Defensive forecasting for optimal prediction with expert advice

Vladimir Vovk
vovk@cs.rhul.ac.uk
http://vovk.net

February 1, 2008

Abstract

The method of defensive forecasting is applied to the problem of prediction with expert advice for binary outcomes. It turns out that defensive forecasting is not only competitive with the Aggregating Algorithm but also handles the case of "second-guessing" experts, whose advice depends on the learner's prediction; this paper assumes that the dependence on the learner's prediction is continuous.

1 Introduction

There are many known techniques in competitive on-line prediction, such as following the perturbed leader (see, e.g., [8, 13, 12]), Bayes-type aggregation (see, e.g., [17, 20, 6]) and the closely related potential methods, gradient descent (see, e.g., [2]) and closely related exponentiated gradient descent [14], and the recently developed technique of defensive forecasting (see, e.g., [27, 24]). Defensive forecasting combines the ideas of game-theoretic probability (see, e.g., [18]) with Levin and Gács's ideas of neutral measure [16, 7] and Foster and Vohra's ideas of universal calibration [5]. See [3] for a general review of competitive on-line prediction.

This paper applies the technique of defensive forecasting to prediction with expert advice in the simple case of binary outcomes The learner's goal in prediction with expert advice is to compete with free agents, called experts, who are allowed to choose any predictions at each step. We will be interested in performance guarantees of the type

$$L_N \le \min_{k=1,\dots,K} L_N^k + a_K \tag{1}$$

where K is the number of experts, a_K is a constant depending on K, L_N is the learner's cumulative loss over the first N steps, and L_N^k is the kth expert's cumulative loss over the first N steps (see §§3–5 for precise definitions).

It has been shown by Watkins ([22], Theorem 8) that the Aggregating Algorithm (implementing Bayes-type aggregation for general loss functions [20, 21], the AA for short) delivers the optimal value of the constant a_K in (1) whenever the goal (1) can be achieved. (Watkins's result was based on earlier results by Haussler, Kivinen, and Warmuth [10], Theorem 3.1, and Vovk [21], Theorem 1, establishing the optimality of the AA for a large number of experts.) Theorem 3 of this paper asserts that, perhaps surprisingly, defensive forecasting also achieves the same performance guarantee.

Whether the goal (1) is achievable depends on the loss function used for evaluating the learner's and experts' performance. The necessary and sufficient condition is that the loss function should "perfectly mixable" (see §5 for a definition). For simplicity, we first consider two specific, perhaps most important, examples of perfectly mixable loss functions: the quadratic loss function in §3 and the log loss function in §4. Those two sections are self-contained in that they do not require familiarity with the AA. In the last section, §5, we establish the general result, for arbitrary perfectly mixable loss functions. In an appendix we state Watkins's theorem in the form needed in this paper.

It is interesting that the technique of defensive forecasting is also applicable to experts who are allowed to "second-guess" the learner: their recommendations can depend (in a continuous manner in this paper) on the learner's prediction. It is not clear that second-guessing experts can be handled at all by the AA.

A result similar to this paper's results is proved by Stoltz and Lugosi in [19], Theorem 14 (a more detailed comparison will be given in [25]). Second-guessing experts are useful in game theory (where competing with second-guessing experts is known as prediction with a small internal regret). For a more down-to-earth example of a useful second-guessing expert, remember that humans tend to give too categorical (i.e., close to 0 or 1) predictions; therefore, a useful second-guessing expert for a human learner would transform his/her predictions to less categorical ones (according to the learner's expected calibration curve [4]).

2 Defensive forecasting

Let E be a topological space ($E = [0,1]^K$ in the application to prediction with expert advice in §§3–5).

THE BINARY FORECASTING PROTOCOL

$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \dots$:
    Expert announces continuous $\gamma_n : [0,1] \to E$.
    Forecaster announces $p_n \in [0,1]$.
    Reality announces $\omega_n \in \{0,1\}$.
END FOR.

