Prediction with expert evaluators' advice

Alexey Chernov and Vladimir Vovk
{chernov,vovk}@cs.rhul.ac.uk

March 23, 2009

Abstract

We introduce a new protocol for prediction with expert advice in which each expert evaluates the learner's and his own performance using a loss function that may change over time and may be different from the loss functions used by the other experts. The learner's goal is to perform better or not much worse than each expert, as evaluated by that expert, for all experts simultaneously. If the loss functions used by the experts are all proper scoring rules and all mixable, we show that the defensive forecasting algorithm enjoys the same performance guarantee as that attainable by the Aggregating Algorithm in the standard setting and known to be optimal. This result is also applied to the case of "specialist" (or "sleeping") experts. In this case, the defensive forecasting algorithm reduces to a simple modification of the Aggregating Algorithm.

1 Introduction

We consider the problem of online sequence prediction. A process generates outcomes $\omega_1, \omega_2, \ldots$ step by step. At each step t, a learner tries to guess the next outcome announcing his prediction γ_t. Then the actual outcome ω_t is revealed. The quality of the learner's prediction is measured by a loss function: the learner's loss at step t is $\lambda(\gamma_t, \omega_t)$.

Prediction with expert advice is a framework that does not make any assumptions about the generating process. The performance of the learner is compared to the performance of several other predictors called experts. At each step, each expert gives his prediction γ_t^n, then the learner produces his own prediction γ_t (possibly based on the experts' predictions at the last step and the experts' predictions and outcomes at all the previous steps), and the accumulated losses are updated for the learner and for the experts. There are many algorithms for the learner in this framework; for a review, see [3].

In practical applications of the algorithms for prediction with expert advice, choosing the loss function is often a problem. The task may have no natural measure of loss, except the vague concept that the closer the prediction to the outcome the better. Thus one can select among several common loss functions, for example, the quadratic loss (reflecting the idea of least squares methods) or

the logarithmic loss (which has an information theory background). A similar issue arises when experts themselves are prediction algorithms that optimize some losses internally. Then it is unfair to these experts when the learner competes with them according to a "foreign" loss function.

This paper introduces a new version of the framework of prediction with expert advice where there is no single fixed loss function but some loss function is linked to every expert. The performance of the learner is compared to the performance of each expert according to the loss function linked to that expert. Informally speaking, each expert has to be convinced that the learner performs almost as well as, or better than, that expert himself.

We prove that a known algorithm for the learner, the defensive forecasting algorithm [4], can be applied in the new setting and gives the same performance guarantee as that attainable in the standard setting, provided all loss functions are proper scoring rules.

Another framework to which our methods can be fruitfully applied is that of "specialist experts": see, e.g., [8], [1], and [11]. We generalize some of the known results in the case of mixable loss functions.

To keep presentation as simple as possible, we restrict ourselves to binary outcomes $\{0,1\}$, predictions from $[0,1]$, and a finite number of experts. We formulate our results for mixable loss functions only. However, these results can be easily transferred to more general settings (non-binary outcomes, arbitrary prediction spaces, countably many experts, second-guessing experts, etc.) where the methods of [4] work.

2 Prediction with simple experts' advice

In this preliminary section we recall the standard protocol of prediction with expert advice and some known results.

Let $\{0,1\}$ be the set of possible *outcomes* ω, $[0,1]$ be the set of possible *predictions* γ, and $\lambda : [0,1] \times \{0,1\} \to [0,\infty]$ be the *loss function*. The loss function λ and parameter N (the number of experts) specify the game of prediction with expert advice. The game is played by Learner, Reality, and N experts, Expert 1 to Expert N, according to the following protocol.

PREDICTION WITH EXPERT ADVICE

$L_0 := 0$.
$L_0^n := 0$, $n = 1, \dots, N$.
FOR $t = 1, 2, \dots$:
  Expert n announces $\gamma_t^n \in [0,1]$, $n = 1, \dots, N$.
  Learner announces $\gamma_t \in [0,1]$.
  Reality announces $\omega_t \in \{0,1\}$.
  $L_t := L_{t-1} + \lambda(\gamma_t, \omega_t)$.
  $L_t^n := L_{t-1}^n + \lambda(\gamma_t^n, \omega_t)$, $n = 1, \dots, N$.
END FOR

The goal of Learner is to keep his loss L_t smaller or at least not much greater than the loss L_t^n of Expert n, at each step t and for all $n = 1, \dots, N$.

We only consider loss functions that have the following properties:

Assumption 1: $\lambda(\gamma, 0)$ and $\lambda(\gamma, 1)$ are continuous in $\gamma \in [0, 1]$ and for the standard (Aleksandrov's) topology on $[0, \infty]$.

Assumption 2: There exists $\gamma \in [0, 1]$ such that $\lambda(\gamma, 0)$ and $\lambda(\gamma, 1)$ are both finite.

Assumption 3: There exists no $\gamma \in [0, 1]$ such that $\lambda(\gamma, 0)$ and $\lambda(\gamma, 1)$ are both infinite.

The *superprediction set* for a loss function λ is

$$\Sigma_\lambda := \left\{(x, y) \in [0, \infty)^2 \mid \exists \gamma \, \lambda(\gamma, 0) \le x \text{ and } \lambda(\gamma, 1) \le y\right\}. \tag{1}$$

By Assumption 2, this set is non-empty. For $\eta > 0$, let $E_\eta : [0, \infty]^2 \to [0, 1]^2$ be the homeomorphism defined by $E_\eta(x, y) := (\mathrm{e}^{-\eta x}, \mathrm{e}^{-\eta y})$. The loss function λ is called *η-mixable* if the set $E_\eta(\Sigma_\lambda)$ is convex. It is called *mixable* if it is η-mixable for some $\eta > 0$.

