corresponding algorithms in the batch setting (provided the philosophy of prequential statistics is accepted).

6 Discussion

It is not completely clear yet how far competitive on-line statistics can be developed; some of its limitations are well understood and others still wait to be disclosed. One serious limitation is that for some interesting games (especially those with non-compact Ω and Γ) the constants $c(\eta)$ and $a(\eta)$ are infinite; as we already mentioned, they become infinite if we remove the assumption that the response variable is bounded in the square-loss game.[5] Another limitation is that Nature is often implicitly assumed to be oblivious. However, the advantages of competitive on-line statistics turned out to be clear enough to generate a lot of interesting research in computational learning community, as described in the previous section. (And, as we mentioned there, there have been attempts to alleviate these limitations.)

The main ideas of competitive on-line statistics can be summarized as follows. The basic "recipe" is:

1. Set a complete protocol of the interaction between the statistician and her environment, including the game $(\Omega, \Gamma, \lambda)$ to be played and possible side information (such as the attributes of the object to be classified).
2. Choose the "yardstick class" $\{\xi_\theta \mid \theta \in \Theta\}$ of decision strategies for the statistician.
3. Decide on the prior probability distribution $P_0(d\theta)$ in Θ.
4. Choose a "learning rate" $\eta \geq 0$.
5. Find the "Bayesian mixture" of the decision strategies in the yardstick class, with the learning rate η and the weights given by the prior P_0.
6. When making a decision, replace the "pseudodecision" generated by the Bayesian mixture with a real prediction.

(When merging algorithms different from the AA are used, the last two steps can be combined.)

Remark 5 There are two main modes of application of merging algorithms such as the AA:

1. Apply the merging algorithm directly to the decision pool.
2. Apply the merging algorithm "on the top" of another algorithm (eg, merging Maximum Likelihood predictions made by polynomials of different degrees).

[5] In such situations one might want to replace the minimax approximation method used by the AA by something different, say to choose an action with the smallest average wr to the distribution $Q(d\omega)\int_\Theta \beta^{\lambda(\omega,\xi(\theta))}P_{t-1}(d\theta)$ in Ω, where Q is some prior distribution in Ω (say, uniform, if it exists); this method is, like the AA, a generalization of both the Weighted Majority Algorithm and the Bayesian mixing scheme. Related ideas can be found in Grünwald [32].

The second option can be applied as soon as some method encounters the problem of overfitting.

Now we will briefly discuss how to implement the steps of the basic recipe, starting with how to choose the prior. In view of inequalities such as (13), a natural idea for choosing μ is: the bigger the better. It would be ideal to take Levin's "universal enumerable semimeasure" M (see Li and Vitanyi [42], 4.3 and 4.4; M is also known as *a priori* semimeasure) as the prior if it were computable. Since under this choice we have

$$M(d\theta) \geq \epsilon P_0(d\theta) 2^{-K(P_0)},$$

where ϵ is a universal constant and $K(P_0)$ is the prefix complexity of any prior P_0, the right-hand side of (13) will increase by at most

$$c(\eta) \log_\beta \left(\epsilon 2^{-K(P_0)}\right) = c(\eta) \left(\frac{C}{\eta} + \frac{K(P_0)}{\eta} \ln 2\right)$$

(C is a universal constant) when we use M in place of P_0. Following Rissanen [47], we can instead take "computable approximations" to the universal enumerable semimeasure. In that paper Rissanen gave a relatively rigorous interpretation of Jeffreys's idea of "noninformative priors" (ie, priors reflecting Statistician's ignorance). He suggested the prior

$$P_0\{\theta\} = \frac{1}{\theta \ln^* \theta}$$

on the integers; we could also use the prior density μ on $[0,1]$ that is uniform in the middle, of the type

$$\mu(\theta) = \frac{1}{\theta \ln^*(1/\theta)}$$

near the left end, and of the type

$$\mu(1-\theta) = \frac{1}{\theta \ln^*(1/\theta)}$$

near the right end. Another popular approach to choosing the prior is to take the "optimal" P_0, typically according to some minimax criterion of optimality; see, eg, Clarke and Barron [14], Barron and Xie [4], Freund [27].