A *process* is any function $S : (E \times [0,1] \times \{0,1\})^* \to \mathbb{R}$. Given the sequence of the players' moves in the binary forecasting protocol, we sometimes write S_N, $N \in \{0,1,\ldots\}$, for $S(\gamma_1(p_1), p_1, \omega_1, \ldots, \gamma_N(p_N), p_N, \omega_N)$. (Notice that S_N depend on γ_n only via $\gamma_n(p_n)$.) We also sometimes interpret S_N as function of the players' moves in the protocol and identify the process S with the sequence of functions S_N, $N = 0, 1, \ldots$, on the set of all histories $(\gamma_1, p_1, \omega_1, \gamma_2, p_2, \omega_2, \ldots)$.

A process S is said to be a *supermartingale* if it is always true that

$$\begin{aligned} p_N S(g_1, p_1, \omega_1, \ldots, g_{N-1}, p_{N-1}, \omega_{N-1}, g_N, p_N, 1) \\ + (1 - p_N) S(g_1, p_1, \omega_1, \ldots, g_{N-1}, p_{N-1}, \omega_{N-1}, g_N, p_N, 0) \\ \le S(g_1, p_1, \omega_1, \ldots, g_{N-1}, p_{N-1}, \omega_{N-1}) \end{aligned} \quad (2)$$

(i.e., it is true for all N, all $g_1, \ldots, g_N$ in E, all $p_1, \ldots, p_N$ in $[0,1]$, and all $\omega_1, \ldots, \omega_{N-1}$ in $\{0,1\}$). In the traditional theory of martingales (when translated into our framework), Expert's move is an element of E (in other words, a constant function), and this would be sufficient for application to the traditional problem of prediction with expert advice; however, the version with second-guessing experts requires the generalization to $\gamma_n : [0,1] \to E$. We say that a supermartingale S is *forecast-continuous* if, for each N, $S(g_1, p_1, \omega_1, \ldots, g_N, p_N, \omega_N)$ is a continuous function of $p_N \in [0,1]$ and $g_N \in E$.

Lemma 1 (Levin, Takemura) *For any forecast-continuous supermartingale S there exists a strategy for Forecaster ensuring that $S_0 \ge S_1 \ge \cdots$ regardless of the other players' moves.*

Proof Set, for $p \in [0,1]$ and $\omega \in \{0,1\}$,

$$\begin{aligned} t(\omega, p) := S(\gamma_1(p_1), p_1, \omega_1, \ldots, \gamma_{N-1}(p_{N-1}), p_{N-1}, \omega_{N-1}, \gamma_N(p), p, \omega) \\ - S(\gamma_1(p_1), p_1, \omega_1, \ldots, \gamma_{N-1}(p_{N-1}), p_{N-1}, \omega_{N-1}). \end{aligned}$$

Our goal is to prove the existence of p such that $t(\omega, p) \le 0$ for both $\omega = 0$ and $\omega = 1$. I will give an argument (from [24], the proof of Lemma 1) that is applicable very generally.

For all $p, q \in [0,1]$ set

$$\phi(q, p) := q t(1, p) + (1 - q) t(0, p).$$

The function $\phi(q, p)$ is linear in its first argument, q, and continuous in its second argument, p. Ky Fan's minimax theorem (see, e.g., [1], Theorem 11.4) shows that there exists $p^* \in [0,1]$ such that

$$\forall q \in [0,1] : \quad \phi(q, p^*) \le \sup_{p \in [0,1]} \phi(p, p).$$

Therefore,

$$\forall q \in [0,1] : \quad q t(1, p^*) + (1 - q) t(0, p^*) \le 0,$$

and we can see that $t(\omega, p^*)$ never exceeds 0. ∎

For generalizations (due to Levin and Takemura) of Lemma 1 in different directions, see, e.g., [26] (Theorem 1) and [24] (Lemma 1). By defensive forecasting we mean using such results in prediction with expert advice.