Theorem 1. *If a loss function λ is η-mixable, then there exists a strategy for Learner that guarantees that in the game of prediction with expert advice with N experts and the loss function λ it holds, for all t and for all $n = 1, \dots, N$, that*

$$L_t \le L_t^n + \frac{1}{\eta} \ln N. \tag{2}$$

The bound is optimal: if λ is not η-mixable, then no strategy for Learner can guarantee (2).

For the proof and other details, see [3], [10], [15], or [16, Theorem 8]; one of the algorithms guaranteeing (2) is the (Strong) Aggregating Algorithm (AA). As shown in [4], one can take the defensive forecasting algorithm instead of the AA in the theorem.

3 Proper scoring rules

A loss function λ is a *proper scoring rule* if for any $\pi, \pi' \in [0, 1]$ it holds that

$$\pi\lambda(\pi, 1) + (1 - \pi)\lambda(\pi, 0) \le \pi\lambda(\pi', 1) + (1 - \pi)\lambda(\pi', 0);$$

it is a *strictly proper scoring rule* if the inequality holds with $<$ in place of $\le$ whenever $\pi' \ne \pi$. The interpretation is that the prediction π is an estimate of the probability that $\omega = 1$. The definition says that the expected loss with respect to a probability distribution is minimal if the prediction is the true probability of 1. Informally, a strictly proper scoring rule encourages a forecaster (Learner or one of the experts) to announce his true subjective probability that the next outcome is 1. (See [6], [9], and [2] for detailed reviews.)

Simple examples of strictly proper scoring rules are provided by two most common loss functions: the log loss function

$$\lambda(\gamma, \omega) := -\ln(\omega\gamma + (1-\omega)(1-\gamma))$$

(i.e., $\lambda(\gamma, 0) = -\ln(1-\gamma)$ and $\lambda(\gamma, 1) = -\ln\gamma$) and the square loss function

$$\lambda(\gamma, \omega) := (\omega - \gamma)^2 .$$

A trivial but important for us generalization of the log loss function is

$$\lambda(\gamma, \omega) := -\frac{1}{\eta}\ln(\omega\gamma + (1-\omega)(1-\gamma)), \tag{3}$$

where η is a positive constant. The generalized log loss function is also a proper scoring rule (in general, multiplying a proper scoring rule by a positive constant we again obtain a proper scoring rule).

We will often say "(strictly) proper loss function" meaning a loss function that is a (strictly) proper scoring rule. Our main interest will be in loss functions that are both mixable and proper. Let $\mathcal{L}$ be the set of all such loss functions.

4 Prediction with expert evaluators' advice

In this section we consider a very general protocol of prediction with expert advice. The intuition behind special cases of this protocol will be discussed in the following sections.

PREDICTION WITH EXPERT EVALUATORS' ADVICE
FOR $t = 1, 2, \ldots$:
  Expert n announces $\gamma_t^n \in [0,1]$, $\eta_t^n > 0$, and η_t^n-mixable $\lambda_t^n \in \mathcal{L}$,
    $n = 1, \ldots, N$.
  Learner announces $\gamma_t \in [0,1]$.
  Reality announces $\omega_t \in \{0,1\}$.
END FOR

The main mathematical result of this paper is the following.

Theorem 2. *Learner has a strategy (e.g., the defensive forecasting algorithm described below) that guarantees that in the game of prediction with N expert evaluators' advice it holds, for all T and for all $n = 1, \ldots, N$, that*

$$\sum_{t=1}^{T} \eta_t^n \left(\lambda_t^n(\pi_t, \omega_t) - \lambda_t^n(\gamma_t^n, \omega_t)\right) \le \ln N.$$

The description of the defensive forecasting algorithm and the proof of the theorem will be given in Section 7.

Corollary 1. *For any $\eta > 0$, Learner has a strategy that guarantees*

$$\sum_{t=1}^{T} \lambda_t^n(\pi_t, \omega_t) \le \sum_{t=1}^{T} \lambda_t^n(\gamma_t^n, \omega_t) + \frac{\ln N}{\eta}, \tag{4}$$

for all T and all $n = 1, \dots, N$, in the game of prediction with N expert evaluators' advice in which the experts are required to always choose η-mixable loss functions λ_t^n.

This corollary is more intuitive than Theorem 2 as (4) compares the cumulative losses suffered by Learner and each expert.

In the following sections we will discuss two interesting special cases of Theorem 2 and Corollary 1.

5 Prediction with constant expert evaluators' advice

In the game of this section, as in the previous one, the experts are "expert evaluators": each of them measures Learner's and his own performance using his own loss function, supposed to be mixable and proper. The difference is that now each expert is linked to a fixed loss function. The game is specified by N loss functions $\lambda^1, \dots, \lambda^N$.

PREDICTION WITH CONSTANT EXPERT EVALUATORS' ADVICE
$L_0^{(n)} := 0$, $n = 1, \dots, N$.
$L_0^n := 0$, $n = 1, \dots, N$.
FOR $t = 1, 2, \dots$:
  Expert n announces $\gamma_t^n \in [0, 1]$, $n = 1, \dots, N$.
  Learner announces $\gamma_t \in [0, 1]$.
  Reality announces $\omega_t \in \{0, 1\}$.
  $L_t^{(n)} := L_{t-1}^{(n)} + \lambda^n(\gamma_t, \omega_t)$, $n = 1, \dots, N$.
  $L_t^n := L_{t-1}^n + \lambda^n(\gamma_t^n, \omega_t)$, $n = 1, \dots, N$.
END FOR

There are two changes in the protocol as compared to the basic protocol of prediction with expert advice in Section 2. The accumulated loss L_t^n of each expert is now calculated according to his own loss function λ^n. For Learner, there is no single accumulated loss anymore. Instead, the loss $L_t^{(n)}$ of Learner is calculated separately against each expert, according to that expert's loss function λ^n. Informally speaking, each expert evaluates his own performance and the performance of Learner according to the expert's own (but publicly known) criteria.