A difficult question is how to choose the loss function. In some situations we might have a clear "true" loss function (when we can precisely evaluate consequences of our predictions); in other situations we do not know where our predictions will be used. In the former, "decision" situations, we are already given a loss function and the problem is to study its properties (is it perfectly mixable? what is the mixability curve?); in the latter, "inference" situations, we can use some conventional loss functions like the square loss.

As concerns choosing the learning rate η, in the situation where our decision pool is finite or countable a natural (though vague) principle [52] is:

- ◇ if we expect that the performance of some strategy with very big weight is not very bad, η should be close to 0 (we should learn well, even if slowly);
- ◇ if we expect that some strategy with not very small weight performs very well, η should be big (we should learn quickly, even if badly).

Sometimes we will want to make η "infinitesimal", without actually setting $\eta = 0$ (eg, in the case of the absolute loss function, where $c(\eta) \to 1$ as $\eta \to 0$); some way of doing this is described in Cesa-Bianchi et al [10]. Cesa-Bianchi et al [11] discuss using a prior distribution in the set of possible values of η (it would be natural to take an approximation to the "universal prior"). In the case of perfectly mixable games the most natural thing to do is to take the largest η such that $c(\eta) = 1$ (in other words, to minimize $a(\eta)$ under the constraint $c(\eta) = 1$).

In conclusion we will place the approach of competitive on-line statistics in the more general framework of the theory of decision making under uncertainty. There are several popular principles of decision making under uncertainty, such as Laplace's principle, Wald's minimax principle and Savage's minimax regret principle, but all of them have been criticized as arbitrary. Competitive on-line statistics is based on the following version of the minimax regret principle: a decision strategy is acceptable if its regret (in the simplest case, the loss of the aggregating strategy minus the loss of the best decision strategy in the pool) is always small (according to (1) the maximum regret of the AA is small if there are not too many experts). This restricted version of the minimax regret principle is far less objectionable than the general principle.

Acknowledgments

I am grateful to Phil Dawid, Meir Feder, Mark Herbster, Kostas Skouras, Manfred Warmuth, and Kenji Yamanishi for useful discussions and illuminating comments (and to Manfred also for finding a mistake in an earlier draft).

This work was partially supported by EPSRC through grants GR/L35812 ("Support Vector and Bayesian learning algorithms"), GR/M14937 ("Predictive complexity: recursion-theoretic variants"), and GR/M16856 ("Comparison of Support Vector Machine and Minimum Message Length methods for induction and prediction").

References

[1] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331, 1995.

[2] Peter Auer and Manfred K Warmuth. Tracking the best disjunction. *Machine Learning*, 32:127–150, 1998.

[3] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. Manuscript. An extended abstract appeared in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 31–40, San Francisco, CA, 1999. Morgan Kaufmann.

[4] Andrew R Barron and Q Xie. Asymptotic minimax regret for data compression, gambling, and prediction. Technical report, Statistics Department, Yale University, 1996. Submitted to *IEEE Trans on Inform Theory*.

[5] Edwin F Beckenbach and Richard Bellman. *Inequalities*. Springer, Berlin, 1965.

[6] Avrim Blum and Carl Burch. On-line learning and the metrical task system problem. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 45–53, New York, 1997. Assoc Comput Mach. Also published in *Mach Learning*.

[7] Avrim Blum and Adam Kalai. Universal portfolios with and without transaction costs. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 309–313, New York, 1997. Assoc Comput Mach.

[8] A A Borovkov. *Mathematical statistics (in Russian)*. Nauka, Moscow, 1984.

[9] Nicolò Cesa-Bianchi and P Fischer. Finite-time regret bounds for the multi-armed bandit problem. In *5th International Conference on Machine Learning*, pages 100–108. Morgan Kaufmann, 1998.

[10] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44:427–485, 1997.

[11] Nicolò Cesa-Bianchi, David P Helmbold, and Sandra Panizza. On Bayes methods for on-line boolean prediction. *Algorithmica*, 22:112–137, 1998.

[12] Nicolò Cesa-Bianchi, Philip M Long, and Manfred K Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Trans Neural Networks*, 7:604–619, 1996.

[13] Bertrand Clarke and Andrew R Barron. Information theoretic asymptotics of Bayes methods. *IEEE Trans on Inform Theory*, 36:453–471, 1990.