3 Algorithm competitive with continuous second-guessers: quadratic loss function

This is the version of the standard protocol of prediction with expert advice under quadratic loss for continuous second-guessing experts:

PREDICTION WITH EXPERT ADVICE UNDER QUADRATIC LOSS

$L_0 := 0$.
$L_0^k := 0$, $k = 1, \dots, K$.
FOR $n = 1, 2, \dots$:
  Expert k announces continuous $\gamma_n^k : [0,1] \to [0,1]$, $k = 1, \dots, K$.
  Learner announces $p_n \in [0,1]$.
  Reality announces $\omega_n \in \{0,1\}$.
  $L_n := L_{n-1} + (p_n - \omega)^2$.
  $L_n^k := L_{n-1}^k + (\gamma_n^k(p_n) - \omega_n)^2$.
END FOR.

To apply Lemma 1 to the problem of prediction with expert advice under quadratic loss, we will need the following result.

Lemma 2 *Suppose $E = [0,1]$ and $\kappa \in [0,2]$. The process*

$$S_N := \exp\left(\kappa \sum_{n=1}^{N} \left((p_n - \omega_n)^2 - (\gamma_n(p_n) - \omega_n)^2\right)\right)$$

is a supermartingale in the binary forecasting protocol.

Proof By (2), it suffices to check that

$$p \exp\left(\kappa\left((p-1)^2 - (g-1)^2\right)\right) + (1-p)\exp\left(\kappa\left((p-0)^2 - (g-0)^2\right)\right) \le 1$$

for all $p, g \in [0,1]$. If we substitute $g = p + x$, the last inequality will reduce to

$$pe^{2\kappa(1-p)x} + (1-p)e^{-2\kappa p x} \le e^{\kappa x^2}, \quad \forall x \in [-p, 1-p].$$

The last inequality is a simple corollary of Hoeffding's inequality ([11], (4.16), which is true for any $h \in \mathbb{R}$: cf. [3], Lemma A.1). Indeed, applying Hoeffding's inequality to the random variable

$$X := \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p, \end{cases}$$

we obtain

$$pe^{h(1-p)} + (1-p)e^{-hp} \le e^{h^2/8},$$

which the substitution $h := 2\kappa x$ reduces to

$$pe^{2\kappa(1-p)x} + (1-p)e^{-2\kappa px} \le e^{\kappa^2 x^2/2} \le e^{\kappa x^2},$$

the last inequality assuming $\kappa \le 2$. ∎

Lemma 2 immediately implies a performance guarantee for the method of defensive forecasting.

Theorem 1 *There exists a strategy for Learner in the quadratic-loss protocol with K experts that guarantees*

$$L_N \le L_N^k + \frac{\ln K}{2} \tag{3}$$

for all $N = 1, 2, \ldots$ and all $k \in \{1, \ldots, K\}$.

Proof Consider the binary forecasting protocol with $E = [0,1]^K$. By Lemma 2, the process

$$\sum_{k=1}^{K} \exp\left(\kappa \sum_{n=1}^{N} \left((p_n - \omega_n)^2 - \left(\gamma_n^k(p_n) - \omega_n\right)^2\right)\right)$$

is a supermartingale. By Lemma 1, Learner has a strategy that prevents this supermartingale from growing. This strategy ensures

$$\sum_{k=1}^{K} \exp\left(\kappa \sum_{n=1}^{N} \left((p_n - \omega_n)^2 - \left(\gamma_n^k(p_n) - \omega_n\right)^2\right)\right) \le K,$$

which implies, for all $k \in \{1, \ldots, K\}$,

$$\exp\left(\kappa \sum_{n=1}^{N} \left((p_n - \omega_n)^2 - \left(\gamma_n^k(p_n) - \omega_n\right)^2\right)\right) \le K,$$

i.e., (3) in the case $\kappa = 2$. ∎

For the proof of (3) being the performance guarantee for the AA, see, e.g., [20], Example 4, or [23], §2.4. It is interesting that even such an apparently minor deviation from the AA as replacing the AA-type averaging of the experts' predictions by the arithmetic mean (with the same exponential weighting scheme) leads to a suboptimal result: the constant 2 in (3) is replaced by 1/2 ([15], reproduced in [23], Remark 3).