In the standard setting of prediction with expert advice it is often said that Learner's goal is to compete with the best expert in the pool. In the new setting, we cannot speak about the best expert: the experts' performance is evaluated

by different loss functions and thus the losses may be measured on different scales. But it still makes sense to consider bounds on the *regret* $L_t^{(n)} - L_t^n$ for each n.

Theorem 2 (or Corollary 1) immediately implies the following performance guarantee for the defensive forecasting algorithm in our current setting.

Corollary 2. *Suppose that every λ^n is a proper loss function that is η^n-mixable for some $\eta^n > 0$, $n = 1, \dots, N$. Then Learner has a strategy (such as the defensive forecasting algorithm) that guarantees that in the game of prediction with N experts' advice and loss functions $\lambda^1, \dots, \lambda^N$ it holds, for all T and for all $n = 1, \dots, N$, that*

$$L_T^{(n)} \le L_T^n + \frac{\ln N}{\eta^n}. \tag{5}$$

The new bound (5) is precisely the same as the bound for the standard setting of Theorem 1. But rigorous comparison of the actual power of these two bounds is not so trivial.

Formally speaking, the task of Learner in the new protocol is not strictly harder and is not strictly easier than in the standard protocol: the task is incomparable. Learner must now compete with different experts by different rules. But this is not necessarily a disadvantage. Consider an example. Suppose that all experts except one are linked to one loss function and the last expert is linked to another loss function. And this last loss function is somehow trivial, say, equals 1 independent of the outcome and the prediction. Then we arrive at the standard protocol with $N - 1$ experts, since the regret against the last expert is zero independent of our predictions. In this example, we can get a better bound than that given by Corollary 2. This non-optimality is especially apparent in the case when we have a huge number of experts, but all except one are linked to a trivial loss function. Then our regret bound is large, being a logarithm of a huge number, whereas one can achieve zero regret against all experts whatever strategy they use—since the loss functions are unfavourable to the experts.

Nevertheless, it is intuitively clear that the new protocol is somewhat harder for Learner in general. And Corollary 2 is really surprising: it is hard to believe that Learner can compete against several arbitrary loss functions as well as against only one of them. The reason why this is possible is that the loss functions are assumed to be proper.

Multiobjective prediction with expert advice

To conclude this section, let us consider another variant of the protocol with several loss functions. As mentioned in the introduction, sometimes we have experts' predictions, and we are not given a single loss function, but have several possible candidates. The most cautious way to generate Learner's predictions is to ensure that the regret is small against all experts and according to all loss functions. The following protocol formalizes this task. Now we have N experts and M loss functions $\lambda^1, \dots, \lambda^M$.

MULTIOBJECTIVE PREDICTION WITH EXPERT ADVICE
$L_0^{(m)} := 0$, $m = 1, \dots, M$.
$L_0^{n,m} := 0$, $n = 1, \dots, N$ and $m = 1, \dots, M$.
FOR $t = 1, 2, \dots$:
    Expert n announces $\gamma_t^n \in [0, 1]$, $n = 1, \dots, N$.
    Learner announces $\gamma_t \in [0, 1]$.
    Reality announces $\omega_t \in \{0, 1\}$.
    $L_t^{(m)} := L_{t-1}^{(m)} + \lambda^m(\gamma_t, \omega_t)$, $m = 1, \dots, M$.
    $L_t^{n,m} := L_{t-1}^{n,m} + \lambda^m(\gamma_t^n, \omega_t)$, $n = 1, \dots, N$ and $m = 1, \dots, M$.
END FOR

Corollary 3. *Suppose that every λ^m is an η^m-mixable proper loss function, for some $\eta^m > 0$, $m = 1, \dots, M$. The defensive forecasting algorithm guarantees that, in the multiobjective game of prediction with N experts and the loss functions $\lambda^1, \dots, \lambda^M$,*

$$L_t^{(m)} \le L_t^{n,m} + \frac{\ln MN}{\eta^m} \tag{6}$$

for all t, all $n = 1, \dots, N$, and all $m = 1, \dots, M$.

Proof. This follows easily from Corollary 2. For each $n \in \{1, \dots, N\}$, let us construct M new experts (n, m). Expert (n, m) predicts as Expert n and is linked to the loss function λ^m. Applying Corollary 2 to these MN experts, we get bound (6). □

The last protocol is harder for Learner than the standard protocol when $M > 1$: Learner must satisfy all old regret bounds and also some new bounds. But the increase in the regret bounds is surprisingly small: only an additive term proportional to $\ln M$. Whether the dependence on M in Corollary 3 is optimal remains an open problem.

A further generalization of our last protocol involves a binary relation R between the N experts and the M loss functions, where nRm, $n \in \{1, \dots, N\}$ and $m \in \{1, \dots, M\}$, is interpreted as Expert n using the loss function λ^m when evaluating Learner's and his own performance. It is assumed that for each n there exists at least one m such that nRm. The relation R is naturally represented as a bipartite graph connecting the vertices in the set $\{1, \dots, N\}$ to vertices in the set $\{1, \dots, M\}$. Equation (6) now becomes

$$L_t^{(m)} \le L_t^{n,m} + \frac{\ln K}{\eta^m},$$

for all $(n, m) \in R$, where K is the cardinality of R (equivalently, the number of edges in the bipartite graph).