[14] Bertrand Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favourable under entropy risk. *J Statist Planning and Inference*, 41:37–60, 1994.

[15] Bertrand Clarke and A Philip Dawid. On-line prediction with experts under a log-scoring rule. Manuscript, 1999.

[16] William W Cohen and Yoram Singer. Context-sensitive learning methods for text categorization. In *SIGIR: Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, 1996.

[17] T Cover. Universal portfolios. *Mathematical Finance*, 1:1–29, 1991.

[18] T Cover and E Ordentlich. Universal portfolios with side information. *IEEE Trans Inform Theory*, 42:348–363, 1996.

[19] A Philip Dawid. Statistical theory: the prequential approach. *J R Statist Soc* A, 147:278–292, 1984.

[20] A Philip Dawid. Probability forecasting. In S Kotz, N L Johnson, and C B Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley-Interscience, New York, 1986.

[21] A Philip Dawid. Fisherian inference in likelihood and prequential frames of reference (with discussion). *J R Statist Soc* B, 53:79–109, 1991.

[22] A DeSantis, G Markowsky, and M N Wegman. Learning probabilistic prediction functions. In *Proceedings of the 29th Annual IEEE Symposium on Foundations of Computer Science*, pages 110–119, Los Alamitos, CA, 1988. IEEE Comput Soc.

[23] N R Draper and H Smith. *Applied Regression Analysis*. Wiley, New York, 2nd edition, 1981.

[24] Meir Feder and Andrew C Singer. Universal data compression and linear prediction. In *Proceedings of the 1998 IEEE Data Compression Conference*, 1998.

[25] Jürgen Forster. On relative loss bounds in generalized linear regression. Manuscript, 1999.

[26] D P Foster. Prediction in the worst case. *Ann Statist*, 19:1084–1090, 1991.

[27] Yoav Freund. Predicting a binary sequence almost as well as the optimal biased coin. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 89–98, New York, 1996. Assoc Comput Mach.

[28] Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *Proceedings, 9th Annual Conference on Computational Learning Theory*, pages 325–332, New York, 1996. Assoc Comput Mach.

[29] Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. Manuscript, 1997.

[30] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[31] Yoav Freund, Robert E Schapire, Yoram Singer, and Manfred Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, New York, 1997. Assoc Comput Mach.

[32] Peter Grünwald. *The Minimum Description Length principle and reasoning under uncertainty.* PhD thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, 1998.

[33] David Haussler, Jyrki Kivinen, and Manfred K Warmuth. Tight worst-case loss bounds for predicting with expert advice. Technical Report UCSC-CRL-94-36, University of California at Santa Cruz, revised December 1994. Short version in Paul Vitányi, editor, *Computational Learning Theory*, Lecture Notes in Computer Science, volume 904, pages 69–83, Springer, Berlin, 1995.

[34] David P Helmbold, D D E Long, and B Sherrod. A dynamic disk spin-down technique for mobile computing. In *Proceedings of the 2nd Annual ACM International Conference on Mobile Computing and Networking*, 1996.

[35] David P Helmbold and Robert E Schapire. Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, 27:51–68, 1997.

[36] Mark Herbster and Manfred Warmuth. Tracking the best expert. In *Proceedings of the 12th International Conference on Machine Learning*, pages 286–294. Morgan Kaufmann, 1995. To appear in *Machine Learning*.

[37] Mark Herbster and Manfred Warmuth. Tracking the best regressor. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 24–31. Assoc Comput Mach, 1998.

[38] Yuri Kalnichkan. General linear relations among different types of predictive complexity. In *Proceedings of the 1990 Annual Conference on Algorithmic Learning Theory*, 1999. To be published.

[39] Yuri Kalnichkan. Linear relations between square-loss and Kolmogorov complexity. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, New York, 1999. Assoc Comput Mach.

[40] Jyrki Kivinen and Manfred K Warmuth. Exponential Gradient versus Gradient Descent for linear predictors. *Inform Computation*, 132:1–63, 1997.

[41] Jyrki Kivinen and Manfred K Warmuth. Averaging expert predictions. Manuscript, 1999.

[42] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications.* Springer, New York, 2nd edition, 1997.