4 Algorithm competitive with continuous second-guessers: log loss function

The log loss function is defined by

$$\lambda(\omega, p) := \begin{cases} -\ln p & \text{if } \omega = 1 \\ -\ln(1-p) & \text{if } \omega = 0, \end{cases}$$

where $\omega \in \{0, 1\}$ and $p \in [0, 1]$; notice that the loss function is now allowed to take value ∞. The protocol of prediction with expert advice becomes:

PREDICTION WITH EXPERT ADVICE UNDER LOG LOSS

$L_0 := 0$.
$L_0^k := 0$, $k = 1, \dots, K$.
FOR $n = 1, 2, \dots$:
    Expert k announces continuous $\gamma_n^k : [0, 1] \to [0, 1]$, $k = 1, \dots, K$.
    Learner announces $p_n \in [0, 1]$.
    Reality announces $\omega_n \in \{0, 1\}$.
    $L_n := L_{n-1} + \lambda(\omega_n, p_n)$.
    $L_n^k := L_{n-1}^k + \lambda(\omega_n, \gamma_n^k(p_n))$.
END FOR.

This is the analogue of Lemma 2 for the log loss function:

Lemma 3 *Suppose $E = [0, 1]$ and $\kappa \in [0, 1]$. The process*

$$S_N := \exp\left(\kappa \sum_{n=1}^{N} \Big(\lambda(\omega_n, p_n) - \lambda(\omega_n, \gamma_n(p_n))\Big)\right)$$

is a supermartingale in the binary forecasting protocol.

Proof It suffices to check that

$$p \exp(\kappa(-\ln p + \ln g)) + (1-p) \exp(\kappa(-\ln(1-p) + \ln(1-g))) \le 1,$$

i.e., that

$$p^{1-\kappa} g^\kappa + (1-p)^{1-\kappa} (1-g)^\kappa \le 1,$$

for all $p, g \in [0, 1]$. The last inequality immediately follows from the inequality between the geometric and arithmetic means when $\kappa \in [0, 1]$. (The left-hand side of that inequality is a special case of what is known as the Hellinger integral in probability theory.) ∎

Lemma 3 implies a performance guarantee for the log loss function as in the previous section.

Theorem 2 *There exists a strategy for Learner in the log loss protocol with K experts that guarantees*

$$L_N \le L_N^k + \ln K \tag{4}$$

for all $N = 1, 2, \ldots$ and all $k \in \{1, \ldots, K\}$.

Proof Take $\kappa := 1$. Lemma 3 guarantees that the process

$$\sum_{k=1}^{K} \exp\left(\kappa \sum_{n=1}^{N} \left(\lambda\left(\omega_n, p_n\right) - \lambda\left(\omega_n, \gamma_n^k(p_n)\right)\right)\right)$$

is a supermartingale in the binary forecasting protocol with $E = [0,1]^K$. Any strategy for Learner that prevents this supermartingale from growing ensures (4) for all $k \in \{1, \ldots, K\}$. ∎

For the proof of (4) being the performance guarantee for the AA, see, e.g., [20], Example 3.

5 Algorithm competitive with continuous second-guessers: perfectly mixable loss functions

In this section we assume that Learner chooses his predictions from a non-empty *decision space* Γ and that his performance is evaluated using a *loss function* $\lambda : \{0,1\} \times \Gamma \to \mathbb{R}$. The triple $(\{0,1\}, \Gamma, \lambda)$ will sometimes be called our *game of prediction* (the first element, the outcome space $\{0,1\}$, is redundant at this time). The loss function will be assumed bounded below; there is no further loss of generality in assuming that it is non-negative.

As mentioned in §1, to have a chance of achieving (1), the loss function has to be assumed "perfectly mixable" (this will be further discussed in the appendix); we start from defining this property.

A point (x, y) of the plane $\mathbb{R}^2$ is called a *superprediction* (with respect to the loss function λ) if there exists a decision $\gamma \in \Gamma$ such that

$$\lambda(0, \gamma) \le x \quad \& \quad \lambda(1, \gamma) \le y.$$

Our next assumption about the game of prediction will be that the superprediction set is closed.