A simple example

Let λ^1 be the log loss function and λ^2 the square loss function. As already mentioned, both loss functions are proper and mixable. It is known (see, e.g., [3], [10], or [14]) that λ^1 is 1-mixable and λ^2 is 2-mixable. Suppose we are competing with N experts producing predictions γ_t^n under these two loss functions. The defensive forecasting algorithm ensures that the regret with respect to the logarithmic loss function is bounded by $\ln(2N) < \ln N + 0.7$, and the regret with respect to the square loss function is bounded by $0.5\ln(2N) < 0.5\ln N + 0.4$—practically the same as the regrets against N experts that are achievable when Learner chooses his predictions with respect to one of the loss functions only.

6 Prediction with specialist experts' advice

The experts of this section are allowed to "sleep", i.e., abstain from giving advice to Learner at some steps. This generalization is important for text-processing applications (see, e.g., [5]). We will be assuming that there is only one loss function λ, although generalization to the case of N loss functions $\lambda^1, \dots, \lambda^N$ is straightforward. The loss function λ does not need to be proper (but it is still required to be mixable).

Let a be any object that does not belong to $[0,1]$; intuitively, it will stand for an expert's decision to abstain.

PREDICTION WITH SPECIALIST EXPERTS' ADVICE
$L_0^{(n)} := 0$, $n = 1, \dots, N$.
$L_0^n := 0$, $n = 1, \dots, N$.
FOR $t = 1, 2, \dots$:
  Expert n announces $\gamma_t^n \in ([0,1] \cup \{a\})$, $n = 1, \dots, N$.
  Learner announces $\gamma_t \in [0,1]$.
  Reality announces $\omega_t \in \{0,1\}$.
  $L_t^{(n)} := L_{t-1}^{(n)} + \mathbb{I}_{\{\gamma_t^n \neq a\}} \lambda(\gamma_t, \omega_t)$, $n = 1, \dots, N$.
  $L_t^n := L_{t-1}^n + \mathbb{I}_{\{\gamma_t^n \neq a\}} \lambda(\gamma_t^n, \omega_t)$, $n = 1, \dots, N$.
END FOR

The indicator function $\mathbb{I}_{\{\gamma_t^n \neq a\}}$ of the event $\gamma_t^n \neq a$ is defined to be 1 if $\gamma_t^n \neq a$ and 0 if $\gamma_t^n = a$. Therefore, $L_t^{(n)}$ and L_t^n refer to the cumulative loss of Learner and Expert n over the steps when Expert n is awake. Now Learner's goal is to do as well as each expert on the steps chosen by that expert.

Corollary 4. *Let λ be a loss function that is η-mixable for some $\eta > 0$. Then Learner has a strategy (e.g., the defensive forecasting algorithm) that guarantees that in the game of prediction with N specialist experts' advice and loss function λ it holds, for all T and for all $n = 1, \dots, N$, that*

$$L_T^{(n)} \leq L_T^n + \frac{\ln N}{\eta}. \tag{7}$$

Proof. Without loss of generality the loss function λ may be assumed to be proper (this can be achieved by reparameterization of the predictions $\gamma \in [0,1]$). The protocol of this section then becomes a special case of the protocol of Section 4 in which at each step each expert outputs $\eta_t^n = \eta$ and either $\lambda_t^n = \lambda$ (when he is awake) or $\lambda_t^n = 0$ (when he is asleep). (Alternatively, in which at each step each expert outputs $\lambda_t^n = \lambda$ and either $\eta_t^n = \eta$, when he is awake, or $\eta_t^n = 0$, when he is asleep.)

□

7 Defensive forecasting algorithm and the proof of Theorem 2

In this section we prove Theorem 2. Our proof is constructive: we explicitly describe the defensive forecasting algorithm achieving the bound in Theorem 2.

The algorithm

For each $n = 1, \dots, N$, let us define the function

$$Q^n : \left([0,1]^N \times (0,\infty)^N \times \mathcal{L}^N \times [0,1] \times \{0,1\}\right)^* \to [0,\infty]$$
$$Q^n\left(\gamma_1^\bullet, \eta_1^\bullet, \lambda_1^\bullet, \pi_1, \omega_1, \dots, \gamma_T^\bullet, \eta_T^\bullet, \lambda_T^\bullet, \pi_T, \omega_T\right) := \prod_{t=1}^{T} \mathrm{e}^{\eta_t^n \left(\lambda_t^n(\pi_t,\omega_t) - \lambda_t^n(\gamma_t^n,\omega_t)\right)}, \tag{8}$$

where γ_t^n are the components of $\gamma_t^\bullet$, η_t^n are the components of $\eta_t^\bullet$, and λ_t^n are the components of $\lambda_t^\bullet$:

$$\gamma_t^\bullet := (\gamma_t^1, \dots, \gamma_t^N),$$
$$\eta_t^\bullet := (\eta_t^1, \dots, \eta_t^N),$$
$$\lambda_t^\bullet := (\lambda_t^1, \dots, \lambda_t^N).$$

As usual, the product $\prod_{t=1}^{0}$ is interpreted as 1, so that $Q^n() = 1$. The functions Q^n will usually be applied to $\gamma_t^\bullet := (\gamma_t^1, \dots, \gamma_t^N)$ the predictions made by all the N experts at step t, $\eta_t^\bullet := (\eta_t^1, \dots, \eta_t^N)$ the learning rates chosen by the experts at step t, and $\lambda_t^\bullet := (\lambda_t^1, \dots, \lambda_t^N)$ the loss functions used by the experts at step t. Notice that Q^n does not depend on the predictions, learning rates, and loss functions of the experts other than Expert n.