[43] Nick Littlestone and Manfred K Warmuth. The Weighted Majority Algorithm. *Inform Computation*, 108:212–261, 1994.

[44] F Pereira and Yoram Singer. An efficient extension to mixture techniques for prediction and decision trees. In *Proceedings, 10th Annual Conference on Computational Learning Theory*, pages 114–121. Assoc Comput Mach, 1997.

[45] William H Press, Brian P Flannery, Saul A Teukolsky, and William T Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1992.

[46] P Raghavan (ed). Special issue on competitive on-line algorithms. *Algorithmica*, 11:1–91, 1994.

[47] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[48] Jorma Rissanen. Stochastic complexity (with discussion). *Journal of Royal Statistical Society B*, 49:223–239 and 252–265, 1987.

[49] Eiji Takimoto, Akira Maruoka, and Volodya Vovk. Predicting nearly as well as the best pruning of a decision tree through dynamic programming scheme. Submitted for publication, 1998.

[50] Volodya Vovk. Probability theory for the Brier game. Accepted for publication in *Theoretical Computer Science*. Preliminary version in Ming Li and Akira Maruoka, editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, volume 1316, pages 323–338, 1997. Full version: Technical Report CSD-TR-97-09, Department of Computer Science, Royal Holloway, University of London, revised February 1998.

[51] Volodya Vovk. On a randomness criterion. *Soviet Mathematics Doklady*, 35:656–660, 1987.

[52] Volodya Vovk. Aggregating strategies. In M Fulk and J Case, editors, *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

[53] Volodya Vovk. Universal forecasting algorithms. *Information and Computation*, 96:245–277, 1992.

[54] Volodya Vovk. Competitive on-line linear regression. In M I Jordan, M J Kearns, and S A Solla, editors, *Advances in Neural Information Processing Systems*, pages 364–370, Cambridge, MA, 1998. MIT Press.

[55] Volodya Vovk. A game of prediction with expert advice. *J Comput Syst Sci*, 56:153–173, 1998.

[56] Volodya Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35(3):247–282, 1999.

[57] Volodya Vovk and Alex Gammerman. Complexity Approximation Principle. *Computer Journal*, 42, 1999. 318–322.

[58] Volodya Vovk and Chris J H C Watkins. Universal portfolio selection. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 12–23, New York, 1998. Assoc Comput Mach.

[59] Volodya Vovk. Competitive on-line statistics. In *Bulletin of the International Statistical Institute. The 52nd Session, Proceedings*, volume LVIII, book 1, pages 231–234, 1999.

[60] Volodya Vovk and Alex Gammerman. Statistical applications of algorithmic randomness. In *Bulletin of the International Statistical Institute. The 52nd Session, Contributed Papers*, volume LVIII, book 3, pages 469–470, 1999.

[61] Chris S Wallace and D M Boulton. An information measure for classification. *Computer Journal*, 11:185–195, 1968.

[62] Chris S Wallace and P R Freeman. Estimation and inference by compact coding (with discussion). *J R Statist Soc* B, 49:240–265, 1987.

[63] Kenji Yamanishi. Generalized stochastic complexity and its applications to learning. In *Proceedings of the 1994 Conference on Information Science and and Systems*, volume 2, pages 763–768, 1994.

[64] Kenji Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans Inform Theory*, 20:100–114, 1998.

[65] Kenji Yamanishi. Minimax relative loss analysis for sequential prediction aalgorithms using parametric hypotheses. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 32–43, New York, 1998. Assoc Comput Mach.

RÉSUMÉ

Cet article décrit une approche nouvelle à modelage statistique combinant les techniques mathematiques de statistique Bayesienne avec la philosophie de la theorie de algorithmes compétitives en ligne. Dans cette approche, qui émergeait durant le décennie dernière dans l'informatique, on ne suppose pas que les données sont produites par une mécanisme stochastique ; au lieu de cela, il est prouvé que les procédures statistiques compétitives en ligne atteignent toujours (et non, par exemple, avec haute probabilité) quelque but desirable (explicitant la bonne performance sur les données réeles). Cet article passe en revue des plusieurs résultats dans cette domaine ; son matériel neuf comprend les preuves pour la performance de le "algorithme agrégeant" dans le problème de la régression linéaire avec la perte carrée.