Let η be a positive constant (the *learning rate* used). A *shift* of the curve $\{(x, y) \mid e^{-\eta x} + e^{-\eta y} = 1\}$ in $\mathbb{R}^2$ is the curve $\{(x, y) \mid e^{-\eta(x+\alpha)} + e^{-\eta(y+\beta)} = 1\}$ for some $\alpha, \beta \in \mathbb{R}$ (i.e., it is a parallel translation of $e^{-\eta x} + e^{-\eta y} = 1$ in any direction and by any distance). The loss function is called *η-mixable* if for each point (a, b) on the boundary of the superprediction set there exists a shift of $e^{-\eta x} + e^{-\eta y} = 1$ passing through (a, b) such that the superprediction set lies completely to one side of the shift (it is clear that in this case the superprediction set must lie to the Northeast of the shift). The loss function is *perfectly mixable* if it is η mixable for some $\eta > 0$.

Suppose λ is η-mixable, $\eta > 0$. Each decision $\gamma \in \Gamma$ can be represented by the point $(\lambda(0,\gamma), \lambda(1,\gamma))$ in the superprediction set. The set of all $(\lambda(0,\gamma), \lambda(1,\gamma))$, $\gamma \in \Gamma$, will be called the *prediction set*; for typical games this set coincides with the boundary of the superprediction set. As far as the attainable performance guarantees are concerned (before we start paying attention to computational issues), the only interesting part of the game of prediction is its prediction set; the game itself can be regarded as an arbitrary coordinate system in the prediction set. It will be convenient to introduce another coordinate system in essentially the same set.

For each $p \in [0,1]$, let (a_p, b_p) be the point (x,y) in the superprediction set at which the minimum of $py + (1-p)x$ is attained. Since λ is η-mixable, the point (a_p, b_p) is determined uniquely; it is clear that the dependence of (a_p, b_p) on p is continuous.

We can now redefine the decision space and the loss function as follows: the decision space becomes $[0,1]$ and the loss function becomes

$$\lambda(0,p) := a_p, \quad \lambda(1,p) := b_p.$$

The resulting game of prediction is essentially the same as the original game (one of the minor differences is that, if the superprediction set has "corners", a decision $\gamma \in \Gamma$ maybe split into several decisions $p \in [0,1]$ in the new game, all leading to the same losses). In the rest of this section, let us assume that the game of prediction has been transformed to this *standard* form. Notice that the new loss function is a "proper scoring rule" (see, e.g., [4]).

The protocol of this section formally coincides with that of the previous section (although λ ranges over a much wider class of loss functions):

PREDICTION WITH EXPERT ADVICE IN A STANDARD PERFECTLY MIXABLE GAME

$L_0 := 0$.
$L_0^k := 0$, $k = 1, \dots, K$.
FOR $n = 1, 2, \dots$:
  Expert k announces continuous $\gamma_n^k : [0,1] \to [0,1]$, $k = 1, \dots, K$.
  Learner announces $p_n \in [0,1]$.
  Reality announces $\omega_n \in \{0,1\}$.
  $L_n := L_{n-1} + \lambda(\omega_n, p_n)$.
  $L_n^k := L_{n-1}^k + \lambda(\omega_n, \gamma_n^k(p_n))$.
END FOR.

Lemmas 2 and 3 carry over to the perfectly mixable loss functions:

Lemma 4 *Let $\eta > 0$, $(\{0,1\}, [0,1], \lambda)$ be a standard η-mixable game of prediction, $E = [0,1]$, and $\kappa \in [0, \eta]$. The process*

$$S_N := \exp\left(\kappa \sum_{n=1}^{N} \Big(\lambda\left(\omega_n, p_n\right) - \lambda\left(\omega_n, \gamma_n(p_n)\right)\Big)\right)$$

is a supermartingale in the binary forecasting protocol.