Set

$$Q := \frac{1}{N} \sum_{n=1}^{N} Q^n$$

and

$$f_t(\pi, \omega) :=$$

$$Q\left(\gamma_1^\bullet, \eta_1^\bullet, \lambda_1^\bullet, \pi_1, \omega_1, \dots, \gamma_{t-1}^\bullet, \eta_{t-1}^\bullet, \lambda_{t-1}^\bullet, \pi_{t-1}, \omega_{t-1}, \gamma_t^\bullet, \eta_t^\bullet, \lambda_t^\bullet, \pi, \omega\right) - Q\left(\gamma_1^\bullet, \eta_1^\bullet, \lambda_1^\bullet, \pi_1, \omega_1, \dots, \gamma_{t-1}^\bullet, \eta_{t-1}^\bullet, \lambda_{t-1}^\bullet, \pi_{t-1}, \omega_{t-1}\right), \quad (9)$$

where (π, ω) ranges over $[0,1] \times \{0,1\}$; the expression $\infty - \infty$ is understood as, say, 0. The defensive forecasting algorithm is defined in terms of the functions f_t.

Defensive forecasting algorithm

FOR $t = 1, 2, \dots$:
  Read the experts' predictions $\gamma_t^\bullet = (\gamma_t^1, \dots, \gamma_t^N) \in [0,1]^N$,
    learning rates $\eta_t^\bullet = (\eta_t^1, \dots, \eta_t^N) \in (0,\infty)^N$,
    and loss functions $\lambda_t^\bullet = (\lambda_t^1, \dots, \lambda_t^N) \in \mathcal{L}^N$.
  Define $f_t : [0,1] \times \{0,1\} \to [-\infty, \infty]$ by (9).
  If $f_t(0,1) \le 0$, predict $\pi_t := 0$ and go to R.
  If $f_t(1,0) \le 0$, predict $\pi_t := 1$ and go to R.
  Otherwise (if both $f_t(0,1) > 0$ and $f_t(1,0) > 0$),
    take any π satisfying $f_t(\pi, 0) = f_t(\pi, 1)$ and predict $\pi_t := \pi$.
  R: Read Reality's move $\omega_t \in \{0,1\}$.
END FOR

The existence of a π satisfying $f_t(\pi,0) = f_t(\pi,1)$ will be proved in Lemma 1 below. We will see that the function $f_t(\pi) := f_t(\pi,1) - f_t(\pi,0)$ takes values of opposite signs at $\pi = 0$ and $\pi = 1$. Therefore, a root of $f_t(\pi) = 0$ can be found by, e.g., bisection (see [12], Chapter 9, for a review of bisection and more efficient methods, such as Brent's).

Reductions

The most important property of the defensive forecasting algorithm is that it produces predictions π_t such that the sequence

$$Q_t := Q(\gamma_1^\bullet, \eta_1^\bullet, \lambda_1^\bullet, \pi_1, \omega_1, \dots, \gamma_t^\bullet, \eta_t^\bullet, \lambda_t^\bullet, \pi_t, \omega_t) \quad (10)$$

is non-increasing. This property will be proved later; for now, we will only check that it implies the bound on the regret term given in Theorem 2. Since the initial value Q_0 of Q is 1, we have $Q_t \le 1$ for all t. And since $Q^n \ge 0$ for all n, we have $Q^n \le NQ$ for all n. Therefore, Q_t^n, defined by (10) with Q^n in place of Q, is at most N at each step t. By the definition of Q^n this means that

$$\sum_{t=1}^{T} \eta_t^n \left(\lambda_t^n(\pi_t, \omega_t) - \lambda_t^n(\gamma_t^n, \omega_t)\right) \le \ln N,$$

which is the bound claimed in the theorem.

In the proof of the inequalities $Q_0 \ge Q_1 \ge \cdots$ we will follow [4] (for a presentation adapted to the binary case, see [17]). The key fact we use is that

Q is a game-theoretic supermartingale. Let us define this notion and prove its basic properties.

Let E be any non-empty set. A function $S : (E \times [0,1] \times \{0,1\})^* \to (-\infty, \infty]$ is called a *supermartingale* (omitting "game-theoretic") if, for any T, any $e_1, \dots, e_T \in E$, any $\pi_1, \dots, \pi_T \in [0,1]$, and any $\omega_1, \dots, \omega_{T-1} \in \{0,1\}$, it holds that

$$\begin{aligned}\pi_T S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1}, e_T, \pi_T, 1) \\ + (1 - \pi_T) S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1}, e_T, \pi_T, 0) \\ \le S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1}). \quad (11)\end{aligned}$$

Remark. The standard measure-theoretic notion of a supermartingale is obtained when the arguments $\pi_1, \pi_2, \dots$ in (11) are replaced by the forecasts produced by a fixed forecasting system. See, e.g., [13] for details. Game-theoretic supermartingales are referred to as "superfarthingales" in [7].

A supermartingale S is called *forecast-continuous* if, for all $T \in \{1, 2, \dots\}$, all $e_1, \dots, e_T \in E$, all $\pi_1, \dots, \pi_{T-1} \in [0,1]$, and all $\omega_1, \dots, \omega_T \in \{0,1\}$,

$$S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1}, e_T, \pi, \omega_T)$$

is a continuous function of $\pi \in [0,1]$. The following lemma states the most important for us property of forecast-continuous supermartingales.

Lemma 1. *Let S be a forecast-continuous supermartingale. For any T and for any values of the arguments $e_1, \dots, e_T \in E$, $\pi_1, \dots, \pi_{T-1} \in [0,1]$, and $\omega_1, \dots, \omega_{T-1} \in \{0,1\}$, there exists $\pi \in [0,1]$ such that, for both $\omega = 0$ and $\omega = 1$,*

$$\begin{aligned}S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1}, e_T, \pi, \omega) \\ \le S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1}) .\end{aligned}$$

Proof. Define a function $f : [0,1] \times \{0,1\} \to (-\infty, \infty]$ by

$$\begin{aligned}f(\pi, \omega) := S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1}, e_T, \pi, \omega) \\ - S(e_1, \pi_1, \omega_1, \dots, e_{T-1}, \pi_{T-1}, \omega_{T-1})\end{aligned}$$

(the subtrahend is assumed finite: there is nothing to prove when it is infinite). Since S is a forecast-continuous supermartingale, $f(\pi, \omega)$ is continuous in π and

$$\pi f(\pi, 1) + (1 - \pi) f(\pi, 0) \le 0 \quad (12)$$

for all $\pi \in [0,1]$. In particular, $f(0,0) \le 0$ and $f(1,1) \le 0$.

Our goal is to show that for some $\pi \in [0,1]$ we have $f(\pi, 1) \le 0$ and $f(\pi, 0) \le 0$. If $f(0,1) \le 0$, we can take $\pi = 0$. If $f(1,0) \le 0$, we can take $\pi = 1$. Assume that $f(0,1) > 0$ and $f(1,0) > 0$. Then the difference

$$f(\pi) := f(\pi, 1) - f(\pi, 0)$$

is positive for $\pi = 0$ and negative for $\pi = 1$. By the intermediate value theorem, $f(\pi) = 0$ for some $\pi \in (0,1)$. By (12) we have $f(\pi, 1) = f(\pi, 0) \le 0$. □

The fact that the sequence (10) is non-increasing follows from the fact (see below) that Q is a supermartingale (when restricted to the allowed moves for the players). The proof of Lemma 1, as applied to the supermartingale Q, is summarized in (9), the pseudocode for the defensive forecasting algorithm, and the paragraph following it.

The weighted sum of finitely many forecast-continuous supermartingales taken with positive weights is again a forecast-continuous supermartingale. Therefore, the proof will be complete if we check that Q^n is a forecast-continuous supermartingale under the restriction that λ_t^n is η_t^n-mixable for all n and t. But before we can do this, we will need to do some preparatory work in the next subsection.

Geometry of mixability and proper loss functions

Assumption 1 and the compactness of $[0,1]$ imply that the superprediction set (1) is closed. Along with the superprediction set, we will also consider the *prediction set*

$$\Pi_\lambda := \left\{(x,y) \in [0,\infty)^2 \mid \exists \gamma\, \lambda(\gamma,0) = x \text{ and } \lambda(\gamma,1) = y\right\}.$$

In many cases, the prediction set is the boundary of the superprediction set. The prediction set can also be defined as the set of points

$$\Lambda_\gamma := (\lambda(\gamma,0), \lambda(\gamma,1)) \tag{13}$$

where γ ranges over the prediction space $[0,1]$. It is clear that the prediction set is compact.

Let us fix a constant $\eta > 0$. The prediction set of the generalized log loss game (3) is the curve $\{(x,y) \mid \mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1\}$ in $\mathbb{R}^2$. For each $\pi \in (0,1)$, the *π-point* of this curve is Λ_π, i.e., the point

$$\left(-\frac{1}{\eta}\ln(1-\pi), -\frac{1}{\eta}\ln\pi\right).$$

Since the generalized log loss function is proper, the minimum of $(1-\pi)x + \pi y$ on the curve $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ is attained at the π-point; in other words, the tangent of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ at the π-point is orthogonal to the vector $(1-\pi, \pi)$.

A *shift* of the curve $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ is the curve $\mathrm{e}^{-\eta(x-\alpha)} + \mathrm{e}^{-\eta(y-\beta)} = 1$ for some $\alpha, \beta \in \mathbb{R}$ (i.e., it is a parallel translation of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ by some vector (α, β)). The *π-point* of this shift is the point $(\alpha,\beta) + \Lambda_\pi$, where Λ_π is the π-point of the original curve $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$. This provides us with a coordinate system on each shift of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ ($\pi \in (0,1)$ serves as the coordinate of the corresponding π-point).

It will be convenient to use the geographical expressions "Northeast" and "Southwest". A point (x_1, y_1) is *Northeast* of a point (x_2, y_2) if $x_1 \geq x_2$ and $y_1 \geq y_2$. A set $A \subseteq \mathbb{R}^2$ is *Northeast* of a shift of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ if each point of A is Northeast of some point of the shift. Similarly, a point is *Northeast*

of a shift of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ (or of a straight line with a negative slope) if it is Northeast of some point on that shift (or line). "Northeast" is replaced by "Southwest" when the inequalities are $\le$ rather than $\ge$, and we add the attribute "strictly" when the inequalities are strict.

It is easy to see that the loss function is η-mixable if and only if for each point (a, b) on the boundary of the superprediction set there exists a shift of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ passing through (a, b) such that the superprediction set lies to the Northeast of the shift. This follows from the fact that the shifts of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ correspond to the straight lines with negative slope under the homeomorphism E_η: indeed, the preimage of $ax + by = c$, where $a > 0$, $b > 0$, and $c > 0$, is $a\mathrm{e}^{-\eta x} + b\mathrm{e}^{-\eta y} = c$, which is the shift of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ by the vector

$$\left(-\frac{1}{\eta}\ln\frac{a}{c}, -\frac{1}{\eta}\ln\frac{b}{c}\right).$$

A similar statement for the property of being proper is:

Lemma 2. *Suppose the loss function λ is η-mixable. It is a proper loss function if and only if for each π the superprediction set is to the Northeast of the shift of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ passing through Λ_π (as defined by (13)) and having Λ_π as its π-point.*

Proof. The part "if" is obvious, so we will only prove the part "only if". Let λ be η-mixable and proper. Suppose there exists π such the shift A_1 of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ passing through Λ_π and having Λ_π as its π-point has some superpredictions strictly to its Southwest. Let s be such a superprediction, let A_2 be the shift of $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$ passing through Λ_π and s, and let A_3 be the tangent to A_1 at the point Λ_π. Then there are points on A_2 between Λ_π and s that lie strictly to the Southwest of A_3 (take any point on A_2 between Λ_π and s that is sufficiently close to Λ_π). By the η-mixability of λ these points must be superpredictions, which contradicts λ being a proper loss function (since A_3 is the straight line passing through Λ_π and orthogonal to $(1 - \pi, \pi)$). □

Notice that we never assume our loss functions to be strictly proper. (Geometrically, the difference between proper mixable loss functions and strictly proper mixable loss functions is that the former's prediction set is allowed to have corners.)