Proof It suffices to check that

$$p\exp\left(\kappa\left(\lambda(1,p)-\lambda(1,g)\right)\right)+(1-p)\exp\left(\kappa\left(\lambda(0,p)-\lambda(0,g)\right)\right)\le 1$$

for all $p, g \in [0,1]$. As λ is η-mixable, it will also be κ-mixable; we will only be using the latter property. Using the notation $(a,b) := (a_p, b_p) = (\lambda(0,p), \lambda(1,p))$ and $(a',b') := (a_g, b_g) = (\lambda(0,g), \lambda(1,g))$, we can slightly simplify this inequality:

$$p\exp\left(\kappa\left(b-b'\right)\right)+(1-p)\exp\left(\kappa\left(a-a'\right)\right)\le 1. \tag{5}$$

It is clear that the superprediction set lies to the Northeast of the shift $e^{-\kappa(x+\alpha)}+e^{-\kappa(y+\beta)}=1$ of $e^{-\kappa x}+e^{-\kappa y}=1$ that passes through (a,b),

$$e^{-\kappa(a+\alpha)}+e^{-\kappa(b+\beta)}=1, \tag{6}$$

and has the tangent at (a,b) orthogonal to $(1-p,p)$,

$$\left(-\kappa e^{-\kappa(x+\alpha)}, -\kappa e^{-\kappa(y+\beta)}\right)_{x:=a,y:=b} \propto (1-p,p) \tag{7}$$

(the expression on the left-hand side is the gradient of $e^{-\kappa(x+\alpha)}+e^{-\kappa(y+\beta)}$ at (a,b)). We can see from (6) and (7) that

$$e^{-\kappa(a+\alpha)}=1-p, \quad e^{-\kappa(b+\beta)}=p.$$

Substituting these values for p and $1-p$ in (5), we transform (5) to

$$e^{-\kappa(b'+\beta)}+e^{-\kappa(a'+\alpha)}\le 1,$$

which is true: the last inequality just says that (a',b') is Northeast of the shift. ∎

Theorem 3 *Let $\eta > 0$ and consider any standard η-mixable game of prediction $(\{0,1\},[0,1],\lambda)$. There exists a strategy for Learner in the prediction protocol with K experts that guarantees*

$$L_N \le L_N^k + \frac{\ln K}{\eta} \tag{8}$$

for all $N = 1,2,\ldots$ and all $k \in \{1,\ldots,K\}$.

Proof Take $\kappa := \eta$ and proceed as in the proof of Theorem 2 (using Lemma 4 instead of Lemma 3). ∎

Inequality (8) as the performance guarantee for the AA is derived in [20], Theorem 1.

Acknowledgments

This work was partially supported by EPSRC (grant EP/F002998/1), MRC (grant G0301107), and the Cyprus Research Promotion Foundation.

References

[1] Ravi P. Agarwal, Maria Meehan, and Donal O'Regan. *Fixed Point Theory and Applications*, volume 141 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, England, 2001.

[2] 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 Transactions on Neural Networks*, 7:604–619, 1996.

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

[4] 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.

[5] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

[6] 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.

[7] Peter Gács. Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science*, 341:91–137, 2005.

[8] James F. Hannan. Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contribution to the Theory of Games, III*, volume 39 of *Annals of Mathematics Studies*, pages 97–139. Princeton University Press, 1957.

[9] G. H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, England, second edition, 1952.

[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] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[12] Marcus Hutter and Jan Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005.

[13] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

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

[15] Jyrki Kivinen and Manfred K. Warmuth. Averaging expert predictions. In Paul Fischer and Hans U. Simon, editors, *Proceedings of the Fourth European Conference on Computational Learning Theory*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 153–167, Berlin, 1999. Springer.

[16] Leonid A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17:337–340, 1976.

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

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

[19] Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59:187–209, 2007.

[20] 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.

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

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

[23] Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[24] Vladimir Vovk. Predictions as statements and decisions. Technical Report `arXiv:cs.LG/0606093`, `arXiv.org` e-Print archive, June 2006.

[25] Vladimir Vovk. Prediction with second-guessers' advice. In preparation, 2007.

[26] Vladimir Vovk, Ilia Nouretdinov, Akimichi Takemura, and Glenn Shafer. Defensive forecasting for linear protocols. Technical Report `arXiv:cs.LG/0506007` (version 2), `arXiv.org` e-Print archive, September 2005.

[27] Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. Technical Report `arXiv:cs.LG/0505083`, `arXiv.org` e-Print archive, May 2005.

Appendix: Watkins's theorem

Watkins's theorem is stated in [22] (Theorem 8) not in sufficient generality: it presupposes that the loss function is perfectly mixable. The proof, however, shows that this assumption is irrelevant (it can be made part of the conclusion), and the goal of this appendix is to give a self-contained statement of a suitable version of the theorem.

By a *game of prediction* we now mean a triple $(\Omega, \Gamma, \lambda)$, where Ω and Γ are sets called the outcome and decision space, respectively, and $\lambda : \Omega \times \Gamma \to \overline{\mathbb{R}}$ is called the loss function ($\overline{\mathbb{R}}$ is the extended real line $\mathbb{R} \cup \{-\infty, \infty\}$ with the standard topology, although the value $-\infty$ will be later disallowed).

Partly following [21], for each $K = 1, 2, \ldots$ and each $a > 0$ we consider the following perfect-information game $\mathcal{G}_K(a)$ (the "global game") between two players, Learner and Environment.

GLOBAL GAME $\mathcal{G}_K(a)$

$L_0 := 0$.
$L_0^k := 0$, $k = 1, \ldots, K$.
FOR $n = 1, 2, \ldots$:
    Environment chooses $\gamma_n^k \in \Gamma$, $k = 1, \ldots, K$.
    Learner chooses $\gamma_n \in \Gamma$.
    Environment chooses $\omega_n \in \Omega$.
    $L_n := L_{n-1} + \lambda(\omega_n, \gamma_n)$.
    $L_n^k := L_{n-1}^k + \lambda(\omega_n, \gamma_n^k)$, $k = 1, \ldots, K$.
END FOR.

Learner wins if, for all $N = 1, 2, \ldots$ and all $k \in \{1, \ldots, K\}$,

$$L_N \le L_N^k + a; \tag{9}$$

otherwise, Environment wins.

It is possible that $L_N = \infty$ or $L_N^k = \infty$ in (9); the interpretation of inequalities involving infinities is natural.

For each K we will be interested in the set of those $a > 0$ for which Learner has a winning strategy in the game $\mathcal{G}_K(a)$ (we will denote this by $\mathrm{L} \smile \mathcal{G}_K(a)$). It is obvious that

$$\mathrm{L} \smile \mathcal{G}_K(a) \;\&\; a' > a \Longrightarrow \mathrm{L} \smile \mathcal{G}_K(a');$$

therefore, for each K there exists a unique *borderline value* a_K such that $\mathrm{L} \smile \mathcal{G}_K(a)$ holds when $a > a_K$ and fails when $a < a_K$. It is possible that $a_K = \infty$ (but remember that we are only interested in finite values of a).

These are our assumptions about the game of prediction (similar to those in [21]):

- Γ is a compact topological space;

- for each $\omega \in \Omega$, the function $\gamma \in \Gamma \mapsto \lambda(\omega, \gamma)$ is continuous;
- there exists $\gamma \in \Gamma$ such that, for all $\omega \in \Omega$, $\lambda(\omega, \gamma) < \infty$;
- the function λ is bounded below.

We say that the game of prediction $(\Omega, \Gamma, \lambda)$ is *η-mixable*, where $\eta > 0$, if

$$\forall \gamma_1 \in \Gamma, \gamma_2 \in \Gamma, \alpha \in [0,1] \; \exists \delta \in \Gamma \; \forall \omega \in \Omega: \\ e^{-\eta\lambda(\omega,\delta)} \geq \alpha e^{-\eta\lambda(\omega,\gamma_1)} + (1-\alpha) e^{-\eta\lambda(\omega,\gamma_2)}. \quad (10)$$

In the binary case, $\Omega = \{0, 1\}$, this condition says that the image of the super-prediction set under the mapping $(x, y) \mapsto (e^{-\eta x}, e^{-\eta y})$ is convex, and it is easy to see that it is equivalent to the definition used in §5.