Proof of the supermartingale property

Let $E \subseteq ([0,1]^N \times (0,\infty)^N \times \mathcal{L}^N)$ consist of sequences

$$\left(\gamma^1, \dots, \gamma^N, \eta^1, \dots, \eta^N, \lambda^1, \dots, \lambda^N\right)$$

such that γ^n is η^n-mixable for all $n = 1, \dots, N$. We will only be interested in the restriction of Q^n and Q on $(E \times [0,1] \times \{0,1\})^*$; these restrictions are denoted with the same symbols.

The following lemma completes the proof of Theorem 2. We will prove it without calculations, unlike the proofs (of different but somewhat similar properties) presented in [4] (and, specifically for the binary case, in [17]).

Lemma 3. *The function Q^n defined on $(E \times [0,1] \times \{0,1\})^*$ by (8) is a supermartingale.*

Proof. It suffices to check that it is always true that

$$\pi_T \exp\left(\eta_T^n \left(\lambda_T^n(\pi_T, 1) - \lambda_T^n(\gamma_T^n, 1)\right)\right) + (1 - \pi_T) \exp\left(\eta_T^n \left(\lambda_T^n(\pi_T, 0) - \lambda_T^n(\gamma_T^n, 0)\right)\right) \le 1.$$

To simplify the notation, we omit the indices n and T; this does not lead to any ambiguity. Using the notation $(a, b) := \Lambda_\pi = (\lambda(\pi, 0), \lambda(\pi, 1))$ and $(x, y) := \Lambda_\gamma = (\lambda(\gamma, 0), \lambda(\gamma, 1))$, we can further simplify the last inequality to

$$(1 - \pi) \exp\left(\eta\left(a - x\right)\right) + \pi \exp\left(\eta\left(b - y\right)\right) \le 1.$$

In other words, it suffices to check that the (super)prediction set lies to the Northeast of the shift

$$\exp\left(-\eta\left(x - a - \frac{1}{\eta}\ln(1-\pi)\right)\right) + \exp\left(-\eta\left(y - b - \frac{1}{\eta}\ln\pi\right)\right) = 1 \qquad (14)$$

of the curve $\mathrm{e}^{-\eta x} + \mathrm{e}^{-\eta y} = 1$. The vector by which (14) is shifted is

$$\left(a + \frac{1}{\eta}\ln(1-\pi), b + \frac{1}{\eta}\ln\pi\right),$$

and so (a, b) is the π-point of that shift. This completes the proof of the lemma: by Lemma 2, the superprediction set indeed lies to the Northeast of that shift. □

A simple special case

In the case where $\lambda_t^n = \lambda$ is the log loss function and $\eta_t^n = 1$ for all n and t, the supermartingale (8) (which is in fact a martingale now) becomes a likelihood ratio process: namely, it becomes the ratio

$$\prod_{t=1}^{T} \frac{\tilde{\gamma}_t^n(\{\omega_t\})}{\tilde{\pi}_t(\{\omega_t\})},$$

where $\tilde{p}$, $p \in [0,1]$, stands for the probability measure on $\{0,1\}$ such that $\tilde{p}(\{1\}) = p$. The mixed martingale Q becomes the likelihood ratio with the Bayes mixture as the numerator, and it is easy to see that in this case defensive forecasting reduces to the Bayes rule.

8 Defensive forecasting for specialist experts and the AA

In this section we will find a more explicit version of defensive forecasting in the case of specialist experts. Our algorithm will achieve a slightly more general version of the bound (7); namely, we will replace the $\ln N$ in (7) by $-\ln p^n$ where p^n is an *a priori* chosen weight for Expert n: all p^n are non-negative and sum to 1. Without loss of generality all p^n will be assumed positive (our algorithm can always be applied to the subset of experts with positive weights). Let A_t be the set of awake experts at time t: $A_t := \{n \in \{1, \dots, N\} \mid \gamma_t^n \neq a\}$.

Let λ be an η-mixable loss function. By the definition of mixability there exists a function $\Sigma(u_1, \dots, u_k, \gamma_1, \dots, \gamma_k)$ (called a *substitution function*) such that:

- the domain of Σ consists of all sequences $(u_1, \dots, u_k, \gamma_1, \dots, \gamma_k)$, for all $k = 0, 1, 2, \dots$, of numbers $u_i \in [0, 1]$ summing to 1, $u_1 + \dots + u_k = 1$, and predictions $\gamma_1, \dots, \gamma_k \in [0, 1]$;
- Σ takes values in the prediction space $[0, 1]$;
- for any $(u_1, \dots, u_k, \gamma_1, \dots, \gamma_k)$ in the domain of Σ, the prediction $\gamma := \Sigma(u_1, \dots, u_k, \gamma_1, \dots, \gamma_k)$ satisfies

$$\forall \omega \in \{0, 1\} : \mathrm{e}^{-\eta\lambda(\gamma,\omega)} \geq \sum_{i=1}^{k} \mathrm{e}^{-\eta\lambda(\gamma_i,\omega)} u_i. \tag{15}$$

Fix such a function Σ. Notice that its value $\Sigma()$ on the empty sequence can be chosen arbitrarily, that the case $k = 1$ is trivial, and that the case $k = 2$ in fact covers the cases $k = 3$, $k = 4$, etc.