It follows from [9] (Theorem 92, applied to the means $\mathfrak{M}_\phi$ with $\phi(x) = e^{-\eta x}$) that if the prediction game is η-mixable it will remain η'-mixable for any positive $\eta' < \eta$. (For another proof, see the end of the proof of Lemma 9 in [21].) Let η^* be the supremum of the η for which the prediction game is η-mixable (with $\eta^* := 0$ when the game is not perfectly mixable). The compactness of Γ implies that the prediction game is η^*-mixable.

Theorem 4 (Chris Watkins) *For any $K \in \{1, 2, \ldots\}$,*

$$a_K = \frac{\ln K}{\eta^*}.$$

In particular, $a_K < \infty$ if and only if the game is perfectly mixable.

It is easy to see that $\mathrm{L} \backsim \mathcal{G}_K(a_K)$: this follows both from general considerations (cf. Lemma 3 in [21]) and from the fact that the AA and this paper's algorithm based on defensive forecasting (the latter assuming $\Omega = \{0, 1\}$) win $\mathcal{G}_K(a_K) = \mathcal{G}_K(\ln K/\eta^*)$.

Proof of Theorem 4 The proof will use Theorem 1 of [21]. Without loss of generality we can, and will, assume $\lambda > 1$ (add a suitable constant to λ if needed); therefore, Assumption 4 of [21] (the only assumption in [21] not directly made in this paper) is satisfied. In view of the fact that $\mathrm{L} \backsim \mathcal{G}_K(\ln K/\eta^*)$, we only need to show that $\mathrm{L} \backsim \mathcal{G}_K(a)$ does not hold for $a < \ln K/\eta^*$. Fix $a < \ln K/\eta^*$.

Since the two-fold convex mixture in (10) can be replaced by any finite convex mixture (apply two-fold mixtures repeatedly), the point $(1, 1/\eta^*)$ belongs to the separation curve (set $\beta := e^{-\eta^*}$ in the definition of $c(\beta)$) whereas the point $(1, a/\ln K)$ is Southwest and outside of the separation curve (use Lemmas 8–12 of [21]). Therefore, E (=Environment) has a winning strategy in the game $\mathcal{G}(1, a/\ln K)$, as defined in [21]. It is easy to see from the proof of Theorem 1 in [21] that the definition of the game $\mathcal{G}$ in [21] can be modified, without changing the conclusion about $\mathcal{G}(1, a/\ln K)$, by replacing the line

E chooses $n \geq 1$ {size of the pool}

in the protocol on p. 153 of [21] by

> E chooses $n^* \geq 1$ {lower bound on the size of the pool}
> L chooses $n \geq n^*$ {size of the pool}

(indeed, the proof in §6 of [21] only requires that there should be sufficiently many experts). Let n^* be the first move by Environment according to her winning strategy.

Now suppose L $\smile \mathcal{G}_K(a)$. From the fact that there exists Learner's strategy $\mathcal{L}_1$ winning $\mathcal{G}_K(a)$ we can deduce: there exists Learner's strategy $\mathcal{L}_2$ winning $\mathcal{G}_{K^2}(2a)$ (we can split the K^2 experts into K groups of K, merge the experts' decisions in every group with $\mathcal{L}_1$, and finally merge the groups' decisions with $\mathcal{L}_1$); there exists Learner's strategy $\mathcal{L}_3$ winning $\mathcal{G}_{K^3}(3a)$ (we can split the K^3 experts into K groups of K^2, merge the experts' decisions in every group with $\mathcal{L}_2$, and finally merge the groups' decisions with $\mathcal{L}_1$); and so on. When the number K^m of experts exceeds n^*, we obtain a contradiction: Learner can guarantee

$$L_N \leq L_N^k + ma$$

for all N and all K^m experts k, and Environment can guarantee that

$$L_N > L_N^k + \frac{a}{\ln K} \ln(K^m) = L_N^k + ma$$

for some N and k. ∎