Defensive forecasting algorithm for specialist experts

$w_0^n := p^n$, $n = 1, \dots, N$.
FOR $t = 1, 2, \dots$:
  Read the list A_t of awake experts
    and their predictions $\gamma_t^n \in [0, 1]$, $n \in A_t$.
  Predict $\pi_t := \Sigma\left(\left(u_{t-1}^n\right)_{n \in A_t}, \left(\gamma_t^n\right)_{n \in A_t}\right)$,
    where $u_{t-1}^n := w_{t-1}^n / \sum_{n \in A_t} w_{t-1}^n$.
  Read the outcome $\omega_t \in \{0, 1\}$.
  Set $w_t^n := w_{t-1}^n \mathrm{e}^{\eta(\lambda(\pi_t, \omega_t) - \lambda(\gamma_t^n, \omega_t))}$ for all $n \in A_t$.
END FOR

This algorithm is a simple modification of the AA, and it becomes the AA when the experts are always awake. In the case of the log loss function, this algorithm was found by Freund et al. [8]; in this special case, Freund et al. derive the same performance guarantee as we do.

Derivation of the algorithm

In this derivation we will need the following notation. For each history of the game, let A^n, $n \in \{1, \dots, N\}$, be the set of steps at which Expert n is awake:

$$A^n := \{t \in \{1, 2, \dots\} \mid n \in A_t\}.$$

For each positive integer k, $[k]$ stands for the set $\{1, \dots, k\}$.

The method of defensive forecasting requires (cf. Corollary 4) that at step T we should choose $\pi = \pi_T$ such that, for each $\omega \in \{0, 1\}$,

$$\begin{aligned}\sum_{n \in A_T} p^n \mathrm{e}^{\eta(\lambda(\pi,\omega)-\lambda(\gamma_T^n,\omega))} \prod_{t \in [T-1] \cap A^n} \mathrm{e}^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^n,\omega_t))} \\ + \sum_{n \in A_T^c} p^n \prod_{t \in [T-1] \cap A^n} \mathrm{e}^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^n,\omega_t))} \\ \le \sum_{n \in [N]} p^n \prod_{t \in [T-1] \cap A^n} \mathrm{e}^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^n,\omega_t))}\end{aligned}$$

where A_T^c stands for the complement of A_T in $[N]$: $A_T := [N] \setminus A_T$. This inequality is equivalent to

$$\begin{aligned}\sum_{n \in A_T} p^n \mathrm{e}^{\eta(\lambda(\pi,\omega)-\lambda(\gamma_T^n,\omega))} \prod_{t \in [T-1] \cap A^n} \mathrm{e}^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^n,\omega_t))} \\ \le \sum_{n \in A_T} p^n \prod_{t \in [T-1] \cap A^n} \mathrm{e}^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^n,\omega_t))}\end{aligned}$$

and can be rewritten as

$$\sum_{n \in A_T} \mathrm{e}^{\eta(\lambda(\pi,\omega)-\lambda(\gamma_T^n,\omega))} u_{T-1}^n \le 1, \tag{16}$$

where $u_{T-1}^n := w_{T-1}^n / \sum_{n \in A_T} w_{T-1}^n$ are the normalized weights

$$w_{T-1}^n := p^n \prod_{t \in [T-1] \cap A^n} \mathrm{e}^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^n,\omega_t))}.$$

Comparing (16) and (15), we can see that it suffices to set

$$\pi := \Sigma\left(\left(u_{T-1}^n\right)_{n \in A_T}, \left(\gamma_T^n\right)_{n \in A_T}\right).$$

Discussion of the algorithm

The main difference of the algorithm of the previous subsection from the AA is in the way the experts' weights are updated. The weights of the sleeping experts are not changed, whereas the weights of the awake experts are multiplied by $\mathrm{e}^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^n,\omega_t))}$. Therefore, Learner's loss serves as the benchmark: the weight of an awake expert who performs better than Learner goes up, the weight of an awake expert who performs worse than Learner goes down, and the weight of a sleeping expert does not change.

Acknowledgements

We are grateful to the anonymous Eurocrat who coined the term "expert evaluator". This work was supported in part by EPSRC grant EP/F002998/1.

References

[1] Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.

[2] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: structure and applications. Manuscript. Available on-line at `http://www-stat.wharton.upenn.edu/~buja/` (accessed on 19 February 2009), 2005.

[3] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, England, 2006.

[4] Alexey Chernov, Yuri Kalnishkan, Fedor Zhdanov, and Vladimir Vovk. Supermartingales in prediction with expert advice. In Yoav Freund, László Györfi, György Turán, and Thomas Zeugmann, editors, *Proceedings of the Nineteenth International Conference on Algorithmic Learning Theory*, volume 5254 of *Lecture Notes in Artificial Intelligence*, pages 199–213, Berlin, 2008. Springer.

[5] William W. Cohen and Yoram Singer. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*, 17:141–173, 1999.

[6] A. Philip Dawid. Probability forecasting. In Samuel Kotz, Norman L. Johnson, and Campbell B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley, New York, 1986.

[7] A. Philip Dawid and Vladimir Vovk. Prequential probability: principles and properties. *Bernoulli*, 5:125–162, 1999.

[8] Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the Twenty Ninth Annual ACM Symposium on Theory of Computing*, pages 334–343, New York, 1997. Association for Computing Machinery.

[9] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[10] David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44:1906–1925, 1998.

[11] Robert D. Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In Rocco A. Servedio and Tong Zhang, editors, *Proceedings of the Twenty First Annual Conference on Learning Theory*, pages 425–436. Omnipress, 2008.

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

[13] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[14] Vladimir Vovk. Aggregating strategies. In Mark Fulk and John Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

[15] Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56:153–173, 1998.

[16] Vladimir Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35:247–282, 1999.

[17] Vladimir Vovk. Defensive forecasting for optimal prediction with expert advice. Technical Report `arXiv:0708.1503` [cs.LG], `arXiv.org` e-Print archive, August 2